V2P-Bench: Evaluating Video-Language Understanding with Visual Prompts for Better Human-Model Interaction

Anonymous ACL submission

Abstract

Large Vision-Language Models (LVLMs) have made significant progress in the field of video understanding recently. However, current benchmarks uniformly lean on text prompts for evaluation, which often necessitate complex referential language and fail to provide precise spatial and temporal references. This limitation diminishes the experience and efficiency of human-model interaction. To address this limitation, we propose the **V**ideo **V**isual **P**rompt **Bench**mark (**V2P-Bench**), a comprehensive benchmark specifically designed to evaluate LVLMs' video understanding capabilities in multimodal human-model interaction scenarios. V2P-Bench includes 980 unique videos and 1,172 QA pairs, covering 5 main tasks and 12 dimensions, facilitating instance-level fine-grained understanding aligned with human cognition. Benchmarking results reveal that even the most powerful models perform poorly on V2P-Bench (65.4% for GPT-4o and 67.9% for Gemini-1.5-Pro), significantly lower than the human experts' 88.3%, highlighting the current shortcomings of LVLMs in understanding video visual prompts. We hope V2P-Bench will serve as a foundation for advancing multimodal human-model interaction and video understanding evaluation.

1 Introduction

In recent years, Large Vision-Language Models (LVLMs) have made significant progress in the field of video understanding, demonstrating powerful capabilities video captioning and question answering tasks, exemplified by recent Gemini-1.5-Pro (Team et al., 2024) and LLaVA-Video (Zhang et al., 2024). Correspondingly, numerous benchmarks have emerged to evaluate these models, covering a diverse range of videos and tasks (Li et al., 2024c; Mangalam et al., 2023; Fu et al., 2024), thereby providing robust support for the assessment of LVLMs from various perspectives.

However, most benchmarks utilize text prompts for human-model interaction, which inevitably introduces certain inherent limitations. As shown in Figure 1, text prompts usually fail to provide precise spatial and temporal references, resulting in difficulties when assessing the ability of LVLMs to understand specific areas or moments in videos, particularly in complex multi-object scenarios. For users, a significant amount of referential language is required to specify targets, which reduces the efficiency of human-model interaction. For the model, it first needs to comprehend the user's referential language, making it prone to confusion at this initial step.

In contrast, as a frontier approach to multimodal human-model interaction, visual prompts offer a simpler and more precise way, facilitating model understanding and aligning more closely with human intuitive cognition. Some previous efforts (Cai et al., 2024; Yang et al., 2023; Lin et al., 2024) conduct initial explorations in image visual prompt areas, demonstrating the superiority of visual prompts over texts. However, existing studies lack research on video modality, limiting the further development of multimodal human-model interaction.

To bridge this gap, we propose **V2P-Bench**, a comprehensive benchmark specifically designed to evaluate the video understanding capabilities of LVLMs in human-model interaction scenarios. As illustrated in Figure 2, V2P-Bench encompasses 5 main tasks, 12 categories, 20 video types and various types of visual prompts. Each query includes at least one visual prompt annotation, focusing on fine-grained spatial and temporal understanding consistent with human cognition, aiming to comprehensively assess the video understanding abilities of LVLMs. Furthermore, V2P-Bench consists of 980 videos and 1,172 question-answer(QA) pairs, with video durations ranging from 5 seconds to 2 hours. All videos are meticulously curated by

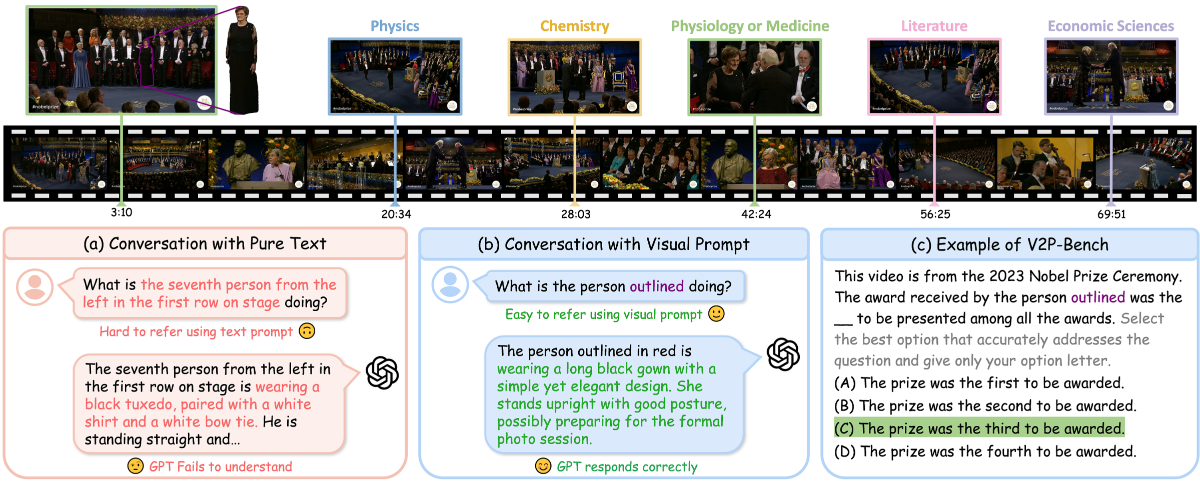

Figure 1: ***(a)(b)*** shows the comparisons of pure text prompting and visual prompting for video understanding. Simply overlaying visual prompts on video frames can enhance the user experience in Human-Model Interaction(HMI) while simultaneously reducing the difficulty for LVLMs to understand user intentions, particularly in complex environments where referential ambiguity is prevalent. ***(c)*** shows an example of V2P-Bench. The ground-truth answer is highlighted in green. Full video could be found at *youtu.be/lDlA7cfNk8A*.

human annotators to ensure high-quality QA pairs and accurate visual prompts.

In our experiments, we first execute the blinding answering evaluation on V2P-Bench to demonstrate that our benchmark avoids extensive prior knowledge of modern LVLMs. Subsequently, we conduct a comprehensive evaluation on 16 LVLMs, including 4 closed-source models and 12 open-source models. Additionally, our assessment incorporates scores from human experts. The evaluation results indicate that even the advanced closed-source models perform poorly on our benchmark (65.4% for GPT-4o (Hurst et al., 2024) and 67.9% for Gemini-1.5-Pro (Team et al., 2024), significantly lower than the human experts' score of 88.3%, revealing the current shortcomings of LVLMs in understanding video visual prompts. In a nutshell, our contributions are as follows:

- V2P-Bench has been meticulously designed, comprising 12 categories covering a wide range of video types and diverse visual prompts. Collection and annotation process undergoes rigorous human validation, aiming to provide the community with a high-quality benchmark for multi-model human-model interaction.

- We conduct extensive experiments and summarize our observations and insights, which demonstrate the current models' shortcomings in understanding video visual prompts and interacting with human.

- V2P-Bench pioneeringly applies visual prompts in video understanding evaluation for multimodal human-model interaction, addressing critical limitations in existing text-based evaluation frameworks. Through V2P-Bench, We seek to advance the field of video understanding evaluation and establish a foundation for more intuitive human-computer interaction.

2 Related Work

2.1 LVLMs for Video Understanding

The rapid development of image-based LVLMs (Liu et al., 2024b, 2023, 2024a; Li et al., 2024a; Chen et al., 2024a,e; Bai et al., 2023) has significantly enhanced the potential of video understanding and question answering tasks, injecting new vitality into the field of artificial intelligence. VideoChat (Li et al., 2023b) and Video-ChatGPT (Maaz et al., 2023) are preliminary attempts in the realm of video understanding. Notable recent works include CogVLM2-Video (Hong et al., 2024), InternVL2 (Chen et al., 2024e) and LLaVA-Video (Zhang et al., 2024), which treat videos as sequences of images and leverage the powerful image comprehension capabilities to process video modality. Furthermore, the high computational and memory demands required for handling high frame rates and long videos have spurred advancements in video compression technologies. For instance, InternVideo2 (Wang et al., 2024c) and Video-LLaMA

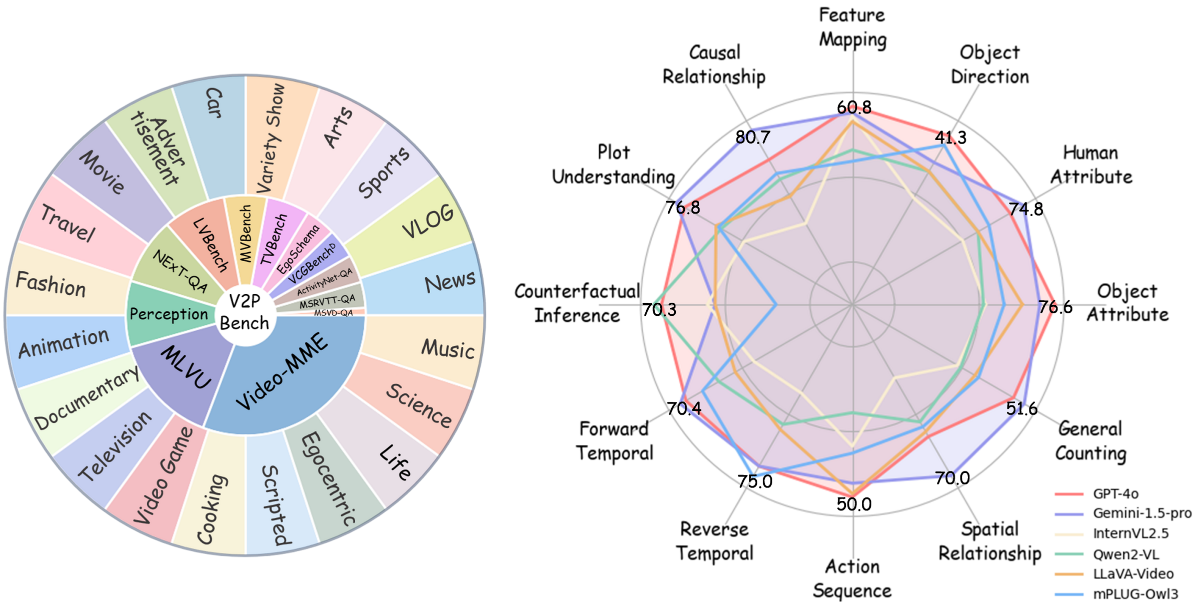

Figure 2: ***(Left)*** **Datasets and categories.** Our dataset is derived from 12 datasets and contains 20 restructured categories. ***(Right)*** **Performance radar chart.** We report the performance of different models on V2P-Bench by dimension. SOTA for each dimension is given.

(Zhang et al., 2023) utilize QFormer (Li et al., 2023a) for efficient video feature extraction, while PLLaVA (Xu et al., 2024) reduces computational load through adaptive pooling. Despite promising results, current video LLMs primarily rely on text prompts and still face challenges in fine-grained spatial and temporal understanding when given visual prompts as input.

2.2 Video Understanding Benchmarks

Traditional video understanding benchmarks, such as MSRVTT-QA (Xu et al., 2017), ActivityNet-QA (Yu et al., 2019), and NExT-QA (Xiao et al., 2021), focus on basic action recognition and video question answering, lack of sufficient detail and narrative to perform a fine-grained evaluation on LVLMs. Recently, more benchmarks have been proposed. MMBench (Liu et al., 2024c), SEED-Bench (Li et al., 2024b), and MVBench (Li et al., 2024c) mainly concentrate on short video clips for evaluation. EgoSchema (Mangalam et al., 2023) and MovieQA (Tapaswi et al., 2016) provide insights into narrative and thematic understanding. LongVideoBench (Wu et al., 2024), Video-MME (Fu et al., 2024), and LVBench (Wang et al., 2024b) offer longer videos and a broader variety of tasks. Additionally, recent works like INST-IT (Peng et al., 2024) and VideoRefer (Yuan et al., 2024) have introduced instance-level video question answering benchmarks. However, constrained by insufficiently robust and comprehensive, they still fail to adequately simulate real-world interactions. To address this limitation, we introduce V2P-Bench, allowing for a comprehensive evaluation of LVLMs that simulates multimodal human-model interaction in realistic settings.

2.3 Visual Prompt as a User-Friendly Solution

Compared to text prompts, visual prompts offer a simple and effective means of facilitating interaction between users and models. Visual prompts have been widely utilized in image understanding. ViP-LLaVA (Cai et al., 2024) enhances the ability of LVLMs to comprehend local image regions by overlaying arbitrary visual prompts on images. Draw-and-Understand (Lin et al., 2024) employs a two-stage training approach to improve performance in pixel-level tasks. Set-of-Mark (Yang et al., 2023) introduces a novel visual prompting method to enhance the performance of LVLMs in visual localization tasks. However, research on visual prompts in the context of video remains limited. INST-IT (Peng et al., 2024) introduces instruction tuning with visual prompts to enhance instance-level understanding in LVLMs. VideoRefer Suite (Yuan et al., 2024) creates a large instance-level video instruction dataset to assist LVLMs in understanding spatiotemporal information in videos.

3 V2P Bench

Table 1 compares the key difference of V2P-Bench with previous benchmarks. The first two blocks list traditional pure text video understanding benchmarks, which primarily understand videos at a

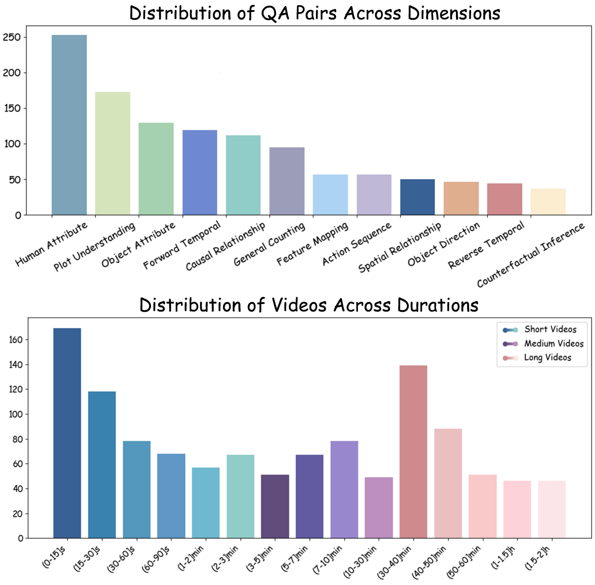

Figure 3: **Distribution of QA dimensions and video durations.** V2P-Bench encompasses 5 main tasks, 12 dimensions and covers a wide range of video lengths, enabling an comprehensive assessment of LVLMs.

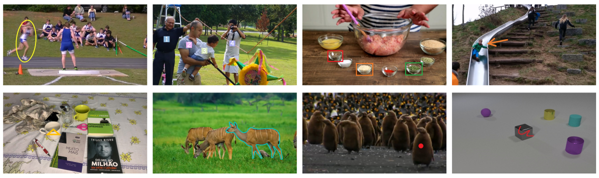

Figure 4: **Various visual prompt types.** We don't consider mask as it would significantly obscure the features of the target, even with a small alpha value (e.g. 0.2).

holistic level and lack instance-level comprehension. Instance-level understanding is crucial as it focuses on the specific elements of greatest interest to us, requiring a more nuanced understanding and consistency.

As shown in the third block, although INST-IT Bench (Peng et al., 2024) and VideoRefer Bench^Q (Yuan et al., 2024) use visual prompts for question-answering, their benchmarks are completely derived from VIS datasets (Yang et al., 2019; Pont-Tuset et al., 2017; Ding et al., 2023), which is insufficiently robust and comprehensive, resulting in: *1)* Shorter video durations(14.2s and 13.8s); *2)* Single continuous shots; *3)* Limited video sources, primarily comprising natural scenes; *4)* Objects of interest may not be suitable for question-answering. To address this limitation, we propose V2P-Bench, a comprehensive benchmark specifically designed to evaluate the video understanding capabilities of LVLMs in human-model interaction scenarios.

3.1 Task Definition

To facilitate fine-grained evaluation of LVLMs from various perspectives, we categorize the questions according to dimensions. Our dimension design strives to ensure both comprehensiveness and orthogonality, and ultimately includes five main tasks and 12 dimensions. Definitions for tasks and dimensions are as follows:

• **Perception** is a fundamental task that tests whether a model can understand visual prompts. This task includes: *1)* Object Attribute (OA); *2)* Human Attribute (HA); *3)* Object Direction (OD); *4)* Feature Mapping (FM).

• **Temporal** focuses on understanding and processing the chronological order of events in the video. This task includes: *1)* Forward Temporal (FT); *2)* Reverse Temporal (RT); *3)* Action Sequence (AS).

• **Reasoning** is an extension of the perception task, requiring logical inference and judgment based on given information to derive new conclusions or answers. This task includes: *1)* Causal Relationship (CR); *2)* Plot Understanding (PU); *3)* Counterfactual Inference (CI).

• **Spatial** focuses on the spatial relationships of the visual prompt targets. Using visual prompts to indicate spatial positions directly avoids the ambiguity and referential difficulties often encountered with text-based prompts, making interactions between users and models more intuitive. This task includes Spatial Relationship (SR).

• **Counting** focuses on the model's accurate identification and counting of the number of objects or events in the video. The model's objective is to perform effective quantity assessment based on the given visual prompt target. This task includes General Counting (GC).

Detailed elaborations and examples of each dimension are provided in Appendix A and B.

3.2 Dataset Construction

The construction process consists of three steps: video collection, QA and visual prompt annotation, post processing. Details of each step are as follows.

3.2.1 Video Collection

To create our dataset, we start from existing video benchmarks, as they already have a wide distribution of durations and diverse video types. We carefully select 12 benchmarks to construct our dataset. We follow Video-MME (Fu et al., 2024) and categorize the video durations into short, medium, and

Table 1: Comparison of different datasets. **Answer Type** indicates whether the QA pair is open-ended(OE) or multiple-choice(MC). **Multi Level** represents whether the videos cover multiple duration levels. **Open Domain** indicates whether the video source is diversified. **Visual Prompt** represents whether the video contains visual prompt annotations.

Benchmarks	Videos	Samples	Tasks	Avg duration	Annotation	Answer Type	Multi Level	Open Domain	Visual Prompt
MSVD-QA(Xu et al., 2017)	504	13157	1	9.8s	Auto	OE	✗	✗	✗
MSRVTT-QA(Xu et al., 2017)	2990	72821	1	15.2s	Auto	OE	✗	✗	✗
ActivityNet-QA(Yu et al., 2019)	800	8000	3	111.4s	Manual	OE	✗	✗	✗
NExT-QA(Xiao et al., 2021)	1000	8564	3	39.5s	Manual	MC	✗	✗	✗
Perception Test(Patraucean et al., 2024)	11600	44000	4	23.0s	Auto&Manual	MC	✗	✗	✗
MLVU(Zhou et al., 2024)	1334	2593	9	~12min	Auto&Manual	OE&MC	✓	✓	✗
VCGBench-Diverse(Maaz et al., 2024)	877	4354	6	217.0s	Auto&Manual	OE	✗	✓	✗
MVBench(Li et al., 2024c)	3641	4000	20	16.0s	Auto	MC	✗	✓	✗
HourVideo(Chandrasegaran et al., 2024)	500	12976	18	45.7min	Auto&Manual	MC	✗	✗	✗
LVBench(Wang et al., 2024b)	103	1549	6	68.4min	Manual	MC	✗	✓	✗
EgoSchema(Mangalam et al., 2023)	5063	5063	1	180.0s	Auto	MC	✗	✗	✗
Video-MME(Fu et al., 2024)	900	2700	12	17.0min	Manual	MC	✓	✓	✗
INST-IT Bench(Peng et al., 2024)	206	1000	1	14.2s	Auto&Manual	OE&MC	✗	✗	✓
VideoRefer Bench^{Q}(Yuan et al., 2024)	198	1000	5	13.8s	Manual	MC	✗	✗	✓
V2P-Bench(ours)	980	1172	12	19.0min	Manual	MC	✓	✓	✓

long videos. Additionally, we reclassify all the videos, resulting in 20 video categories, as shown in Figure 2(left). Our final dataset covers multiple video domains while maintaining a relative balance in video lengths.

3.2.2 QA and Visual Prompt Annotation

After completing the collection process, we conduct the annotation of QA pairs and visual prompts to evaluate the capabilities of LVLMs in video understanding with visual prompts. The annotation work is carried out by researchers proficient in English. To ensure data quality, we provide thorough training for the annotators and conduct pilot annotations to assess their annotation capabilities.

While annotating the QA pairs, annotators are also required to perform visual prompt annotations. To align with real-world distributions, we adopt a fully manual approach for annotating the video frames. We follow ViP-LLaVA (Cai et al., 2024) and predefined various types of visual prompts as follows: rectangle, mask contour, ellipse, triangle, scribble, point, arrow and SoM, just as shown in Figure 4. Additionally, annotators are allowed to exercise their creativity by using any type of visual prompts, not limited to the predefined types mentioned earlier.

3.2.3 Post Processing

To ensure the quality of the dataset, we conduct a rigorous review of the annotated data after completion, including both rule-based and manual review processes.

Blind LLMs Filtering. Inspired by MMStar (Chen et al., 2024b), we perform plain text filtering on the QA pairs to ensure that questions could only be answered correctly by viewing the videos. Specifically, we provide only the pure text QA pairs to the most powerful closed-source models GPT-4o (Hurst et al., 2024) and Gemini-1.5-pro (Team et al., 2024). We set the sampling temperature to 0.2 and conduct two rounds of inference, then exclude samples for which both rounds provided correct answers.

Visual Prompt Filtering. Secondly, we conduct visual prompt filtering on the QA pairs to exclude those questions that could be answered correctly without viewing the visual prompts. Specifically, we provide averaged 8 frames sampled from the video and the QA pairs to GPT-4o (Hurst et al., 2024), without visual prompt frames. We maintain the sampling temperature at 0.2 and perform two rounds of inference, excluding samples for which both rounds yielded correct answers.

Manual and Rule-based Review. After the previous two steps, we perform a rule-based check and manual review of the remaining data. We exclude samples where the length disparity between different options was too significant. Additionally, we shuffle the order of multiple-choice options to ensure a uniform distribution of answer choices, thereby eliminating potential biases of different models toward specific options. The final balanced proportions of the four options are 28.0%, 23.9%, 25.0% and 23.1%. Through this rigorous dataset construction process, we strive to provide a high-quality, diverse, and balanced dataset that will benefit researchers in the field of multimodal human-model interaction.

3.3 Benchmark Statistics

In Table 1, we have already presented the main characteristics of V2P-Bench. Overall, the proposed V2P-Bench defines 5 main tasks and 12 dimensions, encompassing 980 unique videos and 1,172 QA pairs sourced from 12 existing video datasets, covering 20 video categories. The average duration of the videos is 19.0 minutes, which represents a wide range of video lengths, differing from most benchmarks. The format of the QA pairs is multiple-choice with 4 options. Below we introduce a more detailed analysis of our benchmark:

- **Wide distribution of durations.** Figure 3(down) shows the detailed duration distributions on V2P-Bench. We follow Video-MME (Fu et al., 2024) in categorizing video lengths into short (< 3 minutes), medium (3-30 minutes), and long videos (30-120 minutes), with respective proportions of 46.8%, 22.0%, and 31.2%.

- **Diverse video types and comprehensive tasks.** Figure 2(left) shows various datasets and categories on V2P-Bench. We select videos from 12 existing video benchmarks, resulting in a total of 20 reorganized categories. Figure 3(up) shows the detailed distribution of each dimension.

- **Diverse Targets and Visual Prompts.** Figure 4 shows various targets and visual prompts on V2P-Bench, benefiting from diverse video sources.

- **Comprehensive Shot Types.** V2P-Bench includes both continuous and transition videos, the latter of which significantly increases the difficulty of reference, implying that the model must perform temporal and spatial grounding in different scenes.

4 Experiments

4.1 Experiment Setup

Evaluation Models. We evaluate the performance of 12 open-source models that support multi-image or video input. The model list is shown in the third block of Table 2. We sample a fixed number of frames from the original videos at regular intervals to accommodate the context length of the models. Specifically, we sample 16 frames for LLaVA-NeXT (Liu et al., 2024a), PLLaVA (Xu et al., 2024), ShareGPT4Video (Chen et al., 2024c), MiniCPM-V 2.6 (Yao et al., 2024), InternVL2 (Chen et al., 2024e), InternVL2.5 (Chen et al., 2024d) and Qwen2-VL (Wang et al., 2024a), 32 frames for VideoLLaMA2 (Cheng et al., 2024), 64 frames for LLaVA-OneVision (Li et al., 2024a), mPLUG-Owl3 (Ye et al., 2024), LLaVA-Video (Zhang et al., 2024) and LLaVA-NeXT-INST-IT (Peng et al., 2024). All models are evaluated on 8 V100 GPUs. Additionally, we conduct extensive evaluations on 4 closed-source models: GPT-4o (Hurst et al., 2024), GPT-4o-mini (Hurst et al., 2024), Gemini-1.5-Pro (Team et al., 2024), and Gemini-1.5-Flash (Team et al., 2024). For GPT models, we average 64 frames from the original videos; for Gemini-1.5 models, the raw videos were uploaded directly.

Human and Blind Answering. For human evaluations, we divide all the questions equally and assign them to three human experts. To prevent any data leakage, we ensure that the human experts participating in the test have never been involved in the annotation process. Furthermore, considering that LLMs possess extensive prior knowledge, enabling them to answer certain questions without analyzing the video, We report the performance of 4 models: GPT-4o (Hurst et al., 2024), Gemini-1.5-Pro (Team et al., 2024), Qwen2-VL (Wang et al., 2024a), and InternVL2.5 (Chen et al., 2024d) on the blind answering task.

4.2 Results on V2P-Bench

Overall Results. Table 2 and Table 3 presents the evaluation results on V2P-Bench across dimensions and durations, including human performance, blind answering task and 16 models, illustrating the performance of LVLMs in understanding video visual prompts.

As shown in the top of Table 2, human experts reach 88.3%, reporting the upper limit of human performance on V2P-Bench. For the blind answering task, results are shown in the first block of Figure 2. GPT-4o and Gemini-1.5-Pro reach 20.7% and 13.7%, respectively, as they decline to answer 38.9% and 66.7% of the questions,which indicates that our benchmark necessitates access to video content for effective performance. Qwen2-VL and InternVL2.5, adhere strictly to the instructions, even though they could not answer the questions solely through pure text. Consequently, they achieve 30.8% and 27.9%, respectively.

During the evaluation process, we observe that certain models (Cheng et al., 2024; Chen et al., 2024c) cannot generate only options, even when provided with carefully designed prompts. VideoLLaMA2 tends to repeat the entire set of options, while ShareGPT4Video consistently begins with "Answer:". We specifically account for these output characteristics in our analysis of these models.

Table 2: **Evaluation results on V2P-Bench across dimensions.** We report results for 12 open-source models, 4 closed-source models, 4 blind LLMs and human performance on V2P-Bench across dimensions. Gemini-1.5-Pro achieve optimal performance at 67.9%, remaining a significant gap to human performance. The best results are **bold** and the second-best are underlined.

Method	Size	OA	HA	OD	FM	CR	PU	CI	FT	RT	AS	SR	GC	Avg
Human Performance	-	92.2	91.7	84.8	89.5	85.7	83.2	91.9	87.4	84.1	75.4	92.0	95.8	88.3
Pure Text as Input														
GPT-4o(Hurst et al., 2024)	-	20.2	19.4	17.4	14.0	17.9	32.1	21.6	23.5	15.9	19.3	16.0	17.9	20.7
Gemini-1.5-pro(Team et al., 2024)	-	10.1	14.6	2.2	3.5	24.1	13.9	29.7	16.0	20.5	12.3	14.0	5.3	13.7
Qwen2-VL(Wang et al., 2024a)	7B	26.4	31.8	31.5	28.1	33.9	32.9	28.4	33.6	39.8	24.6	25.0	27.4	30.8
InternVL2.5(Chen et al., 2024d)	8B	30.2	26.9	21.7	21.1	30.4	32.9	27.0	29.4	27.2	22.8	24.0	27.4	27.9
Closed-source Models														
GPT-4o(Hurst et al., 2024)	-	**76.6**	68.9	**41.3**	**60.8**	67.0	73.3	67.6	68.1	70.5	**50.0**	54.0	48.4	65.4
GPT-4o-Mini(Hurst et al., 2024)	-	68.8	61.0	30.4	49.0	65.1	63.6	32.4	48.3	56.8	41.1	62.0	45.3	56.3
Gemini-1.5-Pro(Team et al., 2024)	-	70.9	**74.8**	34.8	58.8	**80.7**	**76.8**	48.6	**70.4**	70.5	46.4	**70.0**	51.6	**67.9**
Gemini-1.5-Flash(Team et al., 2024)	-	61.2	64.4	28.3	52.6	72.3	64.2	37.8	58.0	54.5	35.1	56.0	**52.1**	57.3
General Open-source Models														
LLaVA-NeXT(Liu et al., 2024a)	7B	56.6	55.6	34.8	52.5	43.0	48.6	31.6	42.6	42.2	28.1	42.0	30.5	46.0
LLaVA-NeXT-INST-IT(Peng et al., 2024)	7B	57.4	58.4	26.1	42.4	43.0	49.2	31.6	49.2	42.2	26.3	42.0	27.4	46.3
PLLaVA(Xu et al., 2024)	7B	45.7	48.2	21.7	45.6	39.3	54.9	24.3	47.1	45.5	28.1	40.0	28.4	43.0
LLaVA-OV(Li et al., 2024a)	7B	59.7	54.5	32.6	36.8	46.4	59.0	35.1	53.8	59.1	36.8	50.0	32.6	49.9
VideoLLaMA2(Cheng et al., 2024)	7B	47.3	45.8	26.1	45.6	41.1	52.0	35.1	44.5	50.0	29.8	44.0	32.6	43.4
ShareGPT4Video(Chen et al., 2024c)	8B	40.3	43.1	21.7	45.6	40.2	45.7	51.4	43.7	40.9	24.6	46.0	30.5	40.6
mPLUG-Owl3(Ye et al., 2024)	7B	57.4	59.7	39.1	43.9	60.7	58.4	27.0	61.3	**75.0**	38.6	50.0	37.9	54.3
LLaVA-Video(Zhang et al., 2024)	7B	64.3	54.9	32.6	56.1	50.0	59.5	48.6	47.9	54.5	49.1	52.0	36.8	52.6
MiniCPM-V 2.6(Yao et al., 2024)	8B	50.4	51.8	17.4	49.1	53.6	61.8	37.8	49.6	50.0	31.6	48.0	27.4	48.0
InternVL2(Chen et al., 2024e)	8B	48.1	47.8	23.9	35.1	42.9	51.4	59.5	42.0	36.4	28.1	46.0	24.2	42.7
InternVL2.5(Chen et al., 2024d)	8B	50.4	48.2	26.1	57.9	37.5	47.4	51.4	40.3	38.6	36.8	30.0	31.6	43.2
Qwen2-VL(Wang et al., 2024a)	7B	49.6	54.9	32.6	47.4	58.0	57.2	**70.3**	54.6	52.3	28.1	48.0	32.6	50.7

Table 3: **Evaluation results on V2P-Bench across durations.** The best results are **bold** and the second-best are underlined.

Method	Size	Short	Medium	Long	Avg
Human Performance	-	91.6	87.3	84.0	88.3
Pure Text as Input					
GPT-4o(Hurst et al., 2024)	-	18.2	31.6	18.1	20.7
Gemini-1.5-pro(Team et al., 2024)	-	12.0	19.6	12.7	13.7
Qwen2-VL(Wang et al., 2024a)	7B	31.5	38.6	24.7	30.8
InternVL2.5(Chen et al., 2024d)	8B	26.0	35.6	26.2	27.9
Closed Models					
GPT-4o(Hurst et al., 2024)	-	**67.3**	70.8	59.3	65.4
GPT-4o-mini(Hurst et al., 2024)	-	56.2	65.3	51.1	56.3
Gemini-1.5-pro(Team et al., 2024)	-	65.3	**82.3**	**64.1**	**67.9**
Gemini-1.5-Flash(Team et al., 2024)	-	55.1	69.8	52.4	57.3
General Open Models					
LLaVA-NeXT(Liu et al., 2024a)	7B	47.0	47.1	43.8	46.0
LLaVA-NeXT-INST-IT(Peng et al., 2024)	7B	48.6	51.1	39.5	46.3
PLLaVA(Xu et al., 2024)	7B	43.8	48.9	38.1	43.0
LLaVA-OV(Li et al., 2024a)	7B	51.6	57.3	42.7	49.9
VideoLLaMA2(Cheng et al., 2024)	7B	46.8	48.9	34.8	43.4
ShareGPT4Video(Chen et al., 2024c)	8B	45.4	41.8	32.4	40.6
mPLUG-Owl3(Ye et al., 2024)	7B	56.1	65.8	44.3	54.3
LLaVA-Video(Zhang et al., 2024)	7B	57.5	59.1	40.8	52.6
MiniCPM-V 2.6(Yao et al., 2024)	8B	47.5	56.9	43.5	48.0
InternVL2(Chen et al., 2024e)	8B	44.0	50.2	36.2	42.7
InternVL2.5(Chen et al., 2024d)	8B	46.2	43.1	38.4	43.2
Qwen2-VL(Wang et al., 2024a)	7B	53.3	54.6	44.0	50.7

Below we summarize our key findings as follows:

Expert models demonstrate mediocre performance. LLaVA-NeXT-INST-IT fine-tunes on a visual prompt dataset derived from LLaVA-NeXT. However, it achieves only a marginal improvement of 0.3% over LLaVA-NeXT's 46.0%, suggesting that the fine-tuning process is nearly ineffective. We attribute this to the model's inadequate robustness and comprehensiveness in video sources, as well as its reliance on a single type of visual prompt (SoM only), which constrains the model's generalization capabilities.

Even the most powerful closed-source models perform poorly on our benchmark. In the evaluation results, closed-source models GPT-4o (Hurst et al., 2024) and Gemini-1.5-Pro (Team et al., 2024) achieve only 65.4% and 67.0%, respectively, remaining a significant gap compared to human experts, which stand at 88.3%.

Some powerful LVLMs struggle on our benchmark. Some of the state-of-the-art LVLMs, such as the InternVL series, perform poorly on our V2P-Bench, with InternVL2 (Chen et al., 2024e) and InternVL2.5 (Chen et al., 2024d) achieving only 42.7% and 43.2%, respectively. We speculate that this may be due to the InternVL series not being trained on datasets relevant to visual prompts or their inability to adapt to the data organization format that grounds temporal and spatial cues from visual prompt frames to the original video.

General models can comprehend visual prompts without prior training. Excluding closed-source models, all open-source models, except for LLaVA-NeXT-INST-IT, have not undergone specialized training for visual prompts. However, experimental results indicate that they perform reasonably well on V2P-Bench, suggesting that general video understanding capabilities can be partially transferred to understanding video visual prompts. The strongest open-source model mPLUG-Owl3 achieves 54.3%, slightly behind GPT-4o-Mini and Gemini-1.5-Flash; while the weakest model ShareGPT4Video also attains 40.6%, which is 15.6% higher than the baseline of random guessing.

Some dimensions are quite challenging. V2P-Bench requires models to not only understand visual prompts but to make reasoned judgments based on the video and accompanying questions. The results indicate that models underperform in specific areas, such as Object Direction, General Counting, and Action Sequence, with optimal scores of only 41.3%, 52.1%, and 50.0%, respectively. These tasks are relatively abstract and place high demands on the models. Additionally, several open-source models achieve performance levels that approach or even surpass those of closed-source models, highlighting the considerable potential of open-source approaches.

Performance on short videos is unexpectedly poor. In Table 3, we categorize all videos into short, medium, and long durations and report the models' performance across different lengths. We observe that all closed-source models and some open-source models perform worse on short than on medium-length videos. This can be attributed to the fact that over half of the videos in short videos originates from Perception Test (Patraucean et al., 2024), MVBench (Li et al., 2024c), and TVBench (Cores et al., 2024), which feature numerous challenging abstract questions related to sequences and frequencies. Long videos generally exhibit a consistent trend, with all models showing lower performance on long videos compared to both short and medium-length videos.

4.3 Extra Findings

Due to the specificity of visual prompt frames, we conduct an extra experiment on different data formats, as there is currently no research on it. Based on the different positions of the visual prompt frames, data can be organized as ***Retrieval*** and ***Needle***. ***Retrieval*** refers to the sequential input of the original video, questions, and visual prompt frames; while ***Needle*** refers to embedding the visual prompt frames into the video. We annotate 265 timestamped data entries, and replace the visual prompt frames with their nearest neighboring frames based on the timestamps during preprocessing process, thus ensuring the presence of the visual prompt frames.

We conduct evaluations on GPT-4o and Gemini-1.5-Pro. As shown in Figure 5, both models exhibit slightly better performance in the Retrieval format compared to Needle format. Considering model performance and convenience of dataset release process, we opt for Retrieval format to construct V2P-Bench.

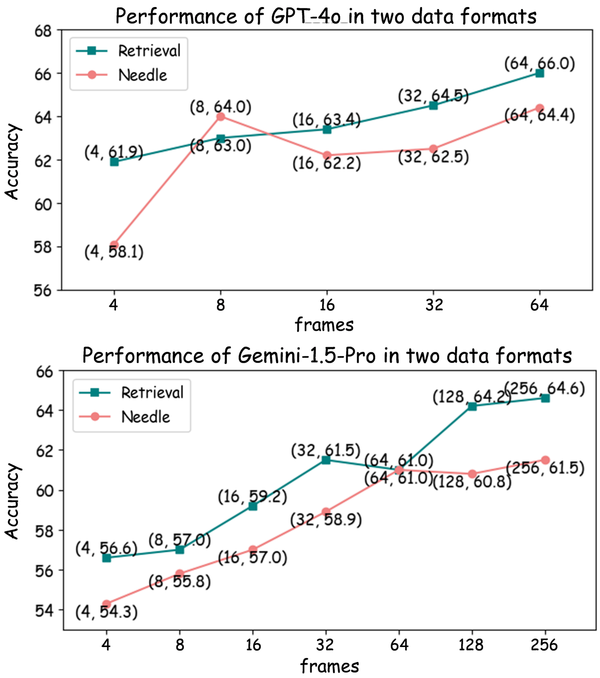

Figure 5: **Results on two data formats.** ***Retrieval*** refers to the sequential input of the original video, questions, and visual prompt frames; ***Needle*** refers to embedding the visual prompt frames into the video.

5 Conclusion

In this study, we introduce V2P-Bench, a comprehensive benchmark to evaluate the video understanding capabilities of LVLMs utilizing visual prompts in human-model interaction scenarios. Our dataset defines 5 main tasks and 12 dimensions, contains 980 unique videos, 1172 QA pairs, covering 20 video categories and a diverse range of videos. We conduct extensive experiments on V2P-Bench with 16 models, including 4 closed-source models and 12 open-source models. The experimental results indicate that even the most powerful model Gemini-1.5-Pro achieves only 67.9%, in stark contrast to the 88.3% demonstrated by human experts, highlighting a significant disparity. We aim to establish the V2P-Bench to advance the development of LVLMs in the field of video understanding and multimodal human-model interaction.

Limitations

Although our V2P-Bench comprehensively evaluates the capabilities of LVLMs in video-language understanding with visual prompts for multimodal human-model interaction, it only focuses on visual and textual inputs, lacking audio input, and supports evaluations only on offline videos, which leaves a gap compared to the ultimate form of multimodal human-model interaction in real world. We plan to develop a fully multimodal real-time human-model interaction benchmark in the future.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. 2024. Vip-llava: Making large multimodal models understand arbitrary visual prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12914–12923.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. 2024. Hourvideo: 1-hour video-language understanding. *arXiv preprint arXiv:2411.04998*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024a. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024b. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. 2024c. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024d. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024e. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees GM Snoek, and Yuki M Asano. 2024. Tvbench: Redesigning video-language evaluation. *arXiv preprint arXiv:2410.07752*.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. 2023. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2694–2703.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. 2024. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024b. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023b. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2024c. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Weifeng Lin, Xinyu Wei, Ruichuan An, Peng Gao, Bocheng Zou, Yulin Luo, Siyuan Huang, Shanghang Zhang, and Hongsheng Li. 2024. Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want. *arXiv preprint arXiv:2403.20271*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2024c. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arXiv preprint arXiv:2406.09418*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. 2024. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36.

Wujian Peng, Lingchen Meng, Yitong Chen, Yiweng Xie, Yang Liu, Tao Gui, Hang Xu, Xipeng Qiu, Zuxuan Wu, and Yu-Gang Jiang. 2024. Inst-it: Boosting multimodal instance understanding via explicit visual prompt instruction tuning. *arXiv preprint arXiv:2412.03565*.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. 2017. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. 2024b. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. 2024c. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653.

Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. 2024. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v, 2023. *URL http://arxiv.org/abs/2310.11441*.

Linjie Yang, Yuchen Fan, and Ning Xu. 2019. Video instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5188–5197.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134.

Yuqian Yuan, Hang Zhang, Wentong Li, Zesen Cheng, Boqiang Zhang, Long Li, Xin Li, Deli Zhao, Wenqiao Zhang, Yueting Zhuang, et al. 2024. Videorefer suite: Advancing spatial-temporal object understanding with video llm. *arXiv preprint arXiv:2501.00599*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2024. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*.

A Elaboration on Dimensions

Table 4 presents detailed information on the five main tasks and twelve dimensions of V2P-Bench.

B Examples of V2P-Bench

We show some examples of V2P-Bench from Figure 6 to Figure 12. The ground-truth answer is highlighted in green.

Table 4: **Our proposed 5 main tasks and 12 dimensions with explanation.**

Perception	
Object Attribute	*This dimension evaluates the model's ability to perceive the visual and motion attributes of objects indicated by visual prompts, such as color, shape, position, and movement.*
Human Attribute	*This dimension evaluates the model's ability to recognize the actions and attributes of individuals indicated by visual prompts, such as their activities, clothing, and appearance.*
Object Direction	*This dimension examines the model's ability to perceive and interpret the motion trajectory of objects pointed by visual prompts, with a particular focus on movement direction.*
Feature Mapping	*This dimension examines the model's capability to extract distinctive features of objects indicated by visual prompts and consistently track them across the entire video.*
Reasoning	
Causal Relationship	*This dimension assesses the model's ability to perceive the causal relationships between actions and events, identifying the underlying intentions of actions and the causes of subsequent events. The visual prompt points to the action executor.*
Plot Understanding	*This dimension examines the model's ability to analyze narrative progression and logically infer subsequent developments based on the given plot. The visual prompt executes the protagonist of the plot.*
Counterfactual Inference	*This dimension evaluates the model's ability to reason under hypothetical scenarios that deviate from the actual video content, with visual prompts guiding the deviation, assessing its capacity to infer potential outcomes based on counterfactual assumptions.*
Temporal	
Forward Temporal	*This dimension assesses the model's ability to accurately locate the visual prompt and track subsequent events that follow the natural chronological order of the video.*
Reverse Temporal	*This dimension evaluates the model's capacity to comprehend the temporal structure of the video by identifying events that precede the visual prompt, demonstrating an understanding of temporal precedence.*
Action Sequence	*This dimension evaluates the model's ability to grasp the overall temporal flow of the video, particularly in understanding and reasoning about the temporal dynamics of multiple action sequences of individuals or objects, as indicated by visual prompts.*
Spatial	
Spatial Relationship	*This dimension assesses the model's ability to discern and comprehend the spatial relationships between instances highlighted by visual prompts within the video scene.*
Counting	
General Counting	*This dimension evaluates the model's ability to perceive and accurately count repeated actions or objects within the video, as indicated by visual prompts, testing its capacity for detailed content understanding and comprehensive scene analysis.*

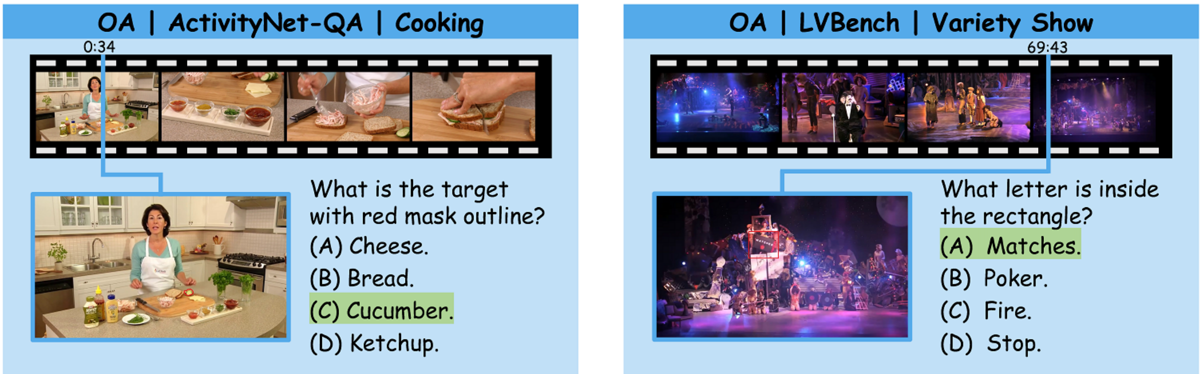

Figure 6: Examples of V2P-Bench in Object Attribute dimension.

Figure 7: Examples of V2P-Bench in Human Attribute dimension.

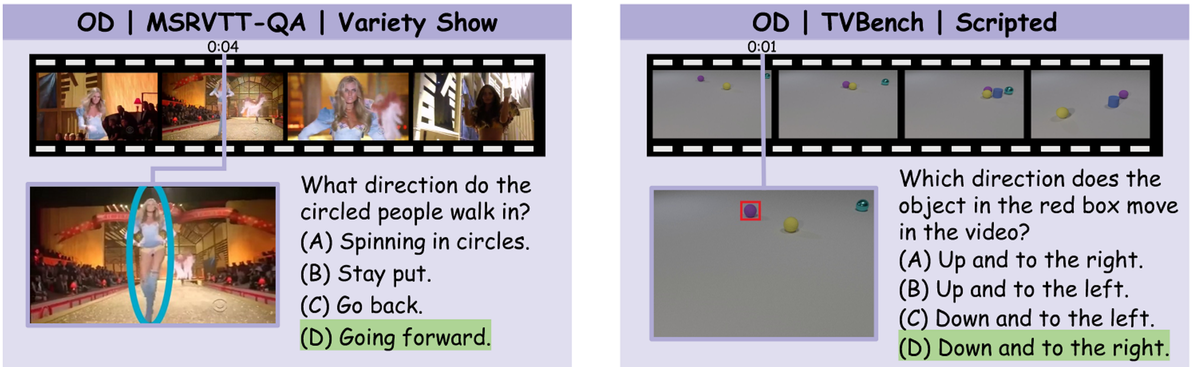

Figure 8: Examples of V2P-Bench in Object Direction dimension.

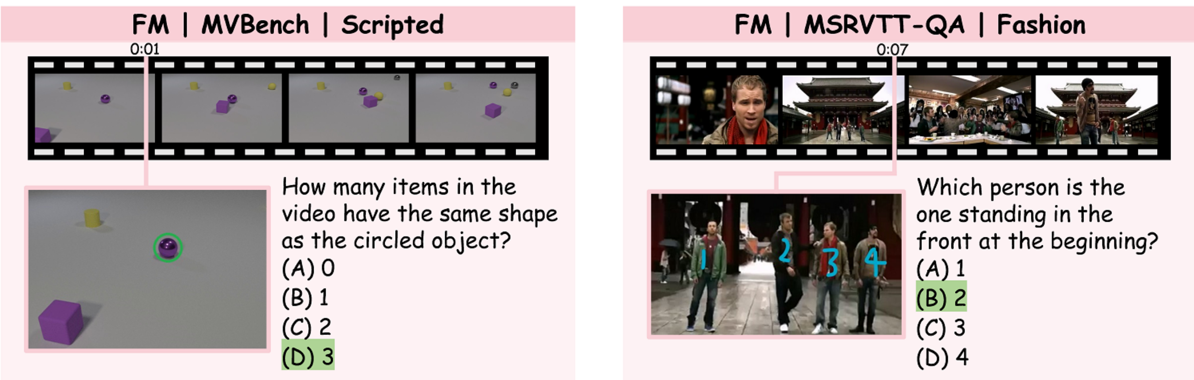

Figure 9: Examples of V2P-Bench in Feature Mapping dimension.

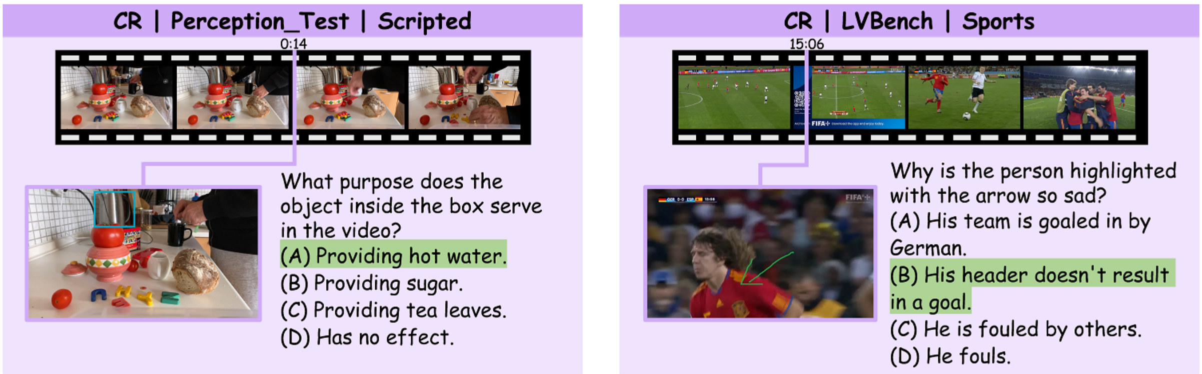

Figure 10: Examples of V2P-Bench in Causal Relationship dimension.

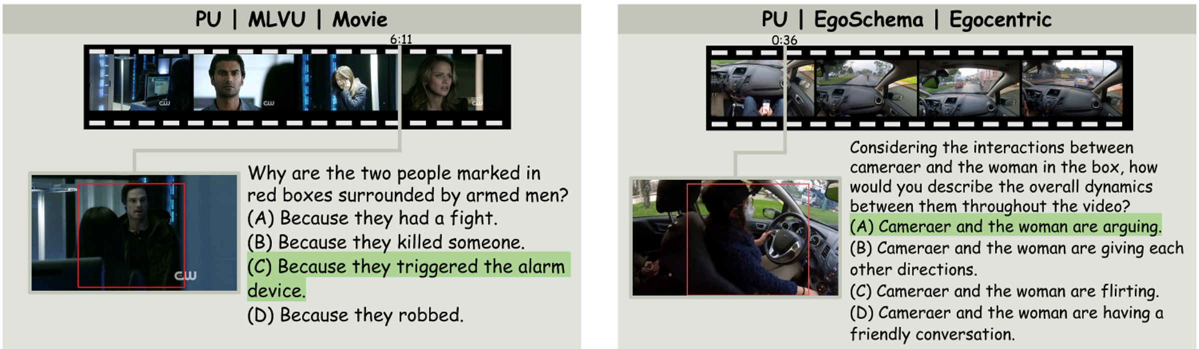

Figure 11: Examples of V2P-Bench in Plot Understanding dimension.

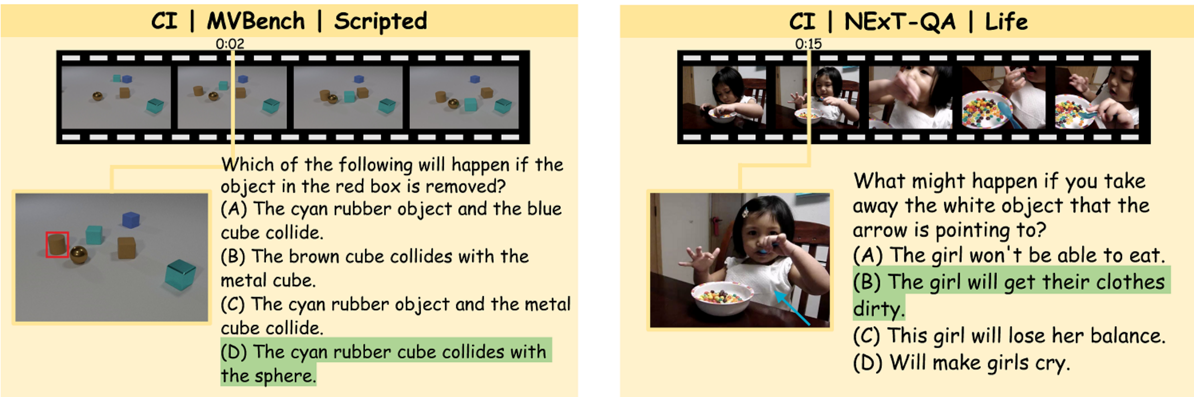

Figure 12: Examples of V2P-Bench in Counterfactual Inference dimension.

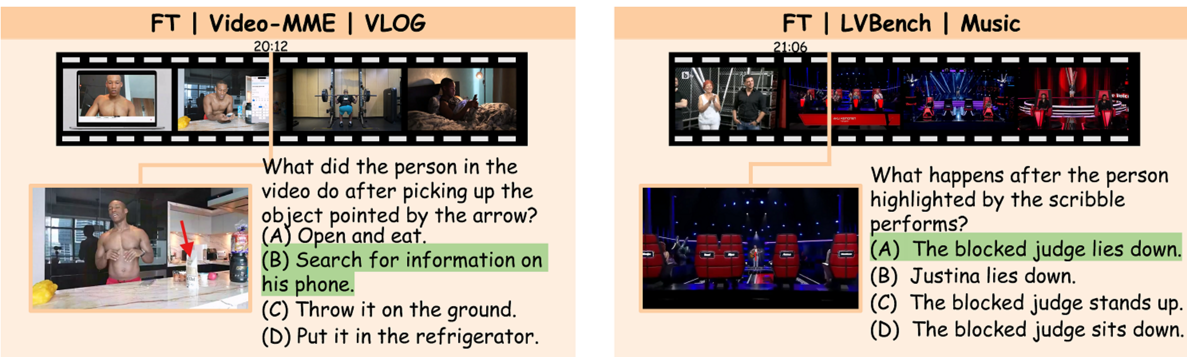

Figure 13: Examples of V2P-Bench in Forward Temporal dimension.

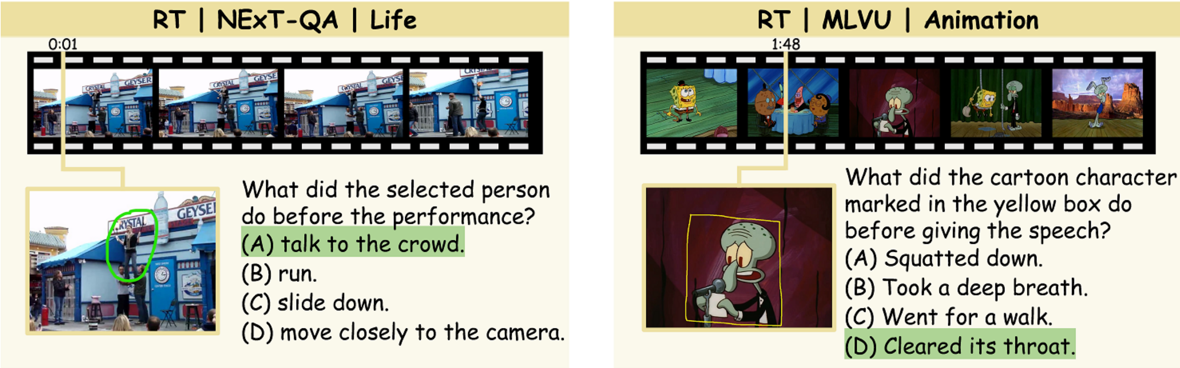

Figure 14: Examples of V2P-Bench in Reverse Temporal dimension.

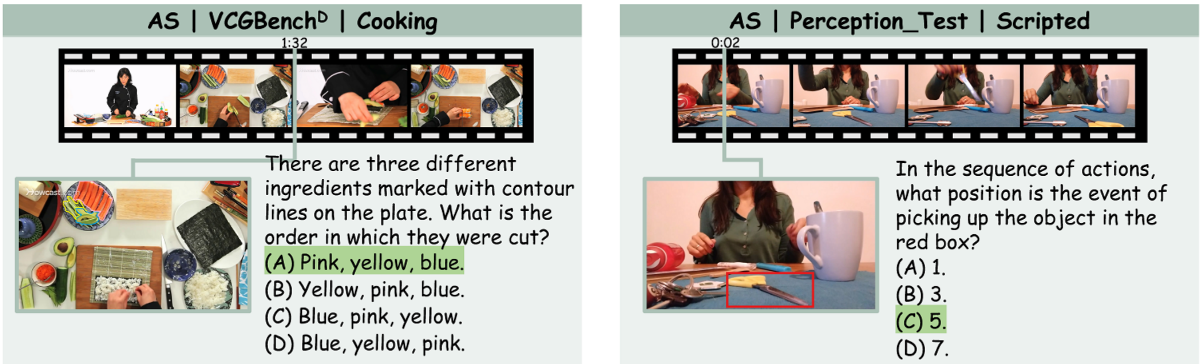

Figure 15: Examples of V2P-Bench in Action Sequence dimension.

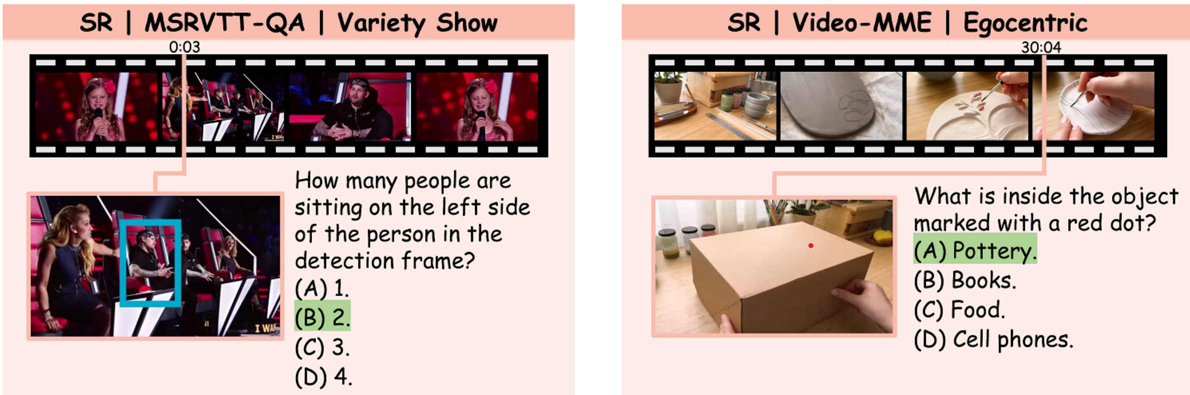

Figure 16: Examples of V2P-Bench in Spatial Relationship dimension.

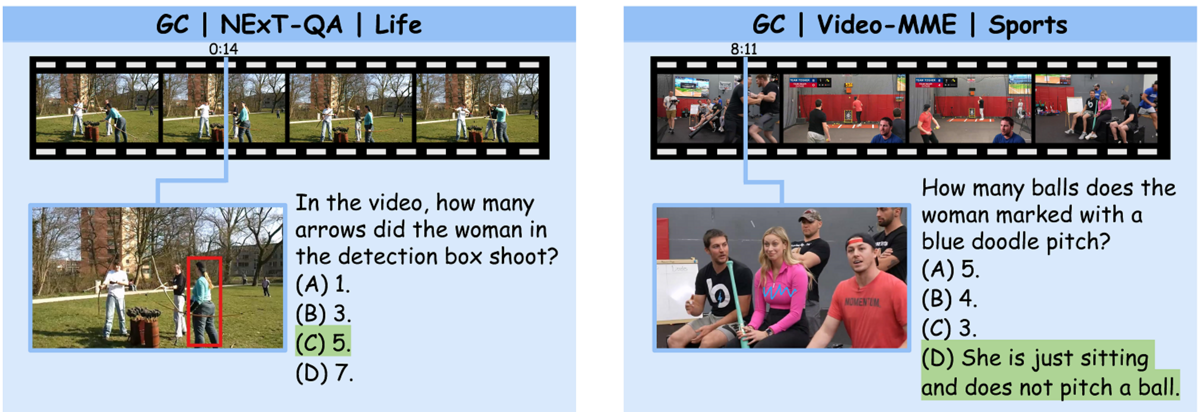

Figure 17: Examples of V2P-Bench in General Counting dimension.