

TVBENCH: ASSESSING VISUAL TEMPORAL REASONING CAPABILITIES IN MULTIMODAL FOUNDATION MODELS

Anonymous authors

Paper under double-blind review

ABSTRACT

Existing benchmarks often highlight the remarkable performance achieved by state-of-the-art Multimodal Foundation Models (MFMs) in leveraging temporal context for video understanding. However, *how well do the models truly perform visual temporal reasoning?* Our study of existing benchmarks shows that this capability of MFMs is likely overestimated as many questions can be solved by using a single, few, or out-of-order frames. To systematically examine current visual temporal reasoning tasks, we propose three principles with corresponding metrics: (1) *Multi-Frame Gain*, (2) *Frame Order Sensitivity*, and (3) *Frame Information Disparity*. Following these principles, we introduce TVBENCH, Temporal Reasoning Video Understanding Benchmark, a novel benchmark crafted to rigorously assess MFMs’ temporal reasoning capabilities in video understanding. TVBENCH comprises 1,484 carefully curated, *human-annotated* questions spanning *six* tasks (*i.e.*, *action count*, *direction*, *rotation*, *shape & trend*, *velocity & frequency*, and *visual cues*), applied to 1,417 videos, including 805 self-recorded and -generated videos, that encompass human-centric, real-world, and simulated scenarios. Our comprehensive evaluation reveals a human-model performance gap of 57.3% with the best-performing model. Moreover, our in-depth analysis uncovers more fundamental limitations beyond this gap in current MFMs. While they can accurately recognize events in isolated frames, they fail to interpret these frames as a continuous sequence. We believe TVBENCH will serve as a crucial testbed for evaluating the next-generation MFMs and as a call to the community to develop AI systems capable of comprehending the human world dynamics through the video modality.

1 INTRODUCTION

Visual temporal reasoning, an important aspect of human perception, refers the cognitive process of understanding and interpreting sequences of visual information over time, such as recognizing patterns of motions, detecting changes in scenery, and following the progression of events (Kahneman et al., 1992). Currently, state-of-the-art methods for addressing visual temporal reasoning are centered on the use of Multimodal Foundation Models (MFMs) (OpenAI, 2024; Anthropic, 2024b; Wang et al., 2024a; Fei et al., 2024), which have shown remarkable performance across numerous temporal reasoning video benchmarks (Li et al., 2023; Liu et al., 2024; Li et al., 2024b; Chen et al., 2024a). However, despite these impressive performances, our study in §5 has shown that the models’ true capabilities in visual temporal reasoning are likely overestimated.

This study examines four existing widely-used temporal reasoning video benchmarks (Li et al., 2023; 2024b; Liu et al., 2024; Chen et al., 2024a), as illustrated in Figure 1. We identify patterns in their question formulation that allow models to exploit shortcuts, enabling them to answer correctly using a single, few, or out-of-order frames. To rigorously evaluate whether a benchmark effectively assesses MFMs’ visual temporal reasoning ability, we propose three principles with corresponding metrics: (1) *Multi-Frame Gain*, where a *higher* value indicates the task is less solvable by a single frame (§3.1), (2) *Frame Order Sensitivity*, where a *higher* value indicates the task is more reliant on the correct order of frames (§3.2), and (3) *Frame Information Disparity*, where a *lower* value indicates information is more evenly distributed across the frames (§3.3). Our analysis reveals that, on average,

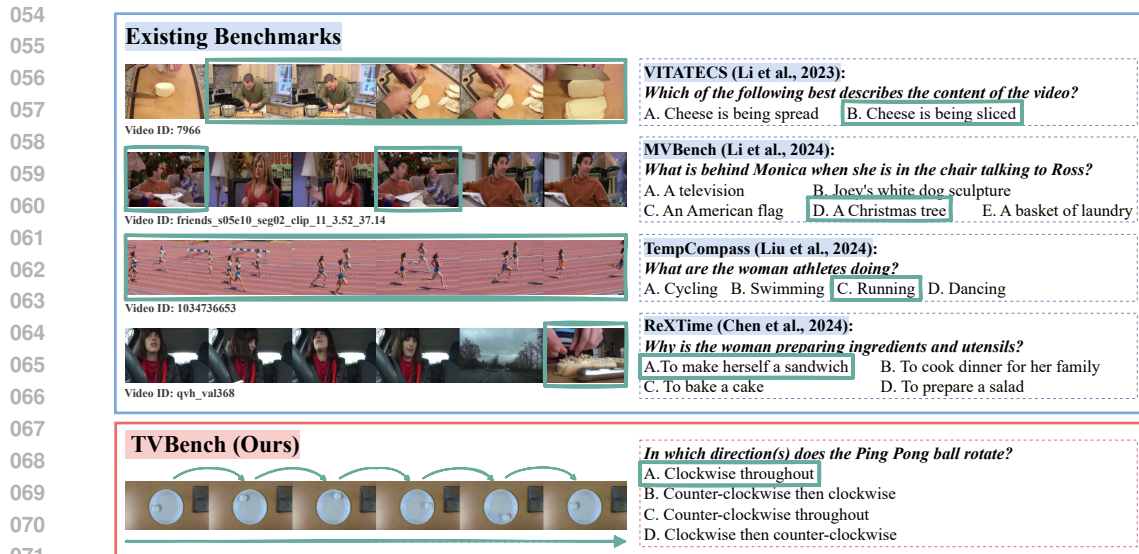


Figure 1: Examples of several VideoQA benchmarks. We examine four existing representative benchmarks and ours: VITATECS, MVBench, TempCompass, ReXTime, and **TVBENCH**. In the four existing benchmarks, we use bounding boxes to highlight frames that models can exploit as shortcuts, allowing questions to be answered using *any one* of highlighted frames. In **TVBENCH**, however, models must reason both between consecutive frames and across all frames as a continuous sequence.

existing benchmark exhibit a *Multi-Frame Gain* of less than 5%, compared to 66.3% for ours (§5.1); a *Frame Order Sensitivity* of less than 8%, compared to 34.1% for ours (§5.2); and a *Frame Information Disparity* of over 27%, while ours is only 4.6% (§5.3). These results suggest that tasks from existing benchmarks are relatively more solvable by a single frame, rely less on the order of the frames, and allow for more unevenly distributed information across the frames. Therefore, models’ visual temporal reasoning capabilities, crucial to MFMs’ video comprehension, are likely overestimated.

Following the three aforementioned principles for building a more effective visual temporal reasoning benchmark, we introduce **TVBENCH**, **T**emporal Reasoning **V**ideo Understanding **B**enchmark, a novel video understanding benchmark designed to explicitly assess MFMs’ temporal reasoning capabilities. **TVBENCH** comprises 1,484 carefully curated, *human-annotated* multiple-choice questions spanning *six* distinct temporal reasoning tasks (*action count*, *direction*, *rotation*, *shape & trend*, *velocity & frequency*, and *visual cues*) applied to 1,417 videos encompassing human-centric, real-world, and simulated scenarios (§4.1). **TVBENCH** features a diverse collection of videos, sourcing from YouTube, four existing video datasets (Jang et al., 2017; Yi et al., 2020; Li et al., 2022; Pătrăucean et al., 2023), as well as self-recorded and -generated videos (§4.2). To enhance diversity, we carefully edit the sourced YouTube and self-recorded videos, incorporating various characteristics such as counter-factual scenarios, composite motions, and zoomed-in views. In creation of the 805 original videos, we explicitly include both human-centric and simulated scenarios to address the lack of existing videos that capture human interactions or generated synthetic scenes. The question-answering (QA) pairs are meticulously designed to ensure that models need to reason about the transitions across all frames, making visual temporal reasoning essential for solving the tasks.

We conduct a comprehensive evaluation of 20 open-source models and 9 proprietary models. Notably, the best-performing open-source model, Qwen2-VL-72B, which achieves 37.9% overall accuracy, outperforms all proprietary models, including GPT-4o, which achieves 37.7% overall accuracy. However, both types of models remain significantly below human-level performance, which reaches 95.2% using full videos and 79.7% with 16 frames. Moreover, our analysis goes beyond highlighting these performance gap, and it exposes deeper, more fundamental limitations in current MFMs’ capabilities (§6.3). Specifically, we show that these models: (1) lack the basic ability to interpret the frames as a continuous sequence, (2) fail to truthfully leverage the visual input while being over-reliant on common sense, and (3) are highly susceptible to noisy information. We hope that our findings can provide useful insights for future work in guiding development of improved MFMs.

In summary, our contributions are:

- We identify an inflated performance of MFMs on existing visual temporal reasoning benchmarks and establish *three* principles with metrics to evaluate the effectiveness of benchmarks in assessing models’ visual temporal reasoning capabilities.
- We introduce **TVBENCH**, a novel benchmark for assessing MFMs’ capabilities on visual temporal reasoning tasks, spanning *six* reasoning types and *three* video scenarios, including 805 self-created and -generated videos.
- We present a comprehensive evaluation of 20 open-source models and 9 proprietary models on **TVBENCH**, revealing a substantial gap between human-level and MFM-enabled visual temporal reasoning capabilities.
- We provide in-depth error case analysis, uncovering more fundamental shortcomings in MFMs’ visual temporal reasoning capabilities that go beyond the human-model performance gap.

2 RELATED WORK

General video understanding benchmarks. Video understanding capability plays a pivotal role in multimodal learning and a key step toward achieving artificial general intelligence. Prior to the era of MFMs, early benchmarks (Heilbron et al., 2015; Goyal et al., 2017; Sigurdsson et al., 2018) focus primarily on action recognition. However, more recent benchmarks (Jang et al., 2017; Yi et al., 2020) represent a shift towards evaluating models’ ability to reason about temporal dynamics and causal events. The rise of MFMs has further propelled the field toward more complex, human-like understanding tasks. These tasks include (1) long-form video understanding (Zhang et al., 2023; Rawal et al., 2024; Zhou et al., 2024; Nagrani et al., 2024), (2) multi-disciplinary video understanding (He et al., 2024), and (3) comprehensive evaluation across various tasks (Ning et al., 2023; Li et al., 2024d;c; Fu et al., 2024b). Building upon this remarkable progress, MMBench-Video (Fang et al., 2024) advocates the needs for more temporal questions since many questions in existing benchmarks are rendered rather “static,” but its reasoning dimension such as Attribute Recognition, Object Recognition, and OCR still remain static. This notion motivates the creation of **TVBENCH**, a benchmark specifically designed to evaluate MFMs’ visual temporal reasoning capabilities.

Visual temporal reasoning benchmarks. Several benchmarks have been developed to specifically evaluate models’ visual temporal reasoning capabilities. For instance, VITATECS (Li et al., 2023) introduces six temporal reasoning tasks (*e.g.*, “A man is putting on a tie or putting off a tie?”) and asks models to distinguish between the correct and counter-factual caption. Addressing the lack of tasks diversity in VITATECS, TempCompass (Liu et al., 2024) expands tasks types to include multiple-choice QA, yes/no QA, caption matching, and caption generation tasks. Aiming to cover a wide-range of temporal-sensitive videos, MVBench (Li et al., 2024b) defines nine core temporal tasks with 20 subtasks, each of which cannot be solved using a single frame. Similarly, ReXTime (Chen et al., 2024a) targets comprehensive temporal reasoning tasks and puts a special emphasis on cause and effect samples. However, despite these efforts, as illustrated in Figure 1, we observe that many questions in these benchmarks can be answered correctly using single, few, or out-of-order frames (§5), limiting their effectiveness to evaluate models’ true visual temporal reasoning capabilities. To address these shortcomings, we introduce **TVBENCH**, a benchmark designed to provide a more rigorous evaluation of visual temporal reasoning.

3 BENCHMARKING PRINCIPLES FOR VISUAL TEMPORAL REASONING TASKS

In this section, we define three key principles and corresponding metrics for assessing how rigorously a benchmark targets visual temporal reasoning rather than static image understanding.

3.1 MULTI-FRAME GAIN

Key principle: A visual temporal reasoning task should require reasoning across multiple frames, making it impossible for models to solve the task using 1 frame.

162 Requiring models to reason across multiple frames ensures that the tasks are distinguished from static
 163 image recognitions. This requirement aligns with prior works (Lei et al., 2022; Fang et al., 2024),
 164 which highlight that many video understanding tasks rely heavily on static visual information.

165 To assess this principle, we define *Multi Frame Gain* κ , which measures the relative performance gain
 166 from using a single frame to m frames as input. When considering a single frame, we examine both
 167 settings: (1) a random frame and (2) a handpicked, highly informative frame specific to the question.
 168 The m frames are uniformly sampled at equal intervals from the videos. Using $Acc(m \text{ frames})$ to
 169 denote the model’s accuracy in solving benchmark tasks using m frames, and ϵ a small positive
 170 constant, we define κ as:

$$171 \kappa = \frac{Acc(m \text{ frames})}{Acc(1 \text{ frame}) + \epsilon} - 1$$

172 A lower κ value indicates that the question can be more accurately answered using a single frame,
 173 while a higher κ value indicates a necessity to reason across multiple frames.

174 3.2 FRAME ORDER SENSITIVITY

175 **Key principle:** A visual temporal reasoning task, when given multiple frames, should enforce
 176 constraints on maintaining the correct order of frames.

177 Shuffling video frames disrupts the temporal information throughout the video, as demonstrated in
 178 the prior works (Misra et al., 2016; Dorkenwald et al., 2022; Hao et al., 2022). If a task is solvable by
 179 shuffled frames, time dependencies across frames are absent, and reasoning along the time dimension
 180 is unnecessary, which disqualifies the task from being temporal.

181 To quantify this principle, we introduce *Frame Order Sensitivity* τ , which measures the relative
 182 performance gain from using the shuffled m frames to the ordered m frames. Using $Acc(m \text{ frames})$
 183 to denote the model’s accuracy in solving benchmark tasks using m frames, and ϵ a small positive
 184 constant, we define τ as:

$$185 \tau = \frac{Acc(m \text{ frames})}{Acc(\text{shuffled } m \text{ frames}) + \epsilon} - 1$$

186 A lower τ value indicates the question can be answered more accurately using out-of-order frames,
 187 while a higher τ value suggests a stronger reliance on the original order of the frames.

188 3.3 FRAME INFORMATION DISPARITY

189 **Key principle:** A visual temporal reasoning task, when given multiple frames, should ensure that
 190 each frame contributes relatively evenly to solving the task.

191 Even contribution suggests that no single frame provides disproportionately more information. Even
 192 in tasks involving sequential events, where the number of events aligns with the number of frames
 193 necessary to answer the question accurately, models should not achieve significantly higher accuracy
 194 by relying on a handpicked single frame over any random single frame. This principle aligns with
 195 prior works (Huang et al., 2018; Liu et al., 2021) in which they emphasize the necessity of leverage
 196 temporal information from all frames to ensure an accurate and robust temporal video understanding.

197 To quantitatively evaluate this principle, we introduce *Frame Information Disparity* ρ , which measures
 198 the relative performance gain achieved by switching from a random single frame to a handpicked
 199 single frame. Using $Acc(m \text{ frames})$ to denote the model’s accuracy in solving benchmark tasks using
 200 m frames, and ϵ a small positive constant, we define ρ as:

$$201 \rho = \frac{Acc(\text{handpicked } 1 \text{ frame})}{Acc(\text{random-sampled } 1 \text{ frame}) + \epsilon} - 1$$

202 A higher ρ value indicates that the question can be more accurately answered by a handpicked single
 203 frame compared to a random single frame, while a lower ρ value indicates a more even distribution

of informativeness across the multiple frames. In other words, ideally, a benchmark with perfectly even distribution of informativeness among all the frames should achieve a ρ value of 0.

4 TVBENCH: A VISUAL TEMPORAL REASONING BENCHMARK

We introduce **TVBENCH**, a new visual temporal reasoning benchmark that satisfies all three aforementioned principles, addressing issues where tasks from existing benchmarks are relatively more solvable by a single frame, rely less on the order of frames, and allow for more unevenly distributed information across the frames (§5). **TVBENCH** comprises 1,484 carefully curated, *human-annotated* questions spanning *six* visual temporal reasoning tasks, applied to 1,417 videos, including 805 self-recorded and -generated videos, that encompass human-centric, real-world, and simulated scenarios. In the following sections, we describe temporal tasks in **TVBENCH** (§4.1), video collection (§4.2), and question annotation §4.3. The main statistics of **TVBENCH** are presented in Table 9.

4.1 TEMPORAL TASKS IN TVBENCH

We introduce *six* visual temporal reasoning tasks, each of them requiring multi-frame visual temporal reasoning: (1) *Rotation*: Determine the rotational direction of of the subject; (2) *Direction*: Identify the direction of the subject’s movement; (3) *Velocity & Frequency*: Detect changes in the subject’s movement speed or variations in the frequency of repeated actions; (4) *Shape & Trend*: Analyze the subject’s trajectory, such as the shape or general trend of its movement; (5) *Visual Cues*: Discern key visual signals to determine the sequence or timing of actions without audio; and (6) *Action Count*: Count how many times a specific action has been performed. Examples are provided in Table 6.

To ensure comprehensive coverage across various scenarios, we categorized each video into one of three *demonstration categories*: (1) *Human-centric*: Involving human interactions, where actions or intentions are observed; (2) *Real-world*: Focuses on actions involving objects in various real-world scenes; and (3) *Simulated*: Depicting simplified, simulated environments representing temporal actions. The distribution of demonstration categories across each task is shown in Figure 3 in §E.2.

4.2 VIDEO COLLECTION

TVBENCH features a diverse range of videos from three distinct sources: *YouTube*, *existing video datasets*, and *self-recorded and -generated benchmark-specific videos*. To enhance diversity, we collect and create videos of three scenarios: *real-world*, *human-centric*, and *simulated*. Additionally, we edit videos to incorporate *counterfactual scenes*, *composite motions*, and *zoomed-in views*, aiming to investigate the impact of these characteristics on the performance of MFMs (§6.3). License information for all videos is detailed in §F.

YouTube videos. We chose YouTube as our primary source to ensure diverse and representative real-world scenarios in our benchmark. Given a specific visual temporal reasoning task, we task human annotators with searching for YouTube videos that best represented the corresponding task definition (§4.1). The selected videos cover a wide array of topics, including science experiments, outdoor activities, educational tutorials, and artistic performances. To prevent models from relying on commonsense knowledge, such as assuming *the second hand of a watch rotates clockwise*, we edit videos to create counterfactual scenarios. These edits, including reversing, concatenating, adjusting speed, and mirroring, ensure that models must fully analyze video content to answer questions correctly. In total, we collect 171 source videos before editing and 398 videos after editing.

Existing video datasets. It is essential for AI agents to comprehend the human world through detecting the point of change in actions (*Visual Cues*) and discern the number of actions occurred (*Action Count*). Therefore, to address the lack of diversity in YouTube videos for these reasoning types, we incorporate four established datasets, each contributing to unique domains and scenarios: (1) *Music-AVQA* (Li et al., 2022) featuring multi-instrument performances; (2) *CLEVRER* (Yi et al., 2020) presenting synthetic scenes with multiple moving objects; (3) *TGIF-QA* (Jang et al., 2017) focusing on action counting in varied scenes ; and (4) *Perception Test* (Pătrăucean et al., 2023) addressing indoor action counting tasks, such as clapping and object moving. For *Music-AVQA* and *CLEVER*, we re-annotate and re-write questions to specifically inquire the temporal contexts of these

270 videos (§4.3). In total, we incorporate 214 videos: 70 from Music-AVQA, 50 from CLEVRER, 50
 271 from TGIF-QA, and 44 from Perception Test.

272
 273 **Self-recorded and -generated benchmark-specific videos.** For an AI agent to truly comprehend
 274 the human world dynamics, it is essential for it to understand scenarios where a person is actively
 275 interacting with the agent, treating the agent as the “other” participant of an interaction. However,
 276 YouTube videos and existing datasets lack such human-centric interactive scenes for various reasoning
 277 types in **TVBENCH** (§4.1). To address this limitation, we record videos such as a human conveying
 278 “turn-around” by twirling their wrist while maintaining a pointing gesture, or drawing a shape in
 279 the air using their arm. Furthermore, to aid data collection and study the effect of simulated videos,
 280 we generate more simple, abstract representations of both real-world objects and human. Thus, we
 281 expand our dataset by (1) creating simulated videos using Keynote¹, featuring objects moving in
 282 different patterns; and (2) generating 3D human model videos through the VIBE (Kocabas et al.,
 283 2019) and SMPL (Bogo et al., 2016) frameworks. To study the effect of different video characteristics,
 284 we edit videos to incorporate counter-factual scenes, composite motions, and zoomed-in views. In
 285 total, we create 298 videos before editing and 805 videos after editing, spanning all *six* tasks.

286 4.3 QUESTION-ANSWER ANNOTATION AND QUALITY CHECK.

287
 288 **QA annotation.** To address the limitation that tasks from existing benchmarks are solvable by
 289 a single, a few, or out-of-order frames (discussed further in §5), **TVBENCH** focuses on crafting
 290 questions that require reasoning about transitions across all frames. Our annotation process varied by
 291 video sources. For YouTube videos, self-created benchmark-specific videos, and their edits, human
 292 annotators composed QA pairs targeting specific temporal reasoning tasks (§4.1). For Music-AVQA
 293 and CLEVRER, we re-annotated QA pairs to emphasize temporal aspects, such as “*which musical*
 294 *instrument plays first*” for Music-AVQA and “*how many collisions are there in the video*” for
 295 CLEVRER. For TGIF-QA and Perception Test, we retained their existing questions but generate
 296 additional numerical answer options close to the groundtruth.

297
 298 **Quality check.** To ensure high-quality QA annotations, we implemented a three-stage process:
 299 initial annotation by annotators, followed by cross-checking and verification among annotators, and
 300 finally, collective resolution of disagreements with a final review (see §G). This rigorous approach
 301 ensured consistency and accuracy across all annotated QAs.

302 5 COMPARISONS AMONG VISUAL TEMPORAL REASONING BENCHMARKS

303
 304 In §3, we defined three key principles with corresponding metrics to assess how effectively a
 305 benchmark addresses visual temporal reasoning. Using these metrics, we compare **TVBENCH** with
 306 four recent visual temporal reasoning benchmarks: VITATECS (Li et al., 2023), MVBench (Li
 307 et al., 2024b), TempCompass (Liu et al., 2024), and ReXTime (Chen et al., 2024a). To conduct
 308 this comparison, we randomly sampled approximately 200 QAs² from these benchmarks, using two
 309 state-of-the-art MFMs: GPT-4o (OpenAI, 2024) and Qwen2-VL-72B (Wang et al., 2024a). For
 310 metrics that requires handpicking frames, we present annotators with the full video and corresponding
 311 question, and ask them to select the most informative frame for each benchmark. For metrics requiring
 312 multiple frames, we set $m = 16$, as our study across $m = 1, 8, 16, 32$ demonstrates that 16 frames
 313 provide a sufficient window for effective analysis (§E.3).

314 5.1 MULTI-FRAME GAIN

315
 316 We present results on *Multi-frame Gain* for both (1) a random frame (Table 1) and (2) a handpicked,
 317 highly informative frame (Table 2). As shown in the two tables, **TVBENCH** achieves a significantly
 318 higher κ value in using both random single frame and handpicked single frame. This significant
 319 relative performance gain from the single frame input setting to the 16 frames input shows the
 320 necessity in our tasks to reason using multiple frames. In comparison of the two tables, we observe
 321

322 ¹<https://support.apple.com/keynote>

323 ²Sampling varies slightly: TempCompass: 200 samples. MVBench: 204 samples (excluding overlapping reasoning types). VITATECS: 203 samples. ReXTime: 180 samples from QVHighlight’s validation set.

Table 1: MFMs performance using single-frame and 16-frame inputs.

# Frames	VITATECS			MVBench			TempCompass			ReXTime			TVBENCH		
	1	16	$\kappa \uparrow$	1	16	$\kappa \uparrow$	1	16	$\kappa \uparrow$	1	16	$\kappa \uparrow$	1	16	$\kappa \uparrow$
GPT-4o	70.0	88.2	26.1	47.1	62.3	32.3	52.5	71.5	36.2	61.7	78.3	26.9	21.2	37.7	78.0
Qwen2-VL	71.4	86.2	20.7	47.1	63.2	34.4	50.0	79.0	58.0	63.9	81.7	27.8	20.6	37.9	84.0
Average	70.7	87.2	23.4	47.1	62.7	33.3	51.3	75.3	47.1	62.8	80.0	27.4	20.9	37.8	81.0

Table 2: MFMs performance using handpicked single-frame (denoted by **1[H]**) and 16-frame inputs.

# Frames	VITATECS			MVBench			TempCompass			ReXTime			TVBENCH (200)		
	1[H]	16	$\kappa \uparrow$	1[H]	16	$\kappa \uparrow$	1[H]	16	$\kappa \uparrow$	1[H]	16	$\kappa \uparrow$	1[H]	16	$\kappa \uparrow$
GPT-4o	84.2	88.2	4.7	59.8	62.3	4.1	64.5	71.5	10.9	86.7	78.3	-9.7	21.5	37.0	72.1
Qwen2-VL	87.7	86.2	-1.7	57.8	63.2	9.3	63.5	79.0	24.4	85.6	81.7	-4.6	24.0	38.5	60.4
Average	86.0	87.2	1.5	58.8	62.7	6.7	64.0	75.3	17.6	86.1	80.0	-7.1	22.8	37.8	66.3

Table 3: MFMs performance using shuffled 16-frame (denoted by **16[S]**) and 16-frame inputs.

# Frames	VITATECS			MVBench			TempCompass			ReXTime			TVBENCH		
	16[S]	16	$\tau \uparrow$	16[S]	16	$\tau \uparrow$	16[S]	16	$\tau \uparrow$	16[S]	16	$\tau \uparrow$	16[S]	16	$\tau \uparrow$
GPT-4o	85.7	88.2	2.9	59.8	62.3	4.1	59.0	71.5	21.2	77.8	78.3	0.6	25.8	37.7	46.2
Qwen2-VL	83.3	86.2	3.6	58.8	63.2	7.5	64.0	79.0	23.4	81.7	81.7	0	31.1	37.9	21.9
Average	84.5	87.2	3.2	59.3	62.7	5.8	61.5	75.3	22.3	79.7	80.0	0.3	28.5	37.8	34.1

Table 4: MFMs performance using handpicked single-frame (denoted by **1[H]**) and 16-frame inputs.

# Frames	VITATECS			MVBench			TempCompass			ReXTime			TVBENCH (200)		
	1	1[H]	$\rho \downarrow$	1	1[H]	$\rho \downarrow$	1	1[H]	$\rho \downarrow$	1	1[H]	$\rho \downarrow$	1	1[H]	$\rho \downarrow$
GPT-4o	70.0	84.2	20.4	47.1	59.8	27.1	52.5	64.5	22.9	61.7	86.7	40.5	20.5	21.5	4.9
Qwen2-VL	71.4	87.7	22.7	47.1	57.8	22.9	50.0	63.5	27.0	63.9	85.6	33.9	23.0	24.0	4.3
Average	70.7	86.0	21.6	47.1	58.8	25.0	51.3	64.0	24.9	62.8	86.1	37.2	21.8	22.8	4.6

an expected decrease in the κ value in using handpicked single frame compared to random single frame on existing benchmarks, indicating the relative ease in answering the benchmark questions using handpicked single frame. Interestingly, the κ values are negative in the handpicked single frame setting on ReXTime; this negative performance gain in using more frames might stem from noise introduced by additional frames.

5.2 FRAME ORDER SENSITIVITY

In our shuffled 16 frames setting, we apply random shuffling on the ordered 16 frames to ensure that the same set of frames are used in both settings. As shown in Table 3, TVBENCH achieves significantly higher τ , demonstrating that our benchmark imposes a stricter requirement on maintaining the order of the frames to accurately answer the question.

5.3 FRAME INFORMATION DISPARITY

As shown in Table 4, the performance gain from random single frame to handpicked single frame is the lowest on TVBENCH, indicating a relatively more consistent informativeness across all frames compared to other existing benchmarks.

6 EVALUATING VISUAL TEMPORAL REASONING IN ADVANCED MFMS

In introducing TVBENCH, we present a comprehensive evaluation of 29 MFMs, including 9 proprietary models and 20 open-source models, to assess their visual temporal reasoning capabilities.

Table 5: Evaluation results on **TVBENCH**. Unless otherwise specified, each model processes 16 frames. **Bold** and underlined numbers indicate the best and second-best performance in each category, respectively. See Table 8 for detailed model configurations. ‡: Models that can process video directly. *: Models that can only process eight frames. Details for these categories are shown in §4.1

Model	Rotation (286)	Direction (403)	Velocity & Frequency (210)	Shape & Trend (223)	Visual Cues (70)	Action Count (292)	All (1,484)
<i>Baselines</i>							
Human (Videos)	93.5	95.4	94.1	100.0	95.0	93.6	95.2
Human (Frames)	60.9	93.9	85.3	91.7	60.0	70.2	79.7
Random (GPT-4o)	16.8	17.4	35.7	29.6	32.9	20.9	23.1
Random Choice (42)	22.0	17.6	22.9	17.9	18.6	13.4	18.5
Frequent Choice	21.5	18.5	24.8	20.2	33.6	18.8	21.0
<i>Proprietary Multimodal Foundation Models (MFMs)</i>							
GPT-4o	24.5	45.2	31.9	42.6	58.6	<u>36.0</u>	37.7
Gemini 1.5 pro ‡	25.9	<u>40.7</u>	35.2	<u>41.3</u>	37.1	36.3	<u>36.1</u>
GPT-4o-mini	19.9	32.8	28.1	29.6	<u>41.4</u>	28.8	28.8
Claude 3 Opus	31.1	23.3	<u>32.4</u>	27.8	28.6	29.5	28.2
Claude 3.5 Sonnet	<u>27.3</u>	25.6	26.2	27.8	32.9	31.2	27.8
Gemini 1.5 Flash ‡	22.4	30.3	31.0	26.9	30.0	27.7	27.8
Claude 3 Haiku	25.9	19.4	31.0	33.2	30.0	26.4	26.2
Reka Flash ‡	19.6	26.6	10.0	21.5	32.9	16.8	20.5
Reka Core ‡	14.3	17.6	9.5	18.8	22.9	12.7	15.3
<i>Open-Source Multimodal Foundation Models (MFMs)</i>							
Qwen2-VL-72B	26.9	38.2	43.8	<u>36.3</u>	<u>48.6</u>	42.8	37.9
Qwen2-VL-7B	23.8	29.5	<u>41.9</u>	29.6	37.1	34.2	<u>31.5</u>
Video-CCAM-v1.1 14B	32.2	26.1	29.5	27.4	44.3	<u>35.6</u>	30.7
InternVL 2 40B	23.4	<u>32.0</u>	15.7	41.7	34.3	29.1	29.0
LLaVA-Video-72B	18.9	30.0	16.7	32.3	50.0	34.6	28.2
Video-CCAM-v1.1 9B	22.4	25.6	25.7	26.0	34.3	33.6	27.0
InternVideo 2 8B	<u>31.8</u>	24.1	23.3	25.6	35.7	25.0	26.4
LLaVA-OneVision 7B	16.8	25.1	23.8	28.7	35.7	31.2	25.5
Video-CCAM-v1.1 4B	21.7	24.3	19.0	27.4	32.9	31.5	25.3
LLaVA-Video-7B	20.1	23.8	15.2	29.1	41.4	29.8	24.9
VILA 13B	29.0	19.6	19.0	27.4	32.9	27.7	24.7
LLaVA-Video-7B-Video-Only	15.4	24.1	19.5	31.4	38.6	25.7	23.9
Video LLaVA 7B *	29.4	17.9	27.1	23.3	34.3	20.9	23.6
VideoLLaMA 2 72B ‡	14.3	24.6	22.4	26.5	27.1	28.8	23.5
InternVL 2 26B	18.5	29.3	10.5	31.4	11.4	25.7	23.3
LLaVA-NeXT-Video-32B	20.6	26.3	12.4	24.2	30.0	24.3	22.7
InternVL 2 8B	17.1	25.1	9.0	28.7	31.4	22.9	21.7
Phi 3.5 Vision	20.3	16.6	14.3	23.3	40.0	24.7	20.7
AuroraCap-7B-VID	19.6	17.1	26.7	20.6	25.7	17.1	20.6
VideoLLaMA 2 7B ‡	10.1	22.8	15.7	18.8	31.4	19.5	18.5

In the following sections, we detail the experimental setup (§6.1), evaluation results (§6.2), and a multi-faceted analysis (§6.3), considering factors such as model architectures, reasoning type correlations, frame counts, and video characteristics.

6.1 EXPERIMENTAL SETUP

Models. We evaluate a diverse set of general-purpose MFMs. For *open-source* MFMs, we test: VILA (Lin et al., 2023b), InternVL 2 (Chen et al., 2024b) Phi 3.5 Vision (Abdin et al., 2024), Video LLaVA (Lin et al., 2023a), InternVideo 2 (Wang et al., 2024b), LLaVA-NeXT-Video (Zhang et al., 2024a), LLaVA-OneVision (Li et al., 2024a), VideoLLaMA2 (Cheng et al., 2024), Qwen2-VL (Wang et al., 2024a), VideoCCAM (Fei et al., 2024), AuroraCap (Chai et al., 2024), and LLaVA-Video (Zhang et al., 2024b). We also evaluate the following *proprietary* MFMs: GPT-4o (OpenAI, 2024), Claude (Anthropic, 2024a;b), Reka (Reka, 2024), and Gemini 1.5 (Gemini, 2024). For all models, we provide generation configuration in Table 8.

Baselines. We include a text-only baseline, denoted as Random (GPT-4o), in which we prompt GPT-4o to guess the answer without access to the videos (prompt in §D.3). Additionally, we report results for Random Guess and Frequent Guess baselines. Furthermore, we evaluate human performance on

TVBENCH, reporting results for humans using video input, *i.e.*, Human (Video), and 16 frames as input, *i.e.*, Human (Frames). Annotator biographies are provided in Table 10 in §G.1.

6.2 EXPERIMENTAL RESULTS

We provide quantitative results on **TVBENCH** for all models in Table 5. To better understand where models fail, we select a set of representative models (OpenAI, 2024; Anthropic, 2024b; Wang et al., 2024a; Fei et al., 2024; Chen et al., 2024b) and present examples of failure cases in §I, §J, §K, §L, §M, and §N. Analysis of video scenarios (counterfactual, zoomed-in, and first-person) is in §B.

Widespread difficulty in visual temporal reasoning. Our evaluation (Table 5) underscores the significant challenges of **TVBENCH** across all tested models. The leading open-source model, Qwen2-VL-72B, achieves 37.9% accuracy, slightly outperforming GPT-4o’s 37.7%. However, this still leaves a substantial 57.3% performance gap compared to human accuracy of 95.2%. While this result demonstrates the competitive potential of open-source models in video understanding, many still fall below 30.0%, indicating weaknesses in **TVBENCH** despite their decent performance in existing benchmarks (Pătrăucean et al., 2023; Li et al., 2024b; Fu et al., 2024a).

6.3 ANALYSIS

Models lack the basic ability to interpret frames as a continuous sequence. While MFMs demonstrate remarkable performance in understanding sequential events in videos (Chen et al., 2024a), our benchmark exposes a more fundamental limitation: models struggle to reason across multiple time steps and to interpret the frames as a continuous sequence. As shown in error case §I.2.1, GPT-4o correctly generates captions for each consecutive change in the moon’s movement, showcasing its ability to reason at individual time steps. However, it fails to infer based on the captions that the overall sequence represents a clockwise rotation. This issue is not limited to *rotation* (§I.1.1, §I.2.1); similar shortcomings are observed in *direction* (§J.3.3), *action count* (§N.2.5), *etc.*

Models fail to truthfully leverage the visual input while being over-reliant on common sense. In our evaluations, despite explicit instructions (§D.1) to rely on the visual input rather than common sense, we find that models frequently hallucinate based on information from single frames rather than utilizing true *visual* reasoning. For instance, in error case §K.2.4, GPT-4o incorrectly concludes that an object is dropping due to the presence of motion blur in some of the frames. However, this is a reversed video where the object is actually only moving upward - a conclusion that can only be reached if the video modality were truthfully utilized. A similar instance of the limitation occurs in error case §J.1.2, where GPT-4o likely assumes the person raises their hand first to reach the posture depicted in the first 8 single frames. In reality, the person’s hand remains relatively stationary throughout these frames, but the model fails to make accurate visual comparisons across these frames.

Models are highly susceptible to noisy information in the input. As demonstrated in error case §J.3.3, where a block moves downwards, models are especially vulnerable to noisy information, such as misleading text on the block. In particular, while GPT-4o correctly describes the block’s downward motion based on its relative position to the screen, it incorrectly concludes that the block is moving upward, likely influenced by the false information presented in the text written on the block. Similarly, in error case §N.2.5, a butterfly-shaped laser spot is moving in a triangle shape. However, the unique butterfly shape likely causes the models to lose focus on the trajectory of the laser spot, and resort to random guessing in their conclusions.

Explicitly incorporating time-aware positional encoding can likely enhance visual temporal reasoning. The Qwen2-VL family (Wang et al., 2024a), which consistently achieves the highest scores across multiple categories, leverages Multimodal Rotary Positional Encoding (M-RoPE) within its visual encoders. M-RoPE explicitly encodes temporal information into visual tokens, allowing models to retain critical temporal context throughout the entire pipeline. In contrast, models without such temporal-aware positional encoding scheme appear to lose critical time-related context after visual encoding, such as Causal Cross-Attention Masks (CCAM) used in vision-language alignment (Fei et al., 2024), Unmasked Video Token Reconstruction during training (Wang et al., 2024b), and Spatial-Temporal Convolution (STC) connector applied after frozen visual encoder (Cheng et al.,

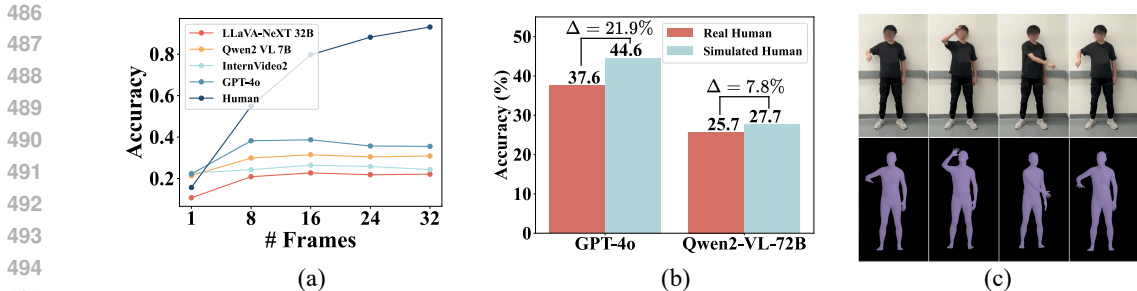


Figure 2: (a) Human and models’ performance on **TVBENCH** with different number of frames. (b) Real human and simulated human’s performance on **TVBENCH**. (c) Example of Real human vs. simulated human.

2024). Performance of Qwen2-VL on **TVBENCH** suggests that explicitly incorporating temporal-aware positional encoding like M-RoPE is likely essential for enhancing MFMs’ visual temporal reasoning capabilities. Future improvements in open-source models could benefit from adopting similar strategies to close the gap between the open-source and proprietary models.

Existing models are limited to understanding events that are interpretable in ≤ 8 frames. We assess four MFMs’ performance across different number of frames on **TVBENCH**, as illustrated in Figure 2 (a). The consistent improvement in accuracy of human performance suggests that our benchmark rely on the additional frames to convey more temporal information. Notably, although models exhibit performance increase transitioning from 1 frame to 8 frames, the performance plateaus beyond this point. This suggests that even if models are able to reason the transitions between the frames before 8 frames, they cannot utilize the additional temporal information obtained by the added frames. Therefore, we conclude that the overall performance of the four MFMs remain suboptimal in their visual temporal reasoning capabilities, and there is still room for improvements in MFMs’ ability to leverage the additional information on the frame transitions introduced by added frames.

Models perform better on simulated human than real human scenarios. To investigate the extent to which a cleaner, more abstract representation of video content (Figure 2 (c)) influences models’ temporal reasoning abilities, we evaluate two leading MFMs. Specifically, this evaluation contrasts real human scenarios and their corresponding simulated counterparts, across five reasoning types (*i.e.*, *action count*, *direction*, *rotation*, *shape & trend*, and *velocity & frequency*), covering the same 101 QA pairs in both scenarios. As a result, GPT-4o marks a noticeable improvement of 21.9% from real human to simulated human scenarios, underscoring the potential of enhancing models’ temporal reasoning capabilities in video understanding through semantic video abstraction. Conversely, Qwen2-VL-72B displays a modest increase of 7.8% transitioning from real human to simulated human. While it marginally outperforms GPT-4o in the overall evaluation on **TVBENCH**, its visual temporal reasoning capabilities in simulated scenarios still show room for improvement (Figure 2 (b)). Future work can target further enhancing models’ temporal reasoning abilities for real human videos by exploring their generalization capabilities through leveraging automatically generated simulated 3D human motion data (Guo et al., 2022).

7 CONCLUSION

Existing benchmarks likely overestimate the true visual temporal reasoning capabilities of MFMs. In response, we establish three key principles and corresponding metrics to systematically examine visual temporal reasoning tasks. Building upon these principles, we introduce **TVBENCH**, a novel video understanding benchmark to rigorously assess MFMs’ true visual temporal reasoning capabilities. Besides revealing a previously underestimated human-model performance gap, our comprehensive evaluation highlights a critical limitation: MFMs fail to interpret videos as continuous sequences, instead resorting to understanding isolated frames, which severely undermines their visual temporal reasoning capabilities. This work sheds the light for developing AI systems capable of comprehending changing scenes in real life through the video modality.

REFERENCES

- 540
541
542 Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany
543 Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Ben-
544 haim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro
545 Mendes, Weizhu Chen, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Yen-Chun Chen, Yi-
546 Ling Chen, Parul Chopra, Xiyang Dai, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon,
547 Ronen Eldan, Victor Fragoso, Dan Iter, Mei Gao, Min Gao, Jianfeng Gao, Amit Garg, Abhishek
548 Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh,
549 Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud
550 Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars
551 Liden, Ce Liu, Mengchen Liu, Weishung Liu, Eric Lin, Zeqi Lin, Chong Luo, Piyush Madan,
552 Matt Mazzola, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel
553 Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sam-
554 budha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital
555 Shah, Ning Shang, Hiteshi Sharma, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini,
556 Xin Wang, Lijuan Wang, Chunyu Wang, Yu Wang, Rachel Ward, Guanhua Wang, Philipp Witte,
557 Haiping Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan
558 Yang, Jianwei Yang, Ziyi Yang, Yifan Yang, Donghan Yu, Lu Yuan, Chengruidong Zhang, Cyril
559 Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou.
560 Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL
<https://arxiv.org/abs/2404.14219>.
- 561 Anthropic. Introducing the next generation of claude, 2024a. URL <https://www.anthropic.com/news/claude-3-family>.
- 562
563 Anthropic. Claude 3.5 sonnet, 2024b. URL <https://www.anthropic.com/news/claude-3-5-sonnet>.
- 564
565
566 Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter V. Gehler, Javier Romero, and Michael J.
567 Black. Keep it SMPL: automatic estimation of 3d human pose and shape from a single image.
568 *CoRR*, abs/1607.08128, 2016. URL <http://arxiv.org/abs/1607.08128>.
- 569
570 Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jeng-Neng
571 Hwang, Saining Xie, and Christopher D. Manning. Auroracap: Efficient, performant video detailed
572 captioning and a new benchmark, 2024. URL <https://arxiv.org/abs/2410.03051>.
- 573
574 Jr-Jen Chen, Yu-Chien Liao, Hsi-Che Lin, Yu-Chu Yu, Yen-Chun Chen, and Yu-Chiang Frank
575 Wang. Rextime: A benchmark suite for reasoning-across-time in videos, 2024a. URL <https://arxiv.org/abs/2406.19392>.
- 576
577 Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi
578 Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial
579 multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024b.
- 580
581 Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi
582 Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal
583 modeling and audio understanding in video-llms, 2024. URL <https://arxiv.org/abs/2406.07476>.
- 584
585 Michael Dorkenwald, Fanyi Xiao, Biagio Brattoli, Joseph Tighe, and Davide Modolo. Scvrl: Shuffled
586 contrastive video representation learning, 2022. URL <https://arxiv.org/abs/2205.11710>.
- 587
588 Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen.
589 Mmbench-video: A long-form multi-shot benchmark for holistic video understanding, 2024. URL
590 <https://arxiv.org/abs/2406.14515>.
- 591
592 Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. Video-ccam: Enhancing
593 video-language understanding with causal cross-attention masks for short and long videos, 2024.
URL <https://arxiv.org/abs/2408.14023>.

- 594 Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu
595 Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li,
596 Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. Video-mme: The first-
597 ever comprehensive evaluation benchmark of multi-modal llms in video analysis, 2024a. URL
598 <https://arxiv.org/abs/2405.21075>.
- 599 Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu
600 Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li,
601 Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. Video-mme: The first-
602 ever comprehensive evaluation benchmark of multi-modal llms in video analysis, 2024b. URL
603 <https://arxiv.org/abs/2405.21075>.
- 604 Gemini. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context,
605 2024.
- 606 Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal,
607 Heuna Kim, Valentin Haenel, Ingo Fründ, Peter N. Yianilos, Moritz Mueller-Freitag, Florian Hoppe,
608 Christian Thureau, Ingo Bax, and Roland Memisevic. The “something something” video database for
609 learning and evaluating visual common sense. *2017 IEEE International Conference on Computer
610 Vision (ICCV)*, pp. 5843–5851, 2017. URL [https://api.semanticscholar.org/CorpusID:
611 834612](https://api.semanticscholar.org/CorpusID:834612).
- 612 Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating
613 diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on
614 Computer Vision and Pattern Recognition*, pp. 5152–5161, 2022.
- 615 Jiachang Hao, Haifeng Sun, Pengfei Ren, Jingyu Wang, Qi Qi, and Jianxin Liao. Can shuffling video
616 benefit temporal bias problem: A novel training framework for temporal grounding, 2022. URL
617 <https://arxiv.org/abs/2207.14698>.
- 618 Xuehai He, Weixi Feng, Kaizhi Zheng, Yujie Lu, Wanrong Zhu, Jiachen Li, Yue Fan, Jianfeng Wang,
619 Linjie Li, Zhengyuan Yang, Kevin Lin, William Yang Wang, Lijuan Wang, and Xin Eric Wang.
620 Mmworld: Towards multi-discipline multi-faceted world model evaluation in videos, 2024. URL
621 <https://arxiv.org/abs/2406.08407>.
- 622 Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet:
623 A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on
624 Computer Vision and Pattern Recognition (CVPR)*, pp. 961–970, 2015. doi: 10.1109/CVPR.2015.
625 7298698.
- 626 De-An Huang, Vignesh Ramanathan, Dhruv Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei,
627 and Juan Carlos Niebles. What makes a video a video: Analyzing temporal information in video
628 understanding models and datasets. In *Proceedings of the IEEE Conference on Computer Vision
629 and Pattern Recognition (CVPR)*, June 2018.
- 630 Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-
631 temporal reasoning in visual question answering, 2017. URL [https://arxiv.org/abs/1704.
632 04497](https://arxiv.org/abs/1704.04497).
- 633 Daniel Kahneman, Anne Treisman, and Brian J Gibbs. The reviewing of object files: Object-
634 specific integration of information. *Cognitive Psychology*, 24(2):175–219, 1992. ISSN 0010-0285.
635 doi: [https://doi.org/10.1016/0010-0285\(92\)90007-O](https://doi.org/10.1016/0010-0285(92)90007-O). URL [https://www.sciencedirect.com/
636 science/article/pii/0010028592900070](https://www.sciencedirect.com/science/article/pii/0010028592900070).
- 637 Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: video inference for human
638 body pose and shape estimation. *CoRR*, abs/1912.05656, 2019. URL [http://arxiv.org/abs/
639 1912.05656](http://arxiv.org/abs/1912.05656).
- 640 Jie Lei, Tamara L. Berg, and Mohit Bansal. Revealing single frame bias for video-and-language
641 learning, 2022. URL <https://arxiv.org/abs/2206.03428>.
- 642 Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei
643 Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024a. URL
644 <https://arxiv.org/abs/2408.03326>.

- 648 Guanyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer
649 questions in dynamic audio-visual scenarios, 2022. URL <https://arxiv.org/abs/2203.14072>.
- 650
- 651 Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping
652 Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding
653 benchmark, 2024b. URL <https://arxiv.org/abs/2311.17005>.
- 654 Shicheng Li, Lei Li, Shuhuai Ren, Yuanxin Liu, Yi Liu, Rundong Gao, Xu Sun, and Lu Hou. Vitatecs:
655 A diagnostic dataset for temporal concept understanding of video-language models, 2023. URL
656 <https://arxiv.org/abs/2311.17404>.
- 657
- 658 Xinhao Li, Zhenpeng Huang, Jing Wang, Kunchang Li, and Limin Wang. Videoeval: Comprehensive
659 benchmark suite for low-cost evaluation of video foundation model, 2024c. URL <https://arxiv.org/abs/2407.06491>.
- 660
- 661 Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and
662 Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description, 2016. URL <https://arxiv.org/abs/1604.02748>.
- 663
- 664 Yunxin Li, Xinyu Chen, Baotian Hu, Longyue Wang, Haoyuan Shi, and Min Zhang. Videovista: A
665 versatile benchmark for video understanding and reasoning, 2024d. URL <https://arxiv.org/abs/2406.11303>.
- 666
- 667
- 668 Bin Lin, Yang Ye, Bin Zhu, Jiayi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning
669 united visual representation by alignment before projection, 2023a. URL <https://arxiv.org/abs/2311.10122>.
- 670
- 671 Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz,
672 Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models, 2023b.
- 673
- 674 Xin Liu, Silvia L. Pintea, Fatemeh Karimi Nejadasl, Olaf Booij, and Jan C. van Gemert. No frame
675 left behind: Full video action recognition, 2021. URL <https://arxiv.org/abs/2103.15395>.
- 676 Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun,
677 and Lu Hou. Tempcompass: Do video llms really understand videos?, 2024. URL <https://arxiv.org/abs/2403.00476>.
- 678
- 679 Ishan Misra, C. Lawrence Zitnick, and Martial Hebert. Shuffle and learn: Unsupervised learning
680 using temporal order verification, 2016. URL <https://arxiv.org/abs/1603.08561>.
- 681
- 682 Arsha Nagrani, Mingda Zhang, Ramin Mehran, Rachel Hornung, Nitesh Bharadwaj Gundavarapu,
683 Nilpa Jha, Austin Myers, Xingyi Zhou, Boqing Gong, Cordelia Schmid, Mikhail Sirotenko, Yukun
684 Zhu, and Tobias Weyand. Neptune: The long orbit to benchmarking long video understanding.
685 2024.
- 686 Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiayi Cui, Lu Yuan, Dongdong Chen, and Li Yuan.
687 Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language
688 models, 2023. URL <https://arxiv.org/abs/2311.16103>.
- 689
- 690 OpenAI. Hello gpt-4o, 2024. URL <https://openai.com/index/hello-gpt-4o/>.
- 691
- 692 Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Contente, Larisa Markeeva,
693 Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch,
694 Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex Frechette, Hanna Klimczak, Raphael
695 Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew
696 Zisserman, and João Carreira. Perception test: A diagnostic benchmark for multimodal video
697 models, 2023. URL <https://arxiv.org/abs/2305.13786>.
- 698
- 699 Ruchit Rawal, Khalid Saifullah, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein.
700 Cinepile: A long video question answering dataset and benchmark, 2024. URL <https://arxiv.org/abs/2405.08813>.
- 701
- 701 Reka. Reka core, flash, and edge: A series of powerful multimodal language models, 2024. URL
<https://arxiv.org/abs/2404.12387>.

702 Gunnar A. Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor
703 and observer: Joint modeling of first and third-person videos, 2018. URL <https://arxiv.org/abs/1804.09627>.
704
705
706 Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung yi Lee, and Yun-Nung Chen. Let
707 me speak freely? a study on the impact of format restrictions on performance of large language
708 models, 2024. URL <https://arxiv.org/abs/2408.02442>.

709 Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu,
710 Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng
711 Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s
712 perception of the world at any resolution, 2024a. URL <https://arxiv.org/abs/2409.12191>.

713 Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Chenting Wang, Guo Chen, Baoqi Pei,
714 Ziang Yan, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li,
715 Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2: Scaling
716 foundation models for multimodal video understanding, 2024b. URL <https://arxiv.org/abs/2403.15377>.
717
718

719 Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B.
720 Tenenbaum. Clevrer: Collision events for video representation and reasoning, 2020. URL
721 <https://arxiv.org/abs/1910.01442>.

722 Hongjie Zhang, Yi Liu, Lu Dong, Yifei Huang, Zhen-Hua Ling, Yali Wang, Limin Wang, and Yu Qiao.
723 Movqa: A benchmark of versatile question-answering for long-form movie understanding, 2023.
724 URL <https://arxiv.org/abs/2312.04817>.

725
726 Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and
727 Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024a. URL
728 <https://llava-vl.github.io/blog/2024-04-30-llava-next-video/>.

729 Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video
730 instruction tuning with synthetic data, 2024b. URL <https://arxiv.org/abs/2410.02713>.

731
732 Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang,
733 Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video
734 understanding, 2024. URL <https://arxiv.org/abs/2406.04264>.
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755

756	CONTENTS	
757		
758	A Tasks Examples of TVBENCH	19
759		
760		
761	B Scenario Analysis	20
762		
763	C Experiment Setup	21
764	C.1 Model Configuration	21
765	C.2 Implementation Details for Model Inference	21
766		
767		
768	D Prompts	22
769	D.1 Evaluation Prompt	22
770	D.2 Answer Extracting Prompt	22
771	D.3 Video-Free Random Guess Prompt	23
772		
773		
774		
775	E Data Analysis	24
776	E.1 Statistics	24
777	E.2 Video Source Distribution	25
778	E.3 Performance Comparisons across Different Number of Frames	26
779	E.4 Model Performance with ≤ 8 Frames	26
780	E.5 Video Duration Distribution	27
781	E.6 Response Length Distribution	28
782		
783		
784		
785	F License Information.	30
786		
787		
788	G Annotation Details	31
789	G.1 Annotator Biographies	31
790	G.2 Pilot Annotation	31
791	G.3 Full-Scale Annotation	31
792		
793		
794	H Data Annotation Platform	32
795		
796		
797	I Common Failure Cases of Rotation	33
798	I.1 Human	33
799	I.1.1 Example 1	33
800	I.2 Object	34
801	I.2.1 Example 1	34
802	I.2.2 Example 2	35
803	I.2.3 Example 3	37
804	I.2.4 Example 4	38
805	I.2.5 Example 5	39
806	I.2.6 Example 6	40
807		
808		
809		

810	I.2.7	Example 7	41
811	I.2.8	Example 8	42
812	I.2.9	Example 9	44
813			
814	I.3	Simulated	46
815			
816	I.3.1	Example 1	46
817	I.3.2	Example 2	47
818	I.3.3	Example 3	48
819	I.3.4	Example 4	49
820	I.3.5	Example 5	51
821			
822			
823			
824	J	Common Failure Cases of Direction	52
825	J.1	Human	52
826			
827	J.1.1	Example 1	52
828	J.1.2	Example 2	54
829	J.1.3	Example 3	56
830			
831	J.2	Object	58
832			
833	J.2.1	Example 1	58
834	J.2.2	Example 2	60
835	J.2.3	Example 3	61
836	J.2.4	Example 4	63
837	J.2.5	Example 5	65
838	J.2.6	Example 6	66
839			
840			
841	J.3	Simulated	67
842			
843	J.3.1	Example 1	67
844	J.3.2	Example 2	68
845	J.3.3	Example 3	69
846	J.3.4	Example 4	70
847	J.3.5	Example 5	71
848	J.3.6	Example 6	72
849			
850			
851			
852	K	Common Failure Cases of Velocity & Frequency	73
853	K.1	Human	73
854			
855	K.1.1	Example 1	73
856	K.2	Object	75
857			
858	K.2.1	Example 1	75
859	K.2.2	Example 2	77
860	K.2.3	Example 3	78
861	K.2.4	Example 4	79
862	K.2.5	Example 5	81
863			

864	K.2.6 Example 6	82
865		
866	K.2.7 Example 7	83
867		
868	K.2.8 Example 8	84
869		
870	K.2.9 Example 9	85
871		
872	K.2.10 Example 10	86
873		
874	K.2.11 Example 11	87
875		
876	K.2.12 Example 12	89
877		
878	K.3 Simulated	90
879		
880	K.3.1 Example 1	90
881		
882		
883		
884	L Common Failure Cases of Shape & Trend	91
885		
886	L.1 Human	91
887		
888	L.1.1 Example 1	91
889		
890	L.2 Simulated	93
891		
892	L.2.1 Example 1	93
893		
894	L.2.2 Example 2	94
895		
896	L.2.3 Example 3	95
897		
898	L.2.4 Example 4	96
899		
900	L.2.5 Example 5	97
901		
902	L.2.6 Example 6	98
903		
904	L.2.7 Example 7	99
905		
906		
907		
908		
909		
910		
911		
912		
913		
914		
915		
916		
917		
	M Common Failure Cases of Visual Cues	100
	M.1 Object	100
	M.1.1 Example 1	100
	M.1.2 Example 2	101
	M.1.3 Example 3	103
	M.1.4 Example 4	104
	M.1.5 Example 5	105
	M.1.6 Example 6	106
	M.1.7 Example 7	107
	M.1.8 Example 8	109
	M.1.9 Example 9	110
	M.1.10 Example 10	112
	M.1.11 Example 11	114
	M.1.12 Example 12	115
	M.1.13 Example 13	116
	M.1.14 Example 14	117
	M.1.15 Example 15	119

918	M.1.16 Example 16	120
919		
920		
921	N Common Failure Cases of Action Count	121
922	N.1 Human	121
923	N.1.1 Example 1	121
924	N.1.2 Example 2	123
925	N.1.3 Example 3	125
926	N.1.4 Example 4	127
927		
928		
929	N.2 Object	129
930	N.2.1 Example 1	129
931	N.2.2 Example 2	131
932	N.2.3 Example 3	132
933	N.2.4 Example 4	134
934	N.2.5 Example 5	136
935		
936		
937	N.3 Simulated	137
938	N.3.1 Example 1	137
939	N.3.2 Example 2	139
940	N.3.3 Example 3	141
941	N.3.4 Example 4	142
942		
943		
944		
945		
946		
947		
948		
949		
950		
951		
952		
953		
954		
955		
956		
957		
958		
959		
960		
961		
962		
963		
964		
965		
966		
967		
968		
969		
970		
971		

972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025

A TASKS EXAMPLES OF TVBENCH

Table 6: Task examples of TVBENCH. Some videos are collected from existing video datasets, including Music-AVQA (Li et al., 2022), CLEVRER (Yi et al., 2020), TGIF-QA (Jang et al., 2017), and Perception Test (Pătrăucean et al., 2023). †: Tasks requiring re-annotation (§4.3).

Temporal Tasks	Video Sources	Examples
Rotation (19.3%)	YouTube & Self-created	<i>In which direction(s) does the object rotate?</i> (A) Clockwise (B) Counter-clockwise (C) Clockwise then counter-clockwise (D) Counter-clockwise then clockwise (E) No rotation
Direction (27.2%)	YouTube & Self-created	<i>In which direction(s) does the person’s hand move?</i> (A) Left (B) Right (C) First to the left then to the right (D) First to the right then to the left (E) No movements
Velocity & Frequency (14.2%)	YouTube & Self-created	<i>What is the speed pattern of the train?</i> (A) Accelerating (B) Decelerating (C) Constant Speed (D) No movement
Shape & Trend (15.0%)	YouTube & Self-created	<i>What is the shape of the object that the person draws in the air?</i> (A) Circle (B) Triangle (C) Square/rectangle (D) Trapezoid (E) Diamond (F) Not drawing at all
Visual Cues (4.7%)	<i>Music-AVQA</i> †	<i>Which musical instrument plays first?</i> (A) Accordion (B) Saxophone (C) Both instruments play simultaneously (D) Neither instrument produces any sound
Action Count (19.7%)	<i>CLEVRER</i> †	<i>How many collision(s) are there in the video?</i> (A) 1 (B) 2 (C) 3 (D) 4 (E) 5 (F) 6
	<i>TGIF-QA</i>	<i>How many times does the cat lick the water tap?</i> (A) 1 (B) 2 (C) 3 (D) 4 (E) 5
	<i>Perception Test</i>	<i>How many times does the person launch the object on the slanted plane?</i> (A) 1 (B) 2 (C) 3 (D) 4 (E) 5 (F) 6
	Self-created	<i>How many trapezoid(s) does the person draw in the air?</i> (A) 1 (B) 2 (C) 3 (D) 4 (E) 5 (F) 6

B SCENARIO ANALYSIS

Table 7: MFMs performance on Counterfactual, Zoomed-in, and First-Person Perspective QAs. “True” and “False” indicate whether the questions meet the criteria specified in the merged cell above the respective column (*e.g.*, Counterfactual, Zoomed-In, and First Person Perspective).

	Counterfactual			Zoomed-In			First-Person Perspective		
	False	True	$\Delta\%$	False	True	$\Delta\%$	False	True	$\Delta\%$
GPT-4o	38.5	24.0	-37.7	37.8	40.1	+6.0	38.8	47.5	+22.5
Qwen2-VL-72B	45.3	32.0	-29.4	31.6	30.3	-4.2	30.4	61.3	+101.6
Qwen2-VL-7B	37.1	33.0	-11.0	24.9	25.4	+1.9	24.7	42.5	+72.1
Video-CCAM-v1.1 14B	37.5	29.0	-22.6	24.7	24.4	-1.0	24.1	35.0	+45.2

In this section, we present our analysis across various video scenarios. While the videos are not explicitly designed to rigorously validate the findings, we believe these insights are valuable for highlighting general trends observed in the experimental results for future studies.

More capable models are more reliant on common sense. In curating the counterfactual QAs, we employ video editing techniques (*e.g.*, reversing, rotating, cropping) to produce contents that are impossible to observe in real life. As detailed in Table 7, although all four models demonstrate similar performance on non-counterfactual QAs, the shift to counterfactual examples reveals a significant performance drop, particularly for the best general-purpose model, GPT-4o, with a decrease of 37.7%, and the leading open-source model, Qwen2-VL-72B, with a 29.4% decrease. These results suggest that more capable models are more prone to exploit shortcuts within the task’s background information and heavily rely on pre-trained knowledge, rather than truthfully understanding the video content, even when explicitly instructed to not rely on commonsense reasoning in solving the tasks.

Zoomed-in views offer limited performance improvement in challenging human scenarios. Although performance on standard views varies across different models (Table 7), providing manual zoomed-in views yields only modest performance gains: 6.0% for GPT-4o and 1.9% for Qwen2-VL-7B. Surprisingly, the zoomed-in views even worsen the performance of Video-CCAM-v1.1 14B and Qwen2-VL-72B by 1.0% and 4.2%, respectively. The limited aid provided by the zoomed-in view indicates that the challenges inherent to the tasks can not be tackled through zooming-in, but require deeper temporal understanding of the videos beyond enhancements in visual focus.

Models excel in first-person over third-person perspective temporal reasoning video understanding. In comparing 80 first-person perspective QAs to a larger set of 668 third-person perspective QAs, our goal is to explore the general trends in model performances based on the presence or absence of a main subject in the video. The results suggest that the absence of a main subject does not hinder model performance. In fact, we observe significantly better model performance on the first-person perspective tasks (Table 7). Notably, the Qwen2-VL models achieve remarkable performance gains of 101.6% and 72.1%, respectively, with Qwen2-VL-72B scoring 61.3—outperforming GPT-4o by 28.9% under the same conditions. These findings also highlight the potential of open-source models to surpass more capable general-purpose proprietary models on temporal reasoning video understanding tasks.

C EXPERIMENT SETUP

C.1 MODEL CONFIGURATION

Table 8: Model configurations for evaluation. Unset values indicate that their default values are being used. ‡: Models that can process video directly. *: Models that can only process eight frames. †: Set token_kept_ratio to 0.61. Configurations are based on official model repositories where available. **Temp.:** temperature.

Model	API Checkpoint / HF Checkpoint	Do Sample	Max New Tokens	Temp.	Top-P	Seed
<i>Proprietary Multimodal Foundation Models (MFMs)</i>						
GPT-4o-mini	gpt-4o-mini-2024-07-18		1024	0	1	215
GPT-4o	gpt-4o-2024-08-06		1024	0	1	215
Claude 3 Opus	claude-3-opus-20240229		1024	0	1	
Claude 3.5 Sonnet	claude-3-5-sonnet-20240620		1024	0	1	
Claude 3 Haiku	claude-3-haiku-20240307		1024	0	1	
Gemini 1.5 Flash	gemini-1.5-flash-001		1024	0	1	
Reka Flash	reka-flash-20240226		1024	0	1	
Reka Core	reka-core-20240501		1024	0	1	
<i>Open-Source Multimodal Foundation Models (MFMs)</i>						
InternVL 2 8B	OpenGVLab/InternVL2-8B	False	1024			
InternVL 2 26B	OpenGVLab/InternVL2-26B	False	1024			
InternVL 2 40B	OpenGVLab/InternVL2-40B	False	1024			
InternVideo 2 Chat 8B	OpenGVLab/InternVideo2-Chat-8B	False	1024			
Video-CCAM-v1.1 4B	JaronTHU/Video-CCAM-4B-v1.1	False	1024			
Video-CCAM-v1.1 9B	JaronTHU/Video-CCAM-9B-v1.1	False	1024			
Video-CCAM-v1.1 14B	JaronTHU/Video-CCAM-14B-v1.1	False	1024			
LLaVA NeXT Video 32B	lmms-lab/LLaVA-NeXT-Video-32B-Qwen	False	1024			
LLaVA OneVision 7B	lmms-lab/llava-onevision-qwen2-7b-ov	False	1024			
Video LLaVA 7B *	LanguageBind/Video-LLaVA-7B	False	1024			
VILA 13B	Efficient-Large-Model/VILA-13b	False	1024			
Phi 3.5 Vision Instruct	microsoft/Phi-3.5-vision-instruct	False	1024			
Qwen2-VL-7B	Qwen/Qwen2-VL-7B-Instruct	False	1024			
Qwen2-VL-72B	Qwen/Qwen2-VL-72B-Instruct-AWQ	False	1024			
VideoLLaMA2 7B ‡	DAMO-NLP-SG/VideoLLaMA2-7B	False	1024			
VideoLLaMA2 72B ‡	DAMO-NLP-SG/VideoLLaMA2-72B	False	1024			
LLaVA-Video-72B	lmms-lab/LLaVA-Video-72B-Qwen2	False	1024			
LLaVA-Video-7B	lmms-lab/LLaVA-Video-7B-Qwen2	False	1024			
LLaVA-Video-7B-Video-Only	lmms-lab/LLaVA-Video-7B-Qwen2-Video-Only	False	1024			
AuroraCap-7B-VID †	wchai/AuroraCap-7B-IMG-xtuner	False	1024			

C.2 IMPLEMENTATION DETAILS FOR MODEL INFERENCE

All MFMs are evaluated using a zero-shot strategy across all benchmarks, including **TVBENCH**, to ensure fair comparison. Whenever possible, we use the provided code from each model’s official page for video preprocessing. Notably, VideoLLaMA 2 7B and VideoLLaMA 2 72B (Cheng et al., 2024) processes video content directly without splitting it into frames, so we input the videos without modification. Gemini 1.5 Flash (Gemini, 2024) also processes video content directly with 1 FPS. As noted by Tam et al. (2024), restricting the output format of MFMs can hinder performance; therefore, we do not impose any output restrictions. However, this approach can complicate answer parsing using simple heuristics. To address this issue, we employ GPT-4o-mini to extract multiple-choice answers from each model’s response when our standard parser fails, where prompts are provided in §D. We use NVIDIA A100 GPUs for all non-API-based evaluation.

1134 D PROMPTS

1135

1136

1137

1138

1139 D.1 EVALUATION PROMPT

1140

1141

1142

1143

1144

1145

1146

1147

1148

1149

1150

1151

1152

1153

1154

1155

1156

1157

1158

1159

1160

1161

1162

1163

1164

1165

1166

1167

1168 D.2 ANSWER EXTRACTING PROMPT

1169

1170

1171

1172

1173

1174

1175

1176

1177

1178

1179

1180

1181

1182

1183

1184

1185

1186

1187

You will be provided with {num_frames} separate frames uniformly sampled from a video, the frames are provided in chronological order of the video. Analyze these frames and provide the answer to the question about the video content. Answer the multiple-choice question about the video content.

You must use these frames to answer the multiple-choice question; do not rely on any external knowledge or commonsense.

<question> {question} </question>

<options> {index2ans} </options>

Even if the information in these separate frames is not enough to answer the question, PLEASE TRY YOUR BEST TO GUESS AN ANSWER WHICH YOU THINK WOULD BE THE MOST POSSIBLE ONE BASED ON THE QUESTION.

DO NOT GENERATE ANSWER SUCH AS 'NOT POSSIBLE TO DETERMINE.'

You are given a response, a list of multiple-choice options, and a index2answer mapping. You are required to extract the letter option from the GPT.

<response> {response} </response>

<all_choices> {all_choices} </all_choices>

<index2answer> {index2ans} </index2answer>

Only output the single parsed letter from the response. No other texts are needed.

If you think no options can match the index2answer dictionary, randomly select one letter.

Your extracted letter is:

1188 D.3 VIDEO-FREE RANDOM GUESS PROMPT
1189

1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241

```
Randomly guess a reasonable answer based on the question only.  
  
<question> {question} </question>  
  
<options> {index2ans} </options>  
  
DO NOT GENERATE ANSWER SUCH AS 'NOT POSSIBLE TO DETERMINE.'
```

1242 E DATA ANALYSIS

1243

1244 E.1 STATISTICS

1245

1246

Table 9: TVBENCH main statistics.

1247

1248

1249

1250

1251

1252

1253

1254

1255

1256

1257

1258

1259

1260

1261

1262

1263

1264

1265

1266

1267

1268

1269

1270

1271

1272

1273

1274

1275

1276

1277

1278

1279

1280

1281

1282

1283

1284

1285

1286

1287

1288

1289

1290

1291

1292

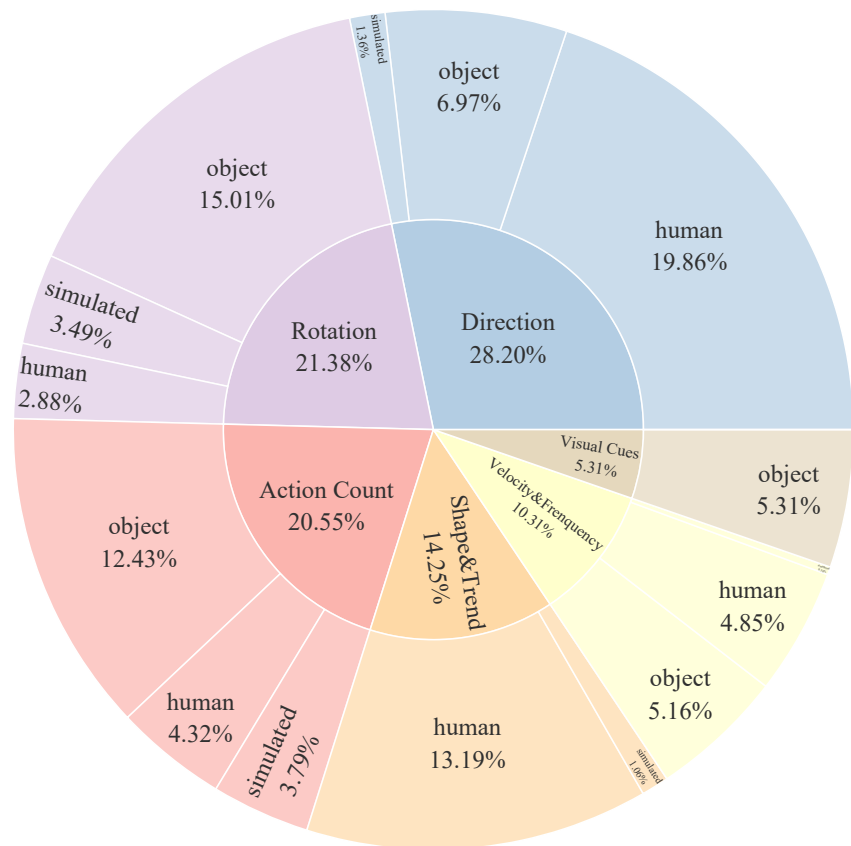
1293

1294

1295

Statistics	Value
Total Questions	1,484
Total Videos	1,417
<i>Demonstration Type</i>	
Human	588 (41.4%)
Object	596 (42.1%)
Simulated	233 (16.4%)
<i>Source</i>	
Self-recorded and -generated	805 (56.8%)
YouTube	398 (28.1%)
Existing Video Datasets	214 (15.1%)
Unique Source Videos	683
Duration (Seconds, avg/max)	9.21 / 72.74
Resolution (avg/max)	1332 × 1076 / 1080 × 1920
Number of Reasoning Tasks	6
Number of Demonstration Categories	3
Question Length (avg/max)	11.71 / 22
Single Choice Length (avg/max)	3.69 / 10
Choices per Question (avg/max)	5.19 / 7

1296 E.2 VIDEO SOURCE DISTRIBUTION
 1297
 1298
 1299
 1300
 1301
 1302
 1303



1316
 1317
 1318
 1319
 1320
 1321
 1322
 1323
 1324
 1325
 1326
 1327
 1328
 1329
 1330
 1331
 1332
 1333
 1334
 1335
 1336
 1337
 1338
 1339
 1340
 1341
 1342
 1343
 1344
 1345
 1346
 1347
 1348
 1349

Figure 3: Video source distribution of TVBENCH.

1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403

E.3 PERFORMANCE COMPARISONS ACROSS DIFFERENT NUMBER OF FRAMES

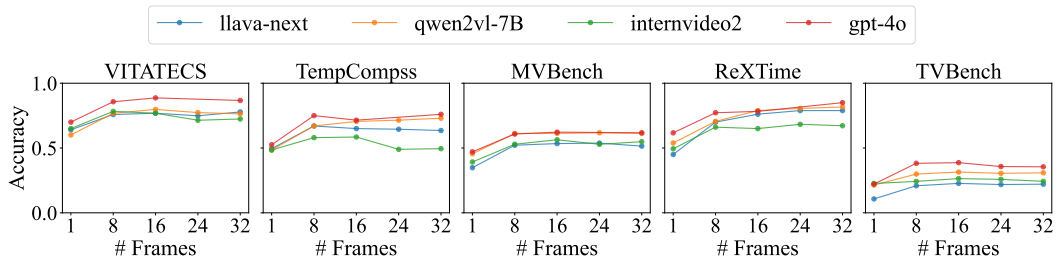


Figure 4: Performance comparisons of VITATECS (Li et al., 2023), TempCompass (Liu et al., 2024), MVBench (Li et al., 2024b), RexTime (Chen et al., 2024a), and TVBENCH across different number of frames.

E.4 MODEL PERFORMANCE WITH ≤ 8 FRAMES

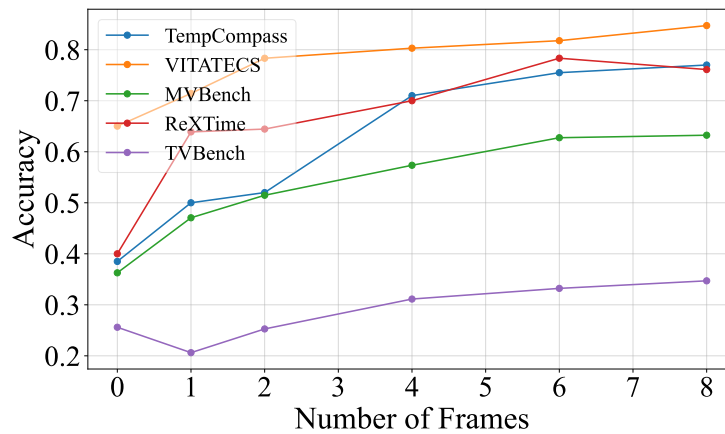


Figure 5: Performance of Qwen2-VL-72B across five benchmarks with input restricted to fewer than 8 frames.

1404 E.5 VIDEO DURATION DISTRIBUTION
 1405
 1406

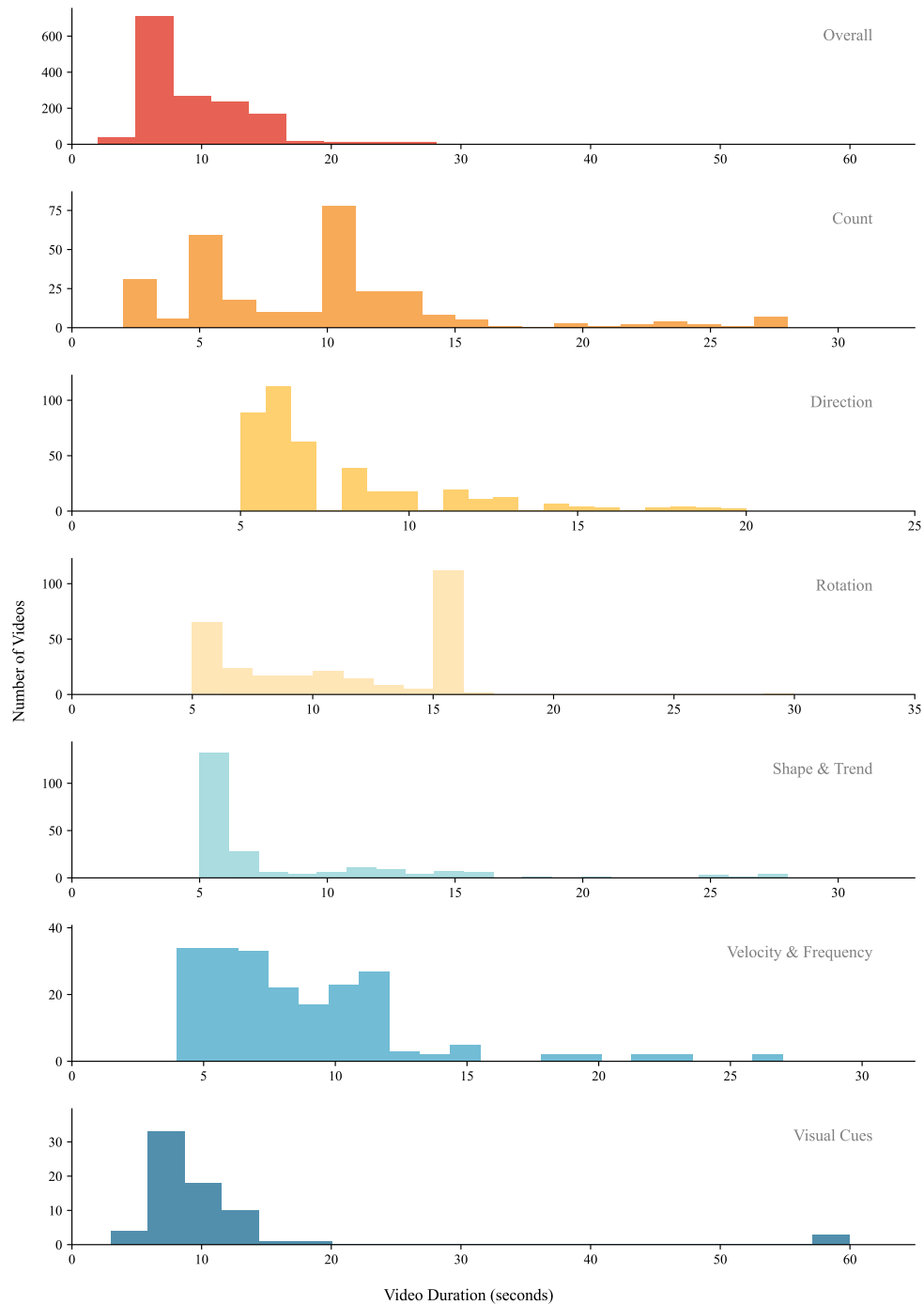


Figure 6: Video duration distribution of TVBENCH

1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511

E.6 RESPONSE LENGTH DISTRIBUTION

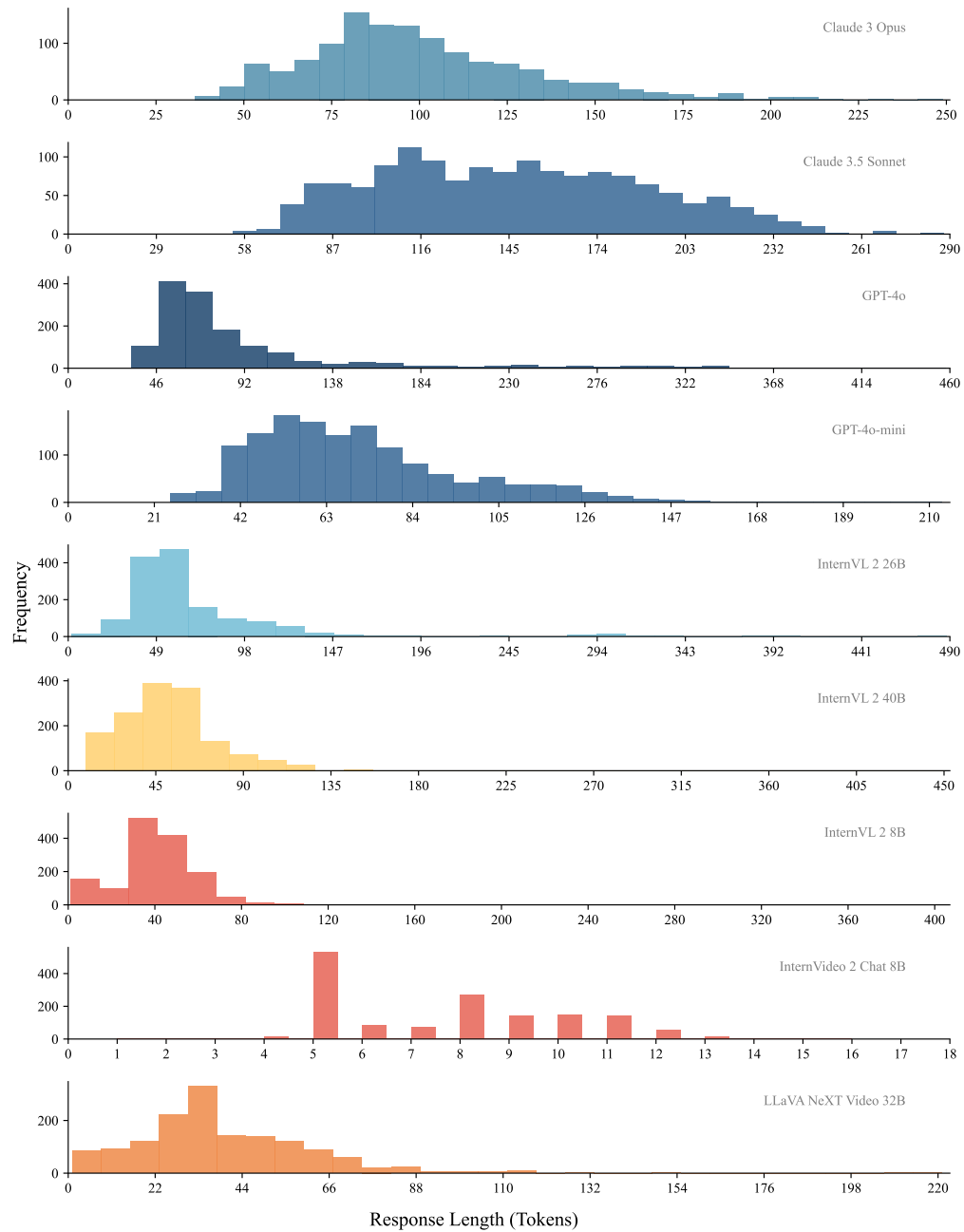


Figure 7: Response length distribution, part 1

1512
 1513
 1514
 1515
 1516
 1517
 1518
 1519
 1520
 1521
 1522
 1523
 1524
 1525
 1526
 1527
 1528
 1529
 1530
 1531
 1532
 1533
 1534
 1535
 1536
 1537
 1538
 1539
 1540
 1541
 1542
 1543
 1544
 1545
 1546
 1547
 1548
 1549
 1550
 1551
 1552
 1553
 1554
 1555
 1556
 1557
 1558
 1559
 1560
 1561
 1562
 1563
 1564
 1565

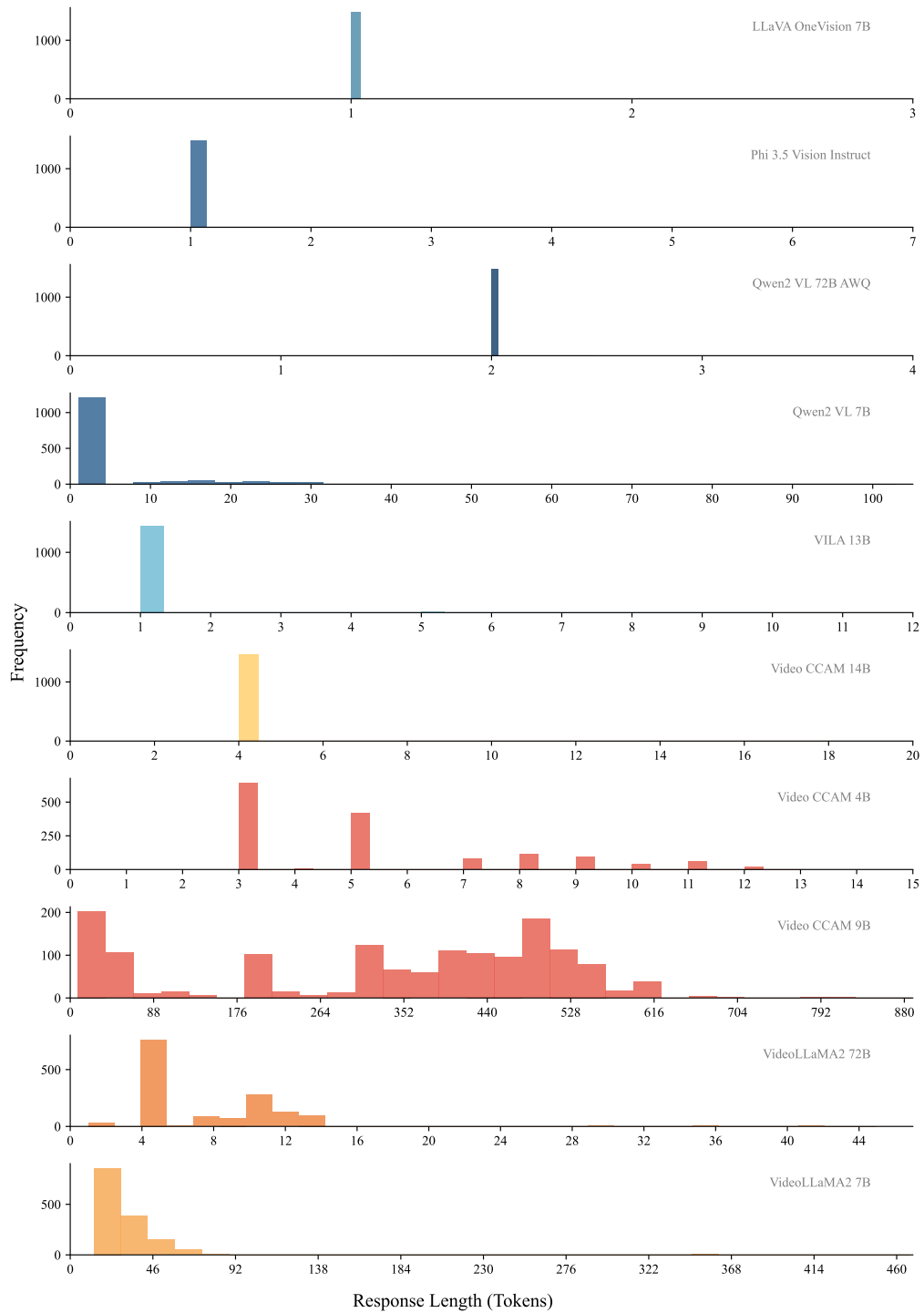


Figure 8: Response length distribution, part 2

1566 F LICENSE INFORMATION.
1567

1568 We, the authors, bear all responsibility in case of violation of rights. **TVBENCH** is intended only used
1569 for academic research purposes. Commercial use of **TVBENCH** in any form is strictly prohibited.
1570 The videos included in **TVBENCH** are subject to the following licensing conditions:
1571

1572 **YouTube videos.** All videos sourced from YouTube³ are licensed under Creative Commons⁴. The
1573 original YouTube video links are provided in our dataset, and proper attribution is given in accordance
1574 with the license terms. We do not hold the copyright to any of the YouTube videos; we ask that you
1575 respect the rights of the original video creators. If you believe that any content in **TVBENCH** infringes
1576 on your rights, please contact the authors immediately, and we will remove the video accordingly.
1577

1578 **Existing video datasets.** CLEVRER (Yi et al., 2020) is licensed under CC0⁵. TGIF-QA Jang et al.
1579 (2017) is built upon TGIF (Li et al., 2016), which is available for non-commercial only⁶. Perception
1580 Test (Pătrăucean et al., 2023) is licensed under Creative Commons Attribution 4.0 International
1581 License (CC-BY)⁷. Music-AVQA (Li et al., 2022) is licensed under Creative Commons Attribution-
1582 NonCommercial 4.0 International (CC BY-NC 4.0)⁸.
1583

1584 **Self-recorded and -generated benchmark-specific videos.** Videos created specifically for
1585 **TVBENCH** are licensed under CC BY-NC-SA⁹. These videos are intended to be freely used for
1586 academic purposes, with proper attribution required. We kindly ask that all individuals appearing in
1587 our self-recorded videos are respected.
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613

1614 ³<https://www.youtube.com/>

1615 ⁴<https://creativecommons.org/>

1616 ⁵<https://creativecommons.org/public-domain/cc0/>

1617 ⁶<https://github.com/raingo/TGIF-Release/blob/master/LICENSE>

1618 ⁷<https://creativecommons.org/licenses/by/4.0/legalcode>

1619 ⁸<https://creativecommons.org/licenses/by-nc/4.0/>

⁹<https://creativecommons.org/licenses/by-nc-sa/4.0/deed.en>

1620 G ANNOTATION DETAILS

1621

1622 G.1 ANNOTATOR BIOGRAPHIES

1623

1624 Table 10: Biographies for human annotators in **TVBENCH**.

1625

1626 ID	1627 Year	1628 Field of Study
1629 1	1630 Second-Year Master’s Student	1631 Mathematics
2	Second-Year Master’s Student	Computer Science & Mathematics
3	Recent Master’s Graduate	Mechanical Engineering
4	Second-Year Ph.D. Student	Computer Science

1632

1633 G.2 PILOT ANNOTATION

1634

1635 To determine the visual temporal reasoning types in **TVBENCH**, we conduct a pilot study aimed at
1636 exploring and refining task definitions. Each annotator is tasked with brainstorming potential task
1637 types relevant to visual temporal reasoning. Once a new reasoning type is proposed, each annotator is
1638 assigned to compose 20 question-answer (QA) pairs of this reasoning type. To ensure that the QA
1639 pairs of this reasoning type adhere to the principles outlined in §3, we utilize GPT-4o to verify each
1640 metrics score. As a result of this pilot process, we identify and finalize *six* distinct visual temporal
1641 reasoning tasks: *Rotation, Direction, Velocity & Frequency, Shape & Trend, Visual Cues*, and *Action*
1642 *Count*. These task types form the foundation for the full-scale annotation of our benchmark, ensuring
1643 comprehensive coverage of visual temporal reasoning.

1644

1645 G.3 FULL-SCALE ANNOTATION

1646

1647 **Initial annotation.** For YouTube videos, annotators are tasked with selecting videos that cor-
1648 responded to a specific reasoning type (as outlined in §4.1). Given the targeted reasoning type,
1649 annotators carefully search for videos from YouTube and compose questions that exemplified the
1650 temporal reasoning patterns (Table 6). For *TGIF-QA* (Jang et al., 2017), annotators are instructed to
1651 select video-question pairs that they could easily answer. This is necessary because many questions
1652 in *TGIF-QA* are challenging for humans to answer due to the low resolution or low frame rate of the
1653 videos. For *CLEVRER* Yi et al. (2020), annotators are responsible for writing the question (*i.e.*, “How
1654 many distinct collision(s) occur throughout the entire video?”), and we extract answers from the meta
1655 data of each video. For *Music-AVQA* (Li et al., 2022), annotators are tasked with curating videos that
1656 exemplified the *Visual Cues* reasoning types and composing corresponding question-answer pairs. For
1657 self-recorded and -generated videos, both annotators and authors are responsible for recording and
1658 generating videos on specific reasoning types, as well as composing corresponding question-answer
1659 pairs.

1659 **Cross-checking and verification.** Each annotator reviews the work of another annotator by care-
1660 fully re-annotating the answer, without being provided with the original annotator’s response. At
1661 the end of the process, the two sets of annotations are compared and checked for consistency. If
1662 discrepancies arise between the two annotators’ answers, both annotators and the authors engage in a
1663 collective resolution to reconcile the differences. During this stage, out of 1,484 questions-answer
1664 pairs, 44 questions require collective resolution during this verification phase.

1665

1666 **Collective resolution.** In cases where the annotators do not reach consensus during the cross-
1667 checking phase, a discussion is held to resolve the disagreement. This phase enables both authors and
1668 annotators to determine whether the discrepancy stems from a lack of careful annotation or from poor
1669 video quality. Out of the 44 mismatched answers, we replace four videos and correct the answers for
1670 the remaining cases.

1671

1672

1673

1674 H DATA ANNOTATION PLATFORM

1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727

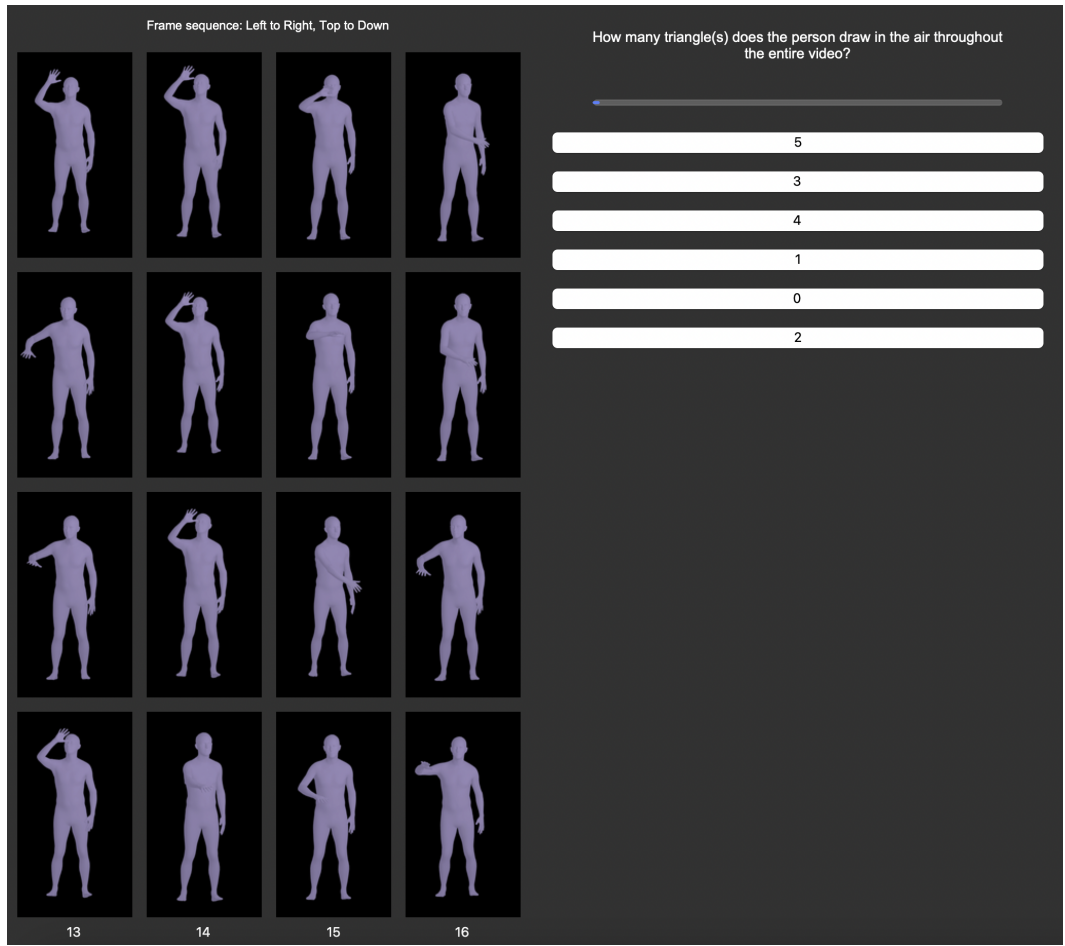


Figure 9: Screenshot of our human evaluation platform. As shown, the annotator is presented with a question (top right) alongside the corresponding video, displayed in a grid of 16 frames (4 × 4). The annotator is then required to select the correct answer from the provided options.

1728 I COMMON FAILURE CASES OF ROTATION

1729

1730 I.1 HUMAN

1731

1732 I.1.1 EXAMPLE 1

1733

1734

rotation 8

1735

1736

1737

1738

1739

1740



1741

Which direction(s) does the person's hand rotate in?

1742

1743

1744

1745

1746

1747

1748

Answer: C. Clockwise throughout.

1749

1750

1751

1752

1753

1754

1755

GPT-4o: Based on the provided frames, the person's hand appears to be rotating in a circular motion. Initially, the hand is pointing upwards, then it moves to the right, downwards, and finally to the left before returning to the starting position. **This suggests a counter-clockwise rotation.**

Therefore, the most likely answer is:

1756

1757

****Option A: Counter-clockwise throughout.****

1758

1759

1760

1761

1762

1763

1764

1765

1766

Claude 3.5 Sonnet: Based on analyzing the sequence of images provided, the most likely answer to the question "Which direction(s) does the person's hand rotate in?" is:

E: Clockwise then counter-clockwise.

The images show the person's hand and arm movement in a series of frames. At the beginning, the hand is raised with the index finger pointing upward. **As the sequence progresses, the hand appears to rotate clockwise, moving down and to the side. Then in later frames, the hand motion reverses, rotating counter-clockwise back up to the initial raised position with the finger pointing up.**

While it's difficult to be certain of the full motion from still images alone, the clockwise followed by counter-clockwise rotation seems to best match the arm and hand positions shown across the frame sequence provided.

1770

1771

Qwen2 VL 72B instruct AWQ: E

1772

1773

Qwen2 VL 7B Instruct: D

1774

1775

Video-CCAM-v1.1 14B: Answer is A.

1776

1777

1778

1779

InternVL 2 40B: Based on the frames provided, the person's hand **appears to rotate in a counter-clockwise direction** throughout the video.

1780

Answer: A

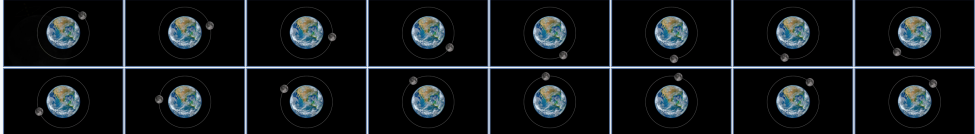
1781

1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835

I.2 OBJECT

I.2.1 EXAMPLE 1

rotation 42 YouTube



<https://www.youtube.com/watch?v=MrzWzMUhQfw>

What direction(s) does the moon rotate relative to the Earth?

- A. Counter-clockwise throughout.
- B. No rotation.
- C. Clockwise throughout.
- D. Clockwise then counter-clockwise.
- E. Counter-clockwise then clockwise.

Answer: C. Clockwise throughout.

GPT-4o: Based on the provided frames, the moon appears to be rotating around the Earth in a counter-clockwise direction. This is evident from the sequential positions of the moon as it moves from the right side of the Earth, downwards, to the left side, and then upwards, completing a counter-clockwise orbit.

Therefore, the most possible answer based on the frames is:

****A': Counter-clockwise throughout.****

Claude 3.5 Sonnet: Based on analyzing the sequence of 16 frames provided, the answer to the question "What direction(s) does the moon rotate relative to the Earth?" appears to be:

A. Counter-clockwise throughout.

The images show the moon orbiting around the Earth in a counter-clockwise direction when viewed from above the North Pole. The moon starts on the right side of the Earth and moves counter-clockwise around the circular orbit, completing a full revolution. Throughout the entire sequence, the moon maintains this counter-clockwise motion relative to the Earth.

Qwen2 VL 72B instruct AWQ: A

Qwen2 VL 7B Instruct: A

Video-CCAM-v1.1 14B: Answer is A.

InternVL 2 40B: Based on the provided frames, the moon appears to rotate in a counter-clockwise direction relative to the Earth throughout the video.

Therefore, the most possible answer is:

****A: Counter-clockwise throughout.****

1836 I.2.2 EXAMPLE 2

1837

1838

1839

1840

1841

1842

1843

1844

1845

1846

1847

1848

1849

1850

1851

1852

1853

1854

1855

1856

1857

1858

1859

1860

1861

1862

1863

1864

1865

1866

1867

1868

1869

1870

1871

1872

1873

1874

1875

1876

1877

1878

1879

1880

1881

1882

1883

1884

1885

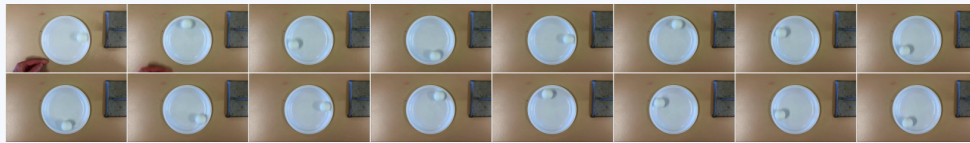
1886

1887

1888

1889

rotation 50 YouTube



https://www.youtube.com/watch?v=Ap06V_oetVI

What direction(s) does the Ping Pong ball rotate in?

- A. Clockwise throughout.
- B. No rotation.
- C. Clockwise then counter-clockwise.
- D. Counter-clockwise throughout.
- E. Counter-clockwise then clockwise.

Answer: D. Counter-clockwise throughout.

GPT-4o: Based on the provided frames, the ping pong ball appears to be rotating in a consistent direction. The ball moves around the plate in a circular motion, which suggests a rotational movement.

From the sequence of frames, it seems that **the ball is rotating in a clockwise direction throughout the video**. There is no indication of a change in the direction of rotation.

Therefore, the most likely answer is:

****Option A: Clockwise throughout.****

Claude 3.5 Sonnet: Based on analyzing the sequence of frames provided, the most likely answer to the question "What direction(s) does the Ping Pong ball rotate in?" is:

C: Clockwise then counter-clockwise.

The reasoning for this conclusion is:

1. In the initial frames, the ball appears to be rotating clockwise as it moves around the edge of the plate in a counter-clockwise direction.
2. **As the ball approaches the bottom of the plate, its rotation seems to slow down and then reverse.**
3. In the later frames, the ball appears to be rotating counter-clockwise as it continues its circular path around the plate.

This change in rotational direction is consistent with the physics of a ball rolling along a curved surface, where friction and the changing angle of contact can cause the ball to reverse its spin. The clockwise rotation followed by counter-clockwise rotation best matches the observed behavior of the ball in the provided image sequence.

Qwen2 VL 72B instruct AWQ: A

Qwen2 VL 7B Instruct: C

Video-CCAM-v1.1 14B: Answer is C.

1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943

InternVL 2 40B: Based on the provided frames, the Ping Pong ball appears to rotate in a clockwise direction throughout the video.

Answer: A

1944 I.2.3 EXAMPLE 3

1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997

rotation 57 YouTube



<https://www.youtube.com/watch?v=zji3Cm.k7go>

In which direction(s) does the person rotate on the horizontal bar?

- A. Clockwise then counter-clockwise.
- B. Clockwise throughout.
- C. Counter-clockwise throughout.
- D. No rotation.
- E. Counter-clockwise then clockwise.

Answer: B. Clockwise throughout.

GPT-4o: Based on the provided frames, the person appears to be **rotating in a counter-clockwise direction** throughout the sequence. The frames show the person starting from a seated position on the bar, moving through an inverted position, and then returning to the seated position, all in a counter-clockwise motion.

Therefore, the most likely answer is:

‘C’: ‘Counter-clockwise throughout.’

Claude 3.5 Sonnet: Based on the sequence of images provided, the answer to the question “In which direction(s) does the person rotate on the horizontal bar?” appears to be:

B: Clockwise throughout.

The images show the person starting in an upright seated position on the horizontal bar, then rotating forward (clockwise) through various positions including hanging upside down, continuing the rotation until returning to the starting position. The rotation seems to be consistently clockwise throughout the entire sequence of frames, without any counter-clockwise movement visible.

Qwen2 VL 72B instruct AWQ: A

Qwen2 VL 7B Instruct: A

Video-CCAM-v1.1 14B: Answer is C.

InternVL 2 40B: Based on the frames provided, the person on the horizontal bar appears to rotate in a counter-clockwise direction throughout the video.

Therefore, the most possible answer is:

‘C’: Counter-clockwise throughout.

1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051

I.2.4 EXAMPLE 4

rotation 58 YouTube



<https://www.youtube.com/watch?v=x1u0tX8mYJo>

In which direction(s) does the gymnastic athlete rotate on the horizontal bar?

- A. Clockwise throughout.
- B. Counter-clockwise then clockwise.
- C. Clockwise then counter-clockwise.
- D. Counter-clockwise throughout.
- E. No rotation.

Answer: A. Clockwise throughout.

GPT-4o: Based on the provided frames, the gymnastic athlete appears to be rotating in a counter-clockwise direction throughout the sequence. The athlete starts from a position above the bar and moves downward and backward, indicating a counter-clockwise rotation.

Therefore, the most likely answer is:

'D': 'Counter-clockwise throughout.'

Claude 3.5 Sonnet: Based on the sequence of images provided, the gymnast appears to be rotating counter-clockwise throughout the routine on the horizontal bar. The athlete starts in a position with their body extended away from the bar, then swings under and around the bar in a counter-clockwise direction. This counter-clockwise rotation continues through subsequent swings and movements shown in the frames.

Therefore, the most likely answer to the question "In which direction(s) does the gymnastic athlete rotate on the horizontal bar?" is:

D: Counter-clockwise throughout.

The images consistently show the gymnast moving in a counter-clockwise direction around the horizontal bar, from the initial swing through to the later parts of the routine captured in the frames.

Qwen2 VL 72B instruct AWQ: D

Qwen2 VL 7B Instruct: C

Video-CCAM-v1.1 14B: Answer is A.

InternVL 2 40B: Based on the frames provided, the gymnastic athlete appears to rotate in a counter-clockwise direction throughout the routine on the horizontal bar.

Therefore, the most possible answer is:

'D': Counter-clockwise throughout.

2052 I.2.5 EXAMPLE 5

2053

2054

rotation 144 YouTube

2055

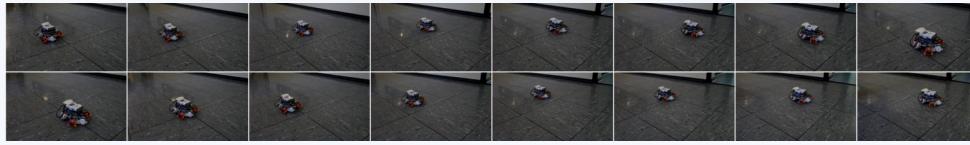
2056

2057

2058

2059

2060



2061

https:

2062

//www.youtube.com/results?search_query=omni+wheel&sp=EgIwAQ%253D%253D

2063

In which direction(s) does the car move?

2064

2065

A. No rotation.

2066

B. Counter-clockwise then clockwise.

2067

C. Counter-clockwise throughout.

2068

D. Clockwise throughout.

2069

E. Clockwise then counter-clockwise.

2070

2071

Answer: D. Clockwise throughout.

2072

2073

2074

GPT-4o: Based on the provided frames, the car appears to be rotating. The direction of rotation can be inferred from the position and orientation of the car in each frame.

2075

2076

From the first frame to the last frame, the car seems to be rotating in a counter-clockwise direction. There is no indication that the car changes its direction of rotation.

2077

2078

2079

Therefore, the most likely answer is:

2080

****C: 'Counter-clockwise throughout.'****

2081

2082

2083

Claude 3.5 Sonnet: Based on the series of images provided, I can observe a small robotic device on a tiled floor. The device appears to be rotating slightly between frames, but it's difficult to determine a consistent direction of rotation from these static images alone.

2084

2085

2086

Given the question and options provided, I will make my best guess based on the limited information available:

2087

2088

The most likely answer seems to be **C: Counter-clockwise throughout.**

2089

2090

This is because in some of the frames, there appears to be a slight counter-clockwise rotation of the device when comparing its orientation to the tile lines on the floor. However, I want to emphasize that this is an educated guess based on limited information, as the rotation is subtle and not definitively clear from these static images.

2091

2092

2093

2094

2095

Qwen2 VL 72B instruct AWQ: E

2096

2097

Qwen2 VL 7B Instruct: B

2098

2099

Video-CCAM-v1.1 14B: Answer is **C.**

2100

2101

InternVL 2 40B: Based on the provided frames, the car appears to move in a counter-clockwise direction throughout.

2102

2103

2104

Answer: C

2105

2106 I.2.6 EXAMPLE 6

2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159

rotation 64 YouTube



<https://www.youtube.com/watch?v=iaauRiRX4do&t=24s>

In what direction(s) does the person rotate?

- A. No rotation.
- B. Clockwise then counter-clockwise.
- C. Counter-clockwise then clockwise.
- D. Counter-clockwise throughout.
- E. Clockwise throughout.

Answer: E. Clockwise throughout.

GPT-4o: Based on the provided frames, the person appears to be **rotating in a counter-clockwise direction initially and then changes direction to rotate clockwise**. This can be inferred from the change in the orientation of the person's body and the direction they are facing.

Therefore, the most likely answer is:

****Option C: Counter-clockwise then clockwise.****

Claude 3.5 Sonnet: Based on analyzing the sequence of images provided, the answer to the question "In what direction(s) does the person rotate?" appears to be:

D: Counter-clockwise throughout.

The images show a person standing on a circular platform and holding a wheel-like object. As the sequence progresses, we can observe that the person's body orientation changes gradually in a counter-clockwise direction. The person starts facing away from the camera and slowly rotates to face different directions, **always turning to their left (counter-clockwise) throughout the sequence of frames**. There is no evidence of clockwise rotation or change in rotation direction in the provided images.

Qwen2 VL 72B instruct AWQ: B

Qwen2 VL 7B Instruct: C

Video-CCAM-v1.1 14B: Answer is D.

InternVL 2 40B: Based on the frames provided, the person **appears to rotate in a counter-clockwise direction** throughout the video.

Answer: D

2160 I.2.7 EXAMPLE 7

2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213

rotation 84 YouTube



<https://www.youtube.com/watch?v=01Htq-vrm6A>

In which direction(s) does the wheel rotate?

- A. Clockwise throughout.
- B. Counter-clockwise throughout.
- C. Counter-clockwise then clockwise.
- D. Clockwise then counter-clockwise.
- E. No rotation.

Answer: B. Counter-clockwise throughout.

GPT-4o: Based on the provided frames, the wheel appears to be rotating in a consistent direction as it moves over the rock. The movement of the wheel suggests that it is rotating in a clockwise direction throughout the sequence of frames.

Therefore, the most likely answer is:

****Option A: Clockwise throughout.****

Claude 3.5 Sonnet: Based on analyzing the sequence of images provided, my best assessment is that the wheel rotates clockwise throughout the video. The images show a Jeep tire on rocky terrain, and while it's difficult to see clear rotation in still frames, the progression suggests the wheel is moving forward over the rocks in a clockwise direction. Therefore, I would select option A: "Clockwise throughout."

The images don't show any obvious change in rotation direction, so the options involving counter-clockwise rotation or changes in direction seem less likely. And the wheel does appear to be moving rather than stationary, ruling out the "No rotation" option.

While there's some uncertainty given the limitations of still images, a clockwise rotation throughout seems the most probable answer based on the visual information provided in the frame sequence.

Qwen2 VL 72B instruct AWQ: A

Qwen2 VL 7B Instruct: D

Video-CCAM-v1.1 14B: Answer is A.

InternVL 2 40B: Based on the frames provided, it is not possible to determine the direction of the wheel's rotation. The frames show the wheel in different positions, but they do not provide enough information to determine the direction of rotation. Therefore, the most possible answer is:

'E': No rotation.

2214 I.2.8 EXAMPLE 8

2215

2216

2217

2218

2219

2220

2221

2222

2223

2224

2225

2226

2227

2228

2229

2230


2231

2232

2233

2234

rotation 74 YouTube



<https://www.youtube.com/watch?v=FJmRQ5iTXKE>

In which direction(s) does the rope rotate?

- A. Clockwise then counter-clockwise.
- B. Clockwise throughout.
- C. Counter-clockwise throughout.
- D. Counter-clockwise then clockwise.
- E. No rotation.

Answer: C. Counter-clockwise throughout.

2235

2236

2237

2238

2239

2240

2241

GPT-4o: Based on the provided frames, the rope appears to be rotating in a consistent direction throughout the sequence. The motion of the rope and the position of the person suggest a continuous rotation.

The most likely answer is:

'B': 'Clockwise throughout.'

2242

2243

2244

2245

2246

2247

2248

2249

2250

2251

2252

2253

2254

2255

2256

2257

Claude 3.5 Sonnet: Based on the sequence of images provided, I can see a shirtless man in gray pants and black shoes performing what appears to be a jump rope exercise. The rope itself is not clearly visible in most frames, but the man's arm and body movements are consistent with jump roping.

Regarding the specific question about the direction of rope rotation, it's difficult to determine with certainty from these still images. However, based on the typical mechanics of jump roping and the man's arm positions in the frames, I would make an educated guess that the rope is likely rotating:

C: Counter-clockwise throughout.

This is because in standard jump rope technique, the rope typically rotates forward over the head and under the feet in a counter-clockwise direction when viewed from the jumper's right side, which is the perspective we see in these images. The man's arm positions and slight forward lean are consistent with this standard counter-clockwise rotation.

However, I want to emphasize that this is an inference based on limited information from still images, and without seeing the actual motion of the rope, it's impossible to be 100% certain of the rotation direction.

2258

2259

Qwen2 VL 72B instruct AWQ: B

2260

2261

Qwen2 VL 7B Instruct: B

2262

2263

2264

Video-CCAM-v1.1 14B: Answer is B.

2265

2266

2267

InternVL 2 40B: Based on the provided frames, it is not possible to determine the exact direction of the rope's rotation. The frames do not provide a clear view of the rope's movement to confidently answer the question. Therefore, the most possible answer would be:

2268
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2320
2321

'E': No rotation.

2322 I.2.9 EXAMPLE 9

2323

2324

2325

2326

2327

2328

2329

2330

2331

2332

2333

2334

2335

2336

2337

2338

2339

2340

2341

2342

2343

2344

2345

2346

2347

2348

2349

2350

2351

2352

2353

2354

2355

2356

2357

2358

2359

2360

2361

2362

2363

2364

2365

2366

2367

2368

2369

2370

2371

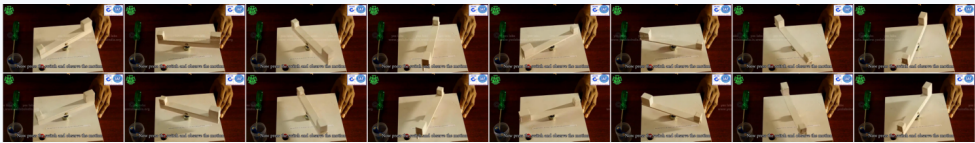
2372

2373

2374

2375

rotation 91 YouTube



<https://www.youtube.com/watch?v=mu9jAHgDjQg>

In which direction(s) does the object rotate?

- A. No rotation.
- B. Counter-clockwise throughout.
- C. Counter-clockwise then clockwise.
- D. Clockwise throughout.
- E. Clockwise then counter-clockwise.

Answer: D. Clockwise throughout.

GPT-4o: Based on the provided frames, the object appears to be rotating in a consistent direction. The frames show a gradual movement of the object in a circular path.

From the sequence of frames, it can be observed that the object is rotating in a counter-clockwise direction throughout the video.

Therefore, the most likely answer is:

****B: Counter-clockwise throughout.****

Claude 3.5 Sonnet: Based on the sequence of images provided, I can analyze the rotation of the white object on the circular platform. The object appears to be a rod with cube-shaped ends that is mounted on a central pivot point.

Looking at the progression of the images, we can observe that the rod rotates in a counter-clockwise direction initially. It starts from a horizontal position and gradually rotates counter-clockwise through various angles.

However, the sequence of images does not show a complete 360-degree rotation or any reversal in the direction of rotation. The rod is seen rotating counter-clockwise throughout the frames provided.

Given the options and the information available from these frames, the most likely answer to the question "In which direction(s) does the object rotate?" would be:

B: Counter-clockwise throughout.

This answer best matches the observed motion in the sequence of images, where the rod consistently rotates in a counter-clockwise direction without any apparent change in rotational direction.

Qwen2 VL 72B instruct AWQ: D

Qwen2 VL 7B Instruct: C

Video-CCAM-v1.1 14B: Answer is B.

2376
2377
2378
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2390
2391
2392
2393
2394
2395
2396
2397
2398
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429

InternVL 2 40B: Based on the provided frames, it appears that **the object rotates in a counter-clockwise direction** throughout the video.

Therefore, the most possible answer is:

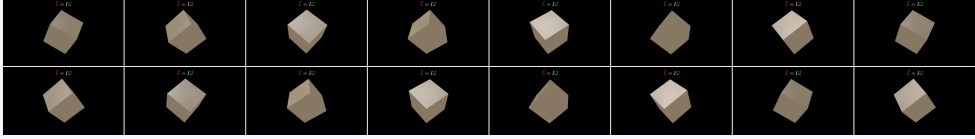
'B': Counter-clockwise throughout.

2430
2431
2432
2433
2434
2435
2436
2437
2438
2439
2440
2441
2442
2443
2444
2445
2446
2447
2448
2449
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2470
2471
2472
2473
2474
2475
2476
2477
2478
2479
2480
2481
2482
2483

I.3 SIMULATED

I.3.1 EXAMPLE 1

rotation 141 YouTube



<https://www.youtube.com/watch?v=fZYfX9VvN8M>

In which direction(s) does cube rotate?

A. Clockwise then counter-clockwise. B. Clockwise throughout.
C. Counter-clockwise then clockwise. D. No rotation.
E. Counter-clockwise throughout.

Answer: E. Counter-clockwise throughout.

GPT-4o: Based on the provided frames, the cube appears to be rotating in a consistent direction throughout the sequence. The rotation seems to be clockwise when observed from a fixed point.

Therefore, the most likely answer is:

'B': 'Clockwise throughout.'

Claude 3.5 Sonnet: Based on analyzing the sequence of 16 frames provided, the cube appears to rotate clockwise throughout the video. The cube starts in one orientation and gradually rotates in a clockwise direction across the frames, without changing direction. Therefore, the most likely answer to the question "In which direction(s) does cube rotate?" is:

B: Clockwise throughout.

The cube maintains a consistent clockwise rotation from the beginning to the end of the frame sequence, without any apparent reversal or change in rotational direction.

Qwen2 VL 72B instruct AWQ: B

Qwen2 VL 7B Instruct: A

Video-CCAM-v1.1 14B: Answer is B

InternVL 2 40B: Based on the provided frames, it appears that the cube does not exhibit any significant rotation. The cube remains in a consistent orientation throughout the frames.


Therefore, the most possible answer is:

'D': 'No rotation.'

2484 I.3.2 EXAMPLE 2

2485
2486
2487
2488
2489
2490
2491
2492
2493
2494
2495
2496
2497
2498
2499
2500
2501
2502
2503
2504
2505
2506
2507
2508
2509
2510
2511
2512
2513
2514
2515
2516
2517
2518
2519
2520
2521
2522
2523
2524
2525
2526
2527
2528
2529
2530
2531
2532
2533
2534
2535
2536
2537

rotation 180 YouTube



<https://www.youtube.com/watch?v=JEJqy1Wloww>

In which direction(s) does the minute hand move on the clock face?

- A. Clockwise then counter-clockwise.
- B. Clockwise throughout.
- C. No rotation.
- D. Counter-clockwise throughout.
- E. Counter-clockwise then clockwise.

Answer: E. Counter-clockwise then clockwise.

GPT-4o: Based on the provided frames, the minute hand of the clock appears to move in a clockwise direction throughout the sequence. There is **no indication of the minute hand moving counter-clockwise in any of the frames.**

Therefore, the most possible answer is:

'B': 'Clockwise throughout.'

Claude 3.5 Sonnet: Based on analyzing the sequence of clock images provided, the answer to the question "In which direction(s) does the minute hand move on the clock face?" is:

B. Clockwise throughout.

The images show the minute hand of the clock moving in a clockwise direction over time, starting near 12 o'clock and progressing around the clock face in a clockwise manner through various positions. There is **no evidence of the minute hand moving counter-clockwise** at any point in the sequence of images. The minute hand consistently moves clockwise as time advances across the frames.

Qwen2 VL 72B instruct AWQ: B

Qwen2 VL 7B Instruct: B

Video-CCAM-v1.1 14B: Answer is **B.**

InternVL 2 40B: Based on the provided frames, the minute hand on the clock face appears to move in a clockwise direction throughout. Therefore, the most possible answer is:

'B': 'Clockwise throughout.'

2538 I.3.3 EXAMPLE 3

2539

2540

2541

2542

2543

2544

2545

2546

2547

2548

2549

2550

2551

2552

2553

2554

2555

2556

2557

2558

2559

2560

2561

2562

2563

2564

2565

2566

2567

2568

2569

2570

2571

2572

2573

2574

2575

2576

2577

2578

2579

2580

2581

2582

2583

2584

2585

2586

2587


2588

2589

2590

2591

rotation 62



In what direction(s) does the minute hand move on the clock face?

- A. Clockwise throughout.
- B. Counter-clockwise then clockwise.
- C. No rotation.
- D. Counter-clockwise throughout.
- E. Clockwise then counter-clockwise.

Answer: D. Counter-clockwise throughout.

GPT-4o: Based on the provided frames, the minute hand of the clock does not appear to move at all. It remains at the 12 o'clock position throughout all the frames.

Given the options, the most appropriate answer is:

'C': 'No rotation.'

Claude 3.5 Sonnet: Based on analyzing the sequence of clock images provided, the answer to the question "In what direction(s) does the minute hand move on the clock face?" is:

A: Clockwise throughout.

The minute hand (the longer red hand) can be observed moving in a clockwise direction across the series of images, starting near 12 o'clock and progressing around the clock face in a clockwise manner. There is no evidence of the minute hand moving counter-clockwise at any point in the sequence.

Qwen2 VL 72B instruct AWQ: A

Qwen2 VL 7B Instruct: D

Video-CCAM-v1.1 14B: Answer is A.

InternVL 2 40B: Based on the provided frames, the minute hand on the clock face appears to move in a clockwise direction throughout.

Therefore, the most possible answer is:

'A': 'Clockwise throughout.'

2592 I.3.4 EXAMPLE 4

2593

2594

2595

2596

2597

2598

2599

2600

2601

2602

2603

2604

2605

2606

2607

2608

2609

2610

2611

2612

2613

2614

2615

2616

2617

2618

2619

2620

2621

2622

2623

2624

2625

2626

2627

2628

2629

2630

2631

2632

2633

2634

2635

2636

2637

2638

2639

2640

2641

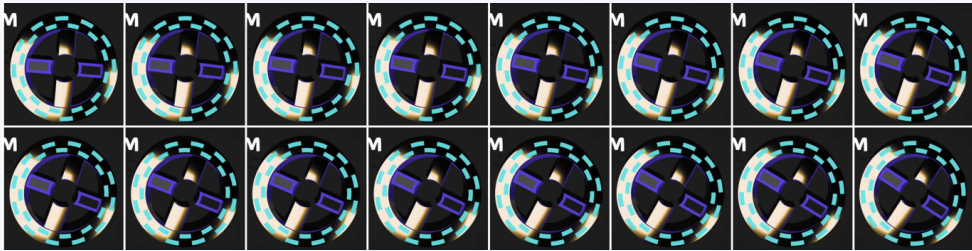
2642

2643

2644

2645

rotation 103 YouTube



<https://www.youtube.com/watch?v=R0xqQ0HNdgc>

In what direction(s) is the wheel rotating?

- A. Counter-clockwise throughout.
- B. Counter-clockwise then clockwise.
- C. Clockwise then counter-clockwise.
- D. Clockwise throughout.
- E. No rotation.

Answer: D. Clockwise throughout.

GPT-4o: Based on the provided frames, it appears that **the wheel is rotating in a counter-clockwise direction** initially and then **changes to a clockwise direction**. This can be inferred from the position of the colored segments and the central cross in each frame.

Therefore, the most likely answer is:

****Option B: Counter-clockwise then clockwise.****

Claude 3.5 Sonnet: Based on the analysis of the 16 frames provided, the most likely answer to the question "In what direction(s) is the wheel rotating?" is:

A: Counter-clockwise throughout.

Reasoning:

1. The wheel in the images consists of a central black circle with four purple rectangular spokes and an outer ring with alternating cyan and black segments.
2. Observing the position of the spokes across the frames, **there appears to be a slight but consistent counter-clockwise movement.**
3. The cyan segments on the outer ring also seem to be moving in a counter-clockwise direction relative to the fixed "M" letter in the top-left corner of each frame.
4. There is no apparent change in the direction of rotation throughout the sequence of frames.

While the movement is subtle and the frames are static, the overall impression from analyzing the sequence suggests a counter-clockwise rotation. Therefore, option A: "Counter-clockwise throughout" seems to be the most appropriate answer based on the given information.

Qwen2 VL 72B instruct AWQ: A

Qwen2 VL 7B Instruct: D

2646
2647
2648
2649
2650
2651
2652
2653
2654
2655
2656
2657
2658
2659
2660
2661
2662
2663
2664
2665
2666
2667
2668
2669
2670
2671
2672
2673
2674
2675
2676
2677
2678
2679
2680
2681
2682
2683
2684
2685
2686
2687
2688
2689
2690
2691
2692
2693
2694
2695
2696
2697
2698
2699

Video-CCAM-v1.1 14B: Answer is **A**.

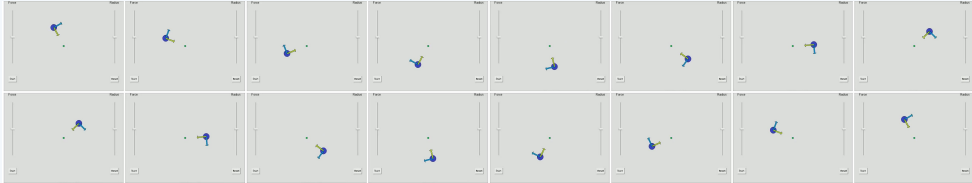
InternVL 2 40B: Based on the provided frames, it is not possible to determine the direction of the wheel's rotation. **The frames do not show a clear indication of the wheel's movement or rotation.** Therefore, the most possible answer is:

'E': No rotation.

2700 I.3.5 EXAMPLE 5

2701
2702
2703
2704
2705
2706
2707
2708
2709
2710
2711
2712
2713
2714
2715
2716
2717
2718
2719
2720
2721
2722
2723
2724
2725
2726
2727
2728
2729
2730
2731
2732
2733
2734
2735
2736
2737
2738
2739
2740
2741
2742
2743
2744
2745
2746
2747
2748
2749
2750
2751
2752
2753

rotation 206 YouTube



<https://www.youtube.com/watch?v=bxhdxLerP9Q>

What direction(s) is the blue circle moving in?

A. Clockwise then counter-clockwise. B. Clockwise throughout.
C. Counter-clockwise then clockwise. D. Counter-clockwise throughout.
E. No rotation.

Answer: C. Counter-clockwise then clockwise.

GPT-4o: Based on the provided frames, the blue circle appears to be rotating in a counter-clockwise direction throughout the sequence. The position of the blue circle relative to the green center point and the orientation of the attached yellow and blue arms indicate a consistent counter-clockwise motion.

Therefore, the most likely answer is:

'D': 'Counter-clockwise throughout.'

Claude 3.5 Sonnet: Based on the sequence of images provided, the blue circle appears to be rotating in a counter-clockwise direction throughout the video. The yellow and blue arms attached to the blue circle consistently rotate counter-clockwise from frame to frame.

Starting from the first image and progressing through the sequence, we can observe that the arms of the blue circle move in a counter-clockwise manner. There is no evidence of the circle changing its direction of rotation at any point in the given frames.

Therefore, the most likely answer to the question "What direction(s) is the blue circle moving in?" is:

D: Counter-clockwise throughout.

Qwen2 VL 72B instruct AWQ: D

Qwen2 VL 7B Instruct: C

Video-CCAM-v1.1 14B: Answer is B.

InternVL 2 40B: Based on the provided frames, the blue circle appears to be moving in a clockwise direction throughout the video.

Answer: B

2754 J COMMON FAILURE CASES OF DIRECTION

2755

2756 J.1 HUMAN

2757

2758 J.1.1 EXAMPLE 1

2759

2760

2761

2762

2763

2764

2765

2766

2767

2768

2769

2770

2771

2772

2773

2774

2775

2776

2777

2778

2779

2780

2781

2782

2783

2784

2785

2786

2787

2788

2789

2790

2791

2792

2793

2794

2795

2796

2797

2798

2799

2800

2801

2802

2803


2804

2805

2806

2807

direction 6



In which direction(s) did the person's hand move?

- A. First to the right then to the left.
- B. First to the left then to the right.
- C. Right.
- D. Not moving at all
- E. Left.

Answer: E. Left

GPT-4o: Based on the provided frames, the person's hand starts pointing to the right and then moves to the left.

Therefore, the most possible answer is:

****B:** 'First to the left then to the right.'**

Claude 3.5 Sonnet: Based on the sequence of images provided, the answer to the question "In which direction(s) did the person's hand move?" is:

C. Right.

The images show the person starting with their hand at their side, then moving it outward to their right and holding it there. There is no clear movement of the hand to the left or back and forth. The hand consistently moves and stays to the right side of the person's body throughout the sequence of frames.

Qwen2 VL 72B instruct AWQ: A

Qwen2 VL 7B Instruct: C

Video-CCAM-v1.1 14B: Answer is C.

2808
2809
2810
2811
2812
2813
2814
2815
2816
2817
2818
2819
2820
2821
2822
2823
2824
2825
2826
2827
2828
2829
2830
2831
2832
2833
2834
2835
2836
2837
2838
2839
2840
2841
2842
2843
2844
2845
2846
2847
2848
2849
2850
2851
2852
2853
2854
2855
2856
2857
2858
2859
2860
2861

InternVL 2 40B: Based on the frames provided, the person's hand appears to move from right to left.
The correct answer is:
'A': 'First to the right then to the left.'

2862 J.1.2 EXAMPLE 2

2863

2864

2865

2866

2867

2868

2869

2870

2871

2872

2873

2874

2875

2876

2877

2878

2879

2880

2881

2882

2883

2884

2885

2886

2887

2888

2889

2890

2891

2892

2893

2894

2895

2896

2897

2898

2899

2900

2901

2902

2903

2904

2905

2906

2907

2908

2909

2910

2911


2912

2913

2914

2915

direction 7



In which direction(s) did the person's hand move?

- A. Right.
- B. First to the right then to the left.
- C. Left.
- D. Not moving at all
- E. First to the left then to the right.

Answer: A. Right.

GPT-4o: Based on the provided frames, the person's hand starts in an upright position and then moves to the left. The hand then returns to the right. This indicates a movement to the left and then back to the right.

Therefore, the most likely answer is:

****Option E: First to the left then to the right.****

Claude 3.5 Sonnet: Based on analyzing the sequence of images provided, the answer to the question "In which direction(s) did the person's hand move?" appears to be:

B: First to the right then to the left.

The images show the person's hand starting in a raised position on their right side, then moving across their body to the left, and then back towards the right. This back-and-forth motion is consistent with option B, which states the hand moved right then left.

Qwen2 VL 72B instruct AWQ: E

Qwen2 VL 7B Instruct: The person's hand moved from the left side to the right side of their body. Therefore, the answer is:

'E': 'First to the left then to the right.'

Video-CCAM-v1.1 14B: Answer is B.

2916
2917
2918
2919
2920
2921
2922
2923
2924
2925
2926
2927
2928
2929
2930
2931
2932
2933
2934
2935
2936
2937
2938
2939
2940
2941
2942
2943
2944
2945
2946
2947
2948
2949
2950
2951
2952
2953
2954
2955
2956
2957
2958
2959
2960
2961
2962
2963
2964
2965
2966
2967
2968
2969

InternVL 2 40B: Based on the frames provided, the person's hand appears to move from left to right.

The most possible answer is:

'E': First to the left then to the right.

2970 J.1.3 EXAMPLE 3

2971

2972

2973

2974

2975

2976

2977

2978

2979

2980

2981

2982

2983

2984

2985

2986

2987

2988

2989

2990

2991

2992

2993

2994

2995

2996

2997

2998

2999

3000

3001

3002

3003

3004

3005

3006

3007

3008

3009

3010

3011

3012

3013

3014

3015

3016

3017

3018

3019

3020

3021

3022

3023

direction 118



What instruction did the person give to the camera?

- A. Move to the right.
- B. Move closer.
- C. Stay still.
- D. Move to the left.
- E. Move further.

Answer: E. Move further.

GPT-4o: Based on the sequence of frames, it appears that the person is making hand gestures that could be interpreted as giving instructions to the camera. **The person raises their hands and moves them slightly forward, which could be interpreted as a signal for the camera to move closer.**

Therefore, the most likely answer based on the frames provided is:

****B**: 'Move closer.'****

Claude 3.5 Sonnet: Based on analyzing the sequence of images provided, the most likely instruction the person gave to the camera appears to be:

C: Stay still.

The reasoning for this conclusion is:

1. The person remains in the same position throughout all 16 frames, standing on the steps in front of the door.
2. The background and surroundings do not change, indicating the camera is not moving.
3. While the person's arms and hands move in various gestures, their overall body position and location stay constant.
4. **There are no indications of the camera moving closer, further away, or to either side.**
5. The consistent framing and stable background strongly suggest the camera remained stationary while capturing these images.

Therefore, the instruction to "Stay still" seems to be the most plausible option given the evidence in the image sequence provided.

Qwen2 VL 72B instruct AWQ: C

Qwen2 VL 7B Instruct: C

Video-CCAM-v1.1 14B: Answer is **B**.

3024
3025
3026
3027
3028
3029
3030
3031
3032
3033
3034
3035
3036
3037
3038
3039
3040
3041
3042
3043
3044
3045
3046
3047
3048
3049
3050
3051
3052
3053
3054
3055
3056
3057
3058
3059
3060
3061
3062
3063
3064
3065
3066
3067
3068
3069
3070
3071
3072
3073
3074
3075
3076
3077

InternVL 2 40B: Based on the frames provided, the person appears to be standing in front of a door, with their hands raised in front of their face. **There is no clear indication of any specific instruction being given to the camera.** However, since the person is stationary and not making any gestures that suggest movement, the most plausible answer would be:

'C': Stay still.

3078 J.2 OBJECT

3079

3080 J.2.1 EXAMPLE 1

3081

3082 direction 242 YouTube

3083

3084

3085

3086

3087

3088

3089

3090

3091

3092

3093

3094

3095

3096

3097

3098



<https://www.youtube.com/watch?v=ygtref12Hks>

In which direction(s) did the drone move during the video?

- A. Leftwards then upwards. B. Rightwards then upwards.
C. Downwards then leftwards. D. Rightwards then downwards.
E. Leftwards then downwards. F. Rightwards then leftwards.

Answer: F. Rightwards then leftwards.

3099

3100

3101

3102

3103

3104

3105

GPT-4o: Based on the provided frames, the drone appears to be **moving rightwards initially and then upwards**. This can be inferred from the drone's position relative to the person and the table. In the first few frames, the drone is moving to the right, and in the later frames, it appears to be rising.

Therefore, the most likely answer is:

****B': Rightwards then upwards.****

3106

3107

3108

3109

3110

3111

3112

3113

3114

3115

3116

3117

3118

3119

3120

3121

3122

3123

3124

3125

3126

3127

Claude 3.5 Sonnet: Based on analyzing the sequence of images provided, the most likely answer to the question "In which direction(s) did the drone move during the video?" is:

D. Rightwards then downwards.

The reasoning for this conclusion is:

1. In the initial frames, the drone appears to be hovering in a relatively stable position on the table in front of the person.
2. As the sequence progresses, there are subtle changes in the drone's position relative to the person's hands and the table edge, suggesting some movement.
3. The drone seems to shift slightly to the right (from the viewer's perspective) in the middle frames.
4. **In the later frames, the drone appears to be lower in position compared to earlier frames, suggesting a downward movement.**
5. The person's hand gestures and gaze direction also provide cues that they are following the drone's movement to the right and then down.

While it's difficult to determine precise movements from still images, the subtle changes in positioning across the frame sequence suggest a rightward movement followed by a downward movement as the most plausible interpretation of the drone's motion.

3128

3129

3130

3131

Qwen2 VL 72B instruct AWQ: B

Qwen2 VL 7B Instruct: B

3132
3133
3134
3135
3136
3137
3138
3139
3140
3141
3142
3143
3144
3145
3146
3147
3148
3149
3150
3151
3152
3153
3154
3155
3156
3157
3158
3159
3160
3161
3162
3163
3164
3165
3166
3167
3168
3169
3170
3171
3172
3173
3174
3175
3176
3177
3178
3179
3180
3181
3182
3183
3184
3185

Video-CCAM-v1.1 14B: Answer is **B**.

InternVL 2 40B: Based on the frames provided, the drone appears to move in the following directions:

1. Initially, the drone is seen moving towards the left side of the frame.
2. *After moving leftwards, the drone is seen ascending upwards.*

Therefore, the most likely answer is:

****A: Leftwards then upwards.****

3186 J.2.2 EXAMPLE 2

3187

3188

3189

3190

3191

3192

3193

3194

3195

3196

3197

3198

3199

3200

3201

3202

3203

3204

3205

3206

3207

3208

3209

3210

3211

3212

3213

3214

3215

3216

3217

3218

3219

3220

3221

3222

3223

3224

3225

3226

3227

3228

3229

3230

3231

3232

3233

3234

3235

3236

3237

3238

3239

direction 269 YouTube



<https://www.youtube.com/watch?v=-Rtd4ZyKZy0>

What directional shifts can be observed in the helicopter's motion from the footage?

- A. Upwards then downwards. B. Leftwards then rightwards.
C. Leftwards then upwards. D. Downwards then rightwards.
E. Rightwards then upwards. F. Rightwards then downwards.

Answer: F. Rightwards then downwards.

GPT-4o: Based on the provided frames, the helicopter appears to be initially in the air and then gradually descends to land. The sequence of frames shows the helicopter moving downwards and slightly to the right as it approaches the ground.

Therefore, the most plausible answer is:

****Option D: Downwards then rightwards.****

Claude 3.5 Sonnet: Based on the sequence of images provided, the most likely answer to the question "What directional shifts can be observed in the helicopter's motion from the footage?" is:

E. Rightwards then upwards.

The images show a yellow ambulance helicopter initially on the ground. In the later frames, the helicopter appears to be lifting off vertically. While it's difficult to discern clear horizontal movement from still images, there seems to be a slight rightward shift in the helicopter's position relative to the background buildings as it begins to rise.

The upward motion is more clearly evident in the later frames where the helicopter is seen hovering above the ground. The rightward shift, while subtle, can be inferred from the changing perspective of the helicopter against the stationary background elements.

It's important to note that without seeing the actual video footage, this analysis is based on the limited information provided by the still frames. The rightward movement is less obvious than the upward motion, but option E best fits the observed changes in the helicopter's position across the image sequence.

Qwen2 VL 72B instruct AWQ: A

Qwen2 VL 7B Instruct: A

Video-CCAM-v1.1 14B: Answer is B.

InternVL 2 40B: The helicopter in the video appears to be moving in a leftwards direction initially and then upwards. Based on the frames provided, the most likely directional shift observed in the helicopter's motion is:

****C: Leftwards then upwards.****

3240 J.2.3 EXAMPLE 3

3241

3242

3243

3244

3245

3246

3247

3248

3249

3250

3251

3252

3253

3254

3255

3256

3257

3258

3259

3260

3261

3262

3263

3264

3265

3266

3267

3268

3269

3270

3271

3272

3273

3274

3275

3276

3277

3278

3279

3280

3281

3282

3283

3284

3285

3286

3287

3288

3289

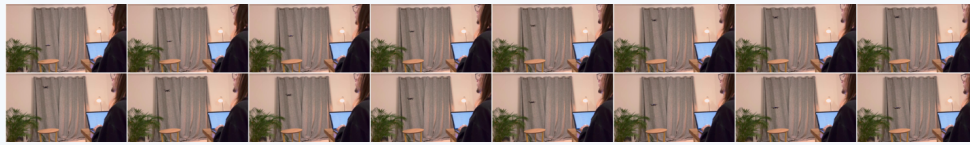
3290

3291

3292

3293

direction 243 YouTube



<https://www.youtube.com/watch?v=ygtreF12Hks>

In which direction(s) did the drone move during the video?

- A. Downwards then rightwards.
- B. Upwards then downwards.
- C. Rightwards then upwards.
- D. Rightwards then leftwards.
- E. Leftwards then upwards.
- F. Rightwards then downwards.

Answer: B. Upwards then downwards.

GPT-4o: Based on the provided frames, the drone's movement can be analyzed as follows:

1. The drone starts near the center of the frame.
2. It moves slightly to the left and upwards.
3. It continues moving upwards and to the left.
4. It reaches the top left corner.
5. It then starts moving downwards.
6. It continues moving downwards.
7. It moves back towards the center.
8. It continues moving downwards and slightly to the right.
9. It moves further downwards.
10. It continues moving downwards and slightly to the right.
11. It moves back towards the center.
12. It continues moving downwards.
13. It moves further downwards.
14. It continues moving downwards and slightly to the right.
15. It moves back towards the center.
16. It continues moving downwards and slightly to the right.

From this analysis, the drone initially moves leftwards and upwards, then moves downwards and slightly to the right.

The most fitting option based on the frames is:

'E': 'Leftwards then upwards.'

Claude 3.5 Sonnet: Based on analyzing the sequence of frames provided, the most likely answer to the question "In which direction(s) did the drone move during the video?" is:

C. Rightwards then upwards.

The reasoning for this conclusion is:

1. In the initial frames, the drone appears as a small blurry object near the center of the curtain.
2. As the frames progress, the drone's position shifts gradually towards the right side of the curtain, indicating rightward movement.
3. In the later frames, the drone's position appears to move higher up on the curtain, suggesting upward movement.

3294
3295
3296
3297
3298
3299
3300
3301
3302
3303
3304
3305
3306
3307
3308
3309
3310
3311
3312
3313
3314
3315
3316
3317
3318
3319
3320
3321
3322
3323
3324
3325
3326
3327
3328
3329
3330
3331
3332
3333
3334
3335
3336
3337
3338
3339
3340
3341
3342
3343
3344
3345
3346
3347

4. The final frames show the drone positioned higher and more to the right compared to its initial location.

This rightward then upward movement pattern best matches option C from the given choices. While there may be some subtle variations or brief movements in other directions, the overall trend visible across the frame sequence supports this answer as the most probable description of the drone's primary movement directions.

Qwen2 VL 72B instruct AWQ: B

Qwen2 VL 7B Instruct: C

Video-CCAM-v1.1 14B: Answer is F.

InternVL 2 40B: Based on the frames provided, the drone appears to move in the direction of rightwards then downwards.

Therefore, the most possible answer is:

'F': Rightwards then downwards.

3348 J.2.4 EXAMPLE 4

3349

3350

3351

3352

3353

3354

3355

3356

3357

3358

3359

3360

3361

3362

3363

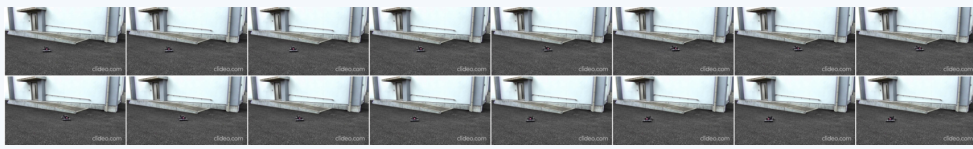
3364

3365

3366

3367

direction 277 YouTube



<https://www.youtube.com/watch?v=bNq60yuTRf8>

What was the directional trend of the toy car's movement throughout the clip?

A. Upwards then rightwards. B. Downwards then leftwards.
C. Leftwards then rightwards. D. Rightwards then leftwards.
E. Leftwards then upwards. F. Downwards then upwards.

Answer: D. Rightwards then leftwards.

3368

3369

3370

3371

3372

3373

3374

3375

3376

3377

3378

GPT-4o: Based on the provided frames, the toy car appears to be moving in a generally leftward direction initially and then changes direction to move rightward.

Here is the analysis of the frames:

1. The car starts from a position closer to the right side of the frame.
2. It moves leftward in the subsequent frames.
3. Around the middle frames, the car starts to change direction.
4. In the later frames, the car appears to be moving rightward.

Given this observation, the most plausible answer is:

****Option C: Leftwards then rightwards.****

3379

3380

3381

3382

3383

3384

3385

3386

3387

3388

3389

3390

3391

3392

3393

3394

3395

3396

3397

3398

3399

3400

3401

Claude 3.5 Sonnet: Based on analyzing the 16 frames provided, the most likely answer to the question "What was the directional trend of the toy car's movement throughout the clip?" is:

C: Leftwards then rightwards.

Reasoning:

1. In the initial frames, the toy car appears to be positioned more towards the right side of the image.
2. As the sequence progresses, the car seems to move slightly towards the left, getting closer to the center of the frame.
3. In the later frames, there's a subtle shift where the car appears to move back towards the right side of the image.

While the movement is not dramatic, there does seem to be a slight leftward trend followed by a rightward trend. This makes option C the most plausible answer among the given choices, based solely on the information provided in these static frames.

It's important to note that the movement is very subtle, and without seeing the actual video, it's challenging to be absolutely certain. However, given the instruction to make a best guess based on the available information, option C appears to be the most fitting answer.

Qwen2 VL 72B instruct AWQ: D

Qwen2 VL 7B Instruct: C

Video-CCAM-v1.1 14B: Answer is B.

3402
3403
3404
3405
3406
3407
3408
3409
3410
3411
3412
3413
3414
3415
3416
3417
3418
3419
3420
3421
3422
3423
3424
3425
3426
3427
3428
3429
3430
3431
3432
3433
3434
3435
3436
3437
3438
3439
3440
3441
3442
3443
3444
3445
3446
3447
3448
3449
3450
3451
3452
3453
3454
3455

InternVL 2 40B: Based on the frames provided, the toy car appears to move in a leftward direction initially and then turns rightward. Therefore, the most possible answer is:

'C': Leftwards then rightwards.

3456 J.2.5 EXAMPLE 5

3457

3458

3459

3460

3461

3462

3463

3464

3465

3466

3467

3468

3469

3470

3471

3472

3473

3474

3475

3476

3477

3478

3479

3480

3481

3482

3483

3484

3485

3486

3487

3488

3489

3490

3491

3492

3493

3494

3495

3496

3497

3498

3499

3500

3501

3502

3503

3504

3505

3506

3507

3508

3509

direction 349 YouTube



<https://www.youtube.com/watch?v=8u0tTaCYqQ>

In what direction did the car move?

- A. Downwards. B. Rightwards. C. Upwards.
D. No movements. E. Leftwards.

Answer: E. Leftwards.

GPT-4o: Based on the sequence of frames provided, the car appears to be moving from the left side of the frame to the right side of the frame. This is evident from the car's initial position on the left and its final position on the right.

Therefore, the most likely answer is:

****Option B: Rightwards.****

Claude 3.5 Sonnet: Based on the sequence of images provided, the answer to the question "In what direction did the car move?" is:

B. Rightwards.

The frames show a red car moving across a grassy field with trees in the background. As the sequence progresses, we can see the car consistently moving from left to right in the frame. The car is kicking up dust and at times even becoming airborne as it travels across the uneven terrain, but its overall direction of movement is clearly rightwards through the field.

Qwen2 VL 72B instruct AWQ: B

Qwen2 VL 7B Instruct: B

Video-CCAM-v1.1 14B: The car is moving to the right. So the answer is rightwards.

InternVL 2 40B: Based on the frames provided, the car appears to be moving upwards, as it is shown in various stages of a jump. Therefore, the most possible answer is:

'C': Upwards.

3510
3511
3512
3513
3514
3515
3516
3517
3518
3519
3520
3521
3522
3523
3524
3525
3526
3527
3528
3529
3530
3531
3532
3533
3534
3535
3536
3537
3538
3539
3540
3541
3542
3543
3544
3545
3546
3547
3548
3549
3550
3551
3552
3553
3554
3555
3556
3557
3558
3559
3560
3561
3562
3563

J.2.6 EXAMPLE 6

direction 294 YouTube



<https://www.youtube.com/watch?v=UgXVVV72dp8>

In what direction did the camera move?

A. Backward B. No movement C. Forward

Answer: A. Backward

GPT-4o: Based on the provided frames, it appears that **the camera is moving forward**. The person in the black outfit and the dog ahead of them are getting closer in each subsequent frame, indicating forward movement.

Therefore, the most likely answer is:

****Option C: Forward****

Claude 3.5 Sonnet: Based on analyzing the sequence of images provided, the most likely answer to the question "In what direction did the camera move?" is:

C: Forward

The images show a consistent perspective of following behind a person walking on a forest path with a dog. **The camera appears to be moving forward along the path**, maintaining a similar distance behind the person throughout the sequence. This gives the impression of forward movement of the camera along with the walking subjects.

Qwen2 VL 72B instruct AWQ: C

Qwen2 VL 7B Instruct: C

Video-CCAM-v1.1 14B: Answer is C.

InternVL 2 40B: Based on the frames provided, it appears that the camera is moving backward. The perspective shows a clear view of the path and the people walking away from the camera, indicating that the camera is moving in the opposite direction of the subjects.

Therefore, the most possible answer is:

'A': 'Backward'

3564 J.3 SIMULATED

3565

3566 J.3.1 EXAMPLE 1

3567

3568

3569

3570

3571

3572

3573

3574

3575

3576

3577

3578

3579

3580

3581

3582

3583

3584

3585

3586

3587

3588

3589

3590

3591

3592

3593

3594

3595

3596

3597

3598

3599

3600

3601

3602

3603

3604

3605

3606

3607

3608

3609

3610

3611

3612

3613

3614

3615

3616

3617

direction 235

What is the direction of the red diamond's movement from the video?

A. Leftwards then rightwards. B. Rightwards then downwards.
C. Downwards then rightwards. D. Rightwards then upwards.
E. Downwards then upwards. F. Upwards then downwards.

Answer: F. Upwards then downwards.

GPT-4o: Based on the provided frames, the red diamond starts at the bottom of the frame and moves upwards. It then moves slightly to the right as it continues to move upwards.

Therefore, the most likely answer is:

****Option D: Rightwards then upwards.****

Claude 3.5 Sonnet: Based on the 16 frames provided, I cannot determine any movement of the red diamond shape. The diamond appears to be in the same position in the center of each frame, without any apparent change in its location or orientation.

However, since I'm instructed to make a guess even if there's not enough information, I'll choose the option that seems most plausible given the question:

D: Rightwards then upwards.

This is purely a guess based on the question asking about movement, even though no movement is visible in the provided frames. I want to emphasize that this answer is not based on any observed motion in the images, but is simply an arbitrary selection from the given options.

Qwen2 VL 72B instruct AWQ: E

Qwen2 VL 7B Instruct: B

Video-CCAM-v1.1 14B: Answer is D.

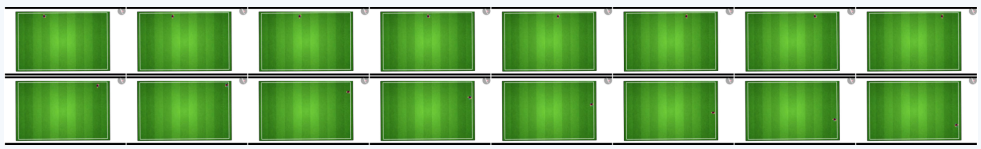
InternVL 2 40B: Based on the frames provided, the red diamond appears to move upwards and then downwards. Therefore, the most possible answer is:

****E: Downwards then upwards****

3618
3619
3620
3621
3622
3623
3624
3625
3626
3627
3628
3629
3630
3631
3632
3633
3634
3635
3636
3637
3638
3639
3640
3641
3642
3643
3644
3645
3646
3647
3648
3649
3650
3651
3652
3653
3654
3655
3656
3657
3658
3659
3660
3661
3662
3663
3664
3665
3666
3667
3668
3669
3670
3671

J.3.2 EXAMPLE 2

direction 238 YouTube



<https://www.youtube.com/watch?v=COHi3qWME0M&t=285s>

What is the direction of the game character's movement from the video?

A. Downwards then leftwards. B. Leftwards then rightwards.
C. Downwards then rightwards. D. Downwards then upwards.
E. Leftwards then downwards. F. Rightwards then downwards.

Answer: F. Rightwards then downwards.

GPT-4o: Based on the provided frames, the game character starts at the top center of the field and moves downwards. As the frames progress, the character then moves towards the right side of the field.

Therefore, the direction of the game character's movement from the video is:

Option 'C': Downwards then rightwards.

Claude 3.5 Sonnet: Based on analyzing the sequence of frames provided, the direction of the game character's movement appears to be:

C: Downwards then rightwards

The character starts near the top of the green field in the early frames, then moves downward towards the center. In the later frames, the character can be seen moving towards the right side of the field. This downward then rightward movement pattern best matches option C from the given choices.

Qwen2 VL 72B instruct AWQ: C

Qwen2 VL 7B Instruct: The direction of the game character's movement from the video is downwards then rightwards.

Video-CCAM-v1.1 14B: Answer is E.

InternVL 2 40B: Based on the frames provided, the game character appears to be moving downwards and then rightwards.

Therefore, the correct answer is:

****C: Downwards then rightwards****

3672 J.3.3 EXAMPLE 3

3673

3674

3675

3676

3677

3678

3679

3680

3681

3682

3683

3684

3685

3686

3687

3688

3689

direction 361

How did the blue text block progress?

A. Upwards. B. Downwards. C. Leftwards.
D. Rightwards. E. No movements.

Answer: B. Downwards.

3690

3691

3692

3693

3694

3695

3696

GPT-4o: Based on the provided frames, the blue text block appears to be moving upwards. In the initial frames, the block is positioned higher on the screen, and as the frames progress, it moves downwards. This indicates that the block is moving upwards in the video.

Therefore, the most possible answer is:

****A': Upwards.****

3697

3698

3699

3700

3701

3702

3703

3704

3705

3706

3707

3708

Claude 3.5 Sonnet: Based on the 16 frames provided, I can answer the question about how the blue text block progressed.

The answer is **A: Upwards.**

In all 16 frames, we see a blue rectangular block containing the text "The block is moving upwards". The block remains in the same position in the center of each frame. While we don't see the block itself moving upwards across the frames, the text within the block explicitly states that it is moving upwards.

Given that we are instructed to use only the information provided in the frames to answer the question, and not rely on external knowledge, the most logical conclusion based on the text in the blue block is that the block is progressing upwards, even though we don't visually see this movement in the static frames provided.

3709

3710

3711

3712

3713

3714

3715

Qwen2 VL 72B instruct AWQ: B

Qwen2 VL 7B Instruct: A

Video-CCAM-v1.1 14B: Answer is A

3716

3717

3718

3719

3720

3721

3722

3723

3724

3725

InternVL 2 40B: The blue text block in the video appears to be moving upwards. This can be inferred from the consistent position of the text block in the frames provided, which shows it progressively moving upwards.

Therefore, the correct answer is:

A: Upwards.

3726 J.3.4 EXAMPLE 4

3727

3728

3729

3730

3731

3732

3733

3734

3735

3736

3737

3738

3739

3740

3741

3742

3743

3744

3745

3746

3747

3748

3749

3750

3751

3752

3753

3754

3755

3756

3757

3758

3759

3760

3761

3762

3763

3764

3765

3766

3767

3768

3769

3770

3771

3772

3773

3774

3775

3776

3777

3778

3779

direction 233

What is the direction of the black sweater icon's movement from the video ?

A. Leftwards then rightwards. B. Leftwards then downwards.
C. Downwards then leftwards. D. Downwards then upwards.
E. Upwards then rightwards. F. Downwards then rightwards

Answer: B. Leftwards then downwards.

GPT-4o: Based on the provided frames, the black sweater icon starts at the top right of the frame and moves downwards and then to the left.

Therefore, the most likely answer is:

****Option C: Downwards then leftwards.****

Claude 3.5 Sonnet: Based on the sequence of frames provided, the black sweater icon appears to redmove in a diagonal direction from the bottom right corner towards the top left corner of the image. The movement seems to be primarily upwards and leftwards.

Given the options provided, the closest match to this observed movement would be:

E: Upwards then rightwards.

While this option is not a perfect description of the movement seen in the frames (as the icon moves leftwards rather than rightwards), it is the closest match among the given choices. The upward movement is correctly captured, and it's the only option that includes an upward direction.

It's important to note that this answer is based solely on the information provided in the image frames and the given options, as instructed. The movement is not entirely clear from the limited frames, but this appears to be the best guess based on the available information and options.

Qwen2 VL 72B instruct AWQ: B

Qwen2 VL 7B Instruct: The direction of the black sweater icon's movement from the video is downwards then rightwards.

Video-CCAM-v1.1 14B: Answer is D.

InternVL 2 40B: The black sweater icon in the video moves from the top-right corner to the bottom-left corner. This movement can be described as moving downwards and then leftwards.

Therefore, the correct answer is:

****C: Downwards then leftwards****

3780 J.3.5 EXAMPLE 5

3781

3782

3783

3784

3785

3786

3787

3788

3789

3790

3791

3792

3793

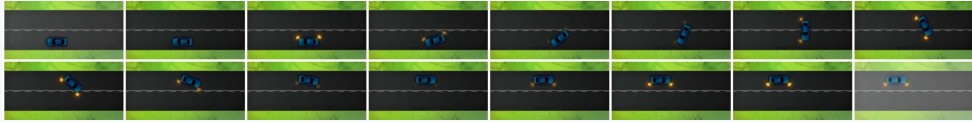
3794

3795

3796

3797

direction 278 YouTube



<https://www.youtube.com/watch?v=YfGIoAim17w>

What was the directional trend of the car's movement throughout the clip?

A. Downwards then leftwards. B. Upwards then downwards.
C. Downwards then upwards. D. Rightwards then leftwards.
E. Leftwards then downwards. F. Downwards then rightwards.

Answer: D. Rightwards then leftwards.

3798

3799

3800

3801

3802

3803

3804

GPT-4o: Based on the provided frames, **the car initially moves downwards and then starts to turn right.** This is evident from the car's position and the activation of the right turn signal lights.

Therefore, the most likely answer is:

Option F: Downwards then rightwards

3805

3806

3807

3808

3809

3810

3811

3812

3813

3814

3815

3816

3817

Claude 3.5 Sonnet: Based on the sequence of frames provided, the most likely answer to the question "What was the directional trend of the car's movement throughout the clip?" is:

F: Downwards then rightwards

Explanation:

1. In the initial frames, the car appears to be moving straight along the road.
2. As the sequence progresses, the car starts to tilt, suggesting a change in direction.
3. **The tilting of the car becomes more pronounced, with its front end pointing downwards and to the right.**
4. In the later frames, the car seems to straighten out again, but it has moved to a position further right on the road compared to its starting position.

This sequence of movements suggests that the car first moved downwards (tilting forward) and then rightwards, making option F the most plausible answer among the given choices.

3818

3819

Qwen2 VL 72B instruct AWQ: C

3820

3821

Qwen2 VL 7B Instruct: C

3822

3823

Video-CCAM-v1.1 14B: Answer is **F**.

3824

3825

3826

3827

3828

3829

3830

3831

3832

3833

InternVL 2 40B: The directional trend of the car's movement throughout the clip can be observed by analyzing the positions of the car in the frames provided.

1. In the first frame, the car is positioned at the bottom of the road.
2. In subsequent frames, the car moves upwards along the road.
3. Finally, the car is seen moving downwards towards the bottom of the road again.

Based on this observation, the most accurate answer is:

'B': Upwards then downwards.

3834 J.3.6 EXAMPLE 6

3835

3836

3837

3838

3839

3840

3841

3842

3843

3844

3845

3846

3847

3848

3849

3850

3851

3852

3853

3854

3855

3856

3857

3858

3859

3860

3861

3862

3863

3864

3865

3866

3867

3868

3869

3870

3871

3872

3873

3874

3875

3876

3877

3878

3879

3880

3881

3882

3883

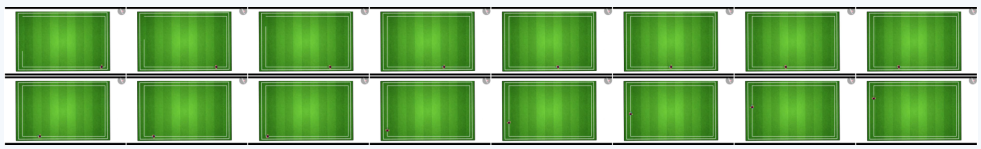
3884

3885

3886

3887

direction 239 YouTube



<https://www.youtube.com/watch?v=COHi3qWME0M&t=285s>

What is the direction of the game character's movement from the video?

A. Leftwards then rightwards. B. Leftwards then upwards.
C. Downwards then rightwards. D. Downwards then leftwards.
E. Upwards then rightwards. F. Leftwards then downwards.

Answer: B. Leftwards then upwards.

GPT-4o: Based on the provided frames, the game character starts at the bottom right corner and moves leftwards along the bottom edge of the field. The character then turns and moves upwards along the left edge of the field.

Therefore, the direction of the game character's movement is:

'B': Leftwards then upwards.

Claude 3.5 Sonnet: Based on analyzing the sequence of frames provided, the direction of the game character's movement appears to be:

A. Leftwards then rightwards.

The character starts in the bottom right corner of the green field. It then moves leftwards along the bottom edge of the field. After reaching the bottom left corner, the character changes direction and moves rightwards back along the bottom edge. This leftwards then rightwards movement pattern matches option A from the given choices.

Qwen2 VL 72B instruct AWQ: D

Qwen2 VL 7B Instruct: The direction of the game character's movement from the video is 'Leftwards then downwards.'

Video-CCAM-v1.1 14B: Answer is F.

InternVL 2 40B: Based on the frames provided, the game character appears to move from the bottom left corner to the bottom right corner. This suggests that the character is moving downwards and then to the right.

Therefore, the most possible answer is:


'D': Downwards then leftwards

3888 K COMMON FAILURE CASES OF VELOCITY & FREQUENCY

3889
3890 K.1 HUMAN

3891
3892 K.1.1 EXAMPLE 1

3893
3894 velocity&frequency 3



3895
3896
3897
3898
3899
3900
3901
3902
3903
3904
3905
3906
3907
3908
3909

3910 What is the pattern of the person's rotation speed?

3911

3912 A. Decelerating. B. Accelerating.

3913 C. Constant speed. D. Not moving at all.

3914

3915 Answer: A. Decelerating.

3916
3917

3918 **GPT-4o:** Based on the provided frames, the person appears to be rotating. The sequence of frames shows the person turning from facing away from the camera to facing the camera and then back to facing away again. **The rotation seems to be happening at a consistent pace throughout the frames.**

3919
3920

3921 Given the options:

3922 {'A': 'Decelerating.', 'B': 'Accelerating.', 'C': 'Constant speed.', 'D': 'Not moving at all.'}

3923

3924 The most plausible answer based on the frames is:

3925 **'C': 'Constant speed.'**

3926

3927

3928 **Claude 3.5 Sonnet:** Based on the provided frames, I cannot definitively determine the pattern of the person's rotation speed. The images show the person in various static poses, facing different directions, but do not capture continuous motion or rotation. However, since I'm asked to make a best guess based on the available information, I would choose:

3929
3930
3931

3932 **C: Constant speed.**

3933

3934 This is because the frames show the person in different orientations that appear to be evenly spaced (front view, side view, back view, etc.), which could **suggest a constant rotation if we assume the frames are sampled at regular intervals.** However, this is purely speculative given the limited information in the static images.

3935
3936
3937

3938 **Qwen2 VL 72B instruct AWQ: C**

3939

3940 **Qwen2 VL 7B Instruct: C**

3941

3942
3943
3944
3945
3946
3947
3948
3949
3950
3951
3952
3953
3954
3955
3956
3957
3958
3959
3960
3961
3962
3963
3964
3965
3966
3967
3968
3969
3970
3971
3972
3973
3974
3975
3976
3977
3978
3979
3980
3981
3982
3983
3984
3985
3986
3987
3988
3989
3990
3991
3992
3993
3994
3995

Video-CCAM-v1.1 14B: Answer is **C**.

InternVL 2 40B: Based on the provided frames, the person appears to be standing in different positions, facing different directions. **There is no clear indication of rotation or movement in the frames.** Therefore, the most possible answer is:

D: Not moving at all.

3996 K.2 OBJECT

3997

3998 K.2.1 EXAMPLE 1

3999

4000

velocity&frequency 42 YouTube

4001

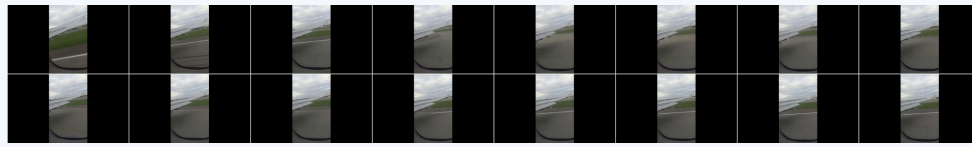
4002

4003

4004

4005

4006



4007

<https://www.youtube.com/watch?v=uwwzofTsMXc>

4008

4009 What is the pattern of the plane's speed on the runway?

4010

4011

- 4012 A. Accelerating. B. Decelerating.
C. Not moving at all. D. Constant speed.

4013

4014

4015 Answer: B. Decelerating.

4016

4017

4018

4019

GPT-4o: Based on the 16 frames provided, the plane appears to be moving on the runway. The background and the position of the wing relative to the ground change slightly between frames, indicating motion. However, **the changes are not drastic**, suggesting that the plane is not accelerating rapidly.

4020

4021

4022

4023 Given the subtle changes in the frames, the most plausible answer is that the plane is moving at a constant speed.

4024

4025

4026 Therefore, the answer is:

4027

4028 ****D: Constant speed.****

4029

4030

4031

4032

Claude 3.5 Sonnet: Based on analyzing the 16 frames provided, the most likely answer to the question "What is the pattern of the plane's speed on the runway?" is:

4033

4034 **A: Accelerating.**

4035

4036

4037 While it's difficult to determine the exact speed from still images, there are a few clues that suggest the plane is likely accelerating:

4038

4039

4040

4041

4042

- 4043 1. The frames show a consistent view of the runway and surrounding area, indicating the plane is moving forward along the runway.
- 4044 2. In typical takeoff procedures, planes accelerate down the runway to reach the necessary speed for liftoff.
- 4045 3. The scenery appears to be moving past the window at an increasing rate in the later frames, though this is subtle.
- 4046 4. **There are no signs of deceleration or stopping**, which would be expected if the plane was slowing down or maintaining a constant speed.

4047

4048

4049

4050 Given these observations and the typical behavior of planes during takeoff, acceleration seems to be the most probable pattern of speed for the plane in this sequence of images. However, I want to emphasize that this is an educated guess based on limited information from still frames, and without more dynamic data, it's impossible to state this with absolute certainty.

4051

4052

Qwen2 VL 72B instruct AWQ: A

4053

4050
4051
4052
4053
4054
4055
4056
4057
4058
4059
4060
4061
4062
4063
4064
4065
4066
4067
4068
4069
4070
4071
4072
4073
4074
4075
4076
4077
4078
4079
4080
4081
4082
4083
4084
4085
4086
4087
4088
4089
4090
4091
4092
4093
4094
4095
4096
4097
4098
4099
4100
4101
4102
4103

Qwen2 VL 7B Instruct: [A](#)

Video-CCAM-v1.1 14B: Answer is [A](#).


InternVL 2 40B: Based on the frames provided, it appears that the plane is moving along the runway. The frames show the plane's wing and the runway in the background, indicating that the plane is in motion. Therefore, the most possible answer is:

[A: Accelerating.](#)

4104 K.2.2 EXAMPLE 2

4105
4106
4107
4108
4109
4110
4111
4112
4113
4114
4115
4116
4117
4118
4119
4120
4121
4122
4123
4124
4125
4126
4127
4128
4129
4130
4131
4132
4133
4134
4135
4136
4137
4138
4139
4140
4141
4142
4143
4144
4145
4146
4147
4148
4149
4150
4151
4152
4153
4154
4155
4156
4157

velocity&frequency 46 YouTube



<https://www.youtube.com/watch?v=dmqkEmdE5nE>

What is the pattern of the roller coaster's speed on the rail?

A. Accelerating. B. Not moving at all.
C. Decelerating. D. Constant speed.

Answer: C. Decelerating.

GPT-4o: Based on the provided frames, the roller coaster appears to be ascending a hill. The frames show a consistent upward movement along the track, which suggests that the roller coaster is being pulled up by a chain lift mechanism. **This typically means the roller coaster is moving at a constant speed during the ascent.**

Given this information, the most likely answer is:

****Option D: Constant speed.****

Claude 3.5 Sonnet: Based on the series of images provided, which show a first-person view from a roller coaster car ascending a steep incline, I would conclude that the most likely answer to the question "What is the pattern of the roller coaster's speed on the rail?" is:

A: Accelerating.

The images consistently show the roller coaster car climbing up a steep incline. As the car moves up this initial hill, it is likely being pulled by a chain lift or similar mechanism. **This would result in a steady acceleration as the car gains both height and speed while ascending the track.**

While we cannot see the actual speed directly, the consistent upward motion and the gradually changing perspective of the track and surrounding structures suggest an increasing velocity. This acceleration is typical for the initial climb of many roller coasters, building up potential energy that will be converted to kinetic energy for the rest of the ride.

Qwen2 VL 72B instruct AWQ: A

Qwen2 VL 7B Instruct: A

Video-CCAM-v1.1 14B: Answer is A.

InternVL 2 40B: Based on the frames provided, **the roller coaster appears to be stationary** at the top of the track. Therefore, the most possible answer is:

B: Not moving at all.

4158 K.2.3 EXAMPLE 3

4159

4160

4161

4162

4163

4164

4165

4166

4167

4168

4169

4170

4171

4172

4173

4174

4175

4176

4177

4178

4179

4180

4181

4182

4183

4184

4185

4186

4187

4188

4189

4190

4191

4192

4193

4194

4195

4196

4197

4198

4199

4200

4201

4202

4203

4204

4205

4206

4207


4208

4209

4210

4211

velocity&frequency 44 YouTube



<https://www.youtube.com/watch?v=eqTZ74aXa68>

What is the pattern of the plane's speed on the runway?

A. Decelerating. B. Not moving at all.
C. Constant speed. D. Accelerating.

Answer: A. Decelerating.

GPT-4o: Based on the provided frames, the plane's position relative to the background and the runway markings does not change significantly. This suggests that the plane is not moving or is moving very slowly.

Given the options:
{ 'A': 'Decelerating.', 'B': 'Not moving at all.', 'C': 'Constant speed.', 'D': 'Accelerating.' }

The most plausible answer based on the frames is:
'B': 'Not moving at all.'

Claude 3.5 Sonnet: Based on the series of images provided, I would say the most likely answer to the question "What is the pattern of the plane's speed on the runway?" is:
C: Constant speed.

The images show a Southwest Airlines plane on a runway, but there is no clear indication of acceleration or deceleration. The plane appears to be in a similar position relative to the background in most frames, suggesting it is moving at a steady pace. While we can't determine the exact speed, the visual evidence points to the plane maintaining a constant speed as it moves along the runway.

Qwen2 VL 72B instruct AWQ: D

Qwen2 VL 7B Instruct: D

Video-CCAM-v1.1 14B: Answer is A.

InternVL 2 40B: Based on the frames provided, it appears that the plane is moving steadily along the runway. There is no visible change in the plane's position relative to the background, suggesting a constant speed.

Therefore, the most possible answer is:
C: Constant speed.

4212 K.2.4 EXAMPLE 4

4213

4214

4215

4216

4217

4218

4219

4220

4221

4222

4223

4224

4225

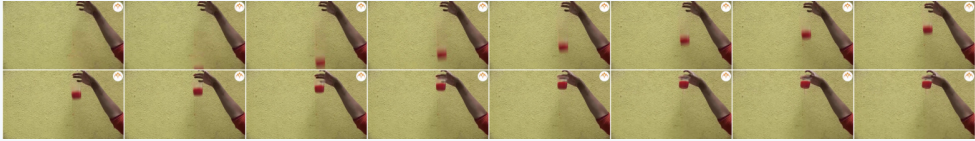
4226

4227

4228

4229

velocity&frequency 66 YouTube



<https://www.youtube.com/watch?v=vpQQJTyoKE>

What is the pattern of the object's speed in the video?

A. Not moving at all. B. Constant speed.
C. Decelerating. D. Accelerating.

Answer: C. Decelerating.

4230

4231

4232

4233

4234

4235

4236

4237

4238

4239

4240

GPT-4o: Based on the provided frames, it appears that the object (a cup with liquid) is **initially moving downward and then upward**. The motion suggests that the object is being dropped and then caught or swung back up. The blur in the images indicates motion, and the distance between the object and the hand changes over time.

Given the sequence of frames, the object seems to be accelerating as it moves downward due to gravity and then decelerating as it is caught or swung back up.

Therefore, the most plausible answer based on the frames is:

****Option D: Accelerating.****

4241

4242

4243

4244

4245

4246

4247

4248

4249

4250

4251

4252

4253

4254

4255

4256

4257

4258

4259

4260

4261

Claude 3.5 Sonnet: Based on analyzing the sequence of frames provided, the pattern of the object's speed in the video appears to be:

D. Accelerating.

The reasoning for this conclusion is:

1. In the initial frames, the red object (which appears to be a small container with liquid) is not visible or barely visible at the bottom of the frame.
2. As the sequence progresses, the object becomes more visible and moves upward in the frame.
3. **The rate at which the object moves upward seems to increase throughout the sequence of frames.**
4. In the final frames, the object is clearly visible near the top of the frame, and appears to be moving more rapidly compared to earlier frames.
5. The blurring of the object in later frames also suggests increased speed.

This pattern of movement, starting slow and gradually increasing in speed as it moves upward, is consistent with acceleration. The object appears to be gaining speed throughout the video sequence, rather than moving at a constant speed or slowing down.

4262

4263

4264

4265

Qwen2 VL 72B instruct AWQ: D

Qwen2 VL 7B Instruct: D

4266
4267
4268
4269
4270
4271
4272
4273
4274
4275
4276
4277
4278
4279
4280
4281
4282
4283
4284
4285
4286
4287
4288
4289
4290
4291
4292
4293
4294
4295
4296
4297
4298
4299
4300
4301
4302
4303
4304
4305
4306
4307
4308
4309
4310
4311
4312
4313
4314
4315
4316
4317
4318
4319

Video-CCAM-v1.1 14B: Answer is C.

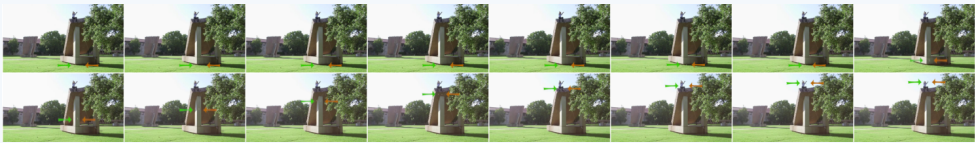
InternVL 2 40B: Based on the frames provided, the object appears to be moving downward in a consistent manner. There is **no indication of sudden changes in speed or direction**. Therefore, the most likely pattern of the object's speed in the video is:

'B': Constant speed.

4320 K.2.5 EXAMPLE 5

4321
4322
4323
4324
4325
4326
4327
4328
4329
4330
4331
4332
4333
4334
4335
4336
4337
4338
4339
4340
4341
4342
4343
4344
4345
4346
4347
4348
4349
4350
4351
4352
4353
4354
4355
4356
4357
4358
4359
4360
4361
4362
4363
4364
4365
4366
4367
4368
4369
4370
4371
4372
4373

velocity&frequency 72 YouTube



https://www.youtube.com/watch?v=AiGM4lrpA_k

What is balls' speed pattern?

A. Decelerating. B. Not moving at all.
C. Accelerating. D. Constant speed.

Answer: A. Decelerating.

GPT-4o: Based on the provided frames, the ball appears to be moving from the left side of the frame to the right side. The distance covered by the ball in each subsequent frame seems to be increasing, indicating that the ball is accelerating.

Therefore, the most likely answer is:

****C': Accelerating.****

Claude 3.5 Sonnet: Based on the sequence of images provided, I can analyze the movement of the green and orange arrows to infer the ball's speed pattern. The arrows appear to represent the ball's motion, with the green arrow indicating forward movement and the orange arrow indicating backward movement.

Throughout the sequence, the relative sizes of the green and orange arrows remain consistent. This suggests that the ball is moving back and forth at a steady rate, without any noticeable acceleration or deceleration.

Given this observation, the most likely answer to the question "What is balls' speed pattern?" is:

D: Constant speed.

The consistent size and positioning of the arrows across the frames indicate that the ball is moving at a constant speed, oscillating back and forth without changing its velocity.

Qwen2 VL 72B instruct AWQ: D

Qwen2 VL 7B Instruct: C

Video-CCAM-v1.1 14B: Answer is A.

InternVL 2 40B: Based on the provided frames, it is not possible to determine the speed pattern of the balls as there are no visible balls in the frames. The frames show a person standing on a structure in a park-like setting. Therefore, the most possible answer would be:

'B': Not moving at all.

4374 K.2.6 EXAMPLE 6

4375

4376

4377

4378

4379

4380

4381

4382

4383

4384

4385

4386

4387

4388

4389

4390

4391

4392

4393

4394

4395

4396

4397

4398

4399

4400

4401

4402

4403

4404

4405

4406

4407

4408

4409

4410

4411

4412

4413

4414

4415

4416

4417

4418

4419

4420

4421

4422

4423

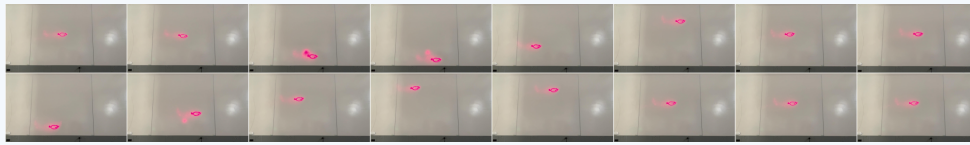
4424

4425

4426

4427

velocity&frequency 135



How does the frequency of the laser pointer movement change?

- A. Decelerating. B. Not moving at all.
C. Accelerating. D. Constant speed.

Answer: A. Decelerating.

GPT-4o: Based on the provided frames, the laser pointer appears to be moving in a circular pattern. The distance between the laser pointer's positions in consecutive frames seems to be increasing, indicating that the speed of the laser pointer is increasing over time.

Therefore, the most likely answer is:

****'C': Accelerating.****

Claude 3.5 Sonnet: Based on analyzing the 16 frames provided, I would say the most likely answer is **C: Accelerating**.

The images show a pink laser pointer drawing of what appears to be a cartoon fish or animal face on a white wall or board. While the fish shape itself remains fairly constant, the blurred trail behind it seems to change in length and intensity across the frames.

In the earlier frames, the blurred trail is relatively short and faint. As the sequence progresses, the trail becomes longer and more pronounced in many of the frames. This suggests the laser pointer is being moved more quickly over time, leaving a more prominent trail.

The increasing length and intensity of the blur trails implies the laser pointer's movement is speeding up or accelerating throughout the video sequence. While there is some variation between individual frames, the overall trend points towards faster movement and longer trails in later frames compared to earlier ones.

Therefore, based solely on the visual information provided in these 16 sample frames, the most likely answer to how the frequency of laser pointer movement changes is **C: Accelerating**.

Qwen2 VL 72B instruct AWQ: C

Qwen2 VL 7B Instruct: C

Video-CCAM-v1.1 14B: Answer is A.

InternVL 2 40B: Based on the provided frames, the laser pointer appears to move in a consistent manner across the whiteboard. There is no significant change in the speed or direction of the movement. Therefore, the most likely answer is:

'D': Constant speed.

4428 K.2.7 EXAMPLE 7

4429

4430

velocity&frequency 136 YouTube

4431

4432

4433

4434

4435

4436

4437



4438

<https://www.youtube.com/watch?v=QA1ZAMtxbBA>

4439

4440

What is the pattern of the pendulum's movement speed throughout the entire video?

4441

4442

A. Not moving at all. B. Constant speed.

4443

C. Accelerating. D. Decelerating.

4444

4445

Answer: C. Accelerating.

4446

4447

4448

GPT-4o: Based on the provided frames, it appears that the pendulum is moving. The pendulum's position changes slightly from frame to frame, indicating movement. However, the changes in position are subtle and do not suggest a significant acceleration or deceleration.

4449

4450

4451

Given the uniformity of the movement across the frames, the most plausible answer is that the pendulum is moving at a constant speed.

4452

4453

4454

Therefore, the answer is:

4455

****B': Constant speed.****

4456

4457

4458

Claude 3.5 Sonnet: Based on the provided frames, I cannot determine the pattern of a pendulum's movement speed, as there is no pendulum visible in these images. The frames show an aerial view of what appears to be a room or hall with furniture and seating.

4459

4460

4461

However, since I'm instructed to provide a guess based on the given options, I would choose:

4462

B: Constant speed.

4463

4464

This is purely a guess and not based on any visual evidence from the provided frames. In reality, there is no information about a pendulum or its movement in these images.

4465

4466

4467

Qwen2 VL 72B instruct AWQ: D

4468

Qwen2 VL 7B Instruct: C

4469

4470

Video-CCAM-v1.1 14B: Answer is B.

4471

4472

InternVL 2 40B: Based on the provided frames, it is not possible to determine the pattern of the pendulum's movement speed throughout the entire video. The frames do not show any pendulum or its movement. Therefore, the most possible answer is:

4473

4474

4475

4476

'NOT POSSIBLE TO DETERMINE.'

4477

4478

4479

4480

4481

4482 K.2.8 EXAMPLE 8

4483

4484

4485

4486

4487

4488

4489

4490

4491

4492

4493

4494

4495

4496

4497

4498

4499

4500

4501

4502

4503

4504

4505

4506

4507

4508

4509

4510

4511

4512

4513

4514

4515

4516

4517

4518

4519

4520

4521

4522

4523

4524

4525

4526

4527

4528

4529

4530

4531

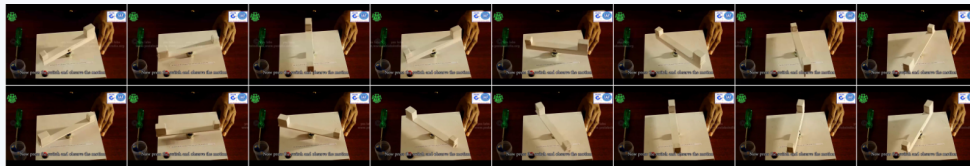
4532

4533

4534

4535

velocity&frequency 139 YouTube



<https://www.youtube.com/watch?v=mu9jAHgDjQg>

What is the object's rotational speed pattern?

- A. Decelerating. B. Not moving at all.
C. Accelerating. D. Constant speed.

Answer: A. Decelerating.

GPT-4o: Based on the 16 frames provided, the object appears to be **rotating at a constant speed**. The positions of the object in each frame suggest a uniform angular displacement between consecutive frames, indicating that the rotational speed is not changing.

Therefore, the most likely answer is:

****Option D: Constant speed.****

Claude 3.5 Sonnet: Based on the sequence of images provided, I can observe a rotating arm on a white platform. The arm **appears to be rotating at a constant speed throughout the frames**. There are no clear indications of acceleration or deceleration in the motion of the arm across the sequence of images.

Given the question "What is the object's rotational speed pattern?" and the available options, the most likely answer based solely on the visual information in these frames is:

D: Constant speed.

The arm seems to maintain a steady rotational movement without any apparent changes in speed from frame to frame. However, it's important to note that making a definitive determination about speed patterns from still images can be challenging, especially without precise timing information between frames.

Qwen2 VL 72B instruct AWQ: D

Qwen2 VL 7B Instruct: D

Video-CCAM-v1.1 14B: Answer is A.

InternVL 2 40B: Based on the provided frames, **it is not possible to determine the exact rotational speed pattern of the object**. However, we can observe that the object is moving in a circular path. Given this information, the most possible answer would be:

'D': Constant speed.

4536 K.2.9 EXAMPLE 9

4537

4538

4539

4540

4541

4542

4543

4544

4545

4546

4547

4548

4549

4550


4551

4552

4553

4554

velocity&frequency 140 YouTube



<https://www.youtube.com/watch?v=sqk4SNdnjRk>

What is the object's rotational speed pattern?

A. Not moving at all. B. Decelerating.
C. Constant speed. D. Accelerating.

Answer: D. Accelerating.

4555

4556

4557

4558

4559

4560

GPT-4o: Based on the provided frames, the object **appears to be rotating at a constant speed**. The position of the object in each frame suggests that the rotational speed does not change significantly over time.

Therefore, the most likely answer is:

'C': 'Constant speed.'

4561

4562

4563

4564

4565

4566

4567

4568

4569

4570

4571

4572

4573

Claude 3.5 Sonnet: used on the series of images provided, I can see a person demonstrating something with a pink object, likely some kind of toy or educational prop. The person is consistently holding and manipulating this object throughout the frames. However, from these static images alone, I cannot definitively determine the rotational speed pattern of the object.

Given the question and available options, I will make my best guess based on the limited information:

The most likely answer seems to be **C: Constant speed**.

My reasoning is that **the person's hand positions and the object's appearance remain fairly consistent across the frames**, suggesting a steady, continuous motion rather than acceleration, deceleration, or no movement at all. However, I want to emphasize that this is an educated guess based on limited information from static images, and may not reflect the actual motion in the video.

4574

4575

4576

4577

4578

4579

4580

4581

4582

4583

4584

4585

4586

4587

4588

4589

Qwen2 VL 72B instruct AWQ: C

Qwen2 VL 7B Instruct: D

Video-CCAM-v1.1 14B: Answer is C.

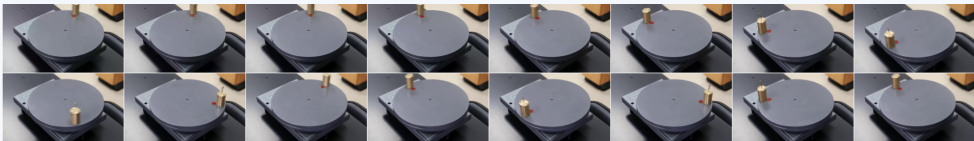
InternVL 2 40B: Based on the provided frames, it appears that the object is a pink object being held by a person. **The object does not seem to be moving or rotating in any of the frames**. Therefore, the most possible answer to the question about the object's rotational speed pattern is:

A: Not moving at all.

4590
4591
4592
4593
4594
4595
4596
4597
4598
4599
4600
4601
4602
4603
4604
4605
4606
4607
4608
4609
4610
4611
4612
4613
4614
4615
4616
4617
4618
4619
4620
4621
4622
4623
4624
4625
4626
4627
4628
4629
4630
4631
4632
4633
4634
4635
4636
4637
4638
4639
4640
4641
4642
4643

K.2.10 EXAMPLE 10

velocity&frequency 151 YouTube



<https://www.youtube.com/watch?v=V00VtEjHqjw>

What is the disk's rotational speed pattern?

A. Constant speed. B. Not moving at all.
C. Decelerating. D. Accelerating.

Answer: D. Accelerating.

GPT-4o: Based on the provided frames, it appears that the disk is slowing down over time. The position of the red mark on the disk changes less between each frame as the sequence progresses, indicating a decrease in rotational speed.

Therefore, the most likely answer is:

Option C: Decelerating.

Claude 3.5 Sonnet: Based on analyzing the sequence of 16 frames provided, the disk's rotational speed pattern appears to be accelerating. In the initial frames, the golden cylindrical object on the disk is stationary or moving slowly. As the sequence progresses, the motion of this object becomes more pronounced and blurred, suggesting the disk is rotating faster. By the final frames, the golden object is significantly blurred due to rapid rotation, indicating an increase in speed over time.

Therefore, the most likely answer to the question "What is the disk's rotational speed pattern?" is:

D: Accelerating.

This conclusion is drawn solely from observing the changing position and blur of the golden object on the disk across the provided image sequence, without relying on any external knowledge.

Qwen2 VL 72B instruct AWQ: A

Qwen2 VL 7B Instruct: A

Video-CCAM-v1.1 14B: Answer is A.

InternVL 2 40B: Based on the provided frames, the disk appears to be rotating at a constant speed. The position of the object on the disk remains consistent throughout the frames, indicating that the disk is not accelerating, decelerating, or stopping. Therefore, the most likely answer is:

A: Constant speed.

4644 K.2.11 EXAMPLE 11

4645

4646

4647

4648

4649

4650

4651

4652

4653

4654

4655

4656

4657


4658

4659

4660

4661

velocity&frequency 157 YouTube



<https://www.youtube.com/watch?v=CP5k2gRabvc>

What is the trend in the camera's movement speed across the entire video?

A. Not moving at all. B. Decelerating.
C. Constant speed. D. Accelerating.

Answer: B. Decelerating.

4662

4663

4664

4665

4666

4667

4668

GPT-4o: Based on the 16 frames provided, it appears that the camera is not moving at all. The position of the zebras, trees, and other elements in the scene remain consistent across all frames, indicating that the camera is stationary.

Therefore, the most likely answer is:

'A': 'Not moving at all.'

4669

4670

4671

4672

4673

4674

4675

4676

4677

4678

4679

4680

4681

4682

4683

4684

4685

4686

4687

4688

4689

4690

4691

4692

4693

4694

4695

4696

4697

Claude 3.5 Sonnet: Based on analyzing the 16 frames provided, the most likely answer to the question "What is the trend in the camera's movement speed across the entire video?" is:

C: Constant speed.

The reasoning for this conclusion is:

1. The overall composition of the scene remains very consistent across all 16 frames. The positioning of key elements like the trees, zebras, and background landscape stays nearly identical.
2. There is no noticeable change in the level of detail or apparent size of objects in the foreground or background, which would typically occur if the camera was accelerating or decelerating significantly.
3. The zebras' positions relative to each other and to fixed elements like trees show only minor changes between frames, consistent with their natural movement rather than rapid camera motion.
4. The clarity and focus of the images remain constant, suggesting steady camera movement rather than acceleration which might cause motion blur.
5. There are no obvious signs of zooming in or out, panning, or other camera movements that would indicate changing speed.

While it's difficult to be absolutely certain without seeing the actual video, the evidence from these static frames strongly suggests that the camera is moving at a constant speed throughout the captured sequence, making option C the most probable answer.

Qwen2 VL 72B instruct AWQ: C

Qwen2 VL 7B Instruct: D

Video-CCAM-v1.1 14B: Answer is B.

4698
4699
4700
4701
4702
4703
4704
4705
4706
4707
4708
4709
4710
4711
4712
4713
4714
4715
4716
4717
4718
4719
4720
4721
4722
4723
4724
4725
4726
4727
4728
4729
4730
4731
4732
4733
4734
4735
4736
4737
4738
4739
4740
4741
4742
4743
4744
4745
4746
4747
4748
4749
4750
4751

InternVL 2 40B: Based on the provided frames, it appears that **the camera's movement speed is relatively constant throughout the video**. The frames show a consistent view of the zebras and the landscape without any noticeable changes in speed or direction.

Therefore, the most possible answer is:

'C': Constant speed.

4752 K.2.12 EXAMPLE 12

4753

4754

velocity&frequency 159 YouTube

4755

4756

4757

4758

4759

4760



4761

<https://www.youtube.com/watch?v=oyvYvruZfmw>

4762

What is the hula hoop's rotational speed pattern?

4763

A. Not moving at all. B. Decelerating.

4764

C. Constant speed. D. Accelerating.

4765

4766

Answer: D. Accelerating.

4767

4768

4769

4770

GPT-4o: Based on the provided frames, the hula hoop **appears to be rotating at a consistent speed** around the person's waist. There is no visible indication of the hula hoop slowing down or speeding up significantly in the sequence of frames.

4771

4772

Therefore, the most likely answer is:

4773

****C': Constant speed.****

4774

4775

4776

4777

Claude 3.5 Sonnet: Based on the series of images provided, I can analyze the hula hoop's rotational speed pattern. The images show a person in a pink shirt using a hula hoop outdoors.

4778

Looking at the sequence of frames, the hula hoop appears to be rotating at a fairly consistent speed throughout. There are no obvious signs of the hoop slowing down or speeding up significantly from frame to frame. The person maintains a steady rhythm and posture, which suggests they are able to keep the hoop spinning at a relatively constant rate.

4779

4780

4781

4782

4783

Given the options provided and based solely on what can be observed in these image frames, the most likely answer to the question "What is the hula hoop's rotational speed pattern?" is:

4784

C: Constant speed.

4785

4786

4787

The hoop seems to maintain a steady rotational speed across the sequence of frames, without any clear indication of acceleration or deceleration. Of course, without more precise measurements or a longer video sequence, it's difficult to rule out minor fluctuations in speed entirely. However, from the visual information provided, a constant speed appears to be the best characterization of the hula hoop's rotational pattern.

4788

4789

4790

4791

4792

4793

4794

Qwen2 VL 72B instruct AWQ: C

4795

4796

Qwen2 VL 7B Instruct: D

4797

4798

Video-CCAM-v1.1 14B: Answer is C.

4799

4800

InternVL 2 40B: Based on the provided frames, the hula hoop **appears to maintain a consistent motion around the girl's waist.** There is no visible change in the speed or pattern of the hoop's rotation. Therefore, the most likely answer is:

4801

4802

4803

'C': Constant speed.

4804

4805

4806 K.3 SIMULATED
4807
4808 K.3.1 EXAMPLE 1
4809

velocity&frequency 81



4811
4812
4813
4814 What is the pattern of the car's speed?

- 4815
4816 A. Decelerating. B. Accelerating.
4817 C. Constant speed. D. Not moving at all.
4818

4819 Answer: A. Decelerating.
4820

4821
4822 **GPT-4o:** Based on the provided frames, the car appears to be moving consistently across the frames.
4823 The position of the car relative to the frame edges does not show significant changes that would indicate
4824 acceleration or deceleration.

4825 Therefore, the most likely answer is:

4826
4827 ****C: Constant speed.****
4828

4829 **Claude 3.5 Sonnet:** Based on the sequence of frames provided, the most likely answer to the question
4830 "What is the pattern of the car's speed?" is:

4831 **C: Constant speed.**

4832 Reasoning:

- 4833
4834 1. The red car appears to be moving from left to right across the frames.
4835 2. The distance the car travels between each frame seems to be relatively consistent.
4836 3. There's no apparent change in the car's position that would suggest acceleration or deceleration.
4837 4. The car is clearly moving, so it's not stationary.

4838 While it's difficult to determine the exact speed without more precise measurements, the consis-
4839 tent movement of the car across the frames suggests it's maintaining a steady pace. Therefore, based on the
4840 visual information provided, the most likely answer is that the car is moving at a constant speed.
4841

4842 **Qwen2 VL 72B instruct AWQ: C**
4843

4844 **Qwen2 VL 7B Instruct: C**
4845

4846 **Video-CCAM-v1.1 14B:** Answer is **A**.
4847

4848 **InternVL 2 40B:** Based on the frames provided, the car appears to be moving from left to right across the
4849 screen. The car's position changes in each frame, indicating that it is moving. However, the frames do not
4850 provide enough information to determine the exact pattern of the car's speed (whether it is accelerating,
4851 decelerating, or moving at a constant speed).
4852

4853 Given the options provided, the most possible answer would be:

4854 **B: Accelerating.**
4855
4856
4857
4858
4859

4860 L COMMON FAILURE CASES OF SHAPE & TREND

4861

4862 L.1 HUMAN

4863

4864 L.1.1 EXAMPLE 1

4865

4866

4867

4868

4869

4870

4871

4872

4873

4874

4875

4876

4877

4878

4879

4880

4881

4882

4883

4884

4885

4886

4887

4888

4889

4890

4891

4892

4893

4894

4895

4896

4897

4898

4899

4900

4901

4902

4903

4904

4905

4906

4907

4908

4909


4910

4911

4912

4913

shape&trend 0



What is the shape of the object that the person drew in the air?

- A. Trapezoid.
- B. Diamond.
- C. Square/rectangle.
- D. Circle.
- E. Triangle.
- F. Not drawing at all.

Answer: A. Trapezoid.

GPT-4o: Based on the sequence of frames, the person appears to be drawing a shape in the air. The movement of the hand suggests a continuous motion that could form a closed shape.

1. The hand starts at the top (Frame 1).
2. Moves to the right (Frame 3).
3. Moves downwards (Frame 5).
4. Moves to the left (Frame 7).
5. Moves upwards (Frame 9).
6. Returns to the starting point (Frame 11).

This sequence of movements is consistent with drawing a square or rectangle.

Therefore, the most likely answer is:

Option 'C': Square/rectangle.

Claude 3.5 Sonnet: Based on the sequence of images provided, it appears the person is not drawing a specific shape in the air. Rather, they are making various gestures with their hand and arm, including pointing upwards, moving their arm horizontally, and positioning their hand near their face. **The movements do not seem to trace out any clear geometric shape like those listed in the options.**

However, since I'm instructed to provide an answer from the given options, I would guess:

4914
4915
4916
4917
4918
4919
4920
4921
4922
4923
4924
4925
4926
4927
4928
4929
4930
4931
4932
4933
4934
4935
4936
4937
4938
4939
4940
4941
4942
4943
4944
4945
4946
4947
4948
4949
4950
4951
4952
4953
4954
4955
4956
4957
4958
4959
4960
4961
4962
4963
4964
4965
4966
4967

C: Square/rectangle.

This is because some of the arm movements, particularly the horizontal and vertical motions, could potentially be interpreted as tracing out straight lines that form corners, which are characteristic of squares or rectangles. But I want to emphasize that this is a very tentative guess based on limited information, as the images do not clearly show the person drawing any specific shape.

Qwen2 VL 72B instruct AWQ: C

Qwen2 VL 7B Instruct: B

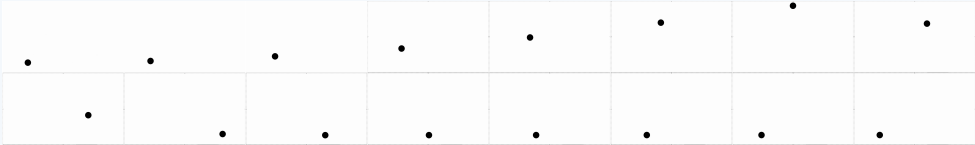
Video-CCAM-v1.1 14B: Answer is B.

InternVL 2 40B: Based on the frames provided, the person appears to be drawing a shape in the air with their finger. The shape seems to be a straight line followed by a curve, which is characteristic of a circle. Therefore, the most possible answer is:

'D': Circle.

4968 L.2 SIMULATED
4969
4970 L.2.1 EXAMPLE 1
4971

shape&trend 174



4972
4973
4974
4975
4976
4977
4978

4979 What shape is formed by the black circle's trajectory as it moves?
4980

4981 A. Diamond. B. Triangle.
4982 C. Circle. D. Square/rectangle.
4983 E. Trapezoid. F. Not drawing at all.
4984

4985 Answer: B. Triangle.
4986
4987

4988 **GPT-4o:** Based on the provided frames, the black circle appears to move in a path that forms a square or
4989 rectangle. The circle starts at the bottom left, moves to the top left, then to the top right, and finally to the
4990 bottom right before returning to the starting position.

4991 Therefore, the shape formed by the black circle's trajectory is:

4992
4993 ****D: Square/rectangle.****
4994

4995 **Claude 3.5 Sonnet:** Based on the sequence of frames provided, the black circle appears to be moving in a
4996 circular trajectory. The circle starts in the bottom left corner, moves up to the top center, then down to the
4997 bottom right, and finally back towards the bottom left. This circular motion is repeated across the frames.

4998 Therefore, the most likely answer to the question "What shape is formed by the black circle's tra-
4999 jectory as it moves?" is:

5000
5001 **C: Circle.**
5002

5003 **Qwen2 VL 72B instruct AWQ: D**
5004

5005 **Qwen2 VL 7B Instruct: C**
5006

5007 **Video-CCAM-v1.1 14B:** Answer is **A**.
5008

5009 **InternVL 2 40B:** The black circle's trajectory forms a shape that resembles a diamond.
5010

5011 **Answer: A**
5012
5013
5014
5015
5016
5017
5018
5019
5020
5021

5022 L.2.2 EXAMPLE 2

5023
5024
5025
5026
5027
5028
5029
5030
5031
5032
5033
5034
5035
5036
5037
5038
5039
5040
5041
5042
5043
5044
5045
5046
5047
5048
5049
5050
5051
5052
5053
5054
5055
5056
5057
5058
5059
5060
5061
5062
5063
5064
5065
5066
5067
5068
5069
5070
5071
5072
5073
5074
5075

shape&trend 175

What shape is formed by the black circle's trajectory as it moves?

A. Diamond. B. Not drawing at all.
C. Trapezoid. D. Circle.
E. Triangle. F. Square/rectangle.

Answer: C. Trapezoid.

GPT-4o: Based on the provided frames, the black circle appears to be moving in a pattern that suggests it is tracing out a circular path. The circle starts at the bottom, moves to the left, then upwards, to the right, and back down, completing a loop.

Therefore, the most likely answer is:

****D: Circle.****

Claude 3.5 Sonnet: Based on the sequence of frames provided, the black circle appears to be moving in a circular pattern. The circle starts in the bottom left corner, moves to the right, then up towards the center, then to the left, and finally back down towards the starting position. This movement pattern suggests that the shape formed by the black circle's trajectory is most likely a circle.

Therefore, the answer to the question "What shape is formed by the black circle's trajectory as it moves?" is:

D. Circle.

Qwen2 VL 72B instruct AWQ: D

Qwen2 VL 7B Instruct: The black circle's trajectory forms a circle.

Video-CCAM-v1.1 14B: Answer is D.

InternVL 2 40B: The black circle's trajectory forms a shape that resembles a diamond.

Answer: A

5076 L.2.3 EXAMPLE 3

5077

5078

5079

5080

5081

5082

5083

5084

5085

5086

5087

5088

5089

5090

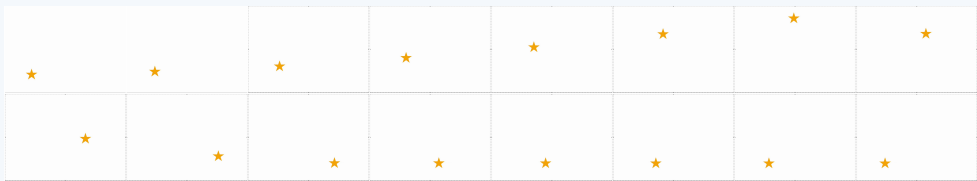
5091

5092

5093

5094

shape&trend 182



What shape is formed by the yellow star's trajectory as it moves?

A. Diamond. B. Triangle.
C. Circle. D. Square/rectangle.
E. Not drawing at all. F. Trapezoid.

Answer: B. Triangle.

5095

5096

5097

5098

5099

5100

5101

GPT-4o: Based on the provided frames, the yellow star appears to move in a pattern that forms a closed shape. The star starts at the bottom left, moves up, then to the right, down, and back to the left, completing a loop. This movement suggests a shape with four sides.

Given the options, the most likely shape formed by the yellow star's trajectory is:

****Option D: Square/rectangle.****

5102

5103

5104

5105

5106

5107

5108

5109

5110

Claude 3.5 Sonnet: Based on the sequence of images provided, the yellow star appears to be moving in a circular trajectory. The star's position changes gradually across the frames, moving from the bottom left, to the center, to the top right, and then back towards the bottom left. This circular motion is consistent throughout the 16 frames.

Therefore, the most likely answer to the question "What shape is formed by the yellow star's trajectory as it moves?" is:

C: Circle.

5111

5112

Qwen2 VL 72B instruct AWQ: C

5113

5114

Qwen2 VL 7B Instruct: C

5115

5116

Video-CCAM-v1.1 14B: Answer is **A**.

5117

5118

5119

5120

InternVL 2 40B: The yellow star's trajectory forms a shape that resembles a diamond.

Answer: A

5121

5122

5123

5124

5125

5126

5127

5128

5129

5130 L.2.4 EXAMPLE 4

5131

5132

5133

5134

5135

5136

5137

5138

5139

5140

5141

5142

5143

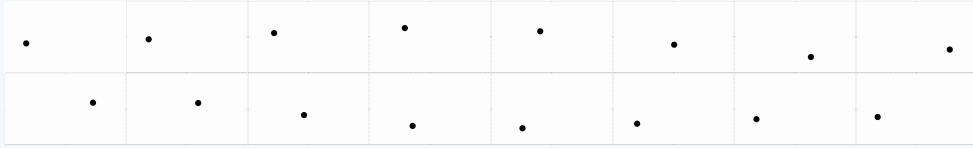
5144

5145

5146

5147

shape&trend 180



What shape is formed by the black circle's trajectory as it moves?

A. Circle. B. Infinity-shape.
C. Square/rectangle. D. Trapezoid.
E. Diamond. F. Not drawing at all.

Answer: B. Infinity-shape.

5148

5149

5150

5151

5152

5153

5154

GPT-4o: Based on the provided frames, the black circle appears to be moving in a pattern that forms a closed loop. The circle moves from the bottom center to the left, then to the top, then to the right, and back to the bottom center. This pattern suggests a circular trajectory.

Therefore, the most likely answer is:

****A': Circle.****

5155

5156

5157

5158

5159

5160

5161

5162

5163

5164

5165

5166

5167

Claude 3.5 Sonnet: Based on analyzing the 16 frames provided, which show the position of a black circle at different points in time, I can make an educated guess about the shape of the circle's trajectory.

The black circle appears to move in a roughly circular or oval path around the center of the frame. It starts in the upper left, moves to the right, then down to the lower right, back to the left, and up again. This circular motion is repeated across multiple frames.

Given the options provided, the most likely answer is:

A: Circle.

The trajectory forms an approximately circular shape as the black circle moves around the frame in a loop. While it may not be a perfect circle, this option best matches the observed motion compared to the other choices given.

5168

5169

Qwen2 VL 72B instruct AWQ: A

5170

5171

Qwen2 VL 7B Instruct: B

5172

5173

Video-CCAM-v1.1 14B: Answer is **A**.

5174

5175

5176

5177

5178

5179

5180

5181

5182

5183

InternVL 2 40B: Based on the provided frames, the black circle appears to move in a way that forms a shape that resembles a diamond. The trajectory of the circle's movement creates the outline of a diamond shape.

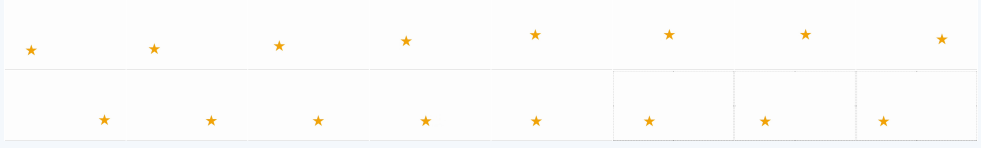
Therefore, the correct answer is:

'E': Diamond.

5184 L.2.5 EXAMPLE 5

5185
5186
5187
5188
5189
5190
5191
5192
5193
5194
5195
5196
5197
5198
5199
5200
5201
5202
5203
5204
5205
5206
5207
5208
5209
5210
5211
5212
5213
5214
5215
5216
5217
5218
5219
5220
5221
5222
5223
5224
5225
5226
5227
5228
5229
5230
5231
5232
5233
5234
5235
5236
5237

shape&trend 184



What shape is formed by the yellow star's trajectory as it moves?

A. Square/rectangle. B. Not drawing at all.
C. Circle. D. Trapezoid.
E. Diamond. F. Triangle.

Answer: D. Trapezoid.

GPT-4o: Based on the provided frames, the yellow star appears to be moving in a circular trajectory. The star starts at the bottom left, moves to the top left, then to the top right, and finally to the bottom right, before returning to the bottom left. This pattern suggests a circular motion.

Therefore, the most likely answer is:

*****C': Circle.****

Claude 3.5 Sonnet: Based on analyzing the 16 frames provided, the most likely answer to the question "What shape is formed by the yellow star's trajectory as it moves?" is:

C. Circle.

The yellow star appears to move in a circular pattern across the frames. It starts in the bottom left corner, moves upward and to the right, then back down to the left, tracing what seems to be a circular trajectory. While we don't see the complete circle in every frame, the overall movement pattern strongly suggests a circular path.

This circular motion is the most consistent with the star's positions across the different frames, compared to the other shape options provided. The star does not appear to move in straight lines that would form a square, rectangle, trapezoid, diamond, or triangle. The smooth curved motion is most indicative of a circular trajectory.

Qwen2 VL 72B instruct AWQ: C

Qwen2 VL 7B Instruct: C

Video-CCAM-v1.1 14B: Answer is **E**.

InternVL 2 40B: The yellow star's trajectory forms a shape that resembles a circle.

Answer: C

5238 L.2.6 EXAMPLE 6

5239

5240

5241

5242

5243

5244

5245

5246

5247

5248

5249

5250

5251

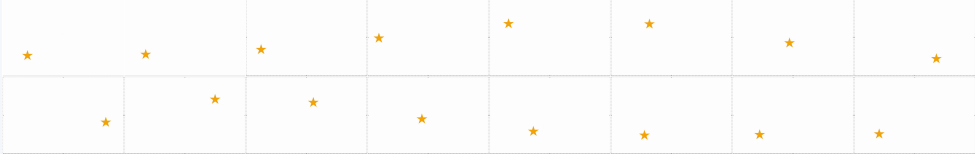
5252

5253

5254

5255

shape&trend 186



What shape is formed by the yellow star's trajectory as it moves?

A. Triangle. B. Infinity-shape.
C. Not drawing at all. D. Circle.
E. Trapezoid. F. Diamond.

Answer: B. Infinity-shape.

5256

5257

5258

5259

5260

5261

5262

GPT-4o: Based on the provided frames, **the yellow star appears to be moving in a circular pattern.** The star starts at the bottom left, moves up, then to the right, down, and back to the left, completing a loop.

Therefore, the shape formed by the yellow star's trajectory as it moves is:

****Option D: Circle.****

5263

5264

5265

5266

5267

5268

5269

5270

5271

5272

Claude 3.5 Sonnet: Based on the sequence of frames provided, **the yellow star appears to be moving in a circular pattern.** The star's position changes gradually from the bottom left corner, to the center, to the top right, and then back towards the bottom left, completing what seems to be a circular trajectory.

Therefore, the most likely answer to the question "What shape is formed by the yellow star's trajectory as it moves?" is:

D. Circle.

The star's movement pattern across the frames strongly suggests a circular path, making this the most plausible option among the given choices.

5273

5274

5275

5276

5277

5278

5279

5280

5281

5282

5283

5284

5285

5286

5287

5288

5289

5290

5291

Qwen2 VL 72B instruct AWQ: **A**

Qwen2 VL 7B Instruct: The yellow star's trajectory **forms a circle.**

Video-CCAM-v1.1 14B: Answer is **A.**

InternVL 2 40B: The yellow star's trajectory forms a shape that **resembles a triangle.**

Answer: **A**

5292 L.2.7 EXAMPLE 7

5293
5294
5295
5296
5297
5298
5299
5300
5301
5302
5303
5304
5305
5306
5307
5308
5309
5310
5311
5312
5313
5314
5315
5316
5317
5318
5319
5320
5321
5322
5323
5324
5325
5326
5327
5328
5329
5330
5331
5332
5333
5334
5335
5336
5337
5338
5339
5340
5341
5342
5343
5344
5345

shape&trend 183

What shape is formed by the yellow star's trajectory as it moves?

A. Not drawing at all. B. Diamond.
C. Circle. D. Trapezoid.
E. Triangle. F. Square/rectangle.

Answer: F. Square/rectangle.

GPT-4o: Based on the provided frames, the yellow star appears to move in a pattern that forms a square or rectangle. The star starts at the bottom left, moves up to the top left, then to the top right, down to the bottom right, and finally back to the bottom left.

Therefore, the shape formed by the yellow star's trajectory is:

****Option F: Square/rectangle.****

Claude 3.5 Sonnet: Based on the sequence of images provided, the yellow star appears to be moving in a circular trajectory. The star's position changes across the frames, starting from the bottom left, moving to the top center, then to the top right, and finally back towards the bottom left. This movement pattern suggests a circular motion.

Therefore, the most likely answer to the question "What shape is formed by the yellow star's trajectory as it moves?" is:

C. Circle.

Qwen2 VL 72B instruct AWQ: C

Qwen2 VL 7B Instruct: The yellow star's trajectory forms a diamond shape.

Video-CCAM-v1.1 14B: Answer is B.

InternVL 2 40B: The yellow star's trajectory forms a shape that resembles a diamond.

Answer: B

5346 M COMMON FAILURE CASES OF VISUAL CUES

5347

5348 M.1 OBJECT

5349

5350 M.1.1 EXAMPLE 1

5351

5352

visual cues 0 Music-AVQA

5353

5354

5355

5356

5357

5358



5359 Which musical instrument sounds last?

5360

5361 A. Cello. B. None of them produces any sound.

5362

5363

5364

5365

5366

5367

5368

5369

5370

5371

5372

GPT-4o: Based on the provided frames, it appears that the three musicians are playing their instruments simultaneously throughout the video. **There is no clear indication that any one instrument stops playing before the others.** Therefore, the most plausible answer based on the frames is:

****Option E: All instruments sound at the same time.****

5373

5374

5375

5376

5377

5378

5379

5380

5381

5382

5383

5384

5385

Claude 3.5 Sonnet: Based on the series of images provided, I can see three musicians performing together outdoors at night. They are playing three different instruments: a flute, a cello, and a violin. Throughout the sequence of images, all three musicians appear to be actively playing their instruments simultaneously.

Given the question "Which musical instrument sounds last?" and the provided options, the most likely answer based on the visual information would be:

E: All instruments sound at the same time.

This is because in all frames, the three musicians are shown in playing positions with their instruments, suggesting they are performing together as an ensemble. **There's no visual indication that any instrument stops playing before the others** or continues after the others have finished. Therefore, the most reasonable assumption from the given information is that all instruments are sounding simultaneously throughout the performance.

5386

5387

5388

5389

5390

5391

5392

5393

5394

5395

5396

5397

5398

5399

Qwen2 VL 72B instruct AWQ: A

Qwen2 VL 7B Instruct: A

Video-CCAM-v1.1 14B: Answer is A.

InternVL 2 40B: Based on the provided frames, it is not possible to determine which musical instrument sounds last. The frames show three women playing different instruments (flute, cello, and violin), but **they do not provide information about the timing of their sounds.** Therefore, the most possible answer is:

'E': All instruments sound at the same time.

5400 M.1.2 EXAMPLE 2

5401

5402

5403

5404

5405

5406

5407

5408

5409

5410

5411

5412

5413

5414

5415

5416

5417

5418

5419

5420

5421

5422

5423

5424

5425

5426

5427

5428

5429

5430

5431

5432

5433

5434

5435

5436

5437

5438

5439

5440

5441

5442

5443

5444

5445

5446

5447

5448

5449

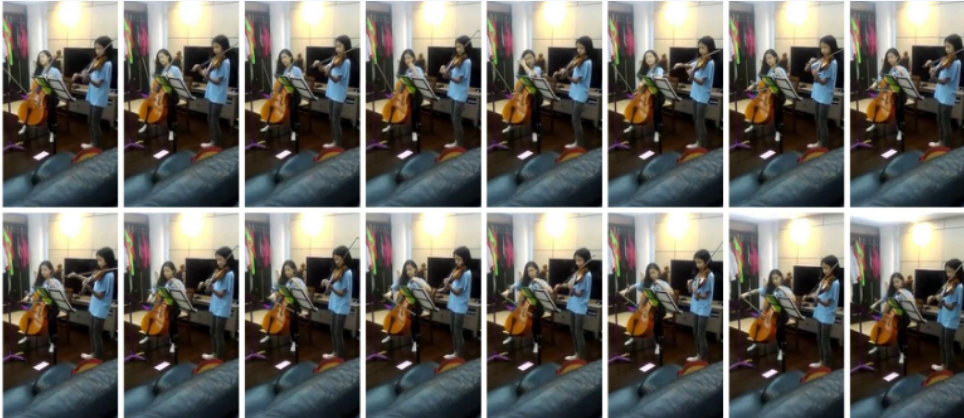
5450

5451

5452

5453

visual cues 28 Music-AVQA



Which musical instrument plays first?

A. Cello. B. Violin.
C. Both instruments begin simultaneously.
D. Neither instrument produces any sound.

Answer: B. Violin.

GPT-4o: Based on the provided frames, both the cello and the violin players appear to be preparing to play their instruments. **There is no clear indication from the frames alone which instrument starts playing first.** However, given that both musicians are in position and appear ready to play, it is reasonable to guess that they might begin simultaneously.

Therefore, the most plausible answer based on the frames is:

****Option C: Both instruments begin simultaneously.****

Claude 3.5 Sonnet: Based on the series of images provided, it's difficult to determine with certainty which instrument plays first. However, I'll make my best guess based on the visual information available.

The images show two individuals, one playing a cello and the other playing a violin. They appear to be in a practice or performance setting, with music stands in front of them. Both musicians seem to be in playing positions throughout the sequence of images.

Given that both musicians appear ready to play from the beginning of the sequence, and **there's no clear indication of one starting before the other**, my best guess would be:

C: Both instruments begin simultaneously.

This answer seems most plausible because both musicians are in playing position throughout the sequence, suggesting they might be performing a duet or playing together. Without audio information or more detailed visual cues about who starts first, it's reasonable to assume they begin playing at the same time.

Qwen2 VL 72B instruct AWQ: A

Qwen2 VL 7B Instruct: A

5454
5455
5456
5457
5458
5459
5460
5461
5462
5463
5464
5465
5466
5467
5468
5469
5470
5471
5472
5473
5474
5475
5476
5477
5478
5479
5480
5481
5482
5483
5484
5485
5486
5487
5488
5489
5490
5491
5492
5493
5494
5495
5496
5497
5498
5499
5500
5501
5502
5503
5504
5505
5506
5507

Video-CCAM-v1.1 14B: Answer is [A](#).

InternVL 2 40B: Based on the provided frames, it appears that both the cello and the violin are being played simultaneously. Therefore, the most possible answer is:

[C](#): Both instruments begin simultaneously.

5508 M.1.3 EXAMPLE 3

5509

5510

5511

5512

5513

5514

5515

5516

5517

5518

5519

5520


5521

5522

5523

5524

visual cues 31 Music-AVQA



Which musical instrument plays first?

A. Both instruments begin simultaneously. B. Flute.
C. Ukulele. D. Neither instrument produces any sound.

Answer: B. Flute.

5525

5526

5527

5528

5529

GPT-4o: Based on the provided frames, it appears that both the flute player and the ukulele player are engaged in playing their instruments throughout the sequence. **There is no clear indication that one starts before the other.** Therefore, the most plausible answer based on the frames is:

****Option A: Both instruments begin simultaneously.****

5530

5531

5532

5533

5534

5535

5536

5537

5538

5539

5540

5541

Claude 3.5 Sonnet: Based on the images provided, I would guess that the answer to the question "Which musical instrument plays first?" is:

C. Ukulele.

My reasoning is that in most of the frames, both musicians appear to be actively playing their instruments simultaneously. However, in the last two frames, we see a close-up of just the ukulele player's hands on the instrument, suggesting a focus on that instrument. **This could indicate that the ukulele starts the piece or has a solo introduction before the flute joins in.**

While I can't be certain without hearing the audio, the visual emphasis on the ukulele in those final frames leads me to believe it's the most likely instrument to play first in this musical performance.

5542

5543

Qwen2 VL 72B instruct AWQ: C

5544

5545

Qwen2 VL 7B Instruct: C

5546

5547

Video-CCAM-v1.1 14B: Answer is **C**.

5548

5549

5550

5551

5552

5553

5554

5555

5556

5557

5558

5559

5560

5561

InternVL 2 40B: Based on the frames provided, **it appears that both the flute and the ukulele are being played simultaneously.** Therefore, the most possible answer is:

A: Both instruments begin simultaneously.

5562 M.1.4 EXAMPLE 4

5563

5564

visual cues 38 Music-AVQA

5565

5566

5567

5568

5569

5570

5571

5572

5573

5574

5575

5576

5577

5578



Which violin is the first sounding instrument?

- A. Both violins sound at the same time. B. The right one.
C. The left one. D. Neither instrument produces any sound.

Answer: B. The right one.

5579

5580

5581

5582

5583

GPT-4o: Based on the provided frames, it appears that both violinists are playing their instruments simultaneously throughout the sequence. **There is no clear indication that one starts playing before the other.** Therefore, the most plausible answer based on the frames is:

'A': 'Both violins sound at the same time.'

5584

5585

5586

5587

5588

5589

5590

5591

5592

5593

5594

5595

5596

5597

5598

5599

Claude 3.5 Sonnet: Based on the series of images provided, it's difficult to determine with certainty which violin is the first sounding instrument. However, I'll make my best guess based on the visual information available.

From the images, we can see two violinists playing together throughout the sequence. They both appear to be actively playing their instruments in all frames, suggesting **they might be playing simultaneously or very close to each other in timing.**

Given that we need to choose an answer, and considering that the question asks specifically about which violin sounds first, I would guess:

A: Both violins sound at the same time.

This answer seems most plausible based on the visual evidence, as both violinists appear to be engaged in playing their instruments from the very first frame. There's no clear indication that one starts before the other. While it's impossible to hear the sound from these images alone, the synchronized movements of both players suggest they might be starting together.

5600

5601

Qwen2 VL 72B instruct AWQ: C

5602

5603

Qwen2 VL 7B Instruct: C

5604

5605

Video-CCAM-v1.1 14B: Answer is C.

5606

5607

5608

5609

5610

5611

5612

5613

5614

5615

InternVL 2 40B: Based on the frames provided, **it is not possible to determine which violin is the first sounding instrument.** The frames show both women playing the violin simultaneously, and there is no clear indication of which one started playing first.

5616 M.1.5 EXAMPLE 5

5617

5618

visual cues 1 Music-AVQA

5619

5620

5621

5622

5623

5624

5625

5626



5627

Which musical instrument plays first?

5628

A. Neither instrument produces any sound.

B. Both instruments begin simultaneously.

5630

C. Violin. D. Piano.

5631

5632

Answer: D. Piano.

5633

5634

5635

5636

GPT-4o: Based on the provided frames, both the boy playing the piano and the girl playing the violin appear to be engaged in playing their instruments simultaneously. There is no clear indication that one starts before the other. Therefore, the most plausible answer based on the frames is:

5637

5638

****Option B: Both instruments begin simultaneously.****

5639

5640

5641

5642

Claude 3.5 Sonnet: Based on the sequence of images provided, it's difficult to determine with certainty which instrument plays first. However, I'll make my best guess based on the information available in the frames.

5643

5644

5645

5646

The images show two children, one playing a violin and the other playing a piano. Both appear to be in playing positions throughout the sequence. **There's no clear indication of one instrument starting before the other.**

5647

5648

5649

Given that both children seem ready to play from the first frame, and their postures remain consistent throughout, I would guess that the most likely answer is:

5650

B: Both instruments begin simultaneously.

5651

5652

5653

5654

This answer assumes that the children start playing together, as they both appear prepared to play from the beginning of the sequence. However, it's important to note that without audio or more detailed visual cues, this is an educated guess based on the limited information provided in the image frames.

5655

5656

Qwen2 VL 72B instruct AWQ: D

5657

5658

Qwen2 VL 7B Instruct: C

5659

5660

Video-CCAM-v1.1 14B: Answer is **C**.

5661

5662

5663

5664

InternVL 2 40B: Based on the provided frames, **it is not possible to determine which musical instrument plays first.** The frames show both the girl playing the violin and the boy playing the piano, but they do not provide enough information to determine the sequence of their playing.

5665

5666

5667

5668

5669

5670 M.1.6 EXAMPLE 6

5671

5672

5673

5674

5675

5676

5677

5678

5679

5680

5681

5682

5683

5684

5685

5686

5687

5688

5689

5690

5691

5692

5693

5694

5695

5696

5697

5698

5699

5700

5701

5702

5703

5704

5705

5706

5707

5708

5709

5710

5711

5712

5713

5714

5715

5716

5717

5718

5719

5720

5721

5722

5723

visual cues 7 Music-AVQA



Which musical instrument plays first?

- A. Neither instrument produces any sound. B. Pipa.
C. Erhu. D. Both instruments begin simultaneously.

Answer: B. Pipa.

GPT-4o: Based on the provided frames, it appears that the person playing the Erhu (on the left) starts moving their bow before the person playing the Pipa (on the right) starts plucking the strings. This suggests that the Erhu might be the first instrument to produce sound.

Therefore, the most likely answer is:

Option C: Erhu.

Claude 3.5 Sonnet: Based on the series of images provided, I'll try my best to answer the question about which musical instrument plays first. However, it's important to note that these are still images, so we can't actually hear any sound or determine the exact order of play. I'll make an educated guess based on the visual cues in the frames.

Looking at the images, we can see two musicians holding traditional Chinese instruments. The person on the left is holding an erhu (a two-stringed bowed instrument), while the person on the right is holding a pipa (a four-stringed plucked instrument).

Throughout the sequence of images, both musicians appear to be in playing positions, holding their instruments ready to perform. There's no clear indication that one starts before the other based solely on these visual cues.

Given the question and the options provided, and considering that both musicians seem equally prepared to play in all frames, I would guess:

D: Both instruments begin simultaneously.

This answer assumes that since both musicians are consistently shown in playing positions, they likely start playing their instruments at the same time. However, I want to emphasize that this is an educated guess based on limited visual information, and without audio, it's impossible to determine with certainty which instrument actually plays first.

Qwen2 VL 72B instruct AWQ: C

Qwen2 VL 7B Instruct: C

Video-CCAM-v1.1 14B: Answer is B.

InternVL 2 40B: Based on the provided frames, it is not possible to determine which musical instrument plays first. The frames do not provide any information about the sequence of the instruments playing.

5724 M.1.7 EXAMPLE 7

5725

5726

5727

5728

5729

5730

5731

5732

5733

5734

5735

5736

5737

5738

5739

5740

5741

5742

5743

5744

5745

5746

5747

5748

5749

5750

5751

5752

5753

5754

5755

5756

5757

5758

5759

5760

5761

5762

5763

5764

5765

5766

5767

5768

5769

5770

5771

5772

5773

5774

5775

5776

5777

visual cues 14 Music-AVQA



Which musical instrument finishes the song?

- A. None of them produces any sound.
- B. Cello on the middle.
- C. All instruments sound at the same time.
- D. Violin on the left.
- E. Violin on the right.

Answer: B. Cello on the middle.

GPT-4o: Based on the 16 frames provided, it appears that the video features a trio of musicians: two violinists (one on the left and one on the right) and a cellist in the middle. The frames show the musicians playing their instruments throughout the video.

Given the options and the visual information from the frames, the most plausible answer is:

****Option D: Violin on the left.****

This is a guess based on the visual cues, as the frames do not provide explicit information about which instrument finishes the song.

Claude 3.5 Sonnet: Based on the provided frames, I will make my best guess to answer the question "Which musical instrument finishes the song?"

From analyzing the 16 frames, we can see a string trio performing - two violins (one on the left and one on the right) and a cello in the middle. All three musicians appear to be actively playing their instruments throughout the sequence of images.

Since the question asks specifically about which instrument finishes the song, and we don't have audio information or a clear indication of when the piece ends, I'll make an educated guess based on common musical practices and the visual cues available.

In many musical compositions, especially for string ensembles, it's common for all instruments to play until the end, often finishing together on a final chord or note. However, if I had to choose one instrument that might be more likely to have the final note or phrase, I would guess:

E: Violin on the right.

This is because:

1. Violins often carry the melody in string ensembles.
2. The right-side violinist seems to be in a slightly more prominent position in the frame.
3. In some compositions, the highest voice (which would typically be a violin) might have the final note or flourish.

However, I want to emphasize that this is a guess based on limited visual information, and without audio or more specific cues about the end of the piece, it's impossible to be certain.

Qwen2 VL 72B instruct AWQ: E

Qwen2 VL 7B Instruct: C

5778
5779
5780
5781
5782
5783
5784
5785
5786
5787
5788
5789
5790
5791
5792
5793
5794
5795
5796
5797
5798
5799
5800
5801
5802
5803
5804
5805
5806
5807
5808
5809
5810
5811
5812
5813
5814
5815
5816
5817
5818
5819
5820
5821
5822
5823
5824
5825
5826
5827
5828
5829
5830
5831

Video-CCAM-v1.1 14B: Answer is B.

InternVL 2 40B: Based on the given frames, it is not possible to determine which musical instrument finishes the song. The frames show the Crystal Strings Trio playing their instruments together, and there is no clear indication of which instrument finishes the song. Therefore, the most possible answer is:

'C': All instruments sound at the same time.

5832 M.1.8 EXAMPLE 8

5833

5834

5835

5836

5837

5838

5839

5840

5841

5842

5843

5844

5845

5846

5847

5848

5849

5850

5851

5852

5853

5854

5855

5856

5857

5858

5859

5860

5861

5862

5863

5864

5865

5866

5867

5868

5869

5870

5871

5872

5873

5874

5875

5876

5877

5878

5879

5880

5881


5882

5883

5884

5885

visual cues 30 Music-AVQA



Which musical instrument plays first?

A. Guitar. B. Both instruments begin simultaneously.
C. Clarinet. D. Neither instrument produces any sound.

Answer: A. Guitar.

GPT-4o: Based on the provided frames, both the clarinet player and the guitarist appear to be playing their instruments simultaneously in all frames. **There is no clear indication that one instrument starts before the other.**

Given this observation, the most plausible answer is:

****B': Both instruments begin simultaneously.****

Claude 3.5 Sonnet: Based on the series of images provided, it appears that both musicians are actively playing their instruments throughout the frames. The woman is consistently shown playing the clarinet, while the man is playing the guitar. Since both musicians seem to be engaged in playing simultaneously across all frames, and **there's no clear indication of one instrument starting before the other**, the most likely answer to the question "Which musical instrument plays first?" is:

B: Both instruments begin simultaneously.

This conclusion is drawn from the consistent positioning and active playing stance of both musicians across all the frames, suggesting they are performing together from the beginning of the captured sequence.

Qwen2 VL 72B instruct AWQ: A

Qwen2 VL 7B Instruct: C

Video-CCAM-v1.1 14B: Answer is C.

InternVL 2 40B: Based on the frames provided, **it appears that both the guitar and the clarinet are being played simultaneously.** Therefore, the most possible answer is:

B: Both instruments begin simultaneously.

5886 M.1.9 EXAMPLE 9

5887

5888

5889

5890

5891

5892

5893

5894

5895

5896

5897

5898

5899

5900

5901

5902

5903

5904

5905

5906

5907

5908

5909

5910

5911

5912

5913

5914

5915

5916

5917

5918

5919

5920

5921

5922

5923

5924

5925

5926

5927

5928

5929

5930

5931

5932

5933

5934

5935

5936

5937

5938

5939

visual cues 55 Music-AVQA



Which musical instrument plays first?

- A. Piano. B. Both instruments begin simultaneously.
C. Neither instrument produces any sound. D. Violin.

Answer: B. Both instruments begin simultaneously.

GPT-4o: Based on the provided frames, both the violinist and the pianist are in position and appear to be playing their instruments. However, there is no clear indication from the frames alone which instrument starts playing first. Given the context of a typical musical performance, it is common for the piano to provide an introduction or accompaniment before the violin joins in.

Therefore, the most plausible answer based on the frames is:

'A': Piano.

Claude 3.5 Sonnet: Based on the images provided, the most likely answer to the question "Which musical instrument plays first?" is:

D. Violin.

The reasoning for this answer is:

1. Throughout all the frames, we see a person actively playing the violin. They are holding the violin and moving the bow across the strings.
2. The person at the piano is seated but does not appear to be actively playing in any of the frames. Their hands are not visible on the keys.
3. The violinist seems to be the main focus of the performance, standing in front of a music stand and microphone.
4. The consistent positioning and active playing of the violin across all frames suggests it is the primary instrument being played in this scene.

While we cannot hear the audio to confirm definitively, based on the visual evidence provided in these frames, it appears that the violin is the instrument playing first or primarily in this performance setting.

Qwen2 VL 72B instruct AWQ: A

Qwen2 VL 7B Instruct: D

Video-CCAM-v1.1 14B: Answer is D.

InternVL 2 40B: Based on the frames provided, it is not possible to determine which musical instrument plays first. The video captures a continuous performance where both the violin and piano are being played simultaneously. Therefore, the most possible answer is:

5940
5941
5942
5943
5944
5945
5946
5947
5948
5949
5950
5951
5952
5953
5954
5955
5956
5957
5958
5959
5960
5961
5962
5963
5964
5965
5966
5967
5968
5969
5970
5971
5972
5973
5974
5975
5976
5977
5978
5979
5980
5981
5982
5983
5984
5985
5986
5987
5988
5989
5990
5991
5992
5993

B: Both instruments begin simultaneously.

5994 M.1.10 EXAMPLE 10

5995

5996

5997

5998

5999

6000

6001

6002

6003

6004

6005

6006

6007

6008

6009

6010

6011

6012

6013

6014

6015

6016

6017

6018

6019

6020

6021

6022

6023

6024

6025

6026

6027

6028

6029

6030

6031

6032

6033

6034

6035

6036

6037

6038

6039

6040

6041

6042

6043

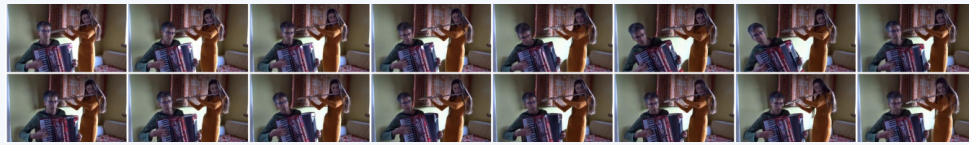
6044

6045

6046

6047

visual cues 62 Music-AVQA



Which musical instrument plays first?

- A. Accordion. B. Flute.
- C. Neither instrument produces any sound.
- D. Both instruments begin simultaneously.

Answer: D. Both instruments begin simultaneously.

GPT-4o: Based on the provided frames, it appears that both the accordion player and the flute player are preparing to play their instruments. However, there is no clear indication from the frames alone which instrument starts playing first.

Given the options and the visual cues, I will make an educated guess:

- The accordion player seems to be actively pressing the keys and moving the bellows in the initial frames.
- The flute player is holding the flute in a playing position but does not show clear signs of starting to play first.

Therefore, the most plausible answer based on the frames is:

****Option A: Accordion.****

Claude 3.5 Sonnet: Based on the sequence of images provided, I would guess that the answer to the question "Which musical instrument plays first?" is:

A: Accordion.

My reasoning for this guess is:

1. Throughout all the frames, we see a man holding and seemingly playing an accordion. His posture and hand positions on the instrument remain consistent, suggesting he is actively playing it.
2. The woman with the flute appears to be in a playing position, but her posture and the position of the flute don't change significantly across the frames. This could indicate she might be preparing to play or joining in later.
3. The accordion player seems more engaged with his instrument from the beginning, which might suggest he starts playing first.
4. While we can't hear the audio, the consistent positioning of the accordion player across all frames gives the impression that he might have begun playing first or is the lead musician in this scenario.

Without being able to hear the audio or see more detailed motion, it's impossible to be certain. However, based on the visual cues in the provided frames, the accordion seems the most likely instrument to have started playing first.

Qwen2 VL 72B instruct AWQ: A

Qwen2 VL 7B Instruct: A

6048
6049
6050
6051
6052
6053
6054
6055
6056
6057
6058
6059
6060
6061
6062
6063
6064
6065
6066
6067
6068
6069
6070
6071
6072
6073
6074
6075
6076
6077
6078
6079
6080
6081
6082
6083
6084
6085
6086
6087
6088
6089
6090
6091
6092
6093
6094
6095
6096
6097
6098
6099
6100
6101

Video-CCAM-v1.1 14B: Answer is **A**.

InternVL 2 40B: Based on the provided frames, it is not possible to determine which musical instrument plays first. Both the accordion and the flute are being played simultaneously in the video. Therefore, the most possible answer is:

D: Both instruments begin simultaneously.

6102 M.1.11 EXAMPLE 11

6103

6104

6105

6106

6107

6108

6109

6110

6111

6112

6113

6114

6115

6116

6117

6118

6119

6120

6121

6122

6123

6124

6125

6126

6127

6128

6129

6130

6131

6132

6133

6134

6135

6136

6137

6138

6139

6140

6141

6142

6143

6144

6145

6146

6147

6148

6149

6150

6151

6152

6153

6154

6155

visual cues 63 Music-AVQA



Which musical instrument plays first?

- A. Both instruments begin simultaneously.
- B. Neither instrument produces any sound.
- C. Guzheng. D. Accordion.

Answer: A. Both instruments begin simultaneously.

GPT-4o: Based on the provided frames, it appears that **the person playing the Guzheng is actively plucking the strings**, while the person with the accordion is holding the instrument but not yet playing it. This suggests that the Guzheng might be the first instrument to produce sound.

Therefore, the most likely answer is:

****C*: Guzheng.****

Claude 3.5 Sonnet: Based on the provided frames, it's difficult to determine with certainty which instrument plays first. However, I'll make my best guess based on the information available in the images.

The frames show two musicians, one playing an accordion and the other playing what appears to be a guzheng (a Chinese plucked string instrument). Both musicians seem to be in performance positions throughout the sequence of images.

Given that the question asks which instrument plays first, and considering that both musicians appear ready to play in all frames, I would guess:

C: Guzheng.

My reasoning for this guess is that **the guzheng player's posture seems slightly more active in the earlier frames**, with their hands positioned over the strings, suggesting they might be about to play or have just started playing. The accordion player's posture remains relatively consistent throughout the frames.

However, I want to emphasize that this is a guess based on limited visual information, as we can't hear the audio or see the exact moment when each instrument begins playing.

Qwen2 VL 72B instruct AWQ: D

Qwen2 VL 7B Instruct: D

Video-CCAM-v1.1 14B: Answer is **D**.

InternVL 2 40B: Based on the provided frames, it is not possible to determine which musical instrument plays first. The frames show both the accordion and the guzheng being played simultaneously, without any clear indication of which one started first. Therefore, the most possible answer is:

'A': 'Both instruments begin simultaneously.'

6156 M.1.12 EXAMPLE 12

6157

6158

visual cues 3 Music-AVQA

6159

6160

6161

6162

6163

6164

6165

6166

6167

6168

6169

6170

6171

6172

6173

6174

6175

6176

6177

6178

6179

6180

6181

6182

6183

6184

6185

6186

6187

6188

6189

6190

6191

6192

6193

6194

6195

6196

6197

6198

6199

6200

6201

6202

6203

6204

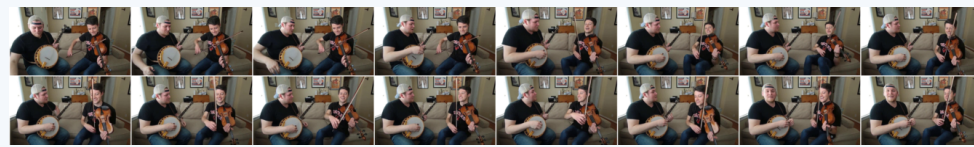
6205

6206

6207

6208

6209



Which instrument plays first?

- A. Both instruments begin simultaneously. B. Banjo.
C. Violin. D. Neither instrument produces any sound.

Answer: C. Violin.

GPT-4o: Based on the provided frames, it appears that the person with the violin is preparing to play first. In the initial frames, the person with the violin is seen positioning the bow and the violin, while the person with the banjo is not yet actively playing.

Therefore, the most likely answer is:

Option C: Violin.

Claude 3.5 Sonnet: Based on the sequence of images provided, I would guess that the answer to the question "Which instrument plays first?" is:

B. Banjo.

My reasoning is as follows:

1. In the early frames, both musicians are holding their instruments (banjo and violin) but not actively playing.
2. As the sequence progresses, **the banjo player appears to become more animated and engaged first.** His posture changes and he seems to be moving his hands on the instrument in a playing motion.
3. The violin player's posture and engagement seems to follow slightly after the banjo player in the sequence of images.
4. In the later frames, both musicians appear to be actively playing and singing/vocalizing together.

While it's impossible to be certain without audio, the visual cues suggest that the banjo player initiates the performance, with the violin joining in shortly after. Therefore, my best guess based solely on these images is that the banjo plays first.

Qwen2 VL 72B instruct AWQ: B

Qwen2 VL 7B Instruct: B

Video-CCAM-v1.1 14B: Answer is **B**.


InternVL 2 40B: Based on the frames provided, **it appears that both the banjo and the violin are being played simultaneously.** Therefore, the most possible answer is:

A: Both instruments begin simultaneously.

6210
6211
6212
6213
6214
6215
6216
6217
6218
6219
6220
6221
6222
6223
6224
6225
6226
6227
6228
6229
6230
6231
6232
6233
6234
6235
6236
6237
6238
6239
6240
6241
6242
6243
6244
6245
6246
6247
6248
6249
6250
6251
6252
6253
6254
6255
6256
6257
6258
6259
6260
6261
6262
6263

M.1.13 EXAMPLE 13

visual cues 19 Music-AVQA



Which guitar is the first sounding instrument?

A. The left one. B. Neither instrument produces any sound.
C. Both guitars sound at the same time. D. The right one.

Answer: D. The right one.

GPT-4o: Based on the provided frames, it appears that both guitarists are actively playing their instruments. However, the guitarist on the right seems to be more animated and engaged in playing, suggesting that he might be the first to produce sound.

Therefore, the most likely answer is:

****D: The right one.****

Claude 3.5 Sonnet: Based on the sequence of images provided, it appears that both guitarists are playing their instruments simultaneously throughout the video. From the first frame to the last, we can see both musicians actively engaged with their guitars, suggesting they are performing together.

While it's impossible to determine the exact timing of when each instrument first produces sound from these still images alone, **the visual evidence suggests that both guitarists are playing concurrently.** Therefore, the most likely answer to the question "Which guitar is the first sounding instrument?" is:

C: Both guitars sound at the same time.

This answer assumes that since both guitarists are shown playing from the beginning, they likely started their performance together. Without audio information, this seems to be the most reasonable conclusion based on the visual information provided in the frames.

Qwen2 VL 72B instruct AWQ: A

Qwen2 VL 7B Instruct: A

Video-CCAM-v1.1 14B: Answer is A.

InternVL 2 40B: Based on the frames provided, **it appears that both guitars are being played simultaneously.** Therefore, the most likely answer is:

'C': Both guitars sound at the same time.

6264 M.1.14 EXAMPLE 14

6265

6266

6267

6268

6269

6270

6271

6272

6273

6274

6275

6276

6277

6278

6279

6280

6281

6282

6283

6284

6285

6286

6287

6288

6289

6290

6291

6292

6293

6294

6295

6296

6297

6298

6299

6300

6301

6302

6303

6304

6305

6306

6307

6308

6309

6310

6311

6312

6313

6314

6315

6316

6317

visual cues 25 Music-AVQA



Which musical instrument plays first?

- A. Guitar. B. Neither instrument produces any sound.
C. Pipa. D. Both instruments begin simultaneously.

Answer: A. Guitar.

GPT-4o: Based on the provided frames, it appears that both musicians are preparing to play their instruments. The person with the guitar is seen singing and strumming the guitar, while the person with the pipa is adjusting and plucking the strings.

Given the visual cues, it seems that the person with the guitar is more actively engaged in playing and singing, suggesting that the guitar might be the first instrument to produce sound.

Therefore, the most likely answer is:

****A': Guitar.****

Claude 3.5 Sonnet: Based on the series of images provided, I would guess that the answer to the question "Which musical instrument plays first?" is:

C. Pipa.

My reasoning for this guess is:

1. The images show two musicians, one playing a pipa (a Chinese lute-like instrument) and the other playing a guitar.
2. In several of the early frames, **the pipa player appears to be more actively engaged with their instrument**, holding it in a playing position and seeming to strum or pluck the strings.
3. The guitar player, while holding the instrument, doesn't seem to be as actively playing in the initial frames.
4. There's a progression in the images where both musicians become more animated, suggesting they both eventually play.
5. Given that the pipa player seems more ready to play in the earlier frames, it's reasonable to guess that this instrument might have started first.

However, it's important to note that this is an educated guess based solely on the visual information provided in the still frames. Without audio or more detailed timing information, it's impossible to be certain which instrument actually played first.

Qwen2 VL 72B instruct AWQ: C

Qwen2 VL 7B Instruct: C

Video-CCAM-v1.1 14B: Answer is C.

6318
6319
6320
6321
6322
6323
6324
6325
6326
6327
6328
6329
6330
6331
6332
6333
6334
6335
6336
6337
6338
6339
6340
6341
6342
6343
6344
6345
6346
6347
6348
6349
6350
6351
6352
6353
6354
6355
6356
6357
6358
6359
6360
6361
6362
6363
6364
6365
6366
6367
6368
6369
6370
6371

InternVL 2 40B: Based on the provided frames, it is not possible to determine which musical instrument plays first. The frames show both the guitar and the pipa being played simultaneously, without any clear indication of which one started first.

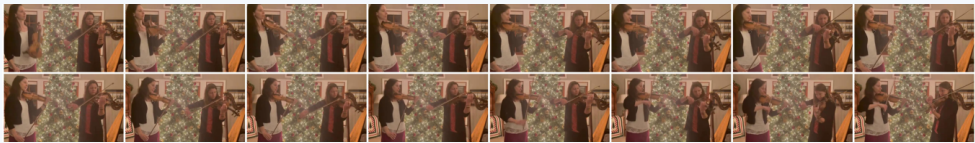
Therefore, the most possible answer is:

D: Both instruments begin simultaneously.

6372
6373
6374
6375
6376
6377
6378
6379
6380
6381
6382
6383
6384
6385
6386
6387
6388
6389
6390
6391
6392
6393
6394
6395
6396
6397
6398
6399
6400
6401
6402
6403
6404
6405
6406
6407
6408
6409
6410
6411
6412
6413
6414
6415
6416
6417
6418
6419
6420
6421
6422
6423
6424
6425

M.1.15 EXAMPLE 15

visual cues 26 Music-AVQA



Which violin is the first sounding instrument?

A. Both violins sound at the same time. B. The right one.
C. Neither instrument produces any sound. D. The left one.

Answer: B. The right one.

GPT-4o: Based on the provided frames, it appears that the person on the right (the right violinist) is consistently in a playing position with the bow on the strings, while the person on the left (the left violinist) is not always in a playing position and sometimes has the bow away from the strings.

Given this observation, the most likely answer is:

****B**: The right one.****

Claude 3.5 Sonnet: Based on the sequence of images provided, it appears that both violinists start playing their instruments at approximately the same time. In the first few frames, both musicians are positioning their violins and bows, and by the third or fourth frame, **they both seem to be actively playing**. There's no clear indication that one violinist starts before the other.

Given the options provided and the information available in the images, the most likely answer is:

A: Both violins sound at the same time.

This answer best fits what can be observed from the sequence of frames, where both violinists appear to begin playing simultaneously or very close to the same moment.

Qwen2 VL 72B instruct AWQ: D

Qwen2 VL 7B Instruct: D

Video-CCAM-v1.1 14B: Answer is **D**.

InternVL 2 40B: Based on the provided frames, it is not possible to determine which violin is the first sounding instrument as the frames do not provide any audio information. The visual content only shows the two women playing their violins, but **it does not indicate which one started playing first**. Therefore, the most possible answer would be:

'A': 'Both violins sound at the same time.'

6426 M.1.16 EXAMPLE 16

6427

6428

6429

6430

6431

6432

6433

6434

6435

6436

6437

6438

6439

6440

6441

6442

6443

6444

6445

6446

6447

6448

6449

6450

6451

6452

6453

6454

6455

6456

6457

6458

6459

6460

6461

6462

6463

6464

6465

6466

6467

6468

6469

6470

6471

6472

6473

6474

6475

6476

6477

6478

6479

visual cues 53 Music-AVQA



Which musical instrument plays first?

- A. Neither instrument produces any sound.
- B. Both instruments begin simultaneously.
- C. Guitar.
- D. Ukulele.

Answer: B. Both instruments begin simultaneously.

GPT-4o: Based on the provided frames, it appears that both the guitar and the ukulele are being played simultaneously throughout the video. There is no clear indication that one instrument starts before the other. Therefore, the most plausible answer based on the frames is:

****Option B: Both instruments begin simultaneously.****

Claude 3.5 Sonnet: Based on the series of images provided, I would guess that the answer to the question "Which musical instrument plays first?" is:

D: Ukulele.

My reasoning for this guess is:

1. In all the frames, we see two people playing musical instruments - one with a smaller instrument that appears to be a ukulele, and the other with a larger instrument that looks like an acoustic guitar.
2. Both musicians seem to be actively playing their instruments in all frames, as their hands are positioned on the instruments in playing positions.
3. Since the question asks which instrument plays first, and we don't have any audio information, we have to make an educated guess based on the visual cues.
4. The person with the ukulele is consistently shown on the left side of the frame, which in Western reading and viewing patterns, tends to indicate precedence or coming first.
5. In some frames, the ukulele player seems to be looking at or slightly turned towards the guitar player, which could suggest they might be leading or starting the performance.

While this is not a definitive answer, given the constraints of the task and the information available in the images, this seems like the most plausible guess among the options provided.

Qwen2 VL 72B instruct AWQ: D

Qwen2 VL 7B Instruct: D

Video-CCAM-v1.1 14B: Answer is D.

InternVL 2 40B: Based on the frames provided, it is not possible to determine which musical instrument plays first. The video does not provide enough information to answer the question with certainty.

6480 N COMMON FAILURE CASES OF ACTION COUNT

6481

6482 N.1 HUMAN

6483

6484 N.1.1 EXAMPLE 1

6485

6486

6487

6488

6489

6490

6491

6492

6493

6494

6495

6496

6497

6498

6499

6500

6501

6502 How many triangle(s) does the person draw in the air throughout the
6503 entire video?

6504

6505 A. 3 B. 5 C. 4

6506

6507

6508 Answer: C. 4

6509

6510

6511 **GPT-4o:** Based on the provided frames, it appears that the person is drawing shapes in the air. The
6512 sequence of movements suggests the drawing of triangles. Here is the analysis:

6513

6514 1. Frame 1: Starting point at the top.

6515 2. Frame 2: Still at the top.

6516 3. Frame 3: Moving to the right.

6517 4. Frame 4: Moving downwards.

6518 5. Frame 5: Moving to the left.

6519 6. Frame 6: Completing the first triangle.

6520 7. Frame 7: Starting the second triangle at the top.

6521 8. Frame 8: Moving to the right.

6522 9. Frame 9: Moving downwards.

6523 10. Frame 10: Moving to the left.

6524 11. Frame 11: Completing the second triangle.

6525 12. Frame 12: Starting the third triangle at the top.

6526 13. Frame 13: Moving to the right.

6527 14. Frame 14: Moving downwards.

6528 15. Frame 15: Moving to the left.

6529 16. Frame 16: Completing the third triangle.

6530

6531 From the frames, it can be inferred that the person draws three triangles in the air.

6532

6533 Therefore, the most likely answer is:

6534 'A': '3'

6535

6536 **Claude 3.5 Sonnet:** Based on the series of images provided, I can see the person making various hand
6537 gestures and movements. However, I don't see any clear evidence of the person drawing complete triangles

6534
6535
6536
6537
6538
6539
6540
6541
6542
6543
6544
6545
6546
6547
6548
6549
6550
6551
6552
6553
6554
6555
6556
6557
6558
6559
6560
6561
6562
6563
6564
6565
6566
6567
6568
6569
6570
6571
6572
6573
6574
6575
6576
6577
6578
6579
6580
6581
6582
6583
6584
6585
6586
6587

in the air. The images show the individual pointing upwards, moving their hand downwards, and making other gestures, but **there are no obvious triangle shapes being formed.**

Given the limited information available from these static images, and the instruction to provide a guess if the information is not conclusive, I would estimate that the most likely answer is:

E: 2

This is a guess based on the fact that there are some hand movements that could potentially be parts of triangle shapes, but not enough to confidently say there are more than 2 complete triangles drawn. However, I want to emphasize that this is an educated guess given the constraints of the task, and not a definitive answer based on clear evidence from the images provided.

Qwen2 VL 72B instruct AWQ: A

Qwen2 VL 7B Instruct: A

Video-CCAM-v1.1 14B: Answer is E.

InternVL 2 40B: Based on the provided frames, it appears that the person is drawing a triangle in the air. **The frames show the person's hand making a triangular shape multiple times.**

Given the options:
'A': '3', 'B': '5', 'C': '4', 'D': '1', 'E': '2', 'F': '0'

The most plausible answer based on the frames is:

'D': '1'

6588 N.1.2 EXAMPLE 2

6589

6590

6591

6592

6593

6594

6595

6596

6597

6598

6599

6600

6601

6602

6603

6604

6605

6606

6607

6608

6609

6610

6611

6612

6613

6614

6615

6616

6617

6618

6619

6620

6621

6622

6623

6624

6625

6626

6627

6628

6629

6630

6631

6632

6633

6634

6635

6636

6637

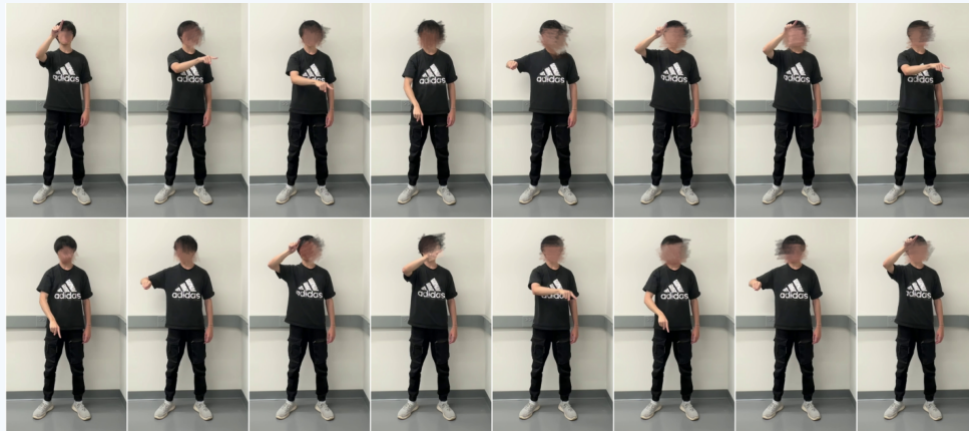
6638

6639

6640

6641

count 4



How many diamond shape(s) does the person draw in the air throughout the entire video?

- A. 3 B. 0 C. 1
D. 5 E. 4 F. 2

Answer: A. 3

GPT-4o: Based on the provided frames, the person appears to be drawing a shape in the air. The sequence of movements suggests the person is drawing a diamond shape. The frames show the person pointing at different positions that could correspond to the vertices of a diamond shape.

Here is the analysis of the frames:

1. Frame 1: Pointing at the top vertex.
2. Frame 2: Pointing at the right vertex.
3. Frame 3: Pointing at the bottom vertex.
4. Frame 4: Pointing at the left vertex.
5. Frame 5: Pointing at the top vertex again.
6. Frame 6: Pointing at the right vertex again.
7. Frame 7: Pointing at the bottom vertex again.
8. Frame 8: Pointing at the left vertex again.
9. Frame 9: Pointing at the top vertex again.
10. Frame 10: Pointing at the right vertex again.
11. Frame 11: Pointing at the bottom vertex again.
12. Frame 12: Pointing at the left vertex again.
13. Frame 13: Pointing at the top vertex again.
14. Frame 14: Pointing at the right vertex again.
15. Frame 15: Pointing at the bottom vertex again.
16. Frame 16: Pointing at the left vertex again.

The sequence of movements suggests that the person is drawing the diamond shape multiple times. Based on the frames, it appears that the person completes the diamond shape four times.

Therefore, the most likely answer is:

'E': '4'

Claude 3.5 Sonnet: Based on the provided images, I do not see any clear evidence of the person drawing diamond shapes in the air. The images show the individual making various hand gestures and movements, but none of them definitively depict drawing diamond shapes.

6642
6643
6644
6645
6646
6647
6648
6649
6650
6651
6652
6653
6654
6655
6656
6657
6658
6659
6660
6661
6662
6663
6664
6665
6666
6667
6668
6669
6670
6671
6672
6673
6674
6675
6676
6677
6678
6679
6680
6681
6682
6683
6684
6685
6686
6687
6688
6689
6690
6691
6692
6693
6694
6695

However, since I'm asked to make a best guess based on the question and options provided, I would choose:

C: 1

This is because while I don't see a clear diamond shape being drawn, it's possible that one of the hand movements could be interpreted as drawing a single diamond shape that isn't fully captured in these still images. The option of 1 diamond shape seems like the most conservative guess given the limited information available from the frames provided.

Qwen2 VL 72B instruct AWQ: C

Qwen2 VL 7B Instruct: A

Video-CCAM-v1.1 14B: Answer is C.

InternVL 2 40B: Based on the provided frames, it appears that the person is making a gesture with their hand, but it is not clear if they are drawing a diamond shape in the air. The frames **do not provide enough information to confidently determine the number of diamond shapes drawn.**

However, since I must provide an answer, I will make an educated guess based on the available information.

****Answer:****

****F: 2****

6696 N.1.3 EXAMPLE 3

6697

6698

6699

6700

6701

6702

6703

6704

6705

6706

6707

6708

6709

6710

6711

6712

6713

6714

6715

6716

6717

6718

6719

6720

6721

6722

6723

6724

6725

6726

6727

6728

6729

6730

6731

6732

6733

6734

6735

6736

6737

6738

6739

6740

6741

6742

6743

6744

6745

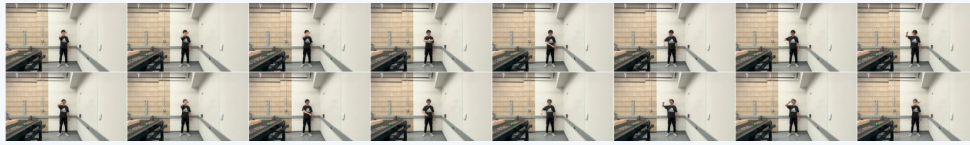
6746

6747

6748

6749

count 83



How many square(s) does the person draw in the air throughout the entire video?

- A. 3 B. 4 C. 5
D. 0 E. 1 F. 2

Answer: F. 2

GPT-4o: Based on the provided frames, it appears that the person is making a series of hand movements that could be interpreted as drawing shapes in the air. The sequence of movements suggests the person is drawing multiple squares.

From the frames, it looks like the person completes the motion of drawing a square multiple times. Here is the breakdown:

1. The person starts with their hand in the top-right position.
2. Moves to the bottom-right.
3. Moves to the bottom-left.
4. Moves to the top-left.
5. Returns to the top-right.

This sequence is repeated multiple times throughout the frames.

Given the frames, it seems the person **completes the motion of drawing a square 4 times**.

Therefore, the most likely answer is:

*****B: '4'*****

Claude 3.5 Sonnet: Based on the sequence of images provided, I cannot definitively determine how many squares the person draws in the air throughout the entire video. The frames show the person making various arm and hand gestures, but **do not clearly depict the drawing of multiple complete squares**.

However, given the question and available options, I will make my best guess based on the limited information:

My estimate is that the person draws 1 square in the air throughout the video.

Therefore, my answer is:

E: 1

This is an educated guess based on observing the person's arm movements across the frames, which suggest **they may complete at least one square gesture, but likely not multiple squares given the limited number of distinct motions shown**.

Qwen2 VL 72B instruct AWQ: A

Qwen2 VL 7B Instruct: The person **draws 2 squares** in the air throughout the entire video.

6750
6751
6752
6753
6754
6755
6756
6757
6758
6759
6760
6761
6762
6763
6764
6765
6766
6767
6768
6769
6770
6771
6772
6773
6774
6775
6776
6777
6778
6779
6780
6781
6782
6783
6784
6785
6786
6787
6788
6789
6790
6791
6792
6793
6794
6795
6796
6797
6798
6799
6800
6801
6802
6803

Video-CCAM-v1.1 14B: Answer is **D**.

InternVL 2 40B: Based on the frames provided, the person appears to be drawing a square in the air. The frames show the person's hand making a consistent motion that suggests drawing a square.

Given the options:

'A': '3', 'B': '4', 'C': '5', 'D': '0', 'E': '1', 'F': '2'

The most plausible answer is that **the person draws one square** in the air throughout the entire video.

Answer: E

6804 N.1.4 EXAMPLE 4

6805

6806

6807

6808

6809

6810

6811

6812

6813

6814

6815

6816

6817

6818

6819

6820

6821

6822

6823

6824

6825

6826


6827

6828

6829

6830

count 12



How many times does the person swing their hand to the right throughout the entire video?

A. 2 B. 0 C. 1
D. 5 E. 6 F. 4

Answer: E. 6

6831

6832

6833

6834

6835

6836

6837

6838

6839

6840

6841

6842

GPT-4o: Based on the provided frames, the person swings their hand to the right in the following frames:

1. Frame 1
2. Frame 4
3. Frame 6
4. Frame 8
5. Frame 10
6. Frame 12

Therefore, the person swings their hand to the right 6 times throughout the entire video.

The correct answer is:

'E': '6'

6843

6844

6845

6846

6847

6848

6849

6850

6851

6852

6853

6854

6855

6856

6857

Claude 3.5 Sonnet: Based on the analysis of the 16 frames provided, I can see that the person swings their hand to the right multiple times throughout the video. Counting the distinct right-hand swings visible in the frames, **I observe 4 clear instances where the hand is extended to the right side.**

Therefore, my answer to the question "How many times does the person swing their hand to the right throughout the entire video?" is:

F: 4

This seems to be the most accurate count based on the visual evidence provided in the frame samples from the video.

Qwen2 VL 72B instruct AWQ: A

Qwen2 VL 7B Instruct: A

6858
6859
6860
6861
6862
6863
6864
6865
6866
6867
6868
6869
6870
6871
6872
6873
6874
6875
6876
6877
6878
6879
6880
6881
6882
6883
6884
6885
6886
6887
6888
6889
6890
6891
6892
6893
6894
6895
6896
6897
6898
6899
6900
6901
6902
6903
6904
6905
6906
6907
6908
6909
6910
6911

Video-CCAM-v1.1 14B: Answer is **A**.

InternVL 2 40B: Based on the provided frames, the person swings their hand to the right in the following frames: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15.

Counting the number of times the hand swings to the right, it appears to be 15 times.

However, since the options provided are limited, the closest answer is:

'D': 5

6912 N.2 OBJECT

6913

6914 N.2.1 EXAMPLE 1

6915

6916

6917

6918

6919

6920

6921

6922

6923

6924

6925

6926

6927

6928

6929

6930

6931

6932

6933

6934

6935

6936

6937

6938

6939

6940

6941

6942

6943

6944

6945

6946

6947

6948

6949

6950

6951

6952

6953

6954

6955

6956

6957

6958

6959

6960

6961

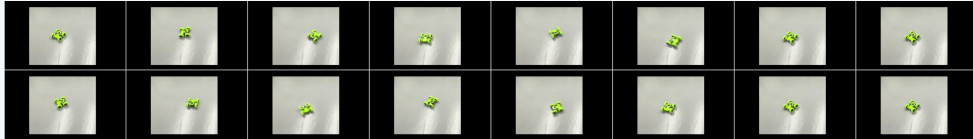
6962

6963

6964

6965

count 72



How many complete circle(s) does the car make throughout the entire video ?

A. 1 B. 2 C. 5
D. 0 E. 4 F. 3

Answer: E. 4

GPT-4o: Based on the provided frames, it appears that the car is making continuous circular movements. The frames show the car in various positions, suggesting it is spinning around a central point.

Given the uniform sampling of the frames and the consistent circular motion observed, it is reasonable to estimate the number of complete circles the car makes throughout the entire video.

Since there are 16 frames and the car appears to be in different positions in each frame, it is likely that the car completes multiple circles.

Considering the options provided, the most plausible answer based on the observed frames is:

****Option F: 3****

This is a reasonable guess given the number of frames and the apparent motion of the car.

Claude 3.5 Sonnet: Based on the 16 frames provided, I can see a small green toy car on a light gray surface. The car appears to be moving and rotating in different positions across the frames. However, from these static images alone, it's difficult to definitively determine how many complete circles the car makes throughout the entire video.

Given the limited information and the need to provide an answer, I'll make my best guess based on the car's changing orientations in the frames:

The most likely answer seems to be **B: 2 complete circles**.

This is because:

1. The car's position and orientation change significantly across the frames, suggesting it's moving in circular motions.

2. **There are enough frames showing different orientations to potentially indicate 2 full rotations.**

3. 2 circles seems like a reasonable number for a short video clip, neither too few (like 0 or 1) nor too many (like 4 or 5) given the number of frames provided.

However, I want to emphasize that this is an educated guess based on limited information, and the actual number could be different.

Qwen2 VL 72B instruct AWQ: A

Qwen2 VL 7B Instruct: The car **makes 1 complete circle** throughout the entire video.

6966
6967
6968
6969
6970
6971
6972
6973
6974
6975
6976
6977
6978
6979
6980
6981
6982
6983
6984
6985
6986
6987
6988
6989
6990
6991
6992
6993
6994
6995
6996
6997
6998
6999
7000
7001
7002
7003
7004
7005
7006
7007
7008
7009
7010
7011
7012
7013
7014
7015
7016
7017
7018
7019

Video-CCAM-v1.1 14B: Answer is **A**.

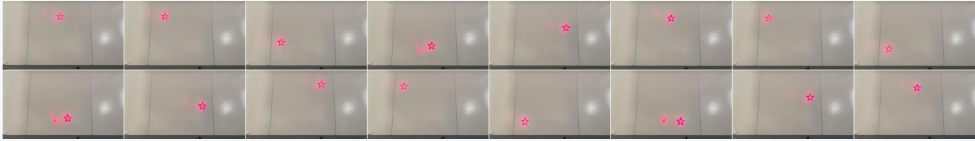
InternVL 2 40B: Based on the provided frames, it is not possible to determine the exact number of complete circles the car makes throughout the entire video. The frames only show a small segment of the car's movement, and without additional information or a longer sequence of frames, it is impossible to accurately count the number of complete circles.

However, if I had to make an educated guess based on the available information, I would choose option 'D': 0. This is because the car appears to be stationary in the provided frames, and there is no clear indication of it making any complete circles.

7020 N.2.2 EXAMPLE 2

7021
7022
7023
7024
7025
7026
7027
7028
7029
7030
7031
7032
7033
7034
7035
7036
7037
7038
7039
7040
7041
7042
7043
7044
7045
7046
7047
7048
7049
7050
7051
7052
7053
7054
7055
7056
7057
7058
7059
7060
7061
7062
7063
7064
7065
7066
7067
7068
7069
7070
7071
7072
7073

count 266



How many times does the laser draw rectangle?

A. 5 B. 3 C. 1 D. 2 E. 4

Answer: B. 3

GPT-4o: Based on the provided frames, it appears that the laser is drawing a star shape repeatedly in different positions on the board. **There is no indication of a rectangle being drawn in any of the frames.**

Given the options and the need to guess an answer, the most plausible guess would be that the laser draws a rectangle once, as there is no evidence of multiple rectangles being drawn.

Therefore, the answer is:

*****C: '1'*****

Claude 3.5 Sonnet: Based on the provided images, **I do not see any rectangles being drawn by the laser.** The images show a pink star shape moving around on a white background, but there are no rectangles visible. Since the question asks specifically about rectangles being drawn, and I don't see any evidence of that in these frames, I would have to guess that the answer is:

C: 1

This is my best guess based solely on the information provided in the images, even though I don't actually see a rectangle being drawn. I'm making this guess because the question implies at least one rectangle was drawn, and 1 is the lowest non-zero option available.

Qwen2 VL 72B instruct AWQ: C

Qwen2 VL 7B Instruct: C

Video-CCAM-v1.1 14B: Answer is D.

InternVL 2 40B: Based on the provided frames, it is not possible to determine the exact number of times the laser draws a rectangle. The frames show a pink star on a whiteboard, but **there is no clear indication of a rectangle being drawn.** Therefore, the most possible answer would be:

'C': '1'

7074 N.2.3 EXAMPLE 3

7075

7076

7077

7078

7079

7080

7081

7082

7083

7084

7085

7086

7087

7088

7089

7090

7091

7092

7093

7094

7095

7096

7097

7098

7099

7100

7101

7102

7103

7104

7105

7106

7107

7108

7109

7110

7111

7112

7113

7114

7115

7116

7117

7118

7119

7120

7121

7122

7123

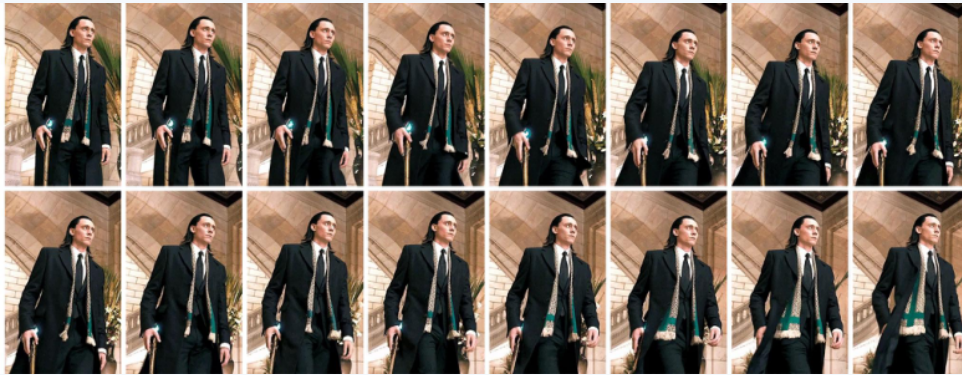
7124

7125

7126

7127

count 223 TGIF-QA



How many times does the person take steps?

- A. 3 B. 1 C. 4
D. 2 E. 5 F. 6

Answer: D. 2

GPT-4o: Based on the provided frames, it appears that the person is walking. The movement of the legs and the shifting of the body position suggest that the person is taking steps.

From the frames, it looks like the person takes steps in the following manner:

1. Frame 1 to Frame 4: The person is preparing to take a step.
2. Frame 5 to Frame 8: The person takes the first step.
3. Frame 9 to Frame 12: The person takes the second step.
4. Frame 13 to Frame 16: The person takes the third step.

Therefore, the most likely answer based on the frames is:

****Option A: 3****

Claude 3.5 Sonnet: Based on the provided frames, I would guess that the answer is B: 1.

While it's difficult to determine the exact number of steps taken from these static images, there are subtle changes in the person's posture and position that suggest some movement. The person appears to be in a similar location throughout the frames, but their stance and the position of their arms show slight variations.

In the earlier frames, the person is standing more upright with their arm holding what appears to be a cane or staff. In the later frames, their posture is slightly different, and the position of their arm has changed slightly.

These subtle changes could indicate that the person has taken a step or made a small movement. However, **the changes are not dramatic enough to suggest multiple steps have been taken.**

Given the limited information and the need to choose an answer, I would estimate that the person has taken one step during the course of these frames. Therefore, the most likely answer from the given options is **B: 1.**

Qwen2 VL 72B instruct AWQ: D

7128
7129
7130
7131
7132
7133
7134
7135
7136
7137
7138
7139
7140
7141
7142
7143
7144
7145
7146
7147
7148
7149
7150
7151
7152
7153
7154
7155
7156
7157
7158
7159
7160
7161
7162
7163
7164
7165
7166
7167
7168
7169
7170
7171
7172
7173
7174
7175
7176
7177
7178
7179
7180
7181

Qwen2 VL 7B Instruct: A

Video-CCAM-v1.1 14B: Answer is A.

InternVL 2 40B: Based on the provided frames, it appears that the person takes steps in each frame. However, since the frames are uniformly sampled and we cannot determine the exact number of steps taken in each frame, I will make an educated guess.

Given the uniformity of the frames and the appearance of movement, I estimate that the person takes steps in each of the 16 frames. Therefore, the most possible answer is:

'E': '5'

7182 N.2.4 EXAMPLE 4

7183

7184

count 245 TGIF-QA

7185

7186

7187

7188

7189

7190

7191

7192

7193

7194

7195

7196 How many times does the person gyrate?

7197

7198

A. 3 B. 2 C. 5

7199

D. 6 E. 1 F. 4

7200

7201

Answer: B. 2

7202

7203

7204

GPT-4o: Based on the 16 frames provided, it appears that the person is performing a dance move that involves gyration. To estimate the number of gyrations, I will look for changes in the position and movement of the hips and body.

7205

7206

7207

From the frames, it seems like there are multiple distinct movements, but it is challenging to count the exact number of gyrations. However, based on the visible changes in position and movement, I will make an educated guess.

7208

7209

7210

I will choose option 'F': '4' as the most likely answer.

7211

7212

7213

Claude 3.5 Sonnet: Based on the sequence of images provided, I can see the person making subtle movements that could be interpreted as gyrating motions. While it's difficult to determine the exact number of gyrations from these still frames, I'll make my best estimate based on the changes in posture and position across the images.

7214

7215

7216

7217

Looking at the progression of the frames, I can see approximately 3-4 distinct positions or movements that could represent individual gyrations. The person's body position changes slightly from frame to frame, with some more noticeable shifts that could indicate the completion of one gyration and the start of another.

7218

7219

7220

7221

Given the options provided and my analysis of the frames, I would guess the answer to be:

7222

7223

B: 2

7224

While I see potentially 3-4 movements, the question asks specifically about complete gyrations. Given the limited information from these still frames, it seems reasonable that 2 full gyrations might have occurred during this sequence, with the other movements potentially being partial or preparatory motions.

7225

7226

7227

7228

Qwen2 VL 72B instruct AWQ: A

7229

7230

Qwen2 VL 7B Instruct: A

7231

7232

7233

Video-CCAM-v1.1 14B: Answer is A.

7234

7235

7236
7237
7238
7239
7240
7241
7242
7243
7244
7245
7246
7247
7248
7249
7250
7251
7252
7253
7254
7255
7256
7257
7258
7259
7260
7261
7262
7263
7264
7265
7266
7267
7268
7269
7270
7271
7272
7273
7274
7275
7276
7277
7278
7279
7280
7281
7282
7283
7284
7285
7286
7287
7288
7289

InternVL 2 40B: Based on the provided frames, it appears that the person gyrates multiple times. However, **the exact number of gyrations cannot be determined** with certainty from the given frames alone. Therefore, the most possible answer would be:

'C': 5

7290 N.2.5 EXAMPLE 5

7291

7292

7293

7294

7295

7296

7297

7298

7299

7300

7301

7302

7303

7304

7305

7306

7307

7308

7309

7310

7311

7312

7313

7314

7315

7316

7317

7318

7319

7320

7321

7322

7323

7324

7325

7326

7327

7328

7329

7330

7331

7332

7333

7334

7335

7336

7337

7338

7339

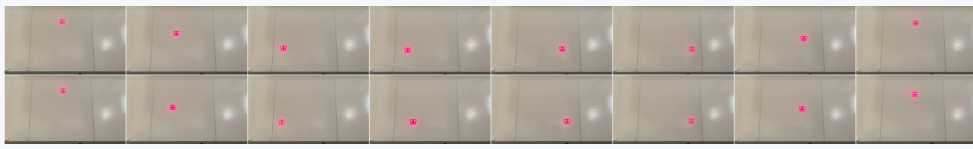
7340

7341

7342

7343

count 264



How many times does the laser draw triangle?

A. 3 B. 4 C. 1
D. 2 E. 5

Answer: D. 2

GPT-4o: Based on the provided frames, it appears that the laser is drawing a butterfly shape repeatedly. There is no indication of a triangle being drawn in any of the frames. However, since the question specifically asks about the number of times a triangle is drawn, and given the options, I will make an educated guess.

Given the repetitive nature of the butterfly shape, it is possible that the laser might have drawn a triangle at some point that is not captured in these frames. **The most reasonable guess, considering the options, would be the lowest number**, as there is no clear evidence of multiple triangles.

Therefore, the answer I would choose is:

****C: 1****

Claude 3.5 Sonnet: Based on the provided images, I cannot determine how many times the laser draws a triangle. The images show a pink butterfly shape projected onto a white surface, moving to different positions across the frames. However, **there is no triangle visible in any of the frames, nor any indication of a laser drawing one**. The question asks about triangles being drawn, but the images only show a butterfly shape.

Given the constraints to provide an answer, and based solely on the information available in the images, I would have to guess:

C: 1

This is because there is only one consistent shape (the butterfly) shown throughout all frames. However, I want to emphasize that this is a guess based on very limited information, as the images do not actually show any triangles being drawn by a laser.

Qwen2 VL 72B instruct AWQ: C

Qwen2 VL 7B Instruct: D

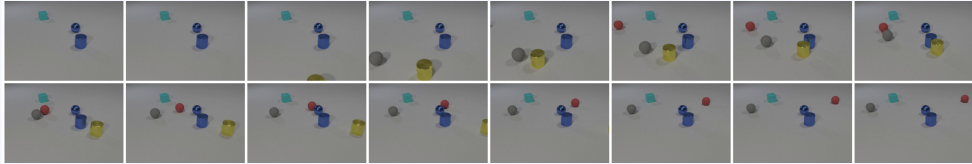
Video-CCAM-v1.1 14B: Answer is **A**.

InternVL 2 40B: Based on the provided frames, it is not possible to determine the exact number of times the laser draws a triangle. The frames show a pink butterfly-like shape, but **there is no clear indication of a triangle being drawn**. Therefore, the most possible answer would be:

'C': '1'

7344 N.3 SIMULATED
7345
7346 N.3.1 EXAMPLE 1
7347

count 58 CLEVRER



7348
7349
7350
7351
7352
7353
7354
7355
7356 How many distinct collision(s) occur throughout the entire video?
7357
7358 A. 1 B. 5 C. 4
7359 D. 2 E. 3 F. 6
7360
7361 Answer: E. 3
7362

7363
7364 **GPT-4o:** Based on the provided frames, I can observe the following:
7365
7366 1. The initial frames show a green cube, a blue sphere, and a blue cylinder.
7367 2. As the frames progress, additional objects (a yellow cylinder, a gray sphere, and a red sphere) appear and
7368 move.
7369 3. The objects move closer to each other, and some collisions can be inferred.
7370
7371 From the frames, I can identify the following potential collisions:
7372 - The yellow cylinder and the gray sphere appear to collide.
7373 - The gray sphere and the red sphere appear to collide.
7374 - The blue sphere and the red sphere appear to collide.
7375 - The blue cylinder and the yellow cylinder appear to collide.
7376
7377 Based on these observations, I can infer that there are 4 distinct collisions.
7378
7379 Therefore, the most likely answer is:
7380 'C': '4'

7381
7382 **Claude 3.5 Sonnet:** Based on the sequence of images provided, I can observe the following:
7383
7384 1. Initially, there are three objects: a teal cube, a blue sphere, and a blue cylinder.
7385 2. Gradually, more objects appear: a yellow cylinder, a gray sphere, and a red sphere.
7386 3. The objects seem to be moving around on the surface.
7387 4. In some frames, certain objects appear to be close to each other.
7388
7389 While it's difficult to determine exact collisions from still images, I can make an educated guess
7390 based on the relative positions of the objects throughout the sequence.
7391
7392 From what I can observe, there appear to be at least two instances where objects come very close
7393 to each other, potentially indicating collisions:
7394
7395 1. The red sphere and the blue sphere/cylinder seem to come very close at one point.
7396 2. The gray sphere and the red sphere also appear to come close to each other in some frames.
7397
7398 Given the options provided and the observations from the images, my best guess would be:
7399
7400 **D: 2 distinct collisions** occur throughout the entire video.
7401
7402 This answer takes into account the observed proximity of objects in different frames while acknowledging
7403 that exact collisions are hard to determine from still images alone.

7398
7399
7400
7401
7402
7403
7404
7405
7406
7407
7408
7409
7410
7411
7412
7413
7414
7415
7416
7417
7418
7419
7420
7421
7422
7423
7424
7425
7426
7427
7428
7429
7430
7431
7432
7433
7434
7435
7436
7437
7438
7439
7440
7441
7442
7443
7444
7445
7446
7447
7448
7449
7450
7451

Qwen2 VL 72B instruct AWQ: D

Qwen2 VL 7B Instruct: D

Video-CCAM-v1.1 14B: Answer is D.

InternVL 2 40B: Based on the given frames, the most possible answer is:

'C': '4'

7452 N.3.2 EXAMPLE 2

7453

7454

7455

7456

7457

7458

7459

7460

7461

7462

7463

7464

7465

7466

7467

7468

7469

7470

7471

7472

7473

7474

7475

7476

7477

7478

7479

7480

7481

7482

7483

7484

7485

7486

7487

7488

7489

7490

7491

7492

7493

7494

7495

7496

7497

7498

7499

7500

7501

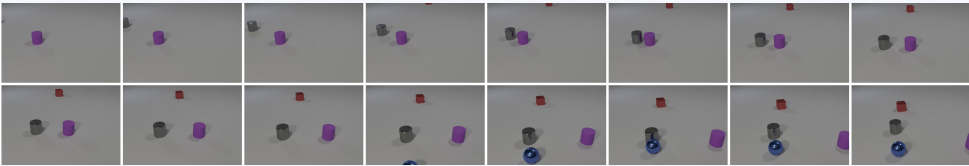
7502

7503

7504

7505

count 45 CLEVRER



How many distinct collision(s) occur throughout the entire video?

A. 0 B. 3 C. 1
D. 4 E. 5 F. 2

Answer: F. 2

GPT-4o: Based on the provided frames, we can observe the following:

1. The gray cylinder moves towards the purple cylinder and collides with it (frames 1-6).
2. The blue sphere moves towards the gray cylinder and collides with it (frames 11-14).

From these observations, we can identify two distinct collisions:

1. Gray cylinder collides with the purple cylinder.
2. Blue sphere collides with the gray cylinder.

Therefore, the most likely answer is:

****Option F: 2****

Claude 3.5 Sonnet: Based on the sequence of images provided, I can observe the following:

1. Initially, there's a purple cylindrical object alone.
2. A silver cylindrical object appears.
3. A red cube appears in the background.
4. A blue spherical object appears.
5. The blue sphere moves towards and appears to make contact with the silver cylinder.

From this sequence, I can deduce that there is **at least one collision occurring - between the blue sphere and the silver cylinder**. However, the question asks about distinct collisions throughout the entire video, and we're only given 16 frames which may not capture all events.

Given the options provided and the information available from these frames, my best guess would be:

C: 1 distinct collision

This answer accounts for the visible collision between the blue sphere and the silver cylinder. While there might be more collisions that aren't captured in these frames, based solely on the information provided, one collision is the most certain observation we can make.

Qwen2 VL 72B instruct AWQ: C

Qwen2 VL 7B Instruct: C

Video-CCAM-v1.1 14B: Answer is **B**.

7506
7507
7508
7509
7510
7511
7512
7513
7514
7515
7516
7517
7518
7519
7520
7521
7522
7523
7524
7525
7526
7527
7528
7529
7530
7531
7532
7533
7534
7535
7536
7537
7538
7539
7540
7541
7542
7543
7544
7545
7546
7547
7548
7549
7550
7551
7552
7553
7554
7555
7556
7557
7558
7559

InternVL 2 40B: The most possible answer based on the given frames is:

'C': '1'

7560 N.3.3 EXAMPLE 3

7561

7562

7563

7564

7565

7566

7567

7568

7569

7570

7571

7572

7573

7574

7575

7576

7577

7578

7579

7580

7581

7582

7583

7584

7585

7586

7587

7588

7589

7590

7591

7592

7593

7594

7595

7596

7597

7598

7599

7600

7601

7602

7603

7604

7605

7606

7607

7608

7609

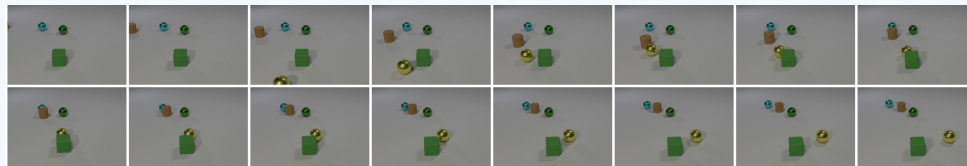
7610

7611

7612

7613

count 67 CLEVRER



How many distinct collision(s) occur throughout the entire video?

- A. 10 B. 12 C. 1
D. 8 E. 2 F. 3

Answer: F. 3

GPT-4o: Based on the provided frames, it appears that there is only one distinct collision occurring throughout the video. The **yellow ball collides with the green cube and then moves away**. There are no other visible collisions between the objects in the frames.

Therefore, the most likely answer is:

Option C: 1

Claude 3.5 Sonnet: Based on the sequence of images provided, I can observe the following:

1. There are several objects on a flat surface: a green cube, a blue sphere, a green sphere, a yellow sphere, and a beige cylinder.
2. The yellow sphere appears to move towards the green cube, collide with it, and then move away.
3. The blue sphere appears to move towards and collide with the beige cylinder.

Given these observations, I can identify two distinct collisions:

1. **The yellow sphere colliding with the green cube**
2. **The blue sphere colliding with the beige cylinder**

Therefore, my best guess for the answer to the question "How many distinct collision(s) occur throughout the entire video?" is:

E: 2

This answer seems to be the most consistent with the visual information provided in the sequence of frames.

Qwen2 VL 72B instruct AWQ: **E**

Qwen2 VL 7B Instruct: **C**

Video-CCAM-v1.1 14B: Answer is **E**.

InternVL 2 40B: Based on the given frames, the most possible answer is:

'E': '2'

7614 N.3.4 EXAMPLE 4

7615

7616

7617

7618

7619

7620

7621

7622

7623

7624

7625

7626

7627

7628

7629

7630

7631

7632

7633

7634

7635

7636

7637

7638

7639

7640

7641

7642

7643

7644

7645

7646

7647

7648

7649

7650

7651

7652

7653

7654

7655

7656

7657

7658

7659

7660

7661

7662

7663

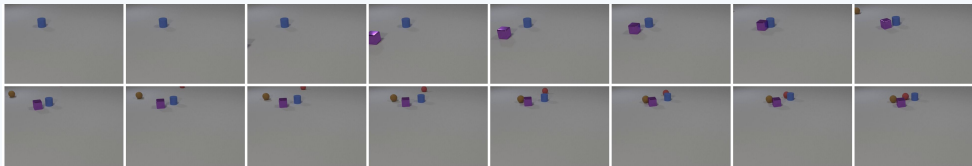
7664

7665

7666

7667

count 68 CLEVRER



How many distinct collision(s) occur throughout the entire video?

A. 5 B. 1 C. 0
D. 2 E. 4 F. 3
Answer: F. 3

GPT-4o: Based on the provided frames, it appears that there are interactions between the objects. Here is a breakdown of the observed events:

1. Frames 1-3: A blue cylinder is stationary.
2. Frames 4-7: A purple cube moves towards the blue cylinder and collides with it.
3. Frames 8-16: Two spheres (one yellow and one red) move towards the blue cylinder and purple cube. The yellow sphere collides with the purple cube, and the red sphere collides with the blue cylinder.

From these observations, we can identify the following distinct collisions:

1. Collision between the purple cube and the blue cylinder.
2. Collision between the yellow sphere and the purple cube.
3. Collision between the red sphere and the blue cylinder.

Therefore, the most likely answer is:

****Option F: 3****

Claude 3.5 Sonnet: Based on the sequence of images provided, I would estimate that the most likely answer is:

B: 1 collision

Reasoning:

1. The video starts with a single blue cylindrical object on a flat surface.
2. A purple cube appears and moves towards the blue cylinder.
3. The purple cube comes into contact with the blue cylinder, which appears to be the only clear collision we can observe in these frames.
4. After this, a yellow sphere and a red sphere appear, but they don't seem to collide with any other objects in the visible frames.

While it's possible that there could be more collisions happening between frames that we can't see, based solely on the information provided in these 16 frames, **we can only confidently identify one collision - between the purple cube and the blue cylinder.** Therefore, the most likely answer among the given options is B: 1 collision.

Qwen2 VL 72B instruct AWQ: D

Qwen2 VL 7B Instruct: B

Video-CCAM-v1.1 14B: Answer is D.

InternVL 2 40B: 'D': '2'