BEYOND VISUAL ENHANCEMENT: ADAPTIVE MULTI-CONTEXT STEERING TO MITIGATE LVLM HALLUCINATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Hallucination remains a significant challenge in Large Vision-Language Models (LVLMs). Existing training-free methods generally mitigate hallucinations through contrastive decoding or visual enhancement, often increasing the relative influence of visual evidence during generation. This raises a fundamental question: **Can LVLMs dynamically regulate the contributions of different context sources to suppress hallucinations?** In this work, we investigate and quantify how LVLMs coordinate multiple context sources during decoding and examine how this intrinsic behavior can guide hallucination mitigation. We find that LVLMs exhibit an intrinsic vision-attending tendency that can guide adaptive visual steering, while textual contexts can also contribute to hallucination mitigation. Motivated by these findings, we propose **AIMS** (Adaptive Information Multi-source Steering), a lightweight training-free framework that adaptively coordinates visual, prefilled textual, and generated contexts during decoding. Specifically, AIMS constructs compact prototypes for the three context domains and estimates their affinities with the current query to determine head-wise steering weights. The resulting multi-source steering direction is applied to the query representation, enabling adaptive context integration without additional model training or auxiliary forward passes. Extensive experiments across multiple LVLMs and decoding strategies demonstrate that AIMS effectively mitigates object hallucination while maintaining competitive general-purpose multimodal capabilities.

1 INTRODUCTION

Recent advances in Large Vision-Language Models (LVLMs) (Chen et al., 2023b; Liu et al., 2024b; Peng et al., 2025; Zhu et al., 2024; Zhang et al., 2024b; Zhao et al., 2024) have enabled remarkable progress in bridging visual perception and language reasoning. By integrating powerful vision encoders with large language models (LLMs), LVLMs demonstrate strong capabilities in tasks such as image understanding (Li et al., 2023), visual question answering (Lee et al., 2024; Wang et al., 2024a), and multimodal reasoning (Deng et al., 2026; Wang et al., 2026; Yang et al., 2025).

Despite their remarkable capabilities, LVLMs still suffer from hallucination (Liu et al., 2024a; Deng et al., 2024), generating content inconsistent with visual inputs, such as non-existent objects, incorrect attributes, or erroneous relationships. Unlike hallucination in text-only LLMs (Shi et al., 2024; Huang et al., 2025), LVLM hallucination is inherently tied to the interaction between visual evidence and language generation, posing a critical challenge to model reliability.

As illustrated in Fig.1, existing training-free methods mainly mitigate hallucinations through contrastive decoding (Leng et al., 2024; Wang et al., 2024b; Huo et al., 2025) or visual enhancement (Liu et al., 2024c; Jiang et al., 2025; Zou et al., 2025). However, they largely overlook the dynamic utilization of contextual information during generation and focus predominantly on visual evidence, leaving textual prompts and generation history underexplored. This raises a fundamental question: **Can LVLMs dynamically regulate different context sources to suppress hallucinations?**

To better explore this issue, we design a series of preliminary experiments to quantify how LVLMs coordinate multiple context sources during autoregressive generation and examine how this intrinsic behavior can guide hallucination mitigation. Our analysis yields three observations. First, LVLMs

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

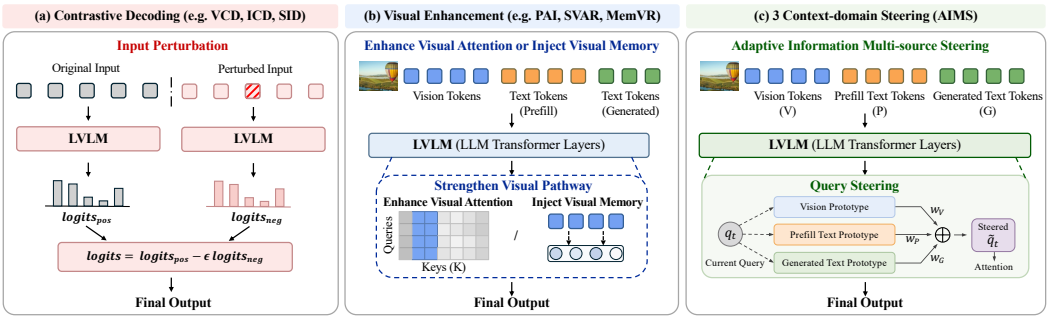


Figure 1: Comparison of training-free hallucination mitigation methods for LVLMs. From left to right: contrastive decoding, visual enhancement, and our method that dynamically steers the query toward three context branches during generation.

exhibit a stage-dependent vision-attending tendency, with visual attention peaking during object grounding and decreasing as generation shifts toward semantic summarization. Second, directly exploiting this intrinsic tendency through query-adaptive steering mitigates hallucination more effectively than uniformly amplifying visual information. Third, prefilled text and generation history also provide corrective signals beyond visual evidence. Together, these findings highlight the need to adaptively coordinate multiple context sources rather than uniformly enhancing visual information.

Building on these observations, we propose **A**daptive **I**nformation **M**ulti-source **S**teering (AIMS), a lightweight training-free framework that adaptively coordinates visual context (V), prefilled text context (P), and previously generated context (G). Specifically, AIMS constructs semantic prototypes for the three contextual sources and estimates their query-dependent affinities at each decoding step. The affinity-weighted prototypes are aggregated to steer the current query representation, enabling dynamic multi-source coordination according to the generation state. Extensive experiments on four benchmarks demonstrate the effectiveness of AIMS, achieving up to a 27.1% relative reduction in CHAIR C_S across different LVLMs and decoding strategies, while preserving general multimodal capabilities. Our main contributions are summarized as follows:

- We investigate the contextual utilization behavior of LVLMs during autoregressive generation and reveal that visual reliance dynamically changes across generation stages. Our analysis further shows that hallucination mitigation requires considering multiple contextual sources beyond visual information alone.
- We propose Adaptive Information Multi-source Steering (AIMS), a training-free framework that dynamically coordinates vision, prefilled text, and generated contexts through query-context affinity. Unlike existing methods that rely on fixed amplification of a single information source, AIMS adaptively integrates evidence during generation.
- We conduct extensive experiments on multiple representative LVLMs, including LLaVA-1.5, Qwen2.5-VL, and Qwen3.5, across four widely used benchmarks. The results demonstrate that AIMS consistently reduces hallucination while effectively preserving overall generation quality compared with existing training-free approaches.

2 Related Work

2.1 Large Vision-Language Models

Modern Large Vision-Language Models (LVLMs) have evolved rapidly through the scaling of architectural connections and language backbones. Early paradigms utilized LLaMA (Touvron et al., 2023a;b) as the linguistic foundation, employing simple MLPs or Q-Formers to project visual tokens, as exemplified by LLaVA (Liu et al., 2023; 2024b) and MiniGPT series (Zhu et al., 2024; Chen et al., 2023a). To handle denser multimodal semantics, subsequent architectures transitioned to highly optimized backbones; the Qwen-VL series (Bai et al., 2023; 2025) introduced VIT-perceiver to internalize fine-grained visual features, while the InternLM-XComposer series (Zhang et al.,

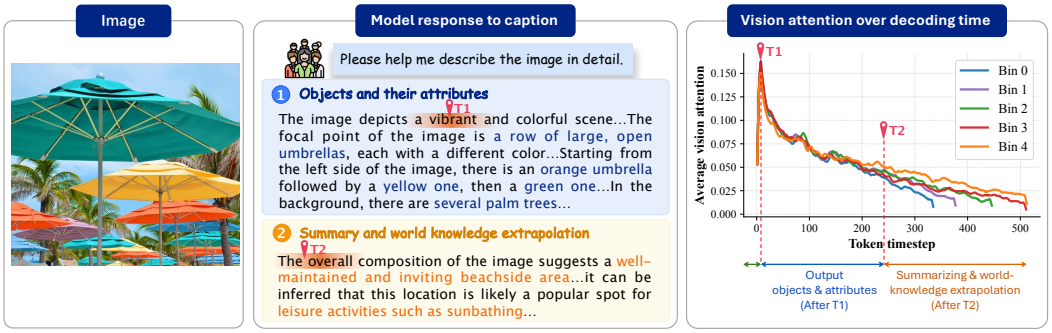


Figure 2: **Visual attention dynamics during decoding.** Results on 500 MSCOCO examples using Qwen2.5-VL-3B, grouped into five sequence-length bins. $T1$ and $T2$ mark the onset of visual description and semantic summarization, respectively.

2023; 2024a) and InternVL (Chen et al., 2024; 2025) scaled up the LLM bases and integrated dynamic high-resolution vision-language interaction. Despite their superior multimodal capabilities, contemporary LVLMs still suffer from severe object hallucination problems (Rohrbach et al., 2018). Effectively mitigating these hallucinations during inference remains a critical challenge, motivating our training-free dynamic steering framework.

2.2 MITIGATING HALLUCINATIONS IN LVLMS

In LVLMs, object hallucination remains the most prevalent form of hallucination, typically manifesting as errors in object category, attribute, and inter-object relation descriptions. To mitigate this issue, existing methods can be broadly divided into several lines. Early studies focused on improving fine-grained vision-language alignment and reducing co-occurrence bias in captioning models (Rohrbach et al., 2018). More recent approaches leverage hallucination-oriented supervision, such as dedicated fine-tuning datasets and RLHF-based optimization (Gunjal et al., 2024; Yu et al., 2024). Another important direction is training-free inference-time intervention. Attentional intervention methods (Jiang et al., 2025; Liu et al., 2024c; Huang et al., 2024) suppress hallucinations by modifying attention behaviors during decoding, but often incur extra inference overhead. In parallel, contrastive-decoding-based methods, such as SID (Huo et al., 2025) and VCD (Leng et al., 2024), steer the decoding distribution by contrasting different visual or textual conditions, though their effectiveness may be unstable due to the additional noise introduced in the contrastive process.

3 PRELIMINARY STUDY

We conduct preliminary experiments on the CHAIR benchmark using 500 MSCOCO examples with Qwen2.5-VL-3B. We examine the temporal dynamics of visual attention during generation, compare fixed and query-adaptive visual steering, and investigate the contributions of different context sources to hallucination mitigation. These analyses yield three key observations.

Table 1: **Preliminary analysis of adaptive and multi-context steering.** (a) Comparison of fixed and adaptive query steering toward the vision prototype. (b) Impact of different context branches under fixed-strength steering. C_S and C_I denote CHAIR_S and CHAIR_I, respectively.

(a) Adaptive Visual Steering

Method	$C_S \downarrow$	$C_I \downarrow$	Recall↑	F1↑	Length↑
Baseline	56.0	9.96	74.05	74.10	404.19
Fixed-V$_{\alpha=0.05}$	51.6	9.48	74.11	74.99	350.88
Fixed-V$_{\alpha=0.15}$	**39.0**	9.30	70.18	75.29	199.70
Adaptive-V	40.2	**8.33**	71.53	75.08	292.41

(b) Multi-context Steering

Method	$C_S \downarrow$	$C_I \downarrow$	Recall↑	F1↑	Length↑
Fixed-V	51.60	9.48	74.11	74.99	350.88
Fixed-V-G	47.80	8.95	73.10	75.15	345.88
Fixed-P-G	47.62	9.35	73.54	74.68	382.81
Fixed-G	50.61	**8.80**	74.68	75.40	385.38
Fixed-V-P-G	**46.00**	8.94	73.35	75.61	317.33

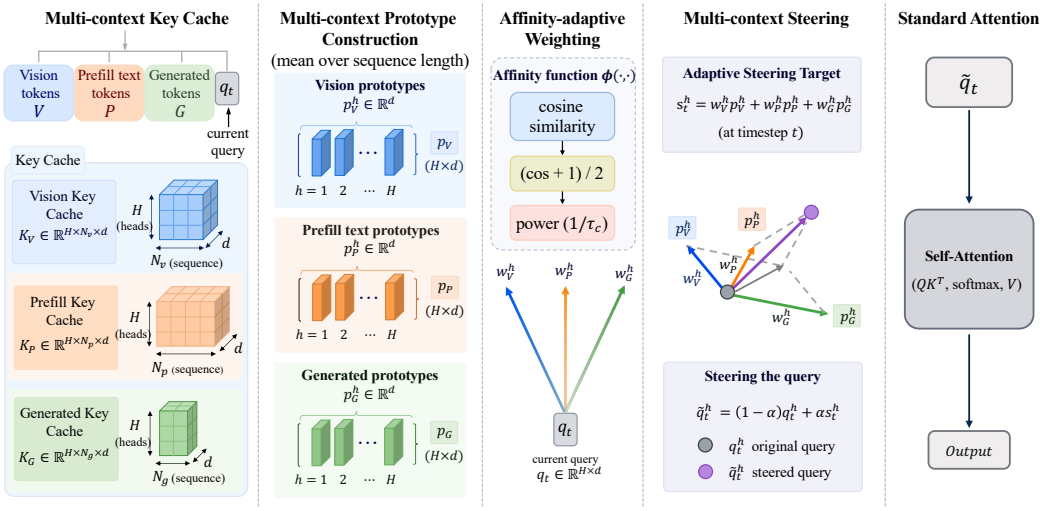


Figure 3: **Overview of AIMS.** Given vision, prefilled text, and generated-token key caches, AIMS constructs head-wise multi-context prototypes and adaptively weights them according to their affinity with the current query. The weighted prototypes form a dynamic steering target that adjusts the query representation before standard self-attention.

Observation 1. Vision attention naturally peaks when object grounding is required. To examine how visual reliance evolves throughout generation, we evenly divide the generated sequences into five length bins according to the number of samples and measure the average attention from the query to visual tokens at each generation timestep. As shown in Fig. 2, despite substantial differences in sequence length, visual attention consistently peaks around $T1 \in \{4, 5\}$, when generation moves beyond generic openings (e.g. "*The image depicts a ...*") and begins grounding concrete objects and attributes. It then gradually declines as generation shifts toward summarization and world-knowledge extrapolation at $T2$, where less direct visual evidence is required. This pattern aligns with human intuition: during summarization and extrapolation, generation can rely on previously established context rather than direct visual evidence.

Observation 2. Adaptive visual steering achieves a better hallucination-generation trade-off than fixed amplification. Table 1 (a) compares fixed visual steering with query–vision affinity-based adaptive steering. Strong fixed steering ($\alpha = 0.15$) reduces C_S to 39.0 but sharply shortens responses to 199.70 tokens. In contrast, $Adaptive\text{-}V$ achieves a comparable C_S of 40.2 while retaining 292.41 tokens and reducing C_I to 8.33, suggesting that visual evidence should be reinforced according to the model's intrinsic vision-attending tendency rather than uniformly amplified.

Observation 3. Hallucination mitigation benefits from multiple context branches beyond vision. As shown in Table 1 (b), textual contexts also provide useful corrective signals: $P+G$ reduces C_S from 56.00 to 47.62 without visual reinforcement, while G alone achieves a C_I of 8.80. Combining V, P, and G further yields the lowest C_S of 46.00 among the branch combinations, motivating multi-source coordination rather than vision-only enhancement.

4 METHOD

The above observations reveal a phenomenon overlooked by existing methods: LVLMs inherently exhibit a semantically meaningful vision-attending tendency aligned with their evolving generation needs, suggesting that hallucination mitigation can benefit from reinforcing, rather than overriding, this intrinsic tendency. Moreover, enhancing prefilled and generated contexts can also mitigate hallucinations, indicating that useful corrective signals extend beyond visual evidence. Motivated by these findings, we propose **AIMS** (Adaptive Information Multi-source Steering), a lightweight training-free framework that adaptively coordinates three context sources inherent to autoregressive generation: vision tokens (V) for perceptual evidence, prefilled text tokens (P) for task intent, and previously generated tokens (G) for contextual coherence. By estimating their query-dependent

affinities, AIMS dynamically determines which contextual information should guide each decoding step, rather than relying on fixed or vision-centric steering.

Multi-context prototype construction. In Transformer attention, the relevance of contextual information to the current generation state is naturally captured through query–key matching. We therefore leverage the existing key cache to construct compact prototypes for each context source. Specifically, we identify vision tokens (V), prefilled text tokens (P), and previously generated tokens (G) by their token indices and mean-pool their key representations within each source. For attention head h, the prototype of context source i is defined as

$$\mathbf{p}_i^h = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbf{k}_{i,j}^h, \qquad i \in \{V, P, G\}, \tag{1}$$

where $\mathbf{k}_{i,j}^h$ denotes the key representation of the j-th token in context source i, and N_i is the number of tokens in that source. The resulting prototypes $\mathbf{p}_V^h$, $\mathbf{p}_P^h$, and $\mathbf{p}_G^h$ provide compact summaries of the three context sources in the model's native query–key representation space, allowing their relevance to the current query to be measured in a unified manner. Since the prototypes are constructed directly from the existing key cache, this process requires no additional encoding or auxiliary forward passes and introduces only lightweight computational overhead.

Affinity-adaptive weighting. Observation 2 suggests that effective intervention should follow, rather than override, the model's intrinsic information-attending tendency. Since the current query encodes the generation state at timestep t, its compatibility with different context prototypes provides a natural internal signal of which information sources are currently relevant. We therefore use query–context affinity to adaptively determine the contribution of each context source.

Specifically, for context source $c \in \{V, P, G\}$ and attention head h, we measure the affinity between the current query $\mathbf{q}_t^h$ and the corresponding prototype $\mathbf{p}_{c,t}^h$ using cosine similarity and map it to $[0, 1]$. The adaptive weight is defined as

$$w_{c,t}^h = \left(\frac{\cos(\mathbf{q}_t^h, \mathbf{p}_{c,t}^h) + 1}{2} \right)^{1/\tau_c}, \qquad c \in \{V, P, G\}, \tag{2}$$

where τ_c controls the sensitivity of source c to its query–context affinity. Consequently, the resulting weights are simultaneously source-, head-, and timestep-adaptive, allowing AIMS to capture fine-grained variations in contextual demand throughout autoregressive generation.

For notational simplicity, we use $\mathbf{p}_{c,t}^h$ to denote the prototype available at timestep t. The visual and prefilled contexts remain unchanged during decoding, and thus $\mathbf{p}_{V,t}^h = \mathbf{p}_V^h$ and $\mathbf{p}_{P,t}^h = \mathbf{p}_P^h$ are fixed. In contrast, the generated context grows autoregressively. Instead of repeatedly pooling an increasingly long generation history, we use the key representation of the latest generated token as $\mathbf{p}_{G,t}^h$. As this representation has already contextualized the preceding tokens through causal self-attention, it provides a naturally updated and lightweight summary of the generation history.

Multi-context steering. Rather than separately reweighting the attention logits of the V, P, and G branches, AIMS first integrates their affinity-weighted prototypes into a unified steering target and uses it to update the current query. The updated query then interacts with all cached keys through the original attention operation, allowing the three context sources to jointly influence the attention assigned to individual tokens. Specifically, we construct a head-wise steering target by integrating the context prototypes according to their adaptive affinities:

$$\mathbf{s}_t^h = \sum_{c \in \{V,P,G\}} w_{c,t}^h \mathbf{p}_{c,t}^h. \tag{3}$$

The current query is then updated as

$$\widetilde{\mathbf{q}}_t^h = (1 - \alpha)\mathbf{q}_t^h + \alpha \mathbf{s}_t^h, \tag{4}$$

where α controls the overall steering strength. The modified query is subsequently passed to the original attention operation:

$$\mathbf{o}_\mathbf{h}^\mathbf{t} = \mathrm{Softmax}\left(\frac{\widetilde{\mathbf{q}}_t^h (\mathbf{K}^h)^\top}{\sqrt{d_h}} \right) \mathbf{V}^\mathbf{h}. \tag{5}$$

AIMS modifies only the current query representation while leaving the key-value cache and the original attention computation unchanged, without requiring an additional model forward pass.

5 EXPERIMENT

Table 2: Evaluation results of different methods on CHAIR.

Method	LLaVA-1.5-7B														
	Greedy					Beam Search					Nucleus				
	$C_S\downarrow$	$C_I\downarrow$	R↑	F1↑	Len↑	$C_S\downarrow$	$C_I\downarrow$	R↑	F1↑	Len↑	$C_S\downarrow$	$C_I\downarrow$	R↑	F1↑	Len↑
Baseline	51.4	14.70	79.42	76.80	97.9	55.2	14.93	80.34	77.10	103.5	50.0	15.31	74.05	73.19	101.5
MemVR	49.8	14.20	78.64	76.81	93.8	53.0	14.30	79.54	77.53	95.7	52.6	15.80	72.91	72.73	93.6
OPERA	–	–	–	–	–	44.6	12.72	79.25	78.40	92.7	–	–	–	–	–
VCD	–	–	–	–	–	–	–	–	–	–	49.9	15.72	75.24	74.00	94.6
VTI	48.8	13.89	78.43	77.06	94.1	48.8	12.87	79.57	78.74	97.9	49.2	14.98	73.86	73.85	95.8
SID	47.6	14.43	76.93	76.76	94.7	–	–	–	–	–	52.0	15.51	75.38	74.60	95.5
AIMS	**43.6**	**13.17**	76.84	77.96	89.9	**43.6**	**12.02**	76.46	77.24	93.5	**47.1**	**15.23**	71.50	72.91	85.6
Method	**Qwen2.5-VL-3B**														
	Greedy					**Beam Search**					**Nucleus**				
	$C_S\downarrow$	$C_I\downarrow$	R↑	F1↑	Len↑	$C_S\downarrow$	$C_I\downarrow$	R↑	F1↑	Len↑	$C_S\downarrow$	$C_I\downarrow$	R↑	F1↑	Len↑
Baseline	56.0	9.96	74.05	74.10	404.2	45.1	8.93	73.41	76.19	367.3	56.6	11.39	72.96	72.50	441.3
MemVR	51.8	9.62	73.90	74.18	412.5	43.2	8.30	73.41	76.70	368.0	57.6	10.66	72.78	72.53	445.5
OPERA	–	–	–	–	–	45.2	8.15	71.70	74.93	361.6	–	–	–	–	–
VCD	–	–	–	–	–	–	–	–	–	–	57.8	9.99	73.86	73.00	477.0
VTI	43.2	**8.02**	70.88	75.02	338.2	41.8	8.47	72.53	76.10	358.6	52.4	9.78	71.51	73.00	430.0
SID	50.3	12.13	74.78	75.08	207.8	–	–	–	–	–	50.6	10.15	73.35	74.25	386.4
AIMS	**40.8**	9.07	71.69	74.66	284.2	**38.2**	**7.67**	71.84	76.27	325.2	**44.8**	**9.33**	70.56	73.62	327.6
Method	**Qwen3.5-9B**														
	Greedy					**Beam Search**					**Nucleus**				
	$C_S\downarrow$	$C_I\downarrow$	R↑	F1↑	Len↑	$C_S\downarrow$	$C_I\downarrow$	R↑	F1↑	Len↑	$C_S\downarrow$	$C_I\downarrow$	R↑	F1↑	Len↑
Baseline	47.4	11.07	74.11	75.87	224.3	50.0	10.82	75.19	76.08	234.0	51.0	12.03	73.86	74.74	227.1
MemVR	61.0	11.39	78.11	74.67	403.2	60.4	11.60	73.41	76.70	368.0	57.6	10.66	78.11	74.67	403.2
VCD	–	–	–	–	–	–	–	–	–	–	55.2	12.61	74.79	73.98	302.7
VTI	47.6	**9.14**	69.23	72.98	355.6	46.0	**8.22**	68.59	72.99	356.8	50.6	**9.82**	68.78	71.95	357.9
SID	59.6	11.74	77.48	74.36	402.9	–	–	–	–	–	57.8	11.29	78.10	74.76	404.5
AIMS	**40.0**	9.91	67.05	73.38	249.9	**42.6**	9.45	69.52	74.52	270.9	**47.0**	10.62	69.56	72.57	256.8

5.1 EXPERIMENTAL SETTINGS

Models and Baselines. We evaluate AIMS on three representative LVLMs, including LLaVA-1.5-7B (Liu et al., 2024b), Qwen2.5-VL-3B (Bai et al., 2025), and Qwen3.5-9B. Since AIMS is a training-free framework, we compare it with representative hallucination mitigation methods, including VCD (Leng et al., 2024), SID (Huo et al., 2025), VTI (Liu et al., 2025), OPERA (Huang et al., 2024), MemVR (Zou et al., 2025) and Baseline.

For decoding strategies, we evaluate AIMS under three generation settings: greedy decoding, beam search with 5 beams, and nucleus sampling with temperature=1, top-p=0.9, and top-k=50. For comparison methods, we follow the decoding strategies supported in their original papers.

Datasets. CHAIR (Rohrbach et al., 2018) evaluates object hallucination in image captioning. Following previous works, we randomly sample 500 MSCOCO (Lin et al., 2014) images and report CHAIR_I (C_I), CHAIR_S (C_S), Recall (R), F1, and generation length (Len). **AMBER-G** (Wang et al., 2023) evaluates object, attribute, and relation hallucinations in free-form descriptions. We follow the official protocol and report $CHAIR$, $Cover$, Hal, and Cog. **FaithScore** (Jing et al., 2024) evaluates free-form hallucinations through atomic-fact verification. We evaluate on LLaVA-1k and report FaithScore and Sentence-FaithScore. **MME** (Fu et al., 2025) evaluates perception and cognition across 14 subtasks. We report perception (Per), cognition (Cog), and total (Sum) scores.

Implementation details. We mainly consider $(\tau_V, \tau_P, \tau_G) \in \{(1.0, 1.0, 1.0), (1.5, 1.2, 1.1)\}$ for the domain-specific sensitivity parameters in Eq. 2. The steering strength α in Eq. 4 is set to 0.05 or

Table 3: Evaluation results of different methods on AMBER-G.

Method	Qwen2.5-VL-3B											
	Greedy				Beam Search				Nucleus			
	CHAIR↓	Cover↑	Hal↓	Cog↓	CHAIR↓	Cover↑	Hal↓	Cog↓	CHAIR↓	Cover↑	Hal↓	Cog↓
Baseline	8.2	69.6	52.7	5.7	6.8	68.3	44.7	4.9	9.9	70.3	63.7	7.4
VTI	6.9	**65.0**	43.8	4.3	5.8	63.8	35.1	3.9	8.5	**65.9**	51.6	5.4
AIMS	**6.6**	64.8	**39.7**	**3.8**	**5.2**	**64.0**	**33.2**	**3.1**	**7.8**	65.2	**47.9**	**4.2**
Method	**Qwen3.5-9B**											
	Greedy				**Beam Search**				**Nucleus**			
	CHAIR↓	Cover↑	Hal↓	Cog↓	CHAIR↓	Cover↑	Hal↓	Cog↓	CHAIR↓	Cover↑	Hal↓	Cog↓
Baseline	10.4	75.7	71.8	5.5	9.6	76.4	69.3	5.3	10.2	75.6	70.5	5.0
VTI	9.6	75.2	69.1	**5.4**	9.3	75.6	69.4	**5.3**	10.0	**75.4**	71.8	5.8
AIMS	**9.4**	**75.3**	**68.9**	5.7	**9.1**	**75.9**	**65.8**	5.6	**9.4**	74.8	**65.9**	**5.4**

Table 4: FaithScore evaluation results on Qwen2.5-VL-3B. FS and S-FS denote FaithScore and sentence-level FaithScore, respectively.

Method	Decoding	FS↑	S-FS↑
Baseline		0.944	0.786
VTI	Greedy	0.942	0.787
AIMS		**0.945**	**0.791**
Baseline		0.942	0.780
VTI	Beam	0.947	0.793
AIMS		**0.948**	**0.854**
Baseline		0.925	0.745
VTI	Nucleus	0.927	0.746
AIMS		**0.929**	**0.789**

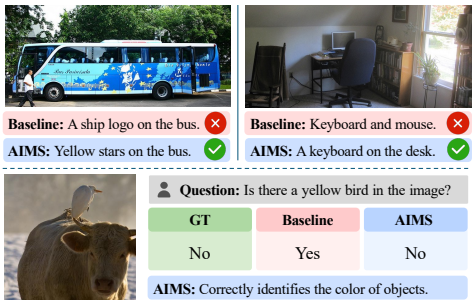


Figure 4: Qualitative comparison between the baseline and AIMS.

0.08 for LLaVA-1.5 and Qwen2.5-VL, and to 0.3 or 0.4 for Qwen3.5. All experiments are conducted on NVIDIA GPUs for consistent hardware evaluation.

Table 5: Evaluation results of different methods on MME.

Method	Qwen2.5-VL-3B								
	Greedy			Beam Search			Nucleus		
	Per↑	Cog↑	Sum↑	Per↑	Cog↑	Sum↑	Per↑	Cog↑	Sum↑
Baseline	1591.38	612.86	2204.24	1558.17	537.14	2095.31	1466.52	513.93	1980.45
MemVR	1591.93	626.42	2218.36	1538.66	544.64	2083.30	1425.96	523.92	1949.88
VCD	–	–	–	–	–	–	1485.20	521.78	2006.98
VTI	1472.86	522.50	1995.36	1520.20	480.35	2000.55	1448.89	518.57	1967.46
SID	1590.13	606.78	2196.91	–	–	–	1450.47	557.14	**2007.61**
AIMS	1591.23	630.35	**2221.58**	1550.71	540.00	**2090.71**	1492.41	487.82	1980.23
Method	**Qwen3.5-9B**								
	Greedy			**Beam Search**			**Nucleus**		
	Per↑	Cog↑	Sum↑	Per↑	Cog↑	Sum↑	Per↑	Cog↑	Sum↑
Baseline	1714.86	678.57	2393.43	1722.59	584.64	2307.23	1587.53	567.85	2155.38
MemVR	1707.45	671.07	**2378.52**	1709.63	574.72	2284.35	1580.36	569.08	2149.44
VCD	–	–	–	–	–	–	1603.84	589.21	2193.05
VTI	1631.79	592.14	2223.93	1652.71	508.21	2160.92	1501.74	535.00	2036.74
SID	1693.31	656.07	2349.38	–	–	–	1620.75	593.92	2214.67
AIMS	1699.97	653.92	2353.90	1710.11	598.21	**2308.32**	1625.41	591.78	**2217.20**

5.2 EXPERIMENTAL RESULTS

Results on CHAIR. As shown in Table 2, AIMS achieves the lowest C_S in all nine model-decoding settings, consistently outperforming the baseline and existing training-free methods. Notably, AIMS maintains a consistent advantage over VTI across all three LVLMs and decoding strategies. While VTI applies pre-computed intervention directions consistently across queries, AIMS

adapts its steering to the current generation state. This query-dependent coordination may explain its stable advantage across different decoding dynamics. Compared with contrastive decoding methods such as VCD and SID, AIMS also achieves stronger hallucination reduction without constructing additional contrastive inputs or auxiliary forward passes.

Results on AMBER-G. As shown in Table 3, AIMS consistently achieves lower Hal than VTI across all six model-decoding settings, indicating fewer hallucinated responses, while maintaining comparable $Cover$ scores and thus similar object coverage. This suggests that its hallucination reduction does not simply result from mentioning fewer objects present in the image. Beyond this general trend, AIMS also reduces Cog across all decoding strategies on Qwen2.5-VL-3B, indicating a lower tendency to generate the human-cognition-associated hallucinated objects captured by AMBER. Together with the consistently lower $CHAIR$ scores, these results show that AIMS improves hallucination control at both the object and response levels without a substantial coverage trade-off.

Results on FaithScore. As shown in Table 4, AIMS achieves the highest FS and S-FS across all three decoding strategies, with particularly pronounced sentence-level gains under beam search and nucleus sampling. Unlike object-centric metrics, FaithScore evaluates descriptive content by decomposing it into atomic facts and verifying their consistency with the image. The improvements therefore provide complementary evidence that AIMS extends beyond reducing hallucinated objects to improving the faithfulness of fine-grained visual facts in free-form responses.

Results on General-purpose Benchmark. Together with Tables 2–4, Table 5 shows that AIMS largely preserves the general-purpose multimodal capabilities of the original models while substantially reducing hallucinations. VTI achieves effective hallucination mitigation but exhibits noticeable MME degradation on the two Qwen models, whereas MemVR better preserves general capabilities but provides more limited hallucination reduction. Overall, AIMS achieves a better balance between hallucination mitigation and general-purpose capability preservation.

Table 6: Ablation of different affinity weighting functions on Qwen2.5-VL-3B.

Method	Greedy					Beam Search					Nucleus				
	$C_S\downarrow$	$C_I\downarrow$	R↑	F1↑	Len↑	$C_S\downarrow$	$C_I\downarrow$	R↑	F1↑	Len↑	$C_S\downarrow$	$C_I\downarrow$	R↑	F1↑	Len↑
AIMS$_{RBF}$	42.8	8.85	71.51	75.31	301.5	38.4	7.87	73.67	77.40	311.2	51.0	9.74	70.56	72.94	339.9
AIMS$_{NCS}$	**40.2**	**8.33**	71.53	75.08	292.4	38.4	7.84	72.84	77.02	325.9	45.6	9.64	70.75	73.74	335.3
AIMS$_{KL}$	46.9	8.48	71.40	74.27	344.7	**36.2**	**7.57**	72.84	77.41	314.7	51.8	9.43	71.83	72.54	399.6
AIMS	40.8	9.07	71.69	74.66	284.2	38.2	7.67	71.84	76.27	325.2	**44.8**	**9.33**	70.56	73.62	327.6

5.3 ABLATION AND ANALYSIS

Ablation on steering configurations. We further examine the steering configurations of AIMS on Qwen2.5-VL-3B using CHAIR under greedy decoding, including the choice of context branches, adaptive weighting, and key hyperparameters. For qualitative illustration, Figure 4 presents several examples, with additional case studies provided in Appendix A.2.

Context Branches. Figure 5 investigates the contributions of context branches in AIMS. Combining the visual branch with either prefilled text context ($V+P$) or generation history ($V+G$) exhibits trade-offs across the evaluation metrics. In contrast, jointly incorporating all three branches ($V+P+G$) achieves the best performance on $CHAIR$ and $F1$. This demonstrates the complementary roles of different contextual sources and supports the use of multi-source steering in AIMS.

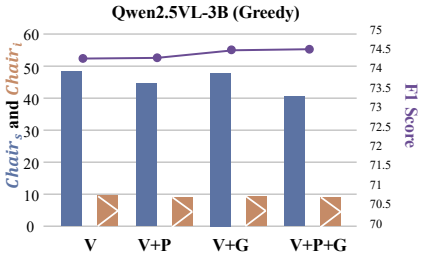


Figure 5: Ablation study of different context branches in AIMS.

Ablation on adaptive weighting functions. As shown in Table 6, AIMS reduces hallucination with different affinity functions, showing its robustness to the weighting function (see Appendix A.1 for function details).

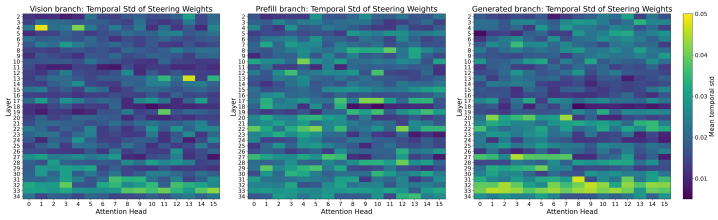


Figure 6: Temporal variability of affinity-adaptive weights across different contextual branches. From left to right: V, P, and G.

Table 7: Fixed vs. Affinity-adaptive weighting.

Method	$C_S \downarrow$	$C_I \downarrow$	F1 $\uparrow$
Baseline	56.0	9.96	74.10
Fixed ($\alpha = 0.1$)	8.0	4.64	56.90
Fixed ($\alpha = 0.01$)	54.0	10.11	74.45
AIMS ($\alpha = 0.08$)	40.8	9.07	74.66

Affinity-adaptive Weighting. Table 7 compares fixed-strength steering with the affinity-adaptive weighting used in AIMS. While strong fixed steering ($\alpha = 0.1$) substantially reduces hallucination, it severely degrades $F1$, indicating excessive intervention in generation. More importantly, under comparable $F1$ scores (the last two rows), AIMS achieves substantially lower C_S and C_I, demonstrating a better trade-off between hallucination mitigation and response quality.

Table 8: Ablation of α and τ.

Method	$\boldsymbol{\alpha}$	$(\boldsymbol{\tau_V}, \boldsymbol{\tau_P}, \boldsymbol{\tau_G})$	$C_S \downarrow$	$C_I \downarrow$	F1 $\uparrow$
AIMS	0.08	(1.5, 1.2, 1.1)	40.8	9.07	74.66
	0.05	(1.5, 1.2, 1.1)	49.8	8.95	75.51
	0.08	(1.0, 1.0, 1.0)	44.4	7.59	75.22

α and τ. Table 8 further studies the sensitivity to the steering strength α and the branch-specific temperatures τ, where a larger τ indicates a broader sensitivity bandwidth to the corresponding context branch. Across different configurations of both α and τ, AIMS consistently reduces hallucination compared with the original baseline while maintaining competitive $F1$ scores, demonstrating its robustness to these hyperparameter choices.

Overhead Analysis. Table 9 reports the overhead measured on a fixed subset of 10 CHAIR images. AIMS introduces nearly no additional peak memory over the baseline and achieves the second-lowest inference latency, closely approaching MemVR, demonstrating a favorable trade-off between effectiveness and computational efficiency.

Table 9: Comparison of computational overhead on Qwen2.5-VL-3B.

Method	Peak Mem. (MB) $\downarrow$	Time (s/token) $\downarrow$
Baseline	7235.87	0.0437
MemVR	<u>7239.53</u>	**0.0698**
VCD	7387.08	0.0905
VTI	7482.31	0.2424
SID	7759.74	0.1075
AIMS	**7236.11**	<u>0.0705</u>

Temporal Adaptivity of Steering Weights. We measure the temporal variability of steering weights using their standard deviation across decoding timesteps. As shown in Figure 6, the three branches exhibit distinct layer- and head-specific patterns. For V, large variations are sparsely concentrated in individual heads at earlier layers (e.g., Layers 4, 13, and 19), but become broadly distributed across heads in Layers 31–33. P shows more dispersed variations across middle and deep layers, whereas G exhibits stronger layer-wise structures around Layers 20, 27, and 31–33. These distinct patterns suggest different dynamic utilization of contextual sources during decoding, motivating fine-grained affinity-adaptive multi-source steering across timesteps.

6 Conclusion

In this work, we revisit object hallucination in LVLMs from the perspective of multi-context interaction. Our analysis systematically reveals an intrinsic vision-attending tendency during generation and shows that complementary contextual information beyond visual evidence can also contribute to hallucination mitigation. Motivated by these findings, we propose AIMS, a training-free method that constructs compact context prototypes and dynamically coordinates their contributions during decoding. Extensive experiments across different LVLM architectures and decoding strategies demonstrate that AIMS consistently reduces object hallucination while preserving strong general-purpose multimodal capabilities. We hope our findings encourage a shift from visual-centric intervention toward adaptive multi-context coordination for hallucination mitigation.

AI USE STATEMENT

Generative AI tools were used for translation and for improving the readability and linguistic clarity of the manuscript. They were not used to develop the research methodology, design experiments, implement the proposed method, analyze or interpret experimental results, or formulate scientific claims. All AI-assisted text was reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work does not involve human subjects or the collection of private or sensitive data. All experiments are conducted using publicly available models and benchmarks. We do not identify any significant ethical concerns arising from this work.

REPRODUCIBILITY STATEMENT

We provide implementation details, hyperparameter settings, and experimental configurations in the main text and Appendix. The source code will be released to facilitate reproduction of our results.

REFERENCES

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL https://arxiv.org/abs/2308.12966.

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

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning, 2023a. URL https://arxiv.org/abs/2310.09478.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023b.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24185–24198, June 2024.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025. URL https://arxiv.org/abs/2412.05271.

Ailin Deng, Zhirui Chen, and Bryan Hooi. Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding. *arXiv preprint arXiv:2402.15300*, 2024.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: Complex vision-language reasoning via iterative sft-rl cycles. *Advances in Neural Information Processing Systems*, 38:123817–123846, 2026.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18135–18143, 2024.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM transactions on information systems*, 43(2):1–55, 2025.

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

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 24272–24295, 2025.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25004–25014. IEEE, 2025.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5042–5063, 2024.

Jusung Lee, Sungguk Cha, Younghyun Lee, and Cheoljong Yang. Visual question answering instruction: Unlocking multimodal large language model to domain-specific visual multitasks. *arXiv preprint arXiv:2402.08360*, 2024.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, June 2024b.

Sheng Liu, Haotian Ye, and James Y Zou. Reducing hallucinations in large vision-language models via latent space steering. In *International Conference on Learning Representations*, volume 2025, pp. 72402–72419, 2025.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pp. 125–140. Springer, 2024c.

Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1437. URL https://aclanthology.org/D18-1437/.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 783–791, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023a. URL https://arxiv.org/abs/2302.13971.

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

Haibo Wang, Chenghang Lai, Yixuan Sun, and Weifeng Ge. Weakly supervised gaussian contrastive grounding with large multimodal models for video question answering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 5289–5298, 2024a.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *Advances in Neural Information Processing Systems*, 38:30865–30891, 2026.

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

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the association for computational linguistics: ACL 2024*, pp. 15840–15853, 2024b.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2376–2385. IEEE, 2025.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment

from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13807–13816, 2024.

Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Haodong Duan, Songyang Zhang, Shuangrui Ding, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition, 2023. URL https://arxiv.org/abs/2309.15112.

Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, Songyang Zhang, Wenwei Zhang, Yining Li, Yang Gao, Peng Sun, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Hang Yan, Conghui He, Xingcheng Zhang, Kai Chen, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output, 2024a. URL https://arxiv.org/abs/2407.03320.

Yi-Kai Zhang, Shiyin Lu, Yang Li, Yanqing Ma, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, and Han-Jia Ye. Wings: Learning multimodal llms without text-only forgetting. In *Advances in Neural Information Processing Systems*, volume 37, pp. 31828–31853. Curran Associates, Inc., 2024b. doi: 10.52202/079017-1001.

Zhen Zhao, Jingqun Tang, Binghong Wu, Chunhui Lin, Shu Wei, Hao Liu, Xin Tan, Zhizhong Zhang, Can Huang, and Yuan Xie. Harmonizing visual text comprehension and generation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 97499–97522. Curran Associates, Inc., 2024. doi: 10.52202/079017-3093.

Deyao Zhu, jun chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 18378–18394, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/50623630a2372839c078474efa6c0cb8-Paper-Conference.pdf.

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, and Xuming Hu. Look twice before you answer: memory-space visual retracing for hallucination mitigation in multimodal large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org, 2025.

A Appendix

A.1 Alternative Affinity Weighting Functions

We evaluate three alternative affinity weighting functions in addition to our default formulation, with results reported in Table 6. Here, $\mathbf{q}_t^h$, $\mathbf{p}_{c,t}^h$, and τ_c follow the notation in Sec. 4.

RBF. The RBF-based weighting measures the Euclidean distance between the query and contextual prototype, assigning larger weights to closer representations:

$$w_{c,t}^h = \exp\left(-\frac{\left\|\mathbf{q}_t^h - \mathbf{p}_{c,t}^h\right\|_2^2}{2\tau_c^2 d_h}\right), \qquad c \in \{V, P, G\}, \tag{6}$$

where d_h denotes the dimensionality of each attention head.

NCS. The nonlinear cosine similarity (NCS) first computes the normalized inner product between the query and contextual prototype, followed by softplus and sigmoid transformations:

$$w_{c,t}^h = \mathrm{sigmoid}\left(\mathrm{softplus}\left(\frac{\cos(\mathbf{q}_t^h, \mathbf{p}_{c,t}^h)}{\tau_c}\right)\right), \qquad c \in \{V, P, G\}. \tag{7}$$



Instruction: Please describe the image in detail.

Baseline

This is a street-level photograph of a city intersection on what appears to be an overcast day. The scene is captured from across the street, looking towards a crosswalk.\n\nIn the foreground, a white-striped crosswalk spans the width of the asphalt road. On the left side of the frame, a man in a blue shirt and jeans walks away from the camera, while another person in a light blue jacket and jeans walks towards the right. In the middle of the crosswalk, a woman in a dark jacket and brown pants is walking to the right, followed by another person in a beige coat pulling a dark suitcase.\n\nThe background is filled with urban elements. On the left, there's a modern building with large glass windows and a curved facade. A tall, vertical sign with red and blue lettering stands prominently. Further down the street, more buildings and signs are visible, including one that reads \"HELLO!\" in English.\n\nA traffic light on a pole in the center of the image shows a red hand signal, indicating \"don't walk.\" The street is lined with green trees, and several other pedestrians can be seen further down the sidewalk on both sides of the street. The overall atmosphere is that of a typical, active day in a Japanese city.\n

Hallucination: the person in beige coat is holding a bag, not a suitcase.

AIMS (Ours)

This is a street-level photograph of a busy urban scene, likely in Japan, captured on a bright, sunny day. The perspective is from the sidewalk, looking across a wide intersection.\n\nIn the foreground, a paved crosswalk with white stripes stretches across the frame. A few pedestrians are crossing the street. On the left, a man in a blue shirt and jeans walks away from the camera. To his right, another person in a light-colored jacket walks away from the viewer. On the right side of the frame, a woman in a beige coat and a man in a dark jacket are walking together across the street.\n\nThe background is filled with tall buildings, shops, and trees. A prominent feature is a tall, red and white sign for a Japanese convenience store or supermarket, with Japanese characters visible. Other signs, including one that appears to be for a bank or financial institution, are also visible. The overall scene captures a typical day in a bustling city, with people going about their daily routines. The lighting suggests it is daytime, and the overall atmosphere is one of everyday urban life.

Correct: avoids incorrectly describing the bag as a suitcase.

Figure 7: Case study 1 on the CHAIR benchmark using Qwen3.5-9B with greedy decoding.

KL. For KL-based weighting, we first transform the query and contextual prototype into probability distributions:

$$\boldsymbol{\pi}^h_{q,t} = \mathrm{softmax}\left(\frac{\mathbf{q}^h_t}{\tau_c}\right), \qquad \boldsymbol{\pi}^h_{c,t} = \mathrm{softmax}\left(\frac{\mathbf{p}^h_{c,t}}{\tau_c}\right). \tag{8}$$

The steering weight is then determined by their KL divergence:

$$w^h_{c,t} = \exp\left(-D_{\mathrm{KL}}\left(\boldsymbol{\pi}^h_{q,t} \,\|\, \boldsymbol{\pi}^h_{c,t}\right)\right), \qquad c \in \{V, P, G\}. \tag{9}$$

A.2 Case Studies

We provide qualitative comparisons between Baseline and AIMS across different LVLMs and benchmarks. Figures 7–10 present examples on CHAIR using Qwen3.5-9B and Qwen2.5-VL-3B, while Figure 11 provides an example on AMBER-G. Figure 12 presents examples on MME using Qwen3.5-9B with nucleus sampling. These cases qualitatively illustrate the differences between Baseline and AIMS outputs.



Instruction: Please describe the image in detail.

Baseline

This is a black and white photograph capturing a quiet moment. An older person with short, light-colored hair and glasses sits at a wooden table, looking down towards a large pizza in front of them. The individual is wearing a dark, collared shirt or jacket.\n\nOn the table, a large, round pizza, possibly cheese or pepperoni, is served on a wire basket lined with parchment paper. It has been cut into eight slices. To the left of the pizza is a glass containing a dark beverage with a straw. A second glass with a straw is visible in the foreground on the right.\n\nThe setting appears to be a restaurant or cafe with a large window. Through the window, a busy outdoor plaza or street scene is visible, filled with numerous people walking around. Some individuals are carrying umbrellas, suggesting it might be raining or has recently rained. The contrast between the solitary figure at the table and the bustling activity outside creates a contemplative mood. The top of the image shows the edge of a dark awning or canopy.\n

Hallucination: incorrectly describes people outside as carrying umbrellas.

AIMS (Ours)

This is a black-and-white photograph with a selective color effect, where most of the image is in grayscale except for a small, vibrant red object.\n\nThe main subject is an older man with short, light-colored hair and glasses, sitting at a table. He is wearing a dark jacket or shirt and is looking down and to his right, seemingly focused on something out of frame. His expression appears contemplative or neutral.\n\nIn the foreground, on a wooden table, sits a large, round pizza on a cardboard base. The pizza appears to be a classic style with a golden-brown crust and a generous topping of melted cheese and possibly other toppings. A glass of dark liquid, possibly a soda or iced tea, sits next to the pizza.\n\nThe background shows a large window or glass wall overlooking a public space, possibly a plaza or courtyard. Several people can be seen walking around in the distance, suggesting an urban or commercial setting. The lighting appears to be natural daylight coming from outside, illuminating the scene from the right side.\n\nThe overall composition places the man slightly off-center, with the pizza prominently featured in the lower right corner, creating a sense of intimacy and everyday life captured in a moment of quiet reflection.

Correct: correctly describes the people in the background without assigning unsupported attributes or actions.

Figure 8: Case study 2 on the CHAIR benchmark using Qwen3.5-9B with greedy decoding.



Instruction: Please describe the image in detail.

Baseline

The image depicts a small, cozy room that appears to be an attic or loft space. The room is well-lit by natural light streaming through a window on the right side of the image. The window has a wooden frame and is covered with a beige curtain that allows some sunlight to filter into the room.\n\nIn the center of the room, there is a desk with a computer setup. The desk is black and has a simple design. On the desk, there is a desktop computer monitor, a keyboard, and a mouse. Next to the monitor, there is a small white object that looks like a speaker or possibly a decorative item. The desk also has a few other items, including what appears to be a stack of books or magazines and a small plant in a pot.\n\nTo the left of the desk, there is a rocking chair with a dark wooden frame and a cushioned seat. The chair is positioned against the wall and seems to be made for comfort and relaxation.\n\nOn the right side of the desk, there is a bookshelf filled with various books. The bookshelf is wooden and has multiple shelves, each holding different books. Some books are stacked vertically, while others are placed horizontally. The top shelf contains a few decorative items, including a small potted plant and a framed picture or artwork.\n\nThe floor of the room is carpeted in a light color, which adds to the warm and inviting atmosphere of the space. The walls are painted in a light color, likely white or a very light shade, which helps to brighten up the room further.\n\nOverall, the room appears to be a comfortable and functional workspace, combining elements of both work and relaxation. The presence of the rocking chair suggests that this space might also serve as a reading or contemplation area. The natural light from the window enhances the room's ambiance, making it feel more open and airy.

Hallucination: incorrectly claims that there is a mouse on the desk, likely driven by the common keyboard–mouse co-occurrence prior.

AIMS (Ours)

The image depicts a cozy, well-organized attic or loft space. The room is characterized by its high ceiling and ample natural light streaming through a large window on the right side of the image. The window allows for a view of trees outside, suggesting that the room is located in a suburban area.\n\n**Room Layout and Furniture:**\n1. **Desk Area:** In the center-left of the image, there is a small desk with a computer monitor placed on it. The desk appears to be made of wood and has a few items on it, including what looks like a keyboard and possibly some papers or a notebook.\n2. **Office Chair:** Positioned directly in front of the desk is an office chair with a blue cushion. This chair seems to be ergonomically designed, providing comfort and support for someone working at the desk.\n3. **Bookshelf:** To the right of the desk, there is a tall bookshelf filled with various books. The books appear to be arranged neatly, indicating that this space is used for reading and studying.\n4. **Side Table:** Next to the bookshelf, there is a small side table with a potted plant on it. The plant adds a touch of greenery to the room, creating a more inviting atmosphere.\n5. **Rocker Chair:** On the left side of the image, there is a rocking chair with a dark wooden frame. The chair is positioned near the desk, offering an additional seating option in the room.\n6. **Additional Furniture:** There is another piece of furniture visible in the image, which appears to be a small table or stand with a plant on it. This adds to the room's decor and provides extra storage or display space.\n\n**Lighting and Ambiance:**\nThe room is well-lit, primarily due to the natural light coming from the large window. The lighting creates a warm and inviting atmosphere, making the space feel cozy and comfortable.\n\n**Decorations and Personal Items:**\n1. **Wall Decorations:** On the wall above the desk, there is a framed picture or artwork, adding a personal touch to the room.\n2. **Personal Items:** Various personal items and decorations are scattered throughout the room, contributing to its lived-in and personalized feel.\n\n**Conclusion:**\nThis attic or loft space is well-organized and features a variety of furniture and decor items. The room is equipped with a desk, office chair, bookshelf, rocking chair, and side table, all of which contribute to a comfortable and inviting living or working environment.

Correct: Remains grounded in the visual evidence without introducing the co-occurring but absent object.

Figure 9: Case study 1 on the CHAIR benchmark using Qwen2.5-VL-3B with greedy decoding.



Instruction: Please describe the image in detail.

Baseline

The image depicts the interior of a bakery or café, showcasing a well-organized display case filled with an assortment of baked goods and pastries. The display case is illuminated by bright lights, making the items inside clearly visible to customers. \n\n**Objects Present in the Image:**\n\n1. **Bakery Display Case:**\n - The display case is made of glass and has multiple shelves.\n - The top shelf contains various types of pastries, including what appears to be croissants, muffins, and possibly some cakes or tarts.\n - The middle shelf holds more pastries, such as cookies, donuts, and other sweet treats.\n - The bottom shelf displays additional baked goods like bread rolls, bagels, and possibly some sandwiches or pastries.\n\n2. **Menu Boards:**\n - There are several chalkboard menus on the wall behind the display case.\n - The leftmost menu board lists \"Weekend Brunch\" offerings, which include items like \"Baked Eggs,\" \"Breakfast Burrito,\" and \"Chocolate Muffin.\"\n - The central menu board features \"Breakfast\" options, such as \"Baked Eggs,\" \"Breakfast Burrito,\" and \"Chocolate Muffin.\"\n - The rightmost menu board provides \"Soup of the Day\" choices, including \"Chicken Noodle Soup,\" \"Beef Stew,\" and \"Vegetable Soup.\"\n\n3. **Decorative Elements:**\n - A large, decorative pumpkin is placed prominently in front of the display case, adding a festive touch to the setting.\n - The pumpkin is painted with the word \"Cookie\" written on it, suggesting that the bakery specializes in cookies.\n\n4. **Lighting:**\n - The lighting in the bakery is bright, ensuring that all items are well-lit and easily visible to customers.\n\n5. **Additional Items:**\n - There are several small containers of condiments or toppings placed near the display case, likely for customers to add to their pastries.\n - A small sign indicating \"LARGE\" is visible, possibly referring to the size of certain items available for purchase.\n\n6. **Wall Decorations:**\n - The walls are adorned with colorful chalkboards displaying the menu items, contributing to a vibrant and inviting atmosphere.\n\n7. **Other Details:**\n - The bakery seems to have a casual and welcoming ambiance, with a focus on fresh and delicious baked goods.

Hallucination: incorrectly identifies several paper cups as containers for condiments or toppings.

AIMS (Ours)

The image depicts a section of a bakery or café, showcasing various items and signs. The main focus is on a display case filled with an assortment of baked goods. The display case has multiple shelves, each containing different types of pastries and breads.\n\nOn the top shelf, there are several types of pastries neatly arranged. These include:\n- Round pastries with a golden-brown crust.\n- Square pastries with a similar golden-brown crust.\n- Some pastries appear to be filled or have a different texture.\n\nOn the middle shelf, there are more pastries, including:\n- A few pastries that look like they might be filled with cream or some other filling.\n- Some pastries that appear to be topped with fruit or nuts.\n\nOn the bottom shelf, there are more pastries, including:\n- Some pastries that look like they might be filled with cream or some other filling.\n- Some pastries that appear to be topped with fruit or nuts.\n\nIn front of the display case, there is a counter with various items displayed. On the left side of the counter, there is a box labeled \"Bouquet\" which appears to contain some kind of gift or product. Next to it, there is a small sign that reads \"Cookie's.\"\n\nBehind the counter, there is a chalkboard menu. The menu is divided into sections, each with its own color-coded writing. The top part of the menu lists various items such as \"Breakfast,\" \"Lunch,\" \"Dinner,\" and \"Specials.\" Each section has a list of items along with their prices. For example, under \"Breakfast,\" there are items listed such as \"Eggs,\" \"Muffins,\" \"Bagels,\" and others.\n\nTo the right of the menu, there is another chalkboard sign. This sign appears to be advertising a special offer or promotion, although the specific details are not clearly visible in the image.\n\nAbove the counter, there is a large window or glass panel, allowing natural light to illuminate the display case and counter area. The lighting inside the bakery is bright, making the pastries and other items easily visible.\n\nOverall, the image captures a well-stocked bakery or café with a variety of baked goods on display, along with a detailed menu board. The setting suggests a welcoming and inviting atmosphere for customers to enjoy their treats.

Correct: avoids assigning unsupported functions to the paper cups.

Figure 10: Case study 2 on the CHAIR benchmark using Qwen2.5-VL-3B with greedy decoding.



Instruction: Describe this image.

Baseline

The image depicts a large blue bus parked in what appears to be an outdoor parking area, possibly near a park or a public transportation hub. The bus is predominantly blue with white and yellow accents. The side of the bus features a design that includes a cartoon-like depiction of a person's face, which seems to be a stylized representation of a famous person, possibly a sports figure given the context of the design elements.\n\nThe bus has multiple windows along its length, and the windows are tinted. The front of the bus displays the text \"Bus Parivizala\" on the lower part of the front door, indicating it might be a public transportation bus serving a specific route or service. The bus also has several advertisements and logos on its body, including **a logo that resembles a ship**, suggesting it might be associated with maritime services or tourism.\n\nIn the background, there are trees and greenery, indicating that the location is likely in a suburban or urban park area. The sky is clear, suggesting good weather conditions at the time the photo was taken. There is no visible traffic or other vehicles around the bus, making it the primary focus of the image.\n\nGiven the design and context, the bus could be used for tourism or public transportation within a coastal or maritime region. The presence of the cartoon-like face and **the ship logo** suggests that the bus might be promoting travel or tourism services related to the sea.\n\nTo summarize:\n- **Bus Type**: Large blue bus.\n- **Design**: Cartoon-like face, **ship logo**, advertisements.\n- **Text**: \"Bus Parivizala\".\n- **Location**: Outdoor parking area, possibly near a park or public transportation hub.\n- **Background**: Trees and greenery, clear sky.\n- **Purpose**: Likely tourism or public transportation related to maritime services.\n\nThis detailed description should provide a comprehensive understanding of the image, allowing a pure text model to answer any related questions effectively.

Hallucination: incorrectly describes a ship logo and infers a maritime/sea-related context.

AIMS (Ours)

The image depicts a large blue bus parked on what appears to be a paved surface, possibly a parking lot or a street. The bus is predominantly blue with white accents and features various graphics and text on its side. \n\n### Description of the Bus:\n1. **Color and Design**: The bus is primarily blue with white accents. The design includes a series of **yellow stars** scattered across the side of the bus.\n2. **Graphics and Text**: There are several graphics and texts on the side of the bus. Notably, there is a large graphic of a person's face in the center, which appears to be a caricature or stylized portrait. Surrounding this central figure are numerous **yellow stars**, adding a dynamic and colorful element to the bus's appearance.\n3. **Text and Branding**: Below the central graphic, there is text that reads \"Bus Parivizala,\" indicating the name or brand of the bus. Additionally, there are other smaller texts and logos, but they are not clearly legible from the image provided.\n\n### Environment:\n1. **Surroundings**: The bus is parked in an area with trees in the background, suggesting it might be in a park, a residential area, or a public space.\n2. **Background Elements**: There are no visible people or vehicles around the bus, giving the impression that the bus is either parked temporarily or waiting for passengers.\n\n### Additional Observations:\n1. **Windows and Doors**: The bus has multiple windows along its length, allowing passengers to see outside. The front door of the bus is visible, and it appears to be a standard entry door.\n2. **License Plate**: Although not clearly visible, the bus likely has a license plate at the back, which is common for buses.\n\n### Analysis:\n- **Purpose**: Given the design and the text \"Bus Parivizala,\" it can be inferred that this bus is used for public transportation, likely serving routes within a city or town.\n- **Functionality**: The presence of multiple windows suggests that the bus is designed to accommodate a significant number of passengers, making it suitable for urban commuting.\n- **Design Elements**: The use of blue and yellow colors, along with **the star graphics**, could be indicative of a specific branding or a particular route associated with the bus company.\n\n### Conclusion:\nThis image shows a blue bus with white accents and **star graphics**, parked in an area with trees in the background. The bus is branded as \"Bus Parivizala\" and features a central graphic of a person's face

Correct: accurately describes the visible elements (e.g. yellow stars) without hallucinated content.

Figure 11: Case study on the AMBER-G benchmark using Qwen2.5-VL-3B with greedy decoding.



Question: Is there a yellow bird in the image? Please answer yes or no.

GT	Baseline	AIMS (Ours)
No	Yes	No

AIMS: Correctly identifies the color of objects.



Question: Is it a good time to drive a car through the road in the picture? Please answer yes or no.

GT	Baseline	AIMS (Ours)
No	Yes	No

AIMS: Correctly interprets the traffic signal.



Question: Is this a picture of Former Tainan Assembly Hall? Please answer yes or no.

GT	Baseline	AIMS (Ours)
No	Yes	No

AIMS: Correctly recognizes the landmark.



Question: Is this movie directed by jonathan liebesman? Please answer yes or no.

GT	Baseline	AIMS (Ours)
Yes	No	Yes

AIMS: Correctly identifies the movie director.



Question: Is this artwork created by Antoniazzo Romano? Please answer yes or no.

GT	Baseline	AIMS (Ours)
Yes	No	Yes

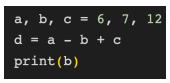
AIMS: Correctly verifies the artwork's attribution.



Question: Is the actor inside the red bounding box named Cher? Please answer yes or no.

GT	Baseline	AIMS (Ours)
Yes	No	Yes

AIMS: Correctly verifies the actor's name.



Question: The image shows a python code. Is the output of the code '11'? Please answer yes or no.

GT	Baseline	AIMS (Ours)
Yes	No	Yes

AIMS: Correctly performs code reasoning.



Question: Is there only one piece of pizza in this image? Please answer yes or no.

GT	Baseline	AIMS (Ours)
No	Yes	No

AIMS: Correctly counts the pizza pieces.



Question: Is there a backpack in this image? Please answer yes or no.

GT	Baseline	AIMS (Ours)
Yes	No	Yes

AIMS: Correctly recognizes the backpack.



Question: Does this image describe a place of pantry? Please answer yes or no.

GT	Baseline	AIMS (Ours)
Yes	No	Yes

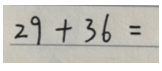
AIMS: Correctly recognizes the scene.



Question: Is it appropriate to translate the Chinese in the image into English 'waiting for a long time' in the picture? Please answer yes or no.

GT	Baseline	AIMS (Ours)
Yes	No	Yes

AIMS: Correctly translates the text.



Question: Is the answer to the arithmetic question in the image 56? Please answer yes or no.

GT	Baseline	AIMS (Ours)
No	Yes	No

AIMS: Correctly solves the arithmetic problem.



Question: Is the word in the logo "excharge hotel"? Please answer yes or no.

GT	Baseline	AIMS (Ours)
No	Yes	No

AIMS: Correctly recognizes the logo text.



Question: Is there a sofa in the middle of potted plants in the image? Please answer yes or no.

GT	Baseline	AIMS (Ours)
Yes	No	Yes

AIMS: Correctly recognizes the spatial relationship.

Figure 12: Qualitative examples on MME using Qwen3.5-9B with nucleus sampling.