VIDHAL: Benchmarking Temporal Hallucinations in Vision LLMs

Anonymous Author(s)
Affiliation
Address
email

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 GPT-4o. Our results uncover significant limitations in existing VLLMs with respect to 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 [36, 34] 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 that appear to be plausible contradict the visual context [1, 59]. 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 [22, 77, 11, 78] and attention calibration [16, 41, 39, 66, 14, 71, 58]. Other efforts have been devoted to the evaluation of hallucinations in VLLMs. For example, CHAIR [47] initially studies object-based hallucination evaluation with the aid of the image captioning task. Subsequent studies [31, 38, 20, 10] instead harness paired $\langle$*positive, hallucinatory*$\rangle$ questions to probe such hallucinations. Additionally, MMHalBench [50] and AMBER [53] 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 [29, 45, 6, 12, 40, 42]. 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-

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

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. Additionally, proprietary models, *e.g.*, GPT-4o [43], often outperform open-source counterparts by significant margins.

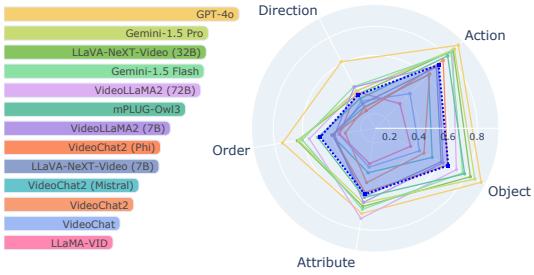

Figure 1: Multiple-Choice Question Answering (MCQA) performance of representative VLLMs on our VIDHAL benchmark. (Left) Overall ranking of VLLMs. (Right) Detailed accuracy results pertaining to each temporal aspect, wherein higher scores indicate fewer hallucinations.

Overall, 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 sourced from public video understanding datasets encompassing a diverse range of temporal concepts and ii) captions with varying levels of hallucination[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 [36, 34, 25, 9, 62, 63, 61]. Typical methods in this category include LLaVA [36], mPLUG-Owl [63, 61, 62], InstructBLIP [9], and MiniGPT-4 [75]. These VLLMs rely on aligning vision encoders with LLMs using connective modules such as Q-Former [9, 26, 25, 67, 8] or MLPs [36, 34, 49] 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 [8] enhances the LLaMA model with video understanding capabilities through a Spatial-Temporal Convolution (STC) module. LLaVA-NeXT-Video [35, 68] 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 [18, 39, 76, 5]. A common demonstration is that the generated responses contain information which is inconsistent with the visual

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

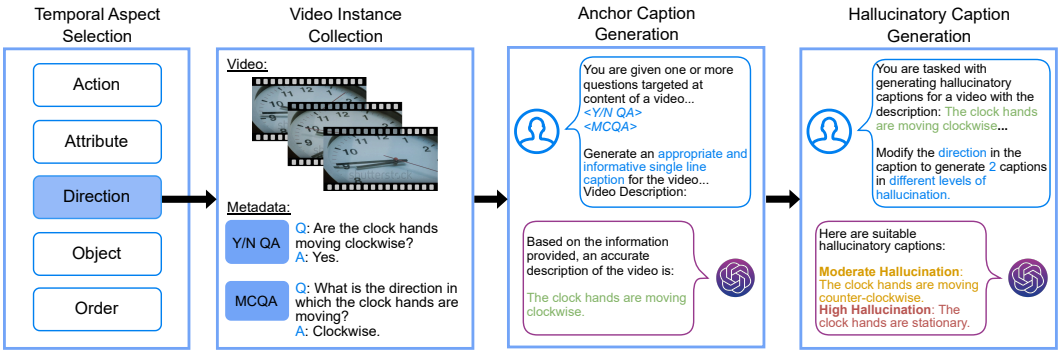

Figure 2: Overview of our VIDHAL benchmark construction pipeline. Using *direction* as an example from the five selected aspects, we begin by sourcing relevant video instances from existing datasets. Next, the anchor (positive) caption is generated from the original video metadata. Finally, GPT-4o is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

content [1, 33, 65, 57]. Most approaches address the hallucination problem with post-hoc techniques. For example, LURE [73] and Woodpecker [64] develop pipelines that assist VLLMs in revising their responses using expert models. To reduce bias from unimodal and statistical priors, contrastive decoding methods, such as VCD [22] and M3ID [11], along with attention calibration techniques like OPERA [16] are employed to refine token predictions. Building on the success of reinforcement learning for preference optimization in LLM development [44], HA-DPO [69], POVID [72] and CSR [74] 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 [23], VideoBench [42], MVBench [29], and VideoMME [12], 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 [6] and Perception Test [45] introduce complex reasoning tasks such as counterfactual and explanatory reasoning, while TempCompass [40] expands temporal concept coverage. Several benchmarks [31, 53, 50, 20, 32, 19, 55, 70, 5, 4, 56, 51, 7] have been constructed to quantify visual hallucinations, primarily targeting object-based hallucinations in images. HallusionBench [15], VideoCon [2], and Vript [60] provides partial coverage of video-based hallucinations, while VidHalluc [24] and VideoHallucer [54] introduces a benchmark 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 [60, 54, 2], 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 [31, 53, 47] and videos [54, 60, 15] 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 and properties 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 [40], Perception Test [45], MVBench [29] and AutoEval-Video [6]. 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 [31, 53] 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 a set of M captions that reflect varying degrees of hallucination in VLLMs. Given the cost and labor intensity of manual annotation, we follow existing studies such as PhD [38] and MVBench [29], 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.

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

	Dataset	Temporal Aspects										Task Formats	Evaluation Metrics
		Action	Attribute					Direction	Object		Order		
			Size	Shape	Color	Count	State-Change		Recognition	Interaction			
Video Reasoning	SEEDBench [23]	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓	MCQA	Accuracy
	VideoBench [42]	✓	✓	✓	✓	✓	✗	✗	✓	✗	✗	MCQA	Accuracy
	MVBench [29]	✓	✗	✗	✗	✗	✗	✓	✓	✓	✓	MCQA	Accuracy
	Video-MME [12]	✓	✓	✓	✓	✓	✗	✗	✓	✗	✗	MCQA	Accuracy
Hallucination Evaluation	Vript [60]	✓	✗	✗	✗	✗	✗	✗	✓	✓	✓	Video Captioning Event Ordering	F1 Score Accuracy
	VideoCon [2]	✓	✓	✓	✓	✓	✗	✗	✓	✗	✓	VL Entailment	ROC-AUC
	HallusionBench [15]	✓	✗	✗	✗	✗	✗	✓	✗	✗	✓	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.

levels of hallucination based on the temporal concepts associated with it. Specifically, $y_-^{i,k}$ exhibits heavier hallucination than $y_-^{i,j}$ for $j < k$. 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

Our VIDHAL benchmark consists of a total of 1,000 video instances. Using our automatic annotation pipeline, each video instance is tagged with $M = 3$ captions. As shown in Table 1, our VIDHAL dataset stands out from other video understanding [23, 42, 29, 12] and hallucination benchmarks [2, 15, 37] 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.

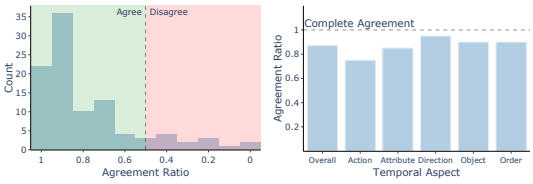

Figure 3: Human agreement on hallucination levels in the VIDHAL dataset. (Left) Distribution of agreement ratios per video sample. (Right) Average agreement ratio for each temporal aspect, with an overall average of 87%.

To ensure the reliability of our generated captions at varying levels, we randomly selected 100 examples for human validation, where each sample is 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% between our automatically generated hallucinatory captions and human annotators, indicating consistency between these two across all temporal aspects.

4 VIDHAL Evaluation Protocol

Aiming 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 model is provided with a video V^i and its corresponding set of captions $\mathcal{Y}^i$ as answer options. The VLLM is then 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.

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.

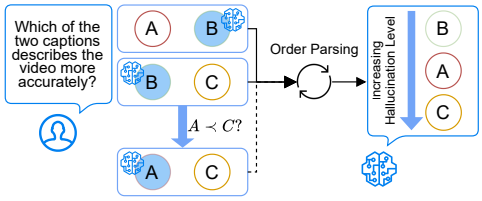

Figure 4: Visual illustration of *relative caption ordering* task in VIDHAL.

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 [13], we adapt the well-established Normalized Discounted Cumulative Gain (NDCG) [17] for hallucination assessment in VIDHAL. Unlike previous metrics like POPE [31], 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 [28]	EVA-CLIP-G	Vicuna	7B	8	0.381	0.475	0.488
LLaMA-VID [30]	EVA-CLIP-G	Vicuna	7B	1fps	0.358	0.486	0.521
VideoChat2 (Vicuna) [29]	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 [61]	SigLIP/SO400M	Qwen2	7B	16	0.596	0.641	0.707
LLaVA-NeXT-Video (7B) [68]	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) [8]	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**
Proprietary Models							
GPT-4o [43]	-	-	-	1fps	<u>0.772</u>	<u>0.840</u>	<u>0.826</u>
Gemini-1.5 (Flash) [46]	-	-	-	1fps	0.657	0.738	0.745
Gemini-1.5 (Pro)	-	-	-	1fps	0.671	0.765	0.753

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 <u>underlined</u> 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 thirteen VLLMs from eight different model families, including six open-source models: VideoChat [28], LLaMA-VID [30], VideoLLaMA2 [8], VideoChat2 [29], mPLUG-Owl3 [61] and LLaVA-NeXT-Video [68], and two proprietary models: GPT-4o [43] and Gemini-1.5 [46]. 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. The input captions in $\mathcal{Y}^i$ were presented in a randomized order 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: i) Proprietary models demonstrate superior results compared to open-sourced models. In particular, GPT-4o achieves the best performance on all tasks, surpassing other models by significant margins. ii) Larger VLLMs generally outperform smaller ones in both tasks. This result is supported by the comparison of different LLM bases for the VideoLLaMA2 and LLaVA-NeXT-Video models. iii) The caption ordering task poses greater difficulty for current VLLMs than MCQA, evidenced by the larger performance margins between the VLLM models and the random baseline. Notably, VideoChat and VideoChat2 (Mistral) show slight to no improvement over the random baseline across both caption ordering tasks. This indicates that current VLLMs greatly suffer from poor fine-grained video understanding and are inclined to generate hallucinations.

Aspect-aware Results. Figure 5 highlights the fine-grained, aspect-specific performance of the evaluated models. 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.

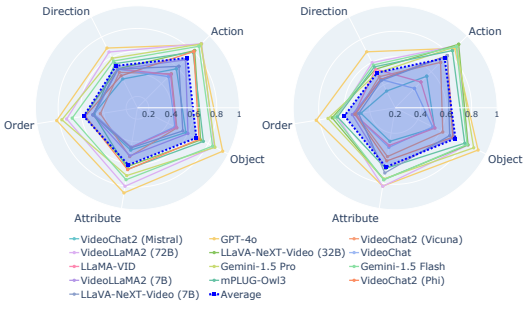

Figure 5: Aspect-specific NDCG scores for the (Left) naive and (Right) relative caption ordering.

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.

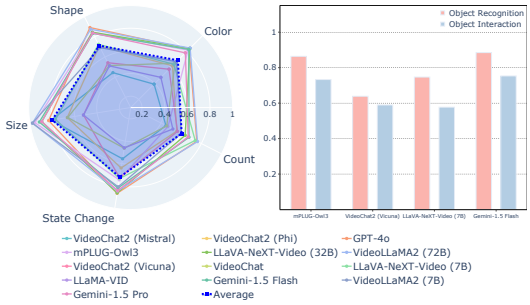

Figure 6: NDCG scores for *Attribute* (Left) and *Object* (Right) sub-aspects in caption ordering.

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 $HM_{3\to1}$, $HM_{3\to2}$, and $HM_{2\to1}$, respectively, with results presented in Figure 7.

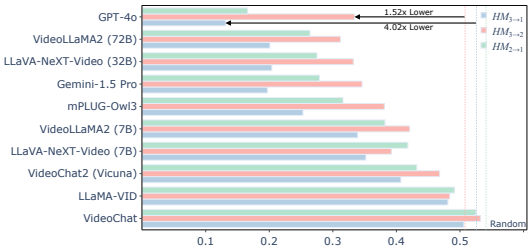

Figure 7: Hallucination misalignment (HM) scores on VIDHAL, with *Random* representing HM scores from the random baseline.

Our findings show that advanced VLLMs, such as VideoLLaMA2 (72B), can generally distinguish positive captions from severely hallucinated ones, as 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 suggests that 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. [31, 53, 54, 15].

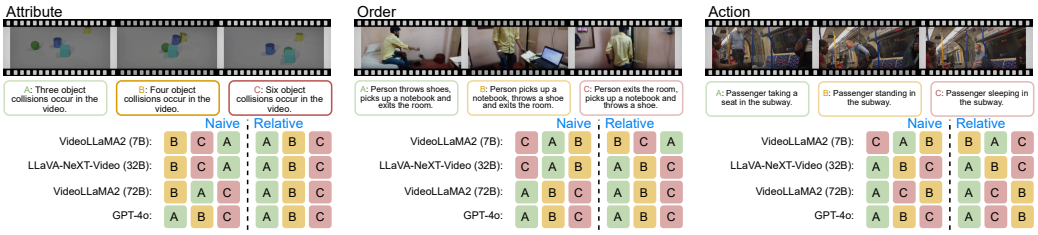

Figure 9: Qualitative examples of VLLM responses on the caption ordering tasks, for the *Attribute*, *Order* and *Action* aspects.

Image Prior Reliance. Previous research shows that VLLMs often rely on image priors for reasoning [21, 3], overlooking key spatiotemporal features. This is exemplified by a few frames having dominant influence on response generation. To examine how this bias affects hallucination generation in video-based VLLMs, we used a video summarization algorithm [48] 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 the 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).

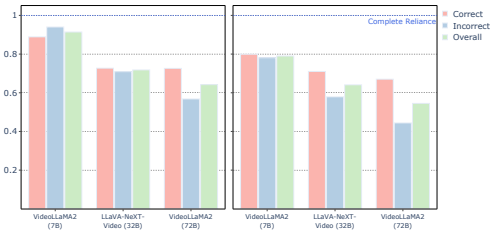

Figure 8: Overlapping ratios of model predictions under single-frame and full-video inputs for correct, incorrect and overall predictions. *Complete Reliance* indicates that the VLLM always produces the same response for both video and single frames.

5.4 Qualitative Results

We conducted a qualitative analysis of responses generated by various VLLMs for the caption ordering task, with examples shown in Figure 9. We observe that: i) Relative caption ordering generally guides VLLMs to produce more accurate responses, as evidenced by improvements from naive to relative caption order predictions in most cases. ii) Advanced VLLMs exhibit more stable performance across both ordering tasks, with lower variation in predictions between both sub-tasks.

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 thirteen 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 [38, 24, 29, 23, 27] or response evaluation [15, 50, 32]. To reduce over-alignment to GPT-4o's preferences, we leverage additional strong LLMs, including Gemini-1.5 [46] and LLaMA2 (70B) [52] to assess and filter generated captions. While this improves the robustness of the annotations, we recognize that fully mitigating LLM-induced biases in caption generation remains an open challenge.

References

[1] 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.

[2] 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*, pages 13927–13937. IEEE, 2024.

[3] 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*, pages 2907–2917. IEEE, 2022.

[4] Qingxing Cao, Junhao Cheng, Xiaodan Liang, and Liang Lin. Visdiahalbench: A visual dialogue benchmark for diagnosing hallucination in large vision-language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 12161–12176. Association for Computational Linguistics, 2024.

[5] 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, editors, *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 3235–3252. Association for Computational Linguistics, 2024.

[6] 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.

[7] 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, 2024.

[8] 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.

[9] 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.

[10] 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.

[11] 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*, pages 14303–14312. IEEE, 2024.

[12] 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.

[13] 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.

[14] 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*, pages 7696–7712. Association for Computational Linguistics, 2024.

[15] 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*, pages 14375–14385. IEEE, 2024.

[16] 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*, pages 13418–13427. IEEE, 2024.

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

[18] 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*, pages 27026–27036. IEEE, 2024.

[19] Chaoya Jiang, Wei Ye, Mengfan Dong, Hongrui Jia, Haiyang Xu, Ming Yan, Ji Zhang, and Shikun Zhang. Hal-eval: A universal and fine-grained hallucination evaluation framework for large vision language models. *CoRR*, abs/2402.15721, 2024.

[20] 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*, pages 27218–27228. IEEE, 2024.

[21] 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, editors, *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 487–507. Association for Computational Linguistics, 2023.

[22] 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*, pages 13872–13882. IEEE, 2024.

[23] 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, 2023.

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

[25] 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*, pages 19730–19742. PMLR, 2023.

[26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR, 2022.

[27] 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*, pages 6449–6464. Association for Computational Linguistics, 2023.

[28] 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, 2023.

[29] 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*, pages 22195–22206. IEEE, 2024.

[30] 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, pages 323–340. Springer, 2024.

[31] 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, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 292–305. Association for Computational Linguistics, 2023.

[32] 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, 2024.

[33] 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, 2024.

[34] 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, 2024.

[35] 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 2024.

[36] 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, editors, *Advances in Neural Information Processing Systems*, 2023.

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

[38] 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, 2024.

[39] 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*, 2024.

[40] 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*, pages 8731–8772. Association for Computational Linguistics, 2024.

[41] 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*, pages 13151–13160. IEEE, 2024.

[42] 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.

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

[44] 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.

[45] 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.

[46] 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.

[47] 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, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045. Association for Computational Linguistics, 2018.

[48] 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*, pages 18847–18856. IEEE, 2024.

[49] 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.

[50] 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, editors, *Findings of the Association for Computational Linguistics*, pages 13088–13110. Association for Computational Linguistics, 2024.

[51] Kim Sung-Bin, Oh Hyun-Bin, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. *arXiv preprint arXiv:2410.18325*, 2024.

[52] 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.

[53] 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.

[54] 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, 2024.

[55] Zhecan Wang, Garrett Bingham, Adams Yu, Quoc V. Le, Thang Luong, and Golnaz Ghiasi. Haloquest: A visual hallucination dataset for advancing multimodal reasoning. *CoRR*, abs/2407.15680, 2024.

[56] 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.

[57] 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*, pages 1167–1181. Association for Computational Linguistics, 2024.

[58] 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*, 2024.

[59] Ziwei Xu, Sanjay Jain, and Mohan S. Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *CoRR*, abs/2401.11817, 2024.

[60] 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.

[61] 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.

[62] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality. *CoRR*, abs/2304.14178, 2023.

[63] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *CoRR*, abs/2311.04257, 2023.

[64] 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.

[65] 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, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1768–1785. Association for Computational Linguistics, 2024.

[66] 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, editors, *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 11766–11781. Association for Computational Linguistics, 2024.

[67] 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, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 543–553. Association for Computational Linguistics, 2023.

[68] 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.

[69] 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.

[70] Weihong Zhong, Xiaocheng Feng, Liang Zhao, Qiming Li, Lei Huang, Yuxuan Gu, Weitao Ma, Yuan Xu, and Bing Qin. Investigating and mitigating the multimodal hallucination snowballing in large vision-language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 11991–12011. Association for Computational Linguistics, 2024.

[71] 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, 2024.

[72] 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, 2024.

[73] 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, 2024.

[74] 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*, 2024.

[75] 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, 2024.

[76] 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*, 2024.

[77] 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, 2024.

[78] 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*, pages 17984–18003. Association for Computational Linguistics, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper introduces a benchmark for evaluating video hallucinations and proposes a novel evaluation task, with analysis supported through experimental results. This reflects the contributions outlined in the introduction and abstract.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Potential biases in the dataset construction process are discussed in the limitations section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: No theoretical results are included in the paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The benchmark dataset, evaluation codebase, and accompanying documentation and instructions are provided with the paper as part of the submission.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The benchmark dataset, evaluation suite, and accompanying documentation and instructions are provided with the paper as part of the submission.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The inference hyperparameters are described in the experimental settings, with full implementation details provided in the accompanying submitted code.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Random variability is minimized through the use of deterministic processes, such as during decoding. No statistical significance tests are conducted in our experimental analysis.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The compute resources used for all experiments are detailed in the experimental settings section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have adhered to the NeuRIPs Code of Ethics when constructing our benchmark dataset.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We briefly highlight the importance of studying video hallucinations to support the development of robust algorithms for video reasoning applications in the conclusion section.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: Extensive manual inspection has been conducted to ensure the dataset does not pose risks of improper or unintended use.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All public datasets and models used in the dataset construction and model evaluation processes are properly cited in our paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The benchmark dataset, evaluation codebase, and accompanying documentation and instructions are provided with the paper as part of the submission.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: Human validation is conducted by a third-party vendor using their proprietary tools and platform. The vendor ensures that all contributors are properly compensated for their work.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: All annotators involved in the human validation process are properly informed of any associated risks by the engaged vendor.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The process of using GPT-4o to prepare annotations for our dataset is detailed in both the main text and supplementary material.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.