Envision, Attend, Then Respond: Counterfactual Hallucination Mitigation in Large Vision-Language Models

Supplementary Material

A. Derivations and Proofs

A.1. Derivation of the Gradient Field

Given the latent representation $\boldsymbol{z}_0$ and the intermediate latent representation $\boldsymbol{z}_T$ in step T, the DDPM forward process (Equation 1) ensures that $\boldsymbol{z}_T = \sqrt{\bar{\alpha}_T}\cdot\boldsymbol{z}_0 + \sqrt{1-\bar{\alpha}_T}\cdot\boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. We define $\alpha_T = \sqrt{\bar{\alpha}_T}$ and $\sigma_T^2 = 1-\bar{\alpha}_T$ to simplify the notation. Therefore, the DDPM forward process is expressed as

$$\boldsymbol{z}_T = \alpha_T \boldsymbol{z}_0 + \sigma_T \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}). \tag{8}$$

Equation 8 indicates that the conditional distribution $p(\boldsymbol{z}_T \mid \boldsymbol{z}_0)$ is Gaussian, i.e.,

$$\begin{aligned} p(\boldsymbol{z}_T \mid \boldsymbol{z}_0) &= \mathcal{N}(\alpha_T \boldsymbol{z}_0, \sigma_T^2 \boldsymbol{I}), \\ \nabla_{\boldsymbol{z}_T} p(\boldsymbol{z}_T \mid \boldsymbol{z}_0) &= -\frac{\boldsymbol{z}_T - \alpha_T \boldsymbol{z}_0}{\sigma_T^2}\, p(\boldsymbol{z}_T \mid \boldsymbol{z}_0). \end{aligned} \tag{9}$$

The marginal distribution $p(\boldsymbol{z}_T)$ is

$$p(\boldsymbol{z}_T) = \int p(\boldsymbol{z}_0)\, p(\boldsymbol{z}_T \mid \boldsymbol{z}_0)\, \mathrm{d}\boldsymbol{z}_0. \tag{10}$$

The gradient field $\boldsymbol{G}$ is

$$\begin{aligned} \boldsymbol{G} &= \nabla_{\boldsymbol{z}_T} \log p(\boldsymbol{z}_T) = \frac{\nabla_{\boldsymbol{z}_T} p(\boldsymbol{z}_T)}{p(\boldsymbol{z}_T)} \\ &= \frac{\int p(\boldsymbol{z}_0) \nabla_{\boldsymbol{z}_T} p(\boldsymbol{z}_T \mid \boldsymbol{z}_0) \mathrm{d}\boldsymbol{z}_0}{\int p(\boldsymbol{z}_0) p(\boldsymbol{z}_T \mid \boldsymbol{z}_0) \mathrm{d}\boldsymbol{z}_0} \\ &= -\frac{1}{\sigma_T^2} \cdot \frac{\int (\boldsymbol{z}_T - \alpha_T \boldsymbol{z}_0) p(\boldsymbol{z}_0) p(\boldsymbol{z}_T \mid \boldsymbol{z}_0) \mathrm{d}\boldsymbol{z}_0}{\int p(\boldsymbol{z}_0)\, p(\boldsymbol{z}_T \mid \boldsymbol{z}_0) \mathrm{d}\boldsymbol{z}_0}. \end{aligned} \tag{11}$$

Given that the expectation of $p(\boldsymbol{z}_T \mid \boldsymbol{z}_0)$ is

$$\begin{aligned} \mathbb{E}[\boldsymbol{z}_0 \mid \boldsymbol{z}_T] &= \frac{1}{p(\boldsymbol{z}_T)} \int \boldsymbol{z}_0 p(\boldsymbol{z}_0) p(\boldsymbol{z}_T \mid \boldsymbol{z}_0) \mathrm{d}\boldsymbol{z}_0 \\ &= \frac{\int \boldsymbol{z}_0 p(\boldsymbol{z}_0) p(\boldsymbol{z}_T \mid \boldsymbol{z}_0) \mathrm{d}\boldsymbol{z}_0}{\int p(\boldsymbol{z}_0) p(\boldsymbol{z}_T \mid \boldsymbol{z}_0) \mathrm{d}\boldsymbol{z}_0}, \end{aligned} \tag{12}$$

we can obtain the gradient field

$$\begin{aligned} \boldsymbol{G} &= -\frac{1}{\sigma_T^2} \left(\boldsymbol{z}_T - \alpha_T \frac{\int \boldsymbol{z}_0 p(\boldsymbol{z}_0) p(\boldsymbol{z}_T \mid \boldsymbol{z}_0) \mathrm{d}\boldsymbol{z}_0}{\int p(\boldsymbol{z}_0)\, p(\boldsymbol{z}_T \mid \boldsymbol{z}_0) \mathrm{d}\boldsymbol{z}_0} \right) \\ &= \frac{\alpha_T \mathbb{E}[\boldsymbol{z}_0 \mid \boldsymbol{z}_T] - \boldsymbol{z}_T}{\sigma_T^2} = \frac{\alpha_T \mathbb{E}[\boldsymbol{z}_0 \mid \boldsymbol{z}_T] - \boldsymbol{z}_T}{1-\bar{\alpha}_T}. \end{aligned} \tag{13}$$

From the DDPM forward process given in Equation 8, we have $\boldsymbol{z}_0 = (\boldsymbol{z}_T - \sigma_T \boldsymbol{\varepsilon})/\alpha_T$. Stable Diffusion v1.5 predicts the injected noise $\boldsymbol{\varepsilon}$ through a noise prediction network $\varepsilon_\theta(\boldsymbol{z}_T, T)$ [57], which gives an estimator of Equation 12:

$$\mathbb{E}[\boldsymbol{z}_0 \mid \boldsymbol{z}_T] \approx \hat{\boldsymbol{z}}_0(\boldsymbol{z}_T, T) = \frac{\boldsymbol{z}_T - \sigma_T \varepsilon_\theta(\boldsymbol{z}_T, T)}{\alpha_T}. \tag{14}$$

With Equation 14, the gradient field $\boldsymbol{G}$ in Equation 13 can be estimated through

$$\begin{aligned} \boldsymbol{G} &\approx \frac{\alpha_T \hat{\boldsymbol{z}}_0(\boldsymbol{z}_T, T) - \boldsymbol{z}_T}{1-\bar{\alpha}_T} \\ &= \frac{\alpha_T \cdot \frac{\boldsymbol{z}_T - \sigma_T \varepsilon_\theta(\boldsymbol{z}_T, T)}{\alpha_T} - \boldsymbol{z}_T}{1-\bar{\alpha}_T} \\ &= -\frac{\sigma_T \varepsilon_\theta(\boldsymbol{z}_T, T)}{1-\bar{\alpha}_T} = -\frac{\varepsilon_\theta(\boldsymbol{z}_T, T)}{\sqrt{1-\bar{\alpha}_T}}. \end{aligned} \tag{15}$$

B. Detailed Experimental Settings

B.1. Benchmarks

VLMBias [68]. The VLMBias benchmark is designed for testing counterfactual visual hallucination in LVLMs. It presents images in which well-known objects are subtly modified (e.g., adding an extra leg to a commonly seen animal or changing stripes on a brand logo) and asks objective tasks (e.g., counting, yes/no identification) to expose whether the model relies on memorised priors rather than actual visual evidence. VLMBias is difficult, probing models' ability to detect counterfactual visual cues.

WHOOPS [7]. WHOOPS is constructed of synthetic images that violate commonsense (e.g., famous athletes playing chess instead of football), designed to challenge visual commonsense reasoning and thus to surface hallucination or misgrounding in multimodal models. WHOOPS contains 26 categories, which we group into three higher-level classes:

- **Social** (10 categories): Age mismatch, Celebrity occupation, Cultural knowledge, Folklore knowledge, Social conventions, Unusual dish, Unsuitable environment, Untypical behavior, Inability to execute, Incorrect usage.
- **Natural** (10 categories): Biological rules, Nature phenomena mismatch, Physics rules, Principles conflict, Refraction mismatch, Object shape, Color inversion, Visual similarity, Safety, Temporal discrepancy.
- **Symbolic** (6 categories): Art knowledge, Symbolic inversion, Geographic mismatch, Unnatural environment, Nutrition mismatch, Health.

PhD [53]. The ChatGPT-Prompted visual hallucination evaluation Dataset introduces a large-scale benchmark to evaluate visual hallucinations in multimodal large-language

models. It spans five recognition tasks (object, attribute, position, sentiment, and counting) and includes over 100k VQA triplets, from standard visual QA to contexts with misleading cues and counterfactual images.

HallusionBench [28]. The HallusionBench benchmark is an advanced diagnostic suite targeting both language hallucination and visual illusion in LVLMs. It consists of 346 images paired with 1,129 human-crafted questions, emphasising nuanced image-context reasoning and exposing sophisticated failure modes of current models.

POPE [44]. Polling-based Object Probing Evaluation is a representative general-purpose dataset for object hallucination detection in LVLMs. It converts the presence or absence of objects into a straightforward binary classification task and comprises three distinct subsets: random, popular, and adversarial according to different negative-sampling strategies. We adopt the MSCOCO-based version of POPE, which is widely used as a canonical object hallucination testbed in LVLMs.

B.2. Other Methods

We compare EnAR with several existing decoding level methods for mitigating hallucination in LVLMs.

VCD [39] compares the predictions generated from a clean visual input and a deliberately corrupted one. Given a textual query $\boldsymbol{x}$ and an image input $\boldsymbol{V}$, VCD constructs a distorted view $\boldsymbol{V}'$ of the image by applying a predefined noise (e.g., Gaussian noise) to $\boldsymbol{V}$. Feeding $\boldsymbol{v}$ and $\boldsymbol{v}'$ ($\boldsymbol{V}$ and $\boldsymbol{V}'$ embedded by the vision encoder) into the LVLM yields two sets of logits, $p(y \mid \boldsymbol{v}, \boldsymbol{x})$ and $p(y \mid \boldsymbol{v}', \boldsymbol{x})$, which are combined to form contrastive decoding logits:

$$p_{\text{vcd}} = (1+\alpha)\, p(y \mid \boldsymbol{v}, \boldsymbol{x}) - \alpha\, p(y \mid \boldsymbol{v}', \boldsymbol{x}). \tag{16}$$

M3ID [23] contrasts the predictions conditioned on the full visual-text input with those from a purely textual prompt that omits the image. Given a textual query $\boldsymbol{x}$ and a visual input $\boldsymbol{v}$, we obtain two sets of logits: $p(y \mid \boldsymbol{v}, \boldsymbol{x})$ from the vision-language model and $p(y \mid \boldsymbol{x})$ from the same model run without visual information. At decoding step t, M3ID forms contrastive decoding logits by gradually amplifying the discrepancy between these two predictions:

$$\begin{aligned} p_{\text{m3id}} &= p(y \mid \boldsymbol{v}, \boldsymbol{x}) + \gamma_t \big(p(y \mid \boldsymbol{v}, \boldsymbol{x}) - p(y \mid \boldsymbol{x})\big), \\ \gamma_t &= \frac{1 - e^{-\lambda t}}{e^{-\lambda t}}. \end{aligned} \tag{17}$$

RITUAL [74] mitigates hallucination by comparing the predictions generated from the original visual input with that under a stochastically transformed version of the same image. Given a textual query $\boldsymbol{x}$ and an image $\boldsymbol{V}$, a transformed view $\boldsymbol{V}'$ is produced using standard image augmentations such as cropping, flipping, or color jitter. Encoding $\boldsymbol{V}$ and $\boldsymbol{V}'$ yields $\boldsymbol{v}$ and $\boldsymbol{v}'$. The model is then queried with both $\boldsymbol{v}$ and $\boldsymbol{v}'$, yielding two sets of logits: $p(y \mid \boldsymbol{v}, \boldsymbol{x})$ and $p(y \mid \boldsymbol{v}', \boldsymbol{x})$. RITUAL blends these two predictions using a weighting coefficient:

$$p_{\text{ritual}} = p(y \mid \boldsymbol{v}, \boldsymbol{x}) + \kappa\, p(y \mid \boldsymbol{v}', \boldsymbol{x}), \tag{18}$$

where κ controls the contribution of the transformed-image prediction.

DeGF [80] introduces an auxiliary visual reference generated from the model's own initial response, enabling the LVLM to self-correct hallucinations during decoding. Given a textual query $\boldsymbol{x}$ and an image $\boldsymbol{V}$, the LVLM first produces an initial response, which is then fed into a text-to-image generative model to synthesize a new visual reference $\boldsymbol{V}'$. Encoding $\boldsymbol{V}$ and $\boldsymbol{V}'$ yields $\boldsymbol{v}$ and $\boldsymbol{v}'$. At decoding step, the model computes two next-token logits: $p(y \mid \boldsymbol{v}, \boldsymbol{x})$ and $p(y \mid \boldsymbol{v}', \boldsymbol{x})$. DeGF measures the discrepancy between these two predictions via token-level Jensen-Shannon divergence $d(\boldsymbol{v}, \boldsymbol{v}')$, and performs *complementary* or *contrastive* decoding according to a threshold γ:

$$\begin{aligned} &p_{\text{degf}} = \\ &\begin{cases} p(y \mid \boldsymbol{v}, \boldsymbol{x}) + \alpha_1\, p(y \mid \boldsymbol{v}', \boldsymbol{x}), & d < \gamma, \\ (1+\alpha_2)\, p(y \mid \boldsymbol{v}, \boldsymbol{x}) - \alpha_2\, p(y \mid \boldsymbol{v}', \boldsymbol{x}), & d \geq \gamma. \end{cases} \\ &d(\boldsymbol{v}, \boldsymbol{v}') = \mathcal{D}_{\text{JS}}\big(p(y \mid \boldsymbol{v}, \boldsymbol{x}) \,\|\, p(y \mid \boldsymbol{v}', \boldsymbol{x})\big), \end{aligned} \tag{19}$$

where the JS divergence between two generic distributions P and Q is defined as

$$\begin{aligned} \mathcal{D}_{\text{JS}}(P \,\|\, Q) &= \frac{1}{2}\mathcal{D}_{\text{KL}}(P \,\|\, M) + \frac{1}{2}\mathcal{D}_{\text{KL}}(Q \,\|\, M), \\ M &= \frac{1}{2}(P+Q). \end{aligned} \tag{20}$$

$\mathcal{D}_{\text{KL}}$ denotes the Kullback-Leibler divergence, α_1 and α_2 modulate the influence of the generated reference.

AGLA [3] suppresses hallucinations by explicitly masking image regions that are irrelevant to the textual query. Given an image $\boldsymbol{V}$ and query $\boldsymbol{x}$, AGLA first employs a matching model to compute the global similarity between the text and image input, and derives a correlation score for each image patch. The estimated irrelevant region is masked to form an augmented image $\boldsymbol{V}'$, which removes visual evidence unrelated to the prompt. The LVLM is then queried with both $\boldsymbol{v}$ and $\boldsymbol{v}'$ ($\boldsymbol{V}$ and $\boldsymbol{V}'$ embedded by the vision encoder), producing two sets of logits: $p(y \mid \boldsymbol{v}, \boldsymbol{x})$ and $p(y \mid \boldsymbol{v}', \boldsymbol{x})$. AGLA fuses these two predictions through a combination:

$$p_{\text{agla}} = p(y \mid \boldsymbol{v}, \boldsymbol{x}) + \alpha\, p(y \mid \boldsymbol{v}', \boldsymbol{x}), \tag{21}$$

where α controls the contribution from the masked view. By suppressing tokens that rely on prompt-irrelevant visual evidence, AGLA encourages the model to focus on regions aligned with the query.

Table 5. **Results on the WHOOPS benchmark.** The results (i.e., the accuracy in %) are obtained across 26 hallucination categories with three LVLM backbones (InternVL3.5-8B, Qwen2.5VL-7B, LLaVA-v1.5-7B) under different strategies. The upper table shows the first 13 task categories; the lower table shows the remaining 13 categories.

Model	Method	Age	Art	Biological	Celebrity	Color	Cultural	Folklore	Geographic	Health	Inability	Incorrect	Nature	Nutrition
	Regular	54.35	83.33	75.47	31.82	57.89	59.57	74.47	**84.21**	30.00	50.00	72.46	81.82	58.62
	+VCD	56.52	94.44	73.58	54.55	61.40	70.21	68.09	78.95	60.00	58.33	81.16	90.91	57.47
	+M3ID	56.52	88.89	77.36	54.55	59.65	74.47	68.09	81.58	70.00	58.33	76.81	90.91	59.77
InternVL3.5-8B	+RITUAL	56.52	88.89	**77.36**	54.55	59.65	74.47	68.09	81.58	70.00	58.33	75.36	90.91	59.77
	+DeGF	52.17	88.89	75.47	54.55	63.16	72.34	68.09	81.58	70.00	55.56	73.91	90.91	55.17
	+AGLA	58.70	94.44	75.47	59.09	63.16	**77.66**	**72.34**	81.58	70.00	58.33	75.36	81.82	60.92
	+EnAR	**67.39**	**100.00**	75.47	**63.64**	**68.42**	74.47	68.09	81.58	**70.00**	**61.11**	**78.26**	**90.91**	**65.52**
	Regular	56.52	50.00	69.81	63.64	61.40	46.81	59.57	21.05	60.00	69.44	43.48	54.55	36.78
	+VCD	54.35	61.11	75.47	68.18	63.16	54.26	68.09	21.05	40.00	69.44	59.42	54.55	42.53
	+M3ID	60.87	61.11	77.36	86.36	68.42	58.51	68.09	21.05	70.00	69.44	60.87	63.64	48.28
Qwen2.5VL-7B	+RITUAL	60.87	55.56	71.70	59.09	68.42	58.51	68.09	23.68	70.00	69.44	55.07	63.64	50.57
	+DeGF	63.04	61.11	77.36	**86.36**	70.18	57.45	68.09	23.68	60.00	72.22	60.87	**63.64**	48.28
	+AGLA	60.87	55.56	69.81	81.82	68.42	54.26	68.09	21.05	60.00	75.00	52.17	54.55	44.83
	+EnAR	**65.22**	**61.11**	**81.13**	81.82	**71.93**	**62.77**	**68.09**	**36.84**	**70.00**	**77.78**	**62.32**	54.55	**59.77**
	Regular	2.17	0.00	3.77	9.09	1.75	9.57	6.38	13.16	0.00	0.00	10.14	0.00	13.79
	+VCD	4.35	0.00	3.77	9.09	1.75	9.57	8.51	5.26	0.00	0.00	10.14	0.00	12.64
	+M3ID	2.17	0.00	3.77	9.09	1.75	9.57	6.38	13.16	0.00	0.00	10.14	0.00	13.79
LLaVA-v1.5-7B	+RITUAL	2.17	0.00	3.77	4.55	1.75	9.57	6.38	13.16	0.00	0.00	8.70	0.00	13.79
	+DeGF	2.17	0.00	0.00	0.00	5.26	10.53	6.38	10.00	10.00	5.56	7.25	0.00	6.90
	+AGLA	2.17	0.00	3.77	4.55	1.75	8.51	6.38	13.16	0.00	0.00	10.14	0.00	13.79
	+EnAR	**15.22**	0.00	**9.43**	**9.09**	**5.26**	**13.83**	**25.53**	**18.42**	**10.00**	**5.56**	**10.14**	**9.09**	**22.99**

Model	Method	Object	Physics	Principles	Refraction	Safety	Social	Symbolic	Temporal	Unnatural	Unsuitable	Untypical	Unusual	Visual
	regular	82.69	64.71	47.83	91.67	73.58	**100.00**	54.44	55.73	54.55	**75.00**	50.00	52.27	76.47
	+VCD	92.31	70.59	46.65	83.33	81.13	83.33	66.67	54.96	72.73	72.22	60.00	70.45	70.59
	+M3ID	88.46	70.59	50.00	91.67	81.13	83.33	70.00	58.02	72.73	72.22	65.00	75.00	70.59
InternVL3.5-8B	+RITUAL	88.46	70.59	50.00	91.67	81.13	83.33	71.11	58.02	68.18	72.22	70.00	72.73	70.59
	+DeGF	88.46	70.59	50.00	83.33	81.13	83.33	70.00	49.62	72.73	69.44	65.00	72.73	70.59
	+AGLA	90.38	70.59	47.83	91.67	79.25	83.33	72.22	63.36	68.18	69.44	70.00	75.00	70.59
	+EnAR	**92.31**	**70.59**	**54.36**	**100.00**	**81.13**	83.33	**76.67**	**63.36**	**81.82**	66.67	**70.00**	**81.82**	**76.47**
	regular	17.31	35.29	60.87	33.33	56.60	100.00	52.22	43.51	40.91	38.89	90.00	70.45	17.65
	+VCD	38.46	41.18	54.35	**66.67**	69.81	100.00	55.56	45.04	45.45	50.00	85.00	63.64	17.65
	+M3ID	40.38	41.18	56.52	41.67	73.58	100.00	63.33	49.62	45.45	58.33	90.00	63.64	17.65
Qwen2.5VL-7B	+RITUAL	40.38	41.18	56.52	41.67	69.81	100.00	62.22	48.85	45.45	55.56	80.00	63.64	17.65
	+DeGF	40.38	41.18	56.52	41.67	**75.47**	100.00	63.33	49.62	45.45	58.33	90.00	**65.91**	17.65
	+AGLA	38.46	35.29	56.52	41.67	**75.47**	100.00	64.44	48.85	40.91	**61.11**	85.00	63.64	17.65
	+EnAR	**44.23**	**52.94**	**60.87**	41.67	73.58	**100.00**	**70.00**	**54.96**	**50.00**	55.56	**95.00**	**65.91**	**47.06**
	regular	19.23	5.80	2.17	0.00	0.00	0.00	3.33	6.87	13.64	11.11	5.00	2.27	41.18
	+VCD	**23.08**	5.88	2.17	0.00	0.00	0.00	4.44	6.87	13.64	8.33	5.00	2.27	**47.06**
	+M3ID	19.23	5.88	2.17	0.00	0.00	0.00	3.33	6.87	13.64	11.11	5.00	2.27	41.18
LLaVA-v1.5-7B	+RITUAL	19.23	5.88	2.17	0.00	0.00	0.00	3.33	6.87	13.64	8.33	5.00	2.27	41.18
	+DeGF	19.23	0.00	2.17	**8.33**	0.00	0.00	4.44	7.63	9.09	5.56	5.00	0.00	35.29
	+AGLA	19.23	5.88	2.17	0.00	0.00	0.00	2.22	9.16	13.64	11.11	5.00	2.27	41.18
	+EnAR	19.23	**5.88**	**10.87**	0.00	**1.89**	0.00	**11.11**	**15.27**	**22.73**	**11.11**	**15.00**	**2.27**	41.18

C. Additional Results

C.1. Detailed Results on WHOOPS

Table 5 presents the results on the complete WHOOPS benchmark, reporting accuracy on all 26 hallucination categories for three representative LVLMs (e.g., InternVL3.5-8B [71], Qwen2.5VL-7B [6], and LLaVA-v1.5-7B [51]) under multiple decoding strategies. The upper block summarizes performance on the first 13 categories, while the lower block reports the remaining 13, together providing a category-wise characterization of how each method behaves on this challenging counterfactual benchmark.

Across all three backbones, EnAR attains the highest aggregate accuracy on WHOOPS over the regular decoding baseline and exceeds other decoding-level approaches on most category groups. The improvements are particularly pronounced in semantically demanding or conceptually sensitive categories such as Age, Nutrition, and Symbolic, where models are more prone to relying on spurious world knowledge rather than actual visual input. These results indicate that EnAR enhances robustness under counterfactual perturbations and offers a more reliable decoding paradigm for mitigating counterfactual hallucinations on WHOOPS.

Table 6. **Ablation study of key components on VLMBias.** We compare the full EnAR across different backbones by removing the uncertainty map, disabling the visual impression, and using the regular decoding process.

Model	Setting	Animals	Chess Pieces	Flags	Game Boards	Logos	Optical Illusion	Patterned Grid	Overall
InternVL3.5-8B	+EnAR	**25.82**	0	**30.00**	**16.07**	**34.78**	**52.27**	**22.32**	**31.36**
	w/o uncertainty map	21.97	0	28.75	12.50	33.57	49.24	20.54	29.03
	w/o visual impression	9.89	0	26.25	10.71	28.57	50.00	21.43	25.80
	w/o ours	0	0	25.0	8.93	5.07	48.48	21.43	19.83
Qwen2.5VL-7B	+EnAR	**18.68**	0	**23.75**	**7.14**	**16.67**	**61.74**	**15.18**	**28.02**
	w/o uncertainty map	15.93	0	21.25	5.36	13.77	60.23	**15.18**	26.25
	w/o visual impression	10.44	0	22.50	5.36	12.32	59.47	14.29	24.50
	w/o ours	0	0	20.00	3.75	12.32	60.23	14.29	22.63
LLaVA1.5-7B	+EnAR	**3.85**	0	**12.50**	10.71	**13.04**	**53.03**	**21.43**	**22.20**
	w/o uncertainty map	2.19	0	10.00	10.71	10.87	51.51	16.96	20.27
	w/o visual impression	1.64	0	10.00	**12.50**	9.42	49.62	2.67	17.78
	w/o ours	0	0	8.75	10.71	8.70	50.0	0	16.92

Table 7. **Ablation study on vision encoder layer selection across different LVLMs.** We evaluate the impact of selecting different layers (0, 1, 6, 10, 16, 20, 23) from the vision encoder on multiple benchmarks, including VLMBias, POPE, and HallusionBench.

Model	Layer	VLMBias	POPE				HallusionBench			Average
			Acc.	Prec.	F1 Score	Average	aACC	qACC	fAcc	
InternVL3.5-8B	0	26.5	88.0	86.0	87.8	87.3	70.3	43.6	45.2	63.91
	1	28.2	88.1	86.3	88.2	87.5	70.8	44.0	46.4	64.57
	6	**31.4**	**88.6**	**87.3**	**89.0**	**88.3**	**71.9**	45.7	**46.8**	**65.81**
	10	29.2	88.3	86.3	88.6	87.7	71.3	**45.8**	46.0	65.07
	16	28.8	88.2	86.4	88.3	87.6	70.4	44.6	46.1	64.69
	20	30.3	88.1	86.2	88.0	87.4	70.6	45.4	46.0	64.94
	23	27.3	88.2	85.9	87.6	87.2	70.1	43.9	45.8	64.11
QwenVL2.5-8B	0	24.1	87.0	96.1	85.7	89.6	71.2	47.9	48.1	65.73
	1	25.4	87.2	96.2	85.7	89.7	71.8	48.0	47.7	66.00
	6	28.0	88.8	**96.6**	87.8	91.1	72.9	48.0	**49.5**	67.37
	10	**28.8**	**88.9**	96.5	**87.9**	**91.1**	**73.0**	**48.3**	48.9	**67.48**
	16	25.1	88.7	96.4	87.4	90.8	72.2	47.5	48.8	66.59
	20	25.2	87.2	96.2	86.0	89.8	72.3	47.1	48.4	66.06
	23	22.1	87.1	96.3	85.8	89.7	72.2	47.2	47.9	65.50
LLaVA1.5-7B	0	17.0	86.2	88.7	85.9	86.9	38.0	17.6	12.4	49.40
	1	17.3	86.7	88.9	86.3	87.3	38.5	18.2	13.0	49.84
	6	**22.2**	**87.1**	**90.5**	**86.6**	**88.1**	**39.3**	**20.0**	**15.4**	**51.59**
	10	19.2	86.5	88.7	86.0	87.1	38.1	19.0	14.1	50.23
	16	18.5	86.1	88.0	85.5	86.5	38.3	18.7	12.8	49.70
	20	19.4	86.3	88.1	85.7	86.7	38.8	18.6	13.8	50.10
	23	16.5	85.7	87.8	85.0	86.2	38.0	18.3	12.1	49.06

D. Additional Ablation Studies

D.1. Key Components Ablation

Table 6 reports the detailed results after removing the two key components in our framework. By comparing EnAR with the setting without the uncertainty map, we observe a consistent drop in the overall score across all backbones, as well as mild but stable decreases in most categories (e.g., *Animals*, *Flags*, *Patterned Grid*). This confirms that the uncertainty map provides a useful auxiliary signal for refining the localization of suspicious regions beyond what the raw attention map can offer.

A more pronounced degradation appears when further removing the visual impression. Under this configuration, the model degenerates into a saliency-driven scheme that focuses on visually prominent regions and lacks an explicit canonical reference view. Lacking a canonical reference view, the model tends to overfit to the original distorted image and fails to reliably distinguish counterfactual elements from normal visual patterns. As a result, performance drops more substantially than without the uncertainty map. Finally, without EnAR corresponds to the baseline that does not apply any processing and simply uses the model's regular decoding, yields the lowest scores, highlighting that uncertainty map and visual impression generation are jointly critical for robust counterfactual hallucination mitigation.

Table 8. **IoU scores on the HKU-IS dataset.** This figure displays the layer-wise IoU scores between the top-K% attention regions and the ground truth salient regions from HKU-IS. The results are reported for K = 5% and K = 10% with two LVLMs, LLaVA-v1.5-7B and InternVL3.5-8B.

Model	K(%)	Layer(0-11)											
		0	1	2	3	4	5	6	7	8	9	10	11
LLaVA-v1.5-7B	5%	0.095	0.158	0.152	0.166	0.163	0.162	0.176	**0.177**	0.137	0.150	0.109	0.060
	10%	0.159	0.272	0.262	0.283	0.284	0.282	**0.313**	0.310	0.245	0.254	0.201	0.120
InternVL3.5-8B	5%	0.117	0.142	0.130	0.124	0.159	0.158	**0.170**	0.140	0.143	0.126	0.118	0.068
	10%	0.200	0.243	0.226	0.218	0.271	0.277	**0.286**	0.254	0.267	0.237	0.233	0.158
Model	**K(%)**	**Layer(12-23)**											
		12	13	14	15	16	17	18	19	20	21	22	23
LLaVA-v1.5-7B	5%	0.057	0.063	0.048	0.058	0.056	0.084	0.074	0.115	0.122	0.111	0.117	0.116
	10%	0.115	0.125	0.127	0.137	0.125	0.185	0.169	0.220	0.224	0.205	0.206	0.211
InternVL3.5-8B	5%	0.007	0.007	0.031	0.038	0.020	0.111	0.051	0.144	0.137	0.114	0.084	0.087
	10%	0.026	0.024	0.071	0.082	0.054	0.209	0.120	0.230	0.225	0.198	0.139	0.185

D.2. Vision Encoder Layer Selection

Table 7 provides a comprehensive ablation study of selecting different vision encoder layers for extracting attention maps. Across all tested LVLMs and benchmarks, layers around the early-mid stage (approximately the 6th layer) consistently achieve the top or near-top overall performance. The trend is remarkably stable. Using very shallow layers (0-1) results in incomplete grounding due to insufficient abstraction, while deeper layers (16-23) gradually deteriorate, reflecting their tendency to encode high-level semantics and global contextual biases rather than localized visual evidence.

To better understand the differences, we compute the IoU score between the top-K% attended regions from each layer and the salient object masks in the HKU-IS dataset [41], as reported in Table 8. The results reveal that early layers achieve substantially higher alignment with ground-truth salient regions. These layers attend more faithfully to concrete object parts, producing spatially precise and semantically localized attention. In contrast, deep layers demonstrate a clear drift toward background areas or peripheral edges, reducing overlap with salient objects. This phenomenon is consistent with prior findings in ViT analysis [17, 33, 46, 81], which show that deeper layers accumulate global information and form tokens that aggregate diffuse context rather than capturing fine-grained visual cues.

Since our framework (EnAR) relies on identifying counterfactual elements embedded in specific visual regions, layers exhibiting stronger spatial localization are preferable. Combining the ablation results and the saliency-IoU evidence, selecting an early-mid layer offers the optimal trade-off. It retains object-centric attention while avoiding the semantic entanglement and drift found in deeper representations. Therefore, we adopt the *6th encoder layer* as the default setting for all experiments. Additional visualizations are provided in Appendix D.4.

D.3. Padding Token Ratio

Table 9 reports the detailed results of varying the padding token ratio from 5% to 50% across three backbones and three benchmarks (VLMBias, POPE, HallusionBench). The last column (i.e., "Average") aggregates all metrics and serves as an overall indicator of hallucination robustness. Across all models, using a small but non-trivial amount of padding (around 10%) yields the best overall performance, while both lower and higher ratios lead to gradual degradation.

InternVL3.5-8B. Increasing the ratio from 5% to 10% improves the average score from 65.40 to 65.81, mainly driven by gains on VLMBias and POPE, while HallusionBench remains stable. Beyond 10%, all metrics start to decline, and the Average drops steadily to 63.84 at 50%. This suggests that a moderate amount of padding is sufficient to cover most counterfactual tokens, whereas excessive padding begins to overwrite useful visual evidence and dilute the model's focus.

QwenVL2.5-8B. QwenVL2.5-8B shows a similar but slightly more robust trend. Performance peaks at a 10% ratio (Average 67.37) and remains relatively flat for 5-20%, with only minor fluctuations in individual metrics. When the ratio is further increased to 30-50%, VLMBias and HallusionBench slowly degrade, leading to an overall Average of 65.39 at 50%. This indicates that QwenVL is less sensitive to the exact padding ratio in the low-to-medium range, but still suffers once padding becomes overly aggressive.

LLaVA1.5-7B. The absolute numbers are lower (reflecting its weaker baseline on counterfactual benchmarks), but the

Table 9. **Ablation study on padding token ratio across different LVLMs.** We evaluate the performance of different padding ratios (5%-50%) on multiple benchmarks including VLMBias, POPE, and HallusionBench.

Model	Ratio	VLMBias	POPE			HallusionBench			Average
			Acc.	Prec.	F1 Score	aACC	qACC	fAcc	
InternVL3.5-8B	5%	30.2	88.5	87.1	88.7	71.6	45.3	46.4	65.40
	10%	**31.4**	**88.6**	87.3	**89.0**	**71.9**	**45.7**	**46.8**	**65.81**
	20%	29.9	88.4	**87.4**	88.9	71.8	45.8	46.5	65.53
	30%	28.7	88.3	87.0	88.5	71.5	45.4	46.5	65.13
	40%	26.9	88.0	86.5	87.9	71.3	45.2	46.1	64.56
	50%	24.6	87.7	85.6	87.0	70.9	44.8	46.3	63.84
QwenVL2.5-8B	5%	27.2	88.7	96.1	87.5	72.3	47.5	48.9	66.89
	10%	**28.0**	**88.8**	96.6	**87.8**	**72.9**	**48.1**	**49.5**	**67.37**
	20%	26.6	88.5	**96.9**	87.7	72.3	**48.1**	48.8	66.97
	30%	24.7	88.6	96.5	87.4	72.1	47.4	48.9	66.51
	40%	23.9	88.0	96.3	86.8	71.6	47.6	48.3	66.07
	50%	23.1	87.2	96.8	85.6	71.0	46.5	47.5	65.39
LLaVA-v1.5-7B	5%	20.1	86.7	90.1	86.4	38.9	19.2	14.1	50.79
	10%	**22.2**	87.1	**90.5**	**86.6**	**39.5**	**20.0**	**15.4**	**51.59**
	20%	20.8	**87.2**	90.4	86.3	**39.5**	19.4	14.2	51.11
	30%	21.1	86.8	89.7	85.9	39.0	19.2	14.3	50.86
	40%	19.4	86.3	88.9	85.0	38.3	19.0	13.7	50.09
	50%	19.2	86.4	88.6	84.5	37.5	18.7	12.7	49.66

relative pattern is the most pronounced. The Average score improves from 50.79 at 5% to 51.59 at 10%, then monotonically decreases as the ratio increases, reaching 49.66 at 50%. Both VLMBias and the HallusionBench metrics benefit from moving from 5% to 10%, confirming that a small amount of padding helps the model reallocate attention away from hallucination-prone regions. However, once too many tokens are padded, the model loses fine-grained visual cues and its answers become less grounded.

Overall, these results show that our framework is *not* extremely sensitive to the exact padding ratio in the 5-20% range, but performance consistently peaks around 10% and degrades when padding becomes too sparse or too dense. Intuitively, too few padding tokens fail to sufficiently mask counterfactual elements, whereas too many tokens behave like aggressive erasing, dispersing information, and hindering visual grounding. Based on this observation, we fix the padding token ratio to 10% for all main experiments.

D.4. Attention Pattern Visualization

Figures 7 and 8 illustrate layer-wise attention maps from LLaVA-v1.5-7B on two representative examples. A clear trend emerges across both cases. In the early layers (0-3), the attention is spatially diffuse and dominated by low-level edges, textures, and background boundaries. These layers provide broad coverage but lack object-specific localization. At layer 3, the attention in Figure 7 heavily concentrates on the distant levee and the outline of the mountaintop, while the attention in Figure 8 focuses on the road curb and large areas of surrounding vegetation. As the depth increases (around layers 4-7), the attention rapidly sharpens and becomes concentrated on semantically meaningful foreground regions, such as the ship in Figure 7 or the vehicles in Figure 8, indicating stronger alignment with salient visual cues.

Beyond the middle layers, the attention gradually drifts away from the main objects and collapses into a few isolated hotspots, often located at peripheral or non-object regions. At the deepest layers (18-23), the maps become almost fully detached from the concrete objects and instead concentrate on background regions. This reflects the emergence of global contextual mixing and register-like tokens reported in prior studies [17]. That is, deep layers aggregate broad semantic information but lose spatial precision. In Figure 7, the deepest layers concentrate their attention on a few background spots scattered across the sea surface.

These visualizations further support the quantitative findings in Appendix D.2. The early-mid layers preserve high-quality grounding and localized object-focused attention, whereas deeper layers exhibit attention drift and reduced alignment with salient regions. This validates our choice of selecting an early-mid layer (i.e., layer 6) for EnAR to ensure robust detection of counterfactual elements.

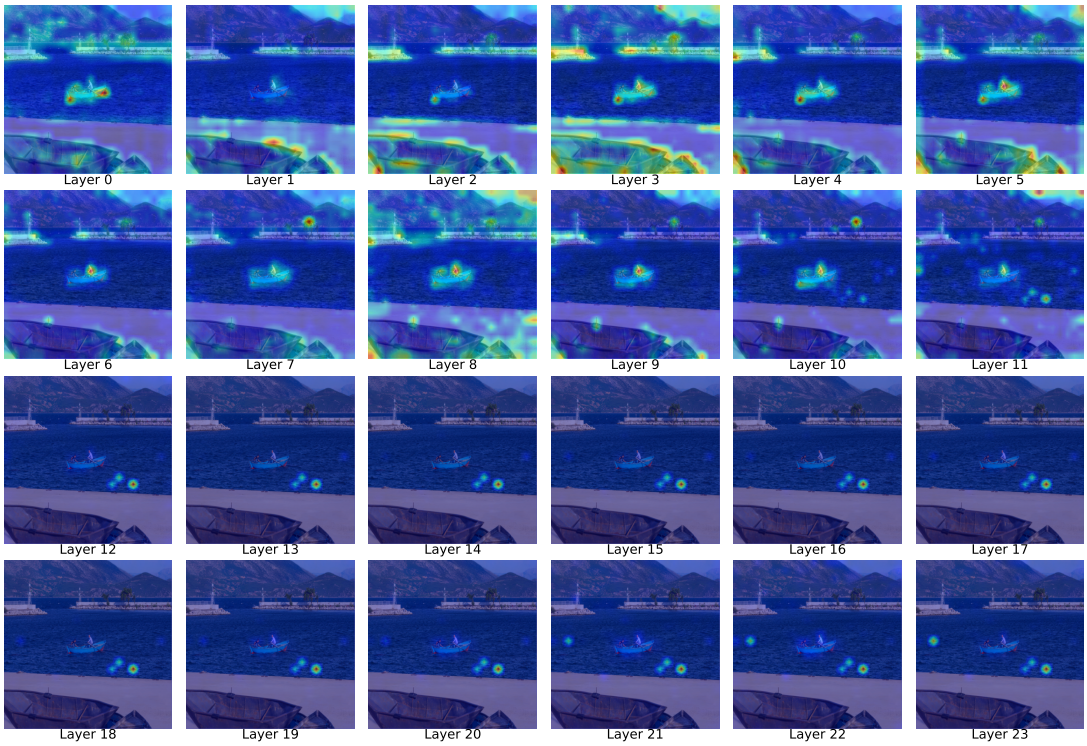

Figure 7. Layer-wise attention maps of the vision encoder on an image of a harbor with ships.

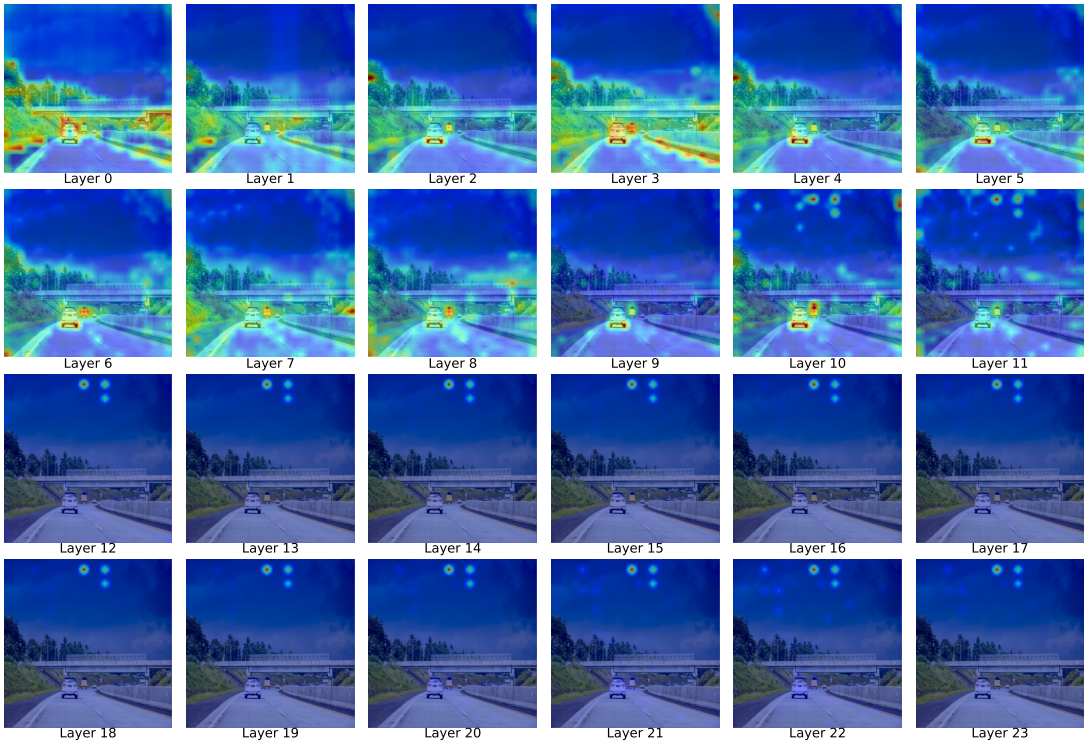

Figure 8. Layer-wise attention maps of the vision encoder on an image of a road with cars.

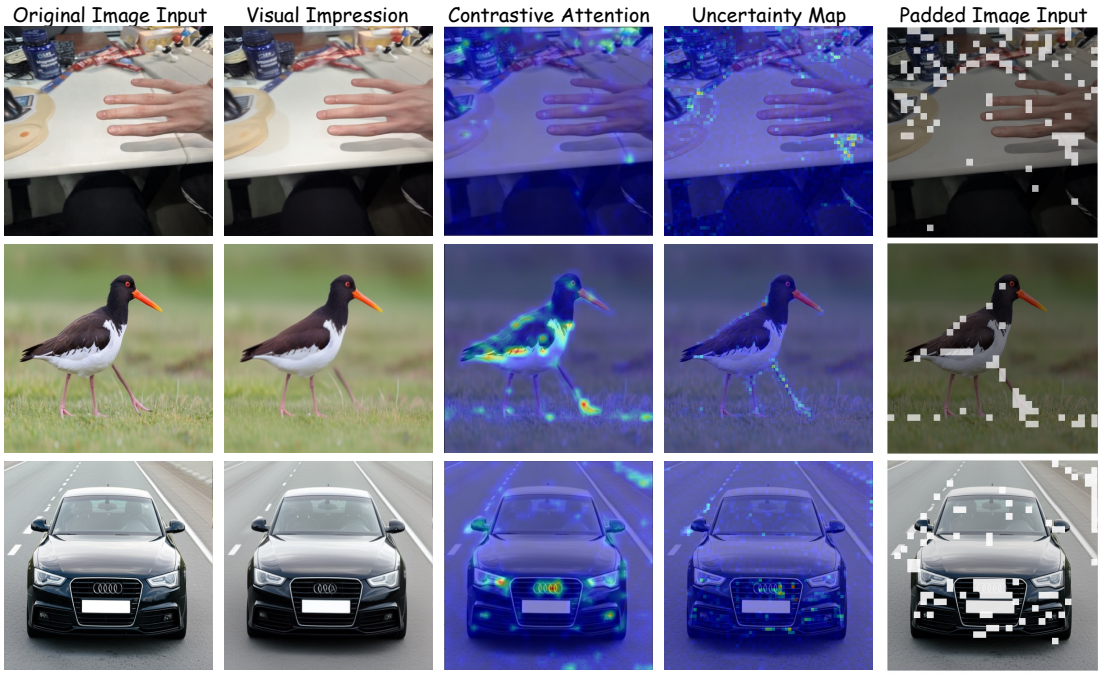

Figure 9. Confounding-detail and failure cases.

Method	Avg. Tokens	Avg. Latency	Peak GPU Memory	Performance
Regular	1846	3.517s(1.00×)	22.44GB	19.83(+0.00)
VCD	3692	5.908s (1.68×)	23.48GB	23.76(+3.93)
DeGF	3692	12.405s (3.53×)	27.19GB	21.29(+1.46)
EnAR	3692	8.386s (2.38×)	27.43GB	31.36(+11.53)

Table 10. **Efficiency comparison across different methods.** We analyze the efficiency and performance of multiple methods on VLMBias, reporting Avg. Tokens, Avg. Latency, Peak GPU Memory, and Performance.

E. Efficiency Analysis and Improvement

As shown in Table 10, VCD achieves 23.76 at 5.908s (1.68×) and 23.48GB, whereas DeGF is much slower (12.405s, 3.53×) with higher memory (27.19GB) and lower performance (21.29). EnAR attains the best performance (31.36, +11.53) with lower latency than DeGF (8.386s, 2.38×) at comparable memory (27.43GB). Although EnAR has higher latency than VCD, it achieves better performance. Compared with diffusion-based DeGF, EnAR delivers higher performance with even lower latency.

To reduce EnAR's inference cost, we consider optimizations on both diffusion and LVLM computation. EnAR already parallelizes visual impression generation by batching DDIM forward and reverse steps with latent perturbations, enabling a single diffusion pass per image. This can be strengthened via larger effective batching, mixed precision, and better reuse of intermediate features. To further alleviate the inference overhead of EnAR, future work could consider replacing the padding of counterfactual region tokens with a direct deletion strategy. By further integrating token pruning techniques to retain only the most critical visual tokens, this approach would effectively streamline the processing pipeline and achieve significant acceleration.

F. Additional Case Study

We add an example with confounding details in the first row of Figure 9, where a person shows four fingers while an extra finger like shadow appears in the background. In such case, the generated visual impression remains close to the original image, while the uncertainty map concentrates around the plausible thumb region, providing a reference for model. In Table 3, EnAR achieves gains on POPE by 6.9%, which is composed of real-world images without counterfactual edits, indicating that the unconditional diffusion prior does not harm performance in such scene.

We also add failure cases where the diffusion prior is imperfect, e.g., an extra leg that is only partly removed and a reconstructed car logo that is blurry. In these cases, the uncertainty map assigns high values exactly on the imperfect regions, and EnAR uses both the visual impression and uncertainty to localize counterfactual elements. EnAR does not rely on a fully corrected visual impression, and the maximum deviation selection, combined with uncertainty, helps avoid amplifying irrelevant noise in complex scenes.