

# TOWARDS MITIGATING HALLUCINATIONS IN LARGE VISION-LANGUAGE MODELS BY REFINING TEXTUAL EMBEDDINGS

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

Hallucinations in Large Vision-Language Models (LVLMs) remain a persistent challenge, often stemming from inadequate integration of visual information during multimodal reasoning. A key cause is the model’s over-reliance on textual priors and underutilization of visual cues, leading to outputs that are linguistically fluent but visually inaccurate. For example, given an image of an empty kitchen countertop, an LVLM might hallucinate a “bowl of fruit” or “cup of coffee,” relying on language associations rather than visual evidence. Most LVLMs incorporate visual features by appending them to the input stream of a pre-trained LLM and training on large-scale vision-language datasets. Our systematic analysis reveals that this strategy often leads to over-dependence on textual information due to the inherent bias of LLMs towards language-dominant representations. This imbalance skews attention towards the text over visual content, weakening the model’s ability to ground outputs in visual inputs. To address this, we propose a simple yet effective visual feature incorporation method that encourages the model to learn visually-informed textual embeddings distinct from those of the base LLM and promotes a more balanced attention distribution. Experimental results across multiple hallucination benchmarks demonstrate that our method significantly reduces hallucinations and fosters more balanced multimodal reasoning. Notably, our approach achieves substantial gains, including **+9.33%** on MMVP-MLLM, **+2.99%** on POPE-AOKVQA, up to **+3.4%** on Merlin, and **+3%** on the hard-data split of HallusionBench.

## 1 INTRODUCTION

The advent of LLMs has transformed tasks like machine translation, dialogue, and content generation with unprecedented accuracy and fluency. Building on this, Large Vision-Language Models (LVLMs) (Lin et al., 2023; Zhang et al., 2023a; Maaz et al., 2024) integrate visual and linguistic understanding in a unified framework, bridging text and visual modalities. This synergy has advanced tasks such as captioning (Chen et al., 2022), question-answering (Li et al., 2023a), multimodal retrieval (Lin et al., 2024), etc. As LVLMs advance, their adoption in domains such as healthcare, autonomous driving, and education is revolutionizing real-world AI application.

Despite this progress, LVLMs remain prone to hallucinations—outputs that are fluent but not grounded in the visual input. These errors, which include fabricating or misinterpreting visual content, undermine reliability and hinder deployment in safety-critical settings. Fig. 1 illustrates failure cases from a LVLM, Video-LLaVA (Lin et al., 2023). In one example, the model captions a scene as “moving it towards a person,” despite the absence of both the person and the action in the video, demonstrating simultaneous object and action hallucination. More broadly, LVLM hallucinations manifest in several forms: attribute hallucinations, where incorrect visual properties are assigned (e.g., describing a blue car as red or denying visible objects); relation hallucinations, which fabricate spatial or contextual relationships (e.g.,

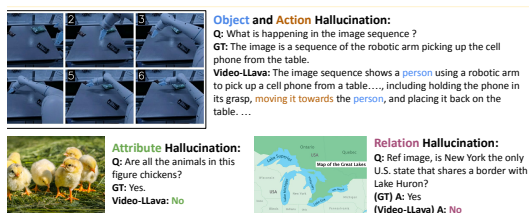


Figure 1: Hallucinations in Video-LLaVA (Lin et al., 2023).

More broadly, LVLM hallucinations manifest in several forms: attribute hallucinations, where incorrect visual properties are assigned (e.g., describing a blue car as red or denying visible objects); relation hallucinations, which fabricate spatial or contextual relationships (e.g.,

054  
055  
056  
057  
058  
059  
060  
061  
062  
063  
064  
065  
066  
067  
068  
069  
070  
071  
072  
073  
074  
075  
076  
077  
078  
079  
080  
081  
082  
083  
084  
085  
086  
087  
088  
089  
090  
091  
092  
093  
094  
095  
096  
097  
098  
099  
100  
101  
102  
103  
104  
105  
106  
107

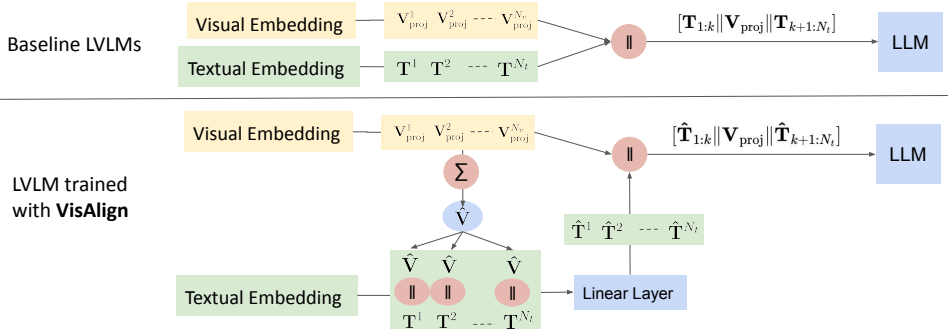


Figure 2: **Top:** Architecture of typical LVLMs like Video-LLaVA, which fuse language and vision embeddings by simple concatenation. **Bottom:** Our modified architecture with a **concatenation block that appends the averaged vision embedding to each token embedding, followed by a projection layer**. This encourages the model to learn visually informed textual embeddings and better attend to visual input during training.

claiming a person is jumping over a fence when they are standing beside it); and, in video settings, temporal hallucinations, where nonexistent dynamics are inferred (e.g., asserting that a person enters the room when no such event occurs).

We hypothesize that a fundamental source of hallucinations in LVLMs arises from the prevailing architectural paradigm in which the visual information is appended as embeddings to the textual embeddings of a pre-trained LLM (Fig. 2, top). This fused input is then passed to the model and fine-tuned on large-scale vision-language datasets, such as image/video captioning, and Visual Question Answer (VQA) (Lin et al., 2023; He et al., 2024; Maaz et al., 2024), etc. While this approach offers modularity, data efficiency, and leverages the strong language generation capabilities of LLMs, it introduces a structural asymmetry: the LLM backbone, trained solely on text, remains inherently biased toward language-driven reasoning (An et al., 2025; Arif et al., 2025). As a result, during fine-tuning, the model may tend to fall back to text priors, under-utilizing the visual embeddings and treating them as secondary in the reasoning process. This modality imbalance may lead to a systematic misalignment between visual evidence and generated text, manifesting during inference as hallucinations: outputs that are linguistically coherent and semantically plausible, yet factually incorrect or unsupported by the visual input.

Motivated by this, we systematically investigate modality imbalance in LVLMs as a potential source of hallucinations, with a focus on the dominant practice of appending visual embeddings to the input textual tokens of pre-trained LLMs (Lin et al., 2023; He et al., 2024; Maaz et al., 2024). We use Video-LLaVA (Lin et al., 2023) as the main model for study this imbalance due to its strong performance and community adoption. We also show results on LLaVA1.5 (Liu et al., 2024a) and Open-Qwen2VL (Wang et al., 2025) to show generalization ability. Our analysis reveals that the prevailing approach of simply appending visual embeddings to the textual input sequence causes the model to over-rely on language while under-utilizing visual information, thereby exacerbating hallucinations. This arises because the backbone LLM, optimized for text, disproportionately emphasizes textual tokens during self-attention operations within the transformer layers.

To address this modality imbalance, we propose VisAlign that integrates visual information into textual embeddings at the token level, thus redefining the input text representations and enabling joint learning of textual and visual information during training. Extensive evaluations across multiple hallucination benchmarks show consistent and statistically significant improvements, demonstrating the effectiveness and generalizability of our approach.

## 2 RELATED WORKS

LVLMs extend pre-trained LLMs to handle visual inputs, typically by appending visual embeddings—extracted from frozen image or video encoders—to the language token sequence. This token-level fusion strategy enables architectural modularity and reusability of LLMs without major modifications. Notable models following this approach include LLaVA (Liu et al., 2024b), MiniGPT-4 (Zhu et al., 2023b), Video-LLaVA (Lin et al., 2023), Video-ChatGPT (Maaz et al., 2023), Bunny (He et al., 2024), Open-Qwen2VL Wang et al. (2025) and Video-LLaMA (Zhang et al., 2023a). Among these, Video-LLaVA has strong benchmark performance, open-source, and

straightforward temporal extension via frame-wise token concatenation (Tang et al., 2025). OpenQwen2VL (Wang et al., 2025) is also a fully open-source multimodal model with SOTA performance which instead of concatenating visual tokens, it fuses it directly into the token embedding space, enabling richer cross-modal interactions and stronger native visual grounding. Models like Flamingo (Alayrac et al., 2022) and BLIP-2 (Li et al., 2023a) use complex cross-attention to integrate modalities dynamically across transformer layers. Although more flexible, they incur higher computational costs and less modularity. Empirical results (Liu et al., 2024b) show simpler token-appending strategies often match or outperform these methods in accuracy and efficiency. For its simplicity, extensibility, and strong performance, we adopt Video-LLaVA (Lin et al., 2023) as the primary model to investigate visual feature integration limitations, focusing on attention distribution, modality alignment, and hallucination. We extend main results to LLaVA1.5 and Open-Qwen2VL to show generalization ability.

**Hallucination Detection and Mitigation in LVLMs** Several approaches have recently been proposed to mitigate hallucinations in LVLMs. M-HalDetect (Gunjal et al., 2023) introduces a dataset of hallucinated captions for training classifiers, while HaELM (Wang et al., 2023b) proposes a fine-tuning framework to distinguish hallucinated from faithful outputs. Reinforcement learning methods such as GAVIE (Liu et al., 2023) penalize ungrounded generations, and ALOHa (Petryk et al., 2023) leverages LLMs to detect hallucinated objects beyond fixed vocabularies. RLHF-based techniques (Sun et al., 2023) further enhance multimodal alignment. CLOCK (Biten et al., 2022) uses attention calibration during training. Inference-time strategies include visual-grounding-enhanced decoding via image descriptions (Ghosh et al., 2024), Instruction Contrastive Decoding (ICD) (Wang et al., 2024b), Self-Introspective Decoding (SID) (Huo et al., 2024), which verifies partial generations, and Visual Contrastive Decoding (VCD) (Leng et al., 2024), which re-ranks outputs to promote visual consistency, ClearSight (Yin et al., 2025) and other attention aligning methods (Zhao et al., 2025; Fazli et al., 2025; Jiang et al., 2025). Together, these methods represent the current state of the art in hallucination mitigation.

Unlike prior approaches that rely during inference-time heuristics, or hallucination-supervised fine-tuning, our method is more principled and addresses hallucination proactively at the input representation during training time.

### 3 BACKGROUND

As noted above, we adopt the widely used and open-source Video-LLaVA as our baseline for major experiments but also show performance on LLaVA 1.5 and Open-Qwen2VL to show generalisability of our approach. Therefore, this section formally outlines the architecture and training pipeline of Video-LLaVA (refer Figure 2 for an overview). It consists of the following components:

**A frozen visual encoder** to extract embeddings from the video (or image), the Video-LLaVA uses the pre-trained LanguageBind (Zhu et al., 2023a).

**A projection layer** that maps the visual embeddings into the textual (base LLM’s) embedding space. The vision-language alignment is carried out via this projection layer. Formally, let  $\mathbf{V} \in \mathbb{R}^{N_v \times d_v}$  denote the visual embeddings, where  $N_v$  is the number of visual tokens and  $d_v$  is the visual embedding dimension. Output from the learnable projection layer  $\mathbf{W}_p \in \mathbb{R}^{d_v \times d_t}$  is denoted as:

$$\mathbf{V}_{\text{proj}} = \mathbf{V}\mathbf{W}_p, \quad \text{where } \mathbf{V}_{\text{proj}} \in \mathbb{R}^{N_v \times d_t} \quad (1)$$

where  $d_t$  is the LLM embedding dimension.

**A backbone LLM:** LVLMs extend upon a pre-trained LLM. Video-LLaVA uses the pre-trained Vicuna-7b (Zheng et al., 2023).

The training consists of two stages:

**Pretraining:** The visual encoder is frozen, and only the projection layer  $\mathbf{W}_p$  is trained to align visual embeddings with the LLM’s input space.

**Finetuning:** The full model, including the LLM, is trained end-to-end to enable effective reasoning over combined visual and textual inputs for visually grounded generation tasks.

### 4 EVALUATING ATTENTION SCORE DISTRIBUTION

Analyzing the attention score distribution across transformer layers provides insight into how information flows from lower to higher layers in LLMs. These scores reveal which tokens most influence the model’s output and offer insight into its learning dynamics (Zhang et al., 2023b). Extending this analysis to LVLMs, we visualize the attention score distributions over both textual and visual tokens to better understand cross-modal interactions.

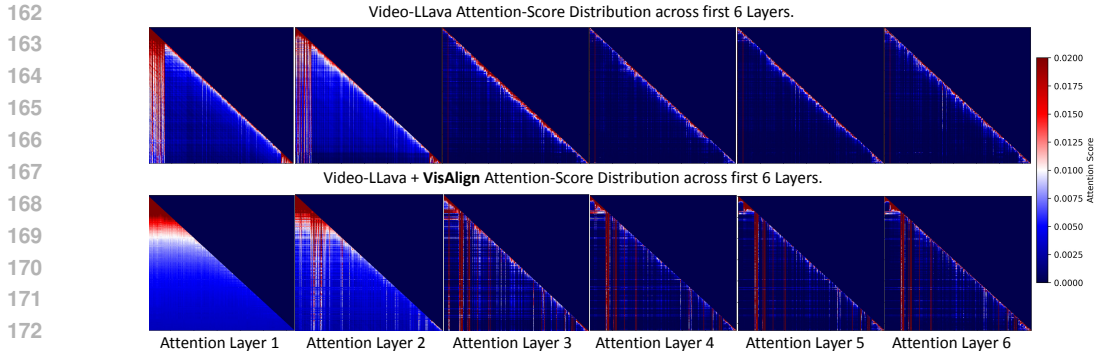


Figure 3: Attention score distributions across the first six attention layers of the baseline Video-LLaVA model (top row) and the VisAlign-enhanced model (bottom row). Video-LLaVA concatenates tokens in a fixed order: 35 initial text tokens, followed by 256 visual embeddings, and then the remaining text tokens. In each map, the x-axis denotes attended tokens (keys), and the y-axis denotes attending tokens (queries). Color intensity reflects attention weight: blue indicates low attention, red/white indicates high attention, and dark (near-black) regions indicate masked or negligible attention due to causal masking in autoregressive LLMs.

Figure 3 shows attention score distributions across multiple transformer layers in Video-LLaVA. In each heatmap, the horizontal axis represents Key tokens (tokens being attended to), and the vertical axis represents Query tokens (tokens performing attention). Color intensity encodes attention strength: cooler tones (e.g., blue) indicate lower scores, while warmer tones (e.g., red) and white indicate stronger attention. Nearly black regions show zero attention due to causal masking. This visualization qualitatively reveals how attention is distributed between visual and textual tokens across the network. Asymmetric or modality-skewed patterns highlight if the model overly favors one modality (typically text) at the expense of the other modality (visual), which can explain hallucination and grounding failures in multimodal tasks.

Figure 3 reveals a pronounced imbalance in how Video-LLaVA distributes attention between textual and visual tokens. In Layer 1 (top row, first plot), attention is heavily concentrated on the initial textual tokens (upper-left red region), sharply declines over the visual tokens, and rises again for the trailing textual tokens—a pattern consistent across layers. As defined in Eq. equation 1, the input sequence  $X$  follows a fixed order: initial textual tokens, followed by visual tokens, and then remaining textual tokens. This results in the model disproportionately focusing on textual tokens at both ends while under-attending to the visual tokens in between.

This asymmetric attention distribution reflects a modality bias rooted in the pre-trained base LLM, which was trained exclusively on text. During fine-tuning, the model relies heavily on linguistic priors and insufficiently leverages the visual embeddings provided by the frozen image or video encoder. This imbalance restricts the effective propagation and integration of visual signals across transformer layers, undermining robust visual grounding. Consequently, the model is prone to generate hallucinations—outputs that are fluent and semantically coherent but factually misaligned or unsupported by the visual input.

## 5 IMPROVING ATTENTION SCORE DISTRIBUTION BY REFINING TEXTUAL EMBEDDINGS

We propose a simple yet principled approach, **VisAlign**, aimed at improving the distribution of attention scores across visual and textual modalities. The underlying hypothesis is that encouraging a more balanced attention pattern—particularly by increasing attention to visual tokens—enables the model to better utilize visual information and reduces hallucinations caused by over-reliance on textual priors. VisAlign operates by refining textual embeddings through the integration of visual context prior to their input to the LLM. This encourages the model to jointly attend to and learn from both textual and visual information during training, leading to more meaningful visual encoding. By fostering a balanced and synergistic interaction between vision and language, VisAlign improves visual utilization without requiring architectural changes or external supervision.

As illustrated in Figure 2, VisAlign first applies average pooling on the projected visual embeddings  $\mathbf{V}_{\text{proj}} \in \mathbb{R}^{N_v \times d_t}$ , resulting in the visual embedding vector  $\hat{\mathbf{V}} \in \mathbb{R}^{1 \times d_t}$ .

$$\hat{\mathbf{V}} = \frac{1}{N_v} \sum_{m=0}^{m=(N_v-1)} \mathbf{V}_{\text{proj}}[m] \quad (2)$$

Next, we fuse  $\hat{\mathbf{V}}$  with the text embeddings  $\mathbf{T}_{1:N_t} \in \mathbb{R}^{N_t \times d_t}$  via concatenation along the  $d_t$  dimension, yielding the fused embeddings  $\mathbf{T}_V$ :

$$\mathbf{T}_V = \left[ \mathbf{T} \parallel \hat{\mathbf{V}} \otimes \mathbf{1}_{N_t} \right] \in \mathbb{R}^{N_t \times 2d_t} \quad (3)$$

Then, we apply a linear projection layer  $\mathbf{W}_d \in \mathbb{R}^{2d_t \times d_t}$  to map the fused representations  $\mathbf{T}_V$  back to the original LLM embedding dimension  $d_t$ , producing the visually-grounded text token sequence,  $\hat{\mathbf{T}} = \mathbf{T}_V \mathbf{W}_d (\in \mathbb{R}^{N_t \times d_t})$ . Unlike the original textual tokens which are derived solely from language embeddings,  $\hat{\mathbf{T}}$  is now a modified version of those language embedding still carrying the textual information. This enforces the model to learn these new embeddings like it tries to learn visual embeddings, thus giving better attention distribution and more effective cross-modal reasoning and visual grounding in downstream tasks. Finally, we append  $\hat{\mathbf{T}}$  to  $\mathbf{V}_{\text{proj}}$  following the original concatenation strategy in Video-LLaVA (Eq. equation 1):

$$\hat{\mathbf{X}} = [\hat{\mathbf{T}}_{1:k} \parallel \mathbf{V}_{\text{proj}} \parallel \hat{\mathbf{T}}_{k+1:N_t}]; \text{ where } \hat{\mathbf{X}} \in \mathbb{R}^{(N_t+N_v) \times d_t} \quad (4)$$

The token sequence  $\hat{\mathbf{X}}$  is then fed into the base LLM. It consists of visually grounded textual embeddings, followed by visual embeddings, and ends with the remaining grounded textual tokens.

**Training Stages:** We use the same datasets and training strategy as used in the baseline Video-LLaVA (Lin et al., 2023) (also discussed in Sec 3). In the *pretraining stage*, we train both the vision-language projection layer and the linear layer, while keeping the LLM frozen (refer to Fig. 2 for an overview). Whereas in the *finetuning stage*, we train the full model end-to-end, including LLM.

### 5.1 ATTENTION SCORE DISTRIBUTION WITH VISALIGN

Figure 3 (bottom row) shows the attention distribution of Video-LLaVA trained with the VisAlign method. As illustrated, attention with VisAlign is more balanced and structured, spanning both visual and textual tokens throughout the sequence. Notably, the vertical attention bands are sharper and more frequent, indicating that the model consistently attends to specific visual regions or tokens that serve as semantic anchors across layers. Additionally, the smoother and more continuous diagonal gradients indicate that tokens attend not only to their local context but also capture long-range dependencies, reflecting a balanced and context-aware attention mechanism. In contrast, the top row (baseline Video-LLaVA) shows less coherent, more fragmented attention patterns. High attention is concentrated at the sequence boundaries, corresponding to textual token positions (Eq. equation 1), revealing a strong bias toward language inputs. The lack of consistent vertical stripes further suggests limited focus on key visual elements, weakening the model’s ability to maintain cross-modal grounding over time. Overall, attention in the baseline appears noisy and scattered across layers, indicating difficulty in forming stable associations between visual content and language queries.

These differences highlight VisAlign’s effectiveness in improving the model’s ability to integrate visual and textual modalities. By promoting more balanced attention, VisAlign improves focus on critical visual cues often overlooked by baseline Video-LLaVA, strengthening temporal and spatial coherence across the transformer layers and boosting overall visual information use.

**Why VisAlign encourages the model to use visual information?** VisAlign introduces no additional visual inputs or training objectives; its contribution is purely representational. By augmenting the LLM’s textual token embeddings with averaged visual embeddings, VisAlign alters the model’s input representation and reshapes the optimization landscape, reducing the reliance on memorized textual priors. As discussed earlier, LVLMs tend to over-attend to text because the underlying LLM is pre-trained exclusively on textual embeddings. During multi-modal training, language tokens remain in-distribution for the backbone LLM, whereas visual tokens are comparatively out-of-distribution, leading to the strong attention bias toward text as observed in Fig. 3 (top). By injecting visual information into every textual embedding, VisAlign deliberately departs from the pre-training distribution, encouraging the model to adapt its early-layer attention patterns during finetuning and to more consistently incorporate visual evidence.

270  
271  
272  
273  
274  
275  
276  
277  
278  
279  
280  
281  
282  
283  
284  
285  
286  
287  
288  
289  
290  
291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302  
303  
304  
305  
306  
307  
308  
309  
310  
311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321  
322  
323

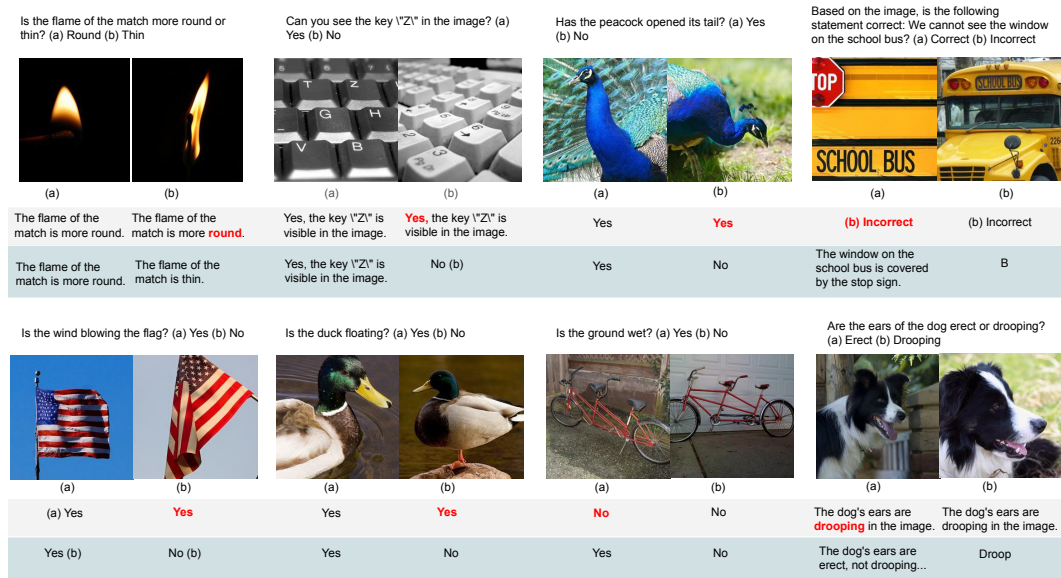


Figure 4: Qualitative results from the **MMVP-MLLM Benchmark**: Below each image, the baseline model’s response is shown first, followed by the response from the model trained with *VisAlign*.

## 6 EXPERIMENTS AND RESULTS

We evaluate *VisAlign* across a diverse set of benchmarks that probe hallucinations and visual grounding from complementary perspectives, including fine-grained visual discrimination, object-level hallucinations, factual consistency under visual edits, sequential visual reasoning, and conflicts between visual input and parametric memory. The results and discussions for each benchmark are presented below (a detailed description of these benchmarks is provided in Appendix Sec.A.2):

**MMVP-MLLM** The results in Table 1 show that Video-LLaVA enhanced with *VisAlign* achieves a substantial **+9.33%** improvement over the baseline. Since MMVP-MLLM is specifically designed to probe bias and hallucination in LVLMs by enforcing fine-grained visual discrimination under minimal semantic variance, this gain is especially significant. It demonstrates that *VisAlign* markedly strengthens the model’s grounding in visual evidence rather than relying on linguistic priors, effectively reducing hallucinations and improving factual consistency. A qualitative comparison is presented in Figure 4. In the first example, the model must distinguish between two flame images—one round and the other elongated. The baseline Video-LLaVA incorrectly classifies both as “round,” indicating over-reliance on memorized language patterns. In contrast, the *VisAlign*-enhanced model correctly differentiates the shapes, demonstrating stronger visual grounding. Similar improvements appear in other examples, underscoring *VisAlign*’s effectiveness in reducing hallucinations and promoting accurate, cross-modal reasoning.

**POPE** Following prior work (Villa et al., 2025), we focus on the most challenging setting: Adversarial SEEM from A-OKVQA, which applies SEEM-based object detection to A-OKVQA images. This subset probes whether models falsely affirm the presence of common yet incorrect objects, revealing object-level hallucinations driven by language bias. Table 1 presents quantitative results on the POPE benchmark, where *VisAlign* consistently surpasses the baseline across key metrics, achieving a **2.99%** increase in accuracy, a **1.76%** boost in precision, and a **1.18%** gain in F1-score. The notable rise in precision indicates a significant reduction in false positives—hallucinated objects—while the improved F1-score reflects a more robust balance between precision and recall. These provide strong evidence that *VisAlign* effectively curtails predictions of frequent yet visually unsupported objects, thereby substantially enhancing object-level visual grounding. Supporting qualitative results in Fig. 5 further reinforce *VisAlign*’s reliability in avoiding erroneous affirmations of absent objects, underscoring its role in advancing cross-modal integration and reducing hallucinations.

	MMVP-MLLM	POPE A-OKVQA			
	Acc	Acc	P	R	F1
Video-LLaVA	14.0	54.1	52.14	99.6	68.45
+ <i>VisAlign</i>	<b>23.33</b>	<b>57.09</b>	<b>53.9</b>	98.33	<b>69.63</b>

Table 1: Results on POPE A-OKVQA (Li et al., 2023b) and MMVP-MLLM (Tong et al., 2024). Acc: accuracy; P: precision; R: recall.

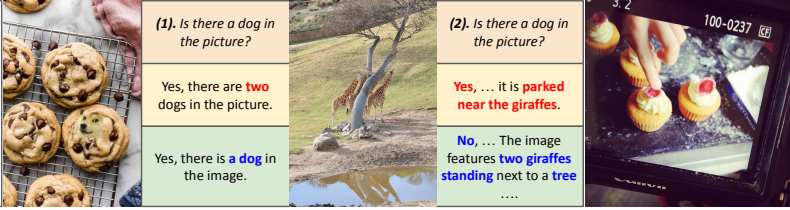
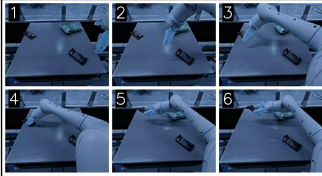
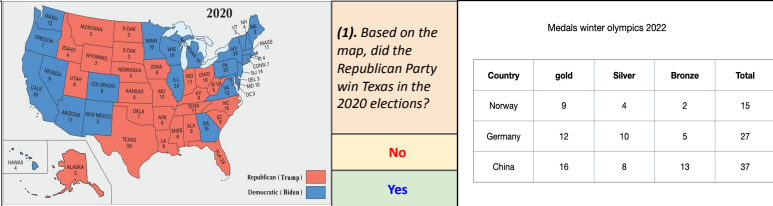
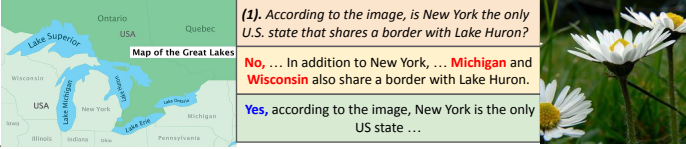
324	(a)		(1). Is there a dog in the picture?	(2). Is there a dog in the picture?	(3). Is there a tree in the image?	
325			Yes, there are <b>two</b> dogs in the picture.	Yes, ... it is <b>parked near the giraffes</b> .	Yes, there is a tree in the image.	
326			Yes, there is <b>a dog</b> in the image.	No, ... The image features <b>two giraffes standing next to a tree</b> ....	No, ... a person holding a <b>camera</b> and taking a picture of <b>cupcakes</b> .	
327			Object Hallucination			
328						
329	Samples taken from (1) HallusionBench (2) POPE (3) POPE					
332	(b)		(1) Write a description for the given image sequence in a single paragraph, what is happening in this episode?			
333			The image sequence shows a person using a robotic arm to pick up a cell phone from a table...., including <b>holding the phone in its grasp, moving it towards the person, and placing it back on the table</b> ....			
334			In the image, there is a robotic arm with a gripper end, which is <b>picking up</b> a cell phone from a table. ... The image is a <b>sequence of the robotic arm picking up the cell phone</b> from the table.			
335			Action Hallucination			
336						
337	Sample taken from (1) Mementos					
340	(c)		(1). Based on the map, did the Republican Party win Texas in the 2020 elections?	(2). According to the table, is the number of gold medals of Norway 9 in winter Olympics in 2022?		
341			No	No		
342			Yes	Yes		
343			Attribute Hallucination			
344						
345	Samples taken from (1) HallusionBench (2) HallusionBench					
348	(d)		(1). According to the image, is New York the only U.S. state that shares a border with Lake Huron?	(2). In the image, there is one daisy that is under the shadow of a taller daisy. (a) Correct (b) Incorrect		
349			No, ... In addition to New York, ... <b>Michigan and Wisconsin</b> also share a border with Lake Huron.	(a) Correct		
350			Yes, according to the image, New York is the only US state ...	(b)		
351			Relation Hallucination			
352						
353	Samples taken from (1) HallusionBench (2) MMVP					

Figure 5: Qualitative examples from **POPE A-OKVQA**, **HallusionBench**, **MMVP**, and **Mementos** benchmarks illustrating various hallucination types. Input prompts are shown in orange, baseline Video-LLaVA outputs in yellow, and VisAlign-enhanced outputs in green. VisAlign consistently improves performance across object, action, attribute, and relation hallucinations.

**MERLIN** evaluates factual consistency and visual grounding in LVLMs through fine-grained object existence verification. Table 2 presents quantitative results for both positive (object present) and negative (object removed) cases, evaluated under two distinct image sampling strategies. VisAlign consistently outperforms the baseline demonstrating superior capability to mitigate hallucinations by enhancing the model’s sensitivity to subtle visual cues, thereby substantially improving visual fidelity and robustness in fine-grained, object-centric reasoning tasks.

**Mementos** evaluates sequential image reasoning in LVLMs across three domains: *Robotics*, *Comics*, and *Daily Life*. It rigorously test object and action hallucinations within dynamic visual contexts, emphasizing temporal coherence and object-behavior relationships. This makes Mementos especially valuable for assessing a multimodal model’s ability to detect hallucinations while accurately understanding complex, evolving visual narratives.

Table 3 shows significant improvements in the *Robotics* domain for both object hallucination (+1.13% accuracy) and action hallucination (+0.97% accuracy). These gains stem from the structured, goal-driven nature of robotic sequences, where predictable temporal patterns and clear visual cues enable VisAlign to maintain coherent attention over time and better align visual tokens with text, enhancing temporal reasoning of object states and behaviors. In contrast, improvements in the

	Curated Images			
	Pos-Orig	Pos-Edited	Neg-Orig	Neg-Edited
Video-LLava	30.9	16.7	71.5	79.6
VisAlign	<b>34.3</b>	<b>20.3</b>	<b>72.7</b>	<b>83.0</b>
	Random Images			
	Pos-Orig	Pos-Edited	Neg-Orig	Neg-Edited
Video-LLava	48.2	33.3	59.5	67.9
VisAlign	<b>48.6</b>	<b>36.7</b>	<b>60.1</b>	<b>71.3</b>

Table 2: Results (%) on the Merlin benchmark Villa et al. (2023). “Pos”: positive, “Neg”: negative.

*Comics* and *Daily Life* domains are more modest, likely due to their greater visual and semantic complexity. Comics often use stylized, symbolic imagery and abstract narratives that disrupt typical visual-linguistic links, while Daily Life scenes involve high variability, subtle object transitions, and complex human actions that hinder consistent temporal alignment. In these unstructured contexts, VisAlign’s attention calibration is limited by noisier, less reliable visual inputs.

**HallusionBench** On this VisAlign achieves an average improvement of approximately 3%, with consistent gains across both VD and VS categories (detailed results in Appendix A.3). The improvements are particularly notable in VS tasks such as Map, OCR, and Table, where reliance on memorized knowledge is most error-prone, indicating that VisAlign effectively encourages visual grounding over language priors. Qualitative examples (Figure 5) further show that VisAlign enables correct interpretation of manipulated visual inputs—such as falsified maps and statistics—where the baseline model hallucinates. Overall, these results demonstrate that VisAlign consistently improves visual grounding and reduces hallucinations under adversarial visual–textual discrepancies.

**Performance on Additional LVLMS:** To further validate the generality and robustness of VisAlign, we evaluate its effectiveness on two SOTA LVLMS, LLaVA 1.5 and Open-Qwen2VL. As shown in Table 4, VisAlign consistently enhances performance and reduces hallucinations highlighting its broad applicability and effectiveness across different LVLMS (further details: refer Sec. A.6).

### Comparison with Existing Hallucination Mitigation Approaches:

This section compares VisAlign with other SOTA hallucination mitigation methods. We focus on Visual Contrastive Decoding (VCD) (Leng et al., 2024), a strong inference-time SOTA method. While model-agnostic and lightweight, such inference-time methods complement VisAlign, which proactively mitigates hallucinations by refining representations during training (refer Sec. A.5 in Appendix for further details on VCD). Table 5 compares the effectiveness of VisAlign and VCD applied to Video-LLaVA, both individually and combined. VisAlign outperforms VCD alone, notably improving accuracy (54.5 → 57.09) and F1-score (68.6 → 69.63). While VCD delivers incremental gains by refining output selection during inference, VisAlign achieves more substantial improvements by addressing modality imbalance during training. When combined, the two methods yield the best overall performance, further boosting accuracy to 58.8 and F1-score to 70.04, demonstrating their complementary strengths.

## 7 CONCLUSION

We systematically analyze attention distributions in LVLMS with respect to hallucinations—outputs that lack grounding in visual input—and find that popular models overemphasize text, leading to increased reliance on linguistic priors. To address this, we propose a simple yet effective method that redefines textual embeddings to rebalance attention during training and improve the use of visual information. This results in significantly reduced hallucinations and more semantically accurate, visually grounded outputs. We validate our approach across multiple challenging hallucination benchmarks, consistently achieving substantial improvements. We hope these findings inspire further research toward better leveraging visual data and enhancing the reliability of multimodal reasoning in LVLMS.

Method	Object				Action			
	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
<i>Robotics Domain</i>								
Video-LLaVA	8.27	16.55	12.40	13.46	5.53	6.99	<b>11.30</b>	8.40
+VisAlign	<b>9.40</b>	<b>19.20</b>	<b>13.61</b>	<b>15.16</b>	<b>6.50</b>	<b>9.60</b>	10.76	<b>9.45</b>
<i>Daily Life Domain</i>								
Video-LLaVA	22.05	38.30	31.90	33.55	13.50	31.70	<b>18.66</b>	<b>22.40</b>
+VisAlign	<b>22.18</b>	<b>38.31</b>	<b>32.31</b>	<b>33.70</b>	12.31	<b>32.10</b>	16.44	20.70
<i>Comics Domain</i>								
Video-LLaVA	11.12	21.00	<b>19.00</b>	<b>18.86</b>	<b>4.48</b>	11.28	<b>6.58</b>	<b>8.08</b>
+VisAlign	<b>12.00</b>	21.00	17.80	18.41	4.00	<b>13.33</b>	5.36	7.10

Table 3: Results on Mementos (Wang et al., 2024a) across object and action hallucinations in three domains.

Model	Acc	Prec.	Recall	F1
LLaVA1.5 (%)	69.0	62.23	97.66	76.02
+ VisAlign (%)	<b>71.0</b>	<b>64.0</b>	97.13	<b>77.01</b>
OpenQwen2VL (%)	53.13	80.12	8.33	15.09
+ VisAlign (%)	<b>55.7</b>	56.8	<b>47.6</b>	<b>51.8</b>

Table 4: Effect of VisAlign on LLaVA1.5 and OpenQwen2VL on POPE-AOKVQA.

Model	Acc	Prec.	Recall	F1
Video-LLaVA	54.1	52.14	<b>99.6</b>	68.45
+ VCD (Leng et al., 2024)	54.5	52.38	99.39	68.60
+ VisAlign	57.09	53.90	98.33	69.63
+ VisAlign + VCD	<b>58.8</b>	<b>55.03</b>	96.33	<b>70.04</b>

Table 5: Performance of Video-LLaVA with hallucination mitigation methods on POPE-AOKVQA.

## REFERENCES

- 432  
433  
434 Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel  
435 Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language  
436 model for few-shot learning. *Advances in neural information processing systems*, 35:23716–  
437 23736, 2022.
- 438 Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xi-  
439 aoqin Zhang, and Shijian Lu. Mitigating object hallucinations in large vision-language models  
440 with assembly of global and local attention. In *Proceedings of the Computer Vision and Pattern  
441 Recognition Conference*, pp. 29915–29926, 2025.
- 442 Kazi Hasan Ibn Arif, Sajib Acharjee Dip, Khizar Hussain, Lang Zhang, and Chris Thomas. Paint:  
443 Paying attention to informed tokens to mitigate hallucination in large vision-language model.  
444 *arXiv preprint arXiv:2501.12206*, 2025.
- 445 Ali Furkan Biten, Lluís Gómez, Marçal Rusiñol, and Dimosthenis Karatzas. Let there be a clock on  
446 the beach: Reducing object hallucination in image captioning. In *Proceedings of the IEEE/CVF  
447 Winter Conference on Applications of Computer Vision*, pp. 2434–2443, 2022.
- 448 Fu Chaoyou, Chen Peixian, Shen Yunhang, Qin Yulei, Zhang Mengdan, Lin Xu, Yang Jinrui, Zheng  
449 Xiawu, Li Ke, Sun Xing, et al. Mme: A comprehensive evaluation benchmark for multimodal  
450 large language models. *arXiv preprint arXiv:2306.13394*, 3, 2023.
- 451 Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian  
452 Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual  
453 language-image model. *arXiv preprint arXiv:2209.06794*, 2022.
- 454 Mehrdad Fazli, Bowen Wei, Ahmet Sari, and Ziwei Zhu. Mitigating hallucination in large vision-  
455 language models via adaptive attention calibration. *arXiv preprint arXiv:2505.21472*, 2025.
- 456 Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Utkarsh Tyagi, Oriol Nieto, Zeyu Jin,  
457 and Dinesh Manocha. Visual description grounding reduces hallucinations and boosts reasoning  
458 in lvlms. *arXiv preprint arXiv:2405.15683*, 2024.
- 459 Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang  
460 Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for  
461 entangled language hallucination and visual illusion in large vision-language models. In *Pro-  
462 ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–  
463 14385, 2024.
- 464 Prasanna Gunjal et al. M-haldetect: Detecting hallucinations in vision-language models. In *Pro-  
465 ceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.
- 466 Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. Efficient  
467 multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530*, 2024.
- 468 Jing Huo et al. Mitigating object hallucinations in large vision-language models via attention cali-  
469 bration. *arXiv preprint arXiv:2502.01969*, 2024.
- 470 Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle  
471 layers of large vision-language models: Interpreting, detecting and mitigating object hallucina-  
472 tions via attention lens. In *Proceedings of the Computer Vision and Pattern Recognition Confer-  
473 ence*, pp. 25004–25014, 2025.
- 474 Long Jing, Zhe Wang, Yichen Zhang, Dacheng Tao, and Mingli Song. Faith: Faithful and  
475 informative textual hallucination detection in image captioning. In *Proceedings of the 2023  
476 Conference on Computer Vision and Pattern Recognition*, pp. 3456–3465. IEEE, 2023. doi:  
477 10.1109/CVPR.2023.00345. URL [https://openaccess.thecvf.com/content/  
478 CVPR2023/html/Jing\\_FAITH\\_Faithful\\_and\\_Informative\\_Textual\\_  
479 Hallucination\\_Detection\\_in\\_Image\\_Captioning\\_CVPR\\_2023\\_paper.html](https://openaccess.thecvf.com/content/CVPR2023/html/Jing_FAITH_Faithful_and_Informative_Textual_Hallucination_Detection_in_Image_Captioning_CVPR_2023_paper.html).
- 480  
481  
482  
483  
484  
485

- 486 Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing.  
487 Mitigating object hallucinations in large vision-language models through visual contrastive de-  
488 coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni-  
489 tion*, pp. 13872–13882, 2024.
- 490  
491 Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image  
492 pre-training with frozen image encoders and large language models. In *International conference  
493 on machine learning*, pp. 19730–19742. PMLR, 2023a.
- 494 Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating  
495 object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023b.
- 496  
497 Bin Lin, Yang Ye, Bin Zhu, Jiayi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning  
498 united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*,  
499 2023.
- 500 Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei  
501 Ping. Mm-embed: Universal multimodal retrieval with multimodal llms, 2024. URL <https://arxiv.org/abs/2411.02571>.
- 502  
503  
504 Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction  
505 tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*,  
506 pp. 26296–26306, 2024a.
- 507 Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction  
508 tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition  
509 (CVPR)*, pp. 26296–26306, June 2024b.
- 510  
511 Yuxiang Liu et al. Gavie: Grounded and verifiable image explanation. In *EMNLP Findings*, 2023.
- 512 Holy Lovenia, Adji Bintang Wibowo, Krisna Kuntoro, Muhammad Firdaus, Radityo Eko Pra-  
513 sojo, Derry Tanti Suhendro, and Kurniawan Kurniawan. Nope: Evaluating and explaining  
514 negative object presence in image captioning. In *Proceedings of the IEEE/CVF International  
515 Conference on Computer Vision*, pp. 1234–1243, 2023. doi: 10.1109/ICCV.2023.00123. URL  
516 [https://openaccess.thecvf.com/content/ICCV2023/html/Lovenia\\_  
517 NOPE\\_Evaluating\\_and\\_Explaining\\_Negative\\_Object\\_Presence\\_in\\_  
518 Image\\_Captioning\\_ICCV\\_2023\\_paper.html](https://openaccess.thecvf.com/content/ICCV2023/html/Lovenia_NOPE_Evaluating_and_Explaining_Negative_Object_Presence_in_Image_Captioning_ICCV_2023_paper.html).
- 519 Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt:  
520 Towards detailed video understanding via large vision and language models. *arXiv preprint  
521 arXiv:2306.05424*, 2023.
- 522  
523 Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt:  
524 Towards detailed video understanding via large vision and language models. In *Proceedings of  
525 the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024.
- 526  
527 Nathan Petryk, Shikhar Sharma, Ali Furkan Biten, Lluís Gomez, Dimosthenis Karatzas, C V Jawa-  
528 har, and Minesh Mathew. Aloha: Assessing language-only hallucinations in image captioning.  
529 In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Process-  
530 ing*, pp. 6789–6798. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.  
531 emnlp-main.567. URL <https://aclanthology.org/2023.emnlp-main.567/>.
- 532  
533 Hao Sun et al. Mmhal-bench: Multimodal hallucination benchmark for vision-language dialogue.  
*arXiv preprint arXiv:2312.00704*, 2023.
- 534  
535 Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An,  
536 Jingyang Lin, Rongyi Zhu, et al. Video understanding with large language models: A survey.  
537 *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.
- 538 Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide  
539 shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF  
Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024.

- 540 Andrés Villa, Juan Carlos León Alcázar, Alvaro Soto, and Bernard Ghanem. Behind the magic,  
541 merlim: Multi-modal evaluation benchmark for large image-language models. *arXiv preprint*  
542 *arXiv:2312.02219*, 2023.
- 543 Andrés Villa, Juan León Alcázar, Motasem Alfarra, Vladimir Araujo, Alvaro Soto, and Bernard  
544 Ghanem. Eagle: Enhanced visual grounding minimizes hallucinations in instructional multimodal  
545 models. *arXiv preprint arXiv:2501.02699*, 2025.
- 547 Weizhi Wang, Yu Tian, Linjie Yang, Heng Wang, and Xifeng Yan. Open-qwen2vl: Compute-  
548 efficient pre-training of fully-open multimodal llms on academic resources. *arXiv preprint*  
549 *arXiv:2504.00595*, 2025.
- 550 Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon,  
551 Taixi Lu, Gedas Bertasius, Mohit Bansal, et al. Mementos: A comprehensive benchmark for mul-  
552 timodal large language model reasoning over image sequences. *arXiv preprint arXiv:2401.10529*,  
553 2024a.
- 554 Yixin Wang, Yuxiang Liu, Chunyuan Chen, Zhe Wang, Shuchang Yan, et al. Mitigating hallucina-  
555 tions in large vision-language models with instruction contrastive decoding. In *Proceedings of*  
556 *the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.
- 558 Zhe Wang et al. Amber: A benchmark for evaluating hallucinations in multimodal models. *arXiv*  
559 *preprint arXiv:2310.12114*, 2023a.
- 560 Zhe Wang et al. Haelm: Hallucination evaluation for large multimodal models. *arXiv preprint*  
561 *arXiv:2305.19162*, 2023b.
- 563 Hao Yin, Guangzong Si, and Zilei Wang. ClearSight: Visual signal enhancement for object halluci-  
564 nation mitigation in multimodal large language models. In *Proceedings of the Computer Vision*  
565 *and Pattern Recognition Conference*, pp. 14625–14634, 2025.
- 566 Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language  
567 model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023a.
- 569 Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo  
570 Zhao. Tell your model where to attend: Post-hoc attention steering for llms. *arXiv preprint*  
571 *arXiv:2311.02262*, 2023b.
- 572 Jianfei Zhao, Feng Zhang, Xin Sun, and Chong Feng. Mitigating hallucination in large vision-  
573 language models through aligning attention distribution to information flow. In *Findings of the*  
574 *Association for Computational Linguistics: EMNLP 2025*, pp. 24849–24863, 2025.
- 576 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang,  
577 Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and  
578 chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.
- 579 Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiayi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang,  
580 Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-  
581 modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023a.
- 582 Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigt-4: En-  
583 hancing vision-language understanding with advanced large language models. *arXiv preprint*  
584 *arXiv:2304.10592*, 2023b.
- 585  
586  
587  
588  
589  
590  
591  
592  
593

## A APPENDIX

### A.1 RELATED WORKS

**Hallucination Detection and Mitigation in LVLMs** is actively studied through a range of benchmarks designed to evaluate diverse hallucination types. POPE-AOKVQA (Li et al., 2023b) and NOPE (Lovenia et al., 2023) focus on object-level hallucinations, while MERLIN (Jing et al., 2023) examines factual consistency via atomic fact decomposition. MMVP-MLLM (Tong et al., 2024) and HallusionBench (Guan et al., 2024) probe model behavior under minimal semantic variation and cross-modal conflicts. Mementos (Wang et al., 2024a) targets temporal hallucinations in sequential visual reasoning. AMBER (Wang et al., 2023a) introduces a unified benchmark for evaluating both discriminative and generative hallucinations. Together, these datasets reveal a broad spectrum of hallucination phenomena—including object, action, attribute, relational, and temporal inconsistencies—highlighting the complexity of achieving reliable visual grounding in LVLMs.

### A.2 DETAILED DESCRIPTION OF VARIOUS BENCHMARKS USED IN EVALUATING OUR METHOD

**MMVP-MLLM** (Tong et al., 2024) benchmark features carefully curated image pairs with highly similar CLIP embeddings, minimizing semantic divergence and emphasizing subtle visual distinctions. Each pair is accompanied by two binary-choice questions targeting fine-grained visual understanding. A model receives credit only if it answers both correctly, enforcing a strict criterion that rewards accurate visual grounding and penalizes reliance on language priors. This makes MMVP-MLLM particularly effective for evaluating hallucinations, as it compels models to rely on actual visual evidence rather than linguistic shortcuts or memorized associations.

**POPE** (Li et al., 2023b) evaluates hallucinations through yes/no questions about object presence in images. “Yes” questions correspond to ground-truth objects, while “No” questions are adversarially crafted from the top-k most frequent object categories absent from the image. This setup exposes the model’s reliance on language priors by testing its ability to reject visually unsupported but common objects. Following prior work (Villa et al., 2025), we focus on the most challenging setting: Adversarial SEEM from A-OKVQA, which applies SEEM-based object detection to A-OKVQA images. This subset probes whether models falsely affirm the presence of common yet incorrect objects, revealing object-level hallucinations driven by language bias. POPE thus offers a fine-grained, targeted measure of visual grounding, serving as a rigorous and complementary benchmark to evaluate VisAlign’s effectiveness in reducing hallucinations.

**MERLIN** Villa et al. (2023) evaluates factual consistency and visual grounding in LVLMs through fine-grained object existence verification. It employs a curated set of original and synthetically edited images to assess whether models can accurately detect the presence or absence of objects. Our evaluation specifically targets a subset of MERLIN where an entire object category, limited to a single instance in the original image, has been removed in the edited version.

**Mementos** (Wang et al., 2024a) evaluates sequential image reasoning in LVLMs across three domains: *Robotics*, *Comics*, and *Daily Life*. It rigorously tests object and action hallucinations within dynamic visual contexts, emphasizing temporal coherence and object-behavior relationships. This makes Mementos especially valuable for assessing a multimodal model’s ability to detect hallucinations while accurately understanding complex, evolving visual narratives.

**HallusionBench** (Guan et al., 2024) is a diagnostic benchmark assessing how parametric memory affects hallucinations in LVLMs. It categorizes questions into Visual-Dependent (VD), requiring visual input, and Visual-Supplement (VS), answerable using world knowledge or training data. VS questions evaluate the model’s ability to resolve conflicts between visual input and parametric memory. The benchmark includes easy and hard splits, with the hard subset featuring human-edited images designed to create modality conflicts.

### A.3 RESULTS ON HALLUSIONBENCH BENCHMARK

**HallusionBench** (Guan et al., 2024) is a diagnostic benchmark assessing how parametric memory affects hallucinations in LVLMs. It categorizes questions into Visual-Dependent (VD), requiring

Method	Visual Dependent				Visual Supplement					
	Figure	Ilusion	Math	OCR	Video	Chart	Map	OCR	Table	Average
	Hard Data Split									
Video-LLaVA	29.27	54.93	35.29	41.30	36.84	24.56	25.00	18.52	28.79	32.72
Video-LLaVA + VisAlign	<b>34.15</b>	49.30	<b>37.25</b>	<b>45.65</b>	<b>36.84</b>	21.05	<b>28.12</b>	<b>33.33</b>	<b>34.85</b>	<b>35.61</b>
	Easy Data Split									
Video-LLava	64.10	40.28	27.78	75.61	15.94	35.11	46.88	53.70	36.36	<b>43.97</b>
Video-LLava + VisAlign	53.85	36.11	<b>37.04</b>	53.66	<b>36.23</b>	25.95	48.44	50.00	28.57	41.1

Table 6: Category-wise results on the HallusionBench benchmark(Guan et al., 2024).

visual input, and Visual-Supplement (VS), answerable using world knowledge or training data. VS questions evaluate the model’s ability to resolve conflicts between visual input and parametric memory. The benchmark includes easy and hard splits, with the hard subset featuring human-edited images designed to create modality conflicts.

Table 6 shows an average improvement of about **3%** on the challenging hard subset. Significant gains are seen in Visual-Dependent (VD) tasks, with improvements of **4.88%**, **1.96%**, and **4.35%** in the “Figure,” “Math,” and “OCR” categories, respectively. Even larger gains occur in Visual-Supplement (VS) tasks, with **3.12%**, **14.81%**, and **6.06%** improvements in “Map,” “OCR,” and “Table.” These results are particularly notable because the hard subset contains human-edited images designed to conflict with common knowledge, forcing the model to rely on visual input rather than memorized facts. The gains indicate that VisAlign substantially improves the model’s ability to ground predictions in visual evidence, reducing over-reliance on language priors. For example, Figure 5 (d)(1) shows a manipulated map where New York is falsely depicted bordering Lake Huron; while baseline Video-LLaVA hallucinates based on memorized geography, Video-LLaVA+VisAlign correctly interprets the altered visual context. Similarly, in (c)(2), a falsified medal count for Norway is accurately detected only by the VisAlign-enhanced model. These examples highlight VisAlign’s effectiveness in enhancing visual grounding and mitigating hallucinations by improving sensitivity to subtle visual inconsistencies.

#### A.4 EFFECT OF VISALIGN ON MME, A GENERAL LVLM BENCHMARK:

In the main paper, we comprehensively evaluated VisAlign’s effectiveness in reducing hallucinations across multiple benchmarks, consistently demonstrating significant and robust improvements. Although our primary focus is on hallucination tasks, to further investigate VisAlign’s broader impact, we also assess how it influences the baseline model’s performance on generic vision-language understanding benchmarks.

To this end, we evaluate on the MME benchmark (Chaoyou et al., 2023), a widely adopted diagnostic suite designed to probe the general capabilities of LVLMs. MME includes various subcategories covering fine-grained visual understanding and textual grounding tasks, such as Existence, Count, Position, Color, Posters, Celebrity, Scene, Landmark, Artwork, and OCR. These categories span a range of difficulty, from low-level visual perception to high-level semantic reasoning, offering a comprehensive lens into overall model competency.

Table 7 reports category-wise performance comparing the baseline Video-LLaVA, VisAlign and VCD augmented versions. VisAlign significantly improves upon the baseline in several key subcategories that are sensitive to visual grounding, such as Existence (**170** → **190**), Count (**121.66** → **131.66**), and Color (**135** → **148.33**). These improvements align with the primary objective of VisAlign—mitigating hallucinations by enhancing the model’s attention to visual evidence—demonstrating its positive influence on tasks that demand precise object recognition and attribute understanding. Moreover, in categories such as OCR and Posters, VisAlign preserves the same level of performance as the baseline, indicating that it does not compromise tasks unrelated to hallucination-prone scenarios. However, some categories—such as Position, Celebrity, Scene, Landmark, and Artwork—show drop in performance. These tasks often require fine-grained spatial reasoning or prior world knowledge, which may be subtly impacted by VisAlign’s architectural shift toward reinforcing visual embeddings over memorized linguistic patterns. This suggests that while VisAlign strengthens core visual grounding, it may introduce minor trade-offs in more specialized or context-dependent tasks.

Another observation from Table 7 is that state-of-the-art hallucination mitigation methods like **VCD cause a universal performance drop or yield no improvements across all MME subcategories.**

In contrast, VisAlign demonstrates a more favorable trade-off: while it introduces minor performance reductions in certain high-level categories, it provides targeted improvements in core grounding tasks without degrading overall reliability. This contrast highlights VisAlign’s orthogonality to inference-time methods and its potential to improve multimodal reasoning in a more integrated and generalizable manner.

In summary, while VisAlign is primarily designed to mitigate hallucinations, it also brings positive side effects on general VLM tasks that benefit from stronger visual grounding. By enriching textual embeddings with visual information, VisAlign promotes faithful grounding in visual inputs and reduces over-reliance on language priors. Unlike inference-time methods like VCD—which often reduce performance on generic benchmarks—VisAlign improves internal representations, preserving or enhancing accuracy in key subcategories like Color, Count, and Existence. However, this stronger grounding can slightly reduce performance in tasks relying on memorized knowledge or abstract reasoning (e.g., Landmark or Celebrity), due to reduced influence from language-driven biases. This trade-off is expected and could potentially be mitigated by training on larger-scale multimodal datasets—an exciting direction for future work. Overall, VisAlign offers a principled, generalizable, and training-efficient approach to hallucination reduction while preserving broader multimodal capabilities.

#### A.5 COMPARISON WITH EXISTING HALLUCINATION MITIGATION APPROACHES

In the main paper, we showed that VisAlign significantly reduces hallucinations in Video-LLaVA by improving the attention score distribution across visual and textual modalities. In this section, we extend our analysis by comparing VisAlign with other state-of-the-art (SOTA) hallucination mitigation methods. As noted in the Related Work section (2), inference-time strategies currently represent the leading approaches for mitigating hallucinations. These methods intervene during the decoding stage to guide the model toward generating outputs that are more aligned with the visual input.

We focus on Visual Contrastive Decoding (VCD) (Leng et al., 2024), a strong inference-time SOTA method. VCD introduces a contrastive re-ranking mechanism, wherein multiple candidate responses are sampled from the model and scored based on both linguistic likelihood and visual alignment. This alignment is computed using a cross-modal similarity function that penalizes syntactically fluent yet visually inconsistent outputs. By re-ranking candidates, VCD encourages the model to favor generations that are both semantically coherent and grounded in the visual input—effectively reducing hallucinations without additional fine-tuning. While model-agnostic and lightweight, such inference-time methods complement VisAlign, which proactively mitigates hallucinations by refining representations during training.

Table 5 compares the effectiveness of VisAlign and VCD applied to Video-LLaVA, both individually and combined. VisAlign outperforms VCD alone, notably improving accuracy (**54.5** → **57.09**) and F1-score (**68.6** → **69.63**). While VCD delivers incremental gains by refining output selection during inference, VisAlign achieves more substantial improvements by addressing modality imbalance during training. When combined, the two methods yield the best overall performance, further boosting accuracy to **58.8** and F1-score to **70.04**, demonstrating their complementary strengths.

These results underscore VisAlign’s orthogonality to inference-time techniques like VCD, allowing it to enhance performance without interference. They also highlight its strong generalizability—VisAlign’s benefits persist even when integrated with other hallucination mitigation strategies, showcasing its robustness across diverse settings.

Model	Existence	Count	Position	Color	Posters	Celebrity	Scene	Landmark	Artwork	OCR
Video-LLaVA	170	121.66	<b>88.33</b>	135	103.74	<b>101.47</b>	<b>163</b>	<b>161</b>	<b>107</b>	87.5
+VCD (Leng et al., 2024)	170	105.00	76.66	125	100.00	100.88	155.75	154.5	99.25	77.5
<b>+VisAlign</b>	<b>190</b>	<b>131.66</b>	53.33	<b>148.33</b>	103.06	78.24	151	125	94	87.5

Table 7: Comparison of baseline Video-LLava with different combination of hallucination mitigation approaches on **MME**.

Table 8: Visual Contribution (%) — Baseline vs. VisAlign. Results clearly favor VisAlign.

Dataset	Baseline	VisAlign
POPE-AOKVQA	63.87	72.26
POPE-MSCOCO	62.90	72.68
TextVQA	59.69	72.27
MME	63.00	72.89
MM-Vet	63.88	72.46

## A.6 PERFORMANCE ON ADDITIONAL BASELINES

### A.6.1 LLAVA1.5

In the main paper, we demonstrated that VisAlign significantly reduces hallucinations in Video-LLaVA by improving attention distribution. To further validate the generality and robustness of VisAlign, we evaluate its effectiveness on another state-of-the-art LVLM, LLaVA 1.5 (Liu et al., 2024a). As shown in Table 4, VisAlign consistently enhances performance and reduces hallucinations when integrated into this baseline as well. These results highlight the broad applicability and effectiveness of the proposed approach across different LVLMs.

### A.6.2 OPEN-QWEN2VL

To further validate the generality and robustness of VisAlign, we evaluate its effectiveness on another state-of-the-art LVLM. We adopt the official Open-Qwen2VL (Wang et al., 2025) training pipeline, a fully open-source multimodal model that tightly integrates visual embeddings with token-level language representations through shared attention layers. Unlike earlier LVLMs that simply append visual tokens to a frozen LLM, Open-Qwen2VL fuses visual features directly into the token embedding space, enabling richer cross-modal interactions and stronger native visual grounding. In our setup, we follow the prescribed two-stage procedure: we complete Stage-1 visual-language alignment pretraining in full, and then continue the full multimodal pretraining only up to 5000 steps due to compute constraints, using this 5000-step checkpoint as the base model for all subsequent VisAlign experiments.

We include Open-Qwen2VL as an additional baseline because it represents a more recent and architecturally distinct family of LVLMs, designed specifically for compute-efficient training on academic resources. Its early and tight fusion of visual information with language tokens provides a complementary testbed to Video-LLaVA, allowing us to examine whether VisAlign remains effective when the underlying model already incorporates stronger visual-text coupling. As shown in Table 4, VisAlign consistently enhances performance and reduces hallucinations when integrated into this baseline as well, indicating that rebalancing attention at the embedding level is beneficial even for modern LVLMs with improved native visual grounding.

## A.7 QUANTITATIVE ASSESSMENT OF VISUAL CONTRIBUTION

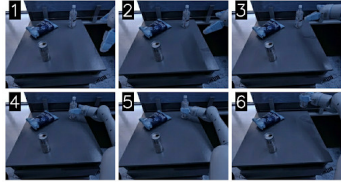
We quantitatively assess “visual contribution” by calculating the proportion of attention allocated to visual tokens (keys) averaged across tokens (queries) and attention heads for 100 randomly selected samples. This analysis is performed only for the first attention layer, as accurate visual-contribution tracking becomes infeasible in deeper layers.

In most LVLMs, visual tokens are inserted in the middle of text tokens, splitting them. Due to the autoregressive nature of the LLM, we skip the first text-token block, since their visual contribution would always be zero.

## A.8 ADDITIONAL QUALITATIVE RESULTS:

Figure 6 (see next page) presents additional qualitative results on the Mementos dataset (Wang et al., 2024a). As illustrated, VisAlign enables the model to produce more visually grounded predictions and significantly reduces hallucinations compared to the baseline Video-LLaVA model.

810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863



**Question:** Write a description for the given image sequence in a single paragraph, what is happening in this episode?

**Video-LLaVA:** The image sequence shows a person using a vacuum cleaner to clean a table. In the first image, the person is holding the vacuum cleaner and preparing to clean the table. In the second image, the person is actively cleaning the table with the vacuum cleaner. In the third image, the person has finished cleaning the table and is holding the vacuum cleaner. In the fourth image, the person is emptying the vacuum cleaner into a trash can. In the fifth image, the person is putting the vacuum cleaner back together. In the sixth image, the person is holding the vacuum cleaner again, possibly preparing to clean another area. Finally, in the seventh image, the person is emptying the vacuum cleaner into a trash can once again.


**Video-LLaVA+VisAlign:** In this image sequence, a robotic arm is seen picking up a bottle and placing it on a table. The scene is captured in four different frames, showing the progression of the action. The robotic arm is seen in different positions in each frame, indicating that it is in the process of picking up the bottle. The bottle is also appears in different positions in each frame, indicating that it is being picked up. The overall scene depicts a robotic arm picking up a bottle and placing it on a table.



**Question:** Write a description for the given image sequence in a single paragraph, what is happening in this episode?

**Video-LLaVA:** The image sequence shows a brown and white dog playing with a stuffed animal. In the first image, the dog is holding the stuffed animal in its mouth, while in the second image, the dog is biting the stuffed animal. The third image shows the dog chewing on the stuffed animal, and the fourth image shows the dog holding the stuffed animal in its mouth again. The fifth image shows the dog biting the stuffed animal, and the sixth image shows the dog chewing on the stuffed animal. The final image shows the dog holding the stuffed animal in its mouth. Overall, the sequence captures the playful and energetic behavior of the dog as it interacts with the stuffed animal.


**Video-LLaVA+VisAlign:** In this image sequence, there are two dogs, a brown and white dog and a black and white dog, playing together in a fenced area. The brown and white dog is biting the black and white dog's ear, and the black and white dog is trying to bite the brown and white dog's face. The two dogs are engaged in a playful interaction, with the brown and white dog being the aggressor and the black and white dog being the defender.



**Question:** Write a description for the given image sequence in a single paragraph, what is happening in this episode?

**Video-LLaVA:** In this image sequence, a group of people is gathered together, posing for a photo. They are standing in front of a building, and some of them are wearing backpacks. The group consists of at least 12 people, with some standing closer to the camera and others further back. The photo captures the essence of a group activity or event, with everyone smiling and enjoying the moment.

**Video-LLaVA+VisAlign:** In this image sequence, there is a man wearing a black jacket and a hat, standing in front of a group of people. The man is taking a selfie with a camera, capturing multiple pictures of himself and the group of people around him. The group consists of people of different ages and genders, and they are all posing for the camera. The man is taking the pictures from different angles and distances, capturing the whole group in each shot. The scene is a lively and fun atmosphere, with everyone enjoying themselves and having a good time.



**Question:** Write a description for the given image sequence in a single paragraph, what is happening in this episode?

**Video-LLaVA:** In this episode, a man is sitting on a chair in a bar, holding a bottle of beer. He is engaged in a conversation with another man who is standing nearby. The scene is depicted in three different drawings, each showing a different moment in the interaction between the two men. The first drawing shows the man sitting on the chair, while the second drawing shows the man standing next to the sitting man. The third drawing shows the man standing in front of the sitting man, indicating that the conversation is ongoing. The bar setting is evident from the presence of a bottle of beer and a chair, and the overall atmosphere appears to be casual and relaxed.

**Video-LLaVA+VisAlign:** In this comic strip, a man is seen sitting at a bar, holding a drink in his hand. Another man is standing behind the bar, pouring a drink into a glass. The man sitting is looking at the drink being poured, and appears to be enjoying the moment. The scene is set in a bar, with bottles and glasses visible in the background. The man pouring the drink is wearing a tie, and the man sitting is wearing a suit. The overall atmosphere of the scene is relaxed and casual, with a hint of humor as the man pouring the drink is exaggerating the pouring motion.

Figure 6: Qualitative results on the Mementos benchmark (Wang et al., 2024a). Text highlighted in red indicates hallucinated content, while text in blue shows the corresponding corrections.