Mitigating Visual Hallucinations for Reliable Multimodal Agents

Sohyeon Kim [1] **Sang Yeon Yoon** [2] **Kyeongbo Kong** [1]

Abstract

Large Vision-Language Models (LVLMs) are increasingly used as perception modules in multimodal agents, yet object hallucination can propagate to false tool use and unsafe downstream decisions. We identify a consistent three-phase attention structure in LVLM vision encoders—diffusion, focus, and rediffusion—and show that hallucination is most sensitive to low-attention tokens in the focus phase. We propose a lightweight training-free intervention that suppresses such tokens using single-pass statistics and DPP-based selection. Across multiple LVLM backbones, our method reduces hallucination with negligible overhead and improves reliability in an object-triggered tool-calling setup.

1. Introduction

Large Vision-Language Models (LVLMs) (Caffagni et al., 2024) are increasingly deployed as perception modules in multimodal agents operating in open-ended, real-world environments (Xie et al., 2024; Durante et al., 2024). In such agent pipelines, visual understanding is not an isolated task; perceptual outputs are directly consumed by downstream reasoning modules to determine tool use, action selection, and safety-related decisions. Consequently, object hallucination (Rawte et al., 2023; Liu et al., 2024b) can propagate beyond a local captioning error and become an agent-level reliability failure. For example, a hallucinated object may cause an agent to invoke unnecessary tools, skip relevant verification steps, or make decisions based on visual evidence that is not actually present (Lin et al., 2025). Mitigating hallucination at the perception stage is therefore a critical step toward safer and more reliable multimodal agents in the wild.

Existing hallucination mitigation methods (Bai et al., 2024) have mainly focused on decoder-stage interventions, such as adjusting token probabilities (Chuang et al., 2024; Huang et al., 2024; Leng et al., 2024; An et al., 2025) or reweighting language-model attention during generation (Jiang et al., 2025b; Liu et al., 2024c). However, these methods intervene after visual representations have already reached the language model, limiting their ability to prevent perception errors from propagating through agent pipelines. More recently, vision-encoder-based approaches suppress unreliable visual signals earlier, but representative methods such as Adversarial Uncertainty Estimation (AUE) (Seo et al., 2025) rely on iterative optimization and incur substantial inference overhead, limiting their practicality in interactive agent settings.

In this work, we revisit hallucination mitigation as a perception-stage reliability problem by analyzing attention dynamics inside the vision encoder. Through layer-wise analysis across multiple LVLM backbones, we observe a consistent three-phase structure of visual processing: diffusion, focus, and rediffusion. We further find that hallucination is sensitive to low-attention tokens during the focus phase, whereas analogous interventions in other phases have much weaker effects. These findings suggest that focus-phase suppression provides an efficient and targeted mechanism for improving visual grounding before unreliable evidence propagates to downstream agent decisions.

Motivated by this observation, we propose a lightweight training-free intervention that selectively suppresses low-attention tokens during the focus phase using statistics from a single forward pass and Determinantal Point Process (DPP)-based token selection. We evaluate the method on four LVLM backbones—LLaVA-1.5-7B, LLaVA-1.5-13B (Liu et al., 2024a), Shikra-7B (Chen et al., 2023), and Qwen-2.5-VL (Bai et al., 2025)—using CHAIR (Rohrbach et al., 2018) and POPE (Li et al., 2023). In addition, to better connect hallucination mitigation with agent reliability, we introduce an object-triggered tool-calling setup to measure how perception errors affect downstream tool invocation behavior. Across these evaluations, our method consistently reduces hallucination while introducing negligible overhead, suggesting that improving perception quality can be an effective step toward safer multimodal agents.

Our contributions are as follows:

[1]Pusan National University, Busan, Republic of Korea [2]Pukyong National University, Busan, Republic of Korea. Correspondence to: Kyeongbo Kong <kbkong@pusan.ac.kr>.

Published at the Second Workshop on Agents in the Wild: Safety, Security, and Beyond (AIWILD) at ICML 2026. Copyright 2026 by the author(s).

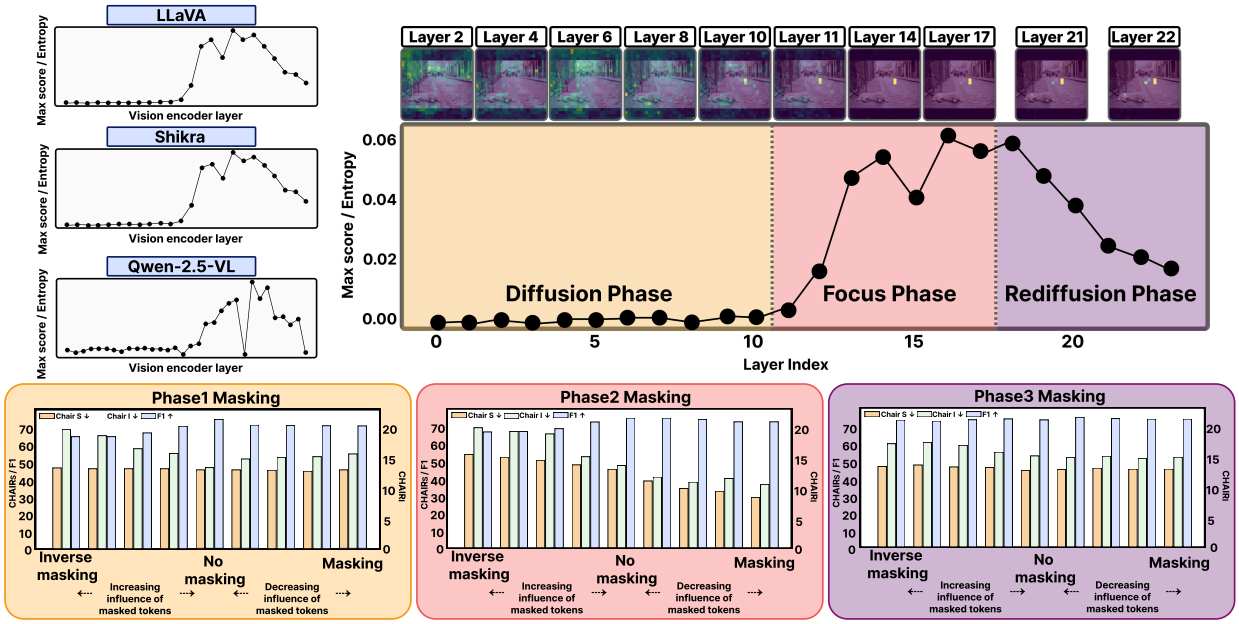


Figure 1. **Layer-wise attention dynamics in LVLM vision encoders and phase-specific hallucination behavior.** Top: the maximum-attention-to-entropy ratio reveals a consistent three-phase structure—diffusion, focus, and rediffusion—across LLaVA, Shikra, and Qwen-2.5-VL. Bottom: modulating low-attention tokens produces the strongest hallucination changes in the focus phase.

- We identify a consistent three-phase structure in LVLM vision-encoder attention and show that hallucination is most sensitive to low-attention tokens in the focus phase.
- We propose a lightweight training-free focus-phase suppression method using single-pass statistics and DPP-based token selection.
- We validate the method across multiple LVLM backbones, showing reduced hallucination with negligible overhead and improved reliability in an agent-style tool-triggering setup.

2. Analysis of Attention Dynamics and Hallucination

2.1. Attention Dynamics in Vision Encoders

To understand how visual information is processed inside LVLMs, we analyze the layer-wise evolution of attention distributions in the vision encoder. We quantify attention concentration using the ratio between the maximum attention score and attention entropy. As shown in Figure 1, this metric reveals a consistent three-phase pattern across multiple LVLM backbones, including LLaVA, Shikra, and Qwen-2.5-VL. In the early layers, attention remains broadly distributed, corresponding to a diffusion phase in which the encoder references a wide range of visual tokens. In the intermediate layers, attention becomes sharply concentrated on a small subset of tokens, forming a focus phase. In the later layers, attention becomes diffuse again, corresponding to a rediffusion phase. The detailed boundary determination procedure is provided in Appendix A. This consistent progression across architectures suggests that the three-phase structure captures a common visual processing pattern in LVLM vision encoders, providing a natural basis for phase-aware perception-stage intervention.

2.2. Phase-Specific Hallucination Analysis

We next examine how the three visual processing phases relate to hallucination behavior. For each intervention phase, we estimate token importance using attention statistics from the immediately preceding layers and construct a low-attention token group via DPP (Macchi, 1975)-based selection, which also accounts for token similarity. We then modulate the contribution of this group under several controlled settings, including masking, inverse masking, and gradual increase or decrease relative to the no-masking baseline. Figure 1 summarizes the intervention setup and the resulting hallucination trends.

The results reveal a clear phase-specific pattern. In the diffusion and rediffusion phases, modulating low-attention tokens produces only minor changes in hallucination metrics. In contrast, focus-phase interventions produce substantially stronger and more directional effects: suppressing low-attention tokens consistently reduces hallucination, whereas amplifying their contribution increases it. This indicates that low-attention tokens during the focus phase are especially associated with hallucination-inducing visual signals. These findings identify the focus phase as an effective intervention point for perception-stage hallucination mitigation. Since hallucinated visual evidence can propagate to downstream

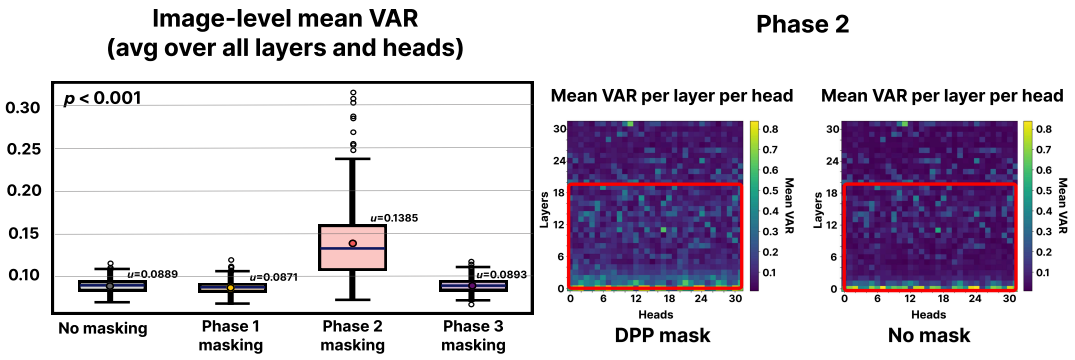


Figure 2. **Visual Attention Ratio (VAR) under different masking conditions.** Focus-phase masking is associated with higher image-level mean VAR and increased visual referencing in intermediate language-model layers relative to the no-masking baseline.

reasoning and tool use, suppressing unreliable tokens at this stage provides a targeted mechanism for improving reliable agent perception.

2.3. Visual Attention Ratio Analysis

To better understand how focus-phase suppression affects downstream generation, we analyze the Visual Attention Ratio (VAR) (Jiang et al., 2025b), which measures how strongly generated tokens attend to visual inputs during decoding. A higher VAR indicates greater reliance on visual evidence, whereas a lower VAR suggests stronger dependence on language priors. In Figure 2, focus-phase masking significantly increases VAR compared to the no-masking baseline, while interventions in the diffusion and rediffusion phases produce only marginal changes. Layer-wise analysis further shows that this increase is most pronounced in intermediate layers of the language model. Together with the phase-specific hallucination results, these observations suggest that suppressing low-attention tokens during the focus phase encourages the decoder to reference visual evidence more strongly during generation. Additional VAR results are provided in Appendix C.This plausibly explains the reduced hallucination behavior and is especially relevant in agent pipelines, where visually grounded outputs determine downstream tool invocation.

3. Method

Based on the analysis above, we design a lightweight phase-aware masking method that intervenes only at the focus phase of the vision encoder. The method requires no training or parameter updates, and aims to suppress unreliable visual signals before they propagate to the language model and downstream agent decisions.

DPP-based token selection. For each input, we estimate token importance using mean attention statistics from the diffusion-to-focus transition layers and model token redundancy using cosine similarity between visual token embeddings. These two signals are combined through a DPP (Macchi, 1975), which selects informative and diverse visual tokens while filtering redundant signals.

Focus-phase masking. The DPP-selected tokens are retained, while the remaining tokens are suppressed through hard masking in the focus-phase layers. Because the masking decision is constructed from statistics obtained in a single forward pass, the method avoids iterative optimization required by prior uncertainty-based approaches, making it practical for interactive agent settings. Detailed DPP masking method is provided in Appendix D.

4. Experiments

We evaluate the proposed method from two complementary perspectives. The first is a quantitative evaluation on standard hallucination benchmarks, and the second is an agent-oriented downstream evaluation. The first set of experiments examines whether the proposed focus-phase masking consistently reduces hallucination across multiple LVLM backbones. The second investigates whether such reductions translate into more reliable downstream behavior when model outputs trigger agent tools. Through these evaluations, we aim to assess not only caption-level hallucination mitigation but also its practical impact on agent reliability.

4.1. Main Quantitative Results on Hallucination Benchmarks

We first report the main quantitative results on four LVLM backbones: LLaVA-1.5-7B, LLaVA-1.5-13B, Shikra-7B, and Qwen-2.5-VL. We evaluate hallucination using CHAIR (Rohrbach et al., 2018) and POPE (Li et al., 2023). CHAIR measures hallucination at the sentence level (CHAIR-S) and instance level (CHAIR-I), while POPE evaluates object-existence judgments on the Random, Popular, and Adversarial splits. Table 1 summarizes the results. Across all backbones, the proposed method consistently reduces CHAIR metrics relative to the baseline

Benchmark		LLaVA-1.5-7B			LLaVA-1.5-13B			Shikra-7B			Qwen-2.5-VL		
		Orig.	AUE	Ours	Orig.	AUE	Ours	Orig.	AUE	Ours	Orig.	AUE	Ours
	Time (s)	3.207	7.307	3.182	6.387	14.992	5.861	3.267	7.677	3.267	2.598	33.462	2.813
CHAIR	CHAIR-S↓	45.0	30.2	**28.8** (-1.4)	41.0	29.8	**29.0** (-0.8)	52.4	47.2	**44.6** (-2.6)	31.4	27.8	**27.8** (0.0)
	CHAIR-I↓	13.3	10.4	**10.2** (-0.2)	11.8	**9.8**	10.2 (+0.4)	16.1	14.5	**14.1** (-0.4)	7.6	7.4	**7.3** (-0.1)
	F1↑	74.2	71.8	**72.0** (+0.2)	75.3	71.4	**72.5** (+1.1)	71.5	72.2	70.7 (-1.5)	75.5	75.8	74.7 (-1.1)
POPE	Random↑	87.4	85.6	**86.4** (+0.8)	86.6	82.4	**83.5** (+1.1)	81.4	81.0	**81.2** (+0.2)	81.3	79.6	**80.3** (+0.7)
	Popular↑	84.3	83.0	**83.1** (+0.1)	83.9	**82.4**	82.3 (-0.1)	80.7	**79.8**	79.5 (-0.3)	80.8	79.2	**79.9** (+0.7)
	Adversarial↑	79.3	78.4	**78.5** (+0.1)	**79.5**	76.5	77.3 (+0.8)	**77.1**	76.1	76.4 (+0.3)	**80.5**	78.7	79.6 (+0.9)

Table 1. **Main quantitative results under greedy decoding on CHAIR and POPE.** Numbers in parentheses in the **Ours** columns denote the AUE-relative delta $\Delta = \text{Ours} - \text{AUE}$. For CHAIR-S / CHAIR-I (lower is better), negative Δ indicates improvement and is highlighted in blue. For F1 and POPE (higher is better), positive Δ indicates improvement and is highlighted in blue.

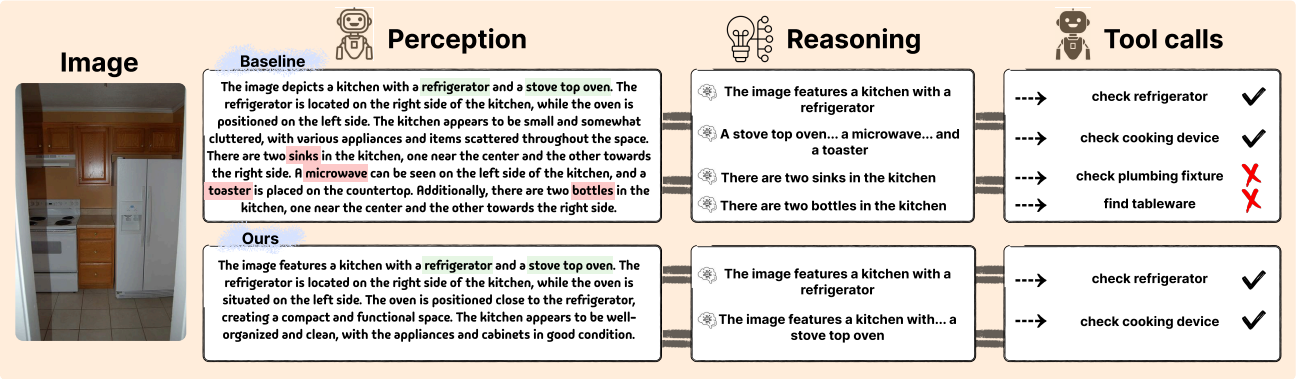


Figure 3. **Agent-oriented tool-call evaluation.** Hallucinated objects in perception outputs can propagate into false downstream tool calls, whereas the proposed masking method reduces such errors by improving perception-stage grounding.

Method	FTCR ↓	False Calls	Total Calls
Baseline	21.47%	275	1281
Ours	**15.19%**	**158**	**1040**

Table 2. **Results of the agent-oriented tool-call evaluation.**

while maintaining competitive POPE performance. Despite avoiding iterative optimization, our method achieves results comparable to or better than AUE in several settings, while remaining close to the original runtime. In contrast, AUE incurs much larger overhead, especially on Qwen-2.5-VL. Additional results in Appendix F.1 show compatibility with existing inference-time mitigation methods, including OPERA (Huang et al., 2024), VCD (Leng et al., 2024), PAI (Liu et al., 2024c), and Devils (Jiang et al., 2025a). Overall, the proposed strategy is both effective and computationally practical across diverse LVLM architectures.

4.2. Agent-Oriented Tool-Call Evaluation

Standard hallucination metrics quantify perception errors, but they do not directly measure whether such errors affect downstream agent behavior. To evaluate this connection, we introduce an agent-oriented tool-call evaluation. As illustrated in Figure 3, the evaluation follows a three-stage pipeline. First, a VLM generates a visual description of the image. Second, a text-only reasoning model reads the description and selects relevant analysis tools. Finally, the selected tools are compared against ground-truth object annotations. This setup captures a common agent failure mode: hallucinated object mentions can be converted into unnecessary tool invocations.

We evaluate on 500 images from MS-COCO val2014 (Lin et al., 2014). To isolate perception quality, we compare baseline and DPP-masked captions under the same reasoning model and evaluation protocol. COCO classes are mapped to 30 semantic tools, and a tool call is false if its object group is absent from the ground-truth annotations. We report the False Tool Call Rate (FTCR), i.e., the fraction of false calls among all tool calls, where lower is better.

As shown in Table 2, our method improves downstream tool invocation reliability. On LLaVA-1.5-7B, the baseline produces 1,281 total tool calls, including 275 false calls, resulting in an FTCR of 21.47%. In contrast, our method produces 1,040 total tool calls with 158 false calls, reducing FTCR to 15.19%. Since the reasoning module and tool-calling pipeline are unchanged, these results indicate that improving perception quality can reduce hallucination-induced agent errors. This supports the role of lightweight vision-encoder intervention as a practical reliability mechanism for multimodal agents.

Acknowledgements

This work was supported by the Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (MOTIE) (RS-2024-00406796, HRD Program for Industrial Innovation), the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2024-00456152), and the "Advanced GPU Utilization Support Program" funded by the Government of the Republic of Korea (Ministry of Science and ICT). The authors also gratefully acknowledge the Cluster Server for Computational Science at Pusan National University for providing computational resources.

References

An, W., Tian, F., Leng, S., Nie, J., Lin, H., Wang, Q., Chen, P., Zhang, X., and Lu, S. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29915–29926, 2025.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., and Lin, J. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., and Shou, M. Z. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

Caffagni, D., Cocchi, F., Barsellotti, L., Moratelli, N., Sarto, S., Baraldi, L., Cornia, M., and Cucchiara, R. The revolution of multimodal large language models: A survey. *Findings of the association for computational linguistics: ACL 2024*, pp. 13590–13618, 2024.

Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., and Zhao, R. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

Chuang, Y.-S., Xie, Y., Luo, H., Kim, Y., Glass, J. R., and He, P. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Durante, Z., Huang, Q., Wake, N., Gong, R., Park, J. S., Sarkar, B., Taori, R., Noda, Y., Terzopoulos, D., Choi, Y., et al. Agent ai: Surveying the horizons of multimodal interaction. *arXiv preprint arXiv:2401.03568*, 2024.

Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., Lin, D., Zhang, W., and Yu, N. 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.

Jiang, Z., Chen, J., Zhu, B., Luo, T., Shen, Y., and Yang, X. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25004–25014, 2025a.

Jiang, Z., Chen, J., Zhu, B., Luo, T., Shen, Y., and Yang, X. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25004–25014, 2025b.

Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., and Bing, L. 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.

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, X., and Wen, J.-R. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305. Association for Computational Linguistics, 2023.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Lin, X., Ning, Y., Zhang, J., Dong, Y., Liu, Y., Wu, Y., Qi, X., Sun, N., Shang, Y., Wang, K., et al. Llm-based agents suffer from hallucinations: A survey of taxonomy, methods, and directions. *arXiv preprint arXiv:2509.18970*, 2025.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., and Peng, W. A survey on hallucination in large vision-language models, 2024b.

Liu, S., Zheng, K., and Chen, W. 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.

Macchi, O. The coincidence approach to stochastic point processes. *Advances in Applied Probability*, 7(1):83–122, 1975.

Rawte, V., Sheth, A., and Das, A. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*, 2023.

Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., and Saenko, K. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, 2018.

Seo, H., Kang, D. U., Cho, H., Lee, J., and Chun, S. Y. On epistemic uncertainty of visual tokens for object hallucinations in large vision-language models. *Advances in neural information processing systems*, 2025.

Xie, J., Chen, Z., Zhang, R., Wan, X., and Li, G. Large multimodal agents: A survey. *arXiv preprint arXiv:2402.15116*, 2024.

Organization of the Appendix

This Appendix provides additional technical details, extended analyses, and qualitative results that complement the main manuscript. In particular, it includes further explanations of the proposed phase-aware token suppression framework, additional empirical evidence supporting the observations in the main paper, and implementation details related to token selection and efficiency.

The document is organized as follows:

- **Appendix A** describes the formulation and empirical procedure used to determine the phase boundaries within vision encoders.
- **Appendix B** provides additional analysis of continuous token modulation, showing how hallucination behavior changes when the influence of low-attention tokens is adjusted across phases.
- **Appendix C** extends the Visual Attention Ratio (VAR) analysis to multiple LLM backbones, examining the consistency of the observed trends across architectures.
- **Appendix D** presents the formulation of the Determinantal Point Process (DPP) kernel and the efficient greedy MAP inference procedure used in our method.
- **Appendix E** describes the qualitative analysis methodology based on ground-truth captions and presents sentence-level examples showing how phase-specific masking affects hallucinated descriptions.
- **Appendix F** reports extended quantitative results on CHAIR and POPE, including compatibility with existing inference-time mitigation methods, and additional evaluation on AMBER.
- **Appendix G** provides extended qualitative examples on CHAIR and POPE and compares the spatial masking patterns of our DPP-based method with AUE.

A. Phase Boundary Determination

In Section 2.1 of the main manuscript, we empirically observed a consistent three-phase visual processing structure (*Diffusion, Focus, and Rediffusion*) across multiple LVLM vision encoders. To clarify how the focus phase boundaries used in our experiments are identified, we describe the procedure for determining the interval $\mathcal{F} = [l_{\text{start}}, l_{\text{end}}]$ based on the dynamics of the attention concentration ratio.

The concentration ratio at layer l is defined as

$$R^{(l)} = \frac{M^{(l)}}{H^{(l)}}, \tag{1}$$

where $M^{(l)}$ denotes the maximum attention score and $H^{(l)}$ denotes the attention entropy at layer l. This ratio increases when attention becomes concentrated on a small subset of tokens and decreases when attention is more evenly distributed.

A.1. Onset of the Focus Phase (l_{start})

During the initial Diffusion Phase, attention is broadly distributed and the layer-wise change of the concentration metric remains relatively small. To identify the transition into the Focus Phase, we examine the discrete forward difference

$$\Delta R^{(l)} = R^{(l)} - R^{(l-1)}. \tag{2}$$

The start layer l_{start} is determined as the first layer where the increase in concentration exceeds the baseline variation observed in the early layers:

$$l_{\text{start}} = \min \left\{ l \mid \Delta R^{(l)} > \mu_{\text{base}} + \lambda \cdot \sigma_{\text{base}} \right\}, \tag{3}$$

where μ_{base} and σ_{base} denote the mean and standard deviation of $\Delta R^{(l)}$ computed from the early diffusion layers (e.g., the first 25% of the network depth), and λ is a confidence multiplier. This criterion identifies the layer where attention concentration begins to increase beyond the relatively stable baseline behavior of the diffusion stage.

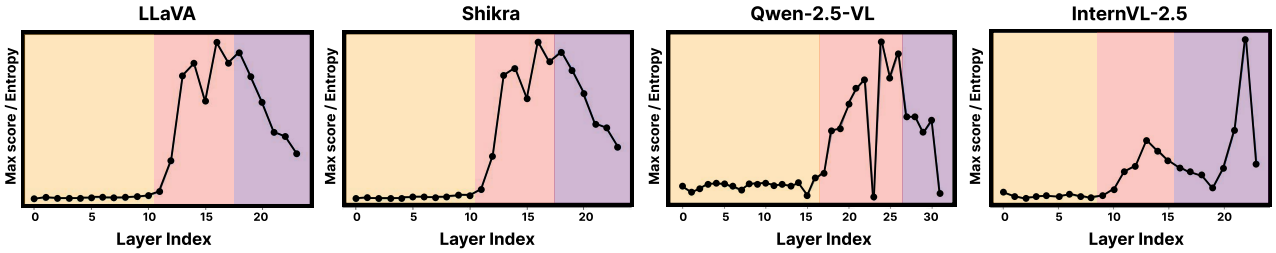


Figure 4. **Layer-wise attention dynamics across evaluated LVLM backbones.** The solid lines indicate the ratio of maximum attention score to entropy ($R^{(l)}$). The shaded regions represent the identified Diffusion (yellow), Focus (red), and Rediffusion (purple) phases based on our gradient-based boundary formulation.

A.2. Offset of the Focus Phase (l_{end})

Empirically, the Focus Phase corresponds to the main region where the concentration ratio remains high after the initial transition. Across the evaluated LVLM architectures (with depths ranging from 24 to 32 layers), this stage typically occupies a contiguous block covering roughly $30\% \sim 40\%$ of the total network depth L.

Rather than enforcing a fixed length, we determine the window length K relative to the shape of the concentration curve following l_{start}. Models exhibiting a sharp concentration peak tend to require a shorter window (approximately 30% of L), whereas models with broader concentration plateaus require slightly longer windows (closer to 40% of L) to capture the main high-concentration region.

The end layer is therefore determined as

$$l_{\text{end}} = l_{\text{start}} + K - 1. \tag{4}$$

This procedure captures the primary region of concentrated attention before the concentration metric begins to decrease again during the Rediffusion Phase.

A.3. Phase Boundaries Across Evaluated Models

Applying the above procedure provides a consistent rationale for the layer intervals used in the main experiments without requiring model-specific hyperparameter tuning, as visualized for each model in Figure 4.

LLaVA-1.5 & Shikra ($L = 24$). The gradient $\Delta R^{(l)}$ remains close to zero in early layers and exhibits a clear increase around layer 11. Setting $l_{\text{start}} = 11$ and using a window length of $K = 7$ ($\approx 29\%$ of L) captures the main concentration region before the metric declines, yielding Layers 11–17.

Qwen-2.5-VL ($L = 32$). The concentration gradient exceeds the early baseline around layer 17 and continues to increase. Using $K = 10$ ($\approx 31\%$ of L) captures the dominant concentration peaks around layers 24 and 26 before the metric decreases near layer 27, resulting in Layers 17–26.

InternVL-2.5 ($L = 24$). The first consistent increase in $\Delta R^{(l)}$ occurs around layer 9. Using the proportional window $K = 7$ ($\approx 29\%$ of L) isolates the primary concentration region, yielding Layers 9–15. Although a secondary spike occasionally appears in the final layers (e.g., in InternVL-2.5), this peak typically corresponds to late-stage representation aggregation rather than the primary attention concentration stage. Therefore, our formulation focuses on the first dominant concentration peak, which consistently marks the transition into the focus phase.

B. In-Depth Analysis of Token Modulation in the Focus Phase

To further examine the observations reported in Section 2.2 regarding the sensitivity of hallucination metrics to token influence within the focus phase, we conducted continuous token modulation experiments. Instead of applying binary masks alone, we systematically modulated the influence of low-attention tokens by directly shifting their pre-softmax attention logits.

As illustrated in Figure 5, increasing the logit values (i.e., boosting the influence) of these low-attention tokens tends to amplify hallucination behavior. When their influence is strengthened, the models increasingly generate descriptions

containing objects that are not present in the image (highlighted in red), and in some cases the generated captions deviate from the actual visual content. Conversely, subtracting from these logits (i.e., suppressing their influence) progressively reduces the occurrence of hallucinated entities, producing descriptions that are more consistent with the visual evidence.

Figure 6 further presents additional qualitative examples demonstrating the same trend across multiple scenes. Across diverse images, strengthening the influence of low-attention tokens consistently introduces hallucinated objects or relationships, whereas suppressing their influence leads to more visually grounded descriptions. This consistent qualitative pattern supports the observation that hallucination behavior is closely related to the influence of low-attention tokens during the focus phase.

Taken together, these results provide empirical evidence that tokens receiving relatively low attention in the focus phase can introduce noisy signals into the generation process when their influence is amplified. The observed monotonic response further indicates that hallucination dynamics are closely tied to token influence within this phase.

C. Extended Visual Attention Ratio (VAR) Analysis

In Section 2.3, we showed that suppressing low-attention tokens during the focus phase increases the Visual Attention Ratio (VAR) (Jiang et al., 2025b), indicating that the language model places greater emphasis on visual inputs during text generation. To further examine whether this behavior is consistent across different language decoder configurations, we extend the VAR analysis to two additional LVLMs: *LLaVA-1.5-13B* and *Shikra-7B*. These models represent two distinct decoder characteristics. First, *LLaVA-1.5-13B* employs a larger 13B parameter language model, allowing us to examine whether the observed effect remains consistent under increased language model capacity. Second, *Shikra-7B* is optimized for referential dialogue and spatial grounding, where the language model is trained to interpret spatial coordinates through natural language expressions. This provides a complementary setting to evaluate whether the intervention remains effective when the decoder is explicitly trained for spatial reasoning.

As illustrated in Figures 7 and 8, both models exhibit a consistent trend. Applying DPP masking during the focus phase (Phase 2) produces a noticeable increase in the image-level mean VAR compared to the no-masking baseline. In contrast, applying the same intervention during the diffusion phase (Phase 1) or the rediffusion phase (Phase 3) results in only minor changes in VAR. The heatmaps further illustrate how this shift occurs within the language model. When focus-phase masking is applied, intermediate layers allocate higher attention weights to visual tokens across multiple heads, indicating that the decoder relies more strongly on visual evidence during generation.

Taken together, these results suggest that suppressing low-attention tokens during the focus phase systematically increases the model's reliance on visual inputs across different decoder architectures and model scales.

D. Detailed Formulation and Efficiency of Phase-Aware DPP Masking

In Section 3 of the main manuscript, we introduced a training-free token suppression strategy based on a Determinantal Point Process (DPP) (Macchi, 1975) applied during the focus phase. This appendix provides the detailed formulation and implementation used to construct the DPP mask and apply it within the vision encoder. The procedure involves three functional components of the encoder layers: the *source layers*, the *feature layer*, and the *target layers*.

D.1. Extraction of Token Importance and Similarity

To determine which visual tokens should be retained, we estimate both the importance of individual tokens and the similarity among them. These quantities are extracted from reference layers immediately preceding the focus phase.

Source Layers (Token Importance). We first estimate the structural importance of each token using attention statistics from the source layers. Specifically, we extract the self-attention matrices from the multi-head attention modules of these layers and average them across heads and layers to obtain

$$\bar{\mathbf{A}} \in \mathbb{R}^{N \times N}, \tag{5}$$

where N denotes the sequence length. The importance score of token i is then computed as the total attention it receives:

$$q_i = \sum_{j=1}^{N} \bar{\mathbf{A}}_{j,i}. \tag{6}$$

Table 3. Inference time per sample (seconds). Values in parentheses indicate the change relative to the origin baseline. Our DPP-based masking introduces only a small overhead, while AUE significantly increases latency.

	Origin		AUE (Seo et al., 2025)		Ours	
Model	Enc. Time	Total	Enc. Time	Total	Enc. Time	Total
LLaVA-1.5	0.008	3.207	4.588 (+4.580)	7.307 (+4.100)	0.198 (+0.190)	3.182 (-0.025)
Shikra	0.006	3.267	4.586 (+4.580)	7.677 (+4.410)	0.188 (+0.182)	3.267 (+0.000)
Qwen-2.5-VL	0.074	2.598	31.244 (+31.170)	33.456 (+30.858)	0.286 (+0.212)	2.813 (+0.215)
InternVL2.5	0.047	2.299	38.768 (+38.721)	41.291 (+38.992)	0.259 (+0.212)	2.722 (+0.423)

Feature Layer (Token Similarity). To measure semantic redundancy between tokens, we extract the visual feature embeddings

$$\mathbf{F} \in \mathbb{R}^{N \times C} \tag{7}$$

from a designated feature layer preceding the focus phase. Each token feature vector is ℓ_2-normalized,

$$\tilde{\mathbf{f}}_i = \frac{\mathbf{f}_i}{\|\mathbf{f}_i\|_2}, \tag{8}$$

and the cosine similarity matrix is computed as

$$\mathbf{S}_{i,j} = \tilde{\mathbf{f}}_i^\top \tilde{\mathbf{f}}_j, \tag{9}$$

yielding $\mathbf{S} \in \mathbb{R}^{N \times N}$.

D.2. DPP Kernel Construction

Using both token importance and token similarity, we construct a positive semi-definite L-ensemble kernel $\mathbf{L} \in \mathbb{R}^{N \times N}$, whose elements are defined as

$$\mathbf{L}_{i,j} = q_i \cdot \mathbf{S}_{i,j} \cdot q_j. \tag{10}$$

Under this formulation, the diagonal entries $\mathbf{L}_{i,i} = q_i^2$ reflect the absolute importance of individual tokens, while the off-diagonal terms penalize the joint selection of semantically redundant tokens.

D.3. Fast Greedy MAP Inference and Efficiency

The goal is to select a subset of K tokens that maximizes the DPP posterior probability (MAP inference). Since exact MAP inference for a DPP is NP-hard, we adopt an efficient greedy approximation algorithm. Starting from an empty set, the algorithm iteratively selects the token that yields the largest marginal gain in the determinant of the corresponding sub-kernel matrix. This process is accelerated using Cholesky factorization updates, which allow efficient incremental updates without recomputing determinants from scratch. Importantly, this selection process is executed using statistics obtained from a single forward pass of the vision encoder. Therefore, the mask generation does not require iterative gradient computations. The resulting inference overhead is summarized in Table 3. Compared with the baseline models, our method introduces only a small increase in vision encoder processing time, while maintaining nearly identical total inference latency. In contrast, the adversarial uncertainty estimation method (AUE) requires iterative gradient-based optimization and therefore introduces substantially higher latency across all evaluated architectures.

D.4. Mask Generation and Target Layer Application

After selecting the subset of K tokens, we construct an additive attention mask $\mathbf{M}$ over the token sequence. To preserve global image semantics, the class token (`[CLS]`) is always retained. Selected tokens and the `[CLS]` token are assigned a mask value of 0, while unselected tokens receive $-\infty$. The mask $\mathbf{M}$ is applied to the self-attention modules of the target layers corresponding to the focus phase. During the forward pass, the mask is added to the pre-softmax attention logits, forcing the attention weights of the masked tokens to become zero after the softmax operation. This phase-aware masking

suppresses low-attention tokens during the focus phase while allowing the remaining tokens to dominate the attention distribution used for downstream generation.

E. Qualitative Analysis Methodology using Ground-Truth Captions

To complement the phase-specific hallucination analysis in the main text, we conduct a qualitative analysis using ground-truth captions to examine how hallucinations appear at the sentence level under different phase-masking conditions. This section details the procedure for constructing and utilizing Ground-Truth (GT) captions for this analysis, the results of which are visualized in Fig. 9 and Fig. 10.

To construct reference descriptions for comparison, we first generated an initial set of eight descriptive sentences for randomly sampled images from the COCO dataset using a large language model. These descriptions were subsequently reviewed and corrected by human annotators to remove inaccurate object mentions or relationships, resulting in GT captions that are consistent with the visual content of the images.

Subsequently, we examined the text generated by the LLaVA-1.5-7B model across four distinct configurations: the baseline (No masking), Phase 1 masking (Diffusion), Phase 2 masking (Focus), and Phase 3 masking (Rediffusion). For each configuration, the model was prompted to generate eight descriptive sentences per image. We then conducted a sentence-by-sentence comparison against the GT captions to qualitatively observe how phase-specific interventions influence the introduction of hallucinated statements.

Please note that while the models generated eight sentences to ensure comprehensive visual coverage, Fig. 9 and Fig. 10 display a representative subset of approximately five sentences per setting. This subset is shown to facilitate clear side-by-side visual comparison of the generated descriptions across different intervention phases while maintaining visual clarity.

F. Extended Results

F.1. Extended Quantitative Results on CHAIR and POPE Datasets

Table 4 reports extended CHAIR and POPE results when the proposed focus-phase masking is combined with existing inference-time hallucination mitigation methods, including OPERA, VCD, PAI, and Devils. These methods mainly intervene at the decoding or language-model stage, whereas our approach suppresses unreliable visual signals earlier in the vision encoder. Therefore, the two types of interventions are largely complementary. Across multiple LVLM backbones and decoding strategies, our method generally maintains or further improves hallucination-related metrics compared with the corresponding AUE variants, while avoiding the iterative optimization cost of AUE. The gains are not uniform across all metrics and settings, which is expected because each decoding method changes the generation distribution differently. Nevertheless, the overall results indicate that focus-phase visual suppression can be integrated with existing inference-time mitigation techniques and can serve as a lightweight perception-stage component for improving hallucination robustness.

F.2. Evaluation on Additional Hallucination Benchmarks

To further validate the generalization of our proposed method beyond the CHAIR and POPE benchmarks, we additionally evaluated our approach on the AMBER dataset. AMBER offers a comprehensive evaluation suite comprising both generative and discriminative tasks, evaluating three types of hallucination: existence, attribute, and relation.

For the generative task, AMBER employs four metrics: CHAIR↓ measures the proportion of hallucinated objects, Cover↑ assesses the coverage of ground-truth objects, Hal↓ captures the fraction of responses containing hallucinations, and Cog↓ quantifies cognitively plausible hallucinations.

For the discriminative task, AMBER assesses whether the model correctly rejects hallucinatory prompts by answering "Yes" or "No" to targeted queries regarding object existence, attributes, and relations. Performance is measured using standard binary classification metrics—Accuracy↑, Precision↑, Recall↑, and F1-score↑—where Precision and Recall treat hallucinatory questions with a ground-truth answer of "No" as the positive class.

Table 5 reports the evaluation results on the AMBER benchmark for the LLaVA-1.5-7B and Shikra-7B architectures. Across both models, the proposed phase-aware DPP masking consistently reduces hallucination-related generation metrics, such as Hal↓ and Cog↓. At the same time, the method maintains or slightly improves the discriminative metrics, including Acc↑, Recall↑, and F1↑. These results indicate that suppressing low-attention tokens during the focus phase can mitigate

Table 4. **Results on CHAIR and POPE benchmarks.** Comparison of Origin (baseline), AUE, and Ours across decoding strategies. Numbers in parentheses in the *ours* columns denote the AUE-relative delta Δ = ours − AUE. For $\text{CHAIR}_S/\text{CHAIR}_I$ (lower is better), negative Δ indicates improvement and is highlighted in blue. For F1 and POPE (higher is better), positive Δ indicates improvement and is highlighted in blue.

			Greedy			OPERA			VCD			PAI			Devils		
Model	**Bench**	**Metric**	**Orig.**	**AUE**	**Ours (Δ)**	**Orig.**	**AUE**	**Ours (Δ)**	**Orig.**	**AUE**	**Ours (Δ)**	**Orig.**	**AUE**	**Ours (Δ)**	**Orig.**	**AUE**	**Ours (Δ)**
LLaVA-1.5-7b	CHAIR	$\text{CHAIR}_S \downarrow$	45.0	30.2	28.8(-1.4)	44.2	29.8	31.6(+1.8)	46.2	34.2	35.6(+1.4)	32.4	19.4	17.4(-2.0)	23.2	13.0	12.8(-0.2)
		$\text{CHAIR}_I \downarrow$	13.3	10.4	10.2(-0.2)	13.5	11.2	11.2(0.0)	14.9	11.8	12.5(+0.7)	9.6	7.6	6.5(-1.1)	7.9	7.2	6.3(-0.9)
	POPE	F1 ↑	74.2	71.8	72.0(+0.2)	74.4	72.3	71.4(-0.9)	71.3	70.5	70.4(-0.1)	74.6	67.7	68.0(+0.3)	72.0	63.5	65.1(+1.6)
		ran. ↑	87.4	85.6	86.4(+0.8)	86.3	84.5	85.2(+0.7)	83.7	78.9	80.6(+1.7)	87.4	85.8	86.2(+0.4)	87.7	84.5	85.4(+0.9)
		pop. ↑	84.3	83.0	83.1(+0.1)	83.3	81.9	81.5(-0.4)	80.9	77.1	77.5(+0.4)	84.5	83.2	83.0(-0.2)	85.3	83.0	83.2(+0.2)
		adv. ↑	79.3	78.4	78.5(+0.1)	79.3	77.9	77.5(-0.4)	77.3	73.3	74.7(+1.4)	79.7	78.8	78.5(-0.3)	79.9	78.7	78.9(+0.2)
LLaVA-1.5-13b	CHAIR	$\text{CHAIR}_S \downarrow$	41.0	29.8	29.0(-0.8)	37.8	27.8	26.4(-1.4)	48.8	32.2	35.4(+3.2)	33.8	21.0	24.0(+3.0)	25.0	17.0	18.8(+1.8)
		$\text{CHAIR}_I \downarrow$	11.8	9.8	10.2(+0.4)	11.5	10.7	9.6(-1.1)	13.5	11.3	11.7(+0.4)	10.7	7.58	7.8(+0.22)	7.2	7.4	6.9(-0.5)
	POPE	F1 ↑	75.3	71.4	72.5(+1.1)	75.0	71.5	73.1(+1.6)	73.1	71.0	70.7(-0.3)	75.2	70.8	70.8(0.0)	73.9	66.3	69.0(+2.7)
		ran. ↑	86.6	82.4	83.5(+1.1)	86.5	78.1	80.4(+2.3)	81.8	75.3	77.7(+2.4)	87.2	82.9	83.8(+0.9)	87.9	82.0	85.5(+3.5)
		pop. ↑	83.9	82.4	82.3(-0.1)	83.4	79.7	79.8(+0.1)	80.0	76.7	76.9(+0.2)	84.1	82.8	82.5(-0.3)	85.8	81.0	83.3(+2.3)
		adv. ↑	79.5	76.5	77.3(+0.8)	79.4	75.1	76.0(+0.9)	77.2	73.1	74.0(+0.9)	79.8	76.8	77.6(+0.8)	81.5	76.1	78.3(+2.2)
Shikra-7b	CHAIR	$\text{CHAIR}_S \downarrow$	52.4	47.2	44.6(-2.6)	35.2	28.8	31.6(+2.8)	54.6	48.8	48.2(-0.6)	34.0	22.6	23.2(+0.6)	27.8	20.8	21.8(+1.0)
		$\text{CHAIR}_I \downarrow$	16.1	14.5	14.1(-0.4)	12.8	10.7	12.6(+1.9)	16.8	16.2	15.8(-0.4)	9.9	7.6	7.2(-0.4)	10.7	10.5	10.3(-0.2)
	POPE	F1 ↑	71.5	72.2	70.7(-1.5)	70.0	68.8	66.5(-2.3)	70.9	70.7	68.2(-2.5)	72.1	69.7	67.5(-2.2)	70.0	66.7	67.0(+0.3)
		ran. ↑	81.4	81.0	81.2(+0.2)	82.5	81.2	81.8(+0.6)	79.5	78.0	79.5(+1.5)	81.2	79.1	78.6(-0.5)	80.5	78.5	78.4(-0.1)
		pop. ↑	80.7	79.8	79.5(-0.3)	80.5	78.6	79.2(+0.6)	78.2	77.5	75.7(-1.8)	80.7	78.0	77.3(-0.7)	77.9	75.1	75.7(+0.6)
		adv. ↑	77.1	76.1	76.4(+0.3)	77.4	75.8	76.5(+0.7)	76.2	74.6	74.1(-0.5)	77.0	75.0	74.5(-0.5)	75.5	73.3	74.4(+1.1)

Table 5. **Evaluation results on the AMBER dataset.** The best result in each row is highlighted in bold.

		LLaVA-1.5-7B		**Shikra-7B**	
Bench	Metric	Orig.	Ours	Orig.	Ours
Gen.	CHAIR ↓	8.2	**6.2**	12.2	**11.5**
	Cover ↑	**48.5**	46.2	**48.9**	47.5
	Hal ↓	34.9	**24.7**	49.7	**44.8**
	Cog ↓	3.5	**2.4**	5.6	**4.6**
Disc.	Acc. ↑	70.0	**71.1**	**73.5**	73.3
	Precision ↑	**90.0**	89.5	**85.7**	85.1
	Recall ↑	61.7	**63.9**	72.8	**74.2**
	F1 ↑	73.2	**74.6**	78.7	**79.3**

hallucination while preserving the model's ability to correctly identify object existence. Overall, the consistent improvements observed on AMBER suggest that the proposed approach generalizes beyond caption-based evaluations and remains effective under the benchmark's object-existence evaluation setting.

G. Extended Qualitative Results and Spatial Mask Comparisons

G.1. Extended Qualitative Results on CHAIR and POPE Datasets

This subsection provides additional qualitative examples from the CHAIR and POPE benchmarks to further illustrate the impact of the proposed phase-aware token suppression. The evaluation encompasses the five distinct LVLM architectures analyzed in the main text: *LLaVA-1.5-7B, LLaVA-1.5-13B, Shikra-7B, Qwen-2.5-VL*, and *InternVL-2.5*.

The provided examples compare the text generation outputs of the original baseline models, the AUE method, and our proposed approach.Specifically, Figs. 11 – 15 present extended text generation examples evaluated under the CHAIR benchmark. Furthermore, Figs. 16 – 20 display qualitative comparisons based on the POPE benchmark, which evaluates models using targeted object-existence queries (e.g., "Is there a [object] in the image?").

Consistent with the quantitative results reported in the main manuscript, the unmasked baseline models frequently output descriptions containing objects absent from the input image. Applying our phase-aware DPP masking suppresses these hallucinated mentions and yields descriptions more tightly grounded to the actual visual content. The qualitative improvements are consistently observable across the descriptive generation tasks (CHAIR) and the targeted object-existence queries

(POPE) for all five evaluated architectures.

G.2. Spatial Mask Comparisons: Ours vs. AUE

To further examine the behavior of the proposed token selection strategy, Fig. 21 provides visual comparisons between the spatial masking patterns produced by our DPP-based method and the PGD-based AUE method.

To ensure a fair and consistent comparison, the visualizations are constructed under similar masked-token budgets for both methods. The examples reveal distinct spatial characteristics of the two selection strategies. The AUE method, which removes tokens based on an absolute uncertainty threshold, tends to produce continuous and dense removal patterns. In contrast, the proposed DPP method jointly models token importance and semantic similarity. Consequently, it produces a more spatially dispersed masking pattern, filtering out redundant visual features while preserving a diverse set of salient visual cues for multimodal reasoning.

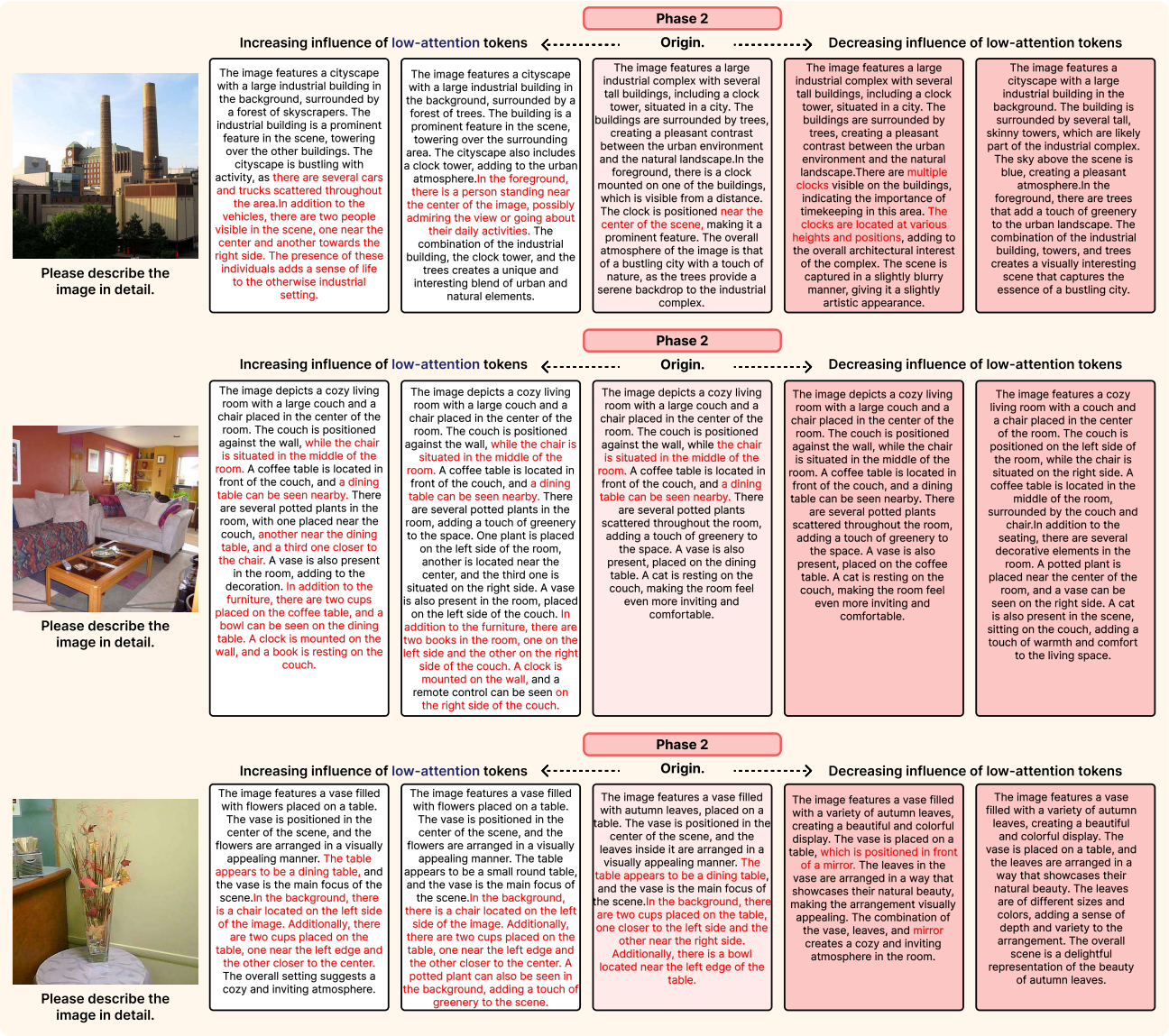


Figure 5. **Qualitative examples of continuous token modulation in the focus phase.** Increasing the influence of low-attention tokens introduces hallucinated objects in the generated descriptions (highlighted in red), while suppressing their influence reduces hallucination and produces captions more consistent with the visual content.

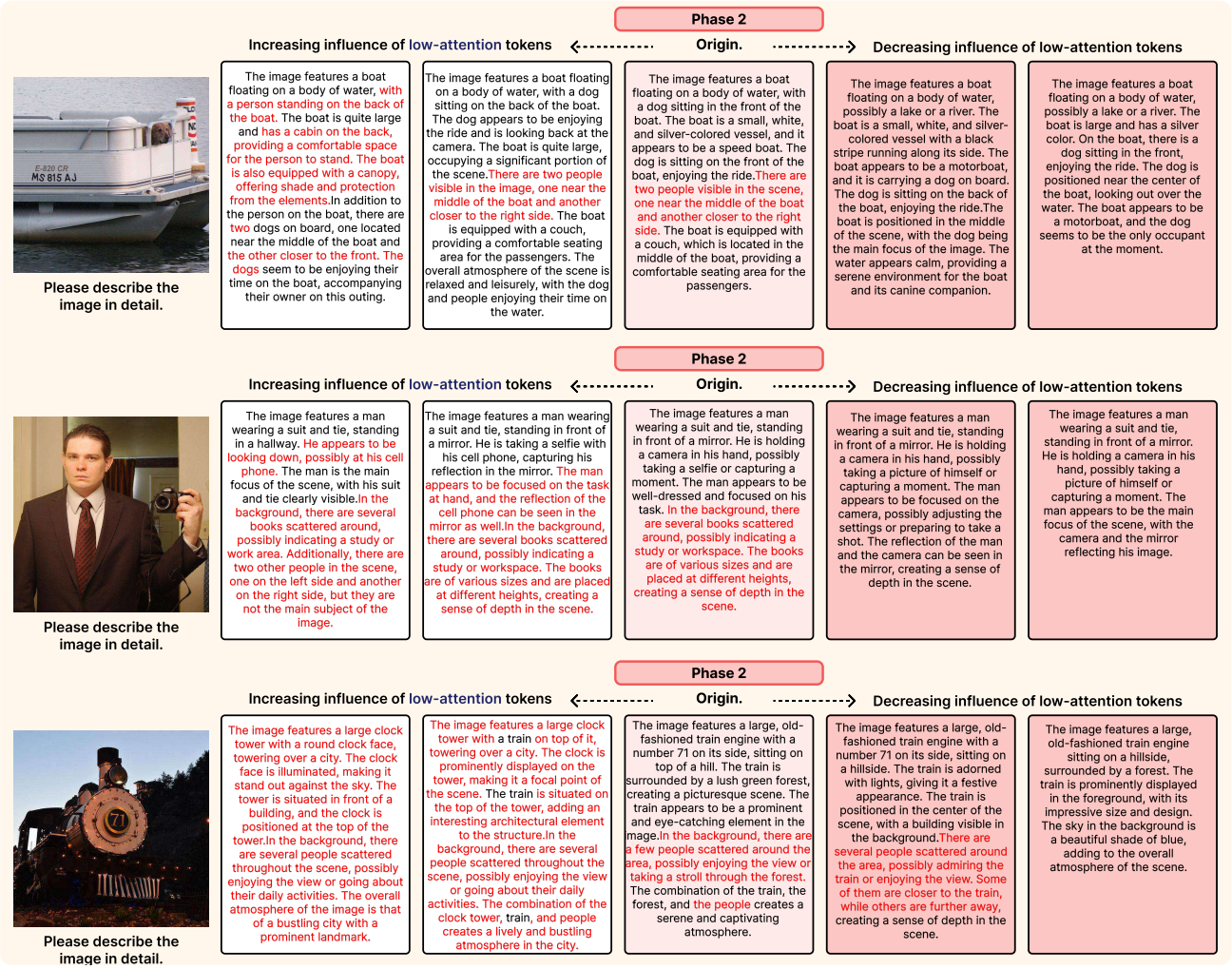


Figure 6. **Additional qualitative examples demonstrating the effect of token influence modulation across diverse scenes.** Strengthening the influence of low-attention tokens tends to introduce hallucinated objects or relationships, whereas suppressing their influence produces descriptions that better align with the visual evidence.

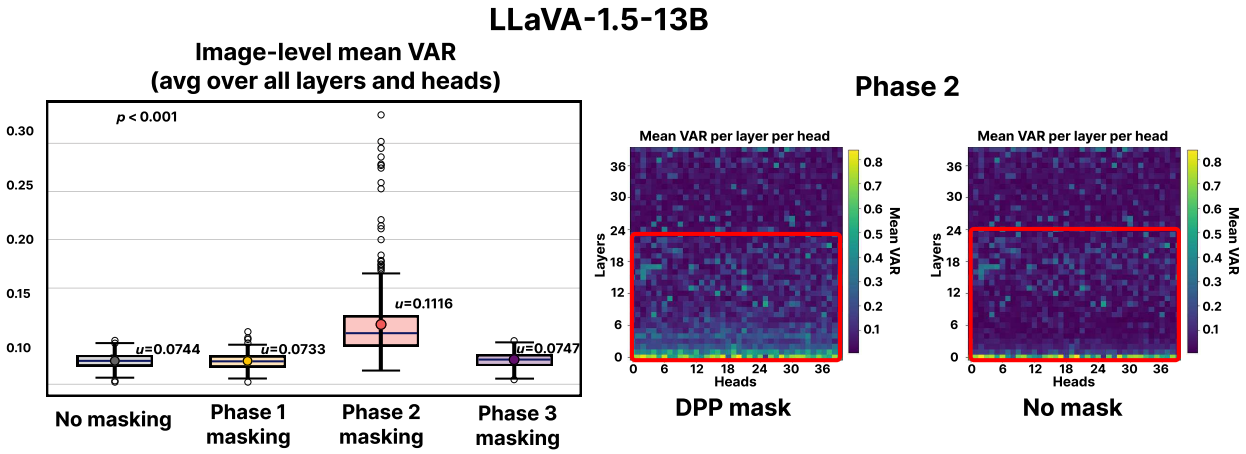


Figure 7. VAR analysis on LLaVA-1.5-13B. **Left:** Distribution of image-level mean VAR across masking strategies. Focus-phase masking (Phase 2) produces a significantly higher VAR than the baseline (no masking), indicating increased reliance on visual tokens during generation. **Right:** Layer-head VAR heatmaps comparing DPP masking and no masking. Applying focus-phase masking increases visual attention across intermediate decoder layers.

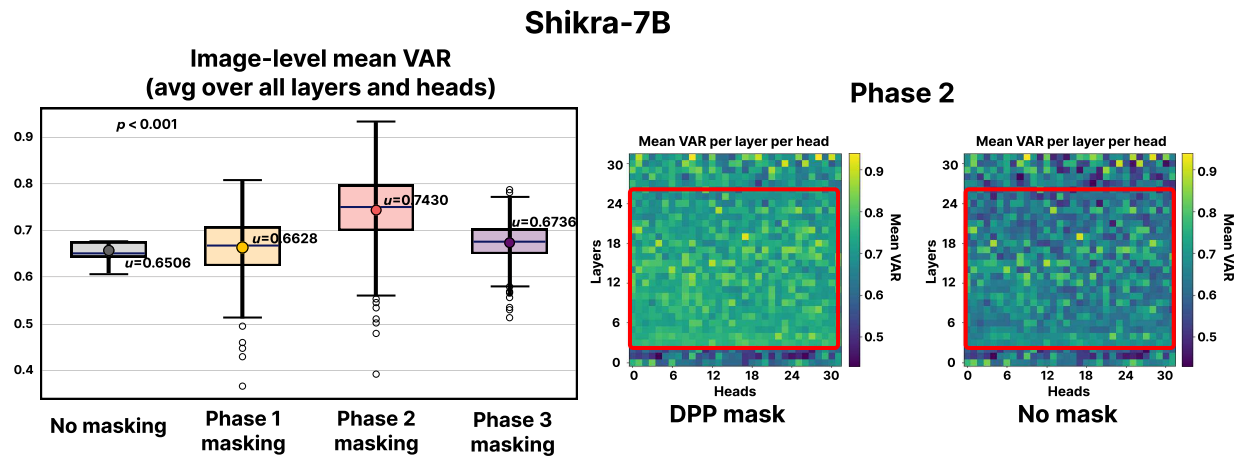


Figure 8. VAR analysis on Shikra-7B. **Left:** Distribution of image-level mean VAR under different masking phases. Focus-phase masking consistently yields higher VAR compared to the baseline. **Right:** Layer-head VAR heatmaps illustrating the increase in visual attention when DPP masking is applied during the focus phase.

Figure 9. **Sentence-level qualitative analysis using Ground-Truth (GT) captions.** Focus-phase masking (Phase 2) consistently mitigates hallucinated statements compared to No masking or interventions in Phase 1 and 3.

GT caption

A plush teddy bear is visible near the center of the image. The teddy bear is beige with a darker nose and eyes. The teddy bear is partially hidden among green ferns and vegetation. A long, angled piece of wood leans upward in front of the teddy bear. There are dense green bushes and trees forming the background.

No masking

A teddy bear is sitting on a pile of sticks. The teddy bear is brown in color. The teddy bear is wearing a brown shirt. The branch is surrounded by leaves. The image is in black and white.

Phase 1 masking

A brown teddy bear is sitting on a pile of branches. The teddy bear is wearing a brown shirt. The bear is surrounded by green leaves. The bear is sitting on a log. The bear is in a forest setting.

Phase 2 masking

A cat is sitting in a toilet bowl. The cat is orange and white. The cat is looking at the camera. The toilet is white. The toilet is open.

Phase 3 masking

A teddy bear is sitting on a pile of sticks. The teddy bear is brown in color. The teddy bear is wearing a brown shirt. The branch is surrounded by leaves. The image is in black and white.

GT caption

Two long boats are floating on a large body of water. The nearer boat contains many people seated closely together. Some of the paddles in the nearer boat are red, while others are darker colors. A second boat with a group of people is visible farther away near the top-center of the image. The rest of the scene is open water with small ripples and no shoreline visible.

No masking

A group of people is riding in a long boat. The boat is filled with people sitting close together. The boat is red and white. The boat is on the water. The people in the boat are enjoying their time together.

Phase 1 masking

A group of people is riding in a boat on the water. The boat is filled with people, and they are all wearing life jackets. The boat is white and red in color. The people in the boat are enjoying their time together. The people in the boat are sitting close to each other.

Phase 2 masking

A group of people are riding in a boat. The boat is filled with people. The boat is on the water. The people are enjoying their time on the boat. The boat is a canoe.

Phase 3 masking

A group of people is riding in a long boat. The boat is filled with people sitting close together. The boat is red and white. The boat is on the water. The people in the boat are enjoying their time together.

GT caption

A man is seated at a table indoors. He is wearing a dark long-sleeve top. He is smiling while looking toward an open laptop. His hands are positioned on the laptop keyboard as if typing. A dark circular sticker with a tree graphic is attached to the laptop lid.

No masking

A group of cows is walking down a road. There are several people standing on the side of the road. The people are watching the cows as they walk by. The cows are brown in color. The people are standing near the cows.

Phase 1 masking

A group of cows is walking down a road. The cows are brown and white. The cows are walking in a line. There are two people in the image, one near the cows and another further away. The people are walking behind the cows.

Phase 2 masking

A group of cows is walking down a road. The cows are brown and white. The cows are walking in front of a house. The cows are walking in a line. The cows are walking on a hill. The people are walking behind the cows.

Phase 3 masking

A group of cows is walking down a road. The cows are walking in front of a car. There are several people standing on the side of the road. The people are watching the cows as they walk by. The image is in black and white.

Figure 10. **Additional qualitative examples using Ground-Truth (GT) captions.** The comparisons highlight that applying the intervention strictly to the focus phase is crucial for ensuring factual alignment without degrading the overall sentence structure.

Please describe the image in detail.

LLaVA-1.5-7B

The image features a red double-decker bus driving down a city street. The bus is quite large and occupies a significant portion of the scene. There are several people walking along the sidewalk, with some of them carrying backpacks. In addition to the bus, there are two cars visible in the scene, one located behind the bus and another further down the street. The presence of pedestrians and vehicles indicates that this is a busy urban area.

LLaVA-1.5-7B+AUE

The image features a red double-decker bus driving down a street, with a large building in the background. The bus is the main focus of the scene, and it appears to be a tour bus, possibly in London. There are several people visible in the image, with one person standing near the bus and others scattered around the scene. The street appears to be busy, with the bus driving down the road and people going about their day.

LLaVA-1.5-7B+Ours

The image features a large red double-decker bus driving down a street. The bus is prominently displayed in the foreground, occupying a significant portion of the scene. There are several people visible in the background, walking or standing along the street. Some of them are closer to the bus, while others are further away. The scene captures the bustling atmosphere of a busy city street.

Please describe the image in detail.

LLaVA-1.5-7B

The image features a bench sitting in a flooded area, surrounded by water. The bench is positioned near a pier, and it appears to be partially submerged in the water. The scene also includes a boat in the background, further emphasizing the flooded environment.There are several people in the scene, with one person standing close to the bench and others scattered around the area. Some of them are closer to the boat, while others are near the water's edge. The presence of people suggests that they might be observing the flooded area or possibly dealing with the aftermath of the flood.

LLaVA-1.5-7B+AUE

The image features a bench situated near a body of water, possibly a lake or a river. The bench is positioned close to the water's edge, providing a relaxing spot for people to sit and enjoy the view. The water appears to be calm and serene, with no visible signs of activity or disturbances. The bench is the main focal point of the scene, and it appears to be the only object in the image.

LLaVA-1.5-7B+Ours

The image features a bench situated near a body of water, possibly a lake or a river. The bench is positioned on the edge of the water, providing a perfect spot for relaxation and enjoying the view. The bench is made of metal and appears to be weathered, adding a sense of history and charm to the scene.In the background, there are several boats visible, some of which are floating on the water. The presence of these boats adds to the overall atmosphere of the scene, making it a picturesque and serene location.

Please describe the image in detail.

LLaVA-1.5-7B

The image depicts a busy subway station with a large crowd of people waiting for the train. Among the crowd, a woman wearing a white headscarf is standing near the edge of the platform, holding a shopping bag. She appears to be waiting for the train to arrive.There are several other people in the scene, some of them carrying handbags and backpacks. The subway station is filled with people of various heights and positions, creating a lively atmosphere. The train is visible in the background, indicating that it is approaching the station.

LLaVA-1.5-7B+AUE

The image depicts a busy subway station with a large number of people waiting for the train. Among the crowd, a woman wearing a headscarf is standing near the edge of the platform, holding a shopping bag. She appears to be waiting for the train to arrive.There are several other people in the scene, some of them carrying handbags and backpacks. The subway station is filled with commuters, creating a bustling atmosphere. The train itself is visible in the background, indicating that it is approaching the station.

LLaVA-1.5-7B+Ours

The image depicts a busy subway station with a large number of people waiting for their train. Among the crowd, a woman wearing a headscarf is standing near the edge of the platform, possibly waiting for her train. She is holding a handbag in her hand.There are several other people in the scene, some of them carrying handbags as well. The subway station is filled with people, creating a bustling atmosphere. The image captures the essence of public transportation and the daily lives of people in a city.

Please describe the image in detail.

LLaVA-1.5-7B

The image features a boat floating on a body of water, with a dog sitting in the front of the boat. The boat is a small, white, and silver-colored vessel, and it appears to be a speed boat. The dog is sitting on the front of the boat, enjoying the ride.There are two people visible in the scene, one near the middle of the boat and another closer to the right side. The boat is equipped with a couch, which is located in the middle of the boat, providing a comfortable seating area for the passengers.

LLaVA-1.5-7B+AUE

The image features a boat floating on a body of water, possibly a lake or a river. The boat is white and has a dog sitting on the front of it, enjoying the ride. The dog is positioned towards the right side of the boat, and it appears to be looking back at the camera. The boat is quite large, covering a significant portion of the water's surface.

LLaVA-1.5-7B+Ours

The image features a boat floating on a body of water, possibly a lake or a river. The boat is large and has a silver color. On the boat, there is a dog sitting in the front, enjoying the ride. The dog is positioned near the center of the boat, looking out over the water. The boat appears to be a motorboat, and the dog seems to be a companion on this journey.

Please describe the image in detail.

LLaVA-1.5-7B

The image features a large blue and yellow train traveling down the tracks. The train is positioned in the middle of the scene, with its front end visible. The train appears to be a passenger train, as it is designed for transporting people.There are several people in the image, with some standing near the train and others further away. They seem to be observing the train as it passes by. The scene captures the essence of a busy train station, with people waiting for their ride or simply enjoying the view.

LLaVA-1.5-7B+AUE

The image features a yellow and blue train car traveling down the tracks. The train is large and occupies a significant portion of the image. The train is moving along the tracks, and it appears to be a passenger train. The train is situated in a rural area, giving it a sense of being in the middle of nowhere. The train is also described as a commuter train, which suggests that it is designed for daily transportation of passengers.

LLaVA-1.5-7B+Ours

The image features a yellow and blue train traveling down the tracks. The train is large and occupies a significant portion of the image. The train is moving swiftly, and its colorful design makes it stand out against the background. The train appears to be a passenger train, possibly a commuter train, as it is designed for public transportation.

Figure 11. **Extended qualitative results on the CHAIR benchmark using LLaVA-1.5-7B.**

Please describe the image in detail.

LLaVA-1.5-13B

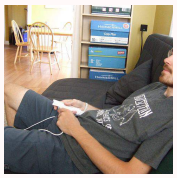
The image features a motorcycle parked on a street, with a sidecar attached to it. The motorcycle is an old-fashioned model, and the sidecar is made of wood. The motorcycle is positioned on the left side of the scene, while the sidecar is located on the right side.There are two people in the scene, one standing near the motorcycle and the other standing further back. The motorcycle and sidecar are the main focus of the image, showcasing their vintage design and unique features.

LLaVA-1.5-13B+AUE

The image features a motorcycle with a black frame and a silver engine. The motorcycle is parked on a street, and the front wheel is visible in the foreground. The motorcycle is old and has a number 60 on the front. The scene is set against a black background, which highlights the motorcycle's details.

LLaVA-1.5-13B+Ours

The image features a motorcycle parked on a street, with a close-up view of the front wheel and the engine. The motorcycle is an old-fashioned model, possibly a vintage or antique, and it appears to be in good condition. The motorcycle is parked next to a bush, adding a touch of greenery to the scene. The overall atmosphere of the image suggests a sense of nostalgia and appreciation for the classic design of the motorcycle.



Please describe the image in detail.

LLaVA-1.5-13B

The image features a man sitting on a couch in a living room, holding a Nintendo Wii remote control in his hand. He appears to be playing a video game, possibly on a Wii console. The couch is positioned in the center of the room, with a dining table located nearby.There are several chairs in the room, with one close to the dining table and another one further away. A few books can be seen scattered around the room, possibly indicating that the man enjoys reading or has been reading before playing the video game.

LLaVA-1.5-13B+AUE

The image features a man lying on a couch, holding a Wii remote in his hand. He appears to be playing a video game, possibly on a Nintendo Wii console. The couch is positioned in the center of the room, with a chair located to the left side of the scene.There are several books scattered around the room, with some placed on the floor and others on a nearby surface. A dining table can be seen in the background, along with a couple of chairs placed around it. The room seems to be a comfortable and cozy space for the man to relax and enjoy his gaming session.

LLaVA-1.5-13B+Ours

The image features a man sitting on a couch, holding a Wii remote in his hand. He appears to be playing a video game, possibly on a Nintendo Wii console. The man is wearing a black shirt and is positioned in the center of the scene.In the background, there is a dining table with chairs placed around it. The table is located on the left side of the image, and the chairs are arranged in a semi-circle formation. The scene suggests a casual and relaxed atmosphere, with the man enjoying his gaming session.

Please describe the image in detail.

LLaVA-1.5-13B

The image features a cityscape with a large industrial building in the foreground, surrounded by trees. The building has two tall smokestacks, which are prominent in the scene. In the background, there are several other buildings, including a clock tower, which adds to the urban atmosphere.The cityscape is bustling with activity, as there are multiple cars and a truck visible in the scene. The cars are scattered throughout the area, with some closer to the foreground and others further in the background. The truck is located near the center of the image, adding to the busy city vibe.

LLaVA-1.5-13B+AUE

The image features a large industrial building with two tall smokestacks in the background. The building is situated in front of a clock tower, which is visible in the middle of the scene. The clock tower appears to be a prominent landmark in the area. The industrial building is surrounded by trees, adding a touch of nature to the scene. The overall atmosphere of the image suggests a blend of urban and natural elements.

LLaVA-1.5-13B+Ours

The image features a cityscape with a large building in the foreground, surrounded by several other buildings. The main building has a clock tower, which stands out among the other structures. In the background, there are two tall smokestacks, likely part of a factory or industrial complex.The scene also includes a few trees scattered throughout the area, adding some greenery to the urban landscape. The overall atmosphere of the image is that of a bustling city with a mix of industrial and architectural elements.

Please describe the image in detail.

LLaVA-1.5-13B

The image features a man sitting on a bench surrounded by a large flock of pigeons. The man appears to be feeding the birds, as they gather around him in various positions. Some pigeons are perched on the bench, while others are standing or flying nearby.In the background, there are a few cars parked, and a handbag can be seen placed close to the bench. The scene captures a moment of interaction between the man and the pigeons, creating a peaceful and lively atmosphere.

LLaVA-1.5-13B+AUE

The image features a man sitting on a park bench surrounded by a large flock of pigeons. The man appears to be feeding the birds, as they gather around him in various positions. Some birds are perched on the bench, while others are standing or flying nearby. The scene captures the interaction between the man and the pigeons, creating a peaceful and lively atmosphere in the park.

LLaVA-1.5-13B+Ours

The image features a man sitting on a bench surrounded by a large flock of pigeons. The man appears to be feeding the birds, as they are all gathered around him. There are at least 13 pigeons in the scene, with some perched on the bench and others flying nearby. The man is seated in the middle of the scene, with the birds surrounding him on all sides. The scene captures a peaceful moment between the man and the birds, as they share a meal together.

Please describe the image in detail.

LLaVA-1.5-13B

The image features a young child standing on a sidewalk, holding a black and white umbrella to protect themselves from the rain. The child is positioned in the center of the scene, with the umbrella covering a significant portion of the image.In the background, there is a car parked on the street, and a truck can be seen further down the road. Additionally, there are two chairs placed near the edge of the scene, possibly belonging to a nearby outdoor seating area.

LLaVA-1.5-13B+AUE

The image features a young child standing on a sidewalk, holding a black umbrella to protect themselves from the rain. The child is positioned in the center of the scene, with the umbrella covering a significant portion of their body.In the background, there is a truck parked on the street, and a bench can be seen further down the sidewalk. The scene captures a typical rainy day, with the child trying to stay dry under their umbrella.

LLaVA-1.5-13B+Ours

The image features a young child standing on a sidewalk, holding a black and white umbrella to protect themselves from the rain. The child is positioned in the center of the scene, with the umbrella covering a significant portion of their body.In the background, there is a truck parked on the street, partially visible on the right side of the image. The scene captures the child's innocence and the simple joy of playing in the rain.

Figure 12. **Extended qualitative results on the CHAIR benchmark using LLaVA-1.5-13B.**

Please describe the image in detail.

Shikra-7B

The image features a tennis court with two tennis players actively engaged in a match. One player is on the ground, possibly taking a swing at a low ball, while the other player is standing next to a tennis racket, possibly preparing for their next move. There are several people in the background, likely watching the game or waiting for their turn to play. Some of them are sitting on a bench, while others are standing around the court. Additionally, there is a backpack placed on the ground near the edge of the court, which might belong to one of the players or spectators.

Shikra-7B+AUE

The image captures a tennis match in progress, with several players and spectators present. A man is leaning over, holding a tennis racket, and preparing to hit the ball. There are multiple people in the stands, watching the match with keen interest.In addition to the players and spectators, there is a tennis ball visible on the court. A backpack can also be seen near the edge of the court, possibly belonging to one of the players. The scene is lively, showcasing the excitement of the tennis match and the dedication of the players and spectators.

Shikra-7B+Ours

The image features a tennis court where a game is taking place. There are multiple people present, including a player holding a tennis racket, likely preparing to hit the sports ball in the air. Another person is seen holding a tennis racket as well, possibly waiting for their turn to play. In addition to the players, there are several spectators watching the game from various positions around the court. A bench is located near the playing area, providing a place for people to sit and enjoy the match. The scene is lively, with everyone engaged in the tennis game or waiting for their turn to play.

Please describe the image in detail.

Shikra-7B

The image depicts a picturesque scene of a herd of sheep grazing in a lush green field. There are at least 13 sheep spread across the field, with some closer to the foreground and others further back towards the edges of the scene. The landscape features a beautiful mountain in the background, adding to the serene atmosphere of the setting.In addition to the sheep, there are a few vehicles visible in the image, including cars and trucks, located near the field, possibly for transportation or farming purposes.

Shikra-7B+AUE

The image features a picturesque scene of a group of white sheep grazing on a grassy hillside. There are at least nine sheep spread across the field, with some closer to the front and others further back. The environment appears to be a blend of a mountain and a green pasture, providing a serene and natural setting for the sheep.

Shikra-7B+Ours

The image features a picturesque scene of a herd of sheep grazing in a lush green field. There are at least 13 sheep spread across the field, with some closer to the foreground and others further back towards the edge of the scene. The landscape is characterized by a gentle slope, and the sheep appear to be enjoying the fresh grass.In the background, there are mountains visible, emphasizing the natural beauty of the landscape. The combination of the vibrant green field and the serene mountainous backdrop creates a tranquil and idyllic atmosphere.

Please describe the image in detail.

Shikra-7B

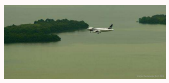
The image features a beach scene with a small metal boat prominently positioned at the edge of the water. The boat appears to be sitting on the shore, partially surrounded by a body of water.There are several other boats scattered throughout the scene, both close to and further away from the main boat. Some of these boats are situated on the left side of the image, while others can be found on the right side.In addition to the boats, there are numerous people in the background enjoying the beach atmosphere. They are dispersed in various groups, likely engaging in different activities or conversations.A bench is also visible in the middle of the scene, providing a place to relax and take in the view.

Shikra-7B+AUE

The image features a sandy beach with a group of boats near the water. There are several small boats, including a metal boat, beached on the shore. The metal boat is positioned closer to the center of the scene, while other boats are scattered around the beach.In addition to the boats, there are a few people enjoying their time on the beach. One person can be seen closer to the left side of the image, another is located further back, and two more people appear on the right side of the scene. The beach and the boats provide a picturesque and relaxing atmosphere.

Shikra-7B+Ours

The image features a beach scene with a small boat sitting on the sand near the water. The boat appears to be surrounded by a body of water, possibly an ocean, and is positioned close to the edge of the shore. There are several people scattered throughout the beach area, enjoying the sun and the water. Some of them are closer to the boat, while others are further away, exploring the beach or engaging in different activities. The overall atmosphere of the scene is relaxed and leisurely, with people enjoying their time by the water.



Please describe the image in detail.

Shikra-7B

The image features a large body of water with a small airplane flying over the water, possibly approaching a landing. The plane appears to be a passenger jet, and it is located towards the right side of the scene. There are also several people in the vicinity of the airplane, possibly enjoying the view or waiting for their flight. Some of these people are located near the center of the image, while others can be found towards the right side.

Shikra-7B+AUE

The image features a large body of water with a small island nearby. There are several people on the island and a plane flying in the sky, seemingly approaching the island. The plane appears to be a large passenger jet, and it is positioned towards the right side of the image. The water surrounding the island and the plane gives the impression of a serene and picturesque setting.

Shikra-7B+Ours

The image features a large body of water with a small island nearby. There is a plane flying in the sky, appearing as a blurry image. The island is located in the middle of the water, surrounded by the lake. The scene creates a sense of serenity and calmness with the presence of the water and the airplane soaring through the sky.

Please describe the image in detail.

Shikra-7B

The image features a man kiteboarding on a large body of water, with a kite attached to his wrists and surfboard in tow. The man is skillfully balancing on the surfboard while being propelled by the kite. There are several other people in the scene, likely enjoying the water activities as well. Some of them can be seen in the background, both near the shore and further out in the water. There is also a boat visible in the scene, situated near the right side of the image.

Shikra-7B+AUE

The image features a man kiteboarding over a large body of water, with a blue sky in the background. The man is skillfully balancing on a board while being propelled by a colorful kite that is flying high above the ocean. There are several other people in the scene, likely enjoying the water activities as well. Some of them are kiteboarding or windsurfing, while others are simply watching the activities from a distance. The image captures a lively and exciting atmosphere with people having a great time on the water.

Shikra-7B+Ours

The image features a man kiteboarding on top of a body of water, with a surfboard attached to his feet. He is skillfully balancing on the surfboard while being pulled by a large kite that is flying in the sky above him. There are several other people in the scene, likely enjoying the water activities as well. Some of them can be seen on the water's surface, either close to the man or further away. The kite is visible in the sky, with its position and size varying across the scene.

Figure 13. **Extended qualitative results on the CHAIR benchmark using Shikra-7B.**

Please describe the image in detail.

Qwen-2.5-VL

The image features a man sitting on a wooden bench in a park, surrounded by a large flock of pigeons. The birds are scattered around the bench, with some sitting on the ground and others flying nearby. The man appears to be enjoying the company of the birds, as he is smiling and looking at them.In the background, there are several cars parked, and a truck can be seen further away. The scene captures a peaceful moment in the park where the man and the birds coexist.

Qwen-2.5-VL+AUE

The image features a man sitting on a bench surrounded by a large flock of pigeons. The man appears to be feeding the birds, as they gather around him, some even sitting on his lap. The birds are scattered all around the bench, with some close to the man and others further away.In the background, there are a few cars parked, and a truck can be seen in the distance. The scene captures a moment of interaction between the man and the pigeons, creating a peaceful and serene atmosphere.

Qwen-2.5-VL+Ours

The image features a man sitting on a wooden bench surrounded by a large flock of pigeons. The man appears to be enjoying the company of the birds, as they gather around him. There are at least 13 pigeons in the scene, with some sitting on the bench and others standing nearby.In addition to the pigeons, there are two birds flying in the sky, one near the top left corner and the other in the top right corner. The scene captures a peaceful moment between the man and the birds, creating a sense of harmony and connection with nature.

Please describe the image in detail.

Qwen-2.5-VL

The image features a young boy wearing a grey shirt and blue jean shorts, standing on a skateboard in a driveway. He appears to be enjoying his time skateboarding, possibly practicing his moves. In the background, there are several potted plants placed around the area, adding a touch of greenery to the scene. A clock can also be seen hanging on the wall, possibly indicating the time for the boy as he skateboards.

Qwen-2.5-VL+AUE

The image features a young boy wearing a grey shirt and blue jeans, standing on a sidewalk and playing with a skateboard. He appears to be enjoying his time outdoors, possibly in a park or a similar setting. The boy is positioned in the center of the scene, with the skateboard placed on the ground in front of him.In the background, there are two potted plants, one on the left side and the other on the right side of the image. Additionally, there are two vases, one located near the center of the scene and the other towards the right side. The scene also includes a clock, which is placed on the left side of the image.

Qwen-2.5-VL+Ours

The image features a young boy wearing a gray shirt and blue jeans, standing on a sidewalk and playing with a skateboard. He appears to be in the middle of a trick, possibly performing a kickflip. The boy is the main focus of the scene, and his skateboard is positioned under his feet.In the background, there are several potted plants placed along the sidewalk, adding a touch of greenery to the scene. The plants are scattered at various distances from the boy, with some closer to the foreground and others further in the background.

Please describe the image in detail.

Qwen-2.5-VL

The image features a baseball sitting on top of a red and orange striped object, which appears to be a baseball glove. The baseball is positioned in the center of the glove, surrounded by the vibrant colors of the striped material. The glove is placed on a surface, possibly a table or a counter, and the baseball seems to be the main focus of the scene.

Qwen-2.5-VL+AUE

The image features a baseball glove with a baseball inside, placed on a surface. The baseball glove is positioned in the center of the scene, and the baseball is nestled within it. The glove appears to be a catcher's mitt, and the baseball is in the process of being caught. The scene captures the essence of a baseball game, with the baseball player ready to make a catch.

Qwen-2.5-VL+Ours

The image features a baseball sitting on top of a baseball glove, which is placed on a surface. The baseball glove is made of leather and is positioned in the center of the scene. The baseball itself is white and appears to be in good condition. The glove and baseball are the main focus of the image, creating a sense of anticipation for a game or practice session.

Please describe the image in detail.

Qwen-2.5-VL

The image features a man standing on a sandy beach, holding a frisbee in his hand. He appears to be preparing to throw the frisbee, possibly engaging in a game of beach frisbee with someone else. The man is wearing a white shirt, and his position suggests that he is focused on the task at hand.In the background, there is a boat visible on the water, adding to the beach atmosphere. The scene captures a moment of leisure and outdoor activity, with the man enjoying the beach and the frisbee game.

Qwen-2.5-VL+AUE

The image captures a man playing with a frisbee on a sandy beach. He is in the middle of a throw, with the frisbee flying through the air. The man is wearing a white shirt and appears to be enjoying the activity.In the background, there is a body of water, possibly the ocean, which adds to the beach atmosphere. The scene is set against a blue sky, creating a pleasant and relaxing environment for the man and his frisbee game.

Qwen-2.5-VL+Ours

The image features a man standing on a beach, wearing a white shirt and black shorts. He is in the process of throwing a frisbee, with his arm raised and the frisbee in the air. The man appears to be enjoying his time outdoors, possibly playing a game of frisbee with friends or family.The beach setting is characterized by the presence of a body of water, which can be seen in the background. The man is the main focus of the scene, with the frisbee and the beach serving as the backdrop.

Please describe the image in detail.

Qwen-2.5-VL

This image captures a snowy urban street scene, likely following a heavy snowfall. The ground and parked cars on both sides of the street are covered with a thick layer of snow, indicating recent snowfall. The street is lined with bare trees, their branches laden with snow, and multi-story residential buildings with brick facades and white window frames.In the foreground, there are two individuals engaged in cross-country skiing. One person, wearing green ski pants, a teal jacket, and a red hat, is carrying a child dressed in winter clothing, including a light-colored hat and a backpack. The child appears to be learning or practicing skiing, as they are holding ski poles and standing on skis. Another individual, dressed in dark clothing, is also skiing nearby, appearing to assist or guide the child.The atmosphere suggests a cold winter day, with people enjoying outdoor activities despite the challenging weather conditions. The scene conveys a sense of community and resilience against the elements.

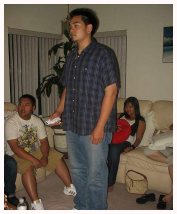
Qwen-2.5-VL+AUE

This image captures a snowy urban street scene, likely after a heavy snowfall. The ground and parked cars on both sides of the street are covered with a thick layer of snow. There are several people in the image, engaged in winter activities.In the foreground, two individuals are skiing on the snow-covered road. One person is wearing green ski pants, a teal jacket, and a red hat, carrying a child dressed in white and brown. The child appears to be learning how to ski, as they are being supported by the adult. Both are using ski poles for balance. Further down the street, another individual is seen skiing alone, wearing a blue outfit with a white hat. They also have ski poles and appear to be enjoying the activity.The background shows a row of residential buildings, some with brick facades and others with light-colored siding. The trees lining the street are bare, indicating it's winter, and their branches are laden with snow. A few other people can be seen further down the street, possibly walking or engaging in similar winter activities. The overall atmosphere is one of a quiet, wintry day in an urban neighborhood.

Qwen-2.5-VL+Ours

This image captures a snowy urban street scene, likely following a heavy snowfall. The ground and the roofs of parked cars are covered with a thick layer of snow, indicating recent snowfall. The street is lined with multi-story brick buildings on both sides, typical of an urban residential area.In the foreground, there are three individuals engaged in cross-country skiing. One person appears to be an adult wearing green ski pants and a red hat, carrying a child dressed in winter clothing. Another adult, dressed in dark clothing, is also skiing, accompanied by a child in blue ski gear. The skiers are using poles for balance and propulsion, and their tracks are visible in the snow.The street itself is partially cleared, with some areas still covered in snow, suggesting that the snowfall has recently occurred or that the area is not heavily trafficked. In the background, other individuals can be seen walking along the street, possibly clearing snow or heading out. The trees lining the street are bare, with snow clinging to their branches, adding to the wintry atmosphere. The overall mood of the image is calm and quiet, with a sense of community activity amidst the snow.

Figure 14. **Extended qualitative results on the CHAIR benchmark using Qwen-2.5-VL.**



Please describe the image in detail.

InternVL-2.5

The image depicts a casual indoor scene where a group of people is gathered in a living room. In the foreground, a man is standing and appears to be playing a video game, holding a white game controller. He is wearing a plaid button-up shirt, blue jeans, and white socks. His expression is focused, and he seems to be concentrating on the game.
Behind him, there are three other individuals seated on a beige couch. To the left, a man is sitting with his legs crossed, wearing a white t-shirt with a graphic design, khaki shorts, and white sneakers. He has a neutral expression and is looking towards the standing man. Next to him, a woman is seated, wearing a black tank top and dark pants. She is holding a red pillow and appears to be watching the game. On the far right, another person is partially visible, wearing a light-colored top and dark pants, with their legs crossed.
The room has a cozy atmosphere, with a large window covered by vertical blinds in the background. There is a potted plant near the window, and a small table with a few items, including a water bottle and a couple of red cups, is visible in the lower left corner. The overall setting suggests a relaxed and social environment, likely a gathering of friends or family enjoying a video game session.

InternVL-2.5+AUE

The image depicts a casual indoor scene where a group of people is gathered in a living room. In the foreground, a man is standing and appears to be playing a video game, holding a white game controller. He is wearing a plaid button-up shirt, blue jeans, and white socks. His expression is focused, and he seems to be concentrating on the game.
Behind him, there are three other individuals seated on a beige couch. To the left, a man is sitting with his legs crossed, wearing a white t-shirt with a graphic design, khaki shorts, and white sneakers. He has a relaxed posture and is looking towards the standing man. Next to him, a woman is seated, wearing a black tank top and dark pants. She is holding a red pillow and appears to be watching the game. On the far right, another person is partially visible, wearing a light-colored top and dark pants.
The room has a cozy atmosphere, with a large window covered by vertical blinds in the background. There is a potted plant near the window, and a small table with a few items, including a water bottle and some red cups, is visible in the lower left corner. The overall setting suggests a casual, social gathering, possibly a game night among friends.

InternVL-2.5+Ours

The image depicts a group of people in a living room setting. A man stands in the center, holding a white Wii remote, suggesting he is playing a video game. He is wearing a plaid button-up shirt, blue jeans, and white socks. His expression is focused, and he appears to be concentrating on the game.
To his left, a man is seated on a beige couch, wearing a white t-shirt with a graphic design, khaki shorts, and white sneakers. He is looking towards the standing man, possibly watching the game.
On the right side of the image, a woman is seated on the same couch, holding a red pillow. She is wearing a black top and blue jeans. Her expression is neutral, and she seems to be observing the scene.
In the background, there is a large window with vertical blinds, and a potted plant is visible near the window. The room is well-lit, and there is a painting on the wall to the right. The overall atmosphere is casual and relaxed.

Please describe the image in detail.

InternVL-2.5

The image shows a black and white cat standing on a bed, reaching into a white cup with colorful designs. The cup is placed on a nightstand next to a lamp with a beige lampshade. The nightstand also holds a box of tissues, a bottle of Dove lotion, and a small container. The bed has a teal bedspread and a white pillow. The room has wooden paneling on the wall, and the overall lighting is warm, creating a cozy atmosphere.

InternVL-2.5+AUE

The image shows a black and white cat standing on a bed, reaching down to a nightstand. The cat is interacting with a white cup that has colorful designs on it. The nightstand is cluttered with various items, including a box of tissues, a candle, a bottle of lotion, and a few other small objects. There is a lamp with a beige shade on the nightstand, providing a warm light. The wall behind the nightstand is paneled with wood, and the bed has a teal-colored bedspread with a white pillow. The overall scene suggests a cozy, domestic setting.

InternVL-2.5+Ours

The image shows a black and white cat standing on a bed, leaning over a nightstand. The cat is drinking from a white cup with colorful designs on it. The nightstand is cluttered with various items, including a box of tissues, a bottle of Dove lotion, a lamp with a beige shade, and a few other small objects. The bed has a teal-colored bedspread, and there is a white pillow behind the cat. The wall behind the nightstand and bed is paneled with light-colored wood. The overall scene is cozy and domestic.

Please describe the image in detail.

InternVL-2.5

The image depicts a tennis match taking place on a clay court. A female tennis player is in the foreground, bending over to pick up a tennis ball. She is wearing a light blue top and white shorts. The court is surrounded by a green fence, and there are several people standing and sitting on the bleachers in the background. The bleachers are part of a larger structure that appears to be a boat or a floating platform, as indicated by the railings and the presence of water in the background. The setting is outdoors, and there are trees visible in the distance. The sky is overcast, suggesting a cloudy day.

InternVL-2.5+AUE

The image depicts a tennis match taking place on a clay court. A female tennis player is in the foreground, bending over to pick up a tennis ball. She is wearing a light blue top and white shorts. The court is surrounded by a green fence, and beyond the fence, there is a large building with multiple levels. The building has a green and white striped awning and several windows. There are people standing and sitting on the upper level of the building, watching the match. The setting appears to be outdoors, with trees visible in the background. The sky is overcast, suggesting a cloudy day.

InternVL-2.5+Ours

The image depicts a tennis match taking place on a clay court. A female tennis player is in the foreground, bending over to pick up a tennis ball. She is wearing a light blue top and white shorts. The court is surrounded by a green fence, and there are several spectators and players visible in the background.
Behind the court, there is a large building with multiple levels, featuring a green and white striped awning. The building has several windows and a balcony area where more people are gathered. The setting appears to be outdoors, with trees visible in the background, suggesting a park or recreational area. The sky is overcast, indicating a cloudy day.

Please describe the image in detail.

InternVL-2.5

The image shows two white bowls filled with a creamy soup. The soup contains pieces of broccoli and spinach, giving it a vibrant green color. The bowls are placed on a wooden surface, and there is a wooden spoon resting on a small wooden plate behind the bowls. To the right, there is a spoon and a fork resting on a beige cloth napkin. The overall setting suggests a simple, rustic meal, possibly homemade. The lighting is soft, highlighting the textures and colors of the food.

InternVL-2.5+AUE

The image shows two white bowls filled with a creamy soup containing broccoli and spinach. The bowls are placed on a wooden surface, and there is a wooden cutting board partially visible behind them. To the right of the bowls, there is a spoon resting on a beige cloth napkin. The overall setting suggests a simple, home-cooked meal, possibly a healthy and nutritious option.

InternVL-2.5+Ours

The image shows two white bowls filled with a creamy soup containing broccoli and spinach. The soup appears to be a light, creamy broth, possibly a broccoli and spinach soup. The broccoli florets are bright green, and the spinach leaves are visible, adding a touch of green to the creamy base. The bowls are placed on a wooden surface, and there is a wooden spoon resting on a cloth napkin beside the bowls. The overall setting suggests a simple, healthy meal, likely served in a rustic or homey environment.

Please describe the image in detail.

InternVL-2.5

The image features a plush teddy bear sitting on a beige surface. The bear is dressed in a cozy, green and white robe with a white fur trim. The robe has a red drawstring tied in a bow at the front. The teddy bear has a light brown fur, with darker brown accents on its ears, paws, and the tip of its nose. Its eyes are small and black, and it has a friendly, inviting expression. The overall appearance is warm and inviting, suggesting a comforting and festive atmosphere.

InternVL-2.5+AUE

The image features a plush teddy bear sitting on a beige surface. The bear is dressed in a cozy, green and white robe with a white fur trim. The robe has a red drawstring tied in a bow at the front. The teddy bear has a light brown fur and is sitting with its legs stretched out in front. The background is slightly blurred, emphasizing the bear as the main subject. The overall scene gives a warm and inviting feel, suggesting a comfortable and relaxed setting.

InternVL-2.5+Ours

The image features a plush teddy bear sitting on a beige surface. The bear is dressed in a green and white robe with a red string tied around its neck. The robe has a fluffy white collar and cuffs, adding to the cozy appearance. The teddy bear has a light brown fur and a friendly expression with black eyes and a small black nose. The overall scene gives a warm and inviting feel, reminiscent of a festive or holiday setting.

Figure 15. **Extended qualitative results on the CHAIR benchmark using InternVL-2.5.**

Is there a tie in the image?

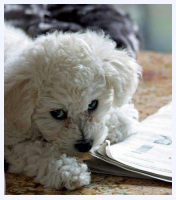
LLaVA-1.5-7B	LLaVA-1.5-7B+AUE	LLaVA-1.5-7B+Ours
yes, there is a tie in the image, which is placed on the coffee table.	no, there is no tie in the image. the image features a couch, a television, and a coffee table.	no, there is no tie in the image. the image features a couch, a table, and a television.

Is there a tie in the image?

LLaVA-1.5-7B	LLaVA-1.5-7B+AUE	LLaVA-1.5-7B+Ours
yes, there is a tie in the image, and it is being worn by a man.	yes, there is a tie in the image, and it is being worn by a man.	no, there is no tie in the image. the image features a woman standing next to a table with a cake on it.

Is there a bicycle in the image?

LLaVA-1.5-7B	LLaVA-1.5-7B+AUE	LLaVA-1.5-7B+Ours
no, there is no bicycle in the image. the image features a man standing in front of a store with a display of soccer jerseys.	no, there is no bicycle in the image. the image features a store with a variety of clothing items displayed outside.	yes, there is a bicycle in the image, and it is parked next to a store.

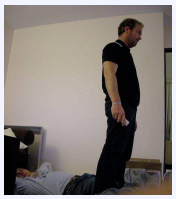


Is there a toaster in the image?

LLaVA-1.5-7B	LLaVA-1.5-7B+AUE	LLaVA-1.5-7B+Ours
yes, there is a toaster in the image, and the small white dog is laying on top of it.	no, there is no toaster in the image. the image features a white dog laying on a table.	no, there is no toaster in the image. the image features a white dog laying on a table.

Is there a scissors in the image?

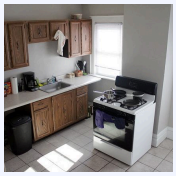
LLaVA-1.5-7B	LLaVA-1.5-7B+AUE	LLaVA-1.5-7B+Ours
yes, there is a pair of scissors in the image, located near the cat.	no, there are no scissors in the image. the image features a cat lying on a bed with blankets.	no, there are no scissors in the image. the image features a cat laying on a bed.



Is there a couch in the image?

LLaVA-1.5-7B	LLaVA-1.5-7B+AUE	LLaVA-1.5-7B+Ours
no, there is no couch in the image. the image shows a person lying on the floor.	yes, there is a couch in the image, and a man is sitting on it.	yes, there is a couch in the image, and a man is standing in front of it.

Figure 16. **Extended qualitative results on the POPE benchmark using LLaVA-1.5-7B.**

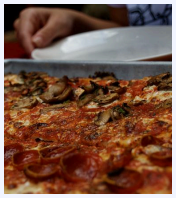


Is there a cake in the image?

LLaVA-1.5-13B	LLaVA-1.5-13B+AUE	LLaVA-1.5-13B+Ours
yes, there is a cake in the image. it is placed on the kitchen counter.	no, there is no cake in the image. the image features a kitchen with a sink, a window, and a dishwasher.	no, there is no cake in the image. the image features a kitchen with a stove and a window.

Is there a dining table in the image?

LLaVA-1.5-13B	LLaVA-1.5-13B+AUE	LLaVA-1.5-13B+Ours
no, there is no dining table in the image. the image shows a living room with a couch, a coffee table, and a television.	yes, there is a dining table in the image.	yes, there is a dining table in the image, located in the living room.



Is there a wine glass in the image?

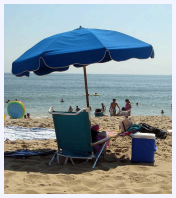
LLaVA-1.5-13B	LLaVA-1.5-13B+AUE	LLaVA-1.5-13B+Ours
yes, there is a wine glass in the image, and it is placed on a table next to a pizza.	yes, there is a wine glass in the image, and it is placed on top of a pizza.	no, there is no wine glass in the image. the image features a close-up of a pizza on a pan.

Is there a kite in the image?

LLaVA-1.5-13B	LLaVA-1.5-13B+AUE	LLaVA-1.5-13B+Ours
yes, there is a kite in the image, which is hanging on the wall.	no, there is no kite in the image. the image features a living room with a fireplace, a couch, and a table.	no, there is no kite in the image. the image features a living room with a fireplace, a couch, and a table.

Is there a tie in the image?

LLaVA-1.5-13B	LLaVA-1.5-13B+AUE	LLaVA-1.5-13B+Ours
yes, there is a tie in the image. it is located near the train tracks and the train.	yes, there is a tie in the image. it is located near the train tracks and is part of the train station.	no, there is no tie in the image. the image features a train and a sign.



Is there a refrigerator in the image?

LLaVA-1.5-13B	LLaVA-1.5-13B+AUE	LLaVA-1.5-13B+Ours
yes, there is a refrigerator in the image, located on the beach.	no, there is no refrigerator in the image.	no, there is no refrigerator in the image. the image features a beach scene with a blue umbrella, a chair, and people enjoying the beach.

Figure 17. **Extended qualitative results on the POPE benchmark using LLaVA-1.5-13B.**

Is there a dining table in the image?

Shikra-7B	Shikra-7B+AUE	Shikra-7B+Ours
no, there is no dining table in the image. the man is holding a pizza in a kitchen setting.	no, there is no dining table in the image. the man is holding a pizza in his hands and placing it in an oven.	yes, there is a dining table in the image, and a man is putting something on it.

Is there a dining table in the image?

Shikra-7B	Shikra-7B+AUE	Shikra-7B+Ours
no, there is no dining table in the image. the people are standing around a food stand.	yes, there is a dining table in the image, with a woman and two other people standing near it.	yes, there is a dining table in the image, and a group of people is standing around it.

Is there a pizza in the image?

Shikra-7B	Shikra-7B+AUE	Shikra-7B+Ours
yes, there is a pizza in the image, and people are waiting at a bus stop to get on a bus.	no, there is no pizza in the image. the image features a city bus and a large building, with people walking on the sidewalk.	no, there is no pizza in the image. the image features a white bus and people standing on the sidewalk near a city street.

Is there a spoon in the image?

Shikra-7B	Shikra-7B+AUE	Shikra-7B+Ours
yes, there is a spoon in the image, and it is covered with food.	yes, there is a spoon in the image.	no, there is no spoon in the image. the image features a dish containing food, but there is no spoon present.

Is there a car in the image?

Shikra-7B	Shikra-7B+AUE	Shikra-7B+Ours
yes, there is a car in the image, and it is driving behind the herd of sheep that are crossing the street.	the answer is no.	no, there is no car in the image. the photo features a herd of sheep running along the road.

Is there a bicycle in the image?

Shikra-7B	Shikra-7B+AUE	Shikra-7B+Ours
no, there is no bicycle in the image. the main subject is a brightly colored bus or double-decker bus with an attached store.	yes, there is a bicycle in the image. the bicycle is green and red, and it is parked next to a street lamp.	yes, there is a bicycle in the image, along with a street scene and a double-decker bus.

Figure 18. **Extended qualitative results on the POPE benchmark using Shikra-7B.**

Is there a backpack in the image?

Qwen-2.5-VL	Qwen-2.5-VL+AUE	Qwen-2.5-VL+Ours
no, there is no backpack in the image. the image shows a black cat standing next to a bottle of wine labeled "notro." there is also a clear glass bowl on the right side of the image, but no backpack is visible.	no, there is no backpack in the image. the image shows a black cat standing next to a bottle of wine labeled "notro." there are no visible backpacks or similar items in the background or foreground.	yes, there appears to be a black backpack in the background of the image, partially visible behind the cat and the wine bottle.

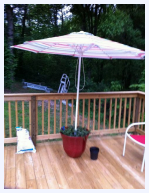
Is there a bottle in the image?

Qwen-2.5-VL	Qwen-2.5-VL+AUE	Qwen-2.5-VL+Ours
yes, there is a bottle visible on the bathroom counter in the background of the image. it appears to be a small, white bottle, possibly containing lotion or another personal care product.	no, there is no bottle visible in the image. the image shows a child standing on a stool in a bathroom, wearing adult-sized boots and holding a toothbrush. there are no bottles present in the scene.	no, there is no bottle visible in the image. the image shows a young child standing on a toilet, wearing adult-sized boots and a diaper, and holding what appears to be a toothbrush. there are no bottles present in the background or elsewhere in the image.

Is there a spoon in the image?

Qwen-2.5-VL	Qwen-2.5-VL+AUE	Qwen-2.5-VL+Ours
no, there is no spoon in the image. the image shows a slice of cake on a plate with whipped cream and orange segments on top, along with a spoon placed next to the plate.	yes, there is a spoon in the image. it is placed on the right side of the plate with the cake.	yes, there is a spoon in the image. it is placed on the right side of the plate with the cake slice.

Is there a potted plant in the image?



Qwen-2.5-VL	Qwen-2.5-VL+AUE	Qwen-2.5-VL+Ours
no, there is no potted plant in the image. The image shows a large red planter with what appears to be soil and some greenery inside it, but it does not have a visible plant growing in it. Instead, an umbrella is placed on top of the planter.	yes, there is a potted plant in the image. it is located under an umbrella on a wooden deck. the pot appears to be red and contains some greenery.	yes, there is a potted plant in the image. it is located on a wooden deck and appears to be a small potted plant with some greenery inside it. the pot is red, and there is an umbrella above it, which might be providing shade for the plant.

Is there a skis in the image?

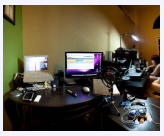
Qwen-2.5-VL	Qwen-2.5-VL+AUE	Qwen-2.5-VL+Ours
no, there are no skis visible in the image. the person in the air appears to be performing a snowboard trick, as indicated by the snowboard under their feet.	yes, there is a pair of skis visible in the image. they can be seen in the background, being used by a person who appears to be performing a jump or trick on the snow.	yes, there is a pair of skis visible in the image. they can be seen in the air, indicating that someone is performing a ski jump or trick. the skier appears to be mid-air, with their skis extended above them.

Is there a chair in the image?

Qwen-2.5-VL	Qwen-2.5-VL+AUE	Qwen-2.5-VL+Ours
yes, there is a chair visible in the image. it appears to be a wooden chair with a woven seat and backrest, positioned behind the group of people.	yes, there is a chair in the image. it appears to be a wooden chair with a cushioned seat and backrest, which the woman and child on the right are sitting on.	no, there is no chair visible in the image. the people appear to be standing or sitting on a surface that is not clearly identifiable as a chair.

Figure 19. **Extended qualitative results on the POPE benchmark using Qwen-2.5-VL.**

Is there a bottle in the image?



InternVL-2.5	InternVL-2.5+AUE	InternVL-2.5+Ours
yes, there is a bottle in the image.	no, there is no bottle in the image.	no, there is no bottle in the image.

Is there a person in the image?

InternVL-2.5	InternVL-2.5+AUE	InternVL-2.5+Ours
yes, there is a person in the image.	no, there is no person in the image.	no, there is no person in the image.

Is there a dining table in the image?

InternVL-2.5	InternVL-2.5+AUE	InternVL-2.5+Ours
yes, there is a dining table in the image. it is located in the background, behind the people who are standing in line.	no, there is no dining table in the image. the image shows a group of people standing in line at a food truck, with no visible dining tables.	no, there is no dining table in the image. the image shows a group of people standing in line at a food truck, with no visible dining tables.

Is there a baseball bat in the image?

InternVL-2.5	InternVL-2.5+AUE	InternVL-2.5+Ours
yes, there is a baseball bat in the image.	no, there is no baseball bat in the image. the image shows a person playing catch with a ball, but there is no baseball bat visible.	no, there is no baseball bat in the image. the image shows a person playing catch with a baseball.

Is there a sink in the image?

InternVL-2.5	InternVL-2.5+AUE	InternVL-2.5+Ours
yes, there is a sink in the image. the man is standing in front of a stainless steel sink.	yes, there is a sink in the image. the man is standing in front of a stainless steel sink.	there is no visible sink in the image. the main focus is on the man, the blender, and the pot. so, the answer is no.

Is there a cup in the image?

InternVL-2.5	InternVL-2.5+AUE	InternVL-2.5+Ours
yes, there is a cup in the image. it is located on the beach, near the surfboard and the umbrella.	no, there is no cup visible in the image. the main objects present are beach chairs, umbrellas, surfboards, and bags.	no, there is no cup in the image. the image features a beach scene with surfboards, a towel, and an umbrella.

Figure 20. **Extended qualitative results on the POPE benchmark using InternVL-2.5.**

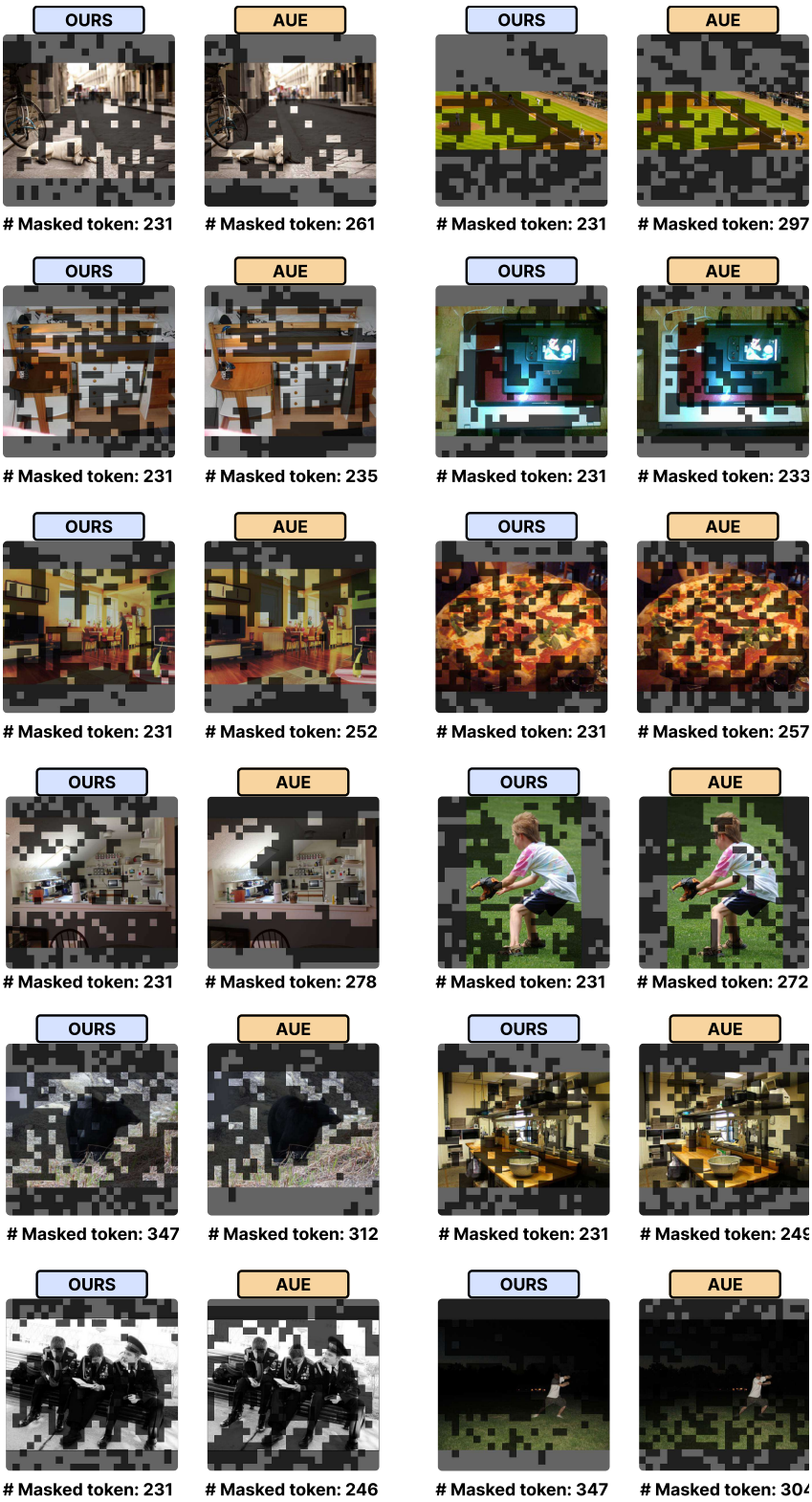


Figure 21. **Spatial mask comparisons between our DPP-based method and the AUE method under similar masked-token budgets.** Our method produces a more spatially dispersed pattern, effectively preserving diverse visual cues while filtering redundancy.