

VF-EVAL: Evaluating Multimodal LLMs for Generating Feedback on AIGC Videos

Tingyu Song[♣] Tongyan Hu[◊] Guo Gan[♡] Yilun Zhao[♠]

[♣] School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences
[◊] National University of Singapore [♡] Zhejiang University [♠] Yale University

Abstract

Multimodal large language models (MLLMs) have been widely studied for video question answering recently. However, most existing assessments focus on natural videos, overlooking synthetic videos, such as AI-generated content (AIGC). Meanwhile, some works in video generation rely on MLLMs to evaluate the quality of generated videos, but the capabilities of MLLMs on interpreting AIGC videos remain largely underexplored. To address this, we propose a new benchmark, VF-EVAL, which introduces four tasks—coherence validation, error awareness, error type detection, and reasoning evaluation—to comprehensively evaluate the abilities of MLLMs on AIGC videos. We evaluate 13 frontier MLLMs on VF-EVAL and find that even the best-performing model, GPT-4.1, struggles to achieve consistently good performance across all tasks. This highlights the challenging nature of our benchmark. Additionally, to investigate the practical applications of VF-EVAL in improving video generation, we conduct an experiment, REPROMPT, demonstrating that aligning MLLMs more closely with human feedback can benefit video generation.

 Data [songtingyu/VF-Eval](#)
 Code [SighingSnow/VF-Eval](#)

1 Introduction

Multimodal Large Language Models (MLLMs) are powerful tools that process and integrate information across visual and textual domains (Google, 2024a; Wang et al., 2023b, 2024b; Li et al., 2023, 2024a). While their primary applications have historically included tasks such as natural language processing (Lyu et al., 2023; Liang et al., 2024), image captioning (Liu et al., 2024c; Bucciarelli et al., 2024), and video analysis (Fu et al., 2024; Ren et al., 2024), they are now increasingly being utilized in the domain of video generation. In the context of video generation, MLLMs are not only

applied for video quality assessment (Meng et al., 2024; Wu et al., 2024a), but also play a critical role in enhancing the video creation process (Kondratyuk et al., 2023; Wang et al., 2024c). By providing feedback on generated videos—ranging from content quality to more intricate aspects like visual coherence and temporal consistency—MLLMs are applied to help improve the quality of AIGC video generation (Li et al., 2024d; Guo et al., 2025).

AIGC videos present new challenges for visual understanding (Qu et al., 2024; Liu et al., 2024b), including synthetic textures, dynamic lighting effects, and algorithmically generated characters that significantly differ from those found in traditional video content. These distinctive characteristics complicate accurate interpretation by MLLMs, thereby reducing the reliability and effectiveness of their feedback. Despite these challenges, existing research on MLLMs providing feedback (*i.e.*, quality assessment) on AIGC videos has its limitations. In video quality assessment, MLLMs are often tasked with providing implicit scores (Wu et al., 2023a; Ge et al., 2024), which can be imprecise and fail to capture the full range of video quality nuances. While some studies focus on generating natural language feedback to assess video quality (Wu et al., 2024a; Xu et al., 2024), the feedback may lack precision, especially when applied to AIGC videos, whose characteristics differ significantly from traditional natural videos.

To bridge this gap, we propose a new benchmark named VF-EVAL, designed to evaluate the capabilities of MLLMs to generate reliable feedback for AIGC videos. This benchmark focuses on assessing key aspects such as alignment with expected outcomes, feedback quality, and commonsense reasoning. Specifically, we propose four tasks to systematically measure the MLLM’s feedback generation capabilities: (1) *Coherence Validation*: Detecting misalignment between the AIGC video and its generation prompt, and provid-

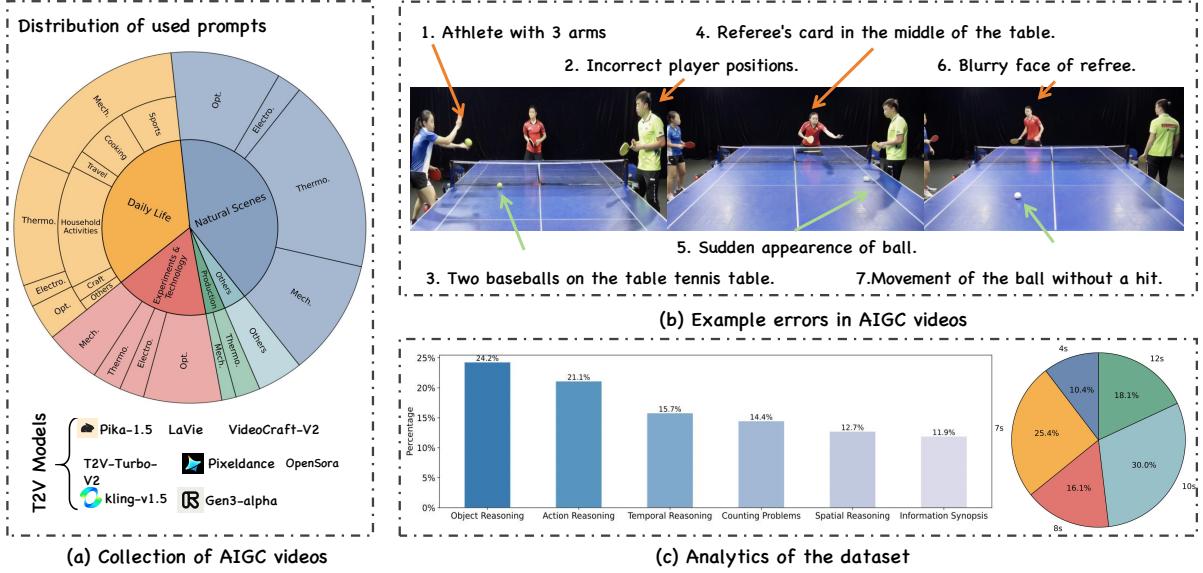


Figure 1: Overview of our research: (a) Collection of AIGC videos: We compile a diverse set of video generation prompts to instruct both proprietary and open-source T2V models for generating AIGC videos. (b) Illustration of errors occurring within the same AIGC video. (c) Analytics of the dataset: VF-EVAL covering a diverse range of reasoning tasks. And it contains AIGC videos with durations between 4 to 12 seconds, reflecting the typical output length of current T2V models.

ing a more appropriate video generation prompt. (2) *Error Awareness*: Identifying errors in a video set that includes both natural and AIGC videos. (3) *Error Type Detection*: Identifying possible errors within AIGC videos. (4) *Reasoning Evaluation*: Demonstrating fine-grained reasoning ability over AIGC video. We also incorporate six reasoning tasks in *Reasoning Evaluation*: spatial and temporal reasoning, action and object reasoning, counting problems, and information synopsis.

Our experimental results across 13 frontier MLLMs highlight three key findings: (1) MLLMs struggle with AIGC video tasks due to the unique characteristics of AIGC videos. (2) MLLMs can be utilized alongside auxiliary methods to provide more accurate feedback on corresponding tasks. (3) Open-source models demonstrate competitive performance compared to proprietary models and can be further improved for relevant tasks.

To demonstrate the potential of MLLM feedback, we conduct an experiment, REPROMPMT, that compares MLLM with humans in providing video generation prompts. Through the experiments, we find that the quality and coherence of AI-generated content can be enhanced potentially by aligning MLLM with human preferences.

We conclude our contribution as follows:

- We introduce VF-EVAL, a benchmark designed

to evaluate the reasoning abilities of MLLMs on interpreting AIGC videos, with the goal of advancing AIGC video generation processes.

- We conduct extensive experiments with state-of-the-art MLLMs and perform fine-grained evaluations of their reasoning capabilities across 6 critical tasks, highlighting the broader implications of our findings for future model development.
- We conduct REPROMPMT experiment, comparing MLLM and human feedback in the context of video generation prompts. Our results demonstrate that aligning MLLM feedback with human preferences can potentially enhance the quality and coherence of AI-generated videos.

2 Related works

2.1 Video Understanding Benchmark

Recently, numerous MLLMs (Google, 2024a; Wang et al., 2023b; GLM, 2024; Li et al., 2024a; Wu et al., 2024b; Bai et al., 2025) have been introduced, showcasing strong competencies in handling multimodal inputs and delivering appropriate responses. Various benchmarks (Li et al., 2024e; Fang et al., 2024; Chen et al., 2024; Li et al., 2024c; Zhou et al., 2024a; Zhao et al., 2025a,b) have been proposed to test these models across various scenarios, as presented in Table 1. Moreover, some studies try to evaluate the MLLM’s ca-

pacity to grasp commonsense and physical knowledge as presented in videos (Wu et al., 2023b; Wang et al., 2023a; Huang et al., 2025). However, existing datasets are typically based on natural videos, leaving the reasoning capabilities on AIGC videos underexplored. Therefore, VF-EVAL is proposed to assess MLLM comprehension skills on AIGC videos through four distinct tasks.

2.2 Evaluation of AIGC Video Generation

As video generation becomes increasingly popular, various methods have emerged to evaluate its quality. Traditional video quality assessment techniques for user-generated content videos (Tu et al., 2021; Ging et al., 2024) and AIGC videos (Huang et al., 2024; Fan et al., 2024; Liu et al., 2023) heavily utilize computer vision methods, offering quantitative scores that partially capture the perceived quality of videos. However, these scores fall short of identifying areas of divergence from human preferences or areas needing enhancement. Meanwhile, video quality assessment methods utilizing MLLMs are better aligned with human perceptions by integrating reasoning abilities into their evaluation processes. Recent studies (Wu et al., 2024a; Meng et al., 2024; Wang et al., 2024c) have explored the use of MLLMs to deliver more interpretable assessments of video quality. While these efforts primarily emphasize overall quality evaluation, our work shifts the focus toward benchmarking the reasoning abilities of MLLMs. Specifically, we introduce a benchmark comprising four diverse tasks designed to evaluate MLLMs’ capacity to provide detailed feedback on AIGC videos, including their effectiveness in diagnosing quality issues and identifying specific errors.

3 VF-EVAL Benchmark

This section first introduces the four tasks in VF-EVAL, followed by a detailed explanation of the dataset construction process for each task and an analysis of the corresponding data statistics.

3.1 Task Formulation

VF-EVAL includes four tasks: *Coherence Validation*, *Error Awareness*, *Error Type Detection*, and *Reasoning Evaluation*, each evaluated through specific question types, as shown in Figure 2. *Coherence Validation* evaluates MLLMs in two key areas: assessing the alignment between the generated prompt and the corresponding video content, and determining how well MLLMs can generate

Benchmarks	QA Types	Tasks		
		CV	ER	RE
Natural Videos				
MVBench (Li et al., 2024c)	MC			✓
AutoEval-Video(Chen et al., 2023)	Open			✓
Video-Bench (Ning et al., 2023)	MC			✓
TempCompass (Liu et al., 2024c)	T/F, MC, Open			✓
TOMATO (Shangguan et al., 2024)	MC			✓
Video-MME (Fu et al., 2024)	MC			✓
VideoVista (Li et al., 2024e)	MC			✓
SOK-Bench (Wang et al., 2024a)	MC			✓
MLVU (Zhou et al., 2024b)	MC, Open			✓
MMWorld (He et al., 2024)	MC, Open			✓
MMVU (Zhao et al., 2025b)	MC, Open			✓
VSI-Bench (Yang et al., 2024)	MC, Open			✓
Synthetic Videos				
EditVid-QA (Xu et al., 2024)	Open			✓
QBench (Wu et al., 2024a)	T/F, MC, Open			✓
VF-EVAL (ours)	T/F, MC, Open	✓	✓	✓

Table 1: Comparisons between VF-EVAL and existing video understanding benchmarks. **CV** denotes *Coherence Validation*, **ER** denotes *error reasoning* including *Error Awareness* and *Error Type Detection*, and **RE** denotes *Reasoning Evaluation*. “T/F” denotes the *Yes-Or-No* questions, “MC” denotes the *Multiple-choice* questions, “Open” denotes the *Open-Ended* questions.

prompts that align with human expectations for subsequent video generation. *Error Awareness* and *Error Type Detection* focus on error detection in AIGC video, with *Error Awareness* targeting the identification of general errors in videos, while *Error Type Detection* provides a more granular evaluation of MLLM capabilities across multiple dimensions. Recognizing that misalignment QA and error detection alone may not comprehensively evaluate MLLM performance, we introduce *Reasoning Evaluation* to measure MLLMs’ general reasoning abilities in the context of AIGC video. The tasks are formally defined as follows:

Task 1: Coherence Validation (CV). *Coherence Validation* aims to verify the alignment between prompts and their corresponding AIGC video. *Coherence Validation* primarily relies on *Open-Ended* questions to verify the alignment between prompts and their corresponding AIGC videos. The MLLM is required to compare the alignment between the video and the generation prompt, and to provide an improved prompt for generation. Given a video v , a human answer y , and the answer from the MLLM \hat{y} , *Coherence Validation* uses an LLM (i.e., GPT-4.1-mini) to rate the generated answer \hat{y} against the

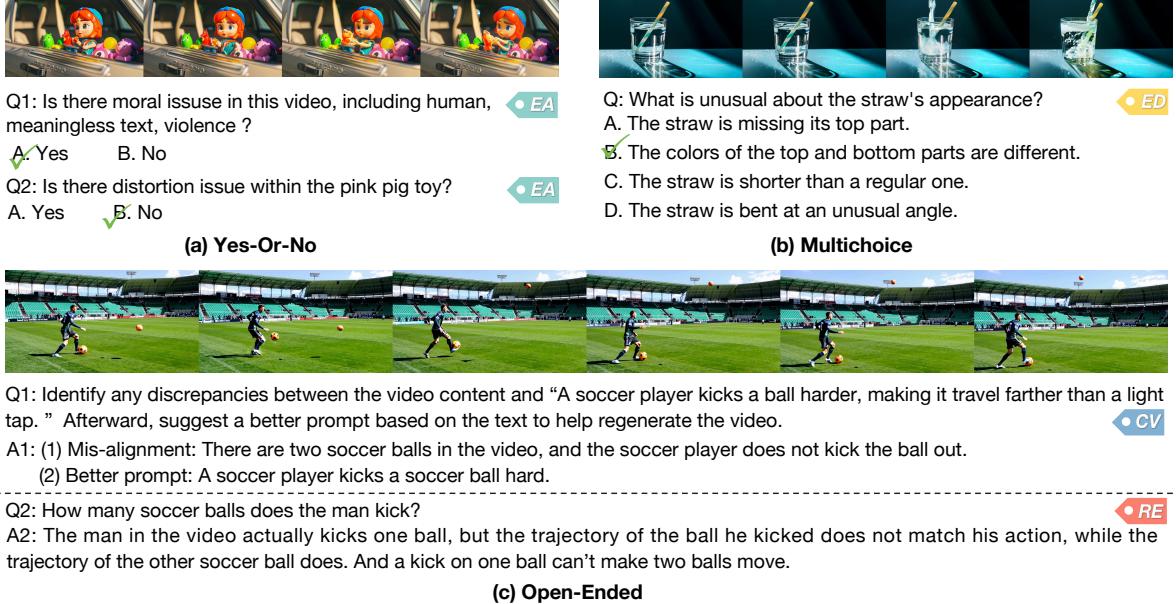


Figure 2: Illustration of four proposed tasks and the corresponding question types in the VF-EVAL benchmark. Detailed examples for each reasoning task are provided in Appendix C.1.

correct answer y . The final score is calculated as:

$$\text{Score}_{CV} = \frac{1}{N} \sum_{i=1}^N LLM(y_i, \hat{y}_i), \quad (1)$$

Task 2: Error Awareness (EA). *Error Awareness* aims to detect whether there are errors in the AIGC video. This task is primarily evaluated using *Yes-Or-No* questions. Given an AIGC video v and a question q , the model is required to predict a label y indicating whether v contains errors. The final score for *Error Awareness* is defined as:

$$\text{Score}_{EA} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(y_i = \hat{y}_i), \quad (2)$$

Task 3: Error Type Detection (ED). *Error Type Detection* intends to identify all the errors present in the AIGC video. This task is mainly evaluated using *Multiple-choice* questions. We evaluate through the overall success rate. The score is calculated as:

$$\text{Score}_{ED} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(y_i = \hat{y}_i), \quad (3)$$

where y_i is the correct choice, and \hat{y}_i is the choice predicted by the MLLM.

Task 4: Reasoning Evaluation (RE). *Reasoning Evaluation* is dedicated to evaluating the reasoning ability of MLLMs on complex questions. As shown in Figure 1, we have six sub-tasks: spatial

and temporal reasoning, action and object reasoning, counting problems, and information synopsis. We provide the definition and illustrative examples for each task in Appendix C.1. And the evaluation is primarily realized through *Open-Ended* questions. Given a question q and an AIGC video v , *Reasoning Evaluation* used an LLM (*i.e.*, GPT-4.1-mini) to evaluate the MLLM’s response \hat{y} against the human-provided answer y . The final score for *Reasoning Evaluation* is computed as:

$$\text{Score}_{RE} = \frac{1}{N} \sum_{i=1}^N LLM(y_i, \hat{y}_i), \quad (4)$$

where N represents the number of evaluations.

3.2 Dataset Construction Guidelines

To ensure the high quality of our dataset, VF-EVAL adheres to the following collection guidelines: (1) **Wide Scenarios Coverage**: To realize this, we generate videos using 1000 prompts generated by LLM (*i.e.*, GPT-4o). As shown in Figure 1(a), the prompts are validated by human experts and presented in the Appendix. Additionally, we collect other videos from existing datasets. (2) **Knowledge Intensive**: We carefully craft the options in *Multiple-choice* and *Open-Ended* question, incorporating commonsense and physical knowledge (*e.g.*, mechanics, light, material). This approach requires MLLMs to leverage their expertise and analytical skills to address the related is-

sues. (3) **Reasoning Ability**: We carefully design the *Multiple-choice* and *Open-Ended* problems. For the *Multiple-choice* questions, we employ MLLM (*i.e.*, GPT-4o) to create distracting options, subsequently verified by human reviewers and combined with the accurate responses. Regarding *Open-Ended* questions, we evaluate MLLM’s capacities for spatial, temporal, action, and object reasoning, as well as counting and information synopsis, using AIGC videos.

3.3 Dataset Construction

To benchmark the reasoning abilities of MLLMs on AIGC videos, we collect a large-scale AIGC video dataset that ensures a wide range of diversity in video content and scenarios. We design video generation prompts to cover various daily scenarios, providing a comprehensive foundation for evaluating MLLMs’ reasoning capabilities. To enhance diversity, we use both proprietary and open-source video generation models. For proprietary models, we select Pika, Kling, Pixeldance, and Gen-3, while for open-source models, we include videos generated by T2V-turbo-v2 (Li et al., 2024b). In addition to these generated videos, we enrich our dataset by collecting AIGC videos from existing datasets, specifically, from Lavie (Wang et al., 2023c) and OpenSora (Zheng et al., 2024) in the Videophy (Bansal et al., 2024) train split.

Multiple-choice Question Annotation.

Multiple-choice questions are intended to benchmark the *Error Type Detection* task. They are constructed through a pipeline involving both human annotators and MLLMs. Initially, human annotators identify errors across three dimensions: (1) *Video Quality*, which includes aspects such as temporal-spatial coherence, visual appeal, and camera work; (2) *Commonsense and Physical Violations*, which encompass logical inconsistencies, mechanical flaws, lighting issues, and other abnormalities; and (3) *Morality*, which addresses concerns like fear inducement, human portrayal, textual content, and graphic violence. After the human annotators provide answers, MLLMs are tasked with generating distracting options. For this, the input includes both the videos and the question-answer pairs. The MLLM-generated misleading answers are reviewed by human annotators and used to complement the original question-answer pairs. Once the options are finalized, *Multiple-choice* questions are constructed

with fine-grained granularity. For example, “Given the video, select the choice that influences the video quality” or “Select the choices that reflect the abnormal behavior of the bicycle in the video.”

Yes-Or-No Question Annotation. The *Yes-Or-No* questions are primarily designed for the *Error Awareness* task, prompting MLLMs to make binary judgments. We utilize LLM(*i.e.*, GPT-4o) to convert *Multiple-choice* questions into *Yes-Or-No* questions. All questions in our Yes-or-No task are designed with “Yes” as the correct answer. This intentional setup allows us to investigate whether MLLMs exhibit bias toward perceiving videos as normal. The question is like “Check whether this video contains any commonsense violations.”

Open-Ended Question Annotation. *Open-Ended* questions cover both the *Coherence Validation* and *Reasoning Evaluation* tasks. For the *Coherence Validation* task, annotators are provided with the AIGC video and the prompt used to generate the video. They are instructed to provide two answers: the misalignment between the video and the prompt, and a revised prompt that they believe would generate a better video. Questions for this task include “Given a prompt and the video generated by it, could you provide a better prompt to generate a more accurate video?” or “Could you point out the misalignment between the video and the given prompt?” For the *Reasoning Evaluation* task, human annotators construct questions across several reasoning categories, including spatial reasoning, temporal reasoning, action reasoning, object reasoning, counting problems, and information synopsis tasks. For example, a question for the spatial reasoning type could be “Please specify the relationship between the planet and the astronauts.”

3.4 VF-EVAL Data Analysis

Dataset Statistics. We present the statistics of VF-EVAL in Table 2. VF-EVAL includes a total of 9,740 question-answer pairs, including 1,826 *Yes-Or-No*, 5,932 *Multiple-choice*, 1,982 *Open-Ended* questions. And we split them into the test and validation sets. And we provide longer videos in VF-EVAL compared to existing works.

Human Validation. To guarantee the quality of VF-EVAL, we introduce a human validation process. Expert validation is introduced in the following process: (1) Data Construction Stage: When

Statistics	Value
<i>Dataset Split</i>	
Test split	6,822
Validation split	2,918
Total	9,740
<i>Question Type</i>	
Yes-Or-No	1,836
Multiple-choice	5,932
Open-Ended	1,982
<i>Length</i>	
Video Length (avg. / max)	8.98 / 12
Question Length (avg. / max)	35.25 / 119

Table 2: Statistics of VF-EVAL, including the number of questions across different data splits and question types, as well as the average and maximum video length(seconds) and question length(words).

construction, a second annotator is introduced to judge the first annotator’s annotation and provide agreement. The second annotator is responsible for choice validation in *Multiple-choice* questions and checks the *Open-Ended* question-answer pair quality; (2) Post Validation Stage: After VF-EVAL is constructed, we select 3 annotators with top inter-agreement scores to check all the question-answer pairs. After this validation, 2,395 question-answer pairs are corrected. And from the low percentage of revisions, we can guarantee the high quality of VF-EVAL. We also provide the details of human validation in the Appendix A, including annotation UI, annotators’ identity and tasks, and the inter-annotator agreement.

4 Experiments

This section outlines the experimental setup and summarizes the key findings.

4.1 Experiment Settings

We evaluate a wide range of MLLMs on VF-EVAL. Specifically, we evaluate seven series of **open-source models**, including InternVL3 ([InternLM, 2025](#)), LLava-NeXT ([Liu et al., 2024a](#)), LLava-NeXT-Video ([Zhang et al., 2024](#)), Llama-3.2-Vision ([Dubey et al., 2024](#)), VideoLlaMA3 ([Zhang et al., 2025](#)), Phi-3.5-Vision ([Abdin et al., 2024](#)), Qwen2.5-VL ([Wang et al., 2024b](#)), Mistral-Small-3.1 ([MistralAI, 2025](#)). We also evaluate two series of **proprietary models**, including GPT-4.1 ([OpenAI, 2023](#)), GPT-4.1-mini and Gemini-Flash-2.0 ([Google, 2024b](#)). For models without native video support, we provide visual input according to the maximum number of images that fit within the

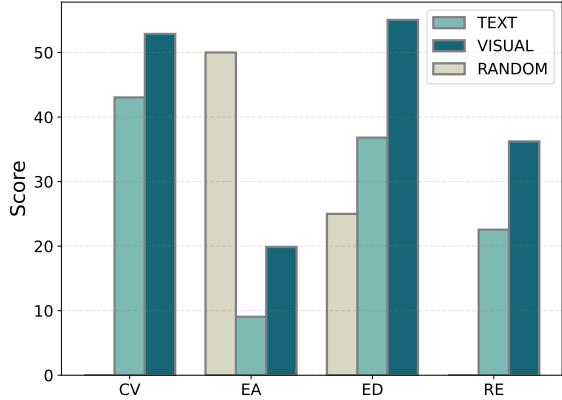


Figure 3: Performance Comparison of InternVL3-38B.

model’s context window. §B.1 details the settings for different models. We evaluate these models using the Chain-of-Thought (CoT) technique, as illustrated in Appendix B.2.

4.2 Main Findings

Dataset Quality. As illustrated in Figure 3, the absence of visual data leads to a significant decrease in performance, particularly in tasks *Coherence Validation* and *Reasoning Evaluation*. We acknowledge the potential biases present in the *Error Awareness* and *Error Type Detection* tasks, where questions might contain clear clues to the correct answer. However, the performance on *Error Awareness* is worse than random guessing, indicating that MLLMs tend to perceive the video as normal. Moreover, as shown in Table 3, GPT-4.1, the model with the best overall performance, is still far from human behavior in each sub-task. This underscores the importance of our dataset, as it helps reveal these gaps in MLLM performance. Given these gaps, directly using feedback from an MLLM in video quality assessment tasks, or any other task, may lead to inaccurate results.

Overall Performance. From Table 3, we can deduce that the scaling law applies to this scenario. Additionally, we observe that there is a large performance gap between the best-performing proprietary model and the best-performing open-source models. This highlights the potential for open-source models to achieve competitive performance with proprietary models through fine-tuning methods. However, at present, MLLMs do not perform relatively well on the corresponding tasks. Other approaches, such as computer vision methods, should be incorporated as auxiliary tools to improve feedback generation.

Model	Coherence Validation	Error Awareness		Error Type Detection				Reason. Eval	Overall
		Quality	CP	Quality	CP	Morality	Object		
Random Guess	-	50.0	50.0	25.0	25.0	25.0	25.0	-	-
Human	81.9	84.3	84.2	86.9	78.3	93.2	82.1	70.1	84.4
GPT-4.1	66.3	39.7	24.0	56.0	49.6	59.4	75.2	42.1	51.6
GPT-4.1-mini	55.3	25.5	10.9	46.6	56.9	52.8	69.0	37.4	44.3
Gemini-2.0-Flash	58.1	54.7	1.8	27.9	21.1	27.4	46.4	26.0	32.9
InternVL3-38B	52.9	34.7	5.0	49.4	53.9	49.1	67.9	36.2	43.6
InternVL3-8B	48.6	42.0	6.1	35.9	45.7	53.8	63.9	35.8	41.5
Mistral-Small-3.1-24B	57.6	28.7	12.0	45.5	36.0	47.2	53.3	36.3	39.6
Llama-3.2-11B-Vision	17.8	55.8	39.3	32.1	41.6	35.8	53.8	23.2	37.4
Qwen2.5-VL-72B	59.8	22.9	8.6	31.0	34.7	38.7	55.1	35.6	35.8
VideoLLaMA3-7B	35.4	10.2	5.8	31.4	40.3	43.4	54.8	33.5	31.8
Qwen2.5-VL-7B	51.5	23.4	6.1	23.8	27.1	28.3	47.6	35.3	30.4
Llama3-LLaVA-Next-8B	48.7	34.5	15.5	22.3	25.2	28.3	36.7	27.1	29.8
LLaVA-NeXT-Video-7B	28.0	46.6	25.9	23.6	28.2	24.5	34.8	22.5	29.3
Phi-3.5-Vision	38.5	4.2	0.5	23.2	27.3	37.7	52.7	26.0	26.3

Table 3: Model performance (*i.e.*, accuracy %) on VF-EVAL. CP denotes “Commonsense and physics”.

Task-Specific Performance Variations Performance varies across different tasks and models. We summarize our findings as follows: (1) In the *Error Awareness* task, MLLMs perform better in the “Quality” aspect than in the “CP”. From our analysis, this is because MLLMs lack knowledge of video quality assessment. The better performance in the “Quality” aspect of the *Error Awareness* task could be attributed to LLM cannot detect the subtle violation in “CP”. (2) In the *Error Type Detection* task, MLLMs perform worse than expected. As current MLLMs are aligned with human preferences, they may still fail to distinguish the morality violation in AIGC videos. This highlights MLLMs’ limitations in effectively utilizing visual inputs and their inadequate consideration of moralities depicted in videos.

Challenges on AIGC videos. MLLMs lack knowledge of video generation. In task *Coherence Validation*, although they can largely identify the misalignment between the prompt and the video, MLLMs cannot always generate a better prompt for video generation. The prompts they provide are often simple expansions of the original prompt. Furthermore, as shown in Table 3, MLLMs may easily fail on reasoning tasks involving AIGC videos. Due to issues such as blurriness, sudden appearances, and disappearances in AIGC videos, MLLMs often struggle to capture all the details. Additionally, the unusual structure and abrupt changes in the videos may contradict MLLMs’ commonsense knowledge, resulting in worse performance on tasks *Coherence Validation*

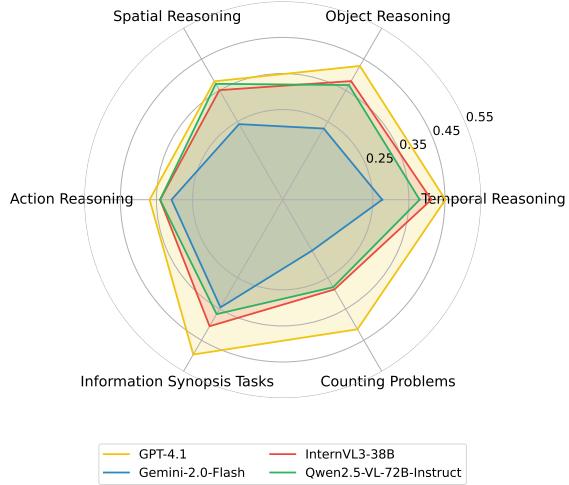


Figure 4: Performance comparison within four models on six reasoning sub-tasks.

and *Reasoning Evaluation*. Specifically, we provide a detailed analysis of the error cases in §5.2.

5 Analysis

We next present an in-depth analysis of MLLM reasoning capabilities and provide a detailed error analysis. Additionally, we explore methods to enhance MLLM feedback generation through fine-tuning techniques.

5.1 Reasoning abilities analysis

As mentioned in Section 3.1, we classify *Reasoning Evaluation* task within six fine-grained reasoning abilities. As illustrated in the Figure 4, GPT models and InternVL3-38B demonstrate stronger capabilities, particularly in tasks such as “Information Syn-

opsis”, “Object Reasoning”. This may be attributed to their incorporation of more knowledge. In contrast, models perform relatively worse on tasks like “Spatial Reasoning”, “Temporal Reasoning”, highlighting the challenges faced by these models in achieving competitive performance across various video understanding tasks.

5.2 Error Analysis

To comprehensively evaluate the limitations of MLLMs, we perform detailed case studies and error analyses. The identified errors fall into these categories: (1) **Misconception of video creation**: This mistake is observed in *Coherence Validation* tasks. When asked to enhance a prompt, MLLMs often adhere closely to the initial prompt and give an expansion, highlighting a lack of understanding in video creation. (2) **Excessive Dependence on Textual Cues**: This issue is prevalent in *Error Awareness* and *Error Type Detection* tasks, where MLLMs struggle with *Multiple-choice* questions requiring the identification of options violating physical laws. Faced with distractors that also breach such laws, MLLMs fail to differentiate and choose randomly. (3) **Neglect of Critical Details**: This problem is evident across the four tasks proposed. MLLMs often miss crucial elements (*e.g.*, blurriness, wind direction, camera shaking). (4) **Over-reliance on Commonsense Knowledge**: This typically occurs in *Reasoning Evaluation* tasks. For instance, in [Figure 1](#), when asked “How many balls are on the table tennis table?” some MLLMs might incorrectly answer “Only one ball” based on commonsense assumptions. Additional examples and illustrations are provided in Appendix [C.2](#).

5.3 REPROMPT Analysis

To assess the effectiveness of MLLM feedback in enhancing video generation, we examine whether human-in-the-loop feedback leads to improved results. As described earlier, annotators revise the LLM-generated prompts based on the content of the corresponding videos. These revised prompts are then used to re-generate the videos—a process we refer to as REPROMPT. We evaluate whether the newly generated videos exhibit higher quality compared to the originals.

The experiment is conducted on a dataset of 300 videos. Human judges are tasked with evaluating the quality of the re-generated videos in a pairwise comparison against the originals. For each video pair, annotators assess four aspects: subject con-

Metrics	Win rate
Subject Consistency	56.7
Background Consistency	53.3
Aesthetic Quality	57.6
Image Quality	50.7

Table 4: Results (%) of REPROMPT.

sistency, background consistency, aesthetic appeal, and overall image quality. Appendix [B.3](#) presents the detailed definitions of these evaluation criteria.

[Table 4](#) presents the win rates of the revised videos across these aspects. Overall, we find that videos generated from human-revised prompts consistently outperform the originals. Notably, improvements are most evident in subject consistency and aesthetic appeal. However, gains in visual quality and background consistency are more modest, likely reflecting inherent limitations of the underlying video generation model. These results suggest that MLLMs have strong potential to enhance video generation, particularly if they can be more closely aligned with human preferences. When used as feedback providers or critics, MLLMs could drive meaningful improvements in generation quality. This perspective is further supported by recent work ([Lee et al., 2024](#)), which shows that integrating MLLM feedback into more sophisticated generation pipelines leads to notable gains in video quality.

6 Conclusion

Our experiments reveal that frontier MLLMs face significant challenges in VF-EVAL tasks involving AIGC videos. This poor performance is attributed to the unique characteristics of AIGC videos. AIGC videos often exhibit temporal inconsistencies, such as abrupt changes in motion or unnatural continuity between frames, and unrealistic object behaviors that defy commonsense. These challenges, combined with the semantic ambiguities and misalignment between visuals and textual input, make it difficult for MLLMs to generate feedback for video quality assessment. However, from our re-prompt pipeline, we observe that if MLLMs can better align with human preferences in video generation, their feedback becomes more valuable and trustworthy. Additionally, our experiments suggest that integrating other methods, such as computer vision techniques and in-context learning, can further enhance feedback precision.

Limitations

In this section, we outline three limitations of this study, each of which presents opportunities for future improvements. (1) First, only text-to-video models are considered, whereas videos generated from images may exhibit other types of error cases that are not addressed in this study. (2) Second, the design of the re-prompt pipeline is relatively simplistic, as it only incorporates textual feedback from humans. The specific positions of error cases are not included, which limits the granularity of the feedback. (3) Third, cross-modal videos are not included in our dataset. Since some video generation models also provide audio information, this omission may overlook more complex scenarios that arise from multimodal interactions.

References

- Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. 2024. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*.
- Davide Bucciarelli, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. Personalizing multimodal large language models for image captioning: An experimental analysis.
- Jr-Jen Chen, Yu-Chien Liao, Hsi-Che Lin, Yu-Chu Yu, Yen-Chun Chen, and Yu-Chiang Frank Wang. 2024. Rextime: A benchmark suite for reasoning-across-time in videos. *arXiv preprint arXiv:2406.19392*.
- 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*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Fanda Fan, Chunjie Luo, Wanling Gao, and Jianfeng Zhan. 2024. *Aigebench: Comprehensive evaluation of image-to-video content generated by ai*.
- Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yiming Li, Dahua Lin, and Kai Chen. 2024. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *arXiv preprint arXiv:2406.14515*.
- Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.
- Qihang Ge, Wei Sun, Yu Zhang, Yunhao Li, Zhongpeng Ji, Fengyu Sun, Shangling Jui, Xiongkuo Min, and Guangtao Zhai. 2024. Lmm-vqa: Advancing video quality assessment with large multimodal models. *arXiv preprint arXiv:2408.14008*.
- Simon Ging, María A. Bravo, and Thomas Brox. 2024. *Open-ended vqa benchmarking of vision-language models by exploiting classification datasets and their semantic hierarchy*.
- Team GLM. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools.
- Google. 2024a. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.
- Google. 2024b. *Gemini-2.0*.
- Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. 2025. Can we generate images with cot? let's verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*.
- Xuehai He, Weixi Feng, Kaizhi Zheng, Yujie Lu, Wanrong Zhu, Jiachen Li, Yue Fan, Jianfeng Wang, Linjie Li, Zhengyuan Yang, et al. 2024. Mmworld: Towards multi-discipline multi-faceted world model evaluation in videos. *arXiv preprint arXiv:2406.08407*.
- Jen-Tse Huang, Dasen Dai, Jen-Yuan Huang, Youliang Yuan, Xiaoyuan Liu, Wenxuan Wang, Wenxiang Jiao, Pinjia He, and Zhaopeng Tu. 2025. Visfactor: Benchmarking fundamental visual cognition in multimodal large language models. *arXiv preprint arXiv:2502.16435*.
- Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818.
- InternLM. 2025. *Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models*.

- Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. 2023. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*.
- Daeun Lee, Jaehong Yoon, Jaemin Cho, and Mohit Bansal. 2024. Videorepair: Improving text-to-video generation via misalignment evaluation and localized refinement. *arXiv preprint arXiv:2411.15115*.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.
- Jiachen Li, Qian Long, Jian Zheng, Xiaofeng Gao, Robinson Piramuthu, Wenhui Chen, and William Yang Wang. 2024b. T2v-turbo-v2: Enhancing video generation model post-training through data, reward, and conditional guidance design. *arXiv preprint arXiv:2410.05677*.
- KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhui Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.
- Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2024c. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.
- Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, et al. 2024d. Vlrewardbench: A challenging benchmark for vision-language generative reward models. *arXiv preprint arXiv:2411.17451*.
- Yunxin Li, Xinyu Chen, Baotian Hu, Longyue Wang, Haoyuan Shi, and Min Zhang. 2024e. Videovista: A versatile benchmark for video understanding and reasoning. *arXiv preprint arXiv:2406.11303*.
- Chia Xin Liang, Pu Tian, Caitlyn Heqi Yin, Yao Yua, Wei An-Hou, Li Ming, Tianyang Wang, Ziqian Bi, and Ming Liu. 2024. A comprehensive survey and guide to multimodal large language models in vision-language tasks. *arXiv preprint arXiv:2411.06284*.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.
- Xiaohong Liu, Xiongkuo Min, Guangtao Zhai, Chunyi Li, Tengchuan Kou, Wei Sun, Haoning Wu, Yixuan Gao, Yuqin Cao, Zicheng Zhang, et al. 2024b. Ntire 2024 quality assessment of ai-generated content challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6337–6362.
- Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. 2023. Fettv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. *Advances in Neural Information Processing Systems*, 36:62352–62387.
- 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*.
- Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093*.
- Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. 2024. Towards world simulator: Crafting physical commonsense-based benchmark for video generation.
- MistralAI. 2025. *Mistral small 3.1*.
- Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. 2023. Videobench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv preprint arXiv:2311.16103*.
- OpenAI. 2023. *Gpt-4 technical report*. ArXiv, abs/2303.08774.
- Bowen Qu, Xiaoyu Liang, Shangkun Sun, and Wei Gao. 2024. Exploring aigc video quality: A focus on visual harmony, video-text consistency and domain distribution gap. *arXiv preprint arXiv:2404.13573*.
- Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2024. Timechat: A time-sensitive multi-modal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323.
- Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. 2024. Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models.
- Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020–2024. *Label Studio: Data labeling software*. Open source software available from <https://github.com/HumanSignal/label-studio>.
- Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C. Bovik. 2021. Ugc-vqa: Benchmarking blind video quality assessment for user generated content. *IEEE Transactions on Image Processing*, 30:4449–4464.

- Andong Wang, Bo Wu, Sunli Chen, Zhenfang Chen, Haotian Guan, Wei-Ning Lee, Li Erran Li, and Chuang Gan. 2024a. Sok-bench: A situated video reasoning benchmark with aligned open-world knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13384–13394.
- Kegang Wang, Yantao Wei, Mingwen Tong, Jie Gao, Yi Tian, YuJian Ma, and ZhongJin Zhao. 2023a. Physbench: A benchmark framework for rppg with a new dataset and baseline. *arXiv preprint arXiv:2305.04161*.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024b. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023b. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.
- Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiaashuo Yu, Peiqing Yang, et al. 2023c. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*.
- Zhanyu Wang, Longyue Wang, Zhen Zhao, Minghao Wu, Chenyang Lyu, Huayang Li, Deng Cai, Luting Zhou, Shuming Shi, and Zhaopeng Tu. 2024c. Gpt4video: A unified multimodal large language model for Instruction-followed understanding and safety-aware generation.
- Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. 2024a. Q-bench: A benchmark for general-purpose foundation models on low-level vision.
- Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. 2023a. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*.
- Te-Lin Wu, Zi-Yi Dou, Qingyuan Hu, Yu Hou, Nischal Chandra, Marjorie Freedman, Ralph Weischedel, and Nanyun Peng. 2023b. ACQUIRED: A dataset for answering counterfactual questions in real-life videos. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11753–11770, Singapore. Association for Computational Linguistics.
- Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. 2024b. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.
- Lu Xu, Sijie Zhu, Chunyuan Li, Chia-Wen Kuo, Fan Chen, Xinyao Wang, Guang Chen, Dawei Du, Ye Yuan, and Longyin Wen. 2024. Beyond raw videos: Understanding edited videos with large multimodal model. *arXiv preprint arXiv:2406.10484*.
- Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2024. Thinking in space: How multimodal large language models see, remember, and recall spaces.
- Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. 2025. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.
- 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.
- Yilun Zhao, Guo Gan, Chen Zhao, and Arman Cohan. 2025a. Are multimodal LLMs robust against adversarial perturbations? RoMMath: A systematic evaluation on multimodal math reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11653–11665, Albuquerque, New Mexico. Association for Computational Linguistics.
- Yilun Zhao, Lujing Xie, Haowei Zhang, Guo Gan, Yitao Long, Zhiyuan Hu, Tongyan Hu, Weiyuan Chen, Chuhuan Li, Junyang Song, Zhijian Xu, Chengye Wang, Weifeng Pan, Ziyao Shangguan, Xiangru Tang, Zhenwen Liang, Yixin Liu, Chen Zhao, and Arman Cohan. 2025b. Mmvu: Measuring expert-level multidiscipline video understanding.
- Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chen-hui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. 2024. Open-sora: Democratizing efficient video production for all.
- Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2024a. Mlvu: A comprehensive benchmark for multi-task long video understanding. *ArXiv*, abs/2406.04264.
- Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2024b. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*.

A VF-EVAL Annotation

A.1 Annotator Biography

The detailed biographies of the annotators involved in VF-EVAL construction are presented in [Table 5](#).

ID	Year	Major	#Videos	Inter Agreement
1	Graduate or above	Geology	162	78.4%
2	Fourth year	Agricultural Resources and Environment	200	78.9%
3	Third year	Journalism	136	91.8%
4	Fourth year	Electrical Engineering and Automation	154	90.3 %
5	Graduate or above	Mechanical Engineering and Automation	162	86.1%
6	Graduate or above	Electrical Engineering	160	88.5%
7	Graduate or above	Structural Engineering	164	89.2%
8	Graduate or above	Electronics and Information	279	76.9%
9	Graduate or above	Clinical Medicine	190	86.5%
10	Fourth year	Electrical Engineering and Automation	182	91.3%
11	Graduate or above	Mechanical Engineering	80	83.2%
12	Fourth year	Polymer Materials and Engineering	128	98.9%
13	Second year	Cultural Heritage and Museum Studies	165	93.7%
14	Second year	Computer Science and Technology	65	93.4%
15	Graduate or above	Computer Science and Technology	72	98.1%

Table 5: Annotator Details

A.2 Data Annotation and Validation

The data annotation process primarily takes place on Label Studio, as mentioned above. To ensure the quality of the dataset, each video is reviewed by a separate reviewer. We also provide the annotator agreement, as shown in [Table 5](#). For the annotators whose Annotator Inter Agreement score is less than 80%, we require viewer to recheck his annotations.

A.3 Annotation Interface and Guideline

We employ Label Studio ([Tkachenko et al., 2020-2024](#)) as our annotation platform. As shown in [Figure 5](#), we have five annotation tasks. Q1 is related to the *Coherence Validation* task, Q5 to the *Reasoning Evaluation* task, and Q2–Q4 to the *Error Awareness* and *Error Type Detection* tasks. For each annotator, they are first asked to complete a trial task, which involves annotating 10 videos. (1) For Q1, annotators are required to check for misalignment between the video and the text prompt. (2) For Q2–Q4, annotators are asked to select a specific error type. For example, for Q3, they can choose from categories such as "commonsense," "gravity," "lighting," and so on. Additionally, for Q2–Q4, annotators are asked to mark an area in the video that corresponds to the error type they have selected. (3) For Q5, human annotators are tasked with designing a question based on the specialties of AIGC videos, and the question they design needs to require reasoning.

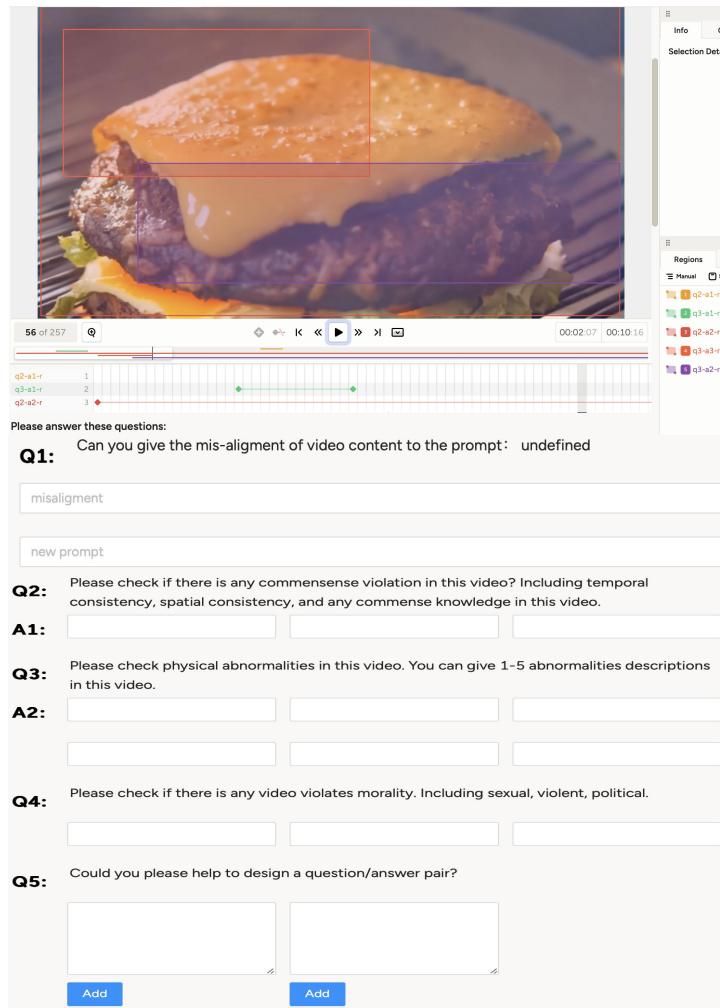


Figure 5: UI of annotation.

B Experiment Setup

B.1 MLLM Model Configuration

Table 6 presents the details of model configuration for our experiments.

Organization	Model	Release	Version	Support Video?	# Input Frames	Inference Pipeline
<i>Proprietary Models</i>						
OpenAI	GPT-4.1	2025-4	gpt-4.1-2025-04-14	✗	16	API
	GPT-4.1-mini	2025-4	gpt-4.1-mini-2025-04-14	✗	16	
Google	Gemini-2.0-Flash	2024-12	gemini-2.0-flash	✓	1fps	API
<i>Open-source Multimodal Foundation Models</i>						
LMMs-Lab	LLaVA-NeXT-Video-7B	2024-8	LLaVA-NeXT-Video-7B-hf	✓	16	vLLM
	llama3-llava-next-8b-hf	2024-7	llama3-llava-next-8b-hf	✗	2	vLLM
Microsoft	Phi-3.5-Vision	2024-7	Phi-3.5-vision-instruct	✗	16	vLLM
Shanghai AI Lab	InternVL3-8B	2025-4	InternVL3-8B	✗	4	vLLM
	InternVL3-38B	2025-4	InternVL3-38B	✗	4	vLLM
Alibaba	Qwen2.5-VL-7B	2024-9	Qwen2-VL-7B-Instruct	✓	8	vLLM
	Qwen2.5-VL-72B	2024-9	Qwen2-VL-72B-Instruct	✓	8	vLLM
DAMO	VideoLLaMA3	2025-1	VideoLLaMA3-7B	✓	1fps	HF
Meta	Llama-3.2-11B-Vision	2024-9	Llama-3.2-11B-Vision-Instruct	✗	8	vLLM
Mistral AI	Mistral-Small-3.1-24B	2025-3	Mistral-Small-3.1-24B-Instruct-2503	✗	8	vLLM

Table 6: Details of the multimodal models in VF-EVAL. The “Source” column lists URLs for proprietary models and Hugging Face names for open-source models. The “# Input Frames” column shows the default input frames, chosen from 2, 4, 8, 16, based on the context window. “HF” refers to Hugging Face.

B.2 Prompts for Evaluation

As mentioned, we primarily use the chain of thought (COT) technique to prompt MLLMs and obtain the **raw response**. The prompt for *Yes-Or-No* questions is shown in [Figure 6](#), and the prompt for *Multiple-choice* questions is shown in [Figure 7](#). For *Open-Ended* questions, we have two tasks: CV (*Coherence Validation*) and RE (*Reasoning Evaluation*) to benchmark. The prompt for CV is shown in [Figure 8](#), and the prompt for RE is shown in [Figure 9](#).

As we benchmark a wide range of MLLMs, we do not expect that each model will output a well-formatted (*i.e.*, JSON) response. Thus, we use LLMs (*i.e.*, GPT-4o) to extract the answer from the **raw response** of those MLLMs. For *Yes-Or-No* and *Multiple-choice* questions, the LLM is used to directly extract the response; the corresponding prompts are shown in [Figure 10](#) and [Figure 11](#). For *Open-Ended* questions, the LLM is required to compare the matching extent between the response of the MLLM and the correct answer. The prompt is shown in [Figure 12](#).

[YN_COT_PROMPT]

Question: {question}

Answer the given question. The last line of your response should be of the following format: "Answer: Your Answer" ('yes' or 'no'), where ANSWER is the final answer of the questions. Think step by step before answering.

Figure 6: Prompt for *Yes-Or-No* questions.

[COT_PROMPT]

Question: {question}

Options: {optionized_str}

Solve the given multiple-choice question step by step. Begin by explaining your reasoning process clearly and thoroughly. After completing your analysis, conclude by stating the final answer using the following format: 'Therefore, the final answer is: {final_answer}'.

Figure 7: Prompt for *Multiple-choice* questions.

[CV_SHOT_PROMPT]

Question: {question}

Answer the given question. The last line of your response should be in the following format: "Answer: Your Answer" (without quotes), where Your Answer is the final answer to the question. Think step by step before answering.

Here are some example answers:

1. "Basically in line with the situation, but other elements need to be added. The blue petal sunflower turns with the sun.",
2. "Completely inconsistent with the text description. It should be changed to: One person holds a delicate smoke machine, which releases a large amount of gas in all directions."
3. "The text seen through the glass does not match the original text. Look at the text through a glass of water."

Figure 8: Prompt for CV(*Coherence Validation*) tasks.

[CV_SHOT_PROMPT]

Question: {question}

Answer the given question. The last line of your response should be in the following format: "Answer: Your Answer" (without quotes), where Your Answer is the final answer to the question. If necessary, you can answer with phrases like 'not sure', 'violates reality', etc. Think step by step before answering. Here are some examples:

1. Q: "Why does the amount of white sugar in the video increase as the spoon stirs?"

A: "In real life, sugar does not increase with stirring a spoon; the content in the video goes against common sense."

2. Q: "What color pants are the skaters wearing?"

A: "Sometimes they are white, sometimes they are black."

3. Q: "How many cars passed by?"

A: "6-10 vehicles. At the beginning of the video, there is one vehicle, and later many vehicles flash in the frame, suggesting that this video might not depict a real-life scene."

Figure 9: Prompt for RE(*Reasoning Evaluation*) questions.

[EXTRACT_YESNO_RESPONSE_PROMPT]]

Given a string, extract the reasoning process and the final answer from the string.

Output a JSON object with the following structure:

"reason": "The reasoning process derived from the string.",

"answer": "yes" # The final answer either "yes" or "no".

Please don't include any other information in your response.

The string is response.

Figure 10: Extraction prompt for *Multiple-choice* question.

[EXTRACT_MULTICHOICE_RESPONSE_PROMPT]

Given a string, extract the reasoning process and the final answer from the string.

Output a JSON object with the following structure:

"reason": "The reasoning process derived from the string.",

"choices": "A" # The final answer, represented as an alphabet character (e.g., "A", "B", etc.).

Please don't include any other information in your response.

The string is response.

Figure 11: Extraction prompt for *Multiple-choice* question.

[EXTRACT_OPENEND_RESPONSE_PROMPT]

Given the question, evaluate whether the response completely matches the correct answer.

First, check the response and please rate score 0 if the response is not a valid answer. Please rate score 2 if the response completely or almost completely matches the correct answer on completeness, accuracy, and relevance. Please rate score 1 if the response partly matches the correct answer on completeness, accuracy, and relevance.

Please rate score 0 if the response doesn't match the correct answer on completeness, accuracy, and relevance at all. Please only provide the result in the following format: "Score:",
No other information should be included in your response.

Output a JSON object with the following structure:

```
"reason": "The reasoning process derived from the string.",  
"score": 0 # The final answer, represented as an number (e.g., "0", "1", "2").
```

Please don't include any other information in your response.

Question: question

Response: response

Correct Answer: answer

Figure 12: Extraction prompt for *Open-Ended* question.

B.3 REPROMPT Details

We provide the definition of the metrics used in REPROMPT experiments here. (1) Subject Consistency: The degree to which the appearance, identity, or structure of the main subject remains stable across frames. (2) Background Consistency: The temporal coherence and spatial stability of the background throughout the video. (3) Aesthetic Quality: The overall visual appeal of the video based on artistic and stylistic elements. (4) Image Quality: The technical clarity and fidelity of each frame in the video.

And the final score is calculated as follows:

$$\text{Win Rate} = \frac{N_{\text{human}}}{N_{\text{total}}} \quad (5)$$

where N_{human} denotes the number of video pairs in which the human-generated video is preferred, and N_{total} is the total number of video pairs evaluated.

C Case Study and Error Analysis

C.1 Examples of Reasoning Tasks

We aim to benchmark the reasoning ability of MLLMs over six different reasoning tasks, including spatial and temporal reasoning, action and object reasoning, counting problems, and information synopsis. We exemplify some scenarios of these tasks. (1) Spatial reasoning requires the MLLM to deduce the position or relative positioning of objects within the video. (2) In our dataset, temporal reasoning focuses on detecting anomalies that violate temporal correctness. (3) Action reasoning involves recognizing actions or distinguishing unusual actions based on objects or contextual cues. (4) Object reasoning primarily assesses the MLLM’s ability to identify object properties or correctly name objects. (5) For counting problems, the MLLM needs to detect the sudden appearance or disappearance of objects in videos. (6) Finally, information synopsis requires the MLLM to summarize the video’s theme or infer its possible background. For clarification, we provide examples of these reasoning tasks as shown in [Figure 13](#).



Figure 13: Detailed examples of reasoning tasks.

C.2 Error Analysis

As described in section 5.2, we have classify the errors of MLLM into four categories. Considering Qwen2-VL-72B-Instruct have a comparable performance against other models, we provide the error cases of it as follows: (1) **Misconception of video creation**: Figure 14 shows that the MLLM doesn't understand how to generate a good video prompt. It may simply add more information and constraints to the original prompt. (2) **Excessive Dependence on Textual Cues**: Figure 15 demonstrates that the MLLM may rely on its prior knowledge rather than observing details in the video. (3) **Neglect of Critical Details**: Figure 16 shows that the MLLM tends to answer questions while overlooking critical details in AIGC videos. (4) **Over-reliance on Commonsense Knowledge**: Figure 17 illustrates the model's over-reliance on textual information.

[Question] : Analyze the mismatch between the video and the original text that was used for its generation. Based on this, propose a revised prompt that can be used to regenerate the video content. The original text is: Cut fruit exposed to air dries and develops a wrinkled texture. If there is no discrepancy, you can answer largely correlated with the video. If there may be no room to improve, you can answer that the original text is fairly well.



[LLM Response]: Answer: The video effectively showcases the process of slicing the fruits and observing the changes in their texture as they are exposed to the air.

New Prompt: \"Create a video that demonstrates the process of cutting citrus fruits and observing how they dry and develop a wrinkled texture when exposed to the air.

[Correct Answer]: The video illustrating the cut action while 'cut' is not needed to shown in the video. New prompt: There is a wrinkled texture on the freshly cut apple.

[Error reason analysis]: (1) The MLLM fails to recognize the wrinkled texture and **thinks the cut is a verb**. (2) The newly prompt **doesn't recognize the human intention** on the wrinkle texture and provide a long video generation prompt, which will confuse the video generation model.

Figure 14: Misconception of Video Creation

[Question]: Identify any quality issues in the video, including low resolution, blurriness, or distortion of the images.

[Choices]:

- A. The grass in the foreground is disappearing intermittently.
- B. The tree shadows are flickering inconsistently.
- C. The sun is pulsating erratically in brightness.



[LLM Response]: A.

[Correct Answer]: D.

[Error reason analysis]: The MLLM fails to analyze the choices by video content **but heavily relies on textual information in the question**. As a result, it randomly chooses answer A.

Figure 15: Excessive Dependence on Textual Cues

[Question] : Given the video, please identify if there is any quality issue in the video. Quality issues include but are not limited to: low resolution, blurry, distorted images, camera problems or aesthetics.



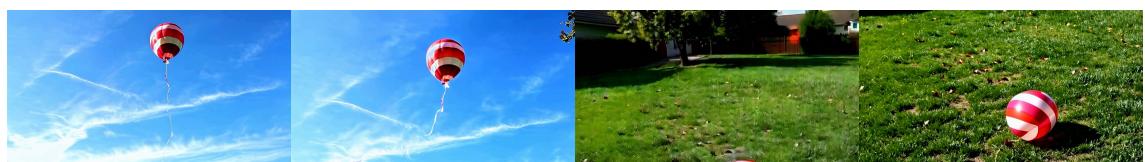
[LLM Response]: No. After analyzing the video, there are no apparent quality issues such as low resolution, blurry, distorted images, camera problems or aesthetic issues. The objects in the video are well-represented, and there are no obvious inconsistencies or anomalies observable. In conclusion, the video appears to be of good quality and can be accurately assessed.

[Correct Answer]: Yes.

[Error reason analysis]: The MLLM fails to analyze the fact that the **size proportion between the hand and the cup is incorrect**. And there exists distortion on the human hand.

Figure 16: Neglect of Critical Details

[Question] : Why does the balloon fall down after flying up in the video?



[LLM Response]: The balloon falls down after flying up in the video because it loses its buoyancy. This can happen for several reasons: the helium inside the balloon escapes, the balloon is punctured or damaged, or the weather conditions change. Balloons rely on the principle of buoyancy, where the lighter-than-air helium lifts the balloon against gravity. If the helium escapes or the balloon is no longer intact, the balloon will lose its buoyancy and descend to the ground. Therefore, the reason why the balloon falls down after flying up in the video is because it loses its buoyancy.

[Correct Answer]: The balloon may fall due to a collision with birds in the air, leading to potential leaks, but it should not turn into a ball once it lands.

[Error reason analysis]: The MLLM fails to analyze the choices by video content **but heavily rely on its commonsense and physical knowledge**. As a result, it fails to produce a satisfying answer.

Figure 17: Over-reliance on Commonsense Knowledge