GLIMPSE: Do Large Vision-Language Models Truly *Think With Videos* or Just Glimpse at Them?

Yiyang Zhou*[1], Linjie Li*[2], Shi Qiu*[1], Zhengyuan Yang[2], Yuyang Zhao[3], Siwei Han[1], Yangfan He[4], Kangqi Li[5], Haonian Ji[1], Zihao Zhao[6], Haibo Tong[7], Lijuan Wang[2], Huaxiu Yao[1]
[1]UNC Chapel-Hill, [2]Microsoft, [3]NUS, [4]University of Minnesota
[5]Case Western Reserve University [6]Rutgers University, [7]UIUC
{yiyangai, shiqiu, huaxiu}@cs.unc.edu

Abstract

Existing video benchmarks often resemble image-based benchmarks, with question types like "What actions does the person perform throughout the video?" or "What color is the woman's dress in the video?" For these, models can often answer by scanning just a few key frames, without deep temporal reasoning. This limits our ability to assess whether large vision-language models (LVLMs) can truly think with videos rather than perform superficial frame-level analysis. To address this, we introduce GLIMPSE, a benchmark specifically designed to evaluate whether LVLMs can genuinely think with videos. Unlike prior benchmarks, GLIMPSE emphasizes comprehensive video understanding beyond static image cues. It consists of 3,269 videos and over 4,342 highly visual-centric questions across 11 categories, including Trajectory Analysis, Temporal Reasoning, and Forensics Detection. All questions are carefully crafted by human annotators and require watching the entire video and reasoning over full video context—this is what we mean by thinking with video. These questions cannot be answered by scanning selected frames or relying on text alone. In human evaluations, GLIMPSE achieves 94.82% accuracy, but current LVLMs face significant challenges. Even the best-performing model, GPT-o3, reaches only 66.43%, highlighting that LVLMs still struggle to move beyond surface-level reasoning to truly think with videos. We publicly release our benchmark and code at https://github.com/aiming-lab/GLIMPSE.

1 Introduction

The rapid advancement of large vision-language models (LVLMs) has enabled sophisticated tasks involving both visual and textual understanding, such as image and video comprehension. Models like GPT-4o, o3 (OpenAI, 2024), Gemini (Team et al., 2023), and several open-source video LVLMs (Lin et al., 2023; Zhang et al., 2024; Wang et al., 2024a; Zhang et al., 2023; Yang et al., 2025; Xu et al., 2024) show strong performance on video understanding and instruction-following tasks. They also perform well on established video benchmarks (Cai et al., 2024; Mangalam et al., 2023; Yu et al., 2019; Xiao et al., 2021; Fu et al., 2024; Yu et al., 2025; Patraucean et al., 2023; Kesen et al., 2023; Li et al., 2024b; Saravanan et al., 2025; Li et al., 2022). However, current benchmarks have notable limitations: many include image-level questions answerable from a single frame (Cai et al., 2024; Li et al., 2024b; Saravanan et al., 2025; Li et al., 2022), and they often lack diversity in video content and types, limiting a comprehensive assessment of whether LVLMs can truly think based on video. For example, in Figure 1, a video where sequential actions occur (e.g., a person skating around an ice rink, moving out of frame and then back into frame), the model can often answer questions like "What is the person doing?" by identifying the general activity as "skating in cir-

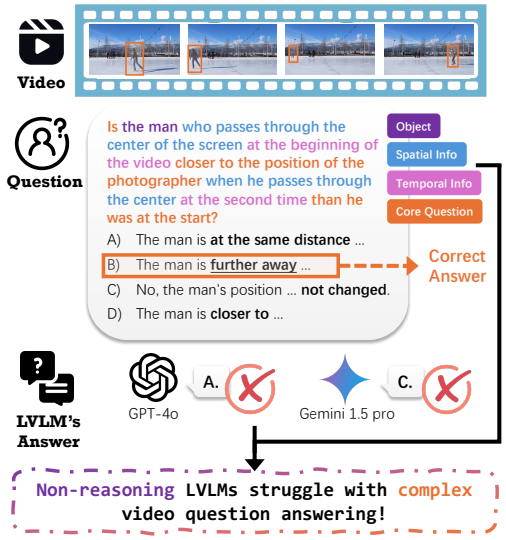


Figure 1: LVLMs without reasoning ability struggle with complex video question answering.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 27831–27845
November 4-9, 2025

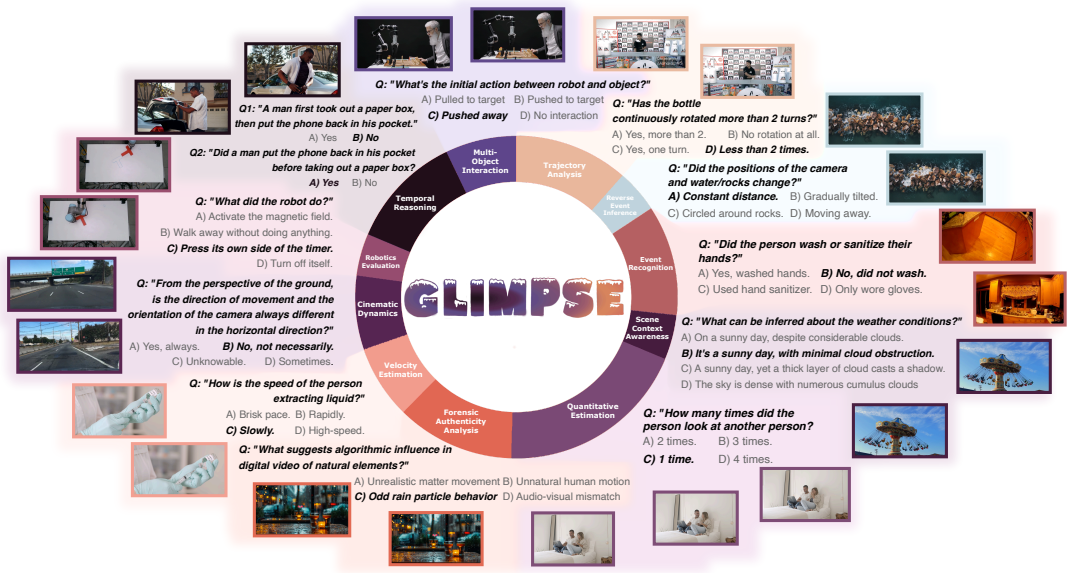


Figure 2: GLIMPSE categorizes visual-centric questions in video data into 11 distinct types, with representative examples from each category illustrated in the figure. Human-annotated questions and answers are reformatted into multiple-choice format for LVLMs. To reduce bias, yes/no questions are presented bidirectionally, requiring correct answers in both directions to be considered accurate.

cles," but this fails to test true video understanding. For instance, the model cannot determine whether the person moves closer to or farther from the camera first, which requires temporal reasoning across multiple frames.

To address these limitations, this paper proposes GLIMPSE (Figure 2), a new video benchmark designed to evaluate the video understanding capabilities of LVLMs. Compared to other video benchmarks (Li et al., 2024a; Ning et al., 2023; Li et al., 2023c; Liu et al., 2024c; Li et al., 2024b; Saravanan et al., 2025; Li et al., 2022), it leverages the temporal characteristics of 3,269 video data and manually annotates 4,342 high-quality visual-centric question pairs across 11 different scenarios. Specifically, GLIMPSE challenges models to truly think with videos rather than perform simple understanding, featuring the following characteristics:

Emphasis on deep reasoning content. Unlike static images, videos contain dynamic information across temporal and spatial dimensions. We manually selected videos with multiple dynamic entities, diverse movements, and complex positional changes that require comprehensive cross-frame analysis. We then constructed questions demanding deep, multi-faceted reasoning rather than superficial pattern recognition. For example, as shown in Figure 2, in the "Forensic Authenticity Analysis" category, the question "What suggests algorithmic influence in digital video of natural elements?" requires models to detect subtle inconsistencies in frame transitions and identify artificial backgrounds through detailed analysis. This challenges models to truly think with videos by integrating spatial relationships, temporal sequences, and contextual understanding to uncover complex visual anomalies.

Finer-grained categorization. Our benchmark spans 11 categories focused on video-specific visual tasks, including trajectory analysis, temporal reasoning, quantitative estimation, event recognition, sequential ordering, and scene context awareness. To enhance coverage, we also include velocity estimation, cinematic dynamics, forensic authenticity analysis, robotics action recognition, and interaction analysis, ensuring GLIMPSE remains comprehensive for evaluating text-to-video generation and embodied environments.

In addition to these major characteristics, to further facilitate automated evaluation and reduce biases during the assessment process, we structured the questions in a multiple-choice format. For yes/no-type questions, we created paired questions by reversing the order of actions or subjects. For example, "Does the person in the video open the cabinet before turning on the light? Answer: yes" is paired with "Does the person in the video turn on the light before opening the cabinet? Answer: no". The model is considered correct only when it answers both questions in the pair accurately.

Using GLIMPSE, we conducted a comprehensive evaluation of several state-of-the-art

LVLMs, including GPT-4o, o3 (OpenAI, 2024), Gemini-1.5 (Team et al., 2024), along with open sourece LVLMs like VideoLLaVA (Lin et al., 2023), LLaVA-NeXT-Video (Zhang et al., 2024), Video-LLaMA (Zhang et al., 2023), Video-LLaMA2 (Cheng et al., 2024),Chat-UniVi-V1.5 (Jin et al., 2024) and Qwen2-vl (Wang et al., 2024a). As a comparison, we also conducted experiments on LVLMs (Liu et al., 2024a; Ye et al., 2024; Bai et al., 2023) fine-tuned on image-based instruction data to demonstrate that the tasks in GLIMPSE cannot be effectively answered using single-frame images. We found that the most advanced LVLMs, Gemini-1.5 pro (Team et al., 2024), achieved an accuracy of only 56.98% on GLIMPSE, significantly lower than the average score 94.82% of human volunteers on a randomly sampled subset. Among the open-source models fine-tuned on video datasets, the best-performing model, Qwen2-VL, also achieved only 52.47% accuracy. Furthermore, the best-performing image-based LVLMs, LLAVA-1.5, achieved an accuracy of 37.48%, indicating that relying solely on single-frame or multi-frame image data in a single modality is insufficient for effectively utilizing and capturing the temporal information in video data. Additionally, we observed that LVLMs perform worse on videos with longer average lengths, highlighting that current LVLMs struggle to handle long video sequences effectively. The issues exposed by these findings provide guidance for the subsequent optimization of the model.

2 Related Work

Large Vision Language Models. Large language models (LLMs) (OpenAI, 2023; Touvron et al., 2023; Taori et al., 2023; Chiang et al., 2023) have demonstrated impressive text comprehension capabilities. With the integration of visual components (Ye et al., 2024; Zhu et al., 2023; Liu et al., 2023b; Qu et al.; Zhou et al., 2024; Wang et al., 2024b), these text models have gained the ability to understand multimodal data, allowing them to interpret real-world images. Notable examples include commercial models like GPT-4V (OpenAI, 2023) and Gemini-1.5 (Team et al., 2024), as well as open-source models such as LLaVA-1.5 (Liu et al., 2024a), Qwen-VL (Bai et al., 2023), and mPlug-Owl (Ye et al., 2023). The development of LVLMs has expanded their capabilities from static image understanding to complex video comprehension. Early work, such as VideoChat (Li et al., 2023b) and Video-ChatGPT (Maaz et al., 2023), integrated visual encoders with language models, laying the foundation for video-based multimodal dialogue. Subsequent research introduced improvements (Zhang et al., 2023; Jin et al., 2023; Ren et al., 2024; Song et al., 2024; Liu et al., 2023a), including the addition of audio modalities, joint training on images and videos, and optimization of feature alignment.

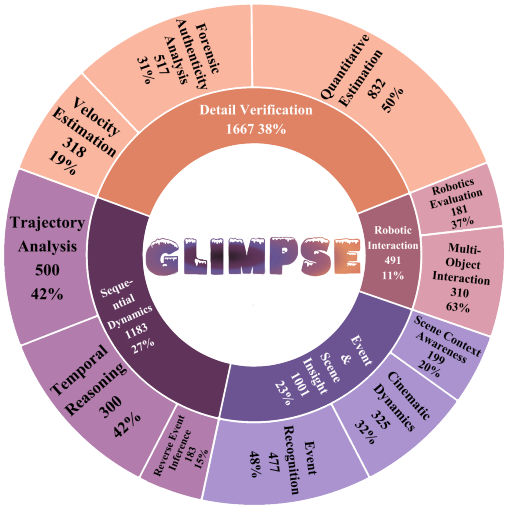


Figure 3: Distribution of categories in the GLIMPSE benchmark. Our benchmark covers 4 key domains and 11 detailed visual-centric question types.

Video Understanding Benchmarks. Benchmarking efforts have increasingly extended from static images to the video domain, where temporal dynamics and content complexity demand new evaluation approaches. Early work like SEED-Bench (Li et al., 2023a) supports multimodal evaluation for both image and video QA, including temporal modeling dimensions. AutoEval-Video (Chen et al., 2023) and Video-Bench (Ning et al., 2023) are specifically designed for video scenarios and generate video QA data using LLMs. MVBench (Li et al., 2024a) introduces an innovative approach by repurposing existing datasets for LVLMs evaluation. ET-Bench (Liu et al., 2024b) targets complex, multi-event, time-sensitive tasks to assess temporal and contextual understanding. Despite this progress, no existing benchmark specifically targets high-quality, vision-centric video QA. To address this gap, this paper introduces GLIMPSE, a benchmark focused on evaluating LVLMs' visual perception and reasoning in video contexts.

3 The GLIMPSE Benchmark

3.1 Overview

In this section, we introduce GLIMPSE, a benchmark focused on visual-centric content. By "visual-

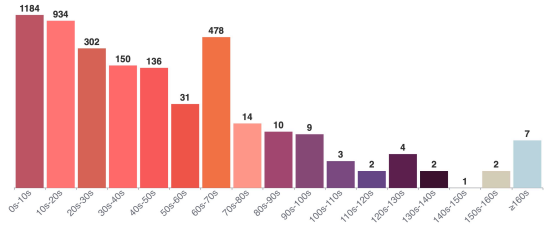


Figure 4: Video length distribution, with the horizontal axis showing duration and the vertical axis indicating the number of videos per range.

centric," we refer to content where understanding requires comprehensive analysis of dynamic visual elements across multiple frames, such as tracking object movements, analyzing complex interactions, and interpreting evolving spatial arrangements, rather than relying on superficial glimpses or single-frame observations. This benchmark challenges models to truly think with videos through sustained analytical engagement. As shown in Figures 2 and 3, GLIMPSE comprises a total of 3,269 carefully selected video instances and 4,342 unique questions, all manually constructed. Answering each question requires a comprehensive understanding of the video. Based on the temporal characteristics of video data, we categorize the questions into 11 subcategories.

3.2 Dataset Curation

The data curation process of the GLIMPSE dataset consists of three key steps: video collection, question-answer annotation, and quality review. Each step is designed to ensure comprehensive coverage and high-quality representation across various categories in our benchmark, while also making certain that each question is visual-centric and closely aligned with the benchmark's theme. We detail the process as follows:

Video Collection. To annotate visual-centric QA pairs, we first select a comprehensive set of dynamic videos covering a wide range of types. We began by identifying the visual content of interest in each video, categorizing it into 11 key areas: **(1) Trajectory Analysis:** This task focuses on analyzing the trajectory or path of objects within the video, involving an understanding of movement direction, displacement over time, and overall motion patterns. The model needs to comprehend the video, identify the objects to be tracked, and make integrated judgments based on the video's context. This helps evaluate the model's fine-grained recognition and temporal reasoning abilities. **(2) Temporal Reasoning:** Involves understanding the timing and sequence of events. Although previous benchmarks, such as (Fu et al., 2024; Li et al., 2024a; Liu et al., 2024b; Ning et al., 2023), have explored similar tasks, they mostly focus on locating a single action, object, or event (e.g., asking when an object appears). In contrast, GLIMPSE aims to provide a more visually-centric evaluation by asking about the temporal order of events, such as whether a specific event occurred before another. This approach effectively assesses the model's temporal reasoning ability. **(3) Quantitative Estimation:** This task involves estimating quantities such as distances or counts within the video, adapted from quantitative tasks in image modalities. In GLIMPSE, we focus on dynamic events—for example, counting how many times a person performs a specific action or how many times an animal appears and disappears. An example question could be, "How many times did the dog and the person clap hands in the video?" This type of question emphasizes the model's ability to understand and track dynamic information in video content. **(4) Event Recognition:** This task aims to determine whether events occur in the video and their sequence relative to other events, assessing the model's understanding of video content and temporal relationships, especially in scenarios where multiple events happen sequentially. The model must accurately identify events and the logical order between them. **(5) Reverse Event Inference:** Determines the correct sequence of events or actions, reconstructing the event flow from partial information. **(6) Scene Context Awareness:** Understands changes in the background of the scene in the video (e.g., location or setting). This task evaluates the model's spatial understanding and context recognition abilities. For example, "How do the traffic lights change throughout the video?" **(7) Velocity Estimation:** Calculates the relative speed of moving objects, analyzing their displacement over time. For example, "How does the speed of the black and white horses change?" **(8) Cinematic Dynamics:** This task focuses on identifying camera motion in video footage, presenting a unique challenge. The model needs to fully understand both the foreground and background of the video and analyze their movement in relation to each other to determine the style of camera motion. For example, "How does the camera move throughout the video?" **(9) Forensic Authenticity Analysis:** By collecting and generating some fake video data using text-to-video models (Podell et al., 2023), we can then use this data for evalu-

ation to test whether current models can identify fake videos. This approach effectively verifies the model's ability in assessing video authenticity. **(10) Robotics Evaluation:** Identifies actions performed by robots, such as grasping, moving, and assembling, to assess the stability, success, and effectiveness of various robotic tasks. **(11) Multi-Object Interaction:** Focuses on analyzing the interactions between multiple objects or entities (such as robots, people, animals, or specific items) within a given scenario, including actions like physical contact, collaboration, or conflict. For instance, "What did the person do with the box when they approached it in the final scene of the video?"

After defining the categories, we carefully selected and collected high-quality video resources that meet the specific requirements of each category to ensure comprehensive coverage across the benchmark. For event recognition, we extracted and reconstructed data from ShareGPTVideo (Chen et al., 2024a). Robotics action recognition included videos sourced from the PushT dataset (Chi et al., 2024) and Cable Routing (Luo et al., 2023). For temporal reasoning, we chose videos from Ego4D (Grauman et al., 2022) and Pexels[1]. Quantitative estimation and velocity estimation tasks utilized selected videos from Panda-70m (Chen et al., 2024b). The remaining videos were curated from Pexels and Pixabay[2]. Additionally, to avoid excessive complexity in the testing tasks and maintain reference value, we controlled the length of the selected videos to be between 20 seconds and 2 minutes. The final dataset includes 3,269 videos with a balanced distribution of video lengths, as shown in Figure 4.

Question-answer Annotation. After collecting the raw video data, we manually annotated high-quality question-answer pairs to evaluate the ability of LVLMs in interpreting video content. To facilitate the annotation process, we enlisted researchers proficient in English to create open-ended questions and answers. Specifically, the annotators first watched the entire video and, by rewatching as needed, generated 1–3 questions per video. For easier testing, we then used GPT-4o API to convert these open-ended question-answer annotations into a multiple-choice format suitable for automated evaluation. For yes/no-type questions, such as those in the "temporal reasoning" category, we aimed to reduce bias during evaluation (e.g., random guessing with a 50% accuracy rate). To address this, we used the GPT API to construct bidirectional question pairs, as exemplified in the "temporal reasoning" samples shown in Figure 2. For example, we initially manually annotated the question-answer pair as: Q1: "A man first took out a paper box, then put the phone back in his pocket." A) Yes B) No. Then we used GPT-4o API to modify it into a reverse question-answer pair: Q2: "Did a man put the phone back in his pocket before taking out a paper box?" A) Yes B) No. The specific prompts used can be found in Appendix A. Only when LVLMs answered both questions correctly was the response considered accurate.

Quality Review. To ensure the quality of the dataset and confirm that the manually constructed questions and selected videos are indeed visual-centric, we implemented a rigorous manual review process. First, different annotators were assigned to check each question-answer pair to ensure that each question is well-defined and answerable based on the video content. For example, questions like "What is the approximate speed of the vehicle in the video?" lack a clear answer and are removed during screening. We also ensured that each question required understanding of the entire video rather than a single frame. For example, questions like "What is the weather in the video?" could be answered with just one frame, so similar questions were filtered out during the review process. Through annotation, manual selection and quality review, we ultimately collected a total of 4342 high-quality question-answer pairs. Detailed examples can be found in Appendix B.

4 Experiment

GLIMPSE systematically evaluates the understanding and perception capabilities of existing LVLMs on video data. In this section, we aim to address the following key questions: (1) Can current LVLMs truly think with video content and answer questions accurately? (2) How large is the gap between current LVLMs and human performance in understanding video content? (3) Do the evaluated LVLMs show a preference for a certain field?

4.1 Experiment Setup

Baseline Models. We first tested (1) three commercial LVLMs: GPT-4o, o3 (OpenAI, 2024), Gemini 1.5 Flash, and Gemini 1.5 Pro (Team et al., 2024) on the GLIMPSE benchmark; (2) eval-

[1] https://www.pexels.com/
[2] https://pixabay.com/

Models	TA	TR	QE	ER	REI	SCA	VE	CD	FAA	RE	MOI	Avg
Random	23.60	25.40	23.92	24.32	16.39	25.63	25.18	24.31	27.08	23.53	22.58	24.14
Open-source Image LVLMs: All models use a randomly sampled single frame as input.												
mPLUG-OWL2 (7B)	39.25	22.40	36.90	31.80	34.00	37.30	28.60	38.40	52.70	33.50	28.34	34.12
Qwen-VL-Chat (7B)	38.24	11.60	35.71	30.72	32.79	36.18	27.48	33.16	51.62	32.04	27.42	30.73
LLAVA-1.5 (7B)	42.03	28.47	30.02	42.96	45.53	28.52	29.48	41.47	67.04	30.01	27.03	37.48
Open-source Video LVLMs: All models use their default numbers of frames as inputs.												
Video-LLaMA (7B)	44.71	28.88	23.30	48.94	38.00	60.46	15.32	53.30	65.09	42.13	51.16	39.71
Video-LLaMA2 (7B)	47.06	30.40	24.53	51.52	40.00	63.64	16.13	56.10	68.52	65.19	53.85	42.60
Chat-UniVi-V1.5 (7B)	39.86	23.68	22.36	54.10	42.00	66.82	16.94	**58.91**	71.95	68.45	56.54	41.47
LLAVA-NeXT-Video (7B)	46.79	42.33	19.55	57.68	44.45	69.04	34.48	53.03	72.87	63.71	57.17	46.80
VideoLLaVA (7B)	40.42	22.56	21.86	52.60	42.99	68.00	17.60	56.48	71.39	65.23	54.95	40.74
Qwen2-VL (7B)	46.15	44.44	28.37	59.32	43.42	67.84	33.18	55.32	73.44	66.85	58.82	52.47
Closed-source LVLMs: All models use fixed interval sampling.												
GPT-4o	48.40	28.80	49.64	59.12	56.83	62.81	40.88	52.92	65.18	70.11	57.10	53.80
GPT-o3	55.42	**53.07**	**65.75**	61.51	**67.39**	**82.00**	**62.90**	56.23	**85.69**	69.55	**71.20**	**66.43**
Gemini 1.5 Flash	54.60	33.60	64.90	62.68	54.10	69.00	44.14	51.38	73.11	73.48	61.62	55.65
Gemini 1.5 Pro	**61.02**	42.84	51.32	**64.10**	56.86	72.12	45.45	53.97	75.24	**77.95**	62.64	56.98
Human Performance												
Human Expert	92.00	96.00	88.00	100.0	92.00	100.0	94.00	96.0	91.00	96.00	98.00	94.82

Table 1: The overall performance of representative LVLMs on the GLIMPSE benchmark across various categories. Here, we report accuracy for each specific category in terms of visual-centric question answering. Specifically, **TA** denotes Trajectory Analysis, **TR** denotes Temporal Reasoning, **QE** denotes Quantitative Estimation, **ER** denotes Event Recognition, **REI** denotes Reverse Event Inference, **SCA** denotes Scene Context Awareness, **VE** denotes Velocity Estimation, **CD** denotes Cinematic Dynamics, **FAA** denotes Forensic Authenticity Analysis, **RE** denotes Robotics Estimation, and **MOI** denotes Multi-Object Interaction. In each column, the best performance, excluding human experts, is bolded, and the second-best is underlined.

uated representative video-based LVLMs, including Video-LLaMA (Zhang et al., 2023), Video-LLaMA2 (Cheng et al., 2024), LLaVA-NeXT-Video (Zhang et al., 2024), VideoLLaVA (Lin et al., 2023), Qwen2-VL (Wang et al., 2024a), and Chat-UniVi-V1.5 (Jin et al., 2024); and (3) to examine the differences between image LVLMs and video-based LVLMs, we also tested representative models such as LLAVA 1.5 (Liu et al., 2024a), Qwen-VL (Bai et al., 2023), and MPlug-Owl 2 (Ye et al., 2024). For video-based models, we followed the frame sampling methods provided by the respective authors during evaluation. For image-based LVLMs, we randomly sampled a single frame as input. For GPT-4o, o3, Gemini-1.5 Flash and Gemini-1.5 Pro, we used fixed interval sampling with parameters set to sample_frequency = 50 and max_frame_num = 16.

Evaluation Metrics. We use accuracy to evaluate the performance of LVLMs on GLIMPSE, where for Temporal Reasoning, questions are constructed in a bidirectional format—meaning each video has two related questions. Thus, when calculating the accuracy for this category, we only count a response as correct if the model answers both questions correctly for a given video.

Human Evaluation. For comparison, we randomly selected 50 questions from each of the 11 classes in GLIMPSE, forming a 550-sample subset, and invited five student volunteers to answer them. Volunteers were instructed to watch each video once without replay and answer immediately. They also recorded the time spent per question (excluding video watching). The total time was 1327 seconds, averaging 2.4 seconds per question.

4.2 Quantitative Results

This section presents the evaluation results on GLIMPSE with detailed LVLM performance in Table 1. To provide a comparison baseline, we include a "Random" row, illustrating task difficulty and whether models perform above chance. "Random" results are computed by generating equal numbers of random A/B/C/D or Yes/No answers based on total question count. Based on Table 1 and human performance, our key findings are:

Gap between video-based LVLMs and image-based LVLMs. We found that image-finetuned models like LLAVA-1.5, mPLUG-OWL2, and Qwen-VL-Chat, while performing well on existing image benchmarks, showed weaker overall performance on the GLIMPSE benchmark compared to LVLMs fine-tuned on video data. This gap is especially noticeable in tasks requiring video context understanding, such as Trajectory Analysis, Temporal Reasoning, Quantitative Estimation,

and Sequential Ordering. For example, in the typically lowest-scoring task of Temporal Reasoning, the best-performing model, GPT-o3, scored 53.07, while the top image LVLMs, LLAVA-1.5, only achieved 28.47. Additionally, video LVLMs outperformed image LVLMs in this task by an average of 56.09% in accuracy. These findings indicate that specialized training on high-quality video data enhances model performance in tasks with high temporal requirements. The results also indicate that single-frame images lack sufficient information to answer questions in GLIMPSE. Improving visual components and developing methods to extract comprehensive and temporal video features may be crucial for further advancing video-based LVLMs performance.

Challenges of Current LVLMs in Temporal Understanding and Fine-Grained Visual Tasks. Based on the results in Table 4, it can be seen that although GPT-o3 performed the best on the GLIMPSE benchmark, its overall average score was only 66.43, significantly below human-level performance. Specifically, in challenging tasks within GLIMPSE such as Temporal Reasoning, Trajectory Analysis, and Quantitative Estimation, the overall performance of the tested models was generally modest. For Temporal Reasoning, the top model, GPT-o3, achieved only 53.07 accuracy, while other models were close to random performance levels. In Quantitative Estimation, apart from GPT-o3, which achieved 65.75 accuracy, other models also showed average performance. This indicates that current LVLMs still face significant challenges in understanding temporal information in videos and in fine-grained visual-centric tasks.

Image Understanding Capabilities Are Equally Important for Enhancing Video Understanding. In open-source video LVLMs, most models outperform open-source image LVLMs overall, especially in tasks requiring temporal information. However, in tasks such as Quantitative Estimation and Velocity Estimation, video LVLMs generally perform worse than image LVLMs and closed-source commercial LVLMs. This discrepancy may be due to video LVLMs focusing more on temporal and dynamic features and lacking sufficient optimization and fine-grained data to handle static quantitative estimation and speed calculation tasks, impacting their performance in these specific areas.

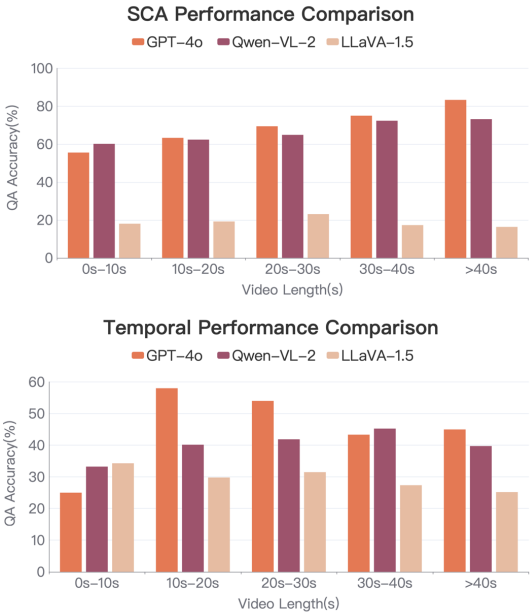


Figure 5: Model performance comparison in Scene Context Awareness (SCA) and Temporal Reasoning across different video durations. The y-axis shows accuracy, and the x-axis shows time intervals.

4.3 Time-Related Performance Analysis

Due to the varying amounts of information covered by videos of different lengths, we conducted an in-depth exploration of the impact of video length on model performance. We selected two categories with relatively even distributions of video duration: Temporal Reasoning and Scene Context Awareness, and tested three models from different categories in Table 4: GPT-4o, Qwen2-VL, and LLAVA-1.5. We divided the original questions into five subsets based on 10-second intervals and tested these models on each subset. The test settings were the same as in Table 4. We calculated the accuracy of these models for questions corresponding to videos within each time interval, and the results are shown in Figure 5. From the results, we can make the following observations:

Video Length	GPT-4o	Qwen2-VL	LLAVA-1.5
0–10s	42.33	34.67	30.00
10–20s	48.67	42.33	38.33
20–30s	55.00	48.00	45.00
30–40s	58.33	52.00	37.67
>40s	64.67	56.00	32.00

Table 2: The overall performance of models on different video lengths across 12 categories.

More information helps improve performance. For Scene Context Awareness tasks, the accuracy of the model's responses significantly improves as video length increases. This is because longer

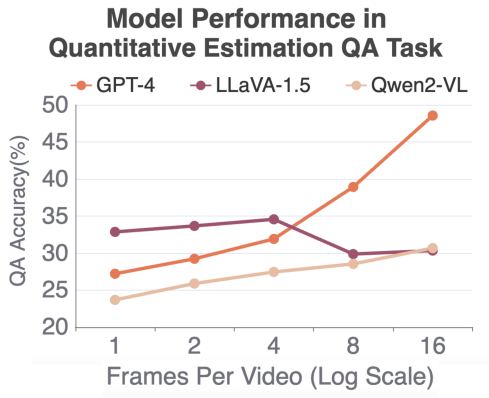


Figure 6: Model performance on quantitative estimation with varying frame counts.

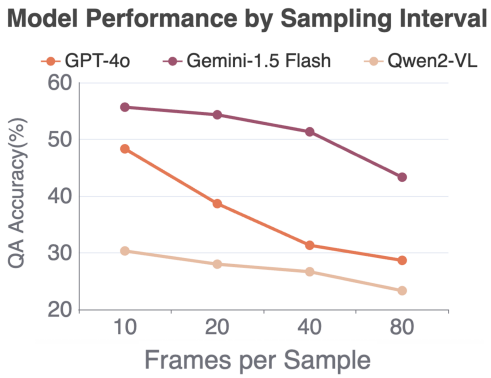


Figure 7: Model performance of Every N Frames sampling.

videos provide more information, allowing the model to capture richer background details and environmental context. This additional contextual information helps the model build a more comprehensive understanding of the scene, including objects, spatial relationships, and the overall atmosphere. With more time to "observe" and "learn" the nuances within the video, the model can more accurately grasp scene characteristics, making it easier to generate responses that reflect the actual context. This indicates that video length directly impacts model performance in scene understanding tasks, suggesting that long video understanding will be an important area of research in the future.

Challenges Remain in Processing Long Videos. For Temporal Reasoning tasks, model accuracy initially increases with video length but then declines. Short videos have limited time spans, and fewer frames are input, making it difficult for the model to capture enough information for accurate reasoning. As the video length increases to between 20 and 30 seconds, the additional information helps the model better understand temporal cues, thus improving reasoning accuracy. However, Temporal Reasoning is inherently challenging; as video length continues to exceed 30 seconds, the complexity and time span of events increase, dispersing the information. This makes it difficult for the model to effectively process long videos, leading to a drop in performance.

4.4 Effect of the Number of Input Frames

In this section, we analyze how frame count affects model performance, focusing on the Quantitative Estimation task. This task is especially sensitive to the amount of visual information, making it well-suited for examining the impact of varying frame counts. Unlike tasks that rely on temporal or sequential understanding, Quantitative Estimation requires detailed snapshots of individual frames, as these provide data points for counting or assessing quantities. By analyzing model performance on this task, we can more clearly observe how frame count affects the model's ability to interpret fine-grained visual information without the added complexity of temporal dependencies. We tested three models—GPT-4o, Qwen2-VL, and LLAVA-1.5—using 1, 2, 4, 8, and 16 frames as input, and plotted their performance as frame count increases. The experimental results are shown in Figure 6. We observe that initially, increasing the frame count improves the accuracy of all models; however, once the frame count surpasses a certain threshold, performance gains diminish and even start to decline. This suggests that, in the Quantitative Estimation task, a moderate number of frames enhances model accuracy, as more frames capture additional key details related to quantity. However, further increasing the frame count leads to a drop in performance, possibly because the model struggles to process excessive frame information effectively, resulting in redundancy and difficulty focusing on relevant features.

5 Error Analysis

This section provides an in-depth error analysis, focusing on two tasks with lower average accuracy: Temporal Reasoning and Quantitative Estimation, which are easy for humans but error-prone for models. Using GPT-4o and Gemini 1.5 Flash as examples, we present case studies in Figure 8.

5.1 Temporal Reasoning

In our experiments, we found that the Temporal Reasoning task had the highest error rate. For this task, to reduce potential biases, we structured each

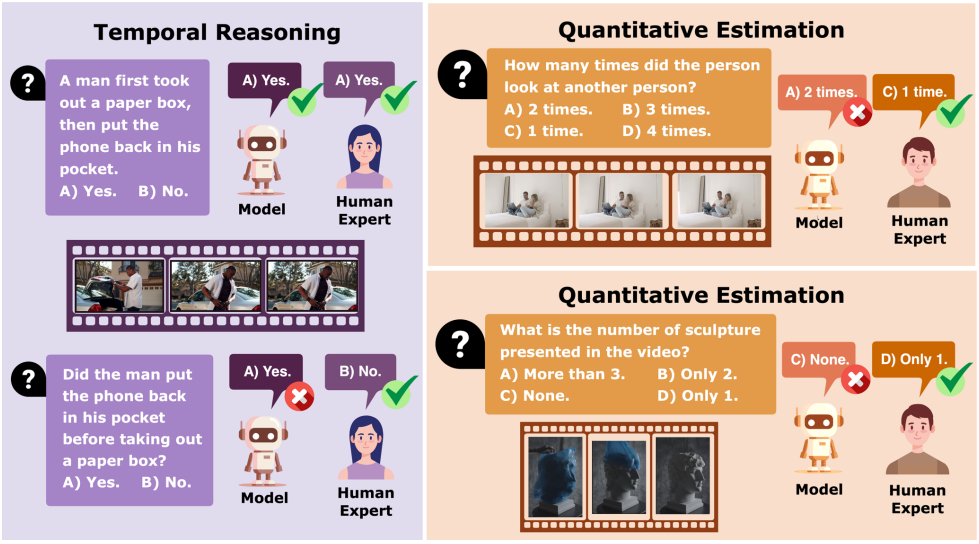


Figure 8: The failure cases of GPT-4o in GLIMPSE. It can be observed that GPT-4o has difficulty accurately understanding quantity-related questions in the video. Additionally, when answering temporal reasoning questions, it makes mistakes when the question is reversed.

temporal reasoning question to include contrasting statements, such as "Question: Does Event A occur before Event B? Answer: yes" and "Question: Does Event A occur after Event B? Answer: no". Only when both answers in a question pair were correct was the response considered accurate. Here, we refer to questions with a "yes" answer as positive questions. Under this setup, we found that when evaluating GPT-4o, if we assessed its accuracy on positive questions alone, it achieved 52.30%, but its accuracy on question pairs was only 28.80%. The primary source of errors in question pairs was an inconsistent selection pattern, where one answer was correct and the other incorrect. An example of this is shown on the right side of Figure 8. This suggests that GPT-4o does not truly understand and integrate the video content and question context when answering, as it may struggle to establish the correct temporal sequence during video encoding and frame extraction, thereby impacting its understanding of event order and sequence. It highlights that GLIMPSE can effectively reduce bias in the evaluation process by constructing bidirectional question-and-answer formats.

5.2 Quantitative Estimation

In the Quantitative Estimation task, there are two main issues: quantity errors and bias toward certain options. Using Gemini-1.5 Flash as an example, our investigation reveals that 62% of errors in Quantitative Estimation task occur when the model selects options like "None" or "No object", leading to incorrect quantity estimation. For instance, as shown in the example on the left of Figure 8, when asked how many sculptures were present in the video, the model responded with "None", despite a stationary sculpture being visible throughout the video. In addition, quantity errors are also significant. For example, in a question about the number of times a man looked at a woman in the video, GPT-4o counted each instance of the man turning his head toward the woman and then back as two separate actions, resulting in an inflated count. This type of quantity error may also stem from the model's difficulty in accurately understanding fine-grained, visually-centric information.

6 Conclusion

This paper presents GLIMPSE, a benchmark dataset for evaluating large vision-language models (LVLMs) on vision-centric video understanding and reasoning. GLIMPSE targets key skills such as trajectory tracking, temporal reasoning, quantity estimation, scene comprehension, and interaction analysis. It features diverse video content and carefully designed question pairs to comprehensively assess models' perceptual and reasoning abilities, as well as whether they can truly think based on video. To ensure accurate and scalable evaluation, GLIMPSE adopts multiple-choice and bidirectional QA formats that mitigate assessment bias. Experimental results reveal that even state-of-the-art multimodal models lag far behind human performance on GLIMPSE, underscoring the challenges and opportunities in deep video understanding.

Limitations

While GLIMPSE advances the evaluation of LVLMs in video understanding, it has several limitations. First, the dataset's focus on pre-selected video categories may underrepresent niche or unconventional scenarios, limiting generalizability. Second, restricting videos to 20 seconds–2 minutes overlooks challenges in ultra-long or extremely short contexts. Additionally, the multiple-choice format simplifies real-world open-ended reasoning, narrowing the assessment of nuanced understanding. Human annotation, though rigorous, introduces potential biases in question design and answer validation. Finally, benchmarked models may not fully encompass emerging or specialized LVLM architectures. Addressing these limitations could improve dataset diversity, evaluation flexibility, and model inclusivity in future work.

Acknowledgement

This research is partially supported by Cisco Faculty Research Award and Amazon Research Awards.

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, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, and 1 others. 2024. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. 2024a. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and 1 others. 2024b. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331.

Xiuyuan Chen, Yuan Lin, Yuchen Zhang, and Weiran Huang. 2023. Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering. *arXiv preprint arXiv:2311.14906*.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and 1 others. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. 2024. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, and 1 others. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012.

Peng Jin, Ryuichi Takanobu, Caiwan Zhang, Xiaochun Cao, and Li Yuan. 2023. Chat-univi: Unified visual representation empowers large language models with image and video understanding. *arXiv preprint arXiv:2311.08046*.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. 2024. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710.

Ilker Kesen, Andrea Pedrotti, Mustafa Dogan, Michele Cafagna, Emre Can Acikgoz, Letitia Parcalabescu, Iacer Calixto, Anette Frank, Albert Gatt, Aykut Erdem, and 1 others. 2023. Vilma: A zero-shot benchmark for linguistic and temporal grounding in video-language models. *arXiv preprint arXiv:2311.07022*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Jiangtong Li, Li Niu, and Liqing Zhang. 2022. From representation to reasoning: Towards both evidence

and commonsense reasoning for video question-answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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, and 1 others. 2024a. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Shuailin Li, Yuang Zhang, Yucheng Zhao, Qiuyue Wang, Fan Jia, Yingfei Liu, and Tiancai Wang. 2023c. Vlm-eval: A general evaluation on video large language models. *arXiv preprint arXiv:2311.11865*.

Yunxin Li, Xinyu Chen, Baotian Hu, Longyue Wang, Haoyuan Shi, and Min Zhang. 2024b. Videovista: A versatile benchmark for video understanding and reasoning. *Preprint*, arXiv:2406.11303.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Haogeng Liu, Qihang Fan, Tingkai Liu, Linjie Yang, Yunzhe Tao, Huaibo Huang, Ran He, and Hongxia Yang. 2023a. Video-teller: Enhancing cross-modal generation with fusion and decoupling. *arXiv preprint arXiv:2310.04991*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning.

Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Ying Shan, and Chang Wen Chen. 2024b. Et bench: Towards open-ended event-level video-language understanding. *arXiv preprint arXiv:2409.18111*.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024c. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*.

Jianlan Luo, Charles Xu, Xinyang Geng, Gilbert Feng, Kuan Fang, Liam Tan, Stefan Schaal, and Sergey Levine. 2023. Multi-stage cable routing through hierarchical imitation learning. *arXiv pre-print*.

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.

Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. 2023. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv preprint arXiv:2311.16103*.

OpenAI. 2023. Chatgpt. https://openai.com/blog/chatgpt/.

OpenAI. 2023. GPT-4V(ision) System Card.

OpenAI. 2024. Gpt-4o. https://openai.com/index/hello-gpt-4o/.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, and 1 others. 2023. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36:42748–42761.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Huaizhi Qu, Xinyu Zhao, Jie Peng, Kwonjoon Lee, Behzad Dariush, and Tianlong Chen. Uq-merge: Uncertainty guided multimodal large language model merging.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2024. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323.

Darshana Saravanan, Varun Gupta, Darshan Singh, Zeeshan Khan, Vineet Gandhi, and Makarand Tapaswi. 2025. VELOCITI: Benchmarking Video-Language Compositional Reasoning with Strict Entailment. In *Computer Vision and Pattern Recognition (CVPR)*.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, and 1 others. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xiyao Wang, Jiuhai Chen, Zhaoyang Wang, Yuhang Zhou, Yiyang Zhou, Huaxiu Yao, Tianyi Zhou, Tom Goldstein, Parminder Bhatia, Furong Huang, and 1 others. 2024b. Enhancing visual-language modality alignment in large vision language models via self-improvement. *arXiv preprint arXiv:2405.15973*.

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.

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*.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2025. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, and 1 others. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051.

En Yu, Kangheng Lin, Liang Zhao, Yana Wei, Zining Zhu, Haoran Wei, Jianjian Sun, Zheng Ge, Xiangyu Zhang, Jingyu Wang, and 1 others. 2025. Unhackable temporal rewarding for scalable video mllms. *arXiv preprint arXiv:2502.12081*.

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.

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, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024. Llava-next: A strong zero-shot video understanding model.

Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, and Huaxiu Yao. 2024. Calibrated self-rewarding vision language models. *arXiv preprint arXiv:2405.14622*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Prompt for Data Format Conversion

In this section, we list the prompts used during the data conversion process, where the entire data format transformation is performed using the GPT-4o API. We employed two types of conversions: (1) converting the original annotated data into a multiple-choice question format, and (2) transforming yes/no type questions into a bidirectional format.

Prompt for multiple-choice questions

[Task]: The task is to modify a question to make it significantly harder and more nuanced. The revised question should focus on specific technical aspects, avoid straightforward clues, and include plausible distractors that are incorrect. Use the correct answer as a subtle hint in the question, while making the other options appear equally viable.

[Guideline]:
Original question: "`{original_question}`"
Correct answer: "`{correct_answer}`"
Options: "`{options}`"

[Requirement]:
Please rewrite this question by following these strict guidelines:

- Focus on a specific, less obvious aspect that indirectly hints at the correct answer.
- Add realistic but irrelevant details to make the question look more complex.
- Avoid directly mirroring the wording of the correct answer, making the question more ambiguous.
- Ensure the distractor options contain minor technical errors or realistic details that seem logical but are incorrect.
- Add only contextual details related to the original question and correct answer; avoid introducing unrelated information.
- Do not change the correct answer.

[Output format]: Your response must be a single line, formatted exactly as follows:
"`...(modified question), Choices: A) xxx B) xxx C) xxx D) xxx`"
Only this format is allowed, and any deviations from this format are strictly forbidden.

Prompt for yes/no type questions

[Task]: Strictly follow the instructions. Rewrite the following question by reversing the temporal sequence and format it as a multiple-choice question with the following choices: A. Yes B. No.

[Guideline]:
Original question: "`{original_question}`"

[Requirement]:

- Reverse the temporal sequence of the original question while keeping it grammatically correct.
- Format the rewritten question as a multiple-choice question with the choices A. Yes and B. No.
- Ensure the output strictly follows the specified format without deviations.

[Output format]: Your response must be a single line, formatted exactly as follows:
"`...(statement).Choices: A.Yes B.No`"

Only this format is allowed, and any deviations from this format are strictly forbidden.

B Dataset Case

In this subsection, we present the constructed examples for each category in GLIMPSE, as shown in Figure 9 and Figure 10.

Trajectory Analysis

Question

Is the position of the person on the screen in the horizontal direction always unchanged?

A) Yes, the position is always unchanged.

B) No, the position can change.

C) The position may sometimes change.

D) The position often changes, but not always.

Answer

B

Temporal Reasoning

Question

The girl first took a picture of the left side of the frame, then took a picture of the right side of the frame.

A) Yes B) No

Did the girl take a picture of the right side of the frame before taking a picture of the left side of the frame?

A) Yes B) No

Answer

A, B

Quantitative Estimation

Question

How many notes are pasted on the upper surface except for the titles?

A) There are 5 notes.

B) There are 3 notes.

C) There are 7 notes.

D) There are 9 notes.

Answer

B

Event Recognition

Question

Did the character in the game not change clothes the entire time?

A) Yes, the character did not change clothes at any point.

B) No, they changed clothes once during the whole time.

C) They changed clothes multiple times during the game.

D) The character's clothes were constantly changing.

Answer

B

Reverse Event Inference

Question

Is the train [entering the platform] or [leaving the platform]?

A) The train is staying at the platform.

B) The train is leaving the platform.

C) The train is entering the platform.

D) The train is passing through the platform.

Answer

C

Scene Context Awareness

Question

Where the lady might be?

A) She is likely at the supermarket.

B) She might be at home.

C) She's probably in the park.

D) She could be at the work.

Answer

C

Figure 9: Dataset cases 1.

Velocity Estimation

Question

How is the speed of the wind turbine in the video?
A) The speed is high.
B) The speed is slow.
C) The speed is moderate.
D) The speed is not observable.

Answer

B

Cinematic Dynamics

Question

With the ground as the reference, does the camera continuously turn more than 90 degrees to the left in the horizontal direction?
A) The camera rotates to the right by 90 degrees.
B) The camera rotates to the left by 90 degrees.
C) The camera does not rotate at all.
D) The camera rotates to the left more than 90 degrees.

Answer

D

Forensic Authenticity Analysis

Question

Why do you think the video is generated by AI?
A) The characters in the video move too fluidly.
B) The color of the city in the video is oversaturated.
C) The video displays historical scenes not possible to capture.
D) There are no humans present in the video.

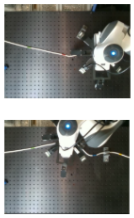
Answer

B

Robotics Evaluation

Question

Does the machine have the capability to bend the wire?
A) No, the machine does not have the ability to bend the wire.
B) Yes, through a rotating mechanism on the machine.
C) Yes, but only by utilizing an additional machine attachment.
D) The machine can only cut the wire, it does not have the capability to bend it.



Answer

B

Multi-Object Interaction

Question

Did the robot and human hands ultimately make contact?
A) Yes, the robot and human hands did make contact.
B) No, the robot and human hands did not make contact.
C) Sure, there was a contact between them.
D) Absolutely, they touched each other.

Answer

B

Figure 10: Dataset cases 2.