MITIGATING OBJECT HALLUCINATION IN LARGE VISION-LANGUAGE MODELS VIA FAITH-GUIDED SELF-CORRECTING DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Vision-Language Models (LVLMs) have demonstrated strong capabilities in open-ended visual understanding, yet remain prone to object hallucinations, generating objects unsupported by the visual input. Existing training-free decoding methods typically correct next-token predictions using auxiliary visual signals, without explicitly assessing the visual support underlying the prediction. We observe that grounded and hallucination-prone predictions respond differently to targeted evidence weakening: grounded predictions tend to lose more predictive support when their attended visual cues are disturbed. Based on this observation, we introduce *visual faithfulness*, an online measure of visual support derived from intervention-induced prediction sensitivity. We further propose **FaiD**, a training-free decoding framework that uses *visual faithfulness* to continuously guide self-correction during generation. Predictions with stronger visual support are preferentially calibrated using positive evidence, while weakly supported ones receive stronger suppression. Extensive experiments across multiple LVLMs and benchmarks show that FaiD consistently reduces object hallucinations while largely preserving grounded content and open-ended response quality.

1 INTRODUCTION

Large Vision-Language Models (LVLMs) have achieved remarkable progress in multimodal tasks such as image captioning, visual question answering, and visual reasoning (Liu et al., 2024; Brown et al., 2020; Chiang et al., 2023). By integrating visual encoders with powerful large language models, LVLMs can generate fluent responses conditioned on visual inputs. However, LVLMs remain prone to object hallucination (Bai et al., 2024; Gunjal et al., 2024; Yin et al., 2024; Wu et al., 2024; Li et al., 2023b), where generated responses mention objects unsupported by the image. Such visually inconsistent outputs undermine model reliability and hinder deployment in applications requiring faithful visual understanding.

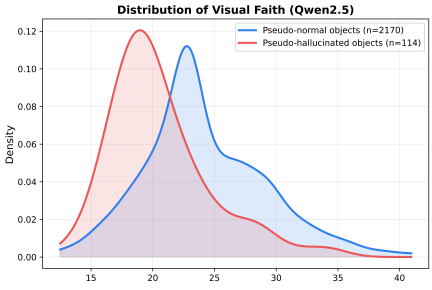


Figure 1: Visual faithfulness distribution on Qwen2.5-VL. Pseudo-grounded tokens tend to exhibit higher faithfulness than pseudo-hallucinated ones, suggesting that visual faithfulness provides a useful signal of visual support.

Recent studies (Huang et al., 2026; Ren et al., 2026; Wang et al., 2026; Zhao et al., 2026; Yin et al., 2026) have attempted to mitigate object hallucinations from both training-based and training-free perspectives. Training-based methods improve visual alignment through preference tuning (Yang et al., 2025c) or post-hoc correction models (Xiao et al., 2025), but often require additional data construction or model updates. Training-free decoding methods (Kim et al., 2024; Li et al., 2023a) instead intervene directly during generation, for example by contrasting predictions under distorted visual inputs (Leng et al., 2024; Park et al., 2025; Wang et al., 2024) or enhancing prompt-relevant visual regions (An et al., 2025). Despite their effectiveness, these methods generally construct an auxiliary signal and use it directly for correction. However, such a signal does not explicitly reveal

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

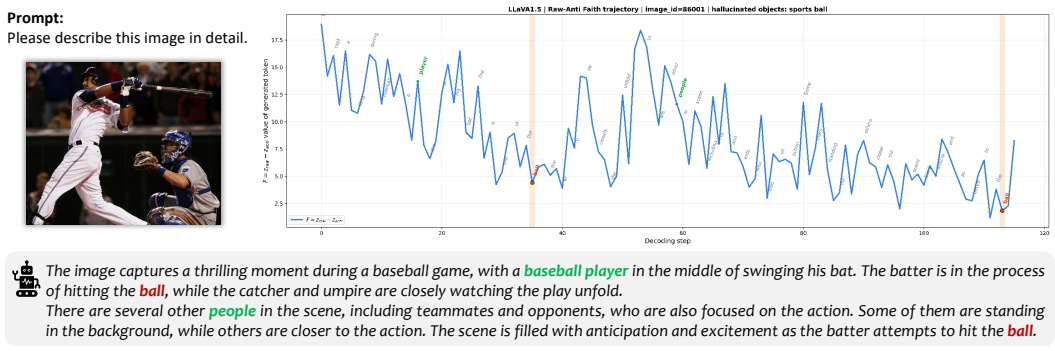


Figure 2: Token-level visual faithfulness trajectory. Faithfulness varies considerably across decoding steps, with hallucinated objects such as *sports ball* often emerging near local valleys, motivating online visual support estimation at each decoding step.

whether the current prediction is actually supported by visual evidence. A high-probability prediction may be visually grounded, weakly supported, or mainly driven by language priors. Without explicitly assessing its visual support, direct correction may either suppress valid predictions or fail to sufficiently correct unsupported ones.

This raises a more fundamental question: *can we estimate how strongly a prediction depends on visual evidence before deciding how to correct it?* Model confidence alone is insufficient, since visually unsupported predictions can still receive high scores due to language priors or contextual associations. Instead, we investigate whether the model's response to targeted visual intervention can reveal its dependence on visual evidence. Intuitively, if a prediction truly relies on specific visual evidence, weakening that evidence should substantially reduce its predictive support. By contrast, predictions driven mainly by non-visual priors should be less sensitive to such intervention. This shifts hallucination mitigation from directly modifying predictions to first probing the visual evidence underlying them.

We empirically examine this hypothesis by selectively weakening the most attended visual features and measuring the resulting prediction change. As shown in Figure 1, pseudo-grounded object predictions generally exhibit larger logit drops than pseudo-hallucinated ones, suggesting that intervention sensitivity reflects the degree of visual support. We term this intervention-induced sensitivity *visual faithfulness*. Moreover, the two distributions overlap substantially, and visual faithfulness varies considerably across decoding steps (Figure 2). These observations suggest modeling visual support as a continuous and dynamic signal rather than a binary grounding decision.

Motivated by these observations, we propose **Fai**th-guided Self-Correcting **D**ecoding (**FaiD**), a training-free decoding framework that uses visual faithfulness to guide correction during autoregressive generation. FaiD dynamically estimates visual support at each decoding step and uses the resulting faithfulness signal as a continuous control variable. Predictions with stronger visual support are primarily calibrated using positive evidence, whereas weakly supported predictions receive stronger faith-based suppression. In this way, FaiD adapts the correction behavior to the estimated degree of visual support rather than applying a uniform correction rule.

Our main contributions are summarized as follows:

- We reveal different intervention sensitivities between grounded and hallucination-prone predictions, suggesting prediction sensitivity as a informative signal of visual support.
- We introduce visual faithfulness and formulate hallucination mitigation as faith-guided correction, where intervention responses continuously guide prediction calibration.
- We propose FaiD, a training-free decoding framework that consistently reduces object hallucination across multiple LVLMs and benchmarks while largely preserving response quality.

2 Related Work

Hallucination in Large Vision-Language Models. Large Vision-Language Models are prone to hallucination, generating objects or attributes unsupported by the visual input. Existing studies in-

vestigate LVLM hallucination through evaluation and detection (Cai et al., 2025; Prabhakaran et al., 2025), generation-time monitoring (Liu et al., 2025; Chang et al., 2025), and internal mechanisms (He et al., 2025; Yang et al., 2025b; Zhou et al., 2025). These works attribute hallucination to factors such as weak visual grounding, biased attention, and modality priors during generation. Recent studies further extend hallucination analysis to fine-grained visual details, videos, and scene text (Sun et al., 2025; Li et al., 2025a; Shu et al., 2025). Beyond explicit hallucination detection, uncertainty and confidence estimation (Li et al., 2025b; Xue et al., 2025; Li et al., 2025c) have also been explored for improving model reliability. However, confidence alone is insufficient for LVLMs, since visually unsupported object predictions may still receive high scores.

Object Hallucination Mitigation. Existing mitigation methods can be broadly divided into training-based and training-free approaches. *Training-based methods* mitigate hallucination through preference alignment, fine-grained supervision, or representation editing (Yang et al., 2025c; Sarkar et al., 2025; Xiao et al., 2025; Yang et al., 2025a; Akl et al., 2026; Wang et al., 2026; Wu et al., 2026). Conversely, *training-free methods* directly intervene inference behavior without updating model parameters. Common approaches include visual contrastive decoding (Leng et al., 2024), attention and activation steering (Cho et al., 2025; Su et al., 2025), and key-value smoothing (Jiang et al., 2026). More recent methods introduce auxiliary generative or diffusion-based references to verify and correct object predictions (Zhao et al., 2025; Zhang et al., 2025; Dastmalchi et al., 2026; Im et al., 2026). Different from feedback-based self-correction methods that rely on auxiliary generative reference, FaiD estimates visual support directly from the model's response to targeted visual intervention. It then uses this signal for continuous faith-guided calibration, rather than making a discrete choice between predefined correction modes.

3 METHOD

We present FaiD, a training-free decoding framework that mitigates object hallucinations by estimating visual support for candidate predictions and adaptively calibrating their logits during autoregressive generation, as illustrated in Figure 3.

3.1 PRELIMINARIES: AUTOREGRESSIVE DECODING

Let I denote the input image, x the textual query, $y_{<t}$ the previously generated tokens, and $\mathcal{V}$ the output vocabulary. For a candidate token $y \in \mathcal{V}$, we define its logit under original visual input as,

$$z_t^{raw}(y) = \text{logit}_\theta(y \mid I, x, y_{<t}), \tag{1}$$

The corresponding next-token probability is,

$$p_\theta(y_t = y \mid I, x, y_{<t}) \propto \exp z_t^{raw}(y). \tag{2}$$

Standard decoding generates the next token from the normalized distribution induced by these logits. However, a high raw logit does not necessarily imply reliable visual support, since visually unsupported candidates may still be favored by language priors or contextual associations. We therefore examine candidate responses to targeted visual intervention.

3.2 TOKEN-CONDITIONED VISUAL SUPPORT PROBING

To estimate the visual support of a candidate prediction, FaiD performs a targeted intervention on the visual features emphasized by the current decoding state. The intuition is that weakening relevant visual cues should reduce the score of predictions that genuinely depend on them, while predictions driven mainly by language priors should be less affected. We therefore construct an evidence-weakened condition at each decoding step and observe the resulting model response.

Let $V = \{v_1, \ldots, v_N\}$ denote the visual feature tokens extracted by the LVLM. At decoding step t, we extract the attention from the current decoding token to all visual tokens from the last transformer layer and average it over all heads, yielding the visual attention scores A_t. We then select the most attended visual tokens as,

$$\mathcal{S}_t = \text{TopK}\left(A_t, \lceil KN \rceil\right), \tag{3}$$

where $K \in (0, 1]$ denotes the perturbation ratio. Importantly, we do not interpret attention magnitude itself as visual faithfulness. Instead, it is used only to localize the visual features on which the current decoding state places greater emphasis, thereby defining a step-dependent intervention region.

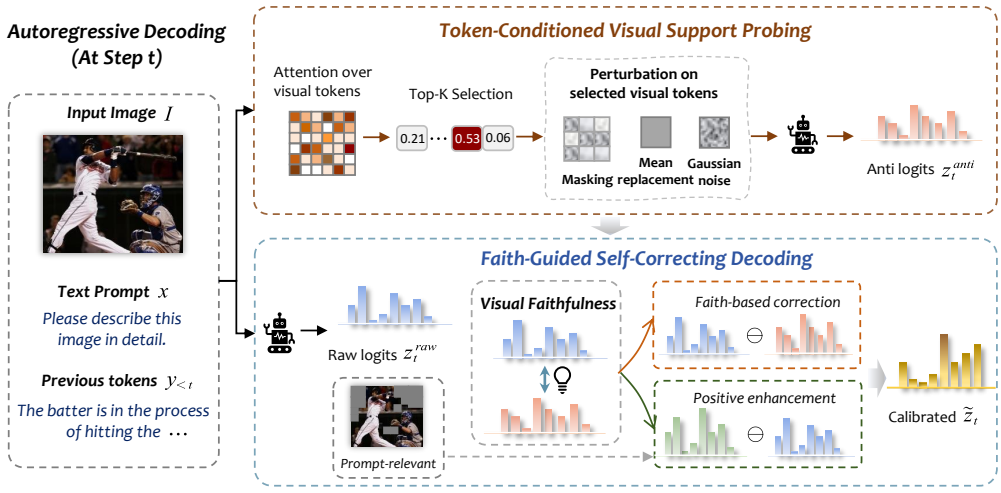


Figure 3: Framework of FaiD. A token-conditioned intervention probes the visual support underlying the current prediction, and the resulting visual faithfulness continuously guides positive-evidence calibration and faith-based suppression during decoding.

We then selectively weaken the visual evidence in $\mathcal{S}_t$ while keeping the remaining visual features unchanged. Specifically, the intervened visual representation $V_t^{anti} = \{v_{t,1}^{anti}, \ldots, v_{t,N}^{anti}\}$ is constructed as,

$$v_{t,n}^{anti} = \begin{cases} \mathbf{0}, & n \in \mathcal{S}_t, \quad \text{Masking}, \\ \bar{v}, & n \in \mathcal{S}_t, \quad \text{Mean Replacement}, \\ v_n + \epsilon, & n \in \mathcal{S}_t, \quad \text{Gaussian Noise}, \\ v_n, & n \notin \mathcal{S}_t, \end{cases} \tag{4}$$

where $\bar{v} = \frac{1}{N}\sum_{n=1}^{N} v_n$ is the mean visual representation and $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. Unless otherwise specified, we adopt Gaussian noise as the default perturbation strategy, as it weakens the selected evidence without completely removing the corresponding visual context. Different intervention operators are compared in Sec. 4.3.

The intervened representation V_t^{anti} is then fed into the LVLM while keeping the textual query x and previously generated tokens $y_{<t}$ unchanged,

$$z_t^{anti}(y) = \text{logit}_\theta(y \mid V_t^{anti}, x, y_{<t}), \tag{5}$$

where $z_t^{anti}(y)$ denotes the logit assigned to candidate y under the weakened visual condition. Since $\mathcal{S}_t$ is recomputed from the current decoding state, the intervention visual region adapts dynamically across decoding steps. Crucially, the evidence-weakened prediction is not directly used for contrastive decoding. Instead, it serves as an intervention probe for assessing how predictions respond to weakened visual evidence. The resulting change forms the basis for visual support estimation in the next section.

3.3 Faith-Guided Self-Correcting Decoding

Given the raw and evidence-weakened logits, we define the intervention-based visual faithfulness score of candidate $y \in \mathcal{V}$ at decoding step t as,

$$\phi_t(y) = z_t^{raw}(y) - z_t^{anti}(y), \tag{6}$$

where $z_t^{raw}(y)$ and $z_t^{anti}(y)$ denote the candidate logits under the original and evidence-weakened visual conditions, respectively. Although the intervention region $\mathcal{S}_t$ is shared by all candidates at step t, the induced logit change is candidate-specific, yielding a distinct faithfulness score $\phi_t(y)$ for each $y \in \mathcal{V}$. A larger $\phi_t(y)$ indicates that weakening the attended visual evidence causes a larger logit drop, suggesting stronger dependence on the original visual input. Conversely, a smaller value indicates weaker sensitivity and potentially greater reliance on non-visual priors. Thus, visual faithfulness measures intervention sensitivity, rather than prediction confidence or attention magnitude. We further compare these control signals in Sec. 4.3.

Since grounded and hallucination-prone predictions exhibit overlapping faithfulness distributions (Figure 1), we treat visual support as a continuous quantity rather than a binary decision. We thus map $\phi_t(y)$ to a continuous faith gate,

$$\rho_t(y) = \sigma\left(\frac{\phi_t(y) - m}{\tau}\right), \tag{7}$$

where m is a faithfulness margin and τ controls transition sharpness. Based on this faith signal, FaiD performs faith-guided logit calibration rather than applying the same correction to all candidates.

For candidates with stronger visual support, we adopt AGLA's prompt-guided enhancement (An et al., 2025) to construct an augmented view I^{aug} as positive visual evidence, whose contribution is adaptively controlled by visual faithfulness. Decoding with I^{aug} yields,

$$z_t^{aug}(y) = \text{logit}_\theta(y \mid I^{aug}, x, y_{<t}), \tag{8}$$

where $z_t^{aug}(y)$ denotes the candidate logit under the augmented visual input. We define the positive-evidence enhancement term as,

$$\Delta_t^{enh}(y) = z_t^{aug}(y) - z_t^{raw}(y). \tag{9}$$

It measures the logit change induced by strengthened visual evidence. FaiD combines positive-evidence calibration with faith-based suppression according to the estimated visual support of each candidate,

$$\tilde{z}_t(y) = z_t^{raw}(y) + \rho_t(y)\Delta_t^{enh}(y) + (1 - \rho_t(y))\Delta_t^{faith}(y), \tag{10}$$

where

$$\Delta_t^{faith}(y) = \phi_t(y) - m. \tag{11}$$

When a candidate receives stronger visual support, $\rho_t(y)$ approaches 1 and its calibration is dominated by the positive-evidence term. Conversely, when the estimated support is weak, $\rho_t(y)$ decreases and the faith-based term becomes more dominant. Since $\phi_t(y) - m < 0$ for insufficiently supported candidates, their logits are selectively suppressed.

This formulation turns visual faithfulness from a diagnostic signal into an explicit control variable for decoding. Rather than directly using the evidence-weakened prediction as an alternative decoding distribution, FaiD uses the intervention response to determine how subsequent correction should be applied. As a result, candidates with stronger visual support are primarily calibrated by positive evidence, while weakly supported candidates receive stronger suppression.

Adaptive Plausibility Constraints. To avoid promoting implausible candidates, we restrict the calibrated distribution to tokens that retain sufficient probability under the raw prediction. Specifically, we define the plausible candidate set as,

$$\mathcal{V}_t^{plaus} = \left\{ y \in \mathcal{V} : p_\theta(y \mid I, x, y_{<t}) \geq \beta \max_\omega p_\theta(\omega \mid I, x, y_{<t}) \right\}, \tag{12}$$

where $\beta \in [0, 1]$ controls the plausibility threshold. The final FaiD decoding distribution is then,

$$p_{FaiD}\left(y_t = y \mid I, x, y_{<t}\right) = \begin{cases} \text{Softmax}(\tilde{z}_t(y)), & y \in \mathcal{V}_t^{plaus}, \\ 0, & y \notin \mathcal{V}_t^{plaus}. \end{cases} \tag{13}$$

The next token is then sampled from this calibrated distribution, ensuring that decoding remains within the plausible candidate set while incorporating faith-guided correction.

4 Experiments

4.1 Experimental Settings

Evaluated LVLMs. We conduct experiments with three publicly available LVLMs: LLaVA1.5-7B (Liu et al., 2024), Qwen2.5-VL-7B (Bai et al., 2025), and InternVL3-9B (Zhu et al., 2025). These models differ in architecture design, providing a diverse evaluation for robustness of our method. We use LLaVA1.5 for the most extensive baseline comparison, as many prior methods provide implementations on this backbone, while Qwen2.5-VL serves as the primary model for analysis. InternVL3 is further included to evaluate generalization across stronger LVLMs.

Datasets. POPE (Li et al., 2023b) and MME-Hallucination (Fu et al., 2025) are discriminative benchmarks, where models are required to answer object-related questions. POPE formulates object existence verification as binary yes or no questions, while MME provides diagnostic evaluation. CHAIR (Rohrbach et al., 2018) and LLaVA-Bench-Wild are adopted as generative benchmarks. CHAIR measures hallucinated object mentions in free-form image captions by matching generated objects with MSCOCO annotations. LLaVA-Bench-Wild evaluates open-ended response quality on real-world images and questions, where both accuracy and description detailedness are considered.

Metrics. For POPE, we report standard classification metrics, including accuracy, precision, and F1 score. For CHAIR, we report the conventional instance-level C_I and sentence-level C_S metrics, together with object recall and response length. For MME-Hallucination, we report the total scores for existence, count, position, and color. For LLaVA-Bench-Wild, we use GPT-4V-based evaluation and report accuracy and detailedness scores ranging from 1 to 10.

Compared methods. Regular Decoding denotes the LVLM generates responses from its original next-token distribution. On LLaVA1.5, we compare with representative training-free methods, including VCD (Leng et al., 2024), AGLA (An et al., 2025), ED (Cho et al., 2025), and DeGF (Zhang et al., 2025). On Qwen2.5-VL and InternVL3, we additionally reproduce VCD and AGLA when their model-agnostic components can be directly transferred.

Implementation. We use the official checkpoints of each evaluated model. The maximum decoding step is 1024 for generative tasks. Unless otherwise specified, we set β=0.5 and τ=1.0. The effects of the faithfulness margin m and perturbation ratio K are further analyzed in the Appendix B.3 and B.4. Experimental analysis are conducted on CHAIR using Qwen2.5-VL.

4.2 Main Results

Results on CHAIR. We first evaluate FaiD on open-ended caption generation, where hallucinated objects can naturally arise during autoregressive decoding. Table 1 shows that FaiD consistently reduces both C_S and C_I across all three LVLMs. On LLaVA1.5, FaiD reduces C_S from 51.0 to 44.8 and C_I from 15.2 to 14.1, while slightly improving recall from 75.2 to 77.0. The improvement is more pronounced on the stronger Qwen2.5-VL and InternVL3 models. In particular, C_S decreases from 24.4 to 19.8 on Qwen2.5-VL and from 32.0 to 23.6 on InternVL3. Compared with AGLA, FaiD also achieves consistently lower hallucination rates on all three backbones. Notably, on Qwen2.5-VL, FaiD generates longer responses than regular decoding while substantially reducing hallucination, indicating that the improvement does not result from conservative shortening. On InternVL3, the reduction in hallucination is accompanied by a moderate decrease in recall, reflecting a more conservative trade-off. Overall, these results show that faith-guided calibration remains effective across different LVLM architectures and open-ended generation behaviors.

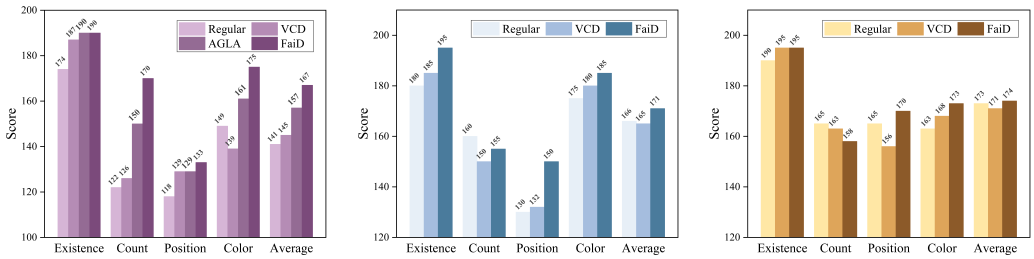
Table 1: Results on CHAIR benchmark. † denotes the results of deploying this method to Qwen2.5-VL and InternVL3.

LVLMs	Method	$C_S\downarrow$	$C_I\downarrow$	Recall↑	Length
LLaVA1.5	Regular	51.0	15.2	75.2	102.2
	VCD	51.0	14.9	77.2	101.9
	AGLA	49.2	<u>14.8</u>	<u>77.6</u>	101.0
	ED	53.0	15.2	**78.7**	101.0
	DeGF	<u>48.8</u>	15.1	73.9	98.4
	FaiD	**44.8**	**14.1**	77.0	98.9
Qwen2.5-VL	Regular	24.4	9.0	55.1	134.5
	VCD†	27.4	9.2	**58.4**	146.8
	AGLA†	<u>23.8</u>	<u>9.0</u>	<u>55.9</u>	140.7
	FaiD	**19.8**	**6.6**	54.4	146.8
InternVL3	Regular	32.0	10.0	66.6	120.5
	VCD†	31.4	10.7	**67.7**	116.8
	AGLA†	<u>25.4</u>	<u>9.5</u>	<u>65.2</u>	110.3
	FaiD	**23.6**	**7.2**	64.5	109.8

Results on POPE. We further evaluate discriminative object-existence hallucination on POPE under random, popular, and adversarial negative sampling. As shown in Table 2, FaiD achieves competitive performance across the three datasets and sampling settings, with particularly consistent improvements in precision. The gains are more evident under popular and adversarial sampling, where negative objects are more likely to be associated with the visual context and thus induce false-positive predictions. FaiD remains effective under these challenging settings, indicating that faith-guided calibration can suppress object responses driven mainly by contextual associations rather than reliable visual evidence. Similar trends are observed on Qwen2.5-VL and InternVL3, further demonstrating the robustness of FaiD across different LVLMs.

Table 2: Results on POPE benchmark. Higher (↑) accuracy, precision, and F1 indicate better performance. The best results are **bolded**, and the second-best are underlined. † represents the results of deploying this method to Qwen2.5-VL and InternVL3.

LVLMs	Setup	Method	Ref.	MS-COCO			A-OKVQA			GQA		
				Acc.↑	Prec.↑	F1↑	Acc.↑	Prec.↑	F1↑	Acc.↑	Prec.↑	F1↑
LLaVA1.5	Ran.	Regular	-	83.13	81.94	83.44	81.90	76.63	83.53	82.23	76.32	84.03
		VCD	CVPR2024	87.00	86.13	87.15	83.83	78.05	85.34	83.23	76.73	85.05
		AGLA	CVPR2025	87.47	97.47	86.00	89.00	93.01	88.46	88.60	92.32	88.08
		ICT	CVPR2025	**89.70**	93.06	**89.28**	88.03	84.98	88.53	87.70	83.05	88.21
		ED	ICLR2025	89.30	96.24	88.43	89.53	91.64	**89.41**	89.00	92.32	88.36
		DeGF	ICLR2025	89.03	91.20	88.74	86.93	84.28	87.42	87.40	83.51	88.09
		Self-Aug	ICLR2026	86.89	96.26	85.42	87.94	90.89	87.49	88.18	91.02	87.75
		FaiD	ours	87.60	**97.64**	86.14	**89.67**	**93.43**	89.20	**89.33**	**92.88**	**88.87**
	Pop.	Regular	-	81.17	78.28	82.08	75.07	68.58	78.77	73.47	66.83	77.84
		VCD	CVPR2024	83.10	79.96	83.94	76.63	69.59	80.19	72.37	65.27	77.58
		AGLA	CVPR2025	86.07	94.13	84.67	85.60	86.53	85.42	83.10	82.39	83.28
		ICT	CVPR2025	86.77	87.49	86.64	80.77	75.01	82.75	74.37	67.31	78.70
		ED	ICLR2025	**88.00**	92.73	**87.30**	86.23	83.74	**86.72**	83.67	81.52	84.20
		DeGF	ICLR2025	86.63	87.75	86.28	80.90	75.68	82.66	78.10	71.65	80.98
		Self-Aug	ICLR2026	85.07	91.99	83.72	83.55	83.06	83.68	81.87	80.14	82.38
		FaiD	ours	86.43	**94.83**	85.03	**86.53**	**87.43**	86.40	**84.30**	**83.69**	**84.44**
	Adv.	Regular	-	77.43	73.31	79.26	67.23	61.56	73.70	68.60	62.43	74.84
		VCD	CVPR2024	77.17	72.18	79.47	67.40	61.39	74.21	68.83	62.26	75.43
		AGLA	CVPR2025	83.83	89.31	82.62	78.87	75.96	79.99	80.50	78.22	81.26
		ICT	CVPR2025	82.57	80.51	83.13	72.13	65.76	76.82	72.17	65.27	77.29
		ED	ICLR2025	**84.93**	87.27	**84.45**	**80.10**	75.27	**81.84**	80.83	77.00	**82.10**
		DeGF	ICLR2025	81.63	80.59	81.94	72.70	66.70	76.86	74.07	67.42	78.22
		Self-Aug	ICLR2026	82.47	86.74	81.39	77.02	73.60	78.57	78.23	75.14	79.49
		FaiD	ours	84.30	**90.16**	83.06	79.33	**76.19**	80.50	**81.17**	**78.84**	81.90
Qwen2.5-VL	Ran.	Regular	-	85.66	98.08	83.53	88.39	94.83	87.49	88.62	94.66	87.80
		VCD†	CVPR2024	**88.26**	98.07	**86.92**	90.49	94.99	89.99	89.77	**94.48**	89.19
		FaiD	ours	88.13	**98.72**	86.69	**90.77**	**95.23**	**90.04**	**90.13**	94.33	**89.64**
	Pop.	Regular	-	84.61	95.24	82.56	86.19	90.14	85.47	84.41	86.51	83.95
		VCD†	CVPR2024	**88.26**	**98.07**	**86.92**	88.23	90.53	87.89	85.47	86.27	85.30
		FaiD	ours	86.83	95.55	85.44	**88.37**	**90.96**	**87.99**	**86.93**	**88.10**	**86.73**
	Adv.	Regular	-	83.49	92.69	81.49	80.02	79.49	80.20	81.92	82.07	81.88
		VCD†	CVPR2024	**88.26**	92.08	**84.31**	81.56	79.72	82.11	83.43	82.61	83.64
		FaiD	ours	85.53	**92.44**	84.25	**82.27**	**80.63**	**82.73**	**83.53**	**82.97**	**83.82**
InternVL3	Ran.	Regular	-	91.84	94.81	91.57	90.83	90.66	91.07	88.86	90.55	88.62
		VCD†	CVPR2024	91.70	93.90	91.49	91.37	90.99	91.41	89.34	90.47	89.19
		FaiD	ours	**93.07**	**95.62**	**92.87**	**93.40**	**93.00**	**93.43**	**91.40**	**92.13**	**91.32**
	Pop.	Regular	-	87.90	87.43	87.98	87.28	84.38	87.79	85.43	84.50	85.63
		VCD†	CVPR2024	**91.70**	87.22	88.20	86.57	83.09	87.24	86.07	84.75	86.33
		FaiD	ours	89.43	**88.79**	**89.52**	**88.77**	**85.18**	**89.31**	**87.63**	**85.57**	**87.98**
	Adv.	Regular	-	83.40	80.34	84.20	81.56	76.33	83.22	81.28	78.00	82.31
		VCD†	CVPR2024	83.94	80.76	84.74	81.78	76.24	83.52	82.27	78.90	83.24
		FaiD	ours	**85.27**	**82.10**	**85.96**	**83.57**	**77.86**	**85.09**	**83.90**	**79.96**	**84.89**



(a) LLaVA1.5 (b) Qwen2.5-VL (c) InternVL3

Figure 4: Results on MME-Hallucination with different decoding methods. Higher scores indicate better performance.

Results on MME-Hallucination. To examine whether FaiD generalizes beyond object-existence hallucination, we further evaluate it on the four MME-Hallucination subsets. As shown in Figure 4, FaiD consistently improves the average score across all three LVLMs. On LLaVA1.5, FaiD achieves

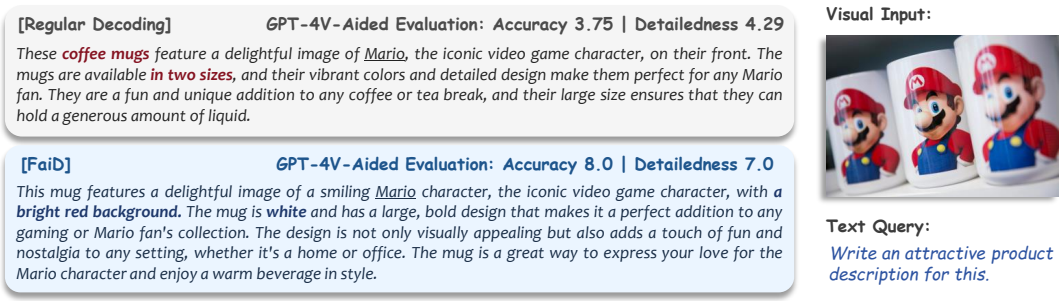


Figure 5: Cases on LLaVA-Bench-Wild using LLaVA-1.5. Hallucinated and accurate captions are highlighted in **red** and **blue**.

clear gains on count and color. These subsets require more fine-grained visual discrimination, suggesting that FaiD can also improve predictions involving object quantity, spatial relations, and visual attributes. The consistent gains across different LVLMs further indicate that the faith-guided calibration is not limited to a specific model architecture or hallucination type.

Results on LLaVA-Bench-Wild. We evaluate whether hallucination reduction comes at the cost of open-ended response quality. GPT-4V-aided evaluation in Table 3 shows that FaiD improves both accuracy and detailedness on LLaVA1.5 and Qwen2.5-VL. On InternVL3, FaiD achieves the highest accuracy, while maintaining competitive detailedness. As illustrated in Figure 5, Regular Decoding introduces unsupported descriptions such as *coffee mugs* and *two sizes*, whereas FaiD generates a more visually grounded description by correctly focusing on the single white Mario mug and the red background. This suggests that FaiD not only reduces object hallucination, but also improves the factuality of open-ended visual descriptions.

Table 3: GPT-4V-aided evaluation on LLaVA-Bench-Wild. † denotes the results of deploying this method to Qwen2.5-VL and InternVL3.

LVLMs	Method	Acc.↑	Detail.↑
LLaVA1.5	Regular	4.65	5.06
	VCD	4.73	5.34
	FaiD	**6.07**	**5.97**
Qwen2.5-VL	Regular	5.59	5.47
	VCD†	5.41	5.32
	FaiD	**6.58**	**6.30**
InternVL3	Regular	7.35	8.27
	VCD†	7.16	**8.49**
	FaiD	**7.38**	8.19

4.3 Mechanism Analysis

Q1: Does Visual Faithfulness Provide an Effective Control Signal?

To examine whether visual faithfulness serves as an effective control signal, we compare it with three commonly used alternatives: *confidence*, *negative entropy*, and *visual attention*. These signals respectively capture model certainty, distributional uncertainty, and attended visual evidence. For a fair comparison, all signals are normalized and mapped through the same gating function while keeping the correction formulation unchanged. As shown in Table 4 and Figure 6, visual faithfulness achieves the lowest C_S and C_I, with only a slight decrease in recall. It also produces the longest responses among all control signals, indicating that the hallucination reduction is not caused by conservative shortening. Overall, these results suggest that intervention-induced visual faithfulness provides a more effective basis for adaptive correction.

Table 4: CHAIR results of different control signals.

Control Signal	C_S↓	C_I↓	Recall↑	Length
Confidence	21.4	7.2	55.2	133.8
Negative Entropy	20.6	7.6	55.4	130.1
Attention	20.4	6.7	54.9	137.2
Visual Faithfulness	19.8	6.6	54.4	146.8

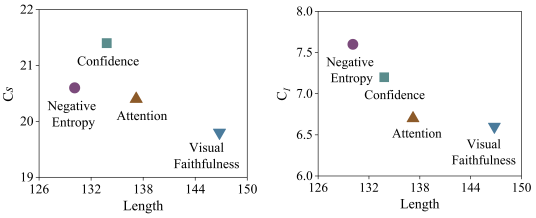


Figure 6: Hallucination vs. Response length.

Q2: Does Targeted Visual Intervention Matter?

❶ *Intervention Localization.* We first examine whether the effectiveness of visual probing depends on where the intervention is applied. Global perturbs all visual tokens, Random-K perturbs randomly selected visual tokens, while Step-wise Top-K dynamically targets the most attended visual tokens at each decoding step. As shown in Table 5, Step-wise Top-K achieves the lowest hallucination rates, indicating that effective visual support probing benefits from dynamically targeting step-specific visual evidence.

Table 5: CHAIR results of visual intervention.

Intervention	$C_S\downarrow$	$C_I\downarrow$	Recall↑	Length
Intervention Localization				
Global	22.2	7.5	56.3	141.6
Random-K	22.4	8.0	55.0	146.3
Step-wise Top-K	19.8	6.6	54.4	146.8
Intervention Operator				
Zero Masking	23.2	6.9	55.8	142.0
Mean Replacement	23.0	7.9	55.2	146.4
Gaussian Noise	19.8	6.6	54.4	146.8

❷ *Intervention Operator.* Given the dynamically selected visual evidence, we further compare different weakening operators. Table 5 shows that Gaussian noise yields the lowest hallucination rates, while zero masking slightly improves recall but results in a higher C_S. This suggests that softly corrupting the evidence provides a more informative probing signal than completely removing or deterministically replacing it. We further analyze the perturbation ratio K in Appendix B.4.

Q3: Why Use Faith-conditioned Correction?

❶ *Direct Correction.* We examine whether the intervention response should be directly used for correction or serve as a control signal. As shown in Table 6, both *Pos. Only* and *Direct Contrast.* reduce hallucination compared with regular decoding. However, their fixed combination performs worse, suggesting that uniformly applying both corrections is suboptimal for candidates with different levels of visual support.

Table 6: Comparison of direct and faith-conditioned correction strategies under the same visual intervention on CHAIR.

Setting	Gate ρ_t	$C_S\downarrow$	$C_I\downarrow$	Recall↑	Length
Regular	-	24.4	9.0	55.1	134.5
Pos. Only	1	22.4	7.7	57.5	127.2
Direct Contrast.	0	22.2	6.9	54.5	143.5
Fixed Dual	0.5	22.8	10.8	55.8	135.8
Hard Gate	$\mathbb{I}(\mathcal{F}_t > m)$	21.8	7.6	55.2	136.6
Linear Gate	Linear	19.0	6.7	54.9	145.5
FaiD	Sigmoid	19.8	6.6	54.4	146.8

❷ *Continuous Gating.* We further compare different faith-conditioned strategies. Hard gating improves over fixed fusion, while the continuous Linear and Sigmoid gates achieve substantially lower hallucination rates. Their comparable performance indicates that the key benefit comes from continuous faith-conditioned calibration rather than a particular gating function.

Q4: What Does FaiD Cost?

FaiD incurs non-negligible latency overhead due to additional probing and enhancement passes. We report the average inference latency and peak GPU memory. Experiments are conducted on a single NVIDIA A800 GPU. Compared with stronger multi-branch decoding methods on LLaVA1.5, Table 7 shows that FaiD achieves the lowest hallucination rate 44.8% with comparable latency to ED. The memory overhead is limited, with peak memory increasing by about $1.07\times\sim1.23\times$. These results demonstrate a favorable hallucination-efficiency trade-off, although latency-sensitive deployment remains a limitation.

Table 7: Efficiency comparison on CHAIR.

LVLMs	Method	Latency (s) ↓	Memory (MB) ↓	$C_S\downarrow$
LLaVA1.5	Regular	5.5(×1.00)	17352(×1.00)	51.0
	VCD	7.4(×1.35)	15954(×0.92)	51.0
	AGLA	9.4(×1.71)	20050(×1.16)	49.2
	ED	18.4(×3.35)	18584(×1.07)	53.0
	FaiD	18.6(×3.38)	21306(×1.23)	44.8
Qwen2.5-VL	Regular	18.1(×1.00)	18334(×1.00)	24.4
	VCD	28.7(×1.59)	18506(×1.01)	27.4
	FaiD	31.9(×1.76)	22032(×1.20)	19.8
InternVL3	Regular	14.5(×1.00)	20204(×1.00)	32.0
	VCD	20.5(×1.41)	18834(×0.93)	31.4
	FaiD	21.2(×1.46)	21560(×1.07)	23.6

5 Conclusion

We propose FaiD, a training-free decoding method for mitigating object hallucination in LVLMs. FaiD estimates token-level visual faithfulness from the model response to dynamically weakened visual evidence and uses it as a soft control signal for decoding. Experiments across multiple LVLMs and benchmarks demonstrate consistent reductions in object hallucination.

AI USE DISCLOSURE

In this work, we used generative AI tools to assist with language polishing, improving the organization and readability of the manuscript, and providing feedback on the presentation of the methodology and experimental design. We also used generative AI to assist with the interpretation and presentation of experimental results, as well as the refinement of figures, section titles, and keywords. All AI-assisted content was carefully reviewed and revised by the authors. The authors independently verified the technical formulations, experimental results, and scientific claims, and take full responsibility for the final content of this work.

REFERENCES

Ahmed Akl, Abdelwahed Khamis, Ali Cheraghian, Zhe Wang, Sara Khalifa, and Kewen Wang. HIME: mitigating object hallucinations in lvlms via hallucination insensitivity model editing. *CoRR*, abs/2602.18711, 2026.

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, Qianying Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *CVPR*, pp. 29915–29926, 2025.

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

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *CoRR*, abs/2404.18930, 2024.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020.

Yishuo Cai, Renjie Gu, Jiaxu Li, Xuancheng Huang, Junzhe Chen, Xiaotao Gu, and Minlie Huang. MHALO: evaluating mllms as fine-grained hallucination detectors. In *ACL (Findings)*, pp. 9197–9222, 2025.

Yurui Chang, Bochuan Cao, and Lu Lin. Monitoring decoding: Mitigating hallucination via evaluating the factuality of partial response during generation. In *ACL (Findings)*, pp. 14574–14587, 2025.

Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

Yeongjae Cho, Keonwoo Kim, Taebaek Hwang, and Sungzoon Cho. Do you keep an eye on what I ask? mitigating multimodal hallucination via attention-guided ensemble decoding. In *ICLR*, 2025.

Hamidreza Dastmalchi, Aijun An, Ali Cheraghian, and Hamed Barzamini. Fighting hallucinations with counterfactuals: Diffusion-guided perturbations for LVLM hallucination suppression. *CoRR*, abs/2603.10470, 2026.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. In *NeurIPS*, 2025.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *AAAI*, pp. 18135–18143, 2024.

Jinghan He, Kuan Zhu, Haiyun Guo, Junfeng Fang, Zhenglin Hua, Yuheng Jia, Ming Tang, Tat-Seng Chua, and Jinqiao Wang. Cracking the code of hallucination in lvlms with vision-aware head divergence. In *ACL (1)*, pp. 3488–3501, 2025.

Yiyang Huang, Liang Shi, Yitian Zhang, Yi Xu, and Yun Fu. Shield: Suppressing hallucinations in lvlm encoders via bias and vulnerability defense. In *ICLR*, volume 2026, pp. 125123–125143, 2026.

Eun Woo Im, Muhammad Kashif Ali, and Vivek Gupta. Self-aug: Query and entropy adaptive decoding for large vision-language models. In *ICLR*, volume 2026, pp. 136345–136368, 2026.

Siyu Jiang, Feiyang Chen, Xiaojin Zhang, and Kun He. Kvsmooth: Mitigating hallucination in multi-modal large language models through key-value smoothing. In *CVPR*, pp. 32561–32571, 2026.

Junho Kim, Hyunjun Kim, Yeonju Kim, and Yong Man Ro. CODE: contrasting self-generated description to combat hallucination in large multi-modal models. In *NeurIPS*, 2024.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *CVPR*, pp. 13872–13882, 2024.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *CVPR*, pp. 13723–13733, 2025a.

Jiaqi Li, Yixuan Tang, and Yi Yang. Know the unknown: An uncertainty-sensitive method for LLM instruction tuning. In *ACL (Findings)*, pp. 2972–2989, 2025b.

Rui Li, Jing Long, Muge Qi, Heming Xia, Lei Sha, Peiyi Wang, and Zhifang Sui. Towards harmonized uncertainty estimation for large language models. In *ACL (1)*, pp. 22938–22953, 2025c.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *ACL (1)*, pp. 12286–12312, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, pp. 292–305, 2023b.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pp. 26286–26296, 2024.

Zihang Liu, Jiawei Guo, Hao Zhang, Hongyang Chen, Jiajun Bu, and Haishuai Wang. Long-form hallucination detection with self-elicitation. In *ACL (Findings)*, pp. 4082–4100, 2025.

Yeji Park, Deokyeong Lee, Junsuk Choe, and Buru Chang. Convis: Contrastive decoding with hallucination visualization for mitigating hallucinations in multimodal large language models. In *AAAI*, pp. 6434–6442, 2025.

Vishnu Prabhakaran, Purav Aggarwal andVinay Kumar Verma, Gokul Swamy, and Anoop Saladi. VADE: visual attention guided hallucination detection and elimination. In *ACL (Findings)*, pp. 14949–14965, 2025.

Hairui Ren, Zixuan Wang, Yibo Yang, He Zhao, Fan Tang, Dandan Guo, and Yi Chang. Imitating the truth: Attention-aware truth-guided enhancement for hallucination mitigation in large vision-language models. In *ICLR*, volume 2026, pp. 43994–44020, 2026.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *EMNLP*, pp. 4035–4045, 2018.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö. Arik, and Tomas Pfister. Mitigating object hallucination in mllms via data-augmented phrase-level alignment. In *ICLR*, 2025.

Yan Shu, Hangui Lin, Yexin Liu, Yan Zhang, Gangyan Zeng, Yan Li, Yu Zhou, Ser Nam Lim, Harry Yang, and Nicu Sebe. When semantics mislead vision: Mitigating large multimodal models hallucinations in scene text spotting and understanding. In *NeurIPS*, 2025.

Jingran Su, Jingfan Chen, Hongxin Li, Yuntao Chen, Li Qing, and Zhaoxiang Zhang. Activation steering decoding: Mitigating hallucination in large vision-language models through bidirectional hidden state intervention. In *ACL (1)*, pp. 12964–12974, 2025.

Xiaoxi Sun, Jianxin Liang, Yueqian Wang, Huishuai Zhang, and Dongyan Zhao. Understanding visual detail hallucinations of large vision-language models. In *IJCAI*, pp. 1900–1908, 2025.

Tianbo Wang, Yuqing Ma, Kewei Liao, Zhange Zhang, Simin Li, Jinyang Guo, and Xianglong Liu. AFTER: mitigating the object hallucination of LVLM via adaptive factual-guided activation editing. *CoRR*, abs/2601.01957, 2026.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *ACL (Findings)*, pp. 15840–15853, 2024.

Jialin Wu, Wei Shi, Han Shen, Peigui Qi, Kunsheng Tang, Zhicong Huang, Binghao Wang, and Zhou Yang. Revis: Sparse latent steering to mitigate object hallucination in large vision-language models. *CoRR*, abs/2602.11824, 2026.

Mingrui Wu, Jiayi Ji, Oucheng Huang, Jiale Li, Yuhang Wu, Xiaoshuai Sun, and Rongrong Ji. Evaluating and analyzing relationship hallucinations in lvlms. *CoRR*, abs/2406.16449, 2024.

Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Fangxun Shu, Hao Jiang, and Linchao Zhu. Detecting and mitigating hallucination in large vision language models via fine-grained AI feedback. In *AAAI*, pp. 25543–25551, 2025.

Boyang Xue, Fei Mi, Qi Zhu, Hongru Wang, Rui Wang andSheng Wang, Erxin Yu, Xuming Hu, and Kam-Fai Wong. Ualign: Leveraging uncertainty estimations for factuality alignment on large language models. In *ACL (1)*, pp. 6002–6024, 2025.

Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection. In *CVPR*, pp. 14635–14645, 2025a.

Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention. In *ICLR*, 2025b.

Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. Mitigating hallucinations in large vision-language models via DPO: on-policy data hold the key. In *CVPR*, pp. 10610–10620, 2025c.

Jianghao Yin, Qin Chen, Kedi Chen, Jie Zhou, Xingjiao Wu, and Liang He. Dynamic multimodal activation steering for hallucination mitigation in large vision-language models. In *ICLR*, 2026.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: hallucination correction for multimodal large language models. *Sci. China Inf. Sci.*, 67(12), 2024.

Ce Zhang, Zifu Wan, Zhehan Kan, Martin Q. Ma, Simon Stepputtis, Deva Ramanan, Russ Salakhutdinov, Louis-Philippe Morency, Katia P. Sycara, and Yaqi Xie. Self-correcting decoding with generative feedback for mitigating hallucinations in large vision-language models. In *ICLR*, 2025.

Jianfei Zhao, Feng Zhang, Xin Sun, Chong Feng, and Zhixing Tan. Tell model where to look: Mitigating hallucinations in mllms by vision-guided attention. In *CVPR*, pp. 32582–32591, 2026.

Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. Mitigating object hallucination in large vision-language models via image-grounded guidance. In *ICML*, 2025.

Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. In *ICLR*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *CoRR*, abs/2504.10479, 2025.

MITIGATING OBJECT HALLUCINATIONS IN LARGE VISION-LANGUAGE MODELS VIA FAITH-GUIDED SELF-CORRECTING DECODING

APPENDIX

In this supplementary document, we provide additional details and experimental results to enhance understanding and insights into our method. This supplementary document is organized as follows:

- Detailed benchmarks and metrics are in Section A.
- Additional experimental results are in Section B.
- More case studies and GPT-4V-aided evaluations are provided in Section C.
- The limitations and broader impacts are discussed in Section D.
- Potential directions for future work are discussed in Section E.

A BENCHMARKS AND METRICS

We conduct experiments on the following benchmarks:

- **POPE** is a widely adopted benchmark for evaluating object hallucination in LVLMs. It assesses whether models can correctly answer binary object-existence questions, such as "Is there a {object} in the image ?" POPE is built upon three datasets, including MS-COCO, A-OKVQA, and GQA. For each dataset, it provides three evaluation settings, i.e., random, popular, and adversarial, which differ in the way negative samples are constructed. Each setting contains 3,000 test instances, generated from 6 questions for each image. We report accuracy, precision, recall, and F1 score on this benchmark.
- **CHAIR** is designed to measure object hallucination in open-ended image captioning. It asks LVLMs to generate descriptions for 500 randomly sampled images from the MSCOCO validation set, and evaluates whether the generated object mentions are supported by the ground-truth annotations. Following the standard protocol, we report two hallucination metrics,

$$\mathrm{C}_I = \frac{|\{hallucinated\ objects\}|}{|\{all\ generated\ objects\}|}$$

$$\mathrm{C}_S = \frac{|\{captions\ with\ hallucinated\ objects\}|}{|\{all\ generated\ captions\}|}$$

  We additionally report recall and the average length of generated responses.
- **MME-Hallucination** provides a fine-grained evaluation of hallucination in LVLMs. It contains existence and count for perception , as well as position and color for attribute-level hallucination. In addition, it also includes posters, celebrity, scene, landmark, artwork, and OCR subsets for perception hallucination, and Commonsense Reasoning, Numerical Calculation, Text Translation, and Code Reasoning for cognition hallucination. Each subset consists of 30 images and 60 yes-or-no questions, with two questions associated with each image. Similar to POPE, the evaluation is based on binary answer accuracy. Following the official protocol, the final score combines accuracy and accuracy+, where accuracy is computed over individual questions, and accuracy+ requires both questions of the same image to be answered correctly.
- **LLaVA-Bench-Wild** contains 24 images covering diverse and challenging visual scenarios, including complex scenes, memes, paintings, and sketches, together with 60 difficult questions. We use this benchmark for qualitative comparison across different decoding strategies. Following VCD, we further adopt GPT-4V as an evaluator to score the generated responses in terms of accuracy and detailedness.

Table 8: Results on CHAIR benchmark under different maximum number of new tokens. Lower (↓) C_S, C_I and higher (↑) recall and length indicate better performance.

LVLMs	Method	C_S ↓	C_I ↓	Recall ↑	Length
Maximum Decoding Number = 128					
LLaVA1.5	Regular	49.8	15.1	75.8	97.5
	VCD	53.6	16.0	75.3	97.4
	FaiD	46.6	13.7	76.0	95.5
Qwen2.5-VL	Regular	22.2	7.8	51.3	103.5
	VCD	23.8	8.8	52.7	103.6
	FaiD	17.6	7.0	50.0	102.4
InternVL3	Regular	28.0	10.6	63.9	99.6
	VCD	27.8	9.9	65.2	97.8
	FaiD	22.4	7.2	63.3	99.9
Maximum Decoding Number = 256					
LLaVA1.5	Regular	50.2	15.3	76.7	100.6
	VCD	57.0	17.8	75.9	103.9
	FaiD	44.8	14.1	77.0	98.9
Qwen2.5-VL	Regular	25.4	7.4	55.1	129.2
	VCD	28.2	9.5	57.9	140.4
	FaiD	20.4	6.6	54.3	128.3
InternVL3	Regular	28.6	9.1	65.5	110.7
	VCD	31.8	10.9	68.0	116.1
	FaiD	24.2	7.4	64.7	110.2

B More Experimental Results and Analysis

B.1 Additional Results on CHAIR

To examine whether FaiD remains effective under different response-length budgets, we conduct experiments on CHAIR with maximum decoding lengths of 128 and 256. As shown in Table 8, FaiD consistently achieves the lowest hallucination rates under both settings. When the maximum length is 128, FaiD reduces C_S and C_I to 46.6 and 13.7 on LLaVA1.5, to 17.6 and 7.0 on Qwen2.5-VL, and to 22.4 and 7.2 on InternVL3, respectively. With a longer decoding budget of 256, FaiD still obtains the best hallucination performance, e.g., reducing C_S from 50.2 to 44.8 on LLaVA1.5 and from 25.4 to 20.4 on Qwen2.5-VL. These results indicate that FaiD can suppress hallucination-prone object tokens across different response lengths while maintaining comparable recall and caption length. To ensure a fair comparison, we uniformly set the maximum decoding length to 1024 in this paper.

B.2 Additional Results on MME-Hallucination

We further evaluate FaiD on a broader set of MME-Hallucination subsets in Table 9. FaiD achieves consistent gains on most perception subsets across different LVLMs, including posters, scene, landmark, artwork, and OCR. For example, on LLaVA1.5, FaiD improves the scores on posters, celebrity, scene, landmark, and artwork over regular decoding, and also obtains the best cognition scores on commonsense reasoning, numerical calculation, and code reasoning. On Qwen2.5-VL and InternVL3, FaiD maintains strong performance on both perception and cognition subsets, with clear improvements on OCR and code reasoning. These results suggest that applying FaiD does not substantially compromise broader perception and reasoning performance, while providing gains on several subsets.

B.3 Effect of Faithfulness Margin

Table 10 studies the influence of the faithfulness margin m. Q denotes the percentile of the Raw-Anti Faith distribution computed from pseudo-grounded object tokens on a calibration set. For example, $Q50$ means that m is set to the 50-*th* percentile. The results show that the optimal margin varies

Table 9: Results on MME-Hallucination subsets. CR: Commonsense Reasoning, NC: Numerical Calculation, TT: Text Translation, CoR: Code Reasoning. Higher scores (↑) indicate better performance.

LVLM	Method	Perception ↑						Cognition ↑			
		Posters	Celebrity	Scene	Landmark	Artwork	OCR	CR	NC	TT	CoR
LLaVA1.5	Regular	128.6	129.1	155.8	160.5	114.0	**132.5**	110.0	45.0	57.5	55.0
	VCD	119.7	122.9	146.0	126.3	109.0	105.0	99.3	45.0	**60.0**	50.0
	FaiD	**143.9**	**133.2**	**158.5**	**166.3**	**116.3**	130.0	**117.9**	**47.5**	57.5	**57.5**
Qwen2.5-VL	Regular	162.9	139.8	154.3	178.3	143.5	**162.5**	152.1	145.0	177.5	162.5
	VCD	169.6	140.0	154.0	**179.5**	139.3	135.7	139.5	**145.0**	177.5	147.5
	FaiD	**169.7**	**144.1**	**155.5**	175.8	**143.5**	155.0	**152.1**	137.5	**177.5**	**170.0**
InternVL3	Regular	172.2	**162.9**	145.8	176.5	156.3	192.5	137.9	155.0	**200.0**	167.5
	VCD	168.7	160.3	149.5	174.8	156.3	192.5	138.6	147.5	177.5	170.0
	FaiD	**172.8**	156.8	**155.0**	**177.5**	**158.0**	**200.0**	**140.0**	**155.0**	192.5	**170.0**

Table 10: Results on CHAIR benchmark under different faithfulness margin. Lower (↓) C_S, C_I and higher (↑) recall and length indicate better performance. Gray background indicates our method's default values.

LVLMs	m	C_S ↓	C_I ↓	Recall ↑	Length
LLaVA1.5	0	50.0	15.2	77.2	100.9
	$Q10$	46.6	13.9	76.8	102.4
	$Q20$	48.2	14.2	76.8	101.9
	$Q30$	46.4	14.6	76.0	100.2
	$Q40$	44.6	14.5	75.5	98.8
	$Q50$	44.8	14.1	77.0	98.9
Qwen2.5-VL	0	22.6	7.2	57.0	138.5
	$Q10$	21.0	7.1	54.4	134.1
	$Q20$	22.6	7.2	57.0	138.5
	$Q30$	22.4	7.7	54.3	142.2
	$Q40$	23.0	7.2	54.9	137.9
	Q50	19.8	6.6	54.4	146.8
InternVL3	0	23.6	7.4	64.9	109.6
	$Q10$	24.2	7.3	65.3	110.8
	$Q20$	23.6	7.2	64.5	109.8
	$Q30$	24.8	7.7	65.2	111.5
	$Q40$	24.6	7.5	66.0	110.4
	$Q50$	25.0	7.5	65.6	109.1

across LVLMs, reflecting their different logit scales and faithfulness distributions. For LLaVA1.5, $Q50$ achieves the lowest C_S and C_I, reducing them to 44.8 and 14.1. For Qwen2.5-VL, $Q50$ also performs best, achieving $C_S = 19.8$ and $C_I = 6.6$. In contrast, InternVL3 obtains the lowest C_S around $m = 0$ or $Q20$, while larger margins tend to increase hallucination rates. These results suggest that the faithfulness margin controls the trade-off between suppressing weakly grounded tokens and preserving valid object mentions, and should be calibrated according to each LVLM.

B.4 Intervention Strength

We finally study the perturbation ratio K, which controls the amount of selected visual evidence being weakened. As shown in Figure 7, $K = 0.5$ yields the lowest C_S while maintaining a low C_I. Smaller perturbations may provide insufficient intervention, whereas overly large ratios can corrupt useful visual context. We therefore use $K = 0.5$ as the default setting.

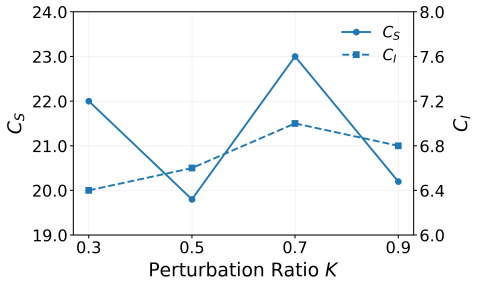


Figure 7: Perturbation ratio results.

JUDGE_PROMPT:

You are an AI designed to evaluate and score the performance of two AI assistants in describing a given image. Your primary focus is on the accuracy and detailedness of their descriptions. You will assess the accuracy by checking for hallucinations—any part of the description that is inconsistent with the image content. For detailedness, you will consider how rich the response is in necessary details, excluding any hallucinated parts. You will first provide scores on a scale from 1 to 10 for each assistant separately, based on these criteria, where a higher score indicates better overall performance. After scoring, you will offer an explanation for your evaluation, ensuring it is free from bias and not influenced by the order of presentation of the responses.Please strictly follow the output format.

Input format:

[Question]
{question}
[Start of Assistant 1' Answer]
{ref_answer}
[End of Assistant 1]
[Start of Assistant 2' Answer]
{model_answer}
[End of Assistant 2]

Output format:

Accuracy:
Scores of the two answers:
Reason:
Detailedness:
Scores of the two answers:
Reason:

Figure 8: Prompt template that we provided to GPT-4V to evaluate the LVLM response based on accuracy and detailedness.

B.5 Construction of the Linear Faithfulness Gate

To provide a controlled baseline for the sigmoid faithfulness gate, we additionally consider a piecewise-linear gate,

$$\rho_t^{linear} = clip\left[\frac{\mathcal{F}_t - m + \delta}{2\delta}, 0, 1\right], \tag{14}$$

where δ controls the transition width around the same faithfulness margin m. Rather than tuning δ independently, we match the local sensitivity of the linear and sigmoid gates at $\mathcal{F}_t = m$. The sigmoid gate has a slope of $1/(4\tau)$ at the margin, whereas the linear gate has a slope of $1/(2\delta)$. Equating the two gives $\delta = 2\tau$. Thus, both gates share the same center and local transition rate, isolating the effect of the gating function itself. Despite its simpler piecewise-linear form, the linear gate performs comparably to the sigmoid gate (Table 6), further indicating that FaiD mainly benefits from continuous faith-conditioned calibration rather than the particular functional form of the gate.

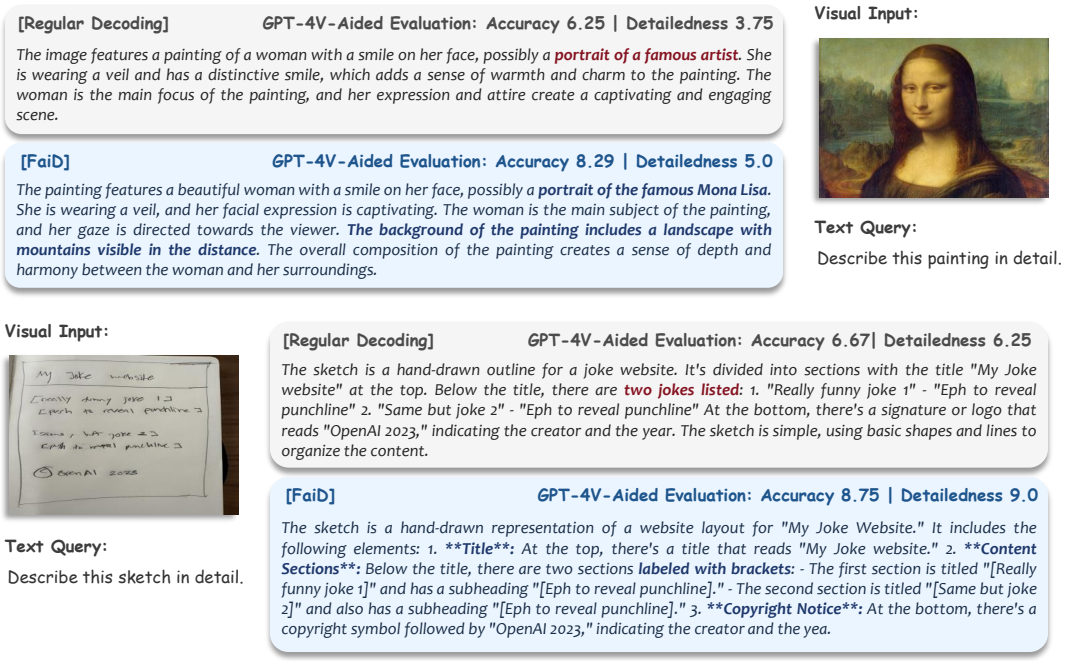


Figure 9: Case studies on the LLaVA-Bench benchmark. We compare the responses generated by regular decoding and our method FaiD. GPT-4V-aided evaluation results are also provided alongside the responses. Hallucinated and accurate captions are highlighted in **red** and **blue**.

C More Case Studies

C.1 Details about GPT-4V-Aided Evaluation

Following the evaluation protocol used in VCD, we adopt GPT-4V as an automatic judge for open-ended generation results. Benefiting from its strong multimodal understanding ability, GPT-4V can identify visual inconsistencies such as incorrect objects, colors, spatial relations, and unsupported descriptions, thus providing a comprehensive assessment of response quality. The prompt used for evaluation is provided in Figure 8. The model is asked to compare the generated responses with the corresponding visual input and assign scores from 1 to 10 in terms of accuracy and detailedness:

- **Accuracy** reflects how well the generated response matches the visual content of the image. A response receives a lower score if it contains image-inconsistent information, such as non-existent objects, wrong attributes, or incorrect relationships.
- **Detailedness** measures the richness and specificity of the generated response. A higher score is assigned when the response provides comprehensive descriptions, includes fine-grained visual details, and offers well-supported explanations. In contrast, vague, overly brief, or insufficiently informative responses receive lower scores.

C.2 More Qualitative Results

As illustrated in Figure 9, we provide additional case studies on LLaVA-Bench to qualitatively demonstrate the effectiveness of our methods in mitigating hallucinations. We also included GPT-4V evaluations of accuracy and detailedness scores for each instance.

D Limitations and Broader Impacts

Limitations. Although FaiD effectively mitigates object hallucinations in LVLMs, it still has several limitations. First, it introduces additional inference overhead, since it requires extra positive-evidence and anti-evidence decoding branches to estimate token-level visual faithfulness. While

our efficiency analysis shows that the overhead is moderate, this may still affect its applicability in latency-sensitive scenarios. Second, the effectiveness of FaiD depends on the quality of visual attention used for constructing the dynamic anti-evidence view. If the attention map fails to localize visual evidence related to the current token, the anti-evidence branch may provide an imperfect faithfulness signal. Third, FaiD may not fully address other hallucination types such as fine-grained attributes, spatial relations, or complex reasoning errors. Extending faith-guided self-correction to broader visual concepts remains an important direction for future work.

Broader Impacts. This work aims to improve the reliability of LVLMs by reducing object hallucinations during generation. By using token-level visual faithfulness as a decoding-time signal, FaiD encourages LVLMs to generate object descriptions that are more consistent with the visual input, which can benefit applications requiring trustworthy multimodal understanding, such as visual assistants, content analysis, and decision-support systems. Since FaiD is training-free and plug-and-play, it can be applied to existing LVLMs without additional data collection or model parameter updates, making it broadly applicable and environmentally more economical than large-scale retraining. However, hallucination mitigation methods should not be viewed as complete safeguards. LVLMs may still produce inaccurate or misleading responses, especially in complex or high-stakes scenarios. Therefore, FaiD should be used as one component of a broader reliability framework that includes careful evaluation and task-specific safety constraints.

E Future Work

In future work, we plan to extend FaiD beyond object-level hallucination to finer-grained visual inconsistencies, such as attributes, spatial relations, counting, and scene text. It would also be valuable to design more efficient faithfulness estimation strategies to reduce the overhead of multi-branch decoding and improve applicability to real-time LVLM systems. In addition, exploring more reliable evidence localization beyond attention maps may further strengthen the quality of dynamic anti-evidence construction.