Mask What Matters: Mitigating Object Hallucinations in Multimodal Large Language Models with Object-Aligned Visual Contrastive Decoding

Boqi Chen[♠] **Xudong Liu**[◇*] **Jianing Qiu**[♣]
[♠]ETH Zurich [◇]Amazon [♣] MBZUAI
boqi.chen@ai.ethz.ch franklxd@amazon.com jianing.qiu@mbzuai.ac.ae

Abstract

We study object hallucination in Multimodal Large Language Models (MLLMs) and improve visual contrastive decoding (VCD) by constructing an object-aligned auxiliary view. We leverage object-centric attention in self-supervised Vision Transformers. In particular, we remove the most salient visual evidence to construct an auxiliary view that disrupts unsupported tokens and produces a stronger contrast signal. Our method is prompt-agnostic, model-agnostic, and can be seamlessly plugged into the existing VCD pipeline with little computation overhead, *i.e.*, a single cacheable forward pass. Empirically, our method demonstrates consistent gains on two popular object hallucination benchmarks across two MLLMs.

https://github.com/ratschlab/OA-VCD

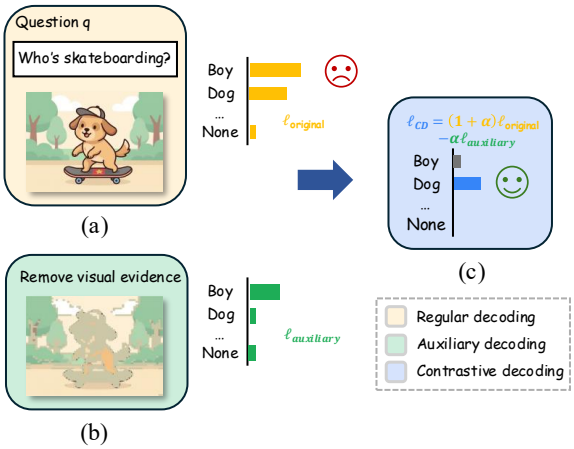

Figure 1: Overview of our method. **(a)** Regular decoding; **(b)** decoding using the auxiliary view where visual evidence is removed; **(c)** contrastive decoding.

1 Introduction

Multimodal Large Language Models (MLLMs) have shown impressive performance acorss various tasks such as image captioning (Li et al., 2023a; Qiu et al., 2024) and visual question answering (Lee et al., 2024; Wang et al., 2024), yet they suffer from object hallucination, *i.e.*, mentioning objects not grounded in the image (Li et al., 2023b). A popular line of work mitigates object hallucination at inference via visual contrastive decoding (VCD). VCD contrasts next-token distributions under the original image and a perturbed auxiliary view to suppress tokens that remain likely without visual support (Leng et al., 2024). Recent works improve VCD by constructing more informative auxiliary views. For instance, VS-CoDe (Kim et al., 2024b) proposes to select the augmentation that maximizes a softmax-distance criterion to strengthen the contrast signal. These perturbations, however, remain heuristic and at image-level, not necessarily aligned with object extents. AGLA (An et al., 2024) targets this alignment by preserving prompt-relevant regions while masking distractors using an image–text matching model, and fuse distributions from original and augmented views. Despite good results, it relies on prompt- and model-dependent cross-modal signals, and can risk circularity when biased attention guides the masking intended to correct it.

Self-supervised Vision Transformer (ViT) attention maps encode rich cues for semantic segmentation (Dosovitskiy et al., 2020; Caron et al., 2021). In this paper, we leverage object-centric attention to localize the most salient visual evidence and generate an auxiliary view by masking it out, yielding prompt-agnostic, semantically meaningful counterfactuals that avoid cross-modal dependencies and provide comprehensive object-level perturbations with a single cacheable forward pass. We empirically show that such auxiliary views produce a stronger contrast signal for VCD and leads to better performance on two popular object hallucination

[*]Work done before joining Amazon.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics Volume 4: Student Research Workshop, pages 9–16
March 26, 2026 ©2026 Association for Computational Linguistics

benchmarks across two different MLLMs.

2 Related Work

Contrastive decoding for reducing object hallucination. Multimodal generation in MLLMs is prone to object hallucination, *i.e.,* models generate responses that include entities not grounded in the image (Rohrbach et al., 2018; Li et al., 2023b; Wang et al., 2023). Recent progress therefore targets guided or constrained decoding to improve visual grounding. For instance, VCD contrasts model output distribution using perturbed images to downweight tokens that are driven by language priors (Leng et al., 2024). Following this, several works have proposed more effective ways to generate auxiliary views of the original image. Retrieval VCD retrieves single-concept positive and negative images from a pre-constructed database (Lee and Song, 2025). VSCoDe proposes to select the perturbation from a pool of augmentations (*e.g.*, blur, crop, color) that maximizes a softmax-distance criterion to strengthen the contrast signal (Kim et al., 2024b). AGLA combines global context with local discriminative features via an image–prompt matching scheme, highlighting relevant content while mitigating distractors (An et al., 2024).

Besides VCD, Instruction Contrastive Decoding contrasts output distributions under standard and disturbance instructions (*e.g.*, adding role prefixes) and subtracts the disturbance-induced distribution to detach hallucinated concepts during inference (Wan et al., 2024). CODE contrasts between the original self-generated caption and its perturbed variants to identify and suppress hallucinated tokens (Kim et al., 2024a). Activation Steering Decoding applies bidirectional contrastive adjustments to the model's hidden-state activations during inference—comparing forward and backward pass representations—to steer generated outputs away from hallucinated objects and toward correct ones (Su et al., 2025).

3 Method

Figure 1 provides an overview of our method. In this section, we first review VCD (Section 3.1), and then detail how to generate auxiliary views by removing salient visual evidence (Section 3.2).

3.1 Visual Contrastive Decoding

We consider an MLLM with parameters θ. Given a textual query x and a visual input v, the model generates a response auto-regressively from

$$\begin{aligned} y_t &\sim p_\theta\big(y_t \mid v, x, y_{<t}\big) \\ &\propto \exp\Big(\operatorname{logit}_\theta\big(y_t \mid v, x, y_{<t}\big)\Big). \end{aligned} \tag{1}$$

VCD obtains a second distribution under a *auxiliary* view v' and forms a contrastive distribution that amplifies the differences between the two:

$$\begin{aligned} p_{\text{vcd}}\big(y \mid v, v', x\big) = \operatorname{softmax}\Big(& \\ (1+\alpha)\operatorname{logit}_\theta\big(y \mid v, x\big) & \\ -\,\alpha \operatorname{logit}_\theta\big(y \mid v', x\big)\Big). & \end{aligned} \tag{2}$$

where $\alpha \geq 0$ controls the contrast strength. To avoid penalizing valid outputs from the original distribution and promoting implausible outputs from the augmented distribution, following (Leng et al., 2024; An et al., 2024), we adopt adaptive plausibility constraints (APC) which selectively consider tokens with high original probabilities and truncate other tokens as follows:

$$\begin{aligned} \mathcal{V}_{\text{head}}(y_{<t}) = \Big\{ y_t \in \mathcal{V} :\ & p_\theta\big(y_t \mid v, x, y_{<t}\big) \\ & \geq \beta \max_{w \in \mathcal{V}} p_\theta\big(w \mid v, x, y_{<t}\big) \Big\}, \end{aligned} \tag{3}$$

and set $p_{\text{vcd}}\big(y_t \mid v', x, y_{<t}\big)$=0 if $y_t \notin \mathcal{V}_{\text{head}}(y_{<t})$, with $\beta \in (0, 1]$. Combining VCD and APC yields the final decoding rule:

$$\begin{aligned} y_t \sim \operatorname{softmax}\Big(& (1+\alpha)\operatorname{logit}_\theta\big(y_t \mid v, x, y_{<t}\big) \\ & -\,\alpha \operatorname{logit}_\theta\big(y_t \mid v', x, y_{<t}\big)\Big), \\ & \text{s.t. } y_t \in \mathcal{V}_{\text{head}}(y_{<t}). \end{aligned} \tag{4}$$

3.2 Generate Auxiliary Views

Given an input image $v \in \mathbb{R}^{H \times W \times 3}$, we extract the attention of the [CLS] token on the heads of the last layer of a self-supervised ViT, *i.e.*, DINO (Caron et al., 2021). We then average the attention from multiple heads, reshape it to the patch grid size, and upsample it to (H, W) to obtain a saliency map $\tilde{\mathbf{S}}$, where higher values indicate more prominent visual evidence without task-specific supervision.

To create the an auxiliary view v', we threshold by quantile to remove the regions with high saliency. Formally, let $\gamma \in (0, 1)$ be the area ratio to remove, we define the quantile threshold as:

$$\lambda_\delta = \operatorname{Quantile}(\tilde{\mathbf{S}},\ 1-\gamma), \tag{5}$$

Table 1: Results (in %) on the three POPE subsets with LLaVA-v1.5 (7B). Best results are in **bold**.

Setting	Method	Accuracy ↑	F1 ↑
Random	Regular	84.7	83.2
	VCD	87.6	86.5
	AGLA	88.0	86.9
	Ours	**89.5**	**88.5**
Popular	Regular	80.8	79.9
	VCD	83.0	82.9
	AGLA	85.1	84.6
	Ours	**85.7**	**85.1**
Adversarial	Regular	77.4	77.4
	VCD	79.4	79.9
	AGLA	81.2	81.3
	Ours	**81.9**	**82.0**

Table 2: Results (in %) on the three POPE subsets with Qwen-VL (7B). Best results are in **bold**.

Setting	Method	Accuracy ↑	F1 ↑
Random	Regular	86.1	84.1
	VCD	86.7	85.0
	AGLA	87.4	85.7
	Ours	**88.0**	**86.5**
Popular	Regular	83.6	82.1
	VCD	84.0	82.5
	AGLA	84.8	83.8
	Ours	**85.5**	**84.3**
Adversarial	Regular	81.1	80.0
	VCD	81.6	80.6
	AGLA	82.6	81.6
	Ours	**82.9**	**82.0**

and the corresponding binary mask as:

$$\mathbf{M}_\delta \;=\; \mathbf{1}\Big\{\, \delta\left(\tilde{\mathbf{S}} - \lambda_\delta\right) \;>\; 0 \,\Big\}, \qquad (6)$$

where $\delta = -1$ for removing the most salient region. Let $\mathbf{B}$ represent a neutral background (*e.g.*, mean color of neighboring pixels), we obtain an auxiliary view

$$v' \;=\; \mathbf{M}_\delta \odot \mathbf{B} \;+\; \left(1 - \mathbf{M}_\delta\right) \odot v, \qquad (7)$$

where $\odot$ denoted element-wise multiplication.

Note that by setting $\delta = 1$, the auxiliary view will have the reserve effect, *i.e.*, removing the least salient region (distractors), highlighting the visual evidence as in (An et al., 2024).

4 Experiments

Settings. We evaluate on POPE (Li et al., 2023b) and the MME hallucination subset (Yin et al., 2024) using two different MLLMs: LLaVA-v1.5 (7B) (Liu et al., 2023) and Qwen-VL (7B) (Bai et al., 2023). We compare our method against three baselines: regular decoding, VCD with noise-based image perturbation (Leng et al., 2024) and AGLA (An et al., 2024). We threshold at $\gamma = 0.8$ and use mean color as neutral background by default (details in Appendix Section A.2). Ablations on different thresholds γ and backgrounds are provided in Section 4. More details on experiment setting can be found in Appendix Section A.1.

Results. Table 1 and 2 reports results on the POPE benchmark using LLaVA-v1.5 (7B) and Qwen-VL (7B), respectively. Across different POPE types and MLLMs, our method consistently improve over baselines. Compared with the strongest baseline, *i.e.*, AGLA, our method achieves higher accuracy and F1 in nearly all settings, with the largest gains on the random subset for LLaVA-v1.5 (7B).

Similar trends hold for the MME benchmark. Figures 3 and 4 visualize category-wise scores and the overall average for LLaVA-v1.5 (7B) and Qwen-VL (7B), respectively. Our method achieves the best average score for both models, with the most noticeable improvements on existence and color, the categories that are most susceptible to to spurious object correlations. Overall, these results suggest that masking salient regions from DINO attention yields a stronger, semantically targeted contrast signal while avoiding the prompt dependence of cross-modal masking. The consistent gains across two distinct MLLMs further support the model-agnostic utility of DINO attention for constructing auxiliary views in VCD.

Ablations We ablate different thresholds and neutral backgrounds on the POPE MSCOCO subset (Lin et al., 2014). Overall, we find our method robust to different parameters, with only modest change in performance when varying the thresholds and background methods.

Case study Figure 2 shows a case study on the LLaVA-Bench (Liu et al., 2023). We can observe that, given the same prompt and image, regular decoding leads to object hallucinations, *e.g.,* "addition dogs" and "two people in the background". We conjecture that these hallucinations stem from the

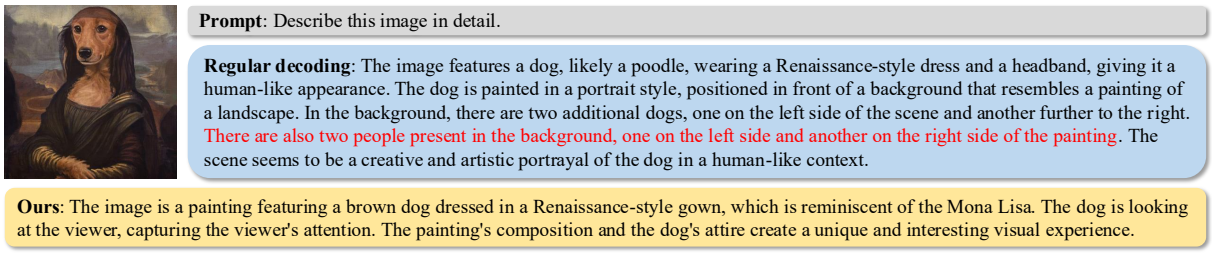

Figure 2: Captions generated by different decoding methods. Hallucinated contents are highlighted in red.

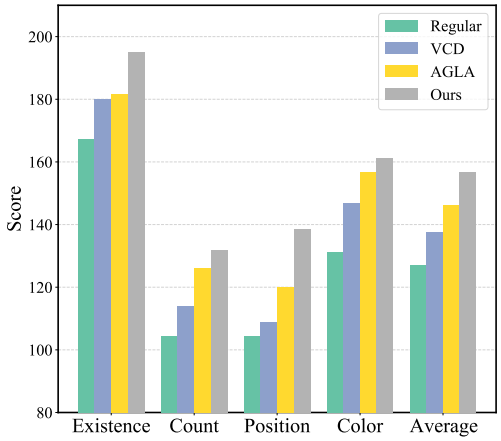

Figure 3: Results averaged across three seeds on the hallucination subset of MME with LLaVA-v1.5 (7B).

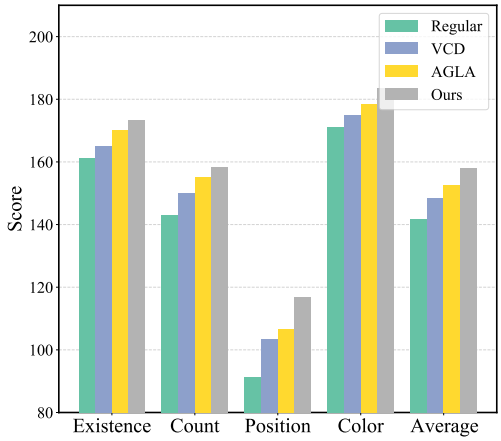

Figure 4: Results averaged across three seeds on the hallucination subset of MME with Qwen-VL (7B).

Table 3: Ablation results (F1, in %) on the POPE MSCOCO subset with LLaVA-v1.5 (7B) using different backgrounds. Threshold $\gamma = 0.8$.

Setting	Blur	Black	Mean
Random	89.4	90.7	90.5
Popular	88.1	88.6	88.6
Adversarial	86.0	86.4	86.2

Table 4: Ablation results (F1, in %) on the POPE MSCOCO subset with LLaVA-v1.5 (7B) using different thresholds γ with mean background.

Setting	0.2	0.4	0.6	0.8
Random	90.0	89.9	90.3	90.5
Popular	88.2	88.0	88.4	88.6
Adversarial	85.8	85.8	86.4	86.2

bias and language priors inherent to pretraining. In contrast, our method successfully mitigates these hallucinations without harming the coherence and informativeness of the output caption.

5 Conclusion

We show that object-aligned auxiliary views, constructed by removing salient visual evidence using a self-supervised ViT's attention, improve VCD for mitigating object hallucinations in MLLMs. Our method is prompt-agnostic, training-free, and model-agnostic, requiring only a single cacheable forward pass while yielding semantically meaningful, object-level perturbations. Empirically, we demonstrate that such auxiliary views yield a stronger contrastive signal than heuristic augmentations or cross-modal masking, and reduce hallucination on two benchmarks across two MLLMs.

Limitations

Our method relies on self-supervised ViT's saliency being well aligned with visual evidence. However, in cluttered scenes, this can lead to under/over-masking. The masking is prompt-agnostic, so in some cases it may suppress regions relevant to the current query, and performance can vary with area ratio and filler choices, though we found a single setting broadly effective. Future work will explore better background filling methods such as diffusion-based image inpainting (Corneanu et al., 2024).

Acknowledgement

This research was primarily supported by the ETH AI Center through an ETH AI Center doctoral fel-

lowship to Boqi Chen.

References

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, and 1 others. 2024. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. *arXiv preprint arXiv:2406.12718*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Ciprian Corneanu, Raghudeep Gadde, and Aleix M Martinez. 2024. Latentpaint: Image inpainting in latent space with diffusion models. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 4334–4343.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Junho Kim, Hyunjun Kim, Kim Yeonju, and Yong Man Ro. 2024a. Code: Contrasting self-generated description to combat hallucination in large multi-modal models. *Advances in Neural Information Processing Systems*, 37:133571–133599.

Sihyeon Kim, Boryeong Cho, Sangmin Bae, Sumyeong Ahn, and Se-Young Yun. 2024b. Vacode: Visual augmented contrastive decoding. *arXiv preprint arXiv:2408.05337*.

Jihoon Lee and Min Song. 2025. Retrieval visual contrastive decoding to mitigate object hallucinations in large vision-language models. *arXiv preprint arXiv:2505.20569*.

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

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. 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 (CVPR)*.

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

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Yannic Neuhaus and Matthias Hein. 2025. Repope: Impact of annotation errors on the pope benchmark. *arXiv preprint arXiv:2504.15707*.

Jianing Qiu, Wu Yuan, and Kyle Lam. 2024. The application of multimodal large language models in medicine. *The Lancet Regional Health–Western Pacific*, 45.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.

Jingran Su, Jingfan Chen, Hongxin Li, Yuntao Chen, Li Qing, and Zhaoxiang Zhang. 2025. Activation steering decoding: Mitigating hallucination in large vision-language models through bidirectional hidden state intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12964–12974.

X Wan and 1 others. 2024. Contrastive response generation for mitigating object hallucination in lvlms. *arXiv preprint arXiv:2405.xxxxx*.

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

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, and 1 others. 2023. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

A Appendix

A.1 Detailed Experiment Settings

A.1.1 Benchmarks

POPE. The Polling-based Object Probing Evaluation (POPE) (Li et al., 2023b) comprises 27,000 binary queries (`Yes/No`) targeting object existence across three sources: MSCOCO (Lin et al., 2014), A-OKVQA (Schwenk et al., 2022), and GQA (Hudson and Manning, 2019). For each source, POPE provides three negative-sampling regimes: random, popular, and adversarial. We report Accuracy, Precision, Recall, and F1. For the MSCOCO subset, we adopt the RePOPE annotations (Neuhaus and Hein, 2025), which correct erroneous labels and remove ambiguous cases.

MME. The MME benchmark (Yin et al., 2024) evaluates broad capabilities of MLLMs, including recognition of object attributes and inter-object relations. In this work, we focus on the hallucination-oriented subset (Existence, Count, Position, and Color), which spans both object- and attribute-level hallucinations. As in POPE, the answers are binary (`Yes/No`). Following the original protocol, our primary metric is Accuracy + Accuracy$^+$, where *Accuracy* is computed per question, and Accuracy$^+$ is computed per image and requires both associated questions for that image to be answered correctly. The latter is therefore a stricter indicator of comprehensive image-level understanding.

LLaVA-Bench. LLaVA-Bench consists of 24 images paired with 60 questions that cover diverse settings, including indoor and outdoor scenes, memes, paintings, and sketches. The benchmark is designed to probe MLLM performance on more challenging problems and out-of-domain scenarios. Following Leng et al. (2024), we present qualitative case studies on this dataset to illustrate the effectiveness of our approach.

A.1.2 Baselines

For all baselines, we use the default parameters.

A.2 Background Methods

Mean color. Let $\mu \in \mathbb{R}^3$ be the per-channel mean of I computed over all pixels in normalized space:

$$\mu_c = \frac{1}{HW} \sum_{x,y} I_{cxy}, \qquad c \in \{1,2,3\}.$$

The background is the constant field

$$B_{\text{mean}}(x,y) = \mu \quad \forall (x,y),$$

i.e., each masked pixel is replaced by the image's global mean color (per channel).

Blur. Let G_σ denote a spatial Gaussian blur (e.g., a 21×21 kernel). The background is the blurred version of the input,

$$B_{\text{blur}} = G_\sigma(I),$$

applied channel-wise. This preserves low-frequency color and illumination while suppressing high-frequency detail inside masked regions.

Black. Given normalized RGB inputs,

$$I_{\text{norm}} = \frac{I_{\text{rgb}} - \text{mean}}{\text{std}},$$

define the per-channel constant corresponding to pure black in RGB as

$$b_c = -\frac{\text{mean}_c}{\text{std}_c}, \qquad c \in \{1,2,3\}.$$

The background is then

$$B_{\text{black}}(x,y) = b \quad \forall (x,y),$$

which replaces masked pixels with a distribution-consistent black in the model's normalized space.

A.3 Dataset License

POPE, MME, and LLaVA-Bench are intended for research usage.

- **POPE.** It has an MIT License, allowing research usage.
- **MME.** It has a Creative Commons Attribution-ShareAlike 4.0 license, allowing research usage.
- **LLaVA-Bench.** It has a Creative Commons Attribution 4.0 license, allowing research usage.

Table 5: Results (in %) on the three POPE subsets with LLaVA-v1.5 (7B) and Qwen-VL (7B). Best results are in **bold**.

Model	Setting	Method	Accuracy ↑	Precision	Recall	F1 Score ↑
LLaVA-v1.5	*Random*	Regular	84.7	87.3	79.4	83.2
		VCD	87.6	89.1	84.0	86.5
		AGLA	88.0	95.1	80.2	86.9
		Ours	**89.5**	92.3	85.0	**88.5**
	Popular	Regular	80.8	81.1	78.7	79.9
		VCD	83.0	82.4	83.4	82.9
		AGLA	85.1	88.1	81.8	84.6
		Ours	**85.7**	86.2	84.1	**85.1**
	Adversarial	Regular	77.4	75.5	79.4	77.4
		VCD	79.4	76.6	83.7	79.9
		AGLA	81.2	81.5	81.7	81.3
		Ours	**81.9**	79.8	84.7	**82.0**
Qwen-VL	*Random*	Regular	86.1	91.9	77.7	84.1
		VCD	86.7	91.9	78.7	85.0
		AGLA	87.4	93.4	79.3	85.7
		Ours	**88.0**	93.4	80.7	**86.5**
	Popular	Regular	83.6	87.4	77.4	82.1
		VCD	84.0	87.8	78.0	82.5
		AGLA	84.8	89.6	78.7	83.8
		Ours	**85.5**	89.3	79.8	**84.3**
	Adversarial	Regular	81.1	82.7	77.5	80.0
		VCD	81.6	83.2	78.1	80.6
		AGLA	82.6	84.4	79.0	81.6
		Ours	**82.9**	84.2	79.9	**82.0**

Table 6: Results averaged across three seeds on the hallucination subset of MME with LLaVA-v1.5 (7B). Mean and standard deviation are reported. Best results are in **bold**.

Method	EXISTENCE	COUNT	POSITION	COLOR
Regular	167.22 ±7.88	104.44 ±1.93	104.45 ±41.41	131.11 ±27.61
VCD	180.00 ±0.00	113.89 ±4.41	108.89 ±11.10	146.67 ±22.13
AGLA	181.67 ±2.89	126.11 ±0.96	120.00 ±1.67	156.66 ±11.35
Ours	**195.00 ±5.00**	**131.67 ±8.82**	**138.33 ±4.41**	**165.00 ±3.47**

A.4 Detailed Results on the POPE Benchmark

The detailed results on the POPE benchmark are shown in Table 5.

A.5 Detailed Results on the MME Benchmark Hallucination Subset

The detailed results on the hallucination subset of the MME benchmark using LLaVA-v1.5 (7B) and Qwen-VL (7B) are present in Table 6 and 7, respectively. Note that since each type in the hallucination subset only contains 60 questions, resulting 240 question in total, we perform three runs with different randomly seeds and report the average performance for a more robust evaluation.

A.6 Visualization of Generated Auxiliary Views with Varying Parameters

We provide visualizations of the generated auxiliary views by removing visual evidence at different thresholds in Figure 6 and with different background inpainting methods in in Figure 5.

Table 7: Results averaged across three seeds on the hallucination subset of MME with Qwen-VL (7B). Mean and standard deviation are reported. Best results are in **bold**.

Method	EXISTENCE	COUNT	POSITION	COLOR
Regular	161.11 ±1.92	142.78 ±5.00	91.11 ±9.18	171.11 ±2.55
VCD	165.00 ±0.00	150.00 ±2.89	103.33 ±2.89	175.00 ±5.00
AGLA	170.00 ±0.00	155.00 ±2.89	106.66 ±2.89	178.33 ±2.89
Ours	**173.33** ±2.89	**158.33** ±2.89	**116.66** ±5.77	**183.33** ±0.00

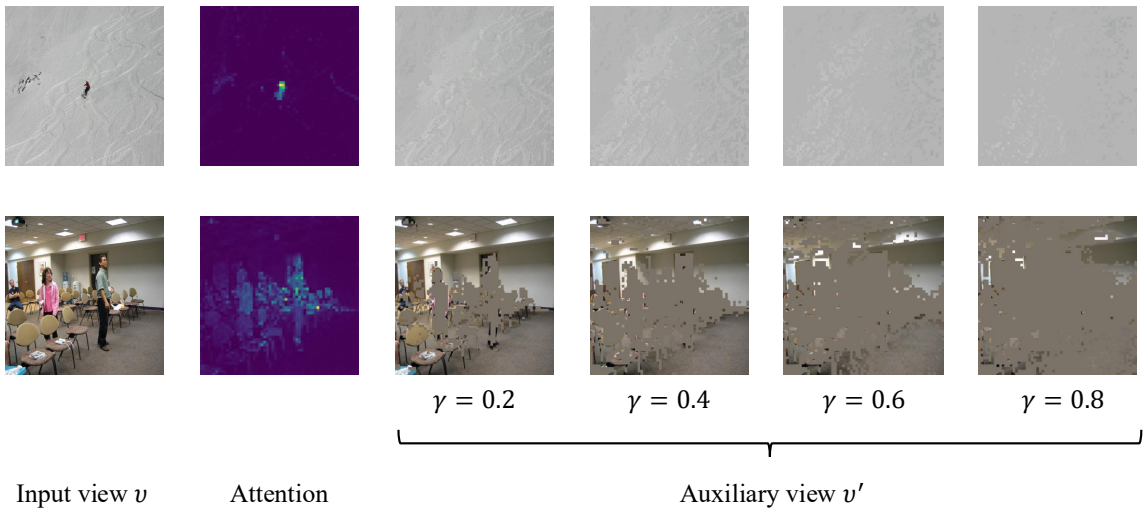

Figure 5: Visualization of generated auxiliary views with different thresholds. Background are all set to mean color.

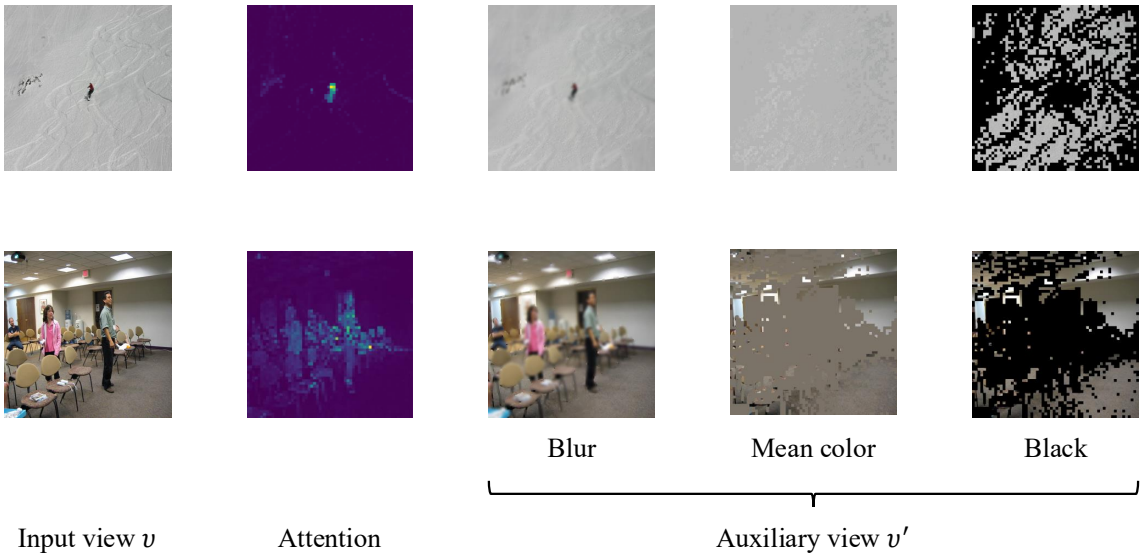

Figure 6: Visualization of generated auxiliary views with different backgrounds. Thresholds are all set to $\gamma = 0.8$.