VIDHAL: BENCHMARKING TEMPORAL HALLUCINATIONS IN VISION LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision Large Language Models (VLLMs) are widely acknowledged to be prone to hallucinations. Existing research addressing this problem has primarily been confined to image inputs, with sparse exploration of their video-based counterparts. Furthermore, current evaluation methods fail to capture nuanced errors in generated responses, which are often exacerbated by the rich spatiotemporal dynamics of videos. To address these two limitations, we introduce VIDHAL, a benchmark specially designed to evaluate video-based hallucinations in VLLMs. VIDHAL is constructed by bootstrapping video instances across a wide range of common temporal aspects. A defining feature of our benchmark lies in the careful creation of captions which represent varying levels of hallucination associated with each video. To enable fine-grained evaluation, we propose a novel caption ordering task requiring VLLMs to rank captions by hallucinatory extent. We conduct extensive experiments on VIDHAL and comprehensively evaluated a broad selection of models, including both open-source and proprietary ones such as GPT4.1 and Gemini 2.5. Our results uncover significant limitations in existing VLLMs regarding video-based hallucination generation. Through our benchmark, we aim to inspire further research on I) holistic understanding of VLLM capabilities, particularly regarding hallucination, and II) advancing VLLMs to alleviate this problem.

1 INTRODUCTION

Building on the advancements of Large Language Models (LLMs), Vision LLMs (VLLMs) have recently gained significant attention. Models such as LLaVA (Liu et al., 2023; 2024c) have shown impressive performance across various visual understanding tasks involving both images and videos. Despite their potential, VLLMs are notably prone to hallucinations, where generated responses appear plausible but contradict visual context (Bai et al., 2024; Xu et al., 2024). This problem significantly compromises the reliability of VLLMs, hindering their practical use in real-world applications.

To tackle this challenge, some methods propose to leverage post-hoc techniques such as contrastive decoding (Leng et al., 2024; Zhu et al., 2024c; Favero et al., 2024; Zhuang et al., 2024) and attention calibration (Huang et al., 2024; Ma et al., 2024; Liu et al., 2024f; Yue et al., 2024; Gong et al., 2024; Zhou et al., 2024a; Xing et al., 2024b). Other efforts have been devoted to the evaluation of hallucinations in VLLMs. For example, CHAIR (Rohrbach et al., 2018) initially studies object-based hallucination evaluation with the aid of the image captioning task. Subsequent studies (Li et al., 2023e; Liu et al., 2024e; Kaul et al., 2024; Ding et al., 2024) instead harness paired ⟨*positive, hallucinatory*⟩ questions to probe such hallucinations. Additionally, MMHalBench (Sun et al., 2024) and AMBER (Wang et al., 2023) expand beyond object-based evaluations by constructing benchmarks that cover attribute and relationship hallucinations.

Unlike their image-based counterparts, video hallucinations pose unique challenges primarily due to the intricate spatiotemporal dynamics of videos (Fu et al., 2024; Liu et al., 2024g; Ning et al., 2023). In particular, video-specific temporal aspects, such as movement direction and chronological order of events, are especially concerning for video-based VLLMs. Furthermore, the richness of video content necessitates a finer-grained understanding, making VLLMs more vulnerable to nuanced hallucinations. Nonetheless, to the best of our knowledge, video-based hallucinations remain underexplored in the existing literature.

To address this research gap, we present VIDHAL, a benchmark specifically designed to evaluate video-based hallucinations of VLLMs. VIDHAL features videos that comprehensively cover a broad range of temporal aspects, such as entity actions and sequence of events. Each video is automatically annotated with multiple captions exhibiting *varying levels* of aspect-specific hallucinations, capturing both subtle and significant discrepancies. In addition, we perform detailed human validation to ensure the robustness and reliability of our annotation process. An additional motivation stems from the limited metrics for quantifying hallucinations in VLLMs. To capture fine-grained hallucinatory errors of these models, we propose a unique caption ordering task that requires models to rank captions by hallucination levels. This consequently leads to a ranking-based NDCG metric and an MCQA accuracy metric, both are distinct from prior ones and specifically tailored to evaluate nuanced hallucinations in video-based VLLMs.

Using our VIDHAL dataset, we benchmark thirteen VLLMs including both open-sourced and proprietary models, with abstracted results summarized in Figure 1. Through these extensive experiments, we identify limitations in nuanced video understanding among all evaluated VLLMs. Specifically, our findings reveal that existing VLLMs struggle to differentiate between captions with varying levels of hallucination. This deficiency is particularly evident when evaluating video-specific aspects, such as *Direction* and *Order*, as illustrated in Figure 1, indicating substantial room for improvement in current video-based VLLMs.

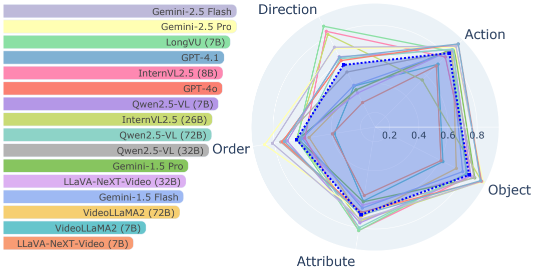

Figure 1: Multiple-Choice Question Answering (MCQA) performance of representative VLLMs on our VIDHAL benchmark. (Left) Overall ranking of VLLMs. (Right) Detailed accuracy results for each temporal aspect, where higher scores indicate fewer hallucinations.

The contributions of this work are three-fold:

- We present VIDHAL, a benchmark dataset dedicated to video-based hallucination evaluation of VLLMs. Our dataset is distinguished by i) video instances encompassing a diverse range of temporal concepts and ii) captions with varying hallucination levels[1].
- We introduce a novel evaluation task of caption ordering along with two metrics designed to evaluate fine-grained hallucination generation in existing VLLMs.
- We conduct extensive experiments on VIDHAL with a variety of VLLMs, uncovering limitations in their fine-grained video reasoning abilities, particularly in their tendency to generate hallucinations.

2 RELATED WORK

Vision Large Language Models. The emergence of powerful LLMs has advanced the development of VLLMs. Typical methods in this category include LLaVA (Liu et al., 2023), MiniGPT-4 (Zhu et al., 2024a), InstructBLIP (Dai et al., 2023), and Qwen-VL (Wang et al., 2024a; Bai et al., 2025). These VLLMs rely on aligning vision encoders with LLMs using connective modules such as Q-Former (Dai et al., 2023; Zhang et al., 2023; Cheng et al., 2024) or MLPs (Liu et al., 2024c; Su et al., 2023) with the instruction tuning stage. Recent methods have extended visual inputs from images to (long) videos, delivering impressive joint spatial-temporal reasoning capabilities. For instance, VideoLLaMA2 (Cheng et al., 2024) enhances the LLaMA model with video understanding capabilities through a Spatial-Temporal Convolution (STC) module. LLaVA-NeXT-Video (Liu et al., 2024d; Zhang et al., 2024) presents an AnyRes approach that enables reasoning with long videos.

Hallucinations in VLLMs. Despite their impressive performance on visual reasoning benchmarks, current VLLMs remain notoriously susceptible to hallucinations (Jiang et al., 2024; Liu et al., 2024f; Zhu et al., 2024b; Chen et al., 2024a). A common demonstration is that the generated responses contain information which is inconsistent with the visual content (Liu et al., 2024b; Yuan et al., 2024; Xing et al., 2024a). Most approaches address the hallucination problem with post-hoc techniques. For example, LURE (Zhou et al., 2024c) and Woodpecker (Yin et al., 2023) develop pipelines that assist VLLMs in revising their responses using expert models. To reduce bias from unimodal and statistical

[1] Our VIDHAL dataset will be made available to the public.

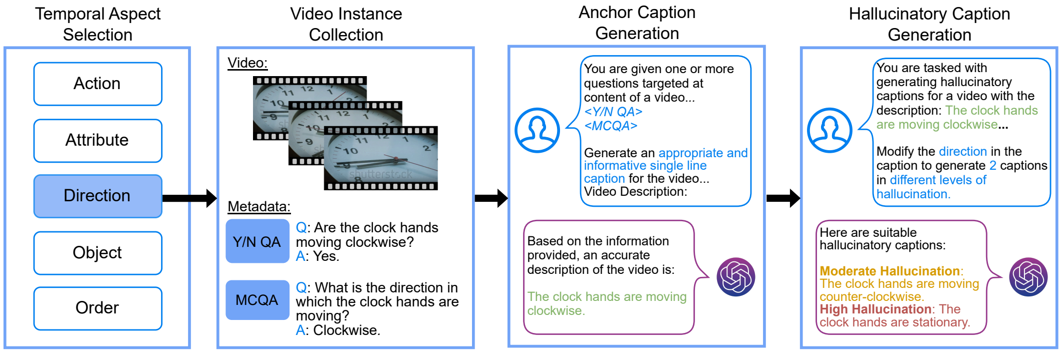
Temporal Aspect Selection
Action
Attribute
Direction
Object
Order
Video Instance Collection
Video:
Metadata:
Y/N QA
Q: Are the clock hands moving clockwise?
A: Yes.
MCQA
Q: What is the direction in which the clock hands are moving?
A: Clockwise.
Anchor Caption Generation
You are given one or more questions targeted at content of a video...
<Y/N QA>
<MCQA>
Generate an appropriate and informative single line caption for the video...
Video Description:
Based on the information provided, an accurate description of the video is:
The clock hands are moving clockwise.
Hallucinatory Caption Generation
You are tasked with generating hallucinatory captions for a video with the description: The clock hands are moving clockwise...
Modify the direction in the caption to generate 2 captions in different levels of hallucination.
Here are suitable hallucinatory captions:
Moderate Hallucination: The clock hands are moving counter-clockwise.
High Hallucination: The clock hands are stationary.

Figure 2: Overview of our VIDHAL benchmark construction pipeline. Using *direction* as an example from the five selected aspects, we begin by sourcing relevant video instances from existing datasets. Next, the anchor (positive) caption is generated from the original video metadata. Finally, GPT-4o is employed to generate hallucinatory captions at varying levels.

priors, contrastive decoding methods, such as VCD (Leng et al., 2024) and M3ID (Favero et al., 2024), along with attention calibration techniques like OPERA (Huang et al., 2024) are employed to refine token predictions. Building on the success of reinforcement learning in LLM development (Ouyang et al., 2022), HA-DPO (Zhao et al., 2023), POVID (Zhou et al., 2024b) and CSR (Zhou et al., 2024d) adopt this paradigm to fine-tune VLLMs, yielding outputs with fewer hallucinations.

Video Reasoning Benchmarks. The rise of video-based VLLMs has driven the development of numerous video benchmarks. Notable examples, such as SEEDBench (Li et al., 2023a), VideoBench (Ning et al., 2023), MVBench (Li et al., 2024b), and VideoMME (Fu et al., 2024), focus on dynamic events requiring temporal reasoning beyond individual frames. However, these benchmarks often lack diversity in reasoning tasks and visual concepts. To address this, AutoEval-Video (Chen et al., 2023) and Perception Test (Patraucean et al., 2023) introduce complex reasoning tasks such as counterfactual and explanatory reasoning, while TempCompass (Liu et al., 2024g) expands temporal concept coverage. Several benchmarks (Li et al., 2023e; Wang et al., 2023; Sun et al., 2024; Kaul et al., 2024; Liu et al., 2024a; Wei et al., 2024; Chen et al., 2024b) have been constructed to quantify visual hallucinations, primarily targeting object-based hallucinations in images. HallusionBench (Guan et al., 2024), VideoCon (Bansal et al., 2024), and Vript (Yang et al., 2024) provides partial coverage of video-based hallucinations, while VidHalluc (Li et al., 2024a) and VideoHallucer (Wang et al., 2024b) introduces benchmarks for hallucination detection in videos. However, these benchmarks provide limited coverage of spatio-temporal concepts, focusing on conventional aspects like actions while neglecting other video-centric elements such as direction. *Additionally, their evaluation strategies primarily follow image-based approaches, which we argue are less effective in capturing nuanced, video-specific hallucinations.*

3 VIDHAL DATASET CONSTRUCTION

We introduce VIDHAL, a unique video-language benchmark designed to evaluate hallucinations of Video-LLMs in a comprehensive manner. As depicted in Figure 2, VIDHAL comprises of video instances which span a diverse spectrum of temporal aspects, including previously unexplored aspects such as directional movement. In contrast to previous studies on video hallucination evaluation (Yang et al., 2024; Wang et al., 2024b; Li et al., 2024a), VIDHAL incorporates multiple hallucinated captions per video, enabling the assessment of video hallucinations at multiple levels of granularity.

3.1 TEMPORAL HALLUCINATIONS IN VIDEOS

Hallucinations in VLLMs occur when the model fabricates details in its responses that contradict the provided visual content. Compared to images, video hallucinations extend beyond static visual elements to include misperceptions of dynamic changes within scenes. We categorize these temporal hallucinations into two semantic levels:

Lexical Semantics (L-Sem) captures instances where VLLMs misinterpret words related to temporal features, including nouns referring to objects or attributes (e.g., misidentifying a color change from green to red as green to orange) and verbs describing actions (e.g., interpreting "kicking a ball" as "throwing a ball").

Clause Semantics (C-Sem) encompasses errors involving event descriptions and their sequences, where the VLLM incorrectly predicts the order of events occurring in the video. For example, given sequentially occurring events A and B in a video, the model may perceive B preceding A.

By addressing these two dimensions of video-based hallucinations, VIDHAL offers holistic coverage over the level of detail in which VLLMs may hallucinate.

3.2 TEMPORAL CONCEPT SELECTION

Prior research on hallucination evaluation for both images (Li et al., 2023e; Wang et al., 2023; Rohrbach et al., 2018) and videos (Wang et al., 2024b; Yang et al., 2024; Guan et al., 2024) has predominantly focused on common visual aspects such as action- and object-based hallucinations. However, video-based hallucinations may involve additional dynamic factors associated with spatio-temporal patterns, which these studies overlook. In light of this, we propose to focus on the following five aspects to ensure comprehensive coverage of temporal concepts. Specifically, the first four aspects address hallucinations based on lexical semantics, while the fifth targets clause semantics.

- **Attribute (L-Sem)** describes the fine-grained characteristics of objects or subjects in the video. We additionally categorize this aspect into sub-aspects of *Size*, *Shape*, *Color*, *Count* and *State Change*.
- **Object (L-Sem)** relates to the interactions between objects and entities within the video. We further delineate this aspect into two fine-grained sub-aspects: *Object Recognition*, identifying the objects engaged in interactions, and *Interaction Classification* which concentrate on how these objects interact with other objects or subjects.
- **Action (L-Sem)** refers to the movements and behaviours exhibited by entities.
- **Direction (L-Sem)** indicates the orientation and movement trajectory of subjects or objects.
- **Event Order (C-Sem)** represents the correct sequence of events in the video. During our collection, we retain videos that contain at least three distinct events.

We present an example that illustrates the direction aspect in Figure 2, with additional examples available in the supplementary material.

3.3 HALLUCINATORY CAPTION GENERATION

Based on the aspects in Section 3.2, we build our benchmark upon four public video understanding datasets: TempCompass (Liu et al., 2024g), Perception Test (Patraucean et al., 2023), MVBench (Li et al., 2024b) and AutoEval-Video (Chen et al., 2023). TempCompass and MVBench extensively cover all five temporal aspects, while Perception Test and AutoEval-Video highlights human-object interactions and attribute changes, respectively.

Existing hallucination benchmarks (Li et al., 2023e; Wang et al., 2023) rely mostly on binary questions for evaluation, limiting their efficacy in detecting subtle video hallucinations, such as minor event inconsistencies. To address this issue, we advocate a novel evaluation protocol incorporating several carefully annotated captions. Specifically, each video will be annotated with M captions that reflect varying degrees of hallucination in VLLMs. Given the cost and labor intensity of manual annotation, we follow existing benchmark studies such as PhD (Liu et al., 2024e) and MVBench (Li et al., 2024b), opting for automatic caption generation using a carefully designed pipeline illustrated in Figure 2.

Anchor Caption Generation. The video instances in VIDHAL are sourced from various public datasets, resulting in distinct associated metadata such as long-form captions in AutoEval-Video and question-answer pairs in MVBench. To ensure structure consistency and information granularity in the respective dataset description across all instances, we automatically generate an anchor caption for each video. Specifically, we input the metadata for each video V^i into GPT-4o and prompt it to generate a concise and accurate description y^i_+ using the provided metadata information.

	Dataset	Temporal Aspects										Task Formats	Evaluation Metrics
		Action	Attribute					Direction	Object		Order		
			Size	Shape	Color	Count	State-Change		Recognition	Interaction			
Video Reasoning	SEEDBench (Li et al., 2023a)	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓	MCQA	Accuracy
	VideoBench (Ning et al., 2023)	✓	✓	✓	✓	✓	✗	✗	✓	✗	✗	MCQA	Accuracy
	MVBench (Li et al., 2024b)	✓	✗	✗	✗	✗	✗	✓	✓	✓	✓	MCQA	Accuracy
	Video-MME (Fu et al., 2024)	✓	✓	✓	✓	✓	✗	✗	✓	✗	✗	MCQA	Accuracy
Hallucination Evaluation	Vript (Yang et al., 2024)	✓	✗	✗	✗	✗	✗	✗	✓	✓	✓	Video Captioning Event Ordering	F1 Score Accuracy
	VideoCon (Bansal et al., 2024)	✓	✓	✓	✓	✓	✗	✗	✓	✗	✓	VL Entailment	ROC-AUC
	HallusionBench (Guan et al., 2024)	✓	✗	✗	✗	✗	✗	✓	✗	✗	✓	Y/N QA	Accuracy
	VIDHAL (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	MCQA Caption Ordering	Accuracy NDCG

Table 1: Comparison of our benchmark dataset with existing video-based reasoning and hallucination evaluation datasets. For datasets with multiple evaluation tasks, only those relevant to hallucination evaluation are included. VL Entailment denotes the task of *video-language entailment*, while *Event Ordering* prompts the model to determine the chronological sequence of scenes in a video.

Hallucinatory Caption Generation. After obtaining the positive caption for each video instance, we augment the dataset with $M - 1$ additional captions containing hallucinated content. For a given video instance V^i, we construct a set $\mathcal{Y}^i_- = \{y^{i,1}_-, \cdots, y^{i,M-1}_-\}$ containing captions with different levels of hallucination based on the temporal concepts associated with it. Specifically, $y^{i,k}_-$ exhibits heavier hallucination than $y^{i,j}_-$ for caption hallucination degree $j < k$. We leverage GPT-4o to generate $\mathcal{Y}^i_-$ by combining the anchor caption y^i_+ and prompting it to create $y^{i,1}_-, \cdots, y^{i,M-1}_-$ progressively in increasing levels of hallucination. The set of captions associated with V^i is then defined as $\mathcal{Y}^i \leftarrow \{y^i_+\} \bigcup \mathcal{Y}^i_-$ consisting of both the anchor and hallucinatory captions.

3.4 Dataset Statistics and Human Validation

Using our automatic annotation pipeline, our VIDHAL benchmark consists of a total of 1,000 video instances each tagged with $M = 3$ captions. As shown in Table 1, our VIDHAL dataset stands out from other video understanding (Li et al., 2023a; Ning et al., 2023; Li et al., 2024b; Fu et al., 2024) and hallucination benchmarks (Guan et al., 2024; Liu & Wan, 2023) in terms of two dimensions: I) VIDHAL encompasses a diverse range of video-centric temporal aspects; and II) We introduce a novel caption ordering task along with two tailored metrics to capture subtle hallucinations previously ignored by paired questions.

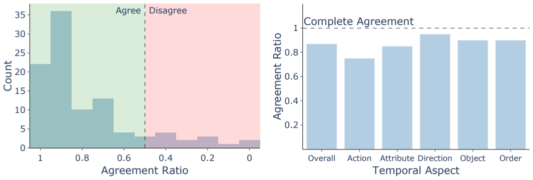

Figure 3: Human agreement on hallucination levels in the VIDHAL dataset. (Left) Distribution of agreement ratios per video sample. (Right) Average agreement ratio for each aspect, with an overall average of 87%.

To ensure the reliability of the generated captions, we randomly selected 100 examples for human validation, with each sample labeled by 15 annotators on average. Our human validation process focuses on verifying that the order of hallucinatory captions generated by our pipeline aligns with human judgment. Figure 3 reflects an overall agreement rate of 87%, indicating consistency with human preferences across all temporal aspects.

4 VidHal Evaluation Protocol

To address the limitations of binary question-based benchmarks, we propose two evaluation tasks: *multiple-choice question answering* and a novel *caption ordering* task, detailed in Section 4.1. We also develop corresponding metrics to comprehensively measure hallucinations in video-based VLLMs, elaborated further in Section 4.2.

4.1 Evaluation Tasks

Multiple-Choice Question Answering (MCQA) assesses the model's spatiotemporal understanding in a coarse-grained manner. Specifically, the VLLM is provided with a video V^i and its corresponding set of captions $\mathcal{Y}^i$ as answer options and instructed to select the most appropriate caption for the video.

Caption Ordering evaluates a model's visual reasoning from a nuanced granularity, instructing VLLMs to order the provided captions based on their hallucination level. Through pairwise comparisons across all captions, this task identifies cases where the model struggles to distinguish varying levels of hallucination severity beyond anchor-hallucination distinctions.

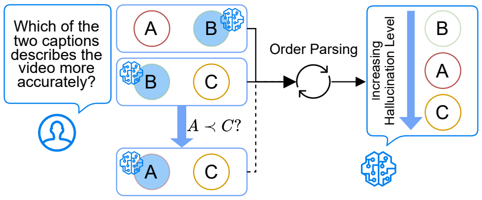
Which of the two captions describes the video more accurately?
A
B
B
C
$A \prec C$?
A
C
Order Parsing
Increasing Hallucination Level
B
A
C

Figure 4: Visual illustration of *relative caption ordering* task in VIDHAL.

Specifically, we design two caption ordering sub-tasks. The first, *naive caption ordering*, requires VLLMs to rank all captions at once. However, this sub-task can confuse several VLLMs due to its inherently challenging nature and the inferior instruction-following capabilities of some models. As a complement, we propose an additional sub-task, *relative caption ordering*, which decomposes the prior task into multiple paired caption ordering tasks. Since each paired ordering task is answered in isolation, the VLLM may produce a non-transitive, cyclic ranking. To circumvent this, we query the model with consecutive caption pairs, prompting the final pair only if multiple orderings are possible. For instance, given captions A, B, and C, if the model predicts $A \prec B$ and $B \prec C$, the overall order $A \prec B \prec C$ can be directly inferred. However, if it instead ranks $B \prec A$, as shown in Figure 4, we additionally include a third comparison between A and C to resolve any ambiguity in determining in the final order.

Notably, our relative caption ordering task is more challenging than previous binary questions. This complexity arises from certain paired questions in VIDHAL where both options are hallucinatory, making them harder to distinguish as opposed to ⟨*positive, hallucinatory*⟩ pairs.

4.2 Evaluation Metrics

Notations For a particular video instance V^i, we define the ground truth caption order for V^i to be $\mathcal{Y}_*^i = (y_+^i, y_-^{i,1}, \cdots, y_-^{i,M-1})$. Further let the j^{th} element in this ordering be indexed as $\mathcal{Y}_*^{i,j}$.

MCQA We employ the standard accuracy metric:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[R_{MCQA}(V^i, \mathcal{Y}^i) = y_+^i\right], \tag{1}$$

where N is the number of video instances, $\mathbb{I}$ denotes the indicator function, and $R_{MCQA}(V^i, \mathcal{Y}^i)$ represents the best matched caption from $\mathcal{Y}^i$ for V^i as predicted by a VLLM.

Caption Ranking Inspired by metrics from the information retrieval domain (Gao et al., 2023), we adapt the well-established Normalized Discounted Cumulative Gain (NDCG) (Järvelin & Kekäläinen, 2002) for hallucination assessment in VIDHAL. Unlike previous metrics like POPE (Li et al., 2023e), our metric awards partial credit for correctly ordered caption pairs even when the optimal ranking is not achieved. As such, we expect the metric to effectively capture and distinguish both subtle and severe hallucinations generated by video-based VLLMs. Formally, we define our adapted NDCG metric as follows:

$$\text{NDCG} = \frac{1}{N}\sum_{i=1}^{N} \frac{\text{DCG}_i - \text{rDCG}_i}{\text{iDCG}_i - \text{rDCG}_i}, \tag{2}$$

where DCG_i is formulated as:

$$\text{DCG}_i = \sum_{j=1}^{M} \frac{r\left(\hat{y}^{i,j}, \mathcal{Y}_*^i\right)}{\log(j+1)}, \tag{3}$$

and $\hat{y}^{i,j}$ represents j^{th} caption in the ranked order predicted by the VLLM. The perfect ordering is achieved when $\hat{y}^{i,1} = y_+^i$ and $\{\hat{y}^{i,j} = y_-^{i,j-1}\}_{j=2 \to M}$. To evaluate predicted caption orders relative to this ideal sequence, a relevance function $r\left(\hat{y}^{i,j}, \mathcal{Y}_*^i\right)$ is designed to assign higher scores to $\hat{y}^{i,j}$ with lower hallucinatory extent.

$$r(\hat{y}^{i,j}, \mathcal{Y}_*^i) = M + 1 - \text{pos}(\hat{y}^{i,j}, \mathcal{Y}_*^i), \tag{4}$$

Model	Vision Encoder	LLM	#Params	#Frames	Accuracy	NDCG	
						Naive	Relative
Baseline							
Random	-	-	-	-	0.326	0.505	0.480
Open-Sourced Models							
VideoChat	EVA-CLIP-G	Vicuna	7B	8	0.381	0.475	0.488
LLaMA-VID	EVA-CLIP-G	Vicuna	7B	1fps	0.358	0.486	0.521
VideoChat2 (Vicuna)	UMT-L	Vicuna	7B	16	0.426	0.486	0.577
VideoChat2 (Mistral)	UMT-L	Mistral	7B	16	0.443	0.503	0.475
VideoChat2 (Phi)	UMT-L	Phi3	3.8B	16	0.514	0.626	0.612
mPLUG-Owl3	SigLIP/SO400M	Qwen2	7B	16	0.596	0.641	0.707
LLaVA-NeXT-Video (7B)	SigLIP/SO400M	Vicuna	7B	32	0.509	0.518	0.620
LLaVA-NeXT-Video (32B)	SigLIP/SO400M	Qwen1.5	32B	32	0.663	0.641	0.747
VideoLLaMA2 (7B)	CLIP ViT-L/14	Mistral	7B	8	0.541	0.564	0.622
VideoLLaMA2 (72B)	CLIP ViT-L/14	Qwen2	72B	8	0.647	0.787	0.760
MiniCPM-V 2.6	SigLIP/SO400M	Qwen2	7B	1fps	0.377	0.530	0.523
LongVU	SigLIP/SO400M	Qwen2	7B	1fps	**0.795**	0.453	**0.846**
InternVL2.5 (8B)	InternViT-300M (V2.5)	InternLM2.5	7B	16	0.773	0.475	0.827
InternVL2.5 (26B)	InternViT-6B (V2.5)	InternLM2.5	20B	16	0.742	0.498	0.775
Qwen2.5-VL (7B)	Qwen2.5-ViT	Qwen2.5	7B	1fps	0.76	**0.825**	0.826
Qwen2.5-VL (32B)	Qwen2.5-ViT	Qwen2.5	32B	1fps	0.732	0.811	0.800
Qwen2.5-VL (72B)	Qwen2.5-ViT	Qwen2.5	72B	1fps	0.74	0.807	0.793
Proprietary Models							
GPT-4o	-	-	-	1fps	0.772	0.840	0.826
GPT-4.1	-	-	-	1fps	0.777	0.845	0.834
Gemini-1.5 (Flash)	-	-	-	1fps	0.657	0.738	0.745
Gemini-1.5 (Pro)	-	-	-	1fps	0.671	0.765	0.753
Gemini-2.5 (Flash)	-	-	-	1fps	0.814	0.875	0.860
Gemini-2.5 (Pro)	-	-	-	1fps	0.814	0.876	0.861

Table 2: Benchmark performance of VLLMs on our VIDHAL dataset. #Params refers to the number of parameters of the base LLM used. The best performance for each task is highlighted in **bold** for open-sourced models, and underlined for closed-sourced models.

where $\text{pos}(\hat{y}^{i,j}, \mathcal{Y}^i_*)$ denotes the position of $\hat{y}^{i,j}$ in $\mathcal{Y}^i_*$. Finally, DCG_i is normalized to a range of $[0, 1]$ using iDCG_i and rDCG_i, with a score of 1 indicating perfect alignment of the predicted order with $\mathcal{Y}^i_*$. Specifically, these terms represent the maximum and minimum DCG_i scores obtained from the optimal ordering $\mathcal{Y}^i_*$ and its reverse, respectively,

$$\text{iDCG}_i = \sum_{j=1}^{M} \frac{r\left(\mathcal{Y}^{i,j}_*, \mathcal{Y}^i_*\right)}{\log(j+1)}, \ \text{rDCG}_i = \sum_{j=1}^{M} \frac{r\left(\mathcal{Y}^{i,M-j}_*, \mathcal{Y}^i_*\right)}{\log(j+1)}. \tag{5}$$

5 EXPERIMENTS

5.1 EXPERIMENTAL SETTINGS

Models. We evaluated twenty-three VLLMs from thirteen different model families, including ten open-source models: VideoChat (Li et al., 2023d), LLaMA-VID (Li et al., 2024c), VideoChat2 (Li et al., 2024b), mPLUG-Owl3 (Ye et al., 2024), LLaVA-NeXT-Video (Zhang et al., 2024), VideoLLaMA2 (Cheng et al., 2024), MiniCPM-V (Yao et al., 2024), LongVU (Shen et al., 2024), InternVL2.5 (Chen et al., 2024c) and Qwen2.5-VL (Bai et al., 2025), and two proprietary models: GPT-4o (OpenAI, 2023), GPT-4.1 and Gemini (Reid et al., 2024; Comanici et al., 2025). These models represent a wide variety of architectural designs and training paradigms. Additionally, we included a random baseline that selects and ranks candidate options randomly.

Implementation Details. All experiments were conducted using four NVIDIA A100 40GB GPUs and inference APIs. The input captions in $\mathcal{Y}^i$ were randomized using a fixed, predefined randomization seed across experiments. We adhered to the inference and model hyperparameters outlined in the respective original models, and employed greedy decoding during generation for a fair comparison.

5.2 OVERALL RESULTS

Benchmark Results. We present the overall results of representative VLLMs in Table 2 across both MCQA and caption ordering tasks. We make three key observations from this table:

Competitive Performance of Open-Source Models. Open-source VLLMs achieve performance comparable to proprietary models, particularly on MCQA and relative caption ordering tasks. Notably, LongVU achieves the highest performance among open-source models and surpasses strong proprietary models such as GPT-4o, GPT-4.1, and Gemini-1.5 on these tasks.

Parameter Scale vs. Performance. Among open-source VLLMs, smaller variants (e.g., 7B parameter models) outperform their larger counterparts within the same model family, as observed with InternVL2.5 and Qwen2.5-VL. This suggests that simply increasing model capacity may provide limited benefits for reducing video-based hallucinations in current VLLM development.

Impact of Architecture Design. Model families that achieve high scores across both tasks often incorporate design efforts specifically targeting visual understanding, such as dynamic resolution scaling (InternVL2.5, Qwen2.5-VL) and temporal reduction techniques (LongVU). These findings may suggest that specialized architectural innovations are key factors in mitigating temporal hallucinations.

Aspect-aware Results. Figure 5 highlights the fine-grained, aspect-specific performance of the notable VLLMs. Notably, VLLMs demonstrate substantially stronger results on the *Action* and *Object* aspects compared to others. This can likely be attributed to current visual instruction tuning datasets predominantly emphasizing object-centric recognition and coarse-grained activity classification, potentially encouraging strong reliance on image-based priors when generating predictions. In contrast, these models tend to underperform on temporally nuanced aspects such as direction and event order, which are inherently unique to the video modality.

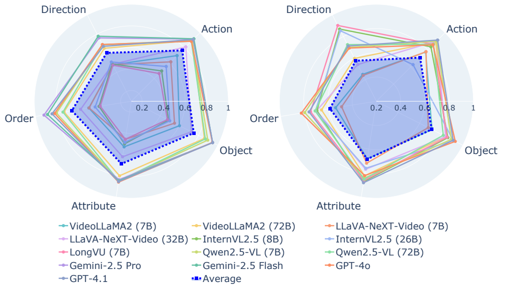

Figure 5: Aspect-specific NDCG scores for the (Left) naive and (Right) relative caption ordering.

We further analyzed the distribution of results for the relative caption ranking task across sub-aspects of the *Attribute* and *Object* aspects in Figure 6. While VLLMs generally maintain consistent performance across *Attribute* sub-aspects, their effectiveness declines slightly when reasoning about *Count* and *Color*, suggesting that reasoning over such fine-grained visual properties remains challenging for VLLMs. For the *Object* aspect, several models performed significantly worse in *Interaction Classification* than in *Object Recognition*, highlighting the need to better model object interactions to bridge the gap between recognition and understanding.

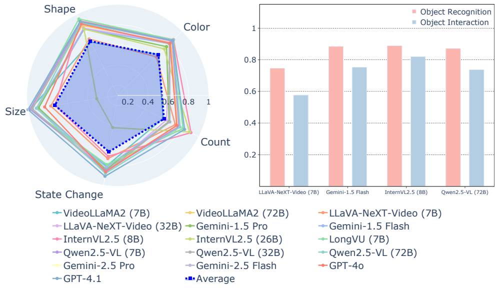

Figure 6: NDCG scores for *Attribute* (Left) and *Object* (Right) sub-aspects in caption ordering.

5.3 Ablation Studies

Hallucination Differentiation Sensitivity. We investigate the tendency of VLLMs to favor captions with higher hallucination over those with lower degree in the relative caption ranking task. For two captions with different hallucination levels j, k where $j > k$, we introduce the following metric to quantify such *hallucination misalignment* cases:

$$HM_{j \to k} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left[\mathcal{Y}_*^{i,j} \prec \mathcal{Y}_*^{i,k} \right]. \tag{6}$$

which reflects the proportion of cases in which the VLLM selects the caption with a higher level of hallucination j over k. Specifically, we examine three key cases: when the most hallucinatory caption is chosen over both the lower-hallucination and anchor captions, and when the lower-hallucination caption is selected over the anchor caption. These cases are represented by $HM_{3 \to 1}$, $HM_{3 \to 2}$, and $HM_{2 \to 1}$, respectively, with results presented in Figure 7.

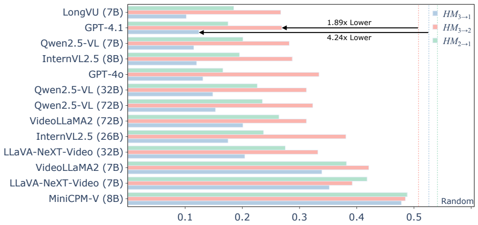

Figure 7: Hallucination misalignment (HM) scores on VIDHAL, with *Random* representing HM scores from the random baseline.

Our findings show that advanced VLLMs, such as VideoLLaMA2 (72B), GPT-4.1 and Qwen2.5-VL models can generally distinguish positive captions from severely hallucinated ones, reflected by their low $HM_{3\to1}$ scores in Figure 7. However, two key observations emerge from our experiments: First, most VLLMs struggle to differentiate the lower hallucinatory caption from the anchor, as evidenced by the gap between $HM_{3\to1}$ and $HM_{2\to1}$. Second, all models exhibit high $HM_{3\to2}$ scores, indicating difficulty in distinguishing between two hallucinatory captions with varying degrees. These results suggest gaps in nuanced video reasoning may contribute to hallucinatory behavior in VLLMs, a challenge not addressed by existing $\langle$*positive*, *hallucinatory*$\rangle$-based evaluation methods. (Li et al., 2023e; Wang et al., 2024b; Guan et al., 2024).

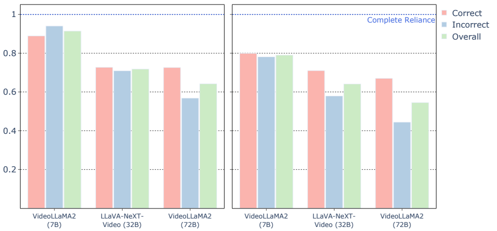

Figure 8: Overlapping ratios of model predictions under single-frame and full-video inputs for correct, incorrect and overall predictions in the (Left) naive and (Right) relative caption ordering tasks. *Complete Reliance* indicates that the VLLM always produces the same response for both video and single frames.

Image Prior Reliance. Previous research shows that VLLMs often rely on image priors for reasoning (Lei et al., 2023; Buch et al., 2022), overlooking key spatiotemporal features. This is exemplified by dominant influence of a few frames on response generation. To examine how this bias affects video-based hallucinations, we used a video summarization algorithm (Son et al., 2024) to extract the most salient frame v^i from V^i. We then generated VLLM responses on VIDHAL using v^i instead of V^i as visual input. The effect of image priors is evaluated by identifying overlapping instances where responses from V^i and v^i remain consistent across both correct and incorrect orderings. As shown in Figure 8, results reveal that VLLMs heavily rely on image priors. This is especially pronounced in smaller models such as VideoLLaMA2 (7B).

6 CONCLUSION

Summary. In this work, we introduce the VIDHAL benchmark to address gaps in the video-based hallucination evaluation of VLLMs. VIDHAL features video instances spanning five temporal aspects. Additionally, we propose a novel caption ordering evaluation task to probe the fine-grained video understanding capabilities of VLLMs. We conduct extensive experiments on VIDHAL through the evaluation of twenty-three VLLMs, exposing their limitations in unexpected hallucination generation. Our empirical results shed light on several promising directions for future work: *e.g.*, incorporating a broader range of temporal features during pretraining and mitigating single-frame priors to enhance temporal reasoning. These advancements will help to address the hallucination problem in video-based VLLMs, enhancing their robustness for real-world video understanding applications.

Limitations. We acknowledge that the VIDHAL evaluation suite relies on synthetic captions generated by GPT-4o, which may contain biases inherently present in the model. We note that this design choice is consistent with prior research, as several established language-only and vision-language benchmarks similarly use GPT-4o for dataset construction (Liu et al., 2024e; Li et al., 2024a;b; 2023a;c) or response evaluation (Guan et al., 2024; Sun et al., 2024; Liu et al., 2024a). To reduce over-alignment to GPT-4o's preferences, we incorporate additional strong LLMs, including Gemini-1.5 (Reid et al., 2024) and LLaMA2 (70B) (Touvron et al., 2023) to assess and filter generated captions. We further conduct a final step of manual verification and editing to address residual misalignments not captured by automated filtering. While these measures enhance annotation robustness, fully eliminating LLM-induced biases in synthetic caption generation remains an open challenge.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *CoRR*, abs/2404.18930, 2024.

Hritik Bansal, Yonatan Bitton, Idan Szpektor, Kai-Wei Chang, and Aditya Grover. Videocon: Robust video-language alignment via contrast captions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13927–13937. IEEE, 2024.

Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the ”video” in video-language understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2907–2917. IEEE, 2022.

Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. Unified hallucination detection for multimodal large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 3235–3252. Association for Computational Linguistics, 2024a.

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

Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David F. Fouhey, and Joyce Chai. Multi-object hallucination in vision-language models. *CoRR*, abs/2407.06192, 2024b.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *CoRR*, abs/2412.05271, 2024c.

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit S. Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric Chu, Toby Boyd, Brad Hekman, Aaron Parisi, Chaoyi Zhang, Kornraphop Kawintiranon, Tania Bedrax-Weiss, Oliver Wang, Ya Xu, Ollie Purkiss, Uri Mendlovic, Ilaï Deutel, Nam Nguyen, Adam Langley, Flip Korn, Lucia Rossazza, Alexandre Ramé, Sagar Waghmare, Helen Miller, Nathan Byrd, Ashrith Sheshan, Raia Hadsell Sangnie Bhardwaj, Pawel Janus, Tero Rissa, Dan Horgan, Sharon Silver, Ayzaan Wahid, Sergey Brin, Yves Raimond, Klemen Kloboves, Cindy Wang, Nitesh Bharadwaj Gundavarapu, Ilia Shumailov, Bo Wang, Mantas Pajarskas, Joe Heyward, Martin Nikoltchev, Maciej Kula, Hao Zhou, Zachary Garrett, Sushant Kafle, Sercan Arik, Ankita Goel, Mingyao Yang, Jiho Park, Koji Kojima, Parsa Mahmoudieh, Koray Kavukcuoglu, Grace Chen, Doug Fritz, Anton Bulyenov, Sudeshna Roy, Dimitris Paparas, Hadar Shemtov, Bo-Juen Chen, Robin Strudel, David Reitter, Aurko Roy, Andrey Vlasov, Changwan Ryu, Chas Leichner, Haichuan Yang, Zelda Mariet, Denis Vnukov, Tim Sohn, Amy Stuart, Wei Liang, Minmin Chen, Praynaa Rawlani, Christy Koh, JD Co-Reyes, Guangda Lai, Praseem Banzal, Dimitrios Vytiniotis, Jieru Mei, and Mu Cai. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261, 2025.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, 2023.

Peng Ding, Jingyu Wu, Jun Kuang, Dan Ma, Xuezhi Cao, Xunliang Cai, Shi Chen, Jiajun Chen, and Shujian Huang. Hallu-pi: Evaluating hallucination in multi-modal large language models within perturbed inputs. *CoRR*, abs/2408.01355, 2024.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14303–14312. IEEE, 2024.

Chaoyou Fu, Yuhan Dai, Yondong 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, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *CoRR*, abs/2405.21075, 2024.

Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. Chat-rec: Towards interactive and explainable llms-augmented recommender system. *CoRR*, abs/2303.14524, 2023. doi: 10.48550/ARXIV.2303.14524. URL `https://doi.org/10.48550/arXiv.2303.14524`.

Xuan Gong, Tianshi Ming, Xinpeng Wang, and Zhihua Wei. DAMRO: dive into the attention mechanism of LVLM to reduce object hallucination. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 7696–7712. Association for Computational Linguistics, 2024.

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. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385. IEEE, 2024.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427. IEEE, 2024.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27026–27036. IEEE, 2024.

Prannay Kaul, Zhizhong Li, Hao Yang, Yonatan Dukler, Ashwin Swaminathan, C. J. Taylor, and Stefano Soatto. THRONE: an object-based hallucination benchmark for the free-form generations of large vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27218–27228. IEEE, 2024.

Jie Lei, Tamara L. Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 487–507. Association for Computational Linguistics, 2023.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882. IEEE, 2024.

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

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. *CoRR*, abs/2412.03735, 2024a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023b.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 6449–6464. Association for Computational Linguistics, 2023c.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *CoRR*, abs/2305.06355, 2023d.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Lou, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206. IEEE, 2024b.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, volume 15104, pp. 323–340. Springer, 2024c.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 292–305. Association for Computational Linguistics, 2023e.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*. OpenReview.net, 2024a.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *CoRR*, abs/2402.00253, 2024b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems*, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2024c.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024d. URL `https://llava-vl.github.io/blog/2024-01-30-llava-next/`.

Hui Liu and Xiaojun Wan. Models see hallucinations: Evaluating the factuality in video captioning. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 11807–11823. Association for Computational Linguistics, 2023.

Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. Phd: A prompted visual hallucination evaluation dataset. *CoRR*, abs/2403.11116, 2024e.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. *arXiv preprint arXiv:2407.21771*, 2024f.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics*, pp. 8731–8772. Association for Computational Linguistics, 2024g.

Fan Ma, Xiaojie Jin, Heng Wang, Yuchen Xian, Jiashi Feng, and Yi Yang. Vista-llama: Reducing hallucination in video language models via equal distance to visual tokens. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13151–13160. IEEE, 2024.

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

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adrià Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems*, 2023.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, and et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530, 2024.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045. Association for Computational Linguistics, 2018.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *CoRR*, abs/2410.17434, 2024.

Jaewon Son, Jaehun Park, and Kwangsu Kim. CSTA: cnn-based spatiotemporal attention for video summarization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18847–18856. IEEE, 2024.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *CoRR*, abs/2305.16355, 2023.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented RLHF. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics*, pp. 13088–13110. Association for Computational Linguistics, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

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

Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *CoRR*, abs/2406.16338, 2024b.

Hongliang Wei, Xingtao Wang, Xianqi Zhang, Xiaopeng Fan, and Debin Zhao. Toward a stable, fair, and comprehensive evaluation of object hallucination in large vision-language models. In *The Annual Conference on Neural Information Processing Systems*, 2024.

Shangyu Xing, Fei Zhao, Zhen Wu, Tuo An, Weihao Chen, Chunhui Li, Jianbing Zhang, and Xinyu Dai. EFUF: efficient fine-grained unlearning framework for mitigating hallucinations in multimodal large language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1167–1181. Association for Computational Linguistics, 2024a.

Yun Xing, Yiheng Li, Ivan Laptev, and Shijian Lu. Mitigating object hallucination via concentric causal attention. In *The Annual Conference on Neural Information Processing Systems*, 2024b.

Ziwei Xu, Sanjay Jain, and Mohan S. Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *CoRR*, abs/2401.11817, 2024. doi: 10.48550/ARXIV.2401.11817. URL `https://doi.org/10.48550/arXiv.2401.11817`.

Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. Vript: A video is worth thousands of words. In *Advances in Neural Information Processing Systems*, 2024.

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

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

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *CoRR*, abs/2310.16045, 2023.

Fan Yuan, Chi Qin, Xiaogang Xu, and Piji Li. HELPD: mitigating hallucination of lvlms by hierarchical feedback learning with vision-enhanced penalty decoding. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1768–1785. Association for Computational Linguistics, 2024.

Zihao Yue, Liang Zhang, and Qin Jin. Less is more: Mitigating multimodal hallucination from an EOS decision perspective. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 11766–11781. Association for Computational Linguistics, 2024.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In Yansong Feng and Els Lefever (eds.), *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 543–553. Association for Computational Linguistics, 2023.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024. URL `https://llava-vl.github.io/blog/2024-04-30-llava-next-video/`.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *CoRR*, abs/2311.16839, 2023.

Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. *CoRR*, abs/2410.04780, 2024a.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. *CoRR*, abs/2402.11411, 2024b.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *The International Conference on Learning Representations*. OpenReview.net, 2024c.

Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, and Huaxiu Yao. Calibrated self-rewarding vision language models. In *Advances in Neural Information Processing Systems*, 2024d.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The International Conference on Learning Representations*. OpenReview.net, 2024a.

Jiawei Zhu, Yishu Liu, Huanjia Zhu, Hui Lin, Yuncheng Jiang, Zheng Zhang, and Bingzhi Chen. Combating visual question answering hallucinations via robust multi-space co-debias learning. In *ACM Multimedia 2024*, 2024b.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. IBD: alleviating hallucinations in large vision-language models via image-biased decoding. *CoRR*, abs/2402.18476, 2024c.

Xianwei Zhuang, Zhihong Zhu, Zhanpeng Chen, Yuxin Xie, Liming Liang, and Yuexian Zou. Game on tree: Visual hallucination mitigation via coarse-to-fine view tree and game theory. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17984–18003. Association for Computational Linguistics, 2024.

APPENDIX

A USE OF LARGE LANGUAGE MODELS

Large language models were utilized in this work solely for two specific purposes: enhancing the coherence and style of the written manuscript, and generating dataset annotations using GPT-4o with methodologies detailed in both the main paper and appendix following established practices from prior benchmark studies. All other research components, such as experimental design and analysis, were conducted without involving LLMs.

B BENCHMARK CONSTRUCTION DETAILS

B.1 DATASET STATISTICS

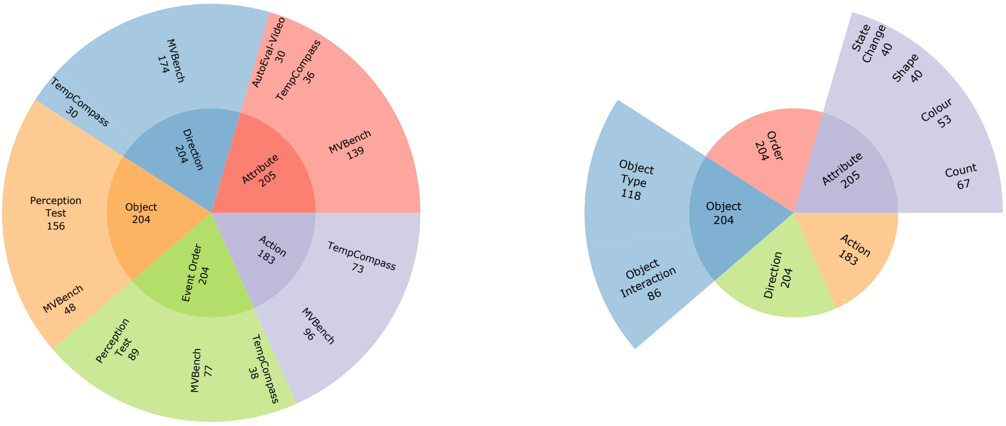

Figure 9: Distribution of visual instances in VIDHAL by (Left) public dataset source, categorized by the five temporal aspects, and (Right) temporal aspects and their sub-aspects.

Figure 9 presents the distribution of visual instances in VIDHAL by public dataset sources and temporal aspects. Additionally, Figure 10 further shows the distribution of ground truth answers for the MCQA and caption ordering tasks. One can observe that both temporal aspects and ground truth options are uniformly distributed across our benchmark. The distribution of video caption lengths and video durations is also presented in Figure 11.

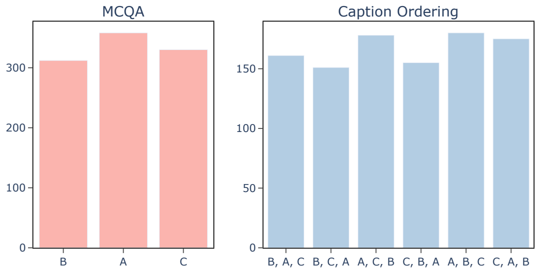

Figure 10: Distribution of (Left) correct answer options for the MCQA task and (Right) optimal option orders for the caption ordering task.

B.2 DATASET DEVELOPMENT PIPELINE

Visual Instance Selection To ensure a rich coverage of temporal aspects and visual diver-

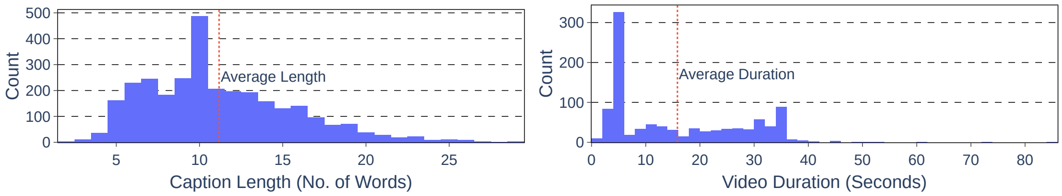

Figure 11: Distribution of (Left) caption lengths with an average of 11.2 words, and (Right) duration of videos in VIDHAL with an average of 15.8s.

Object Recognition [Object]:
What object does the person use to hit other objects?
What ingredients did the person put in the bowl or on the plate?
Which object was removed by the person from the tabletop?
What geometric shapes did the person put on the table?
What objects did the person hit?
What is the order of the letters on the table at the end?
What letters did the person type on the computer in order?
Distractor Action [Action]:
What is the person preparing?
Motion [Action]:
What happens with the object after being placed on the slanted plane?
What happened once the person removed an object from the tabletop?
Action Recognition [Action]:
What object does the person use to hit other objects?
What objects did the person hit?
What is the person preparing?
Which statement describes better the actions done by the person?
Sequencing [Event Order]:
What letters did the person show in order?
What is the order of the letters at the end?",
In what order did the person put the objects in the backpack?
What is the order of the letters on the table at the end?

Figure 12: Specific skills and corresponding questions from the Perception Test dataset chosen for VidHal instance selection, with the matched aspects indicated in brackets.

sity, we methodically selected video instances from four public datasets: TempCompass Liu et al. (2024g), Perception Test Patraucean et al. (2023), MVBench Li et al. (2024b), and AutoEval Video Chen et al. (2023). Given the unique characteristics of each dataset, we outline the specific guidelines adopted for each dataset below:

- **TempCompass** encompasses five temporal aspects: *Action*, *Speed*, *Direction*, *Event Order*, and *Attribute Change*. As most of these aspects align with those chosen to construct VidHal, we retain all video instances except those related to speed. TempCompass includes four evaluation tasks: *MCQA*, *Yes/No QA*, *caption matching*, and *caption generation*. Given the conciseness of captions in the latter two tasks, their information can often be subsumed within the more detailed QA-based annotations. Therefore, we focus exclusively on MCQA and Yes/No QA annotations to create an informative anchor caption.
- **Perception Test** spans various skill and reasoning domains to thoroughly evaluate VLLMs' perception and understanding abilities. Our inspection of these evaluation dimensions reveals alignment between the *semantics*, *physics*, and *memory* skill areas, as well as *descriptive* and *explanatory* reasoning dimensions, with the temporal aspects of action, order, and event order. Accordingly, we limit our video selection in Perception Test to these specific pillars. Additionally, we review the question templates adopted in these areas and select video instances with question-answer pairs that support VidHal's evaluation objectives. The specific skills and associated questions chosen are detailed in Figure 12.
- **MVBench** includes twenty video understanding tasks with question-answer pairs designed to challenge the reasoning capabilities of VLLMs. Similar to the Perception Test, we identify the tasks relevant to the temporal aspects in VidHal and focus on collecting videos belonging from these tasks. The specific tasks for each aspect are presented in Figure13. We observe that MVBench contains repeated use of certain scenarios across tasks, indicated by similar question templates. To enhance caption diversity and minimize redundancy, we limit the number of examples for each unique scenario. The collected instances cover all five temporal aspects of VidHal.
- **AutoEval-Video** evaluates open-ended response generation in VLLMs through questions with detailed answers across nine skill dimensions. We focus on instances related to the *state transition* area, specifically assessing changes in object and entity attributes. For each instance, we retain the only answers to associated questions as they act as informative, long-form captions for the video.

Incorrect Anchor Captions A minority of videos contain anchor captions misaligned with their content, often due to noisy metadata. Such discrepancies subsequently lead to undesirable hallucinatory captions. To remove such instances, we use BLIP2 Li et al. (2023b) to calculate frame-text matching scores across all video frames, selecting the maximum score as the representative video-text alignment score. Examples with incorrect anchor captions typically achieve low alignment scores, which are discarded

Action: Action Sequence, Fine-Grained Action and Fine-Grained Pose
Direction: Moving Direction.
Object: Object Interaction, Object Existence.
Attribute: Moving Attribute, Moving Count.
Order: Action Sequence

Figure 13: Evaluation tasks in MVBench aligned with temporal aspects in VidHal, categorized by aspect.

You are a chatbot tasked with generating hallucinatory captions for a video given the input ground truth caption provided. Your objective is to modify the <aspect> present in the provided caption to generate 2 incorrect captions of different levels of hallucination. <aspect_definition>. The extent of hallucination of each caption is measured on a scale of 1 to 3 in increasing levels of hallucination, with 1 denoting no hallucinations present and 3 denoting a large extent of hallucination. A description of the extent of hallucination represented by each score is given as follows:

1. The caption contains no hallucination. The caption that representing this score is the ground truth caption.
2. The caption includes moderate hallucination, describing an event that is different from the ground truth, yet possible given the context of the video
3. The caption contains high hallucination, describing an event that is realistic, but typically unlikely to happen given context reflected by the original caption.

The generated hallucinated captions should follow the guidelines below.

Guidelines:
1. Focus only on modifying the temporal aspect provided in the instruction. Do not change any other temporal aspect associated with objects or subjects in the video.
2. Keep your modifications brief but coherent. Your generated captions should be of similar length to the original caption.
3. Ensure that your generated captions depict realistic and believable scenarios even as they deviate from the original context. For example, avoid creating fictitious scenarios such as "Person flying on a broomstick" and "Monkey painting a picture".
4. You may rephrase the provided caption to maintain consistent sentence structure across all captions. However, make sure the factual content of the ground truth caption remains unchanged.
5. Each generated hallucinatory caption should be of the form <score> : <caption>, <score> takes a value from the hallucination scale defined and <caption> represents your provided hallucinatory caption.
6. No two generated <caption> should share the same <score>, and each caption should take on a unique level of hallucination from 2 to 3.

Here are some examples of how hallucinatory captions are expected to be constructed.

<in_context_examples>

Now, generate hallucinatory captions for the following video description.

Original Caption:
<anchor_caption>
Hallucinated Captions:

Figure 15: Prompt for generating aspect-specific hallucinatory captions based on anchor captions and in-context examples.

as noisy instances.

LLM-based Caption Generation We utilize GPT-4o's OpenAI (2023) text processing and generation capabilities to generate an anchor caption for each selected video, based on metadata from its original public dataset source. This metadata includes QA-based annotations for TempCompass, Perception Test, and MVBench, along with long-form answers for AutoEval-Video. The anchor caption is subsequently used as input for GPT-4o to generate corresponding hallucinatory captions.

To ensure the generated hallucinatory captions meet high-quality standards, we employ a detailed prompt adopting the following strategies to guide GPT-4o's output:

- Aspect-specific definitions which outline the characteristics of each aspect to be varied, prompting GPT-4o to modify anchor captions accordingly.
- Caption construction guidelines that define the structure, format, and hallucination levels required for the generated captions.
- In-context examples to illustrate the desired form of each hallucinatory caption for each aspect.

You are given a long caption describing the content of a video. Your task is to provide a summarised and concise version of this caption. Ensure that you keep all essential detail in the original caption.

<metadata>

Video description:

Figure 14: Prompts used for generating the anchor caption from long-form captions.

The prompts for generating anchor and hallucinatory captions are shown in Figures 14 to 17a, respectively, with definitions for each aspect are provided in Figure 16. Aspect-specific in-context examples are detailed in Figures 17b to 21. Separate in-context examples are provided for each *Attribute* subaspect of *Shape*, *Size*, *Color*, *Count*, and *State Change* to account for their distinct natures.

Caption Quality Scoring To identify video instances with the high quality generated captions, we utilize powerful LLMs to evaluate the quality of generated captions. The captions are assessed is based on three specific criteria:

- **Realism** determines whether generated scenarios are plausible.
- **Ordering Quality** evaluates whether the hallucination level ordering is appropriate.
- **Relevance** ensures that deviations from the anchor caption align with the designated aspect.

Binary questions are used to evaluate captions for each criterion, assigning a score of 1 for positive responses, *i.e.*, ”yes”, and 0 otherwise. The scores for each criterion are averaged across all models

Action: Actions refer to observable movements or activities performed by entities that may involve interaction with objects or the environment in the video.
Direction: Direction refers to the course or path along which objects or subjects move in the video.
Order: Order refers to the sequential arrangement of events that occur in the video.
Object: Objects refer to inanimate, physical entities or items present within the video.
State: State refers to the condition or status of an object or subject, indicating its current properties, position or the phase of action the subject is taking or phase of process the object is undergoing.
Count: Count refers to the frequency of an action being performed or an event occurring. It may also refer to the number of objects or subjects involved in an event or interaction.
Color: Color refers to the hue or shade of an object or subject.
Shape: Shape refers to the form or outline of an object or subject.
Size: Size refers to the dimensions or magnitude of an object or subject.

Figure 16: Definitions incorporated into the prompt for generating hallucinatory captions for each aspect, with separate definitions provided for each sub-aspect in the *Attribute* aspect.

You are given one or more questions targeted at content of a video and their corresponding answers. You are tasked with generating an appropriate and informative single line caption for the video using this information given to you. Ensure that you restrict yourself to only information present in the question-answer pairs provided. If the answers to the questions provide various types of information, concentrate on the color related to the subjects and objects in the video in your caption. Focus on providing clear and concise descriptions without using overly elaborate language.

`<metadata>`

Video description:

(a) Prompt used for generating the anchor caption from QA-based annotations.

Original Caption:
1 : A red bucket of liquid goes from empty to half full.
Hallucinated Captions:
2 : A red bucket of liquid goes from empty to completely full.
3 : A red bucket of liquid goes from completely full to empty.

Original Caption:
1 : The light in the room is slowly dimming.
Hallucinated Captions:
2 : The light in the room slowly dims, then brightens again.
3 : The light in the room is slowly getting brighter.

Original Caption:
1 : The sky changes from clear to partly cloudy.
Hallucinated Captions:
2 : The sky changes from clear to completely overcast.
3 : The sky changes from partly cloudy to clear.

(b) In-context examples for the *State* sub-aspect under the *Attribute* aspect.

Figure 17: (Left) Prompts used for generating the anchor caption, and (Right) in-context examples for the *State* sub-aspect.

Original Caption:
1 : A boy inflates the balloon, which grows vertically.
Hallucinated Captions:
2 : A boy inflates the balloon, which grows horizontally.
3 : A boy deflates the balloon, which shrinks horizontally.

Original Caption:
1 : The bag expands in height as items are being placed inside.
Hallucinated Captions:
2 : The bag expands in width as items are being placed inside.
3 : The bag shrinks in height as items are being placed inside.

Original Caption:
1 : The size of the puddle of water is increasing.
Hallucinated Captions:
2 : The size of the puddle of water is decreasing.
3 : The size of the puddle of water remains unchanged.

Original Caption:
1 : A circle shaped block is placed in a wooden box.
Hallucinated Captions:
2 : A square shaped block is placed in a wooden box.
3 : A star shaped block is placed in a wooden box.

Original Caption:
1 : Cubes are transforming into cylinders.
Hallucinated Captions:
2 : Cubes are transforming into cones.
3 : Cubes are transforming into spheres.

Original Caption:
1 : The clouds form a fluffy circle in the sky.
Hallucinated Captions:
2 : The clouds form a fluffy square in the sky.
3 : The clouds form a fluffy triangle in the sky.

Figure 18: In-context examples for the *Size* (Left) and *Shape* (Right) sub-aspects.

Original Caption:
1 : A leaf with holes turns green to red.
Hallucinated Captions:
2 : A leaf with holes turns from green to orange.
3 : A leaf with holes turns from yellow to orange.

Original Caption:
1 : A yellow ball bounces on the ground, and lands in the pool.
Hallucinated Captions:
2 : A red ball bounces on the ground, and lands in the pool.
3 : A blue ball bounces on the ground, and lands in the pool.

Original Caption:
1 : A stationary purple cup appears at the beginning of the video.
Hallucinated Captions:
2 : A stationary blue cup appears at the beginning of the video.
3 : A stationary green cup appears at the beginning of the video.

Original Caption:
1 : The man wearing a jacket performed three backflips.
Hallucinated Captions:
2 : The man wearing a jacket performed four backflips.
3 : The man wearing a jacket performed five backflips.

Original Caption:
1 : Four birds perched on the wire.
Hallucinated Captions:
2 : Five birds perched on the wire.
3 : Six birds perched on the wire.

Original Caption:
1 : One car drove down the road.
Hallucinated Captions:
2 : Two cars drove down the road.
3 : Three cars drove down the road.

Figure 19: In-context examples for the *Color* (Left) and *Count* (Right) sub-aspects.

Original Caption:
1 : The man hits another object with a bat.
Hallucinated Captions:
2 : The man hits another object with a racket.
3 : The man hits another object with a broom.

Original Caption:
1 : The ball bounces down the slanted plane.
Hallucinated Captions:
2 : The ball rolls down the slanted plane.
3 : The ball zigzags down the slanted plane.

Original Caption:
1 : A person puts two rectangles and one circle into the bag.
Hallucinated Captions:
2 : A person puts a rectangle, a square and a circle into the bag.
3 : A person puts two squares and a circle into the bag.

Original Caption:
1 : A person puts a bottle in the bag. Then, he puts a book in the bag. Lastly, he puts a pencil case into the bag.
Hallucinated Captions:
2 : A person puts a book in the bag. Then, he puts a bottle in the bag. Lastly, he puts a pencil case into the bag.
3 : A person puts a pencil case in the bag. Then, he puts a book in the bag. Lastly, he puts a bottle into the bag.

Original Caption:
1 : A man writes letters in the following order: A, V, T, Y.
Hallucinated Captions:
2 : A man writes letters in the following order: A, Y, T, V.
3 : A man writes letters in the following order: Y, T, V, A.

Original Caption:
1 : A woman with white coat places a book on the table. She takes two vials of liquid and mixes them together.
Hallucinated Captions:
2 : A woman with white coat places a book on the table. She takes off her coat. Then, she takes two vials of liquid and mixes them together.
3 : A woman with white coat takes two vials of liquid and mixes them together. She then places a book on the table.

Figure 20: In-context examples for the *Object* (Left) and *Event-Order* (Right) aspects.

Original Caption:
1 : The people are cooking in the video.
Hallucinated Captions:
2 : The people are chopping in the video.
3 : The people are washing in the video.

Original Caption:
1 : A car is driving down the road.
Hallucinated Captions:
2 : A car is reversing down the road.
3 : A car is being repaired along the road.

Original Caption:
1 : A dog is digging a hole near the tree.
Hallucinated Captions:
2 : A dog is scratching the tree.
3 : A dog is barking at the tree

Original Caption:
1 : An eagle is flying from left to right diagonally upwards.
Hallucinated Captions:
2 : An eagle is flying from left to right horizontally.
3 : An eagle is flying from left to right diagonally downwards.

Original Caption:
1 : The car drives forward and makes a right turn.
Hallucinated Captions:
2 : The car drives forward and continues driving straight.
3 : The car drives forward and makes a left turn.

Original Caption:
1 : The ball on the table rolls away from the camera.
Hallucinated Captions:
2 : The ball on the table rolls from left to right.
3 : The ball on the table rolls towards the camera.

Figure 21: In-context examples for the *Action* (Left) and *Direction* (Right) aspects.

and prompts, and then summed across all criteria to produce a final quality assessment score for the generated captions of a video instance.

We evaluate each set of captions using three LLMs: GPT-4o, Gemini-1.5 Flash Reid et al. (2024), and LLaMA3 (70B) Dubey et al. (2024) along with three variants for each binary question. This ensemble of both models and prompts enhances the robustness of our evaluation.. Figures 22 and 23 provide details of the criterion-specific quality assessment queries and the prompt templates employed for each LLM. We select the top 1,000 examples with the highest quality assessment scores to construct VIDHAL.

GPT-4o & Gemini-1.5 Flash:
You are provided with a ground truth description of a video, and 2 other captions that contain hallucinations in the aspect of <aspect>. The hallucinated captions are displayed in increasing order of hallucination, where the first caption contains the least amount of hallucinated elements and the last caption having significant hallucination. You are tasked with answering a question regarding the quality of the hallucinated captions. Provide your answer as detailed in the question, without further explanation of your answer.

Ground truth caption:
<anchor_caption>

Hallucinated captions:
<hallucinatory_captions>

Question:
<quality_assessment_question>

Answer:

LLaMA3 (70B):
<|begin_of_text|><|start_header_id|>system<|end_header_id|>
You are provided with a ground truth description of a video, and 2 other captions that contain hallucinations in the aspect of <aspect>. The hallucinated captions are displayed in increasing order of hallucination, where the first caption contains the least amount of hallucinated elements and the last caption having significant hallucination. You are tasked with answering a question regarding the quality of the hallucinated captions. Provide your answer as detailed in
the question, without further explanation of your answer.
<|eot_id|>
<|start_header_id|>user<|end_header_id|>
Ground truth caption:
<anchor_caption>

Hallucinated captions:
<hallucinatory_captions>

Question:
<quality_assessment_question>

Answer:
<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>

Figure 22: Prompt template for evaluating the quality of generated captions for the GPT-4o, Gemini-1.5 Flash, and LLaMA3 (70B) models.

Realism:
1. Is the scenario presented in caption <option> realistic? Provide your answer only as a single "yes" or "no".
2. Is the event in caption <option> believable? Provide your answer only as a single "yes" or "no".
3. Is the setting present in caption <option> plausible? Provide your answer only as a single "yes" or "no".

Order Quality:
1. Which caption better matches the ground truth description: Caption <option_A> or <option_B>? Provide your answer only as a single number (<option_A> or <option_B>)
2. Which caption aligns more closely with the ground truth description: Caption <option_A> or <option_B>? Provide your answer only as a single number (<option_A> or <option_B>)
3. Which caption is more faithful to the ground truth description: Caption <option_A> or <option_B>? Provide your answer only as a single number (<option_A> or <option_B>)

Relevance:
1. Does hallucinated caption <option> differ from the ground truth caption only in the <aspect>? Provide your answer only as a single "yes" or "no".
2. Is the only difference between hallucinated caption <option> and the ground truth caption the <aspect>? Provide your answer only as a single "yes" or "no".
3. Did hallucinated caption <option> change the ground truth caption only with respect to the <aspect>? Provide your answer only as a single "yes" or "no".

Figure 23: Question prompts for evaluating caption quality based on the three assessment criteria. Prompts with the placeholder <option> are applied individually to the anchor and hallucinatory captions. For question associated with *order quality*, <option_A> and <option_B> are replaced with the corresponding hallucinatory caption options shown to the LLMs.

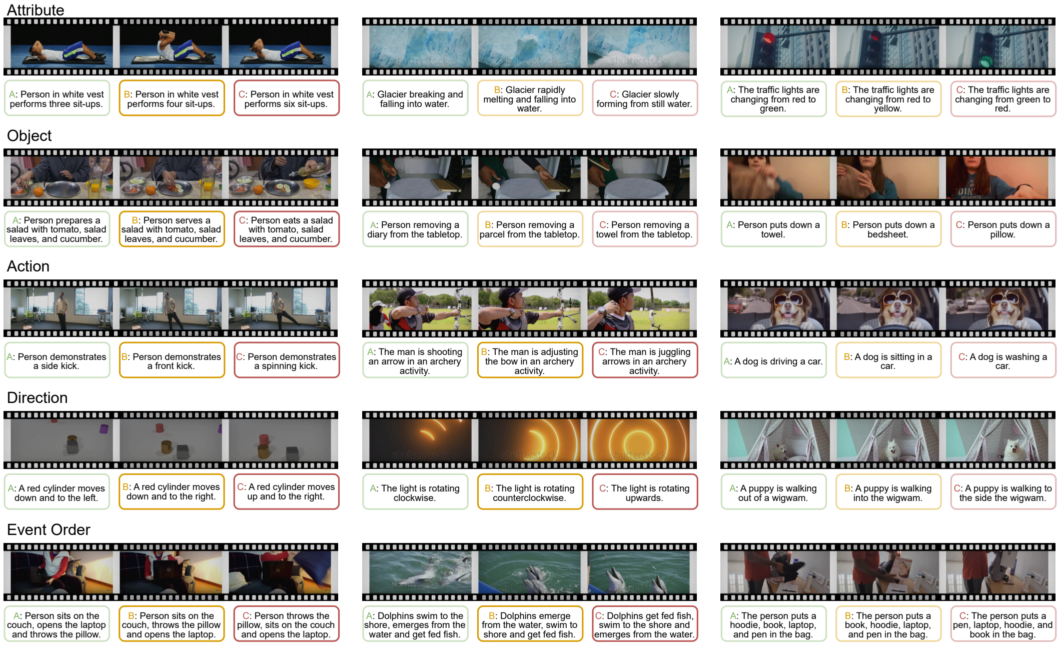
Attribute
A: Person in white vest performs three sit-ups.
B: Person in white vest performs four sit-ups.
C: Person in white vest performs six sit-ups.
A: Glacier breaking and falling into water.
B: Glacier rapidly melting and falling into water.
C: Glacier slowly forming from still water.
A: The traffic lights are changing from red to green.
B: The traffic lights are changing from red to yellow.
C: The traffic lights are changing from green to red.
Object
A: Person prepares a salad with tomato, salad leaves, and cucumber.
B: Person serves a salad with tomato, salad leaves, and cucumber.
C: Person eats a salad with tomato, salad leaves, and cucumber.
A: Person removing a diary from the tabletop.
B: Person removing a parcel from the tabletop.
C: Person removing a towel from the tabletop.
A: Person puts down a towel.
B: Person puts down a bedsheet.
C: Person puts down a pillow.
Action
A: Person demonstrates a side kick.
B: Person demonstrates a front kick.
C: Person demonstrates a spinning kick.
A: The man is shooting an arrow in an archery activity.
B: The man is adjusting the bow in an archery activity.
C: The man is juggling arrows in an archery activity.
A: A dog is driving a car.
B: A dog is sitting in a car.
C: A dog is washing a car.
Direction
A: A red cylinder moves down and to the left.
B: A red cylinder moves down and to the right.
C: A red cylinder moves up and to the right.
A: The light is rotating clockwise.
B: The light is rotating counterclockwise.
C: The light is rotating upwards.
A: A puppy is walking out of a wigwam.
B: A puppy is walking into the wigwam.
C: A puppy is walking to the side the wigwam.
Event Order
A: Person sits on the couch, opens the laptop and throws the pillow.
B: Person sits on the couch, throws the pillow and opens the laptop.
C: Person throws the pillow, sits on the couch and opens the laptop.
A: Dolphins swim to the shore, emerges from the water and get fed fish.
B: Dolphins emerge from the water, swim to shore and get fed fish.
C: Dolphins get fed fish, swim to the shore and emerges from the water.
A: The person puts a hoodie, book, laptop, and pen in the bag.
B: The person puts a book, hoodie, laptop, and pen in the bag.
C: The person puts a pen, laptop, hoodie, and book in the bag.

Figure 24: Qualitative examples of video instances and their corresponding generated captions in the VIDHAL Benchmark, across the five temporal aspects.

B.3 Additional Dataset Examples

We provide additional qualitative examples of video instances and their corresponding captions in Figure 24 for each of the five temporal aspects.

C Human Validation Details

C.1 Human Validation Process

As varying hallucination levels are a distinctive feature of our benchmark, we prioritize validating the robustness of caption ordering produced by our annotation pipeline. Each anchor caption is derived from the original video metadata, making it the most accurate reflection of the video content. Our primary objective is to ensure that the ordering of hallucinatory captions aligns with human judgment. To achieve this, human annotators are shown the video instance along with both hallucinatory captions and are tasked with selecting the caption that better aligns with the video content, as illustrated in Figure 25. Each video instance is reviewed by multiple annotators, with the final human-aligned order determined through a majority vote and compared with our automatically generated order.

C.2 Misaligned Instances

Table 3 lists video instances that fail to meet the majority agreement threshold established by our annotation process. We additionally provide the corresponding human agreement scores for each instance.

D Evaluation Pipeline Details

D.1 Model and Inference Hyperparameters

We provide additional details on the inference and generation settings used across all evaluated models in Table 4, as well as hyperparameters specific to LlaVA-NeXT-Video models in Table 5.

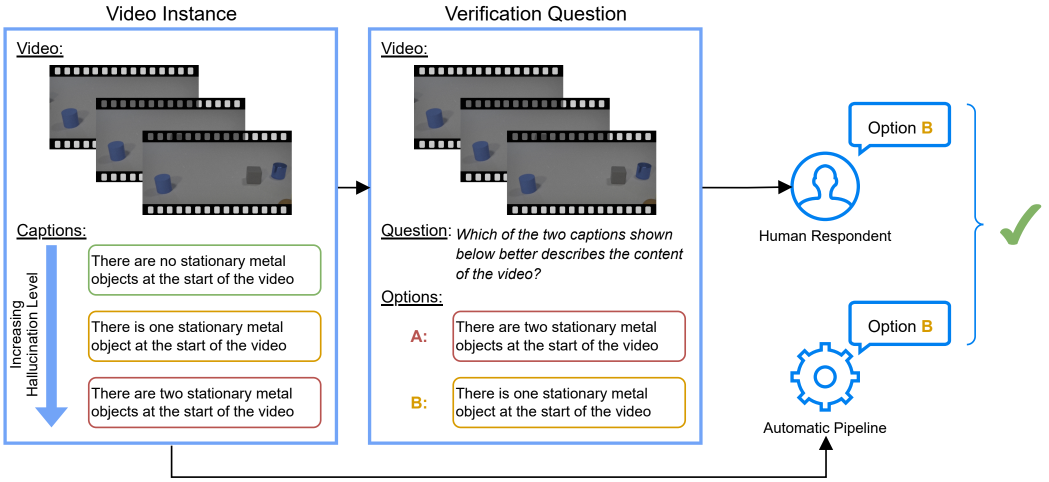
Video Instance
Video:
Captions:
Increasing Hallucination Level
There are no stationary metal objects at the start of the video
There is one stationary metal object at the start of the video
There are two stationary metal objects at the start of the video
Verification Question
Video:
Question: Which of the two captions shown below better describes the content of the video?
Options:
A:
There are two stationary metal objects at the start of the video
B:
There is one stationary metal object at the start of the video
Option B
Human Respondent
Option B
Automatic Pipeline

Figure 25: Pipeline for validating the quality of generated caption orders in VidHal. For each instance, human annotators are provided with the video and its associated hallucinatory captions. The annotators then select the caption that best aligns with the video content. The selected response is subsequently checked for consistency with the caption with lower hallucination according to our annotation process.

Video ID	Agreement Score
action_55	0.429
action_88	0
action_90	0.308
action_118	0.200
action_153	0.250
order_60	0.500
order_109	0.154
attribute_90	0.400
attribute_180	0.071
attribute_192	0.188
object_25	0.375
object_170	0
direction_188	0.400

Table 3: Instances where generated caption orders diverge from human preference in quality checks. The agreement score reflects the proportion of respondents who chose our annotated order.

Hyperparameter	**Value**
Data Processing	
Video Sampling Rate (FPS)	30
Generation	
do_sample	False
temperature	0.0
repetition_penalty	1.0
max_new_tokens	128
Computation	
Precision	FP16

Table 4: Hyperparameter configuration used in VIDHAL evaluation across all models.

Hyperparameter	LLaVA-NeXT-Video (7B)	LLaVA-NeXT-Video (32B)
mm_spatial_pool_mode	average	average
mm_newline_position	no_token	grid
mm_pooling_position	after	after

Table 5: Model-specific hyperparameters for LLaVA-NeXT-Video models.

> You are provided with a video and a set of several captions. Your task is to watch the video provided carefully, and select the caption that best describes the video. Provide your answer only as a single letter representing the option whose caption that best describes the video, without any explanation.
>
> Watch the video provided, and choose the option whose caption describes the video most accurately.
>
> A. <caption_A>
> B. <caption_B>

Figure 26: Prompt template for the MCQA and relative caption ordering evaluation tasks.

> Watch the video provided, and rank the captions below in order from the most accurate to the least accurate in describing the video. Provide your response only as a sequence of comma separated option letters matching the corresponding captions. Do not give any additional explanation for your answer.
>
> For example, if option B contains the caption that best describes the video, option A contains the caption that describes the video second best and option C contains the caption that describes the video least accurately, provide your response as: B, A, C.
>
> A. <caption_A>
> B. <caption_B>
> C. <caption_C>

Figure 27: Prompt template for the naive caption ordering evaluation task.

D.2 Evaluation Task Prompts

Figures 26 and 27 present the prompts used for the MCQA and naive caption ordering tasks, respectively. The same prompt used for both the MCQA task and the paired questions in the relative caption ordering task. Our manual inspection of these instances reveals that these videos often feature visually complex content, making them challenging even for human annotators.

D.3 Relative Order Parsing

Prompting the VLLM to predict the order of captions based on their hallucinatory level in the relative caption ordering task involves asking a series of paired questions derived from different caption combinations. However, providing the model with all possible pairs at once may result in cyclic and non-transitive orderings. To address this, we present each caption pair to the VLLM in a systematically selected sequence, beginning with two paired questions. The final paired question is presented to the model to resolve inconsistencies if the multiple possible orderings can be derived from the responses to the first two paired questions. The responses across all paired questions presented to the VLLM is then parsed according to the workflow illustrated in Figure 28.

E Additional Experiments

E.1 Input Order Sensitivity

To assess the robustness of VLLM responses to the order of displayed captions, we conducted additional experiments by evaluating three VLLMs using a fixed static display order across all instances. We repeated this process across all different permutations of input caption order, presenting the results of these models in Figure 29. We observe that the performance of these VLLMs is highly sensitive to the order in which captions are displayed, reflected by their varying results across different order permutations. This instability intensifies with smaller model sizes, with VideoLLaMA2 (7B) showing the highest variance in evaluation results and VideoLLaMA2 (72B) the lowest. Our findings suggest that VLLMs may be particularly vulnerable to input caption order, potentially confounding their performance.

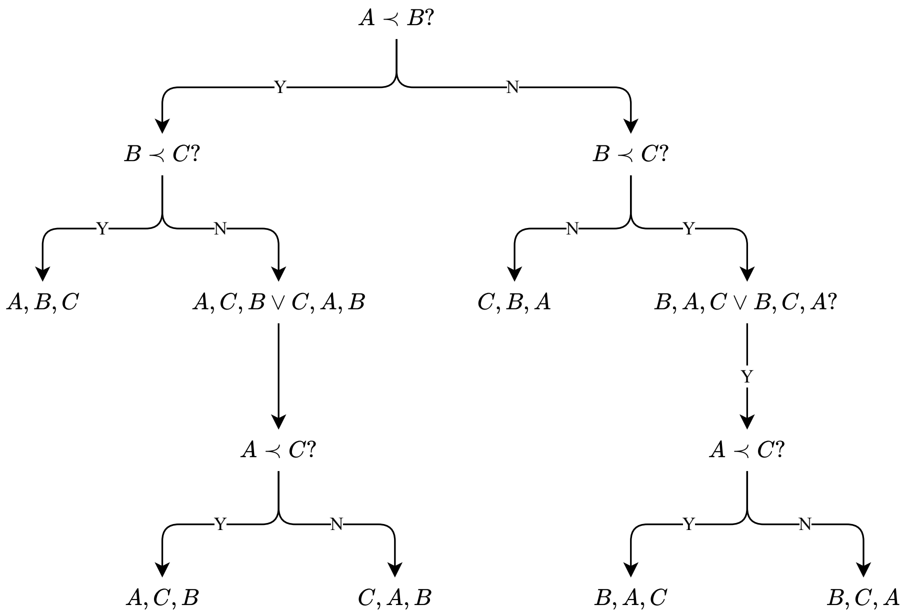

Figure 28: Decision tree for determining the final caption order based on VLLM responses to paired questions in the relative caption ordering evaluation task.

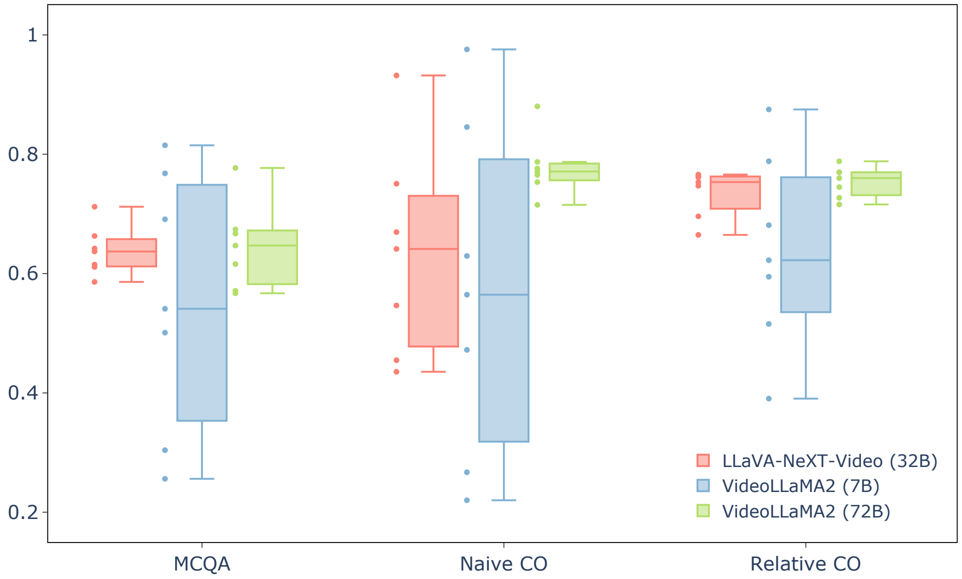

Figure 29: Distribution of results of VLLMs across varied input caption orders for the three evaluation tasks.

E.2 Naive Caption Ordering Response Quality

To analyze VLLMs' ability to handle naive caption ordering tasks, which possess unique task structures compared to conventional video understanding tasks, we employ two quantitative metrics. Regurgitation Rate (RR) captures the model's propensity to consistently generate identical responses regardless of input, defined as the maximum proportion of instances in VidHal where a specific caption order is predicted across all possible orderings. Invalid Response Rate (IRR) measures the proportion of responses that fail to provide valid caption orders for the naive ordering task. Figure 30 presents IRR and RR scores for all evaluated models, revealing two key observations. First, many models exhibit high IRR scores, frequently outputting incomplete caption orders (e.g., generating only a single option). Second, despite formulating responses with correct structure, many VLLMs produce identical caption orders regardless of the input video V^i, as reflected by high RR scores, a behavior observed even in models performing well on MCQA and relative caption ordering tasks, such as InternVL2.5.

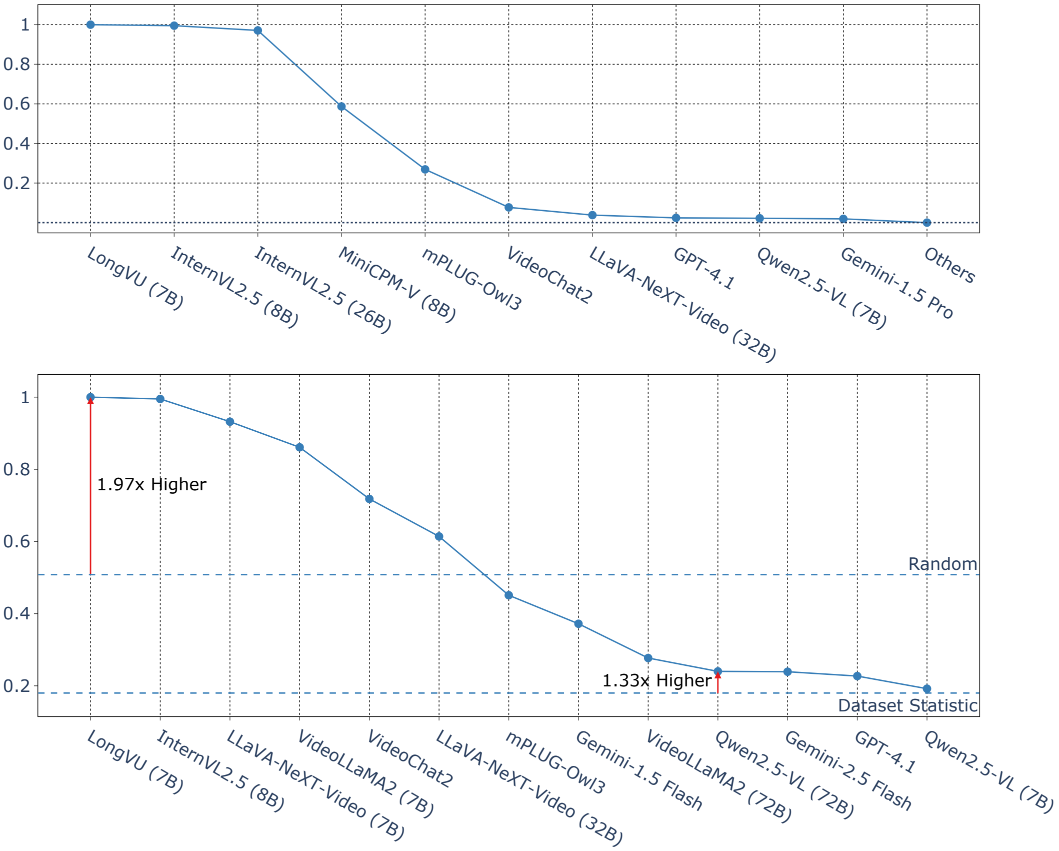

Figure 30: (Top) Invalid response rates across all models. VLLMs with no invalid responses are grouped under *Others*. (Bottom) Regurgitation rates of VLLMs on VIDHAL. *Random* and *Dataset Statistic* indicate the regurgitation rates of the random baseline and ground truth answers, respectively. For both metrics, a lower value indicates better model performance.

E.3 IMAGE PRIOR RELIANCE - ABLATION STUDY ON VIDEO SUMMARIZATION ALGORITHM

We conduct additional single-frame bias experiments using uniform and motion-based sampling strategies with varying clip lengths (1, 2, and 4 frames), with results presented in Tables 6 and 7. The overlap ratios demonstrate consistency across all three video summarization methods (saliency-based, uniform, and motion-based sampling) for extracting frames v^i. In particular, single-frame outputs substantially overlap with full-video inputs regardless of the summarization algorithm employed. These additional results confirm that our single-frame bias study is robust across different frame selection methods, with VLLMs relying on single-frame information for over half of the queries in VIDHAL.

	1 Frame			2 Frames			4 Frames		
Model	**C**	**I**	**O**	**C**	**I**	**O**	**C**	**I**	**O**
VideoLLaMA2 (7B)	0.674	0.708	0.700	0.781	0.798	0.794	0.846	0.829	0.833
LLaVA-NeXT-Video (32B)	0.680	0.570	0.620	0.735	0.649	0.688	0.831	0.706	0.763

Table 6: Overlapping ratios of model predictions under single-frame and full-video inputs for (C)orrect, (I)ncorrect and (O)verall predictions using uniformly sampled frames v^i, across multiple frame sampling rates.

	1 Frame			2 Frames			4 Frames		
Model	**C**	**I**	**O**	**C**	**I**	**O**	**C**	**I**	**O**
VideoLLaMA2 (7B)	0.521	0.495	0.515	0.558	0.507	0.519	0.670	0.653	0.657
LLaVA-NeXT-Video (32B)	0.634	0.550	0.558	0.658	0.546	0.597	0.675	0.563	0.614

Table 7: Overlapping ratios of model predictions under single-frame and full-video inputs for (C)orrect, (I)ncorrect and (O)verall predictions using motion-based sampled frames v^i, across multiple frame sampling rates.