SmartSight: Mitigating Hallucination in Video-LLMs Without Compromising Video Understanding via Temporal Attention Collapse

Yiming Sun, Mi Zhang, Feifei Li, Geng Hong†, Min Yang†

College of Computer Science and Artificial Intelligence, Fudan University, Shanghai, China
ymsun24@m.fudan.edu.cn, mi_zhang@fudan.edu.cn, ffli23@m.fudan.edu.cn,
ghong@fudan.edu.cn, m_yang@fudan.edu.cn

Abstract

Despite Video Large Language Models (Video-LLMs) having rapidly advanced in recent years, perceptual hallucinations pose a substantial safety risk, which severely restricts their real-world applicability. While several methods for hallucination mitigation have been proposed, they often compromise the model's capacity for video understanding and reasoning. In this work, we propose SmartSight, a pioneering step to address this issue in a training-free manner by leveraging the model's own introspective capabilities. Specifically, SmartSight generates multiple candidate responses to uncover low-hallucinated outputs that are often obscured by standard greedy decoding. It assesses the hallucination of each response using the *Temporal Attention Collapse* score, which measures whether the model over-focuses on trivial temporal regions of the input video when generating the response. To improve efficiency, SmartSight identifies the *Visual Attention Vanishing* point, enabling more accurate hallucination estimation and early termination of hallucinated responses, leading to a substantial reduction in decoding cost. Experiments show that SmartSight substantially lowers hallucinations for Qwen2.5-VL-7B by 10.59% on VRIPT-HAL, while simultaneously enhancing video understanding and reasoning, boosting performance on VideoMMMU by 8.86%. These results highlight SmartSight's effectiveness in improving the reliability of open-source Video-LLMs.

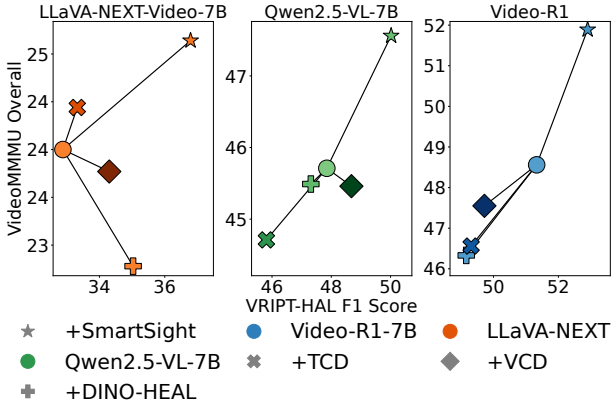

Figure 1: Comparison between SmartSight and existing methods in terms of hallucinations suppression, video understanding and reasoning. Lighter colors indicate lower computational cost. Existing methods suffer from two main limitations: (1) inferior transferability across models, showing effectiveness only on LLaVA-NEXT-Video; and (2) reducing hallucinations at the cost of impaired video understanding. SmartSight simultaneously suppresses hallucinations and enhances video comprehension, achieving a more favorable balance between accuracy and efficiency.

1 Introduction

Recently, Video Large Language Models (Video-LLMs) have demonstrated remarkable capabilities in visual perception, semantic understanding, and complex reasoning (Hu et al. 2025; Sun et al. 2024). This has led to their widespread deployment in real-world applications such as autonomous driving and robotic manipulation (Zhao et al. 2025). Despite their impressive performance, Video-LLMs still face a major challenge: **perception hallucinations**. Specifically, they may misrepresent object attributes, fabricate nonexistent content, or overlook critical video segments. This poses serious risks to the use of Video-LLMs, especially in safety-critical scenarios.

Although a few works have attempted to mitigate perception hallucinations in video perception by fine-tuning Video-LLMs (Yang et al. 2024) or applying training-free methods such as Visual Contrastive Decoding (VCD) (Leng et al. 2024), we found that these methods significantly compromise the model's essential ability for video understanding. As shown in Figure 1, they suppress hallucinations only on LLaVA-Next-Video, while severely impairing the model's ability to understand and reason. One contributing factor is that inappropriate fine-tuning leads to catastrophic forgetting (Li et al. 2024b), while training-free methods are prone to misidentifying expected outputs as hallucinations and introduce significant computational overhead during inference (Huo et al. 2025). In this work, we take an initial step toward bridging the gap between (1) mitigating perception hallucinations and (2) preserving understanding and reasoning capabilities in Video-LLMs.

Our method is grounded in several key observations. We reveal that hallucinated outputs are associated with a collapse of attention onto a single frame or a semantically triv-

†Corresponding authors: Geng Hong and Min Yang.

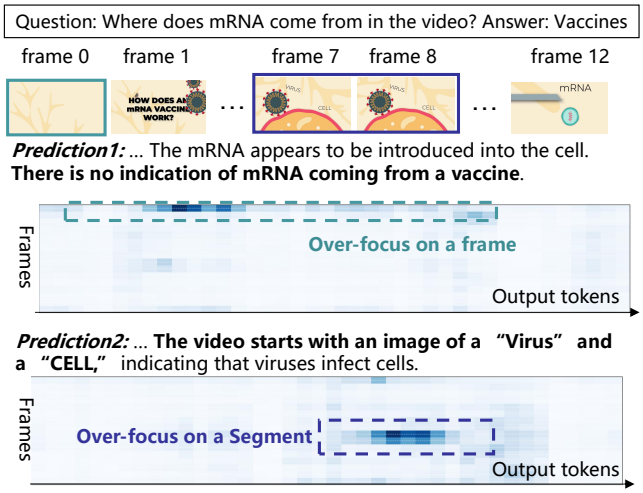

Figure 2: Illustration of hallucination in Video-LLM outputs. Prediction 1 exhibits an over-reliance on the first frame, causing it to overlook crucial information in frame 12 and leading to an incorrect output. Prediction 2 concentrates excessively on visually similar segments (frames 7–8), which results in a misinterpretation of the video content.

ial segment – a phenomenon we term **Temporal Attention Collapse (TAC)**. As shown in Figure 2, in *Prediction 1*, model allocates excessive attention to a single frame, causing it to overlook other important parts of the video. In *Prediction 2*, it disproportionately focuses on a segment with limited variation in object appearance or motion, leading to a misinterpretation of event order. It is worth noting that over-focused frames or segments are not necessarily equivalent to key frames, as models need to leverage the video context to understand semantics before identifying key frames. One potential cause of TAC is the image bias inherent in Video-LLMs, where certain frames resemble images in the training data, thereby drawing disproportionate attention from the model (Yu et al. 2025). Furthermore, sampling multiple responses uncovers the model's capacity to generate outputs with *fewer hallucinations*, which is often obscured by greedy decoding. As shown in Figure 3, for each question, at least one of the ten sampled responses contains fewer hallucinations than the response generated with greedy decoding. In addition, we observe a turning point during generation from which the model's attention to visual information in the input video drops sharply – referred to as **Visual Attention Vanishing (VAV) Point**. It marks a specific token after which attention to visual inputs has already substantially diminished. The majority of attention to visual tokens is allocated before the VAV point. This reflects an inherent property of rotary positional encoding, where attention scores naturally decay as the relative distance between tokens increases (Su et al. 2024).

Based on these findings, we propose a training-free method, dubbed *SmartSight*, that effectively mitigates hallucinations while enhancing the model's understanding and reasoning abilities. SmartSight generates multiple candidate

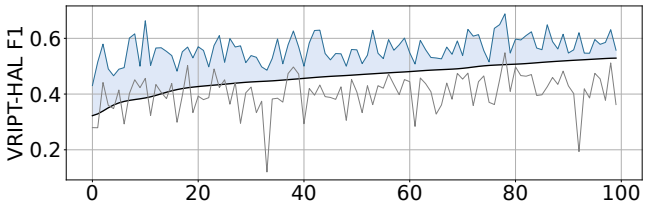

Figure 3: Comparison between greedy decoding and sampling N=10 responses per query. We visualize 100 randomly selected responses from VRIPT-HAL. The black curve shows the hallucination level of responses generated by greedy decoding. The blue and gray curves indicate the least and most hallucinated among the sampled responses. The shaded region illustrates that sampling can produce responses with lower hallucination.

responses to elicit less hallucinated outputs that are often obscured by standard greedy decoding (Wan et al. 2025). To facilitate the selection of a less hallucinated response, we propose the *Temporal Attention Collapse* Score (TAC score) as an introspective metric for evaluating the hallucination severity of each candidate. Building on the observation that hallucinated outputs tend to over-focus on semantically trivial video segments or frames, the TAC score evaluates hallucinations without relying on external models. To improve efficiency, we further assess response quality at the proposed *Visual Attention Vanishing* (VAV) Point, which enables accurate estimation of hallucination severity via the TAC score without waiting for full generation. Responses with low TAC score at this point are terminated early, reducing the computational overhead in the decoding stage by up to 79.6%, with negligible impact on overall response quality. Experimental results demonstrate that our method consistently enhances the understanding and reasoning capabilities of 10 diverse Video-LLMs. With SmartSight, Qwen2.5-VL-7B achieves performance competitive with Qwen2.5-VL-32B and proprietary Gemini 1.5 Pro. As the model size increases, SmartSight exhibits a favorable scaling property not observed in existing methods. Overall, these findings highlight SmartSight as a robust and plug-and-play solution for improving the reliability of Video-LLMs in more challenging scenarios, such as harmful video detection, video-based jailbreak attack defense, and video quality assessment.

In summary, our main contributions are threefold:

- We reveal that current hallucinations mitigation methods may unintentionally impair the semantic understanding and reasoning capabilities of Video-LLMs.
- We uncover the Temporal Attention Collapse phenomenon in hallucinated responses, characterized by excessive attention to trivial temporal regions and consistently observable across a variety of Video-LLMs.
- We propose SmartSight, a pioneering step to alleviate perception hallucinations and improve video understanding and reasoning capabilities in a training-free manner, with impressive scalability and compatibility across a wide range of Video-LLMs.

2 Related Works

2.1 Video Large Language models

In recent years, Video Large Language Models (Video-LLMs) have made rapid progress, leading to a wide range of real-world applications (You et al. 2025; Sun, Li, and Wang 2023; Chen et al. 2025). Video-LLMs encode video frames using a vision encoder, concatenate their embeddings across frames, and project them into the LLM's linguistic embedding space via a cross-modal adapter (Bai et al. 2025). This design enables Video-LLMs to perceive video and understand its semantics (Hu et al. 2025). To further improve their capacity for complex reasoning, recent efforts such as Video-R1 (Feng et al. 2025), VideoChat-R1 (Li et al. 2025b), and TinyLLaVA-Video-R1 (Zhang et al. 2025b) have introduced reinforcement learning and test-time scaling strategies, inspired by the success of large-scale reasoning models like DeepSeek-R1 (Guo et al. 2025). Our method further improves the understanding and reasoning capabilities of these models by suppressing hallucinations.

2.2 Evaluation for Hallucination in Video-LLMs

Despite their impressive performance, Video-LLMs are more prone to hallucinations than image-based models, as the temporal dimension introduces unique challenges. In addition to hallucinations in image tasks, Video-LLMs may misidentify actions, confuse the sequence of events, omit essential segments, or fabricate scenes that are not present in the video. Existing hallucination evaluation methods for Video-LLMs primarily fall into two categories: VQA-based (Zhang et al. 2025a) and caption-based approaches (Yang et al. 2024). VQA-based methods formulate multiple-choice questions that query the model about the presence of plausible yet incorrect objects or events in a given video, thereby assessing whether the model hallucinates content not grounded in the visual input.

While straightforward, Video-LLMs tend to exhibit more hallucinations in open-ended text generation tasks (e.g., video captioning) than in multiple-choice verification (Rawal et al. 2025). To address this limitation, caption-based evaluation methods quantify hallucination occurrence by comparing model-generated captions with human-provided references (Yang et al. 2024). However, these methods typically fail to account for the impact of hallucination mitigation strategies on the model's utility, such as its capacity for semantic understanding and reasoning. Our proposed method demonstrates competitive results under both evaluation protocols, effectively reducing hallucinations and enhancing the model's usability.

2.3 Mitigation for Hallucination in Video-LLMs

Existing approaches for mitigating hallucinations in Video-LLMs can be broadly categorized into training-based and training-free methods. Training-based methods, such as Vriptor (Yang et al. 2024), retrain models with densely annotated video captions to enhance fine-grained visual perception. MASH-VLM (Bae et al. 2025) and Vista-LLaMA (Ma et al. 2023) modify the self-attention mechanism to better capture temporal dependencies, which requires model retraining. PaMi-VDPO (Ding et al. 2025) constructs preference data via data augmentation and applies Direct Preference Optimization to fine-tune Video-LLMs. While effective, these methods involve architectural changes and fine-tuning, which limit the generalizability of the fine-tuned model across domains such as mathematical reasoning.

To overcome these limitations, recent efforts have explored training-free hallucination mitigation. Temporal Contrastive Decoding (Zhang et al. 2025a), inspired by Visual Contrastive Decoding (Leng et al. 2024), perturbs the input video to amplify potential hallucinations, then compares the outputs before and after perturbation to identify hallucinated tokens. However, improper perturbations may misidentify correct content as hallucinations. DINO-HEAL (Li, Im, and Fazli 2025) addresses this by leveraging DINOv2 (Oquab et al. 2023) saliency maps to highlight key regions and suppress attention to background content. However, we find that both training-based and training-free methods may undermine model usability. Furthermore, their effectiveness does not generalize well to reasoning-oriented models such as Video-R1 (Feng et al. 2025).

3 Methodology

Overview: As shown in Figure 4 (A), SmartSight elicits the model's self-introspective capabilities to generate outputs with fewer hallucinations by sampling multiple responses. To evaluate the faithfulness of each response to the video content, we introduce Temporal Attention Collapse (TAC) Score (Sec. 3.2, Figure 4 (C)), which assesses hallucination severity without relying on external models. We propose the Visual Attention Vanishing (VAV) Point in Sec. 3.3 (Figure 4 (B)), which facilitates early estimation of hallucination severity before full generation. Responses with higher hallucination levels are terminated in advance, thereby improving overall efficiency. For clarity, we first introduce the notations used throughout this section in Sec. 3.1.

3.1 Preliminary

Input Formulation. Video-LLMs[1] receive a sequence of T frames $\mathcal{V} = \langle f^t \rangle_{t=1}^{T}$ together with a textual question $\mathcal{X} = \langle x_l \rangle_{l=1}^{L}$ as input. They firstly divide each frame f^t into M patches $f^t = \langle p_m^t \rangle_{m=1}^{M}$, where each patch is then encoded into a visual token v_m^t. Consequently, a video with T frames yields $T \times M$ visual tokens in total. The textual input is also encoded into textual tokens: $\mathcal{F}^{text} = \langle t_l \rangle_{l=1}^{L}$. The visual tokens $\mathcal{F}^{vid} = \langle v_m^t \rangle_{m=1,t=1}^{M,T}$ and the text tokens are concatenated to form the full input $\mathcal{F} = \mathcal{F}^{vid} \cup \mathcal{F}^{text}$ to the language model.

Stochastic Sampling. Let p be a language model. Given input $\mathcal{F}$, the model generates a response $y = (y_1, y_2, \ldots, y_l)$ auto-regressively, where l is the number of tokens. At each decoding step k, the next token is sampled from: $y_{k+1} \sim p(\cdot \mid \mathcal{F}, y_{\leq k})$. Unlike standard greedy decoding, which deterministically selects the token with the highest probabil-

[1]Our study focuses on visual information, excluding subtitles or audio tracks.

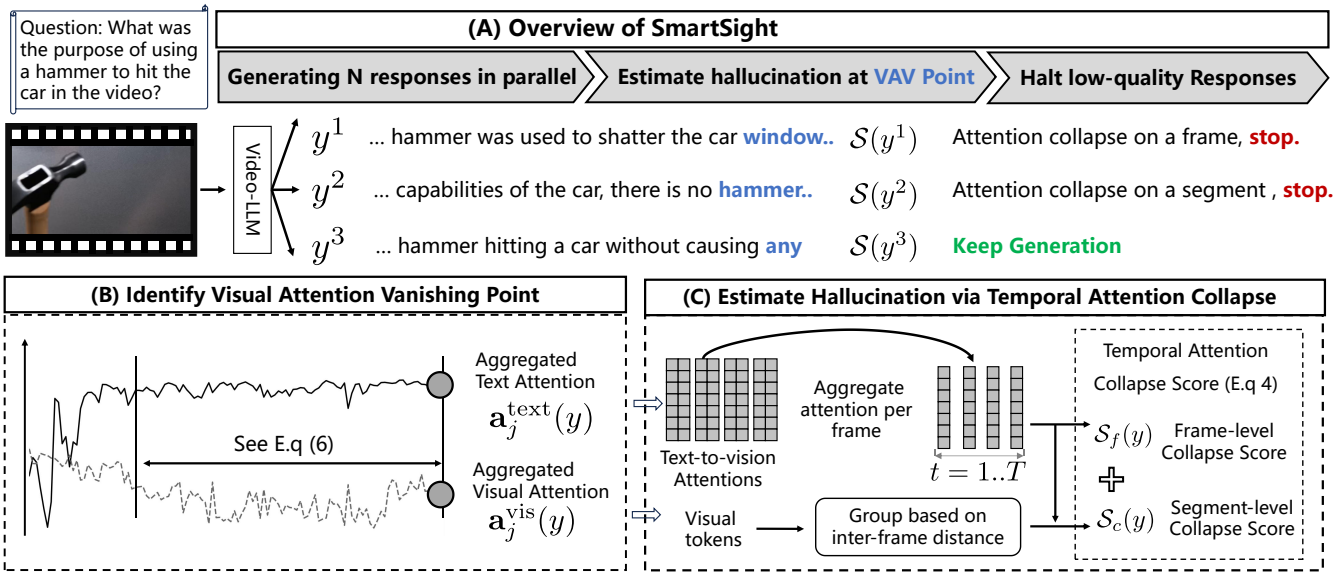

Figure 4: Overview of the proposed SmartSight. Given an input video and a textual query, SmartSight generates N responses in parallel. During generation, it dynamically detects the Visual Attention Vanishing Point and estimates hallucination severity using the proposed Temporal Attention Collapse Score. Only one high-quality candidate is retained for continued generation, thereby achieving a favorable balance between efficiency and effectiveness.

ity, sampling from the distribution promotes diversity and avoids linguistic bias. We denote the full response sampling process as $y \sim p(\cdot \mid \mathcal{F})$. We then sample N i.i.d. responses: $y^1, \ldots, y^N \sim p(\cdot \mid \mathcal{F})$, where y_i^n is the i-th token of the n-th response (Renze 2024).

3.2 Self-Introspective Hallucination Estimation via Temporal Attention Collapse

In this section, we leverage the observed phenomenon of Temporal Attention Collapse to evaluate hallucinations solely on the model's internal attention distribution. Given a generated response $y = (y_1, \ldots, y_L)$ with L tokens, we compute the average attention received by each frame during the generation of y, denoted as: $\mathbf{a}_t(y) = \frac{1}{ML}\sum_{m=1}^{M}\sum_{j=1}^{L} a_{j,(t,m)}(y)$. To simplify notation, we refer to $\mathbf{a}_{j,(t,i)}(y)$ and $\mathbf{a}_t(y)$ as $\mathbf{a}_{j,(t,i)}$ and $\mathbf{a}_t$, respectively, throughout this section. Next, we evaluate hallucination severity from two aspects: Frame-level Collapse and Segment-level Collapse.

Frame-level Collapse. To quantify whether attention is disproportionately concentrated on a single frame, we define the Frame-level Collapse Score $\mathcal{S}_f(y)$ for the response y :

$$\mathcal{S}_f(y) = -\sum_{t=1}^{T} \hat{\mathbf{a}}_t \log \hat{\mathbf{a}}_t, \quad \hat{\mathbf{a}}_t = \frac{\mathbf{a}_t}{\sum_{t'=1}^{T} \mathbf{a}_{t'}} \tag{1}$$

The entropy of the normalized attention scores is used to assess the distribution of attention across frames. A lower $\mathcal{S}_f(y)$ indicates a sharp concentration of attention on one or a few isolated frames, suggesting a stronger tendency toward Frame-level Collapse.

Segment-level Collapse. To assess whether attention is disproportionately allocated to segments with little variation in object appearance or motion, we introduce the Segment-level Collapse Score $\mathcal{S}_c(y)$. We begin by analyzing inter-frame similarity based on video tokens $\mathcal{F}^{\text{vid}}$ to identify the trivial video segments. However, a direct use of cosine similarity between adjacent frames struggles to account for motion and fine-grained appearance differences. To address the issue, we define a Motion-Aware Cost Matrix:

$$C_{ij}^{(t)} = \left(1 - \cos\left(v_i^{t-1}, v_j^{t}\right)\right) + \left\|\boldsymbol{u}_{t,j} - \boldsymbol{u}_{t-1,i}\right\|_2 \tag{2}$$

Here, v_j^t denotes the visual embedding of the j-th patch in frame t, and $\boldsymbol{u}_{t,j}$ is its normalized spatial coordinate. The first term measures appearance dissimilarity between patch i in frame $t-1$ and patch j in frame t via cosine distance, while the second term captures their relative motion based on spatial displacement. Using this cost matrix, we define the inter-frame distance d_t as the minimum total matching cost under a one-to-one correspondence between patches:

$$d_t = \sum_{i=1}^{N} C_{i,\pi^*(i)}^{(t)}, \quad \pi^* = \arg\min_{\pi \in \mathcal{B}_N} \sum_{i=1}^{N} C_{i,\pi(i)}^{(t)} \tag{3}$$

where $\mathcal{B}_N$ denotes all possible one-to-one matchings between patches in frame $t-1$ and frame t.

A higher d_t suggests a larger degree of dissimilarity between the two frames. We segment the video into K temporally contiguous parts $\{s_k | k \in \{1, \ldots, K\}\}$ by selecting frame t as a boundary if its inter-frame distance d_t exceeds a predefined threshold, i.e., $d_t > \gamma$. For each segment, the attention it receives is aggregated and normalized as $\hat{\mathbf{a}}_{s_k}(y) = \mathbf{a}_{s_k}(y) / \sum_{\ell=1}^{K} \mathbf{a}_{s_\ell}(y)$, where $\mathbf{a}_{s_k}(y) = \sum_{t \in S_k} \mathbf{a}_t(y)$. The Segment-level Collapse Score $\mathcal{S}_c(y)$ is defined as the entropy of the resulting attention distribu-

tion: $\mathcal{S}_c(y) = -\sum_{k=1}^{K} \hat{\mathbf{a}}_{s_k}(y) \log \hat{\mathbf{a}}_{s_k}(y)$. Finally, the overall Temporal Attention Collapse Score is defined as:

$$\mathcal{S}(y) = \mathcal{S}_f(y) + \mathcal{S}_c(y) \tag{4}$$

A lower $\mathcal{S}(y)$ indicates that the model's attention is more sharply concentrated on isolated frames and semantically trivial segments, suggesting a higher degree of hallucination.

3.3 Efficient Generation via Early Stopping at Visual Attention Vanishing Point

We begin by generating N responses in parallel, denoted as $Y = \{y^1, y^2, \ldots, y^N\}$, where each $y^n \sim p(\cdot \mid \mathcal{F})$ is a response independently sampled from the model. This helps the model to explore diverse reasoning paths and break free from over-reliance on its biased prior, resulting in outputs that better align with the visual content.

To improve efficiency, we assess the quality of each sampled response based on its partial output truncated at the Visual Attention Vanishing (VAV) Point, instead of completing the entire generation. The VAV Point marks the stage when attention to visual tokens has substantially diminished. To identify this point on-the-fly during the generation of n-th response y^n, we compute $\mathbf{a}_j^{\text{vis}}(y^n)$ and $\mathbf{a}_j^{\text{text}}(y^n)$, which denote the attention that the j-th token assigns to visual tokens and preceding text tokens, respectively.

$$\mathbf{a}_j^{\text{vis}}(y^n) = \sum_{t=1}^{T} \sum_{m=1}^{M} a_{j,(t,m)}(y^n),$$
$$\mathbf{a}_j^{\text{text}}(y^n) = \sum_{l=1}^{j-1} a_{j,l}(y^n) \tag{5}$$

Here, $a_{j,(t,m)}(.)$ denotes the attention weight from the j-th output token to the m-th visual token v_m^t in t-th frame, averaged across all attention heads and layers. Similarly, $a_{j,l}(.)$ represents the averaged attention from the j-th token to the l-th generated token.

We define the VAV Point of response y^n as the earliest generation step $j_{\text{vav}}(y^n)$ at which the model's attention to visual tokens remains low for w consecutive steps:

$$j_{\text{vav}}(y^n) = \min \left\{ j \;\middle|\; \frac{\mathbf{a}_k^{\text{vis}}(y^n)}{\mathbf{a}_k^{\text{text}}(y^n)} < \alpha,\ \forall k \in [j-w+1, j] \right\} \tag{6}$$

The ratio $\mathbf{a}_k^{\text{vis}}(y^n)/\mathbf{a}_k^{\text{text}}(y^n)$ indicates the model's relative attention to visual input versus previously generated text at step k. By requiring the ratio to remain below the threshold α for w consecutive tokens, we ensure the VAV Point is not triggered by short-term fluctuations in attention. We then compute the TAC Score (Eq. 4) for each sampled response truncated at its VAV Point and select the top-1 candidates:

$$n^* = \arg\max_n \left(\mathcal{S}(y^n_{1:j_{\text{vav}}(y^n)}) \;\middle|\; n = 1, \ldots, N \right) \tag{7}$$

The selected candidate response y^{n^*} is then allowed to complete generation and is used as the final output.By terminating the generation of suboptimal candidates early,we achieve a favorable trade-off between performance and efficiency.

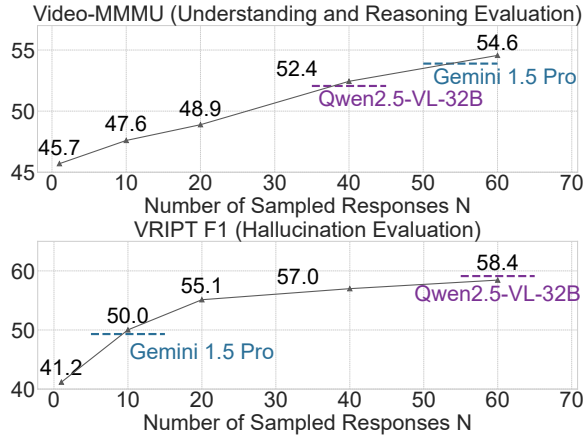

Figure 5: Results of applying SmartSight to Qwen2.5-VL-7B with different values of N. The figure shows that increasing N enables test-time scaling, which is challenging to achieve with prior methods. When $N = 60$, the 7B model achieves performance comparable to the Qwen2.5-VL-32B model and the proprietary Gemini 1.5 Pro.

4 Experiments

4.1 Experimental Settings

Implementation Details. We use temperature sampling (Renze 2024) with a temperature of 0.7 to generate N=10 candidate responses before early stopping. Hyperparameters α and w are set to 1.2 and 10, respectively. All methods use 32 sampled frames with a fixed resolution of 392 for fair comparison.

Benchmarks. We assess hallucination mitigation on the caption-based VRIPT-HAL (Yang et al. 2024) and VQA-based EventHallusion (Zhang et al. 2025a) benchmarks. To evaluate video understanding and reasoning capabilities, we utilize two large-scale, long-form, and open-domain benchmarks: Video-MME (Fu et al. 2025) and Video-MMMU (Hu et al. 2025). For the VQA task, all methods adopt the same chain-of-thought (CoT) prompt and the official evaluation code. Following standard practice for isolating the contribution of the visual modality (Fu et al. 2025), we do not include subtitles during evaluation.

Baseline Methods. We compare SmartSight against three representative hallucination mitigation methods: two approaches specifically designed for video understanding, TCD (Zhang et al. 2025a) and DINO-HEAL (Li, Im, and Fazli 2025), as well as VCD (Leng et al. 2024), a method initially developed for images but transferable to video tasks. All methods are implemented using the default settings as described in their original papers.

4.2 Main Results

We evaluate our method on three representative Video-LLMs: LLaVA-NEXT-Video (Li et al. 2025a), Qwen2.5-VL (Bai et al. 2025), and Video-R1 (Feng et al. 2025).

Effectiveness in Hallucination Mitigation. Our method achieves SOTA hallucination mitigation across both VQA-based and caption-based evaluations. In contrast, exist-

	VRIPT-HAL			EventHallu.	Video-MME				Video-MMMU			
	F1 (Δ)	R	P	Acc (Δ)	Overall (Δ)	Long	Med.	Short	Overall (Δ)	Perc.	Compr.	Adap.
Qwen2.5-VL	47.9	35.6	75.0	60.4	54.8	45.4	52.5	66.5	45.7	56.0	50.0	31.0
+Dinoheal	47.3 (-0.54)	35.0	75.9	56.8 (-3.53)	54.6 (-0.19)	47.4	51.9	64.5	45.5 (-0.22)	58.3	48.0	30.1
+TCD	45.8 (-2.04)	33.4	76.2	61.4 (+1.01)	54.7 (-0.09)	44.1	53.1	66.8	44.7 (-1.00)	54.7	47.7	31.6
+VCD	48.7 (+0.83)	36.0	75.4	61.9 (+1.51)	54.8 (-0.02)	45.1	52.3	67.0	45.5 (-0.25)	59.3	46.5	30.5
+SmartSight	50.0 (**+2.17**)	38.5	73.8	63.1 (**+2.78**)	56.2 (**+1.39**)	49.3	53.2	66.1	47.6 (**+1.85**)	59.7	51.3	31.7
Video-R1	51.3	40.3	71.8	62.6	59.8	50.8	59.7	69.0	48.6	64.0	47.3	34.3
+Dinoheal	49.2 (-2.17)	37.8	71.3	62.6 (+0.01)	59.2 (-0.59)	50.1	58.6	69.0	46.3 (-2.23)	62.0	46.3	30.7
+TCD	49.3 (-2.02)	37.6	72.9	64.9 (+2.27)	59.7 (-0.07)	50.1	59.1	70.0	46.6 (-2.00)	63.7	44.3	31.7
+VCD	49.7 (-1.61)	38.1	72.9	65.4 (+2.78)	59.8 (+0.03)	50.2	59.9	69.4	47.5 (-1.01)	64.0	45.7	33.0
+SmartSight	52.9 (**+1.57**)	43.0	69.7	65.5 (**+2.89**)	60.2 (**+0.41**)	51.1	59.1	70.4	51.9 (**+3.33**)	66.0	50.7	39.0
LLaVA-NEXT	32.9	22.0	69.0	53.4	28.8	25.9	32.1	28.4	24.0	26.7	18.9	26.3
+Dinoheal	35.0 (+2.15)	24.3	66.6	53.9 (+0.51)	28.3 (-0.50)	28.1	28.1	28.6	22.8 (-1.19)	27.0	18.0	23.3
+TCD	34.3 (+1.42)	23.2	69.2	55.7 (+2.36)	31.1 (+2.39)	27.9	30.9	34.7	23.8 (-0.20)	27.3	16.7	27.3
+VCD	33.3 (+0.44)	22.5	68.7	53.1 (-0.25)	30.6 (+1.80)	27.6	30.6	33.6	24.4 (+0.47)	23.7	20.3	29.3
+SmartSight	36.8 (**+3.91**)	26.0	67.6	55.8 (**+2.43**)	34.0 (**+5.22**)	31.2	34.8	35.9	25.1 (**+1.17**)	27.1	22.0	26.3

Table 1: Comparison of state-of-the-art mitigation methods including DINO-HEAL (CVPR'25), TCD (arXiv'25), and VCD (CVPR'24) on three Video-LLMs: Qwen2.5-VL-7B, Video-R1, and LLaVA-NEXT-Video-7B. VRIPT-HAL and EventHallusion (binary QA subset) evaluate hallucination suppression, while VideoMME and VideoMMMU assess understanding and reasoning capabilities. Higher scores indicate better performance. The most significant improvements are highlighted in **bold**.

Qwen2.5-VL		Base	TCD	VCD	SmartSight
3B	VRIPT-HAL	28.3	28.8	29.0	30.7
	Video-MMMU	37.3	34.9	35.0	38.3
7B	VRIPT-HAL	47.9	45.8	48.7	50.0
	Video-MMMU	45.7	44.7	45.5	47.6
32B	VRIPT-HAL	59.1	58.4	57.2	59.4
	Video-MMMU	52.1	49.9	45.3	52.2

Table 2: Performance across different model sizes on VRIPT-HAL and VideoMMMU benchmarks. Base denotes standard greedy decoding. SmartSight maintains its effectiveness and efficiency trade-off as model size increases.

Model	Base	+ Ours	Gain
Gemini1.5 Pro	49.3	-	-
Claude3.5-Sonnet	44.6	-	-
Video-LLaVA	24.6	25.7	1.1
LLaVA-OneVision-7B	25.8	31.4	5.6
InternVL3-8B	30.4	34.1	3.7
Qwen2.5-VL-7B	47.9	50.0	2.1
VideoChat-R1	48.4	52.1	3.7

Table 3: SmartSight is applied to a range of video-language models and consistently mitigates hallucinations on the VRIPT-HAL benchmark. Notably, several open-source models equipped with SmartSight achieve performance on par with proprietary models.

ing approaches improve performance only on closed-ended VQA tasks, while exhibiting modest or even negative gains in open-ended captioning. Moreover, these methods typically fail to generalize across different Video-LLMs.

Effectiveness in Complex Reasoning. On the challenging video understanding and reasoning benchmarks Video-MMMU and Video-MME, almost all baseline mitigation methods consistently degrade performance. Conversely, by leveraging reasoning paths that are more faithful to the video's content, our method improves the reasoning capabilities of nearly all evaluated Video-LLMs.

Generalization on Different Video-LLMs. As shown in Table 3, SmartSight demonstrates effective hallucination mitigation across various video-language models, including Video-LLaVA (Lin et al. 2023), LLaVA-OneVision-7B (Li et al. 2024a), and VideoChat-R1 (Li et al. 2025b). Notably, several open-source models achieve performance on par with proprietary counterparts after applying SmartSight. For example, VideoChat-R1 and Qwen2.5-VL-7B surpass Gemini 1.5 Pro. These results underscore the strong generalization capability of SmartSight across diverse architectures.

4.3 Computational and Scaling Efficiency

Computational Efficiency. As shown in Table 2, SmartSight achieves efficiency by optimizing both the Prefill and Decode stages of Video-LLM inference. In the Prefill Stage, which accounts for a large portion of the total inference cost, SmartSight processes only the original video input. This design avoids the extra computation introduced by contrastive decoding methods, which typically forward both clean and perturbed inputs during this stage. In the Decode Stage, where cost grows linearly with the number of generated tokens (Kwon et al. 2023), SmartSight halts low-quality re-

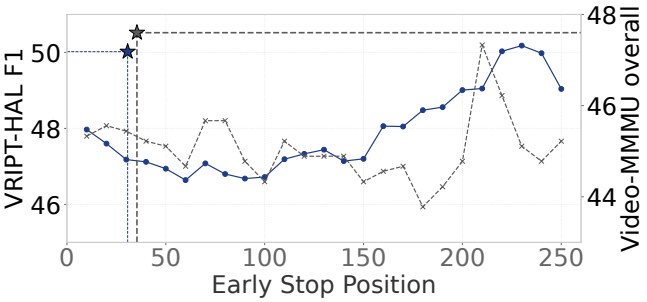

Figure 6: Comparison of early stopping strategies. Blue and gray stars indicate the VAV early-stop points for VRIPT-HAL and Video-MMMU, respectively. The proposed VAV Point yields more accurate stopping decisions, resulting in superior hallucination suppression and reasoning performance compared to fixed-position stopping.

sponses early and continues only the most promising one. This strategy results in early termination at an average of 20.8% and 12% of the output length on Qwen2.5-VL-7B and Video-R1, respectively. Contrastive decoding methods, on the other hand, perform predictions at every decoding step using both clean and perturbed inputs, which results in substantially higher computational costs.

Scaling with Model Size. As shown in Table 2, SmartSight consistently mitigates hallucination and improves understanding across different model sizes. Unlike competing methods, which are only effective on smaller models (e.g., 3B or 7B) and degrade at larger scales, SmartSight remains effective as model size increases, demonstrating favorable scalability.

Scaling with the Number of Sampled Responses N. Figure 5 illustrates the effect of applying SmartSight to Qwen2.5-VL-7B under different numbers of sampled responses N. Increasing N consistently improves understanding on Video-MMMU and mitigates hallucination on VRIPT-HAL, demonstrating the ability of SmartSight to perform test-time scaling, which is difficult to achieve with existing methods. Notably, when $N = 60$, Qwen2.5-VL-7B achieves performance comparable to the Qwen2.5-VL-32B while requiring significantly less computation, consuming 79.2 TFLOPs on VideoMMMU compared to 200.4 TFLOPs. It also performs competitively with the proprietary Gemini 1.5 Pro (Team et al. 2024).

4.4 Ablation Study

Effectiveness of the VAV Point. As shown in Figure 6, fixed stopping positions result in unstable performance on both VRIPT-HAL and Video-MMMU, highlighting the challenge of applying a one-size-fits-all truncation strategy. In contrast, early stopping at the VAV Point yields consistently better results, underscoring its effectiveness in precisely determining the optimal stopping position for each response.

Component-wise Analysis. As shown in Table 4, we conduct a component-wise analysis on Qwen2.5-VL. Compared to the Baseline and Greedy settings, using either the Frame-level or Segment-level Collapse Score independently enables more accurate selection of high-quality responses. Notably, early stopping via the VAV Point reduces decoding cost without significantly affecting performance.

	Greedy	*Baseline*	*+FLC*	*+SLC*	*+TAC*	*+TAC+ES*
VRIPT -HAL	47.9	41.2	48.2 +7.0	48.3 +7.1	50.1 +8.9	50.0 +8.8
Video MME	54.8	51.8	55.2 +3.4	55.7 +3.9	56.3 +4.5	56.22 +4.4

Table 4: Component-wise analysis on Qwen2.5-VL-7B. *Greedy* uses greedy decoding without sampling, while *Baseline* randomly selecting one response from $N = 10$ sampled candidates. *+FLC*, *+SLC* , and *+TAC* select responses based on Frame-level, Segment-level, and TAC Scores, respectively. *+ES* applies early stopping at the VAV Point.

w	1	3	5	7	9	11	13	15	17	19
VRIPT	48.6	49.7	50.0	50.1	50.5	50.1	50.2	50.3	50.1	50.1
VMMMU	40.4	41.0	42.3	42.8	42.1	42.7	42.4	43.1	43.1	42.9
α	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9
VRIPT	47.3	49.2	50.0	49.7	49.2	48.8	48.4	47.8	48.6	48.7
VMMMU	45.9	46.1	47.6	47.3	47.7	46.8	47.9	46.7	47.1	46.8

Table 5: Performance comparison under different w and α settings on VRIPT and VMMMU. VRIPT is short for VRIPT-HAL, and VMMMU is short for Video-MMMU.

Ablations on Hyper-parameters. As shown in Table 5, increasing α from 1.0 to 1.2 and w from 1 to 15 leads to consistent performance gains across both benchmarks, after which the metrics stabilize. These results suggest that SmartSight is robust to the choice of α and w.

5 Conclusion

In this work, we present SmartSight, a training-free method that effectively mitigates perception hallucinations in Video-LLMs without compromising their understanding and reasoning abilities. By introducing the Temporal Attention Collapse score and the Visual Attention Vanishing Point, SmartSight not only suppresses hallucinated outputs but also achieves a favorable trade-off between computational efficiency and accuracy. Extensive experiments across ten diverse Video-LLMs validate its effectiveness, demonstrating notable gains in both hallucination reduction and video understanding performance. For the limitation, we note that our method yields modest improvements on short videos with simple scenes. This is primarily due to the fact that hallucinations are less likely to occur in such videos. Overall, these findings position SmartSight as a promising and generalizable solution for enhancing the reliability of Video-LLMs in more challenging and realistic settings.

Acknowledgements

We are thankful to the shepherd and reviewers for their careful assessment and valuable suggestions, which have

helped us improve this paper. This work was supported in part by the National Natural Science Foundation of China (62472096, 62172104, 62172105, 62102093, 62102091, 62302101, 62202106). Min Yang is a faculty of the Shanghai Institute of Intelligent Electronics & Systems and Engineering Research Center of Cyber Security Auditing and Monitoring, Ministry of Education, China.

References

Bae, K.; Kim, J.; Lee, S.; Lee, S.; Lee, G.; and Choi, J. 2025. MASH-VLM: Mitigating Action-Scene Hallucination in Video-LLMs through Disentangled Spatial-Temporal Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13744–13753.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Chen, C.; Shu, J.; He, G.; Wang, C.; and Li, Y. 2025. Motion-Zero: Zero-Shot Moving Object Control Framework for Diffusion-Based Video Generation. arXiv:2401.10150.

Ding, X.; Zhang, K.; Han, J.; Hong, L.; Xu, H.; and Li, X. 2025. PaMi-VDPO: Mitigating Video Hallucinations by Prompt-Aware Multi-Instance Video Preference Learning. *arXiv preprint arXiv:2504.05810*.

Feng, K.; Gong, K.; Li, B.; Guo, Z.; Wang, Y.; Peng, T.; Wang, B.; and Yue, X. 2025. Video-R1: Reinforcing Video Reasoning in MLLMs. *arXiv preprint arXiv:2503.21776*.

Fu, C.; Dai, Y.; Luo, Y.; Li, L.; Ren, S.; Zhang, R.; Wang, Z.; Zhou, C.; Shen, Y.; Zhang, M.; Chen, P.; Li, Y.; Lin, S.; Zhao, S.; Li, K.; Xu, T.; Zheng, X.; Chen, E.; Shan, C.; He, R.; and Sun, X. 2025. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 24108–24118.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hu, K.; Wu, P.; Pu, F.; Xiao, W.; Zhang, Y.; Yue, X.; Li, B.; and Liu, Z. 2025. Video-MMMU: Evaluating Knowledge Acquisition from Multi-Discipline Professional Videos. arXiv:2501.13826.

Huo, F.; Xu, W.; Zhang, Z.; Wang, H.; Chen, Z.; and Zhao, P. 2025. Self-Introspective Decoding: Alleviating Hallucinations for Large Vision-Language Models. In *The Thirteenth International Conference on Learning Representations*.

Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J. E.; Zhang, H.; and Stoica, I. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 1–18.

Leng, S.; Zhang, H.; Chen, G.; Li, X.; Lu, S.; Miao, C.; and Bing, L. 2024. Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding. In *CVPR*, 13872–13882.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; and Li, C. 2025a. LLaVA-OneVision: Easy Visual Task Transfer. *Transactions on Machine Learning Research*.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Li, C.; Im, E. W.; and Fazli, P. 2025. VidHalluc: Evaluating Temporal Hallucinations in Multimodal Large Language Models for Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13723–13733.

Li, H.; Ding, L.; Fang, M.; and Tao, D. 2024b. Revisiting Catastrophic Forgetting in Large Language Model Tuning. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 4297–4308. Miami, Florida, USA: Association for Computational Linguistics.

Li, X.; Yan, Z.; Meng, D.; Dong, L.; Zeng, X.; He, Y.; Wang, Y.; Qiao, Y.; Wang, Y.; and Wang, L. 2025b. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*.

Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; and Yuan, L. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Ma, F.; Jin, X.; Wang, H.; Xian, Y.; Feng, J.; and Yang, Y. 2023. Vista-llama: Reliable video narrator via equal distance to visual tokens. *arXiv preprint arXiv:2312.08870*.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Rawal, R.; Shirkavand, R.; Huang, H.; Somepalli, G.; and Goldstein, T. 2025. ARGUS: Hallucination and Omission Evaluation in Video-LLMs. *arXiv preprint arXiv:2506.07371*.

Renze, M. 2024. The Effect of Sampling Temperature on Problem Solving in Large Language Models. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 7346–7356. Miami, Florida, USA: Association for Computational Linguistics.

Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; and Liu, Y. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063.

Sun, Y.; Li, Y.; and Wang, C. 2023. Multi-Source Templates Learning for Real-Time Aerial Tracking. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Sun, Y.; Yu, F.; Chen, S.; Zhang, Y.; Huang, J.; Li, Y.; Li, C.; and Wang, C. 2024. ChatTracker: Enhancing Visual Tracking Performance via Chatting with Multimodal Large Language Model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS) 2024*.

Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Wan, G.; Wu, Y.; Chen, J.; and Li, S. 2025. Reasoning Aware Self-Consistency: Leveraging Reasoning Paths for Efficient LLM Sampling. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 3613–3635. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-189-6.

Yang, D.; Huang, S.; Lu, C.; Han, X.; Zhang, H.; Gao, Y.; Hu, Y.; and hai zhao. 2024. Vript: A Video Is Worth Thousands of Words. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

You, L.; Huang, W.; Xie, X.; Wei, X.; Li, B.; Lin, S.; Li, Y.; and Wang, C. 2025. TimeSoccer: An End-to-End Multimodal Large Language Model for Soccer Commentary Generation. *arXiv preprint arXiv:2504.17365*.

Yu, E.; Lin, K.; Zhao, L.; Wei, Y.; Zhu, Z.; Wei, H.; Sun, J.; Ge, Z.; Zhang, X.; Wang, J.; and Tao, W. 2025. Unhackable Temporal Reward for Scalable Video MLLMs. In *The Thirteenth International Conference on Learning Representations*.

Zhang, J.; Jiao, Y.; Chen, S.; Zhao, N.; Tan, Z.; Li, H.; and Chen, J. 2025a. EventHallusion: Diagnosing Event Hallucinations in Video LLMs. arXiv:2409.16597.

Zhang, X.; Wen, S.; Wu, W.; and Huang, L. 2025b. Tinyllava-video-r1: Towards smaller lmms for video reasoning. *arXiv preprint arXiv:2504.09641*.

Zhao, G.; Wang, X.; Zhu, Z.; Chen, X.; Huang, G.; Bao, X.; and Wang, X. 2025. DriveDreamer-2: LLM-Enhanced World Models for Diverse Driving Video Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(10): 10412–10420.