

Exploring Hallucination of Large Multimodal Models in Video Understanding: Benchmark, Analysis and Mitigation

Hongcheng Gao^{*1}, Jiashu Qu^{*2}, Jingyi Tang^{*1,3}, Baolong Bi¹, Yue Liu⁴, Hongyu Chen⁵

Liang Li^{†3}, Li Su^{†1}, Qingming Huang^{1,3}

¹University of Chinese Academy of Sciences ²University of Cincinnati

³Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS

⁴National University of Singapore ⁵Beijing Jiaotong University

{gaohongcheng23, tjy23}@mails.ucas.ac.cn, quju@mail.uc.edu

Abstract

Hallucinations in large multimodal models (LMMs), where responses appear correct but are factually wrong, limit their reliability in practical applications. This paper focuses on hallucinations in video understanding, a more dynamic and challenging domain than static modalities. We present **HAVEN**, a comprehensive benchmark with 6K questions spanning three dimensions: hallucination causes, aspects, and question formats. Then, we systematically evaluate 18 LMMs and analyze seven influential factors affecting hallucinations, including video duration, model size, and reasoning capabilities. Additionally, we explore two reasoning-based strategies to reduce hallucinations in LMMs: (1) text-based reasoning via rejection fine-tuning on reasoning datasets and segment-weighted Direct Preference Optimization, and (2) video-based reasoning through a tool-augmented frame-level operations on video optimized via supervised tool learning and Hallucination-aware Group Relative Policy Optimization (HGRPO). Extensive experiments and analyses demonstrate the effectiveness. Remarkably, video-based reasoning strategy improves the Qwen2.5-VL-7B-Instruct by 10.16% in accuracy and reduces the bias score by 4.89%.

1 Introduction

In recent years, the success of Large Language Models (LLMs) (Touvron et al., 2023a,b; Jiang et al., 2024a, 2023) has significantly advanced the development of Large Multimodal Models (LMMs) (Yin et al., 2023; Liu et al., 2023b; OpenAI, 2023; Anthropic, 2024; Reid et al., 2024; Liu et al., 2024b; Dai et al., 2023a). By integrating vision and audio modules with LLMs, LMMs are capable of generating accurate responses based on

users’ textual prompts, visual inputs, and even audio data (Shao et al., 2023; Sammani et al., 2022; Fu et al., 2025a; Zhou et al., 2024; Gupta and Kembhavi, 2023). However, LLMs are found to suffer from hallucinations—providing responses that appear correct but are actually incorrect—especially when faced with questions that conflict with their training data (e.g., contradictions with factual knowledge, where presented cartoons depict cats as smaller than mice while in reality cats are larger than mice) (Li et al., 2023a; Zhang et al., 2023c). Despite the advancements in LMMs, they also experience similar hallucination issues in both text and image understanding (Huang et al., 2025b; Ji et al., 2023; Dai et al., 2023b). This phenomenon undermines user trust in LMMs and limits their applicability in high-stakes areas such as healthcare (Liu et al., 2023a) and autonomous driving (Tian et al., 2024). In response to these challenges, a series of studies (Zhu et al., 2024a; Zhang et al., 2023b; Han et al., 2024; Xu et al., 2024b; Chen et al., 2024a; Lee et al., 2023) have been proposed to evaluate and mitigate hallucinations in both LLMs and LMMs.

However, previous research on hallucinations of LMMs has primarily focused on image understanding (Li et al., 2023d; Liu et al., 2025; Guan et al., 2024a; Jiang et al., 2024b; Lovenia et al., 2024; Huang et al., 2024; Kaul et al., 2024; Fieback et al., 2024), as earlier LMMs could not process video inputs. These benchmarks are designed to evaluate hallucinations involving factors such as objects, relationships and attributes in a single image. With advancements in multimodal technologies, numerous LMMs now support video processing. Although many of these models do not incorporate audio inputs from videos, most can effectively process the visual content of video. Unlike image understanding, videos consist of sequences of multiple image frames over time, making video understanding more complex. It requires the anal-

^{*}Equal contribution.

[†]Corresponding authors: Li Su and Liang Li.

ysis of continuous temporal dynamics, including sequential changes in human actions, object movements, and scene transitions. Hence, hallucinations in video understanding also differ from those in images.

To address the concern above, we first proposed a benchmark for **H**Allucination in **V**ideo **U**ndersta**N**ding (**HAVEN**). Fundamentally differing from Track2Act (Bharadhwaj et al., 2024), which predicts point tracks from videos for generalizable robot manipulation, HAVEN is meticulously designed to quantitatively evaluate hallucinations in video understanding. To this end, HAVEN is constructed based on the following dimensions: (i) Three *causes* of hallucinations: conflict with prior knowledge, in-context conflict, and inherent capability deficiencies of LMMs. (ii) Three types of hallucination *aspects* in a video: object, scene, and event. (iii) Three *formats* of questions: binary-choice, multiple-choice, and short-answer. HAVEN represents a systematic distillation of hallucination dimensions into a unified framework rather than a simple combination of existing resources. By synthesizing scattered dimensions into a coherent taxonomy and enabling cross-dimensional comparison under consistent protocols, HAVEN reveals insights and interactions difficult to observe from isolated benchmarks.

Our evaluation of 18 existing LMMs reveals that GPT-5.1 and OpenAI o1 demonstrated the lowest hallucination rates and superior response consistency among all models tested. The analysis further indicated that model performance is influenced by several factors: accuracy initially increases then decreases with longer video duration, decreases with more complex questions, improves with increased frame sampling, and generally shows reduced hallucinations and higher consistency with larger model sizes. Besides, chain-of-thought (CoT) reasoning can reduce hallucinations in LMMs.

In addition, we explore two reasoning-based strategies to mitigate hallucinations inspired by recent advances in CoT. We develop (1) text-based reasoning, which strengthens long CoT to encourage iterative thinking and self-revision; and (2) video-based reasoning, which encourages LMMs to actively leverage external tools for frame-level video processing and visual exploration. For text-based reasoning, we perform Rejection Fine-tuning (RFT) with distilled reasoning data from OpenAI o1, then apply segment-weighted Direct Preference

Optimization (DPO) to reduce hallucinations in the reasoning process. For video-based reasoning, we construct a multi-round tool-augmented reasoning pipeline with three operations—temporal interval segmentation, replay, and bidirectional temporal sliding—and optimize tool utilization through Supervised Fine-tuning (SFT) and our proposed Hallucination-aware Group Relative Policy Optimization (HGRPO) algorithm. Extensive experiments on Qwen2.5-VL series models demonstrate that our reasoning-enhanced approaches yield substantial improvements in both reducing hallucinations and enhancing response consistency.

2 Related Works

Large Multimodal Models (LMMs). Recent LMMs primarily focus on visual-language understanding, which can be categorized into image-language (Liu et al., 2023b; OpenAI, 2023; Anthropic, 2024; Liu et al., 2024b) and video-language models (Lin et al., 2024; Maaz et al., 2024a; Zhang et al., 2023a; Reid et al., 2024). Early visual-language models, designed for single-image inputs (Radford et al., 2021; Jia et al., 2021; Li et al., 2019; Kim et al., 2021), followed two main approaches: self-supervised learning with image-text pairs (Radford et al., 2021; Jia et al., 2021), and adding image adapters to existing LLMs followed by alignment fine-tuning (Li et al., 2019; Kim et al., 2021). Subsequent research extended to multi-image processing capabilities, naturally evolving into video-language models, which leverage frame-by-frame processing inherited from image-language models to achieve video understanding through text-video alignment (Lin et al., 2024; Maaz et al., 2024a; Zhang et al., 2023a, 2024c; Xu et al., 2024a; Li et al., 2024c; Cheng et al., 2024; Chen et al., 2024b). VideoLLaVA (Lin et al., 2024), for example, incorporates a Language-Bind encoder to align visual features from both images and videos into a unified space for joint video-text encoding. Recent developments have further expanded to modalities including audio (Cheng et al., 2024; Huang et al., 2025a) and 3D point clouds (Xu et al., 2025).

Hallucination Benchmarks. Hallucination is widespread in large models (Touvron et al., 2023a,b; Jiang et al., 2024a, 2023), and the relevant benchmarks primarily target LLMs and VLMs. Hallucination benchmarks for LLMs (Han et al., 2019; Li et al., 2023a; Lin et al., 2022; McKenzie et al., 2023; Wei et al., 2024; Ravichander et al.,

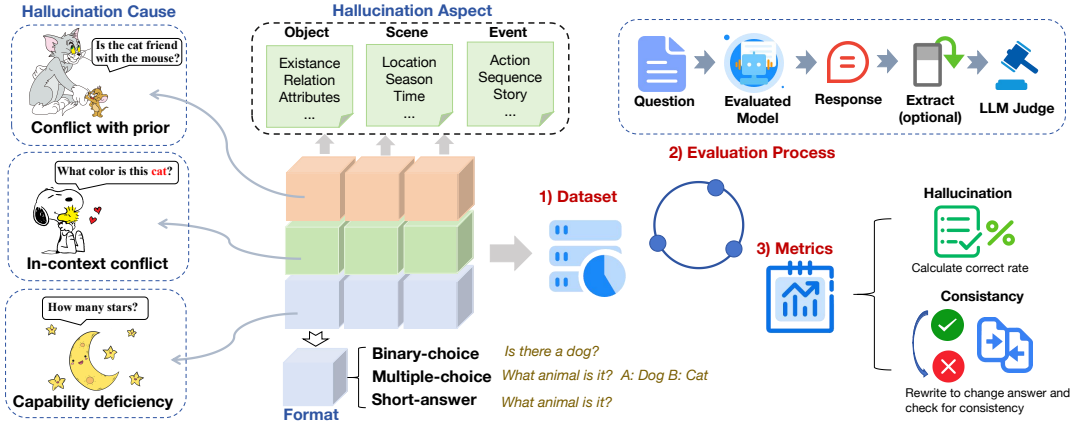


Figure 1: **Construction protocol of HAVEN.** The left section outlines the three dimensions of data construction and the associated categories within each, while the right section details the evaluation process and metrics. “Conflict with Prior” illustrates scenarios where visual facts (e.g., a cat befriending a mouse) contradict the model’s internal priors (e.g., predator-prey relationship).

2025) mainly focus on identifying factual inaccuracies, deviations from original contexts and the self-awareness. For VLMs, hallucination refers to the phenomenon that the generated text responses are inconsistent with the corresponding visual content (Bai et al., 2024). Numerous benchmarks (Li et al., 2023d; Liu et al., 2025; Guan et al., 2024a; Jiang et al., 2024b; Lovenia et al., 2024; Huang et al., 2024; Kaul et al., 2024; Fieback et al., 2024; Jing et al., 2024; Cui et al., 2023; Wang et al., 2023, 2024a; Chen et al., 2024c; Cha et al., 2024; Sun et al., 2024) have been developed to evaluate these hallucinations. Early benchmarks (e.g., CHAIR (Rohrbach et al., 2018) and POPE (Li et al., 2023c)) primarily target object presence, while later ones like HallusionBench (Guan et al., 2024b) and MMHal-Bench (Sun et al., 2024) also evaluate object relationships and scene understanding. Recent studies (Jing et al., 2024; Wang et al., 2024c; Liu et al., 2025; Cui et al., 2023) have begun addressing factual hallucinations in image-language models, yet most approaches focus on single image, neglecting continuous content. To bridge this gap, several benchmarks on video have been proposed from different perspectives—general video hallucination (Wang et al., 2024d), event hallucination (Zhang et al., 2024a), motion hallucination (Kong et al., 2025), and caption hallucination (Choong et al., 2026)—but they typically focus on a single dimension or a specific task format, lacking a comprehensive multi-dimensional, cross-task evaluation framework.

Hallucination Mitigation. Approaches to mitigating hallucinations can be broadly classified into two categories based on the stage at which they

intervene. The first category deals with the training phase, which focused on creating datasets specifically for hallucination-related tasks (Gunasekar et al., 2023; Gunjal et al., 2024; Liu et al., 2024a; Yu et al., 2024a; Tang et al., 2025; Shen et al., 2021; Raunak et al., 2021) and developing novel training objectives or methodologies aimed at mitigating hallucinations (Chen et al., 2024a; Lee et al., 2023; Shi et al., 2024). Although these methods have demonstrated effectiveness, they often require extensive training processes, making them both time-consuming and resource intensive. The second category focuses on the inference phase, usually involves CD-based (Li et al., 2023b) novel decoding strategies (Chen et al., 2024d; Chuang et al., 2024; Kim et al., 2024a; Liu et al., 2024c; Phan et al., 2024; Kim et al., 2024b; Woo et al., 2024b; Park et al., 2024; Zhao et al., 2024; Zhu et al., 2024b). Another approach is to detect potential hallucinations while generating and correct them (Kuhn et al., 2023; Nikitin et al., 2024), among others (Li et al., 2024a; Yue et al., 2024; Woo et al., 2024a; Zhang et al., 2024b). In addition, some researchers tackle hallucinations by manipulating attention weight assigned to the image (Zhu et al., 2024a; Zhang et al., 2023b; Xiao et al., 2024; Huo et al., 2024; Wu et al., 2024b), and prompt-based methods (Han et al., 2024; Xu et al., 2024b; Wang et al., 2024b; Qu et al., 2024; Wu et al., 2024a; Wang et al., 2024d). Concurrent work (Wang et al., 2026) explores an agentic framework for active video perception in long videos to improve VLM video understanding performance.

Benchmark	Hallucination Causes			Hallucination Aspects			Question Types			Position Bias	# of Ques
	Prior Conf.	In-context Conf.	Capability	Object	Scene	Event	T/F	MC	SA		
VideoHalluciner (Wang et al., 2024d)	✓	✓	✓	✓	✗	✓	✓	✗	✗	✗	1800
HallusionBench (Guan et al., 2024d)	✓	✗	✓	✓	✗	✗	✓	✓	✗	✓	1129
EventHallusion (Zhang et al., 2024a)	✓	✓	✗	✗	✗	✓	✓	✗	✓	✗	711
MHBench (Kong et al., 2025)	✗	✗	✓	✓	✗	✓	✓	✓	✗	✓	3600
VIDHAL (Choong et al., 2026)	✓	✓	✓	✓	✓	✓	✗	✓	✗	✗	4000
HAVEN	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	6497

Table 1: Comparison of Video Hallucination Benchmarks

3 HAVEN Benchmark

Recent video understanding benchmarks (Fang et al., 2024; Fu et al., 2025b; Li et al., 2024b) have advanced the evaluation of the general capability of video perception and understanding. However, they are not designed to assess hallucinations in video understanding, motivating us to introduce HAVEN, a dedicated benchmark for evaluating and analyzing video hallucinations.

3.1 Data Sources

Our video data comprises videos from three public video datasets (COIN (Tang et al., 2019), ActivityNet (Caba Heilbron et al., 2015), and Sports1M (Kang and Wildes, 2016)) and 260 manually collected video clips from YouTube (YouTube). All the questions and answers are manually annotated by human experts. To ensure correctness, each example is independently checked by at least two annotators for answer correctness, category assignment, and video-grounding validity; disagreements are adjudicated before inclusion. Further annotation details and quality-control statistics are provided in Appendix A.

3.2 Construction Protocol

As shown in Fig. 1, we first constructed a dataset based on three dimensions—*hallucination causes*, *hallucination aspects*, and *question formats*—before implementing an automated evaluation process along with two metrics.

3.2.1 Hallucination Causes

Drawing on and synthesizing insights from prior hallucination studies in both LLMs and VLMs (Choong et al., 2026; Guan et al., 2024d; Han et al., 2019; Li et al., 2023a; Lin et al., 2022; McKenzie et al., 2023; Ravichander et al., 2025; Wang et al., 2024d; Wei et al., 2024; Zhang et al., 2024a), we organize HAVEN around three operational hallucination-source categories: conflict with prior, in-context conflict, and capability deficiencies. These categories summarize and unify commonly discussed hallucination factors in existing

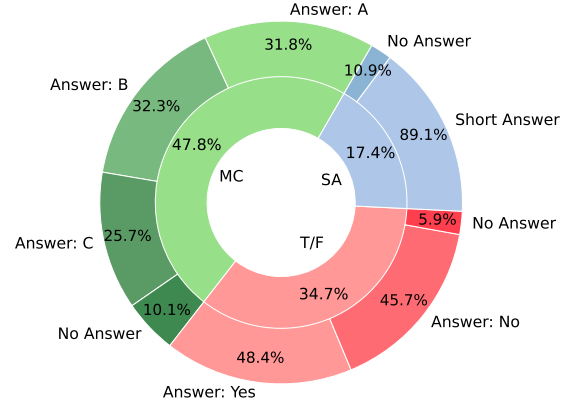


Figure 2: **Question format distribution.** Percentage share of each format-binary-choice (T/F), multiple-choice (MC), and short-answer (SA)—and the proportion occupied by the detailed answer.

literature into a taxonomy tailored for video understanding.

Conflict with Prior. When the content extracted from the video contradicts the model’s inherent prior knowledge acquired during training, the model may rely on its prior knowledge learned from established facts or widely accepted knowledge, leading to incorrect responses that don’t align with the provided information. Following prior work on hallucination evaluation in VLMs (Guan et al., 2024c; Li et al., 2023c), we approximate model priors using commonly shared world knowledge, as such knowledge is widely represented in large-scale pretraining corpora. Based on this assumption, prior-conflict examples are manually annotated as cases where the video evidence contradicts expectations derived from world knowledge.

In-context Conflict. When task components conflict—such as discrepancies between the question and options or between the video and the questions/answers—valid answers cannot be derived from the given material. In such instances, the model should indicate uncertainty by responding with “I don’t know/no answers”. However, models prone to hallucination often produce fabricated responses by focusing on limited contextual cues.

Capability Deficiencies. Most LMMs struggle with numerical tasks, often leading to quantifica-

tion errors in videos. For example, they might label a scene as having four cars when there are actually three, or miscount the number of people present. Such inaccuracies can undermine the reliability of video analysis due to incorrect quantitative details.

3.2.2 Hallucination Aspects

According to the components of a video, hallucination aspects can be categorized into three types: object, scene, and event hallucinations.

Object. Object hallucinations involve inaccuracies or fabrications related to the entities present in the video. This primarily includes the presence of objects, their inherent attributes such as size, color, and shape, as well as the emotions of characters.

Scene. Scene hallucinations involve inaccuracies related to the overall setting and environment of the video. This includes errors concerning the background, location, season, time of day, lighting conditions, and the general context in which events take place. For example, a scene-level hallucination might mistakenly depict an indoor setting as outdoor, alter the time of day, or misrepresent environmental factors such as weather conditions.

Event. Event hallucinations involve distortions or false representations of the actions and occurrences within the video. This includes inaccuracies in the sequence of events, the nature of interactions between characters, and the specific actions taking place. For instance, an event-level hallucination might depict a character performing an action that does not actually occur in the video or misrepresent the timing and order of events.

3.2.3 Question Types

We classify questions into three types: binary-choice(T/F), multiple-choice(MC), and short-answer(SA). Binary-choice questions are phrased as interrogative sentences (e.g., “Is there a dog?”) requiring a “yes” or “no” answer. Multiple-choice question present a set of options, while short-answer question need concise reply. For any question that lacks a definitive answer, the LMM should respond with “no answer” or “I don’t know.”

3.3 Data Post-processing

To evaluate the inherent response biases of language models—namely their predisposition toward particular responses in binary-choice questions and multiple-choice questions—we implement a systematic question transformation protocol. For each binary question, we generate two complementary

Cause/Aspect	Object	Scene	Event	#Total
Prior Conflict	2162	686	1763	4569
In-context Conflict	94	82	404	538
Capability	1107	121	78	1156
#Total	3363	889	2245	6497

Table 2: The Statistics of HAVEN.

variants: one where “yes” is the correct answer and another where “no” is correct. Similarly, we transform each multiple-choice question into three parallel versions, with each version designating a different option (A, B, or C) as the correct answer. This approach aims to neutralize potential positional biases in model responses.

3.4 Dataset Statistics

As shown in Table 2 and Fig. 2, HAVEN contains 6,497 questions. The three question formats are distributed approximately in a 5:2:3 ratio. The prior-conflict category accounts for the largest share, while the three hallucination aspects are distributed roughly as 3:1:2. This imbalance mainly stems from the benchmark construction process. Prior-conflict examples can often be identified from naturally occurring inconsistencies between visual evidence and commonly shared world knowledge. In contrast, in-context conflict and capability-deficiency examples typically require more deliberate manual construction and tend to become increasingly repetitive when scaled up. To account for this bias, we report both overall and category-level results.

Figure 3 shows the distribution of our data in terms of duration, frame count, and question length. The duration spans from 0 to 70 seconds, with most instances falling within the 0–20 second interval. The frame count ranges from 0 to 1200 frames, with the majority concentrated between 0 and 400 frames. The question length is computed based on the token count obtained via GPT-4 tokenizer, with a dominant range of 10–15 tokens.

The main comparison between HAVEN and existing benchmarks is presented in Table 1. As it shows, HAVEN is the largest and most comprehensive benchmark, covering different hallucination causes, hallucination aspects, and question types. Importantly, HAVEN is not merely an aggregation of existing resources but a unified evaluation framework that integrates multiple hallucination dimensions within a single cohesive benchmark. This systematic approach allows us to conduct cross-dimensional analyses under unified protocols, un-

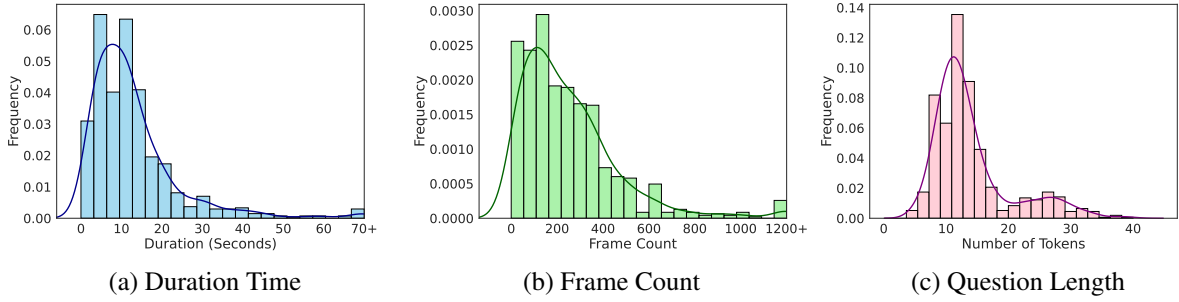


Figure 3: Distribution of duration time, frame count, and question length.

covering novel patterns and interactions between different hallucination types that were previously obscured when studied in isolation.

4 Analysis of Video Hallucination

4.1 Settings

We evaluated 17 LMMs in total: one 3B model, nine 7B models, three 13B models, one 34B model, and three API models. To ensure a fair comparison, we maintained the original settings for all baselines, including frame counts and generation hyperparameters. Details of these LMMs can be found in Appendix D.

4.2 Hallucination Evaluation

Table 3 demonstrates a performance comparison of various LMMs in hallucination evaluation across different dimensions with accuracy. Higher scores represent lower levels of hallucination. The evaluation results are categorized into three causes (prior conflict, in-context conflict, and capability deficiencies), with each category assessing three video aspects (object, scene, and event).

GPT-5.1 achieves the best overall performance with 74.11% accuracy, outperforming OpenAI o1 (68.24%) and GPT-4o (63.34%), particularly on prior-conflict and capability-related questions. In contrast, Video-LLaMA-2-13B and VideoChatGPT-7B show the weakest performance, particularly struggling with in-context conflict situations. We present some failure cases in Fig. 11 and 12 of Appendix.

4.3 Consistency Evaluation

Table 4 shows consistency evaluation with bias scores, where lower values indicate better performance. GPT-5.1 achieves the lowest overall bias score (13.33%), outperforming OpenAI o1 (27.03%) and Qwen2.5-VL-3B-Instruct (27.66%). It also performs best on binary-choice tasks, while Qwen2.5-VL-3B-Instruct achieves the lowest bias

on multiple-choice tasks. In contrast, VideoChat2-7B and PLLaVA-7B showed the weakest performance in consistency evaluations.

4.4 Influential Factors of Hallucination

We analyzed seven influential factors affecting hallucinations, as follows:

Length of Videos and Questions. To analyze the relationship between LMM hallucination and various factors of input data, we conducted an analysis examining how accuracy varies with respect to video duration, frame count, and question length. As illustrated in Fig. 4, an increase in video duration or frame count initially enhances performance, but beyond a certain point, the performance declines. This trend may be attributed to the fact that longer videos provide additional information; however, if the video is too long, the model may fail to sample the most relevant frames, leading to information loss.

Conversely, as the length of the question text increases, the model’s performance consistently deteriorates, possibly because the additional textual information shifts the model’s multimodal attention toward the text, diverting it away from the visual content. To empirically verify this, we conducted a further analysis using Qwen2.5-VL averaged over 10 samples. As shown in Fig. 6a, visual attention notably decays as prompt length increases, while Fig. 6b reveals that the output similarity between real and blank videos rises to a high saturation level. These trends indicate that excessive text causes the model to marginalize visual inputs in favor of textual priors, driving the observed performance drop.

Sampling Number of Frames. As video information is frame-dependent, we examine the relationship between the number of sampled frames and model performance. We choose three models to demonstrate overall trends with mean values and variances. Fig. 7 shows that model hallu-

Model/Type	Prior			In-context			Capability			Total
	Object	Scene	Event	Object	Scene	Event	Object	Scene	Event	
VideoChatGPT-7B(Maaz et al., 2024b)	34.78	43.88	38.00	17.02	13.41	17.57	32.52	47.11	20.51	34.69
Valley-Eagle-7B (Wu et al., 2025)	68.55	<u>75.95</u>	63.52	24.47	43.90	15.10	57.36	55.37	47.43	61.29
VideoLLaVA-7B (Lin et al., 2024)	47.55	57.29	50.59	14.89	12.19	10.15	35.68	33.88	34.61	43.73
VideoChat2-7B (Li et al., 2025a)	43.20	48.98	43.00	30.85	26.83	26.00	26.74	28.92	33.33	39.11
ShareGPT4Video (Chen et al., 2024b)	51.48	62.24	48.27	17.02	19.51	8.91	43.00	38.84	32.05	46.28
LLaMA-VID-7B (Li et al., 2024c)	48.75	56.71	50.09	20.21	21.95	21.78	37.76	42.98	29.49	45.31
LLaMA-VID-13B (Li et al., 2024c)	47.82	54.23	49.29	12.76	29.26	5.20	40.65	38.01	32.05	43.91
PLLaVA-7B (Xu et al., 2024a)	44.26	60.93	41.41	20.21	37.80	9.90	43.63	15.70	32.05	40.17
PLLaVA-13B (Xu et al., 2024a)	62.02	69.39	56.55	21.28	50.00	15.10	48.60	46.28	44.87	54.87
Qwen2.5-VL-3B-Instruct (Team, 2025)	52.36	65.31	49.12	20.21	37.80	5.20	43.63	15.70	32.05	46.85
Qwen2.5-VL-7B-Instruct (Team, 2025)	55.97	60.05	50.25	20.21	32.93	8.66	53.39	44.63	34.61	50.19
LLaVA-NeXT-Video-DPO-7B (Zhang et al., 2024c)	49.35	57.58	49.86	15.96	31.71	12.62	38.12	43.80	41.03	45.25
LLaVA-NeXT-Video-DPO-34B (Zhang et al., 2024c)	52.59	60.79	47.42	22.34	34.15	7.67	45.17	41.32	26.92	46.81
Video-LLaMA-2-13B (Cheng et al., 2024)	28.95	40.96	30.80	11.70	9.76	12.62	16.71	21.49	26.92	26.97
GPT-4o-mini (Hurst et al., 2024)	52.87	59.47	54.79	58.51	64.63	62.13	63.50	52.89	60.26	56.80
GPT-4o (Hurst et al., 2024)	64.62	65.25	59.19	63.37	67.46	<u>64.12</u>	66.45	55.97	63.97	63.34
OpenAI o1 (Jaech et al., 2024)	<u>69.38</u>	72.59	<u>63.61</u>	65.86	<u>69.27</u>	65.28	72.85	<u>60.05</u>	67.73	<u>68.24</u>
GPT-5.1 (OpenAI, 2026)	77.85	76.58	77.87	<u>64.29</u>	70.00	61.82	64.96	77.78	53.85	74.11

Table 3: **Hallucination Evaluation.** Accuracy of LMMs on questions of different hallucination targets across three causes. The **bold** indicates the best performance, and the underline indicates the second-best.

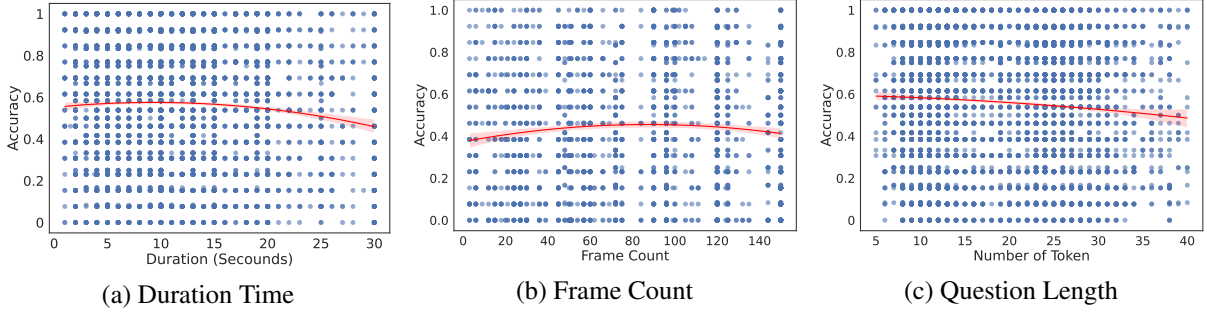


Figure 4: The impact of video duration, frame count, and question length on LMM hallucination.

Model/Type	Binary-choice	Multiple-choice	Total
VideoChatGPT-7B	57.58	39.19	46.97
Valley-Eagle-7B	42.41	<u>17.99</u>	28.31
VideoLLaVA-7B	59.08	43.62	50.16
VideoChat2-7B	52.43	49.75	50.88
ShareGPT4Video	60.11	20.90	37.48
LLaMA-VID-7B	52.15	40.20	45.25
LLaMA-VID-13B	59.83	30.75	43.05
PLLaVA-7B	61.14	42.51	50.38
PLLaVA-13B	52.62	24.02	36.11
Qwen2.5-VL-3B-Instruct	41.67	17.39	27.66
Qwen2.5-VL-7B-Instruct	48.50	19.90	31.99
LLaVA-NeXT-Video-DPO-7B	51.13	41.41	45.52
LLaVA-NeXT-Video-DPO-34B	49.06	18.90	31.65
Video-LLaMA-2-13B	66.48	24.32	42.14
GPT-4o-mini	34.08	29.95	31.69
GPT-4o	35.46	31.54	33.19
OpenAI o1	<u>32.36</u>	23.14	27.03
GPT-5.1	8.51	30.77	13.33

Table 4: Consistency Evaluation

cinations first decrease then increase with more frames, as performance gains from extra context are eventually offset by the train-test frame-count mismatch. This phenomenon is particularly pronounced in models trained with fixed frame counts (e.g., VideoLLaVA-7B with 8 frames, Video-LLaMA-2 with 16 frames), while models using dynamic frame sampling strategies show greater robustness to frame count variations. Fig. 8

demonstrates that Qwen2.5-VL achieves consistent gains as frame counts increase, showing no performance decay and peaking with 32 frames.

Model Size. Fig. 9 illustrates how accuracy and bias scores change as parameter number increases across different model categories. We selected four distinct types of models across various sizes and presented the regression lines. Most LMMs fit the trend that larger models tend to experience fewer hallucinations and exhibit lower levels of bias, although a few LMMs don’t align perfectly.

Co-impact of Each Two Dimensions. In Fig. 5, we visualize the co-impact of each pair of dimensions by calculating the average accuracy across all models for corresponding categories in each dimension pair. The results demonstrate that hallucinations caused by in-context conflicts consistently show the most significant impact across all combinations. Additionally, short-answer questions consistently exhibit lower performance scores across all combinations.

Chain-of-Thought Reasoning. To explore whether enhanced reasoning capabilities could

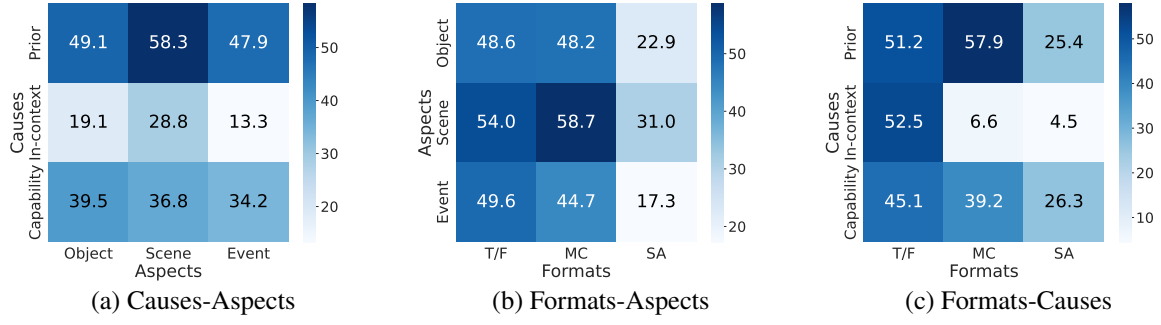
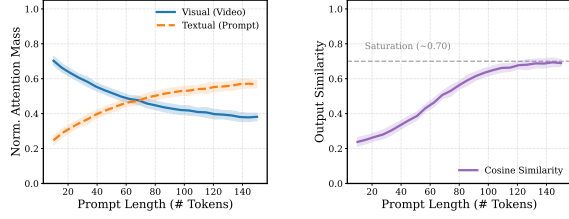


Figure 5: Accuracy heatmap along two dimensions.



(a) Visual Attention Decay. As prompt length increases, the model’s attention on video tokens decreases, indicating a persistent but diminished visual grounding.

(b) The Blindness Curve. The similarity between outputs given real vs. blank videos rises with text length, saturating at ~ 0.70 , suggesting high textual dependency.

Figure 6: **Quantitative Analysis of Textual Divergence.** We simulate the impact of prompt length on (a) visual attention allocation and (b) the model’s sensitivity to visual input removal. Shaded areas represent standard deviation across 50 runs.

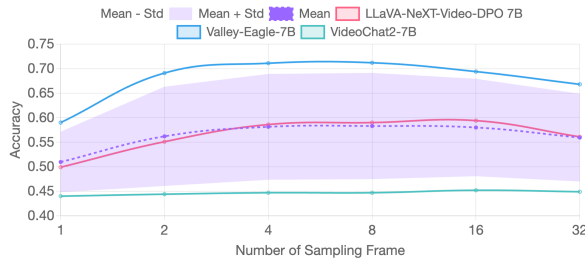


Figure 7: Relationship between number of sampling frames and model performance for model trained with **fixed frame counts**.

reduce model hallucinations, we conducted a comparative study of three LMMs, evaluating their performance before and after implementing Chain-of-Thought (Wei et al., 2022)(CoT) by appending the prompt “Let’s think step by step.” to the original question. As shown in Table 5, all LMMs demonstrated improved accuracy via CoT, indicating reduced hallucinations. Additionally, bias scores decreased for the majority of LMMs, suggesting improved consistency.

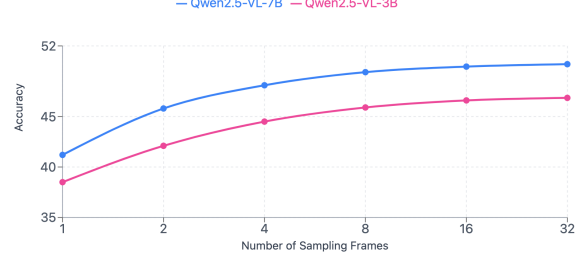
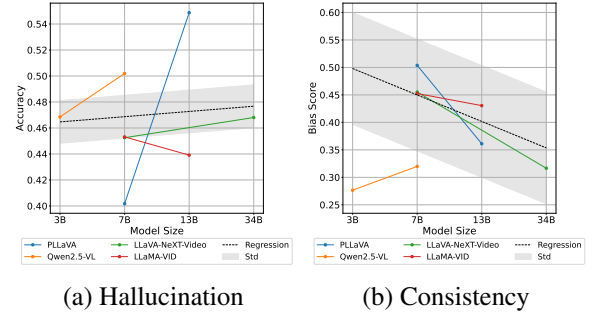


Figure 8: Relationship between number of sampling frames and model performance for model trained with **dynamic frame counts**.



(a) Hallucination (b) Consistency

Figure 9: Model size vs. performance.

5 Mitigating Video Hallucination

As shown in Section 4.4, CoT may offer a path to enhance the reliability of LMMs by promoting step-wise, verifiable inference. We thereby explore two reasoning strategies for video understanding: (1) *text-based reasoning*, which strengthens long CoT to encourage iterative thinking and self-revision; and (2) *video-based reasoning*, which encourages the LMMs to actively leverage external tools to perform frame-level video processing for visual exploration and reasoning. Figure 10 provides an overview of our approach.

5.1 Text-based Reasoning

The o1-like large reasoning models (LRMs) (Jaech et al., 2024; Guo et al., 2025; Team et al., 2025) have demonstrated that long text CoT can significantly improve the reasoning capabilities of large models by revealing their thinking process. Hence,

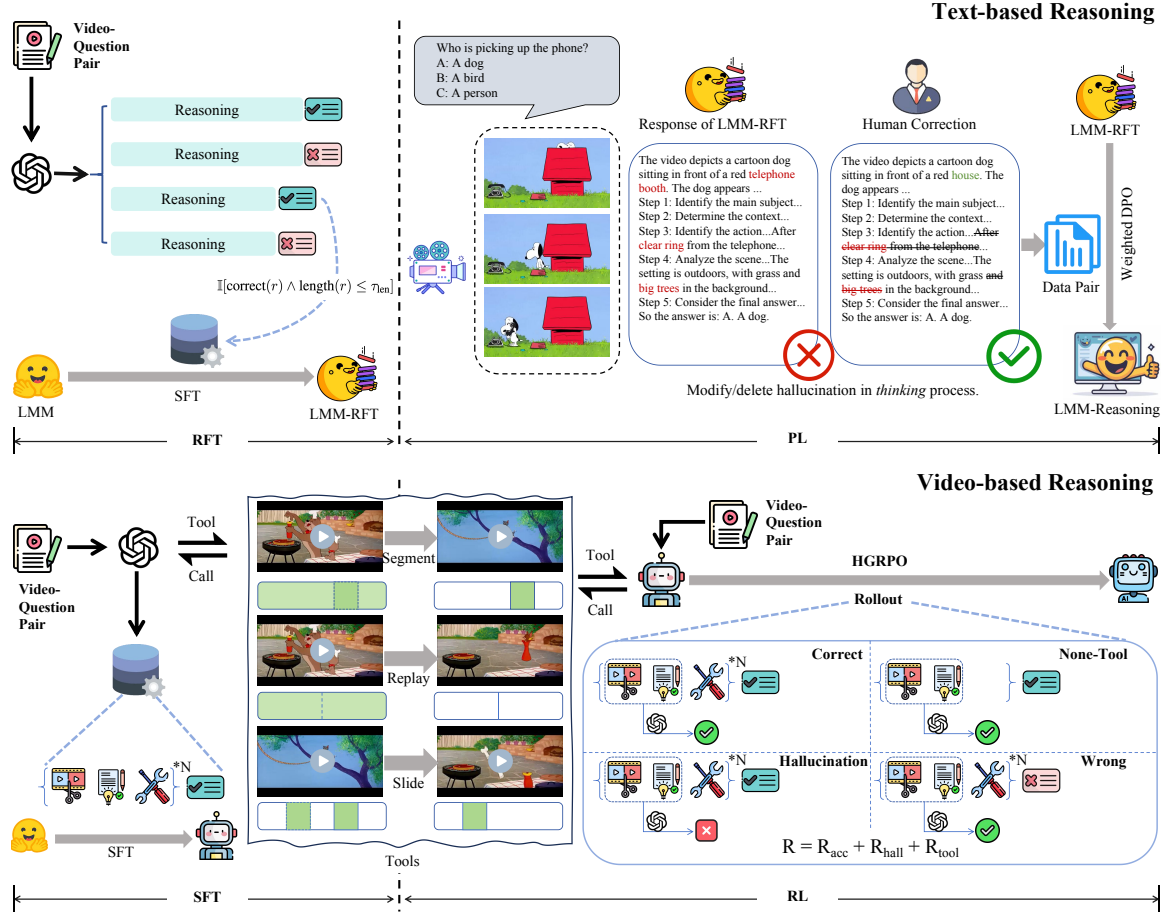


Figure 10: **Our two reasoning-based strategies for mitigating video hallucinations.** *Top:* text-based reasoning via rejection fine-tuning (RFT) and segment-weighted DPO for fine-grained preference learning (PL). *Bottom:* video-based reasoning with frame-level tool augmentation—SFT on tool-usage trajectories followed by Hallucination-aware Group Relative Policy Optimization (HGRPO) with accuracy, hallucination, and tool-use rewards.

Model/Type	Accuracy $\uparrow$	Bias Score $\downarrow$
PLLaVA-7B	40.17	50.38
+CoT	44.95	40.38
LLaMA-VID-7B	45.31	45.25
+CoT	46.05	50.20
LLaVA-NEXT-Video-7B	45.25	45.52
+CoT	49.56	45.15
Qwen2.5-VL-3B-Instruct	46.85	27.66
+CoT	50.49	31.89
Qwen2.5-VL-7B-Instruct	50.25	31.99
+CoT	53.98	29.97

Table 5: Comparison of the model before/after incorporating CoT.

we begin by unlocking reasoning abilities via RFT on reasoning datasets distilled from LRMs, and subsequently minimize hallucinations through Preference Learning (PL).

Unlocking Reasoning Ability. To unlock the reasoning capabilities of LMMs, a straightforward

and efficient method is RFT with distilled reasoning data (Guo et al., 2025) from video o1-like models (Feng et al., 2025; Li et al., 2025b; Zhang et al., 2025). The reasoning dataset $\mathcal{D} = \{(v_i, x_i, r_i)\}_{i=1}^N$ is constructed by feeding video-question pairs (v_i, x_i) to o1-like models to obtain appropriate responses r_i through reject sampling. Specifically, we generate K candidate responses $\{r_k\}_{k=1}^K$ per sample using temperature sampling and select the first candidate response satisfying $\mathcal{F}(r) = \mathbb{I}[\text{correct}(r) \wedge \text{length}(r) \leq \tau_{\text{len}}]$. In our work, we leverage OpenAI o1 (Jaech et al., 2024) to generate reasoning outputs for video-text inputs, enhancing the reasoning capabilities of LMMs. During RFT, we optimize model parameters θ by minimizing the cross-entropy loss $\mathcal{L}_{\text{RFT}} = -\sum_{i=1}^N \log p_{\theta}(r_i|v_i, x_i)$ over the video-text-reasoning triplets in $\mathcal{D}$.

PL on Hallucination. Although RFT endows the LMMs with reasoning capability, the reasoning

Model/Type	Prior			In-context			Capability			Total
	Object	Scene	Event	Object	Scene	Event	Object	Scene	Event	
GPT-4o-mini	52.87	59.47	54.79	58.51	64.63	62.13	63.50	52.89	60.26	56.80
GPT-4o	64.62	65.25	59.19	63.37	67.46	64.12	66.45	55.97	63.97	63.34
OpenAI o1	69.38	72.59	63.61	65.86	69.27	65.28	72.85	60.05	67.73	68.24
GPT-5.1	77.85	76.58	77.87	64.29	70.00	61.82	64.96	77.78	53.85	74.11
PLLaVA	44.26	60.93	41.41	20.21	37.80	9.90	43.63	15.70	32.05	40.17
+CoT	49.44	53.50	44.13	35.10	58.54	38.37	36.77	33.88	29.49	44.95
LLaMA-VID-7B	48.75	56.71	50.09	20.21	21.95	21.78	37.76	42.98	29.49	45.31
+CoT	50.04	58.45	50.54	24.47	25.61	26.98	35.50	39.67	30.77	46.05
LLaVA-NeXT-Video-DPO-7B	51.48	62.24	48.27	17.02	19.51	8.91	43.00	38.84	32.05	45.25
+CoT	52.08	64.28	50.88	36.17	51.22	34.41	41.01	47.93	37.18	49.56
Qwen2.5-VL-3B-Instruct	52.36	65.31	49.12	20.21	37.80	5.20	43.63	15.70	32.05	46.85
+CoT	53.91	67.51	51.46	31.06	54.86	32.34	42.36	36.09	34.81	50.49
+Text Reasoning 🔥	56.42	69.26	52.75	36.47	56.68	38.14	44.16	37.64	35.08	52.66
+Video Reasoning 🔥	58.16	68.14	52.36	35.28	56.14	40.14	46.78	38.08	36.85	53.59
Qwen2.5-VL-7B-Instruct	55.97	60.05	50.25	20.21	32.93	8.66	53.39	44.63	34.61	50.19
+CoT	58.07	62.93	52.98	39.06	56.48	36.06	51.74	46.29	37.36	53.98
+Text Reasoning 🔥	62.63	64.57	58.52	58.56	59.48	41.62	56.15	52.64	46.43	58.83
+Video Reasoning 🔥	63.45	63.56	60.65	56.43	58.27	42.65	60.43	52.18	49.62	60.35

Table 6: Hallucination evaluation of base models trained by two strategies compared with other LMMs.

process still exhibits hallucinations. Drawing inspiration from fine-grained PL in RLHF-V (Yu et al., 2024b), we consider this approach particularly well-suited for reducing hallucinations in reasoning steps for two key reasons: hallucinations typically appear as isolated incorrect items within the CoT, and reasoning steps have natural segmentation that allows us to assign different preference weights based on each segment’s hallucination severity. Therefore, we use segment-weighted DPO method tailored for the reasoning process for fine-grained PL. Given a video v and a corresponding question x , we initially employ the LMM to generate a response y . We then construct preference data $\mathcal{D} = (x_i, v_i, y_i^+, y_i^-)_{i=1}^N$ where y^- is the original model-generated response, and y^+ is the manually revised response with hallucinated reasoning steps removed or modified. We score the response as a weighted aggregation of the fine-grained thinking segments. Specifically, the likelihood of generating an output response y given an input question x and video v is defined as:

$$\log \pi(y|x, v) = K \left[\sum_{y_i \in y_o} \log p(y_i | x, v, y_{<i}) + \gamma \sum_{y_i \in y_h} \log p(y_i | x, v, y_{<i}) \right] \quad (1)$$

where y_o and y_h respectively denote the token sets corresponding to the original and human-corrected segments. Unlike the standard DPO—which uses

the likelihood $\log \pi(y | x, v) = \sum_{y_i \in y} \log p(y_i | x, v, y_{<i})$ as action score—treats all segments equally, the segment-weighted DPO assigns greater weight to the corrected reasoning steps, thereby reinforcing the human preference for responses that are factually grounded. Tradeoff hyperparameter γ can increase the emphasis on the corrected segments. As with RLHF-V, we also use normalization constant, $K = 1/(|y_o| + \gamma|y_h|)$, where $|y_o|$ and $|y_h|$ denotes length of the original and human corrected segments, preventing longer responses from receiving disproportionately high scores.

5.2 Video-based Reasoning

Text-only CoT can not compensate for missing or noisy visual evidence, especially in long videos. Prior work (e.g., OpenAI o3 (OpenAI, 2025)) shows that augmenting LMMs with tools like cropping and zooming improves visual understanding. We extend this idea to video by adding frame-level tool augmentation to reduce hallucinations via targeted frame extraction. Concretely, we equip the model with a frame-level tool interface supporting three operations—temporal interval segmentation, replay, and bidirectional temporal sliding—to actively retrieve and reason with visual evidence.

Learning to Use Tool. We construct a multi-round tool-augmented reasoning pipeline for data generation with three operations: (i) temporal interval segmentation (select a subclip by start–end timestamps), (ii) replay (re-examine

a specific frame in the video sequence), and (iii) bidirectional temporal sliding (shift the interval forward/backward to locate missing cues). Given a video-question pair (V_0, T_0) , the agent iteratively produces a trajectory $\tau = \{T_1, C_1, V_1, \dots, T_n, A_n\}$ across n rounds, where T_i is textual reasoning, C_i is a structured tool call (operation + parameters), and V_i is the returned visual evidence. At round k , the LMM consumes the history $\{T_i, C_i, V_i\}_{i=0}^k$ and outputs $O_k = f_{\text{LMM}}(\cdot)$, which is parsed into the next reasoning-tool tuple $(T_{k+1}, C_{k+1}) = p(O_k)$; the toolbox executes C_{k+1} and yields $V_{k+1} = g_{\text{tool}}(C_{k+1})$. For each sample, we generate $K = 5$ responses with temperature $\tau = 1.0$ to diversify trajectories, and retain only moderately difficult instances by discarding cases where all responses succeed or all fail. This data generation process creates a supervised fine-tuning dataset $\mathcal{D}_{\text{tool}} = \{(V_i, T_i, \tau_i)\}_{i=1}^N$ containing video-question pairs paired with successful tool-usage trajectories. We then perform supervised fine-tuning on this dataset to teach the LMM when and how to leverage the frame-level tools for enhanced video understanding, optimizing the cross-entropy loss $\mathcal{L}_{\text{SFT}} = -\sum_{i=1}^N \log p_{\theta}(\tau_i | V_i, T_i)$ to align the model’s behavior with the demonstrated tool-usage patterns.

Reinforce Tool Using. Unlike CoT, tool response provides accurate and reliable visual evidence without introducing excessive hallucinations. Moreover, SFT alone is insufficient to endow LMMs with the capability to actively explore and leverage visual evidence for enhanced reasoning. Therefore, we further optimize tool utilization through reinforcement learning with our HGRPO algorithm. For each sample, we generate $G = 8$ rollouts and compute rewards $R(\tau) = R_{\text{acc}}(\tau) + R_{\text{hall}}(\tau) + R_{\text{tool}}(\tau)$. The accuracy reward $R_{\text{acc}}(\tau)$ measures how well the model’s final answer matches the ground truth. The hallucination penalty $R_{\text{hall}}(\tau)$ detects and penalizes reasoning steps that cannot be fully supported by the corresponding visual evidence. We employ an ensemble of LMMs as hallucination evaluators that collectively assess whether each reasoning step can be completely justified by its associated video segment. The evaluation process uses majority voting among multiple LMMs to determine hallucination instances, computing: $R_{\text{hall}}(\tau) = -\alpha \sum_{i=1}^{|S|} \mathbb{I} \left[\frac{1}{M} \sum_{j=1}^M h_j(s_i, v_i) > 0.5 \right]$, where s_i represents the i -th reasoning step, v_i is its corre-

sponding extracted video segment, M is the number of evaluator LMMs, $h_j(s_i, v_i) \in \{0, 1\}$ indicates whether the j -th evaluator judges step s_i as hallucinated given visual evidence v_i , and α controls the penalty strength. The indicator function triggers when the majority of evaluators flag a reasoning step as unsupported by its visual context. The tool reward $R_{\text{tool}}(\tau)$ incentivizes successful tool calls that retrieve relevant visual evidence to ground the reasoning. LMM is optimized by:

$$J_{\text{HGRPO}}(\theta) = \mathbb{E}_{(v,x) \sim \mathcal{D}, \{\tau^k\}_{k=1}^G \sim \pi_{\theta_{\text{old}}}} \left[\frac{1}{G} \sum_{k=1}^G \frac{\pi_{\theta}(\tau^k | v, x)}{\pi_{\theta_{\text{old}}}(\tau^k | v, x)} A_k - \beta D_{\text{KL}}(\pi_{\theta} \| \pi_{\text{ref}}) \right] \quad (2)$$

where advantages A_k are computed from these hallucination-aware rewards, enabling the model to learn when and how to leverage tools for reducing hallucinations in video understanding.

5.3 Experiment Result

Main Result. We adopt the 3B and 7B versions of Qwen2.5-VL-Instruct as our base models to evaluate the proposed reasoning strategies. As shown in Table 6, compared with the base models, video-based reasoning strategy achieves 6.74% and 10.16% improvement for the 3B and 7B versions respectively. Table 7 presents the consistency evaluation results measured by bias scores, where video-based reasoning strategy achieves 0.28% and 4.89% bias score reduction compared with the base models. Notably, video-based reasoning strategy consistently outperforms text-based reasoning strategy across both metrics and model sizes. Both proposed methods demonstrate substantial superiority over CoT. Furthermore, video-based reasoning strategy also maintains a competitive advantage over API models like GPT-4o-mini.

Ablation on Text-based Reasoning. To investigate the individual contributions of different components in our text-based reasoning strategy, we conduct ablation studies examining the necessity of both RFT and PL stages. As shown in Table 8, applying only RFT without PL achieves lower performance than the full strategy. This demonstrates that while RFT equips models with reasoning capabilities, the PL stage is essential for eliminating hallucinated content within the reasoning process through segment-weighted preference optimization.

Model/Type	Binary-choice	Multiple-choice	Total
GPT-4o-mini	34.08	29.95	31.69
GPT-4o	35.46	31.54	33.19
OpenAI o1	32.36	23.14	27.03
GPT-5.1	8.51	30.77	13.33
PLLaVA	61.14	42.51	50.38
+CoT	48.32	34.56	40.38
LLaMA-VID-7B	52.15	40.20	45.25
+CoT	65.70	38.86	50.20
LLaVA-NeXT-Video-DPO-7B	51.13	41.41	45.52
+CoT	56.84	36.58	45.15
Qwen2.5-VL-3B-Instruct	41.67	17.39	27.66
+CoT	43.46	23.42	31.89
+Text Reasoning 🔥	40.12	21.26	29.23
+Video Reasoning 🔥	39.41	18.56	27.38
Qwen2.5-VL-7B-Instruct	48.50	19.90	31.99
+CoT	45.14	18.86	29.97
+Text Reasoning 🔥	42.57	18.04	28.41
+Video Reasoning 🔥	39.94	17.69	27.10

Table 7: Consistency evaluation of base models trained by two strategies vs. other LMMs.

Method	Stage	Accuracy ↑	Bias Score ↓
Qwen2.5-VL-7B-Instruct	Base	50.19	31.99
Text Reasoning	+RFT	55.26	32.28
	+RFT&DPO	58.83	28.41
Video Reasoning	+Toolbox	52.41	33.26
	+SFT	55.58	30.85
	+SFT&RL	60.35	27.10

Table 8: Ablation of both reasoning strategies.

Ablation on Video-based Reasoning. We conduct ablation studies on our video-based reasoning approach to examine the effectiveness of supervised tool usage learning and HGRPO. As shown in Table 8, using only supervised tool usage learning achieves limited improvements compared with the complete pipeline, confirming that supervised learning establishes fundamental tool usage patterns while HGRPO refines these behaviors through hallucination-aware rewards, making both components indispensable. Interestingly, simply providing tool usage without training can yield measurable performance gains, highlighting the inherent value of tool availability in our strategy.

6 Conclusion

In this study, we present HAVEN, a comprehensive benchmark for evaluating hallucinations in video understanding for LMMs, featuring 6,497 questions across three key dimensions. Extensive evaluation across 18 models reveals insights into how factors such as video/question length, frame sampling and model size collectively impact model hallucinations. Furthermore, We propose two reasoning-based mitigation strategies: (1) text-based reasoning with long-CoT RFT and segment-weighted DPO, and (2) video-based rea-

soning using a tool-augmented reasoning pipeline with HGRPO optimization. Both strategies effectively improve hallucination evaluation accuracy and reduce bias scores effectively.

Acknowledgments

We thank the Action Editor and the anonymous reviewers for their careful reading, constructive feedback, and helpful suggestions, which improved the quality and clarity of this paper.

References

- Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.
- Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.
- Homanga Bharadhwaj, Roozbeh Mottaghi, Abhinav Gupta, and Shubham Tulsiani. 2024. Track2Act: Predicting point tracks from internet videos enables generalizable robot manipulation. In *European Conference on Computer Vision*, pages 306–324. Springer.
- Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970.
- Sungguk Cha, Jusung Lee, Younghyun Lee, and Choljong Yang. 2024. Visually dehallucinative instruction generation: Know what you don’t know. *arXiv preprint arXiv:2402.09717*.
- Beitao Chen, Xinyu Lyu, Lianli Gao, Jingkuan Song, and Heng Tao Shen. 2024a. **Alleviating hallucinations in large vision-language models through hallucination-induced optimization**.
- 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. 2024b. **Sharegpt4video: Improving video understanding and generation with better captions**. In *Advances in Neural Information Processing Systems*, volume 37, pages 19472–19495. Curran Associates, Inc.
- Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. 2024c. **Unified hallucination detection for multimodal large language models**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3235–3252, Bangkok, Thailand. Association for Computational Linguistics.
- Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024d. **HalC: Object hallucination reduction via adaptive focal-contrast decoding**.
- Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.
- Wey Yeh Choong, Yangyang Guo, and Mohan Kankanhalli. 2026. **VidHal: Benchmarking Hallucinations in Vision LLMs**. *Transactions on Machine Learning Research*.
- Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2024. **Dola: Decoding by contrasting layers improves factuality in large language models**.
- Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. 2023. Holistic analysis of hallucination in GPT-4V (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*.
- Wenliang Dai, Junnan Li, DONGXU LI, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023a. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS*, pages 49250–49267.
- Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. 2023b. Plausible may not be faithful: Probing object hallucination in vision-language pre-training. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL*, pages 2128–2140.
- Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. 2024. **Mmbench-video: A long-form multi-shot benchmark for holistic video understanding**.
- Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. 2025. **Video-r1: Reinforcing video reasoning in mllms**. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 99114–99137. Curran Associates, Inc.
- Laura Fieback, Jakob Spiegelberg, and Hanno Gottschalk. 2024. Metatoken: Detecting hallucination in image descriptions by meta classification. *arXiv preprint arXiv:2405.19186*.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. 2025a. **Mme: A comprehensive evaluation benchmark for multimodal large language models**. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc.

- Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiaowu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. 2025b. **Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis**.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024a. **Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024b. **Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024c. **Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models**.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. 2024d. **HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.
- Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. **Textbooks are all you need**.
- Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. **Detecting and preventing hallucinations in large vision language models**.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 14953–14962.
- Moonsu Han, Minki Kang, Hyunwoo Jung, and Sung Ju Hwang. 2019. **Episodic memory reader: Learning what to remember for question answering from streaming data**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4407–4417, Florence, Italy. Association for Computational Linguistics.
- Zongbo Han, Zechen Bai, Haiyang Mei, Qianli Xu, Changqing Zhang, and Mike Zheng Shou. 2024. **Skip: A simple method to reduce hallucination in large vision-language models**.
- Ailin Huang, Boyong Wu, Bruce Wang, Chao Yan, Chen Hu, Chengli Feng, Fei Tian, Feiyu Shen, Jingbei Li, Mingrui Chen, et al. 2025a. Step-audio: Unified understanding and generation in intelligent speech interaction. *arXiv preprint arXiv:2502.11946*.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025b. **A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions**. *ACM Transactions on Information Systems*, 43(2):1–55.
- Wen Huang, Hongbin Liu, Minxin Guo, and Neil Gong. 2024. **Visual hallucinations of multi-modal large language models**. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9614–9631, Bangkok, Thailand. Association for Computational Linguistics.
- Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. 2024. **Self-introspective decoding: Alleviating hallucinations for large vision-language models**.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.
- Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, pages 1–38.

- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024a. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.
- Chaoya Jiang, Wei Ye, Mengfan Dong, Hongrui Jia, Haiyang Xu, Ming Yan, Ji Zhang, and Shikun Zhang. 2024b. **Hal-Eval: A universal and fine-grained hallucination evaluation framework for large vision language models**. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3152–3162.
- Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024. **FaithScore: Fine-grained evaluations of hallucinations in large vision-language models**. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5042–5063, Miami, Florida, USA. Association for Computational Linguistics.
- Soo Min Kang and Richard P Wildes. 2016. Review of action recognition and detection methods. *arXiv preprint arXiv:1610.06906*.
- Prannay Kaul, Zhizhong Li, Hao Yang, Yonatan Dukler, Ashwin Swaminathan, CJ Taylor, and Stefano Soatto. 2024. Throne: An object-based hallucination benchmark for the free-form generations of large vision-language models. *arXiv preprint arXiv:2405.05256*.
- Junho Kim, Hyunjun Kim, Yeonju Kim, and Yong Man Ro. 2024a. **Code: Contrasting self-generated description to combat hallucination in large multi-modal models**.
- Sihyeon Kim, Boryeong Cho, Sangmin Bae, Sumyeong Ahn, and Se-Young Yun. 2024b. **Vacode: Visual augmented contrastive decoding**.
- Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR.
- Ming Kong, Xianzhou Zeng, Luyuan Chen, Yadong Li, Bo Yan, and Qiang Zhu. 2025. Mhbench: Demystifying motion hallucination in videollms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 4401–4409.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. **Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation**.
- Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2023. **Factuality enhanced language models for open-ended text generation**.
- Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464.
- Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024a. **Inference-time intervention: Eliciting truthful answers from a language model**.
- Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2025a. **Videochat: chat-centric video understanding**. *Science China Information Sciences*, 68(10):200102.
- Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2024b. **Mvbench: A comprehensive multi-modal video understanding benchmark**.
- Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.
- Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023b. **Contrastive decoding: Open-ended text generation as optimization**.
- Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. 2025b. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*.
- Yanwei Li, Chengyao Wang, and Jiaya Jia. 2024c. Llama-vid: An image is worth 2 tokens in large language models.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023d. Evaluating object hallucination in large vision-language models. In *Proceedings of the Empirical Methods in*

- Natural Language Processing, EMNLP*, pages 292–305.
- Bin Lin, Yang Ye, Bin Zhu, Jiayi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. **Video-LLaVA: Learning united visual representation by alignment before projection**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, Miami, Florida, USA. Association for Computational Linguistics.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. **TruthfulQA: Measuring how models mimic human falsehoods**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.
- Fenglin Liu, Tingting Zhu, Xian Wu, Bang Yang, Chenyu You, Chenyang Wang, Lei Lu, Zhangdaihong Liu, Yefeng Zheng, Xu Sun, et al. 2023a. A medical multimodal large language model for future pandemics. *NPJ Digital Medicine*, 6(1):226.
- Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024a. **Mitigating hallucination in large multi-modal models via robust instruction tuning**.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. **Improved baselines with visual instruction tuning**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Proceedings of the Neural Information Processing Systems, NeurIPS*.
- Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. 2025. **PHD: A ChatGPT-prompted visual hallucination evaluation dataset**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28306–28316.
- Shi Liu, Kecheng Zheng, and Wei Chen. 2024c. **Paying more attention to image: A training-free method for alleviating hallucination in vlms**.
- Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. 2024. **Negative object presence evaluation (NOPE) to measure object hallucination in vision-language models**. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 37–58, Bangkok, Thailand. Association for Computational Linguistics.
- Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024a. **Video-ChatGPT: Towards detailed video understanding via large vision and language models**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1114–1124, Bangkok, Thailand. Association for Computational Linguistics.
- Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024b. **Video-chatgpt: Towards detailed video understanding via large vision and language models**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*.
- Ian R McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, et al. 2023. Inverse scaling: When bigger isn’t better. *arXiv preprint arXiv:2306.09479*.
- Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. 2024. **Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities**.
- OpenAI. 2023. Gpt-4v (ision) system card.
- OpenAI. 2025. o3-o4-mini-system-card.
- OpenAI. 2026. **Openai gpt-5 system card**.
- Yeji Park, Deokyeong Lee, Junsuk Choe, and Buru Chang. 2024. **Convis: Contrastive decoding with hallucination visualization for mitigating hallucinations in multimodal large language models**.
- Phuc Phan, Hieu Tran, and Long Phan. 2024. **Distillation contrastive decoding: Improving llms reasoning with contrastive decoding and distillation**.
- Xiaoye Qu, Jiashuo Sun, Wei Wei, and Yu Cheng. 2024. **Look, compare, decide: Alleviating hallucination in large vision-language models via multi-view multi-path reasoning**.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. **The curious case of hallucinations in neural machine translation**.
- Abhilasha Ravichander, Shruti Ghela, David Wadden, and Yejin Choi. 2025. Halogen: Fantastic llm hallucinations and where to find them. *arXiv preprint arXiv:2501.08292*.
- Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

- Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045.
- Fawaz Sammani, Tanmoy Mukherjee, and Nikos Deligiannis. 2022. NLX-GPT: A model for natural language explanations in vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8322–8332.
- Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 14974–14983.
- Lei Shen, Haolan Zhan, Xin Shen, Hongshen Chen, Xiaofang Zhao, and Xiaodan Zhu. 2021. **Identifying untrustworthy samples: Data filtering for open-domain dialogues with bayesian optimization**. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management, CIKM '21*, page 1598–1608. ACM.
- Weijia Shi, Sewon Min, Maria Lomeli, Chunting Zhou, Margaret Li, Gergely Szilvasy, Rich James, Xi Victoria Lin, Noah A. Smith, Luke Zettlemoyer, Scott Yih, and Mike Lewis. 2024. **In-context pretraining: Language modeling beyond document boundaries**.
- Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Yixuan Tang, Shayne Longpre, Duyu Lu, Mitchell Wortsman, Dongxin Jiang, Jie Wang, Lu Wang, Trevor Darrell, Sanja Fidler, Luke Zettlemoyer, Ali Farhadi, Yejin Choi, and Hannaneh Hajishirzi. 2024. **Aligning large multimodal models with factually augmented RLHF**. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13106–13125, Bangkok, Thailand. Association for Computational Linguistics.
- Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. 2019. COIN: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1207–1216.
- Zilu Tang, Rajen Chatterjee, and Sarthak Garg. 2025. Mitigating hallucinated translations in large language models with hallucination-focused preference optimization. *arXiv preprint arXiv:2501.17295*.
- Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.
- Qwen Team. 2025. **Qwen2.5-vl**.
- Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. 2024. DriveVLM: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023. An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation. *arXiv preprint arXiv:2311.07397*.
- Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. 2024a. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. In *International Conference on Multimedia Modeling*, pages 32–45. Springer.
- Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024b. **Mitigating hallucinations in large vision-language models with instruction contrastive decoding**.
- Yueqian Wang, Xiaojun Meng, Jianxin Liang, Yuxuan Wang, Qun Liu, and Dongyan Zhao. 2024c. Hawkeye: Training video-text LLMs for grounding text in videos. *arXiv preprint arXiv:2403.10228*.
- Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. 2024d. Videohalluciner: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv preprint arXiv:2406.16338*.
- Ziyang Wang, Honglu Zhou, Shijie Wang, Junnan Li, Caiming Xiong, Silvio Savarese, Mohit Bansal, Michael S. Ryoo, and Juan Carlos Niebles. 2026. **Active video perception: Iterative evidence seeking for agentic long video understanding**.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

- Jiaheng Wei, Yuanshun Yao, Jean-Francois Ton, Hongyi Guo, Andrew Estornell, and Yang Liu. 2024. Measuring and reducing LLM hallucination without gold-standard answers via expertise-weighting. *arXiv preprint arXiv:2402.10412*.
- Sangmin Woo, Jaehyuk Jang, Donguk Kim, Yubin Choi, and Changick Kim. 2024a. **Ritual: Random image transformations as a universal anti-hallucination lever in large vision language models**.
- Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. 2024b. **Don't miss the forest for the trees: Attentional vision calibration for large vision language models**.
- Junfei Wu, Qiang Liu, Ding Wang, Jinghao Zhang, Shu Wu, Liang Wang, and Tieniu Tan. 2024a. **Logical closed loop: Uncovering object hallucinations in large vision-language models**.
- Kai Wu, Boyuan Jiang, Zhengkai Jiang, Qingdong He, Donghao Luo, Shengzhi Wang, Qingwen Liu, and Chengjie Wang. 2024b. **Noiseboost: Alleviating hallucination with noise perturbation for multimodal large language models**.
- Ziheng Wu, Zhenghao Chen, Ruipu Luo, Can Zhang, Yuan Gao, Zhentao He, Xian Wang, Haoran Lin, and Minghui Qiu. 2025. Valley2: Exploring multimodal models with scalable vision-language design. *arXiv preprint arXiv:2501.05901*.
- Xin Xiao, Bohong Wu, Jiacong Wang, Chunyuan Li, Xun Zhou, and Haoyuan Guo. 2024. **Seeing the image: Prioritizing visual correlation by contrastive alignment**.
- Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. 2024a. **Pillava : Parameter-free llava extension from images to videos for video dense captioning**.
- Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. 2025. PointLLM: Empowering large language models to understand point clouds. In *European Conference on Computer Vision*, pages 131–147. Springer.
- Xiaohan Xu, Chongyang Tao, Tao Shen, Can Xu, Hongbo Xu, Guodong Long, Jian guang Lou, and Shuai Ma. 2024b. **Re-reading improves reasoning in large language models**.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.
- YouTube. <https://www.youtube.com/>.
- Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. 2024a. **Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data**.
- Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwan He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2024b. RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.
- Zihao Yue, Liang Zhang, and Qin Jin. 2024. **Less is more: Mitigating multimodal hallucination from an eos decision perspective**.
- Hang Zhang, Xin Li, and Lidong Bing. 2023a. **Video-LLaMA: An instruction-tuned audio-visual language model for video understanding**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 1–9, Singapore. Association for Computational Linguistics.
- Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Na Zhao, Zhiyu Tan, Hao Li, and Jingjing Chen. 2024a. EventHallusion: Diagnosing event hallucinations in video LLMs. *arXiv preprint arXiv:2409.16597*.
- Shaolei Zhang, Tian Yu, and Yang Feng. 2024b. **Truthx: Alleviating hallucinations by editing large language models in truthful space**.
- Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. 2025. TinyLLaVA-Video-R1: Towards smaller LMMs for video reasoning. *arXiv preprint arXiv:2504.09641*.
- Yi-Fan Zhang, Qingsong Wen, Xue Wang, Weiqi Chen, Liang Sun, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. 2023b. **Onenet: Enhancing time series forecasting models under concept drift by online ensembling**.
- Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangeke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024c. **Llava-next: A strong zero-shot video understanding model**.
- Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023c. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.
- Zheng Zhao, Emilio Monti, Jens Lehmann, and Haytham Assem. 2024. **Enhancing contextual understanding in large language models through contrastive decoding**.
- Gengze Zhou, Yicong Hong, and Qi Wu. 2024. NavGPT: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI, pages 7641–7649.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. 2024a. [Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding](#).

Wenhao Zhu, Sizhe Liu, Shujian Huang, Shuaijie She, Chris Wendler, and Jiajun Chen. 2024b. [Multilingual contrastive decoding via language-agnostic layers skipping](#).

A Annotation Details and Quality Control

Annotation Team. The annotation process was conducted by a team of six annotators with research backgrounds in multimodal large language models (MLLMs), computer vision, and natural language processing. The team consisted of two undergraduate annotators, two graduate annotators (Master’s and Ph.D. students), and two senior annotators who served as adjudicators during the annotation process.

- **Primary Annotators.** The primary annotation team consisted of two undergraduate annotators and two graduate annotators (Master’s and Ph.D. students). All annotators completed the same training program before formal annotation. The undergraduate annotators had prior research experience in multimodal learning, while the graduate annotators each had more than two years of research experience in multimodal learning and prior experience in vision-language dataset construction, annotation, and evaluation.
- **Senior Annotators.** The two senior annotators had extensive experience in multimodal dataset construction, large-scale data quality control, and MLLM research. Throughout the annotation process, they were responsible for reviewing difficult cases, resolving annotation disagreements, adjudicating conflicting labels, and maintaining consistency across hallucination categories and annotation criteria.

Annotator Training. All annotators underwent guideline training and participated in pilot annotation sessions to align their understanding of hallucination definitions, category boundaries, and video-grounding requirements. Representative examples from all hallucination categories were reviewed collectively, and annotators discussed challenging cases to establish consistent annotation standards before large-scale annotation began.

Video Preprocessing. Before annotation, all source videos underwent a preprocessing stage. Videos with severe visual corruption, missing content, duplicated segments, or insufficient visual evidence were removed. To reduce distractions from irrelevant context, each video was temporally trimmed to retain only the segment necessary for answering the target question.

Annotation Procedure. Each example in HAVEN was constructed following a predefined hallucination annotation framework. Annotators first reviewed the video and identified visual evidence corresponding to a target hallucination category and hallucination-aspect category. The identified evidence served as the basis for subsequent question construction.

After grounding the relevant evidence, annotators designed a probing question intended to expose a specific type of hallucination behavior. Questions were required to be answerable solely from the video content and to focus on the targeted hallucination phenomenon. Annotators then manually wrote a reference answer strictly grounded in the observed visual evidence, avoiding the use of external knowledge or subjective interpretation.

Following the initial annotation stage, all question-answer pairs entered a multi-stage quality control pipeline.

Quality Control. To ensure annotation quality and category consistency, every question-answer pair was independently reviewed by at least two annotators.

After the initial annotation, each example was cross-checked by a second annotator, who independently verified the visual evidence, reference answer, hallucination category, and hallucination-aspect category. Reviewers verified the correctness of the visual evidence, reference answer, hallucination category, and hallucination-aspect category. When disagreements were identified during cross-checking, the example was forwarded to a senior annotator for adjudication.

Final labels were determined through discussion and consensus. Examples that remained ambiguous after adjudication, or for which a unique category assignment could not be established with sufficient confidence, were excluded from the benchmark.

Annotation Statistics. The average inter-annotator agreement was 88.3%. During adjudication, 972 examples (13%) were removed because they could not be assigned to a category unambiguously.

B Evaluation Details

Our evaluation is conducted using GPT-4o (Hurst et al., 2024) by comparing human-generated answers with model responses to determine correctness.

For models using CoT, we found that long reasoning process might affect GPT-4o’s judgment in multiple-choice and binary-choice questions. Therefore, we first use GPT-4o to extract a concise answer from the model’s reasoning process based on the given options in the question. Then, we can conduct our evaluation based on the extract answer. For multiple-choice questions, we let GPT-4o directly assess whether the model’s response aligns with the intended meaning of the correct answer based on its reasoning process. Multiple-choice questions are evaluated using the prompt in Table 10, binary-choice questions using the prompt in Table 11, and short-answer questions using the prompt in Table 12.

For models without CoT, responses can be directly compared using GPT-4o. Multiple-choice questions are evaluated using the prompt in Table 14, binary-choice questions using the prompt in Table 13, and short-answer questions using the prompt in Table 15.

C Reliability of Automatic Evaluation

To assess the reliability of GPT-4o as an automatic evaluator, we conduct a human audit on a held-out subset of model responses. The goal is to verify whether GPT-4o judgments are consistent with human assessment when evaluating hallucination-related responses in HAVEN.

Human Audit Setup. We sample responses from three representative MLLM families, namely Qwen, LLaVA-NeXT-Video, and GPT-4o-mini. For each model family, we randomly select 1,200 responses, resulting in a total of 3,600 audited samples. To ensure that the audit subset remains representative of the benchmark, the sampled questions follow the same distribution as the full benchmark. Specifically, binary-choice (BC), multiple-choice (MC), and short-answer (SA) questions are sampled according to the original 5:2:3 ratio.

Human Evaluation Protocol. The sampled responses are independently evaluated by the two graduate annotators. Annotators are instructed to determine whether a model response is semantically consistent with the corresponding reference answer under the same evaluation rubric used by GPT-4o.

For BC and MC questions, responses are considered correct if they clearly express the semantics of the correct option, regardless of the exact wording

used. For SA questions, annotators assess whether the response is factually supported by the video evidence and semantically equivalent to the reference answer.

Disagreements between human annotators are resolved through discussion, and the resulting consensus labels are treated as the reference judgments for reliability analysis.

Agreement with Human Evaluation. Table 9 reports the agreement between GPT-4o and human evaluation. GPT-4o achieves an overall agreement of 96.31% with human evaluation. Agreement remains consistently high across all question formats, reaching 96.78% for BC questions, 98.47% for MC questions, and 94.07% for SA questions. Most disagreements arise from short-answer responses that are partially correct, underspecified, or expressed using uncommon wording.

Overall, these results support the use of GPT-4o as a scalable evaluator for HAVEN while highlighting the importance of answer extraction for reliable assessment of free-form reasoning outputs.

Subset	#Samples	Agreement (%)
Overall	3600	96.31
BC	1800	96.78
MC	720	98.47
SA	1080	94.07

Table 9: Agreement between GPT-4o evaluation and human auditing on the held-out subset. BC, MC, and SA denote binary-choice, multiple-choice, and short-answer questions, respectively.

D Implementation Details of the LMMs

The LMMs evaluated in our experiment are listed as follows:

The open-source models Video-ChatGPT (Maaz et al., 2024b), Valley-Eagle (Wu et al., 2025), Video-LLaVA (Lin et al., 2024), VideoChat2 (Li et al., 2025a), ShareGPT4Video (Chen et al., 2024b), LLaMA-VID (Li et al., 2024c), PLLaVA (Xu et al., 2024a), LLaVA-NEXT-Video-DPO (Zhang et al., 2024c), and Video-LLaMA2 (Cheng et al., 2024), all of which strictly follow the protocol proposed in VideoHalluciner (Wang et al., 2024d): each model uses the *default* frame-sampling rate and input resolution reported in its original paper, and no additional prompt engineering or post-processing is applied to guarantee fair comparison.

In addition, we evaluate both the 3B and 7B variants of Qwen2.5-VL-Instruct ([Team, 2025](#)), which sample 2 frames per second from each video as visual input.

For the API models GPT-4o ([Hurst et al., 2024](#)) and OpenAI o1 ([Jaech et al., 2024](#)), we uniformly sample 32 frames from each video and resize them to 768 pixels while preserving the original aspect ratio, with all other parameters kept at their default settings.

Binary-Choice:

(1) Extracting Answer:

This is a multiple-choice question. Based on the given question and reasoning process, extract the corresponding answer of the reasoning process.

Question: {question[idx]}

Reasoning Process: {res[idx]}

Instructions:

1. Identify the correct answer based on the reasoning process.
2. If the reasoning process explicitly states "yes" or "no", return direct "yes" or "no"
3. If the reasoning process concludes that the correct answer is "no answer" or "I don't know", return "no answer".
5. Return only the final "yes" or "no", without explanation or additional text.

(2) Judging:

You are a professional homework grading tool. I will provide you with four rules for grading

1. This is a yes/no question. The Standard Answer is only Yes/No, please directly compare the standard answer with 'yes' or 'no' in the predicted answer.
2. No matter what kind of questions, only response with one of 1 or 0. No more explanation.
3. 1 means correct (they are the same), 0 means wrong (they are different).

For example, if the prediction conveys the same meaning as the standard answer, you should respond with 1.

Based on the question, is the prediction correct? If yes, only return 1, otherwise only return 0.

Question: {question[idx]}

Standard Answer: {answer[idx]}

The Predicted Answer: {extract_answer}

Table 10: The evaluation prompts for Binary-Choice questions employed in the experiment for models with CoT.

Multiple-Choice:

(1) Extracting Answer:

This is a multiple-choice question. Based on the given question and reasoning process, extract the corresponding answer of the reasoning process.

Question: {question[idx]}

Reasoning Process: {res[idx]}

Instructions:

1. Identify the correct answer based on the reasoning process.
2. If the reasoning process directly mentions one of the given choices (A, B, or C), return the corresponding letter along with the full text of that option (e.g., "A. Option text").
3. If the reasoning process provides an answer that does not explicitly mention A, B, or C, compare its meaning to the given choices and return the best-matching option in the format "Letter. Option text".
4. If the reasoning process concludes that the correct answer is "no answer" or "I don't know", return "no answer".
5. Return only the final answer, without explanation or additional text.
6. Focus more on the final summary sentence.

(2) Judging:

You are a professional homework grading tool. I will provide you with four rules for grading:

1. This is a multiple-choice question. Judge the correctness based on selected letter and actual content of the provided answers.
2. Regardless of the question type, respond only with either 1 or 0, without any additional explanation.
3. 1 means the prediction is correct, and 0 means it is incorrect.
4. If the predicted answer matches the correct answer in meaning, even if it is phrased differently, consider it correct.

For example, if the prediction conveys the same meaning as the standard answer, you should respond with 1.

Based on the question and its standard answer, is the prediction correct? If yes, return only 1; otherwise, return only 0.

Question: {question[idx]}

Standard Answer: {answer[idx]}

The Predicted Answer: {extract_answer}

Table 11: The evaluation prompts for multiple-choice questions employed in the experiment for models with CoT.

Short-Answer:

You are a professional homework grading tool. I will provide you with four rules for grading

1. The Standard Answer is a sentence. Compare the provided predicted answers based on their meaning rather than exact wording. The prediction is correct if it conveys the same intent.
2. If the question asks about identity or species, the predicted answer is correct as long as the core identity or species is correct, even if descriptive adjectives differ.
3. If the question asks about a scene, the predicted answer is correct if the described scene exists in the standard answer or is similar.
4. If the answer indicates that the asked character, object or event is not visible or does not exist, it should be considered as "No answer."
5. Regardless of the question type, respond only with either 1 or 0, without any additional explanation.
6. 1 means correct, and 0 means incorrect.

For example, if the prediction conveys the same meaning as the standard answer, even if phrased differently, you should respond with 1.

Based on the question, is the prediction correct? If yes, return only 1; otherwise, return only 0.

Question: {question[idx]}

Standard Answer: {answer[idx]}

The Predicted Answer: {res[idx]}

Table 12: The evaluation prompts for Short Answer questions employed in the experiment for models with CoT.

Binary-Choice:

You are a professional homework grading tool. I will provide you with four rules for grading:

1. This is a yes/no question. The Standard Answer is only Yes/No, please directly compare the standard answer with 'yes' or 'no' in the predicted answer.
2. No matter what kind of questions, only response with one of 1 or 0. No more explanation.
3. 1 means correct (they are the same), 0 means wrong (they are different).

For example, if the prediction conveys the same meaning as the standard answer, you should respond with 1.

Based on the question, is the prediction correct? If yes, only return 1, otherwise only return 0.

Question: {question[idx]}

Standard Answer: {answer[idx]}

The Predicted Answer: {extract_answer}

Table 13: The evaluation prompts for Binary-Choice questions employed in the experiment for models without CoT.

Multiple-Choice:

You are a professional homework grading tool. I will provide you with four rules for grading:

1. This is a multiple-choice question. Judge the correctness based on the selected letter and actual content of the provided answers.
2. Regardless of the question type, respond only with either 1 or 0, without any additional explanation.
3. 1 means the prediction is correct, and 0 means it is incorrect.
4. If the predicted answer matches the correct answer in meaning, even if it is phrased differently, consider it correct.

For example, if the prediction conveys the same meaning as the standard answer, you should respond with 1.

Based on the question and its standard answer, is the prediction correct? If yes, return only 1; otherwise, return only 0.

Question: {question[idx]}

Standard Answer: {answer[idx]}

The Predicted Answer: {extract_answer}

Table 14: The evaluation prompts for multiple-choice questions employed in the experiment for models without CoT.

Short-Answer:

You are a professional homework grading tool. I will provide you with four rules for grading

1. The Standard Answer is a sentence. Compare the provided predicted answers based on their meaning rather than exact wording. The prediction is correct if it conveys the same intent.
2. If the question asks about identity or species, the predicted answer is correct as long as the core identity or species is correct, even if descriptive adjectives differ.
3. If the question asks about a scene, the predicted answer is correct if the described scene exists in the standard answer or is similar.
4. If the answer indicates that the asked character, object or event is not visible or does not exist, it should be considered as "No answer."
5. Regardless of the question type, respond only with either 1 or 0, without any additional explanation.
6. 1 means correct, and 0 means incorrect.

For example, if the prediction conveys the same meaning as the standard answer, even if phrased differently, you should respond with 1.

Based on the question, is the prediction correct? If yes, return only 1; otherwise, return only 0.

Question: {question[idx]}

Standard Answer: {answer[idx]}

The Predicted Answer: {res[idx]}

Table 15: The evaluation prompts for Short Answer questions employed in the experiment for models without CoT.

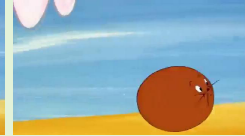
VideoChatGPT				
				
 Question What is the cat holding in the video? A. Sword. B. Fish. C. Mouse.	 Answer A. Sword.	 Response The cat is holding a mouse in the video.	 Extracted C.Mouse	
				
 Question What is the color of the man's car?	 Answer I don't know/No answer	 Response The color of the man's car is black.	 Extracted Black	
Videollama-7B				
				
 Question What item did the man immerse in the water?	 Answer Alarm clock	 Response The man immersed a glass of water in the water.	 Extracted Water	
				
 Question Is the object in the video a brown balloon?	 Answer No	 Response Yes, the object in the video is a brown balloon.	 Extracted Yes	
LlamaVid-7B				
				
 Question Who is holding a bouquet of flowers? A. Mouse. B. Blue cat. C. Boy.	 Answer B.Blue cat	 Response A. Mouse	 Extracted A. Mouse	
				
 Question Did the poker tower on the table in the video collapse?	 Answer No	 Response Yes, the poker tower on the table in the video collapsed.	 Extracted Yes	

Figure 11: Case of response from evaluated baseline LMMs.

PLLaVA				
				
 Question What is placed on the wooden table? A. Bookshelf. B. Cup. C. Lamp tube.	 Answer C. Lamp tube.	 Response A	 Extracted A. Bookshelf.	
				
 Question Which is larger in the video? A. The person. B. The airplane.	 Answer A. The person.	 Response B	 Extracted B. The airplane.	
Video-LLaVA				
				
 Question Is there an individual in a wheelchair?	 Answer No	 Response Yes, there is an individual in a wheelchair present in the video.	 Extracted Yes	
				
 Question What did the man place in the suitcase? A. Clothes. B. Suitcase. C. Shoes.	 Answer B. Suitcase	 Response A. Clothes.	 Extracted A. Clothes.	
VideoChat2				
				
 Question What did the man wear on his hands? A. Bag. B. Socks. C. Gloves.	 Answer B. Socks.	 Response The man in the video wore gloves on his hands.	 Extracted C. Gloves.	
				
 Question What is the item positioned on the boy's head?	 Answer A pot.	 Response The thing that is placed on the boy's head is a hat.	 Extracted A hat.	

Figure 12: Case of response from evaluated baseline LMMs.