A Detailed Results on FAVOR-Bench

In this section, we provide comprehensive analysis of MLLMs' performance on FAVOR-Bench, including detailed results across different tasks and perspectives, as well as in-depth examination of common failure patterns.

A1 Task-Specific Performance Analysis

We report the detailed results of MLLMs on MCQA and our proposed LLM-free evaluation framework in Table S1 and Table S2.

For MCQA, MLLMs' performances on each task are provided. Notably, the first-person perspective MCQA evaluation does not include Camera Motion and Non-Subject Motion. In such videos, the subject's turning is assessed instead, and the focus is on the interaction between the camera wearer and objects or other subjects, rather than background elements. For the LLM-free evaluation, we provide specific results in three aspects: camera motion, subject action list, and temporal action list. The camera motion metric assesses the model's ability to identify and describe viewpoint changes and camera movements accurately. The subject action list metric focuses on the actions of individual subjects, while the temporal action list evaluates the understanding and description of all subjects in terms of chronological order. For the subject and temporal action lists, we further consider the action matching degree and action sequence separately. Specifically, action matching measures the accuracy and detailedness of models' descriptions of the actions each subject performs. Action sequence checks whether these actions are reported in the correct temporal order.

Task Difficulty Hierarchy. Through analyzing the results in Table S1 and Table S2, we observe a clear difficulty hierarchy among the six subtasks:

- **Camera Motion (CM) and Action Sequence (AS)**: These represent the most challenging tasks in our benchmark. CM proves to be the most difficult, with consistently lowest scores across all models. For instance, even top-performing models like Gemini-1.5-Pro achieve only 41.49% on third-person videos, significantly below their overall performance. This difficulty stems from the subtle nature of viewpoint dynamics and focus shifts, which require indirect understanding through visual changes. AS also presents high difficulty as it requires both accurate identification of individual actions and correct temporal ordering.
- **Single/Multiple Action Details (SAD/MAD)**: These tasks show mixed difficulty levels depending on the specific motion characteristics. MAD generally proves more challenging than SAD as it requires tracking and comparing multiple temporal points, with performance gaps particularly evident in third-person videos.
- **Holistic Action Classification (HAC) and Non-Subject Motion (NSM)**: These represent the relatively easier tasks in our benchmark. HAC benefits from similarity to existing coarse-grained action recognition tasks, with top models achieving over 65% accuracy on ego-centric videos. NSM, focusing on obvious environmental movements, also shows higher success rates compared to other subtasks.

Perspective-based Differences. Comparing ego-centric and third-person perspectives reveals interesting patterns. Most models achieve higher MCQA scores on ego-centric videos (e.g., Qwen2.5-VL-72B: 58.48% vs. 46.67%), suggesting that the complex camera motions and multi-subject interactions in third-person videos pose greater challenges. However, this trend reverses in open-ended tasks, where models generally perform better on third-person videos, indicating difficulties in describing ego-centric motion dynamics despite understanding them.

A2 Analysis of MLLM Failures

By examining the detailed performance metrics and common error patterns, we identify several critical failure modes of current MLLMs on fine-grained motion understanding:

Motion Recognition vs. Temporal Ordering. A striking pattern emerges from the LLM-free evaluation results: while most models achieve high scores (>85%) in both subject-specific and temporal action sequences, they struggle significantly with action matching (typically below 40%). This discrepancy reveals that MLLMs can reasonably determine the temporal order of actions when

correctly identified, but fail to comprehensively recognize all fine-grained motions occurring in videos. For example, Gemini-1.5-Pro achieves 94.37% in subject sequence but only 41.87% in subject match for third-person videos.

Camera Motion Understanding Deficits. The consistently low performance on Camera Motion tasks across all models highlights a fundamental limitation in current MLLMs. Models with strong comprehensive capabilities like Gemini-1.5-Pro, LLaVA-Video-72B-Qwen2, and Qwen2.5-VL-72B show CM scores significantly below their overall performance. This suggests that understanding viewpoint dynamics and focus shifts remains particularly challenging, possibly due to insufficient training data with explicit camera motion annotations or architectural limitations in capturing such subtle visual cues.

Scale and Architecture Effects. Clear scaling effects are observed within model families. For instance, Qwen2.5-VL-72B outperforms its 7B variant by approximately 8% in overall MCQA scores. However, scaling alone does not guarantee better fine-grained understanding. Some smaller models like VideoChat-Flash-Qwen2-7B (43.52%) outperform larger alternatives like InternVL2.5-78B (38.42%) in overall MCQA, suggesting that specialized architectures or training strategies may be more crucial than model size for motion understanding.

These analyses reveal that while current MLLMs have made progress in video understanding, significant gaps remain in fine-grained motion comprehension. The difficulty hierarchy and failure patterns identified here provide valuable insights for future model development.

B Discussions about the LLM-Free Framework

B1 Correlation with GPT-assisted Evaluation

In this section, we calculate the Spearman correlation coefficient between LLM-free and GPT-assisted metrics to demonstrate the reliability of our proposed LLM-free framework for open-ended fine-grained motion description. As visualized in Figure S1, the LLM-free scores and GPT-assisted scores show a strong and statistically significant correlation, achieving a Pearson correlation of 0.86 ($p < 0.001$) and a Spearman correlation of 0.77 ($p < 0.001$). This suggests that our proposed LLM-free framework is reasonable and reliable for open-ended evaluation.

B2 Examples of Structured Motion Elements Extraction

We provide the results of structured motion elements extraction for several responses. As can be seen from Figure S2, the extraction tool we developed can capture most subjects and their motions. For cases where subjects appear alternately temporally (like the upper part of the figure), actions can also be correctly assigned to their corresponding subjects, which further validates the effectiveness.

C More Details of FAVOR-Bench

C1 More Data Statistics

In this subsection, we provide more data statistics of FAVOR-Bench. In Figure S3, we provide the distribution of video durations and the number of questions for each video, as well as the index distribution of correct answers. Figure S4 exhibits the statistics of motion words with the highest frequency in FAVOR-Bench.

C2 Process of Data Curation

The detailed data curation process for different video sources is as follows. The purpose of this process is to get videos with high quality and dynamic motions.

- **Daily-life record**: For this subset, we sample videos from Charades [37]. This dataset is publicly available and contains various types of human actions and interactions in daily life. Video lengths, number of subjects and density of motions, and interactions of this dataset are relatively moderate. 868 videos are sampled from the subset with the highest quality score (quality scores are offered by Charades).

Table S1: Ego-centric performances of 21 MLLMs on FAVOR-Bench, including close-ended multiple choice and LLM-free evaluation. Random selecting and human performance are also compared. The highest and second-highest results among all MLLMs are indicated in bold and underlined. Due to the API response limitations, the video input of proprietary MLLMs is restricted to 16 frames if the video is longer than 16 seconds (demoted as "1 fps*".) Tarsier2-Recap-7B is a model specially designed for captioning and it fails to fulfill the close-ended evaluation.

Methods	**MCQA**				**LLM-Free**				
	AS	HAC	SAD	MAD	Camera	Subject Match	Subject Seq.	Temporal Match	Temporal Seq.
Full mark	100	100	100	100	100	100	100	100	100
Random	20	20	20	20	–	–	–	–	–
Human	90.90	89.22	92.65	91.75	85.54	67.32	95.76	71.37	97.17
Proprietary MLLMs									
Gemini-1.5-Pro [40]	50.76	**65.87**	55.12	59.79	57.72	28.63	84.86	30.14	88.87
GPT-4o [19]	45.90	53.29	50.24	48.45	52.78	28.36	84.75	32.31	95.06
Claude-3.7-Sonnet [1]	47.72	46.11	44.39	47.94	48.28	15.91	80.42	18.05	86.09
Open-source MLLMs									
Video-LLaVA-7B [27]	24.32	22.16	23.41	29.90	47.20	19.56	81.75	18.49	90.08
LLaVA-NeXT-Video-7B [58]	20.06	21.56	32.20	27.32	43.27	10.19	75.07	8.93	75.04
LLaVA-NeXT-Video-34B [58]	31.31	31.14	34.15	18.04	48.26	20.64	84.46	20.22	92.60
Tarsier-7B [43]	15.81	22.16	21.46	19.59	77.23	19.90	80.60	18.36	86.61
Tarsier-34B [43]	30.40	37.72	26.83	32.99	80.98	25.76	84.91	24.05	91.62
Aria [23]	29.18	45.51	22.93	36.60	42.41	19.18	81.53	19.59	90.21
InternVL2.5-2B [6]	18.54	29.94	23.90	21.65	72.67	20.08	79.91	18.30	86.53
InternVL2.5-8B [6]	32.22	46.11	40.98	38.66	48.40	22.37	82.52	22.73	88.97
InternVL2.5-78B [6]	39.21	50.30	43.90	55.67	68.58	20.27	82.37	20.61	90.83
Tarsier2-Recap-7B [55]	–	–	–	–	80.47	**36.66**	**85.92**	**39.18**	**95.95**
LLaVA-Video-7B-Qwen2 [59]	37.99	50.90	46.83	51.55	43.37	22.17	82.33	22.88	90.59
LLaVA-Video-72B-Qwen2 [59]	51.98	52.69	50.73	**64.43**	42.58	20.64	83.38	21.75	88.84
VideoChat-Flash-Qwen2-7B [25]	40.43	50.90	46.83	51.55	**89.76**	12.71	75.96	9.33	80.35
VideoLLaMA3-2B [56]	29.18	45.51	37.07	41.75	51.36	16.38	79.49	15.42	79.22
VideoLLaMA3-7B [56]	39.21	53.29	48.78	49.48	56.07	22.73	82.85	22.83	91.48
Qwen2.5-VL-3B [52]	38.91	51.50	43.90	46.39	64.32	26.56	82.85	25.76	91.09
Qwen2.5-VL-7B [52]	38.30	50.30	53.66	47.42	50.51	27.16	84.17	28.06	92.55
Qwen2.5-VL-72B [52]	**57.75**	60.48	**56.10**	60.31	41.94	30.86	84.56	34.80	94.29
Qwen2.5-VL-7B+FAVOR-Train	43.16	52.10	52.68	48.45	89.74	43.99	82.20	52.07	92.76

- **TV-series & animation**: For more diverse distributions of characters, motions and scenarios, we collect a corpus of data from TV series and animations to form new subsets. Raw videos with short side resolution greater than 480 are first cropped into segments of a few to tens of seconds. In order to avoid scene transitions in clips, an adaptive detector algorithm from the scene detector library is adopted to generate a list of scenes before generating the clip start and end times. Then, the optical flow calculation with OpenCV is utilized to filter out segments with little motion. Additionally, we manually curate the segments, selecting approximately one out of every twenty clips from a large pool of candidates. Our selection criteria focus on avoiding clips with overly simplistic motions (for example, a person talking throughout the whole video clip) or an excessive number of dynamic subjects in the frame to ensure high-quality fine-grained annotations. Specifically, we exclude segments that only featured simple motions, such as talking heads or basic walking movements. In addition, the video selection process considers the balanced distribution of single-subject and multi-subject scenarios. Finally, 574 clips from TV-series and 138 clips from animations are selected.

- **Egocentric**: In order to further expand the distribution of motions and present new challenges, a certain number of egocentric videos have been included in our benchmark. These videos capture unique interaction patterns and contextual information, occupying a certain proportion in real-world scenarios. We select EgoTaskQA [20] as the data source and randomly select 196 egocentric videos featuring kitchen and daily activities.

To get high quality structured manual annotation, we build a comprehensive annotation guideline as shown in Figure S5.

Table S2: Third-person performances of 21 MLLMs on FAVOR-Bench, including close-ended multiple choice and LLM-free evaluation. Random selecting and human performance are also compared. The highest and second-highest results among all MLLMs are indicated in bold and underlined. Due to the API response limitations, the video input of proprietary MLLMs is restricted to 16 frames if the video is longer than 16 seconds (demoted as "1 fps*".) Tarsier2-Recap-7B is a model specially designed for captioning and it fails to fulfill the close-ended evaluation.

Methods	**MCQA**						**LLM-Free**				
	AS	HAC	SAD	MAD	CM	NSM	Camera	Subject Match	Subject Seq.	Temporal Match	Temporal Seq.
Full mark	100	100	100	100	100	100	100	100	100	100	100
Random	20	20	20	20	20	20	–	–	–	–	–
Human	91.46	92.43	91.12	89.12	88.19	96.88	70.07	68.42	98.10	70.28	98.66
Proprietary MLLMs											
Gemini-1.5-Pro [40]	**48.79**	**52.18**	**47.91**	**53.81**	**41.49**	**55.56**	49.39	41.87	94.37	43.55	94.23
GPT-4o [19]	38.60	44.03	39.88	43.82	35.53	47.62	51.71	38.15	**94.71**	40.40	94.62
Claude-3.7-Sonnet [1]	44.32	42.50	41.32	47.08	38.70	44.44	52.03	27.95	91.51	31.43	91.72
Open-source MLLMs											
Video-LLaVA-7B [27]	23.18	19.36	22.65	27.00	25.86	22.22	50.04	25.84	90.86	26.30	88.94
LLaVA-NeXT-Video-7B [58]	20.49	21.91	23.47	24.13	22.98	14.29	46.60	14.65	85.70	14.14	65.59
LLaVA-NeXT-Video-34B [58]	31.11	30.42	30.61	22.65	29.30	46.03	51.33	24.27	90.14	23.99	85.30
Tarsier-7B [43]	10.27	15.50	10.98	12.36	20.84	28.57	60.62	30.74	91.22	29.82	88.54
Tarsier-34B [43]	26.52	30.57	19.15	26.90	29.49	34.92	61.07	31.29	91.83	30.26	88.83
Aria [23]	29.59	37.34	17.57	26.41	27.81	50.79	45.08	28.64	92.07	30.00	90.23
InternVL2.5-2B [6]	18.20	28.02	23.68	28.29	19.16	22.22	48.94	28.41	92.25	29.81	89.85
InternVL2.5-8B [6]	31.50	37.77	37.68	37.09	26.14	34.92	53.38	29.57	92.30	31.43	91.02
InternVL2.5-78B [6]	37.95	39.45	38.37	41.05	29.40	38.10	55.56	29.02	92.11	31.18	91.14
Tarsier2-Recap-7B [55]	–	–	–	–	–	–	61.28	**44.26**	94.63	**47.46**	**95.37**
LLaVA-Video-7B-Qwen2 [59]	35.57	40.10	40.43	42.43	29.40	46.03	49.00	31.63	92.86	32.83	91.39
LLaVA-Video-72B-Qwen2 [59]	47.27	46.72	44.34	48.76	33.02	52.38	48.40	32.26	93.02	34.84	92.75
VideoChat-Flash-Qwen2-7B [25]	41.51	48.11	42.28	49.95	34.98	50.79	**63.07**	25.39	88.80	24.49	77.43
VideoLLaMA3-2B [56]	28.60	35.52	34.52	36.40	28.56	39.68	50.84	25.53	90.14	25.99	77.91
VideoLLaMA3-7B [56]	40.08	43.01	41.46	47.18	31.26	41.27	52.83	34.32	93.62	35.76	92.65
Qwen2.5-VL-3B [52]	37.87	36.61	35.62	37.88	29.77	31.75	55.71	32.10	92.97	33.93	92.15
Qwen2.5-VL-7B [52]	38.91	42.43	41.66	42.43	33.49	38.10	55.89	35.49	93.04	35.72	90.83
Qwen2.5-VL-72B [52]	48.74	45.34	47.01	49.75	40.28	50.79	55.41	37.48	94.40	40.20	94.56
Qwen2.5-VL-7B+FAVOR-Train	40.47	44.25	39.19	40.85	38.98	38.10	63.69	42.51	92.72	43.55	92.27

C3 Prompt for Automatic QA Generation

This subsection provides the prompt templates for automatically generating QA pairs. These templates are carefully designed to ensure high-quality questions that thoroughly evaluate MLLMs' fine-grained video motion understanding capabilities. For each of the six close-ended tasks in FAVOR-Bench, we use distinct prompts to guide DeepSeek-R1 [13] in generating diverse and challenging multiple-choice questions. Each template is adjusted according to the task type and emphasizes different aspects of fine-grained motion understanding. Tables S3-S7 present the detailed prompt templates for each task type.

C4 Prompt for GPT-Assisted Evaluation

For the GPT-assisted evaluation, we design a comprehensive prompt to instruct the powerful GPT-4o to directly compare and evaluate model responses from two critical dimensions of fine-grained video motion understanding: correctness and detailedness. The correctness dimension assesses whether the model accurately describes the motions, activities, interactions, and camera movements that occur in the video. The detailedness dimension evaluates how comprehensively the model captures the temporal dynamics, including the execution manner of actions, camera movements, and interactions. To improve the robustness of the evaluation, we develop detailed criteria for different ratings respectively. Tables S9 and S10 present the complete prompt template used in this paper.

C5 Limitations

While FAVOR-Bench advances fine-grained video motion understanding evaluation, several limitations warrant acknowledgment. First, despite our efforts to include diverse video content spanning ego-centric and third-person perspectives, it might still not address all dimensions of motion understanding necessary for real-world applications. Second, although our benchmark construction process incorporates multiple filtering stages and manual verification, some inaccuracies and incon-

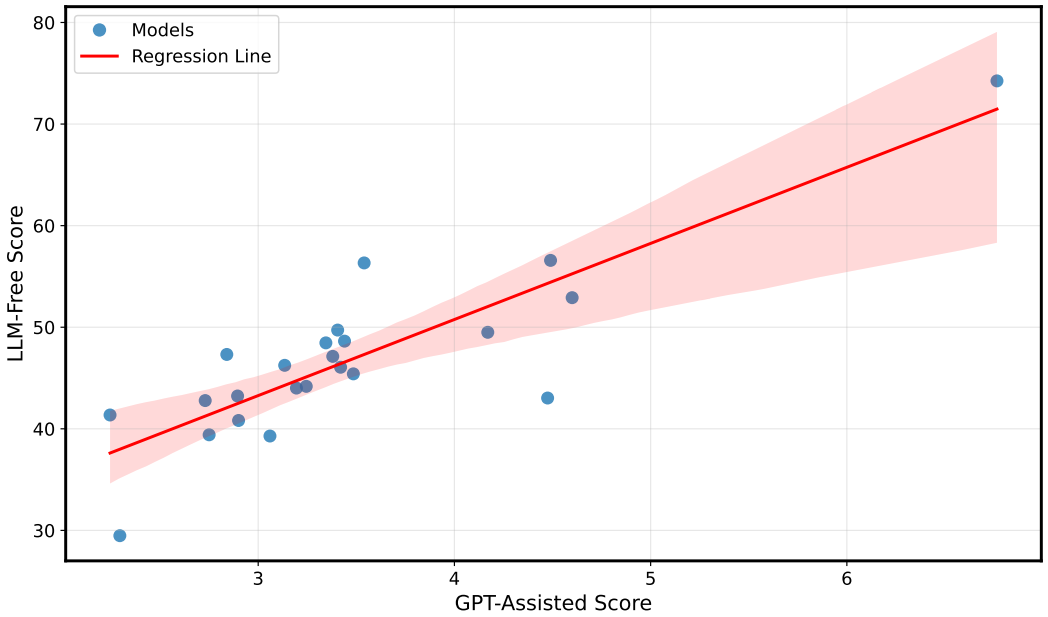


Figure S1: Visualization of the correlation between GPT-assisted metrics and LLM-free metrics in the open-ended evaluation of FAVOR-Bench.

sistencies may persist due to the inherent subjectivity in fine-grained motion interpretation and the complexity of temporal dynamics. Third, while our proposed LLM-free evaluation framework offers improved accessibility and reproducibility, it may not fully capture the nuanced aspects of motion understanding that human evaluators might recognize. These limitations highlight opportunities for future work in creating more diverse, culturally inclusive benchmarks with even more refined evaluation methodologies.

D More Details of FAVOR-Train

For third-person perspective videos, all videos are sampled from Koala-36M [46]. Before sampling, an automatic filtering is performed using the clarity score, aesthetic score, motion score, and VTSS provided by Koala-36M. We use three different sampling strategies to create three subsets from the filtered 7M videos, which collectively make up the third-person component of our training set:

- **Random sampling**: Simple random sampling is performed to get the first one-third of videos.
- **Videos with long-tail actions**: Before sampling of this step, we use Qwen2-72B to extract motion phrases in the caption from Koala-36M of each video. We then traverse the filtered subset video list and maintain a set of motions (each from a video) and keep their distance in the semantic embedding space as far as possible. Thus, a set of long-tail motions is obtained.
- **Videos with typical motions**: We select nearly 1600 motions from the motion set of previous datasets for caption recognition, such as Kinetics [21], ActivityNet [3], and MIT [31]. Videos whose captions contain these motions are accumulated to form another one-third of the videos.

For first-person perspective videos, videos from EgoTaskQA and Charades-Ego can be directly sampled as training data. However, videos from Egoexo4D and EgoexoLearn are much longer. We adopt a cropping strategy based on their temporal annotations and ensure that all cropped video clips are shorter than 40 seconds.

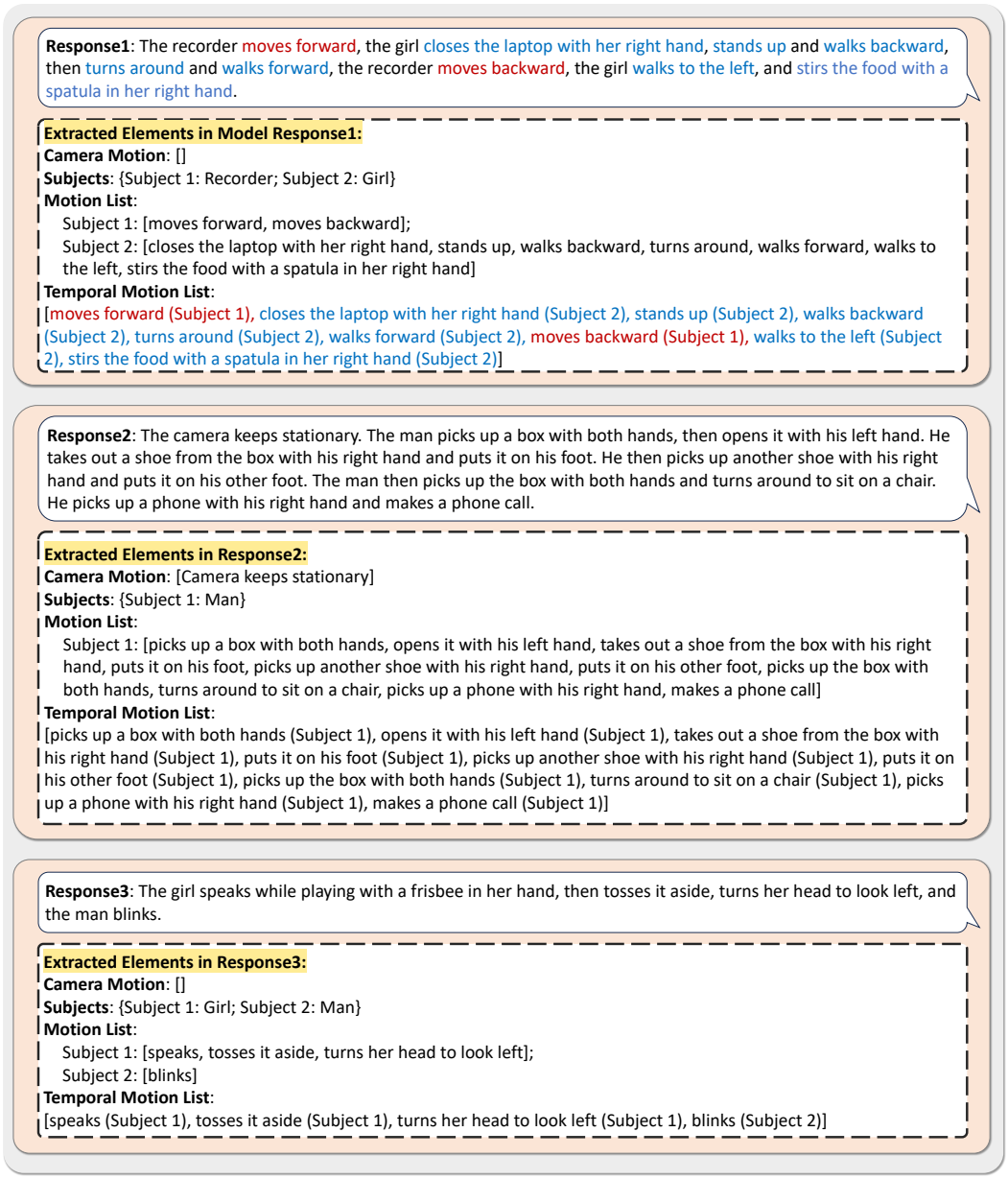


Figure S2: Results of structured motion elements extraction for several responses. For responses involving multiple subjects, we differentiate the motions of different subjects with distinct colors in the Temporal Motion List.

E More Samples of FAVOR-Bench

For better demonstration, we show more samples of videos and their corresponding close-ended questions, options and correct answers for all six tasks in Figures S6 to S10.

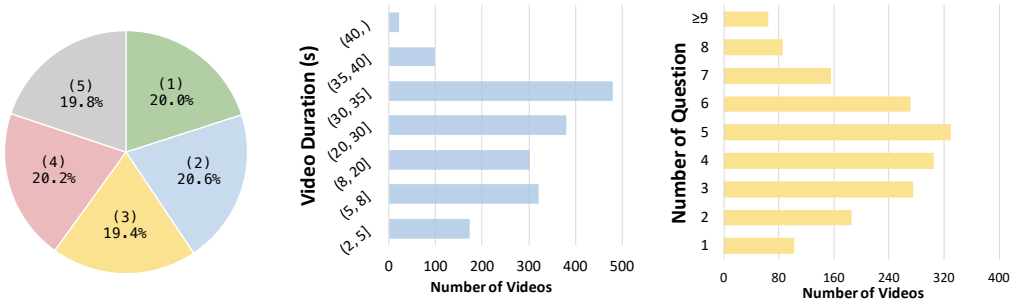


Figure S3: More data statistics of FAVOR-Bench. **Left:** Index distribution of correct answers for the close-ended tasks. For example, "(1)" indicates that the correct option is ranked first. **Middle:** Video duration distribution of FAVOR-Bench. **Right:** Question number distribution for videos of FAVOR-Bench.

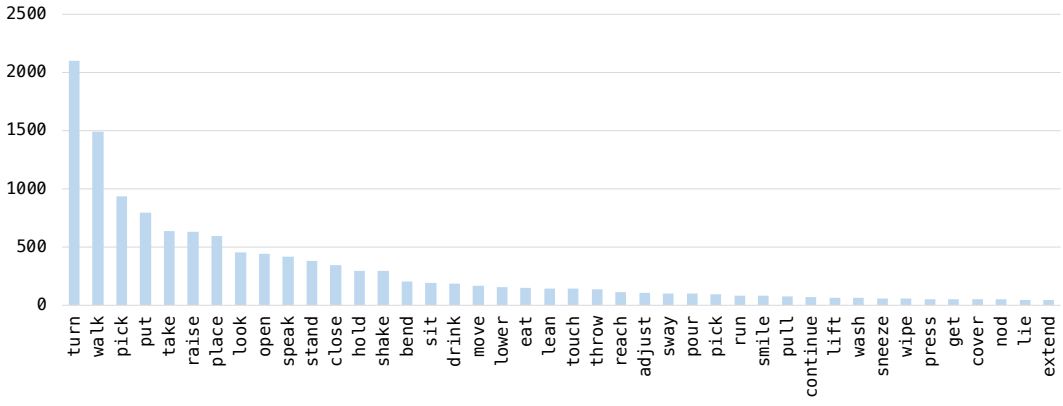


Figure S4: Statistics of motion words with the highest frequency in FAVOR-Bench.

Table S3: Prompt template for Action Sequence (AS) task.

Prompt Template: Generating QA Pairs for Action Sequence (AS) Task

You are a professional question designer focusing on temporal dynamics in videos, including camera movements, motions, activities, and interactions, rather than static content. You will receive detailed annotations about the temporal details of the entire video, with duration markers in parentheses after “camera_motion" and “motion_list". Based on these annotations, design 3 multiple-choice questions around the “Action Sequence" theme to test models’ fine-grained video motion understanding, particularly:

- Understanding and analysis of temporal logic, requiring precise identification of action sequences and comprehension of dynamic continuity.

Additionally, follow these question design guidelines:

1. Multiple-choice questions should include 5 options (without A,B,C option labels), ensuring exactly one correct answer. If descriptions are ambiguous (e.g., timeline or content), prioritize answer uniqueness.
2. Focus on representative and significant events or motions, avoiding excessive details or traps.
3. Distractor generation standards:
   - Avoid excessive similarity between options.
   - Exclude options requiring subjective inference (emotions, intentions, etc.).
   - Distractors can be outside the annotated motion list. Design relevant distractors, but ensure they don’t affect the uniqueness of the correct answer.
4. Subject identification standards:
   - When involving multiple subjects, use subject attributes (e.g., “man wearing black clothes with gold patterns") for reference.
   - Avoid abstract identifiers like “subject 1" or “person A".
   - Extract key visual features (color, position, clothing, etc.) from the original annotations.
5. Avoid questions about specific moments, such as “When...?" or “At which second...?"
6. Different questions should cover more video content, avoiding repetitive questioning.
7. Avoid questions that can be answered correctly just from the question wording, or from a single frame, without watching the entire video.
8. Include 1 multiple-choice question asking about a subject’s complete behavioral sequence in the video, creating distractors by rearranging the sequence. Connect different behaviors/actions with arrows.

Table S4: Prompt template for Camera Motion (CM) task.

Prompt Template: Generating QA Pairs for Camera Motion (CM) Task

You are a professional question designer focusing on temporal dynamics in videos, including camera movements, motions, activities, and interactions, rather than static content. You will receive detailed annotations about the temporal details of the entire video, with duration markers in parentheses after “camera_motion" and “motion_list". Based on these annotations, design 3 multiple-choice questions around the “Camera Motion" theme to test models’ fine-grained video motion understanding, particularly:

- Understanding camera movement direction and focus changes in the video.

Additionally, follow these question design guidelines:

1. If a video’s “camera_motion" has only one element, such as “camera_motion": “static", or “camera_motion": “camera shaking (0-22)", skip this video and don’t generate any content.
2. Multiple-choice questions should include 5 options (without A,B,C option labels), ensuring exactly one correct answer. If descriptions are ambiguous, prioritize answer uniqueness.
3. Focus only on camera movements and focus changes. Concentrate on representative and significant events or motions, avoiding excessive details or traps.
4. Distractor generation standards:
   - Avoid excessive similarity between options.
   - Exclude options requiring subjective inference (emotions, intentions, etc.).
   - Distractors can be outside the annotated motion list. Design relevant distractors, but ensure they don’t affect the uniqueness of the correct answer.
5. Subject identification standards:
   - When involving multiple subjects, use subject attributes (e.g., “man wearing black clothes with gold patterns") for reference.
   - Avoid abstract identifiers like “subject 1" or “person A".
   - Extract key visual features (color, position, clothing, etc.) from the original annotations.
6. Avoid questions about specific moments, such as “When...?" or “At which second...?"
7. Different questions should cover more video content, avoiding repetitive questioning.
8. Questions and answers should not include any specific time information, such as “during the camera shake phase (0-3 seconds)"
9. If “camera_motion" doesn’t mention focus, designed questions should not include focus.
10. Avoid questions that can be answered correctly just from the question wording, or from a single frame, without watching the entire video.

Table S5: Prompt template for Holistic Action Classification (HAC) task.

Prompt Template: Generating QA Pairs for Holistic Action Classification (HAC) Task

You are a professional question designer focusing on temporal dynamics in videos, including camera movements, motions, activities, and interactions, rather than static content. You will receive detailed annotations about the temporal details of the entire video, with duration markers in parentheses after “camera_motion" and “motion_list". Based on these annotations, design 3 multiple-choice questions around the “Holistic Action Classification" theme to test models’ fine-grained video motion understanding, particularly:

- Emphasizing overall summarization ability, requiring distillation of core behaviors from the entire video.

Additionally, follow these question design guidelines:

1. Multiple-choice questions should include 5 options (without A,B,C option labels), ensuring exactly one correct answer. If descriptions are ambiguous, prioritize answer uniqueness.
2. Focus on the main behaviors and motions throughout the entire video, which should cover most of the video duration or occupy more than half of the caption length. Don’t ask about motions occurring only in local time segments of the video, avoiding expressions like “in the first half of the video" or “at the end of the video."
3. Distractor generation standards:
   - Avoid excessive similarity between options (e.g., “chin moving up and down" vs. “slight opening and closing").
   - Exclude options requiring subjective inference (emotions, intentions, etc.).
   - Distractors can be outside the annotated motion list. Design relevant distractors, but ensure they don’t affect the uniqueness of the correct answer.
4. Subject identification standards:
   - When involving multiple subjects, use subject attributes (e.g., “man wearing black clothes with gold patterns") for reference.
   - Avoid abstract identifiers like “subject 1" or “person A".
   - Extract key visual features (color, position, clothing, etc.) from the original annotations.
5. Avoid questions about specific moments, such as “When...?" or “At which second...?"
6. Different questions should cover more video content, avoiding repetitive questioning.
7. Avoid questions that can be answered correctly just from the question wording, or from a single frame, without watching the entire video.

Table S6: Prompt template for Multiple Action Detail (MAD) task.

Prompt Template: Generating QA Pairs for Multiple Action Detail (MAD) Task

You are a professional question designer focusing on temporal dynamics in videos, including camera movements, motions, activities, and interactions, rather than static content. You will receive detailed annotations about the temporal details of the entire video, with duration markers in parentheses after “camera_motion" and “motion_list". Based on these annotations, design 3 multiple-choice questions around the “Multiple Action Detail" theme to test models’ fine-grained video motion understanding, particularly:

- Multi-moment information comparison and analysis ability, requiring understanding of object state changes at different times, subject interactions with multiple objects at different times, or subject multiple interactions with the same object, emphasizing contrast and associative reasoning.

Additionally, follow these question design guidelines:

1. Multiple-choice questions should include 5 options (without A,B,C option labels), ensuring exactly one correct answer. If descriptions are ambiguous, prioritize answer uniqueness.
2. Focus on representative and significant events or motions, avoiding excessive details or traps.
3. Distractor generation standards:
   - Avoid excessive similarity between options (e.g., “chin moving up and down" vs. “slight opening and closing").
   - Exclude options requiring subjective inference (emotions, intentions, etc.).
   - Distractors can be outside the annotated motion list. Design relevant distractors, but ensure they don’t affect the uniqueness of the correct answer.
4. Subject identification standards:
   - When involving multiple subjects, use subject attributes (e.g., “man wearing black clothes with gold patterns") for reference.
   - Avoid abstract identifiers like “subject 1" or “person A".
   - Extract key visual features (color, position, clothing, etc.) from the original annotations.
5. Avoid questions about specific moments, such as “When...?" or “At which second...?"
6. Different questions should cover more video content, avoiding repetitive questioning.
7. If a subject interacted with multiple objects, include 1 multiple-choice question asking about which items a subject interacted with at different times; otherwise, this is not needed.
8. Avoid questions that can be answered correctly just from the question wording, or from a single frame, without watching the entire video.

Table S7: Prompt template for Single Action Detail (SAD) task.

Prompt Template: Generating QA Pairs for Single Action Detail (SAD) Task

You are a professional question designer focusing on temporal dynamics in videos, including camera movements, motions, activities, and interactions, rather than static content. You will receive detailed annotations about the temporal details of the entire video, with duration markers in parentheses after “camera_motion" and “motion_list". Based on these annotations, design 3 multiple-choice questions around the “Single Action Detail" theme to test models’ fine-grained video motion understanding, particularly:

- Testing detail capturing ability, understanding subject interaction with a specific object, or subject/object state at a specific moment.

Additionally, follow these question design guidelines:

1. Multiple-choice questions should include 5 options (without A,B,C option labels), ensuring exactly one correct answer. If descriptions are ambiguous, prioritize answer uniqueness.
2. Focus on representative and significant events or motions, avoiding excessive details or traps.
3. Distractor generation standards:
   - Avoid excessive similarity between options (e.g., “chin moving up and down" vs. “slight opening and closing").
   - Exclude options requiring subjective inference (emotions, intentions, etc.).
   - Distractors can be outside the annotated motion list. Design relevant distractors, but ensure they don’t affect the uniqueness of the correct answer.
4. Subject identification standards:
   - When involving multiple subjects, use subject attributes (e.g., “man wearing black clothes with gold patterns") for reference.
   - Avoid abstract identifiers like “subject 1" or “person A".
   - Extract key visual features (color, position, clothing, etc.) from the original annotations.
5. Avoid questions about specific moments, such as “When...?" or “At which second...?"
6. Different questions should cover more video content, avoiding repetitive questioning.
7. Avoid questions that can be answered correctly just from the question wording, or from a single frame, without watching the entire video.

Table S8: Prompt template for Non-subject Motion (NSM) task.

Prompt Template: Generating QA Pairs for Non-subject Motion (NSM) Task

You are a professional question designer focusing on temporal dynamics in videos, including camera movements, motions, activities, and interactions, rather than static content. You will receive detailed annotations about the temporal details of the entire video, with duration markers in parentheses after “camera_motion" and “motion_list". Based on these annotations, design 3 multiple-choice questions around the “Non-subject Motion" theme to test models’ fine-grained video motion understanding, particularly:

- Testing environmental awareness ability, requiring attention to secondary elements (such as background objects, background characters) movement and behavior, as supplementary information for fine-grained video motion understanding.

Additionally, follow these question design guidelines:

1. Multiple-choice questions should include 5 options (without A,B,C option labels), ensuring exactly one correct answer. If descriptions are ambiguous, prioritize answer uniqueness.
2. Focus on representative and significant events or motions, avoiding excessive details or traps.
3. Distractor generation standards:
   - Avoid excessive similarity between options (e.g., “chin moving up and down" vs. “slight opening and closing").
   - Exclude options requiring subjective inference (emotions, intentions, etc.).
   - Distractors can be outside the annotated motion list. Design relevant distractors, but ensure they don’t affect the uniqueness of the correct answer.
4. Subject identification standards:
   - When involving multiple subjects, use subject attributes (e.g., “man wearing black clothes with gold patterns") for reference.
   - Avoid abstract identifiers like “subject 1" or “person A".
   - Extract key visual features (color, position, clothing, etc.) from the original annotations.
5. Avoid questions about specific moments, such as “When...?" or “At which second...?"
6. Different questions should cover more video content, avoiding repetitive questioning.
7. Avoid questions that can be answered correctly just from the question wording, or from a single frame, without watching the entire video.

Table S9: Prompt template for GPT-assisted evaluation of FAVOR-Bench (Part 1).

Prompt Template for GPT-Assisted Evaluation (1/2)

Please act as a professional video motion analysis expert to evaluate models' fine-grained motion understanding capabilities in videos. You will compare the model-generated description (Response) with human-annotated standard description (Caption), and rate the model's performance on two dimensions, "Correctness" and "Detailedness", each on a scale from 1 to 10. Remember that the model received this instruction: "Please analyze and describe the temporal dynamics in this video, focusing on camera movements, motions, activities, and interactions, rather than static content."

Evaluation Dimension 1: Correctness (1-10 points)
Evaluate whether the model's description of motions, activities, interactions, and camera movements that actually appear in the video is accurate.

Correctness Rating Criteria:

9-10 points (Extremely High Correctness)
- Completely correct description of all core motions, activities, and interactions in the video;
- Correctly identified camera movements and changes;
- Completely accurate description of motion temporal relationships and directions;
- No errors or only negligible minor inaccuracies.

7-8 points (High Correctness)
- Correctly described most core motions, activities, and interactions in the video;
- Basically correctly identified the main camera movements;
- Generally accurate description of motion temporal relationships and directions;
- 1-2 minor errors that don't affect the overall understanding of video dynamic content.

5-6 points (Medium Correctness)
- Correctly described some core motions, activities, and interactions;
- Partially identified camera movements;
- Some confusion in motion temporal sequence or direction description;
- Several obvious errors, but core motion descriptions remain partially correct.

3-4 points (Low Correctness)
- Correctly described only a few motions or activities;
- Incorrect or missing description of camera movements;
- Numerous errors in motion temporal sequence or direction description;
- Multiple obvious errors, significant misunderstanding of video content.

1-2 points (Extremely Low Correctness)
- Almost no correct description of any actual motions or activities;
- Severe misunderstanding of video content, numerous errors or fabricated content;
- Completely confused motion temporal sequence;
- Description almost completely inconsistent with actual video content.

Evaluation Dimension 2: Detailedness (1-10 points)
Evaluate whether the model comprehensively and thoroughly describes the dynamic content in the video, including temporal dynamics, camera movements, motions, activities, and interaction details.

Detailedness Rating Criteria:

9-10 points (Extremely High Detailedness)
- Comprehensively captured details of all key motions and activities in the video;
- Detailed description of how motions are executed (e.g., speed, force, amplitude);
- Complete capture of temporal dynamics and motion transitions;
- Precise description of various camera movements and changes (e.g., panning, pushing/pulling, rotation);
- In-depth analysis of interaction relationships and dynamic changes in the scene.

7-8 points (High Detailedness)
- Captured details of most key motions and activities in the video;
- Described the execution manner of most motions;
- Good capture of temporal dynamics and main motion transitions;
- Described the main camera movements;
- Analyzed the main interaction relationships.

Table S10: Prompt template for GPT-assisted evaluation of FAVOR-Bench (Part 2).

Prompt Template for GPT-Assisted Evaluation (2/2)

5-6 points (Medium Detailedness)
- Captured details of some key motions and activities;
- Partially described how motions are executed;
- Basic capture of temporal dynamics;
- Mentioned some camera movements;
- Included some interaction relationship descriptions.

3-4 points (Low Detailedness)
- Provided only basic descriptions of motions and activities, lacking details;
- Rarely described how motions are executed;
- Brief description of temporal dynamics;
- Almost no mention of camera movements;
- Insufficient description of interaction relationships.

1-2 points (Extremely Low Detailedness)
- Extremely brief, mentioning only the most basic motions;
- No description of how motions are executed;
- Missing temporal dynamics;
- Completely ignored camera movements;
- No description of interaction relationships.

Scoring Guiding Principles:
1.**Ignore Static Content Assessment:** Scoring should focus on dynamic content; points should not be heavily deducted for missing static scene descriptions.
2.**Tolerate Expression Differences:** Different expressions (e.g., “moves to the left" vs. “walks toward the window") should be considered equivalent if they refer to the same motion.
3.**Correctness First:** If a description is seriously incorrect, it should not receive high scores even if detailed.
4.**Distinguish Between Omissions and Errors:** Not mentioning certain content (omission) should not be treated the same as incorrect descriptions; errors should more severely impact correctness scores.
5.**Distinguish Primary from Secondary:** Correct descriptions of core motions/primary activities are more important than minor details.

Output Format:

Correctness Analysis
[Detailed analysis of how well the model’s description matches the actual video content, pointing out correct aspects and errors]

Detailedness Analysis
[Detailed analysis of the comprehensiveness and richness of details in the model’s description, pointing out detailed aspects and shortcomings]

Ratings
Correctness Rating: [Integer from 1-10]
Detailedness Rating: [Integer from 1-10]

Human-Annotated Standard Description (Caption)
{caption}

Model Response
{response}

Annotation Guidelines for Structured Video Motion Annotation

1. **Task Description:** Your task is to distinguish subjects with temporal dynamics and provide comprehensive structured annotations including the attributes of all subjects, camera motion sequence, motion sequence of each subject, and the complete caption of the video.
2. **Video Discarding Guideline:** If the following situations occur, no annotation is required and the video will be discarded:
   (1). The video contains 4 or more subjects with temporal dynamics.
   (2). The camera is tilted for more than 45 degrees for more than a half of the video time.
   (3). Multiple subjects that can't be distinguished by their appearances.
   (4). The video has no temporal dynamics for more than half of the time.
3. **Detailed Annotation Guideline for each Components:**
   (1). Attributes of all subjects: For people, please consider the following attributes: clothes, shoes, hairstyle, belongings, mask, glasses, and accessories. provide at least 3 attributes for one person if visible. For none-person subjects like animals, at least 1 attribute is required if visible.
   (2). Camera motion sequence: Please provide camera motions with corresponding start and end time (accurate to the second). Some reference camera motions: Camera shake, pan <direction>, zoom in/out, camera focus change from xx to xx. First person videos do not require camera movements.
   (3). Motion sequence of each subject: Please provide all actions/motions done by each subject and their start and end time (accurate to the second).
   (4). Complete caption of the video: Please provide a fluent description about the video. All the above motions in motion sequence should be included. At least one attribute of each subject (if has) should be included.
4. **Annotation Format:** All annotations should follow the format below, brackets are necessary.
5. **Example of Structured Video Motion Annotation:**

camera_motion: Move camera to the lower left (0-3), move camera to the upper right (4-6), camera shake (7-8), move camera to the left (9-13), move camera to the right (14-16), camera shake (17-31),

subject_attributes: Subject 1: Man [wearing black clothes, black pants, black hair] Subject 2: Man [wearing blue clothes, black pants, black hair],

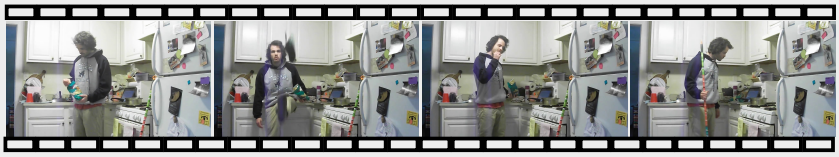
motion_list: Subject 1: Man [bend down to pick up a bowl (1-2), stand up (3-4), pick up a spoon and stir in the bowl (5-7), put down the spoon (8-9), hold the bowl with both hands (9-11)] Subject 2: Man [body shaking (11-29), holding and shaking clothes (12-29), repeatedly throwing clothes (17-30), raising hand and walking to the right (29-31)],

caption: The camera moves to the lower left, the man in black bends down to pick up a bowl and stands up, the camera moves to the upper right, the man in black picks up a spoon and stirs it in the bowl, the camera shakes, the man in black puts down the spoon and holds the bowl with both hands, the camera moves to the left, the man in blue clothes shakes his body, the camera moves to the right, the man in blue clothes holds and shakes clothes, the camera shakes, the man in blue clothes repeatedly throws the clothes onto a rope, then raises his hand and walks to the right.

Figure S5: Annotation guidelines of structured video motion annotation and a case of our testing video and corresponding annotations.

Task type: Action Sequence (AS)

Which series of continuous actions did the woman in the blue shirt complete before bending down to dust with the vacuum cleaner?
(1) Hands vase to woman while speaking →Turns right and raises hand →Removes glasses and moves right;
(2) Removes glasses and moves right →Turns right and raises hand →Hands vase to woman while speaking;
(3) Turns right and raises hand → Removes glasses and moves right → Woman takes vase;
(4) Woman moves right while he removes glasses → Turns and raises hand before handing vase → speaking;
(5) Moves right while speaking → Hands vase when turning → Woman removes glasses after receiving.



Regarding the chronological order of the gray-clothed man's shoe flicking action, which of the following descriptions is correct?
(1) The shoe flicking occurs before eating the first potato chip;
(2) The shoe flicking occurs after eating the second potato chip;
(3) The shoe flicking occurs after putting down the bag;
(4) The shoe flicking occurs after the first potato chip and before the second potato chip;
(5) The shoe flicking happens simultaneously with the broom swaying action.

The man (first-person protagonist) wearing a black outfit with golden patterns, after completing the rinsing of the bottle cap, which two consecutive actions does he perform next?
(1) Take the bottle cap with the left hand and pour water → pick up the bottle body with the right hand;
(2) Pick up the bottle cap with the right hand → rinse the bottle cap;
(3) Turn around and walk to the cooking counter → put down the items;
(4) Turn on the faucet with the right hand → place the bottle body under the faucet with the left hand;
(5) Put down the bottle body with the left hand → pick up the bottle cap with the right hand.

Figure S6: Examples of Action Sequence (AS) task in the close-ended evaluation.

Task type: Holistic Action Classification (HAC)

What is the main action of the woman in black pants throughout the video?
(1) Holding a broom;
(2) Covering with a blanket;
(3) Turning around;
(4) Waving the broom;
(5) Putting down the broom.

What is the core behavior of the man wearing a white top and black-and-white checkered pants throughout the entire video?
(1) Repeatedly opening and closing the wardrobe door to check its structure;
(2) Cleaning ceiling dust with a broom;
(3) Organizing clothes and cleaning the area around the wardrobe;
(4) Swiftly waving a broom for exercise;
(5) Adjusting clothing positions multiple times.

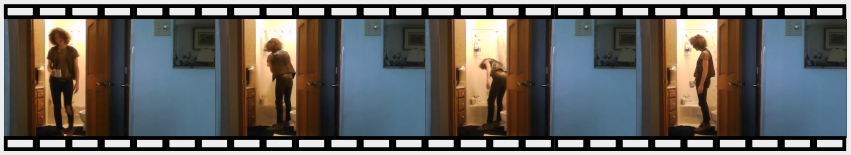
What is the main action performed by the man in black suit and black pants in the video?
(1) Open the cabinet and take out items;
(2) Pace back and forth in the room;
(3) Carry boxes to the other side;
(4) Talk to a man wearing a necklace;
(5) Close the cabinet and walk forward.

Figure S7: Examples of Holistic Action Classification (HAC) task in the close-ended evaluation.

Task type: Single Action Detail (SAD)

At the end of the video, what action does the girl perform on the item of the black-haired man?
(1) Suddenly snatch the book from the man's hands;
(2) Shake the book vertically above her head;
(3) Flip through the pages rapidly to make noise;
(4) Place the book flat on the table and press it down;
(5) Tap the man's shoulder three times with the book.

When the man in the gray shirt interacts with the bread on the plate, which action most precisely describes the interaction details?
(1) takes a bite from another piece of bread on the table;
(2) tears the bread into halves and places it back on the plate;
(3) dips the bread into the sauce in the pot;
(4) picks up the bread from the plate and puts it into his mouth to chew;
(5) hands the bread to the person wearing checkered pants.

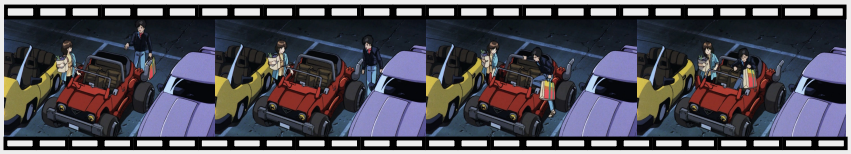


When the man is pouring water, where does the liquid in the cup end up being poured?
(1) Poured into an area behind the curtain;
(2) Poured into a trash can;
(3) Poured into a sink;
(4) Splashed directly onto the ground;
(5) Poured onto the windowsill edge.

Figure S8: Examples of Single Action Detail (SAD) task in the close-ended evaluation.

Task type: Multiple Action Details (MAD)

What interactions did the man in gray clothes have with the cup at different times?
(1) Repeatedly picked up the cup to drink water and blew on the cup midway;
(2) Continuously wiped the surface of the cup with his hand;
(3) Handed the cup to the boy in yellow clothes and then took it back;
(4) Placed the cup upside down on the table while drinking water;
(5) Briefly picked up the cup to drink water and immediately put it down.



At the end of the video, what actions do the man in black clothes and the woman with brown hair correspond to respectively?
(1) The man closes the trunk, the woman starts walking;
(2) The woman organizes the trunk, the man adjusts the steering wheel;
(3) The man picks up a bag and speaks, the woman holds the car door;
(4) Both simultaneously open left and right car doors;
(5) The man gets into the car and sits down, the woman opens the car door.

In the video, which items did the first-person protagonist directly interact with at different moments?
(1) Spray kettle, countertop, blue tool, scissors;
(2) Blue tool, cabinet door, cup, basin;
(3) Spray kettle, countertop, cabinet door, basin;
(4) Spray kettle, blue tool, cabinet door, basin;
(5) Spray kettle, blue tool, scissors, basin.

Figure S9: Examples of Multiple Action Details (MAD) task in the close-ended evaluation.

Task type: Camera Motion (CM)

In the video, what is the main directional change in camera movement?
(1) The camera remains stationary throughout;
(2) The camera first zooms in and then zooms out;
(3) The camera continuously shakes without directional change;
(4) The camera first rotates clockwise then counterclockwise;
(5) The camera moves vertically and returns to its original position.

What is the main movement direction of the camera in the video?
(1) Downwards;
(2) Upwards;
(3) Left and right shaking;
(4) Continuous rotation;
(5) Completely stationary.

Task type: Non-Subject Motion (NSM)

Which element in the video primarily demonstrates environmental dynamics unrelated to the main subject?
(1) The curtain being pulled and opened/closed multiple times;
(2) The ceiling fan rotating at high speed;
(3) The poster on the wall suddenly falling off;
(4) The rear curtain naturally sways;
(5) Water stains spreading on the floor.

Figure S10: Examples of Camera Motion (CM) and Non-Subject Motion (NSM) tasks in the close-ended evaluation.