ATTENTION-ENTROPY DECODING WITH VISUAL GROUNDING FOR HALLUCINATION MITIGATION IN MULTIMODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) have made significant progress in vision–language reasoning, yet they remain prone to hallucination even when relevant visual information is already present in intermediate layers. Existing mitigation methods either require additional training or address the problem during decoding through visual reinforcement, contrastive prediction, or layer-wise correction. However, accurately identifying which intermediate layer contains useful visual information remains challenging. We hypothesize that the concentration of visual attention provides an effective signal: low-entropy attention indicates concentrated and reliable visual information, whereas high-entropy attention suggests dispersed and unreliable information. Based on this insight, we propose AEDG, a training-free attention-entropy decoding framework that uses visual information to guide the decoding process. At each decoding step, AEDG selects a low-entropy layer as a positive anchor and a high-entropy layer as a negative anchor, and uses them to adjust the final-layer prediction. AEDG requires neither parameter updates nor additional forward computation. Extensive experiments on representative MLLMs show that AEDG effectively reduces hallucination while maintaining strong multimodal capability with only a small inference overhead.

1 INTRODUCTION

Multimodal large language models (MLLMs) have advanced vision–language reasoning, but they still hallucinate objects, attributes, and relations that are absent from the image (Bai et al., 2024). This failure limits their reliability in image captioning and visual question answering. As illustrated in Figure 1, the correct answer can already be decoded from an intermediate layer, yet the final layers may override it with a wrong answer. The visual evidence is therefore not necessarily missing; it may simply fail to control the final prediction. This failure is especially common when the image contains multiple objects, background regions, or ambiguous visual cues. In such cases, the model may attend to the image while still forming a diffuse and unreliable visual representation, and the language prior can then fill in the missing details with a plausible but unsupported answer. Thus, the key problem is not only whether the model looks at the image, but whether its visual evidence is organized around the relevant content at the moment of prediction.

Existing hallucination mitigation strategies can be broadly divided into training-time and inference-time methods. Training-based approaches improve cross-modal alignment through additional supervision or preference optimization (Wang et al., 2024a; Chen et al., 2025; Yu et al., 2025; Zhang et al., 2025), but require extra data and computation. Inference-time approaches avoid modifying the checkpoint and include contrastive decoding, response revision, and logit calibration (Leng et al., 2024; Wang et al., 2024b; Qu et al., 2025; Dong et al., 2025). Recent training-free methods instead strengthen visual evidence during decoding: MemVR (Zou et al., 2024) retraces visual information in memory space, AIR (Zhu et al., 2026) selectively reinforces relevant visual tokens, and DeCo (Wang et al., 2025) corrects the final prediction with logits from a dynamically selected intermediate layer. These methods are efficient and annotation-free, but their effectiveness depends on identifying which visual evidence should be trusted at each decoding step.

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

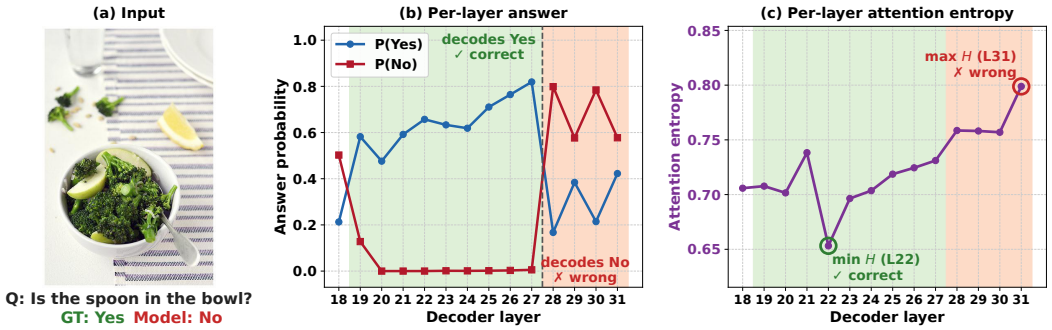


Figure 1: **(a)** The input and the answer. **(b)** The per-layer answer read out with the shared head: the mid layers decode the correct answer, the answer flips at layer 28, and the final output is the wrong one. **(c)** The normalized visual attention entropy H_l of the same step, with its minimum and maximum marked. The lowest-entropy layer ($l = 22$) is the layer that decodes the correct answer, and the highest-entropy layer ($l = 31$) is the one that produces the wrong one, consistent with dispersed visual attention indicating weakened grounding.

Layer-wise decoding offers a natural way to recover visual evidence by combining the final-layer prediction with an intermediate-layer prediction (Chuang et al., 2024; Wang et al., 2025). Unlike DeCo, which selects an intermediate layer from vocabulary-level probabilities, AEDG reads the structure of visual attention itself. A layer may assign substantial attention to the image while spreading it across irrelevant regions; therefore, attention magnitude alone does not reveal whether the layer has formed a reliable visual representation.

Based on this observation, we propose **AEDG** (**A**ttention-**E**ntropy **D**ecoding with visual **G**rounding), a training-free framework that highlights relevant visual evidence while suppressing dispersed and redundant signals. At each decoding step, AEDG computes the normalized entropy of attention within the visual token span. It selects the lowest-entropy layer as a positive anchor and the highest-entropy layer as a negative anchor, then uses both to correct the final-layer logits. A confidence gate prevents unreliable positive evidence from being over-amplified, while top-k-then-top-p filtering limits the correction to plausible candidate tokens. The layer pair is recomputed at every step, with no parameter updates or additional forward pass.

To validate AEDG, we conduct extensive experiments on three representative MLLMs: LLaVA-1.5-7B (Liu et al., 2024), Qwen-VL-Chat (Bai et al., 2023), and GLM-4V-9B (Glm et al., 2024). AEDG consistently reduces hallucination on CHAIR (Rohrbach et al., 2018) and POPE (Li et al., 2023), while preserving or improving general multimodal capability on object- and attribute-level tasks, as summarized in Figure 4. These results demonstrate that AEDG provides an effective, efficient, and training-free solution for hallucination mitigation across diverse MLLMs.

Our contributions are:

- We identify visual-attention concentration as a reliable layer-wise signal for visual grounding and hallucination detection.
- We propose AEDG, an efficient entropy-driven dual-anchor decoding method with confidence-aware gating, requiring no training or additional forward pass.
- Through extensive experiments on three representative MLLMs and multiple hallucination and general-purpose benchmarks, we demonstrate that AEDG consistently reduces hallucination while preserving broad multimodal capability.

2 Related Work

2.1 Hallucination Mechanisms in MLLMs

Object hallucination, namely describing objects absent from the image, is a well-established failure mode in MLLMs (Li et al., 2023; Rohrbach et al., 2018). Prior work attributes it mainly to the

progressive dominance of language priors. DeCo (Wang et al., 2025) shows that correct objects are decodable in middle layers, whereas hallucinated tokens emerge near the output. OPERA (Huang et al., 2024) connects hallucination to attention aggregation, and VCD (Leng et al., 2024) shows that models can remain confident even when visual input is corrupted or removed. These findings suggest that visual evidence may remain available internally but fail to control the final prediction.

2.2 Hallucination Mitigation in MLLMs

Existing methods can be grouped into three categories. Training-based methods improve supervision or cross-modal alignment (Wang et al., 2024a; Chen et al., 2025; Yu et al., 2025; Zhang et al., 2025), but require additional data and training. Post-hoc methods revise or filter responses (Qu et al., 2025; Dong et al., 2025; Zhang et al., 2026), at the cost of extra models or generations. Training-free decoding methods intervene directly in generation: VCD and ICD use contrastive inputs (Leng et al., 2024; Wang et al., 2024b), OPERA penalizes over-confident tokens (Huang et al., 2024), and MemVR and AIR reinforce visual evidence (Zou et al., 2024; Zhu et al., 2026). Our method belongs to the last category and requires neither training nor an auxiliary model.

2.3 Layer-wise Decoding Methods

Layer-wise decoding combines the final-layer prediction with intermediate-layer predictions that may retain stronger visual evidence (Chuang et al., 2024; Wang et al., 2025; Huo et al., 2024; Wang et al., 2026). Existing methods select layers from vocabulary-level signals, as in DoLa and DeCo, or from attention-map statistics. LaSCD (Cao et al., 2026), a concurrent work, uses the Laplacian energy of the attention map, while AEDG uses the normalized entropy of attention within the visual token span. Both methods move beyond attention magnitude, but they capture different properties of the attention distribution. AEDG then uses the selected high- and low-entropy layers as negative and positive anchors in a single-pass logit correction. The shared layer-wise decoding framework enables a direct comparison of these selection signals.

3 Method

3.1 Preliminaries: Layer-wise Decoding in MLLMs

A decoder-only MLLM maps an image v and a text query x to a sequence of visual tokens $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K]$ and to text embeddings, which are concatenated into a single input sequence

$$X = [\, \boldsymbol{Z},\ x_1, \ldots, x_L \,], \tag{1}$$

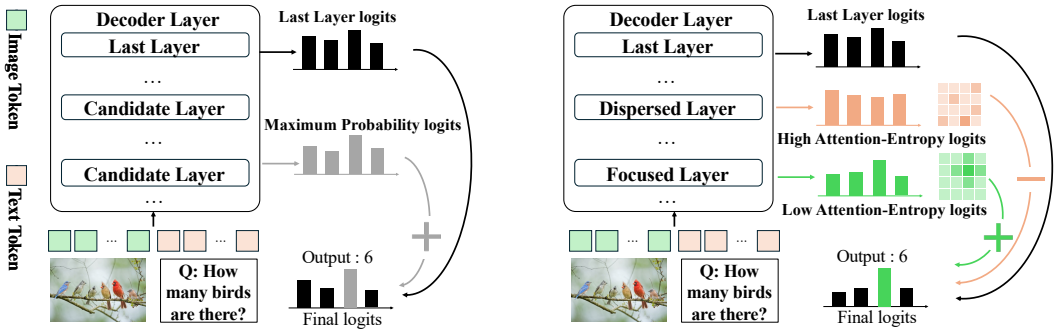
where K is the number of visual tokens and L the query length. The backbone is a stack of N Transformer layers. At decoding step t, let $\boldsymbol{h}_t^{(l)}$ denote the hidden state of the last input position produced by layer l. All layers share a single language head, so the logits and the next-token distribution read off layer l are

$$\boldsymbol{z}_t^{(l)} = \boldsymbol{W}\, \boldsymbol{h}_t^{(l)}, \qquad p_t^{(l)} = \mathrm{softmax}\big(\boldsymbol{z}_t^{(l)}\big), \tag{2}$$

with $l \in \{1, \ldots, N\}$ and $\boldsymbol{W}$ the (shared) unembedding matrix. Standard decoding uses only $l = N$; layer-contrastive methods additionally read one or more intermediate layers with $l < N$.

3.2 Motivating Observation

Figure 1 illustrates the phenomenon that motivates AEDG. At a single decoding step, the correct answer is already decoded by an intermediate layer, but is overridden by the final layer. The layer producing the correct answer has concentrated visual attention and low entropy, whereas the layer producing the wrong answer has dispersed visual attention and high entropy. This observation suggests that visual attention entropy indicates how well grounded the decoding state is at a given depth, rather than how much attention the layer allocates to the image. Section 4.4 quantifies this over the whole benchmark (Figure 6).



(a) Dynamic Correction Decoding with preCeding-Layer Knowledge(DeCo)

(b) Attention Entropy Decoding with Visual Grounding(AEDG)

Figure 2: Overview of the two decoding schemes. (a) DeCo corrects the final distribution with the most confident preceding layer, chosen by the candidate-token probability. (b) AEDG instead reads the visual attention entropy of the candidate layers and selects a dispersed (high-entropy) layer and a focused (low-entropy) layer, then fuses both into the final distribution.

3.3 Visual Attention Entropy

Let $\boldsymbol{a}_t^{(l)} \in \mathbb{R}^{|\mathrm{KV}|}$ be the attention row of the last query token at layer l and step t, averaged over heads. We identify the contiguous span $\mathbb{V} = \{s, \ldots, s+K-1\}$ of visual tokens in the key–value sequence, where s is the position of the first visual token. For each candidate layer we renormalize the attention inside this span,

$$p_i^{(l)} = \frac{a_i^{(l)}}{\sum_{j\in\mathbb{V}} a_j^{(l)}}, \qquad i \in \mathbb{V}, \tag{3}$$

and define the visual attention entropy together with its normalized form

$$H_l = -\sum_{i\in\mathbb{V}} p_i^{(l)} \log p_i^{(l)}, \qquad \bar{H}_l = \frac{H_l}{\log |\mathbb{V}|} \in [0,1], \tag{4}$$

where $|\mathbb{V}| = K$ is the number of visual tokens. For $K > 1$, $\bar{H}_l$ equals 0 when all visual attention is assigned to a single token and 1 when it is uniform. Dividing by $\log K$ normalizes the range to $[0,1]$ but does not eliminate the effects of visual-token count or image content. We separately record the visual attention mass

$$M_l = \sum_{i\in\mathbb{V}} a_i^{(l)}, \tag{5}$$

which measures how much attention is allocated to the image at all. M_l and $\bar{H}_l$ describe different properties and, as we show in Section 4.4, they behave differently: a layer can have a large M_l while its visual attention is nearly uniform.

3.4 Entropy-Driven Dual-Anchor Selection

We restrict the search to a candidate window $\mathcal{C} = [w_0, w_1) \subseteq \{1, \ldots, N\}$ of intermediate layers, which excludes the earliest layers (weak visual evidence) and the final layers (language priors). At every decoding step t, AEDG selects

$$l_{\mathrm{disp}} = \arg\max_{l\in\mathcal{C}} \bar{H}_l, \qquad l_{\mathrm{focus}} = \arg\min_{l\in\mathcal{C}} \bar{H}_l, \tag{6}$$

i.e. the most dispersed and the most focused visual-attention layer among the candidates. The dispersed layer is treated as a *negative* anchor: it is the state in which visual evidence is spread out and grounding is weakest, so its prediction is subtracted. The focused layer is treated as a *positive* anchor: it is the state in which the layer attends to a small, decisive set of visual tokens. Both indices depend on t and are recomputed at every step, so the anchors follow the evolution of the decoding state instead of being fixed for the whole sample.

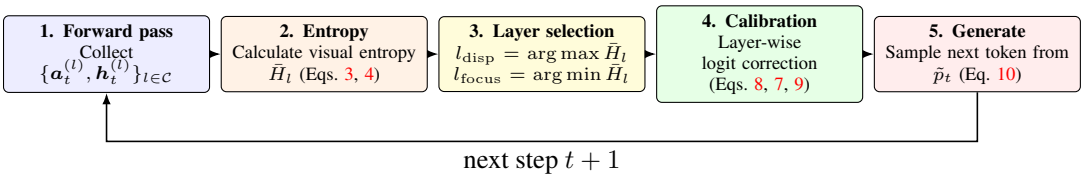


Figure 3: AEDG decoding flow. The layer pair is selected from the normalized visual attention entropy and used to calibrate the final-layer candidate logits before the next token is generated.

3.5 Dual-Anchor Calibration of the Final Distribution

Let $\boldsymbol{z}_t^{(N)}$, $\boldsymbol{z}^{\text{disp}}$ and $\boldsymbol{z}^{\text{focus}}$ denote the logits of the final, dispersed and focused layers at step t (Eq. 2). AEDG calibrates the final distribution by

$$\tilde{\boldsymbol{z}} = (1+\alpha)\,\boldsymbol{z}_t^{(N)} - \alpha\,\boldsymbol{z}^{\text{disp}} + \beta_{\text{eff}}\,\boldsymbol{z}^{\text{focus}}, \qquad \beta_{\text{eff}} = \beta\, g_{\text{focus}}, \tag{7}$$

which can be read as adding a layer-contrastive increment $\alpha(\boldsymbol{z}_t^{(N)} - \boldsymbol{z}^{\text{disp}})$ to the final logits and then adding the focused anchor. The scalars α and β control the strength of the negative and the positive term, respectively. The gate g_{focus} is the maximum probability assigned by the positive anchor over the vocabulary,

$$g_{\text{focus}} = \max_v \operatorname{softmax}\big(\boldsymbol{z}^{\text{focus}}\big)_v. \tag{8}$$

When g_{focus} is small, the positive-anchor term is downweighted. This confidence-based scaling is intended to limit the influence of uncertain positive anchors on the corrected distribution.

The correction is applied only to a candidate set built from the final-layer distribution. We take the top-k tokens of $p_t^{(N)}$ in descending probability order and keep the shortest prefix whose cumulative probability reaches at least p,

$$\mathcal{T} = \operatorname{TopK}\big(p_t^{(N)},\, k\big) \quad \text{truncated at cumulative mass } p, \tag{9}$$

where cumulative probabilities are taken from the original final-layer distribution without renormalizing within the top-k set. If their total mass is below p, all k tokens are retained. We define the calibrated distribution as

$$\tilde{p}_t = \operatorname{softmax}\big(\tilde{\boldsymbol{z}}\big) \text{ on } \mathcal{T}, \qquad \tilde{p}_t = 0 \text{ otherwise}, \tag{10}$$

i.e. all non-candidate logits are set to $-\infty$. The next token is then sampled from $\tilde{p}_t$ with any standard decoding rule (greedy, sampling, or beam search).

Per-step procedure. At each decoding step, AEDG follows the flow shown in Figure 3: it reads the backbone states and attention, identifies the focused and dispersed layers, calibrates the candidate logits, and generates the next token. The same procedure is repeated at step $t+1$.

3.6 Discussion and Complexity

AEDG is training-free and model agnostic: it only requires that the backbone exposes per-layer attention and a language head shared across layers, both of which hold for LLaVA-1.5, Qwen-VL, and GLM-4V. It adds no parameters, no auxiliary model, and no second forward pass. After head averaging, computing entropy over the visual spans costs $O(|\mathcal{C}|K)$ per step, and calibration requires two shared-head projections. Accessing attention maps and performing these operations still incur computation and memory overhead, which we report in Section 4.5.

We argue that the *relative* ordering of the candidate layers by visual attention concentration is an informative and stable signal for locating the layers in which visual evidence is still usable. Section 4.4 provides the controlled comparison against alternative signals.

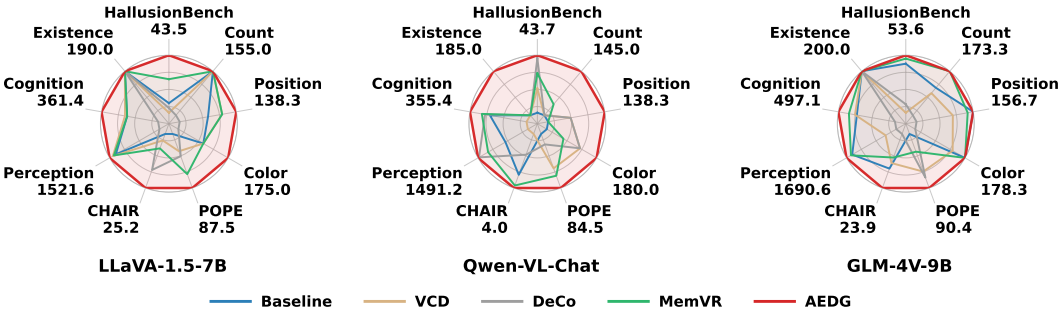


Figure 4: Overview of the results across the three backbones. Each axis is contrast-normalised within a backbone, so the best method on an axis reaches the outer ring; the nine axes cover the MME hallucination sub-scores, POPE, CHAIR, MME Perception/Cognition, and HallusionBench.

4 Experiments

4.1 Experimental Setup

Backbones. We evaluate on three open-source MLLMs that differ in visual encoder, connector, language backbone, visual-token budget, and depth: LLaVA-1.5-7B (Liu et al., 2024) (32 layers, 576 visual tokens), Qwen-VL-Chat (Bai et al., 2023) (256 visual tokens), and GLM-4V-9B (Glm et al., 2024) (40 layers, 1600 visual tokens). AEDG is applied at inference time only: no backbone parameter is updated, no auxiliary model is trained, and the released checkpoints are used unchanged.

Baselines. We compare against the vanilla backbone and three training-free, inference-time baselines: VCD (Leng et al., 2024), which contrasts the predictions obtained from clean and visually distorted inputs; DeCo (Wang et al., 2025), which dynamically adds the logits of a preceding layer selected by the candidate-token probability; and MemVR (Zou et al., 2024), which retraces memory-space visual information. We use the official implementation and default hyper-parameters of every baseline, and keep the backbone, prompt, and decoding configuration identical across all methods.

Benchmarks and metrics. For object hallucination we use POPE (Li et al., 2023) on MSCOCO (Lin et al., 2014) under the random, popular, and adversarial splits, reporting accuracy and F1; CHAIR (Rohrbach et al., 2018) on a 500-image COCO Val2014 subset with the standard captioning prompt, reporting sentence-level CHAIR_S and instance-level CHAIR_I as well as object Recall; and HallusionBench (Guan et al., 2024) , reported with its official metrics , where the open-ended judging follows the official protocol with GPT-4o-mini as the judge. For general multimodal capability we use MME (Fu et al., 2026), reporting the hallucination-sensitive MME-Hall subset, Perception, and Cognition. Dataset descriptions, metric definitions, and evaluation protocols are detailed in Appendix B.

Implementation details. To ensure a fair comparison, we adopt the experimental setup of prior work (Zou et al., 2024; Wang et al., 2025). Specifically, we use greedy decoding, where the do_sample =False, the temperature is set to 0. All baselines (VCD (Leng et al., 2024), MemVR (Zou et al., 2024), and DeCo (Wang et al., 2025)) are run under exactly the same decoding conditions with their officially recommended hyperparameters, so that the evaluation remains accurate. In our method, we set $\alpha = 0.1$ and $\beta = 0.6$, and build the candidate-token set from the top-$k = 10$ tokens of the final layer under a cumulative-probability cutoff of $p = 0.9$. Unless otherwise specified, all experiments are conducted on a single NVIDIA A100 GPU.

4.2 Performance on Hallucination Benchmarks

As shown in Table 2, AEDG consistently achieves the lowest CHAIR_S and CHAIR_I across three representative MLLMs, demonstrating its effectiveness in suppressing hallucinations. For example, on LLaVA-1.5-7B, AEDG reduces CHAIR_S from 50.0 to 39.8 and CHAIR_I from 15.4 to 10.6, while maintaining a comparable Recall. These results support the effectiveness of entropy-driven layer selection and dual-anchor logit correction in reducing hallucinations. Table 1 further evaluates AEDG on the POPE benchmark. On MSCOCO, AEDG achieves the best or near-best accuracy

Method	Random		Popular		Adversarial		Average	
	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑
LLaVA-1.5-7B	83.49 ↑0.0	82.28 ↑0.0	79.98 ↑0.0	79.34 ↑0.0	76.03 ↑0.0	76.26 ↑0.0	79.83 ↑0.0	79.29 ↑0.0
+VCD	86.84 ↑3.4	86.83 ↑4.6	82.65 ↑2.7	83.37 ↑4.0	77.31 ↑1.3	79.28 ↑3.0	82.27 ↑2.4	83.16 ↑3.9
+DeCo	86.66 ↑3.2	89.15 ↑6.9	84.63 ↑4.7	85.84 ↑6.5	77.63 ↑1.6	80.64 ↑4.4	82.97 ↑3.1	85.21 ↑5.9
+MemVR	88.50 ↑5.0	87.34 ↑5.1	86.07 ↑6.1	85.60 ↑6.3	81.97 ↑5.9	81.33 ↑5.1	85.51 ↑5.7	84.75 ↑5.5
+AEDG	**90.43** ↑6.9	**90.17** ↑7.9	**88.27** ↑8.3	**88.21** ↑8.9	**83.70** ↑7.7	**84.34** ↑8.1	**87.47** ↑7.6	**87.57** ↑8.3
Qwen-VL-Chat	84.17 ↑0.0	81.48 ↑0.0	83.73 ↑0.0	81.09 ↑0.0	82.97 ↑0.0	80.43 ↑0.0	83.62 ↑0.0	81.00 ↑0.0
+VCD	85.00 ↑0.8	83.08 ↑1.6	84.30 ↑0.6	**82.31** ↑1.2	83.23 ↑0.3	81.40 ↑1.0	84.18 ↑0.6	82.26 ↑1.3
+DeCo	84.47 ↑0.3	81.77 ↑0.3	83.93 ↑0.2	81.26 ↑0.2	82.97 ↑0.0	80.35 ↓0.1	83.79 ↑0.2	81.12 ↑0.1
+MemVR	85.10 ↑0.9	82.80 ↑1.3	**84.60** ↑0.9	82.20 ↑1.1	83.30 ↑0.3	81.00 ↑0.6	84.30 ↑0.7	82.00 ↑1.0
+AEDG	**85.60** ↑1.4	**83.50** ↑2.0	84.50 ↑0.8	82.20 ↑1.1	**83.40** ↑0.4	**81.50** ↑1.0	**84.50** ↑0.9	**82.40** ↑1.4
GLM-4V-9B	92.50 ↑0.0	91.22 ↑0.0	88.47 ↑0.0	87.45 ↑0.0	86.53 ↑0.0	86.75 ↑0.0	89.17 ↑0.0	88.47 ↑0.0
+VCD	93.33 ↑0.8	93.06 ↑1.8	89.37 ↑0.9	89.32 ↑1.9	87.43 ↑0.9	87.62 ↑0.9	90.04 ↑0.9	89.95 ↑1.5
+DeCo	93.53 ↑1.0	93.25 ↑2.0	89.47 ↑1.0	89.46 ↑2.0	87.53 ↑1.0	87.76 ↑1.0	90.18 ↑1.0	90.10 ↑1.6
+MemVR	92.87 ↑0.4	91.69 ↑0.5	88.68 ↑0.2	87.91 ↑0.5	87.20 ↑0.7	86.88 ↑0.1	89.58 ↑0.4	88.83 ↑0.4
+AEDG	**94.13** ↑1.6	**94.09** ↑2.9	**89.53** ↑1.1	**89.51** ↑2.1	**87.60** ↑1.1	**87.81** ↑1.1	**90.42** ↑1.3	**90.47** ↑2.0

Table 1: Performance comparison on POPE with LLaVA-1.5-7B, Qwen-VL-Chat, and GLM-4V-9B. The best value among the compared methods in each column is bolded. the small superscript arrow gives the change relative to that block's own vanilla backbone.

Method	CHAIR				HallusionBench				
	$CHAIR_S$↓	$CHAIR_I$↓	Average↓	Recall↑	fACC↑	qACC↑	easyA↑	hardA↑	aACC↑
LLaVA-1.5-7B	50.0 ↓0.0	15.4 ↓0.0	32.7 ↓0.0	77.1 ↑0.0	17.9 ↑0.0	8.1 ↑0.0	36.0 ↑0.0	36.7 ↑0.0	41.5 ↑0.0
+VCD	48.6 ↓1.4	14.9 ↓0.5	31.8 ↓0.9	77.3 ↑0.2	13.9 ↓4.0	**11.4** ↑3.3	33.0 ↓3.0	34.7 ↓2.0	41.1 ↓0.4
+DeCo	42.8 ↓7.2	12.6 ↓2.8	27.7 ↓5.0	76.7 ↓0.4	15.3 ↓2.6	8.1 ↑0.0	**42.1** ↑6.1	33.8 ↓2.9	41.3 ↓0.2
+MemVR	47.6 ↓2.4	13.8 ↓1.6	30.7 ↓2.0	**79.6** ↑2.5	**17.9** ↑0.0	9.0 ↑0.9	36.9 ↑0.9	37.7 ↑1.0	42.5 ↑1.0
+AEDG	**39.8** ↓10.	**10.6** ↓4.8	**25.2** ↓7.5	77.2 ↑0.1	16.8 ↓1.1	10.8 ↑2.7	40.4 ↑4.4	**39.5** ↑2.8	**43.5** ↑2.0
Qwen-VL-Chat	6.8 ↑0.0	5.3 ↑0.0	6.1 ↑0.0	53.4 ↑0.0	19.9 ↑0.0	12.3 ↑0.0	39.3 ↑0.0	32.0 ↑0.0	41.8 ↑0.0
+VCD	13.0 ↑6.2	12.3 ↑7.0	12.6 ↑6.5	47.9 ↓5.5	21.7 ↑1.8	14.1 ↑1.8	39.1 ↓0.2	**34.7** ↑2.7	42.6 ↑0.8
+DeCo	12.2 ↑5.4	6.4 ↑1.1	9.3 ↑3.2	**57.3** ↑3.9	20.2 ↑0.3	14.0 ↑1.7	**42.8** ↑3.5	34.6 ↑2.6	43.6 ↑1.8
+MemVR	5.2 ↓1.6	3.6 ↓1.7	4.4 ↓1.7	52.4 ↓1.0	19.7 ↓0.2	12.9 ↑0.6	42.4 ↑3.1	33.0 ↑1.0	43.1 ↑1.3
+AEDG	**5.0** ↓1.8	**3.0** ↓2.3	**4.0** ↓2.1	52.6 ↓0.8	**22.8** ↑2.9	**14.7** ↑2.4	42.0 ↑2.7	33.5 ↑1.5	**43.7** ↑1.9
GLM-4V-9B	41.8 ↑0.0	8.7 ↑0.0	25.3 ↑0.0	73.5 ↑0.0	30.6 ↑0.0	21.7 ↑0.0	52.3 ↑0.0	41.4 ↑0.0	53.1 ↑0.0
+VCD	41.8 ↑0.0	9.5 ↑0.8	25.7 ↑0.4	70.3 ↓3.2	26.9 ↓3.7	18.7 ↓3.0	45.9 ↓6.4	41.9 ↑0.5	50.2 ↓2.9
+DeCo	44.8 ↑3.0	10.8 ↑2.1	27.8 ↑2.5	69.1 ↓4.4	26.6 ↓4.0	18.9 ↓2.8	**52.3** ↑0.0	37.4 ↓4.0	50.7 ↓2.4
+MemVR	43.0 ↑1.2	9.1 ↑0.4	26.1 ↑0.8	72.4 ↓1.1	31.5 ↑0.9	22.4 ↑0.7	52.1 ↓0.2	43.4 ↑2.0	53.4 ↑0.3
+AEDG	**39.2** ↓2.6	**8.5** ↓0.2	**23.9** ↓1.4	**73.7** ↑0.2	**32.1** ↑1.5	**22.6** ↑0.9	51.9 ↓0.4	**43.8** ↑2.4	**53.6** ↑0.5

Table 2: Performance comparison on CHAIR and HallusionBench with LLaVA-1.5-7B, Qwen-VL-Chat, and GLM-4V-9B. The best value among the compared methods in each column is bolded.

and F1-score under Random, Popular, and Adversarial settings. Notably, AEDG sustains strong performance even under adversarial prompts, outperforming prior defenses such as MemVR.

4.3 Performance on General-Purpose Benchmarks

As shown in Table 3, AEDG preserves strong performance on MME, achieving results comparable to or better than existing methods across object- and attribute-level tasks. Across all three models, AEDG improves both Perception and Cognition totals relative to the vanilla backbone. These results suggest that AEDG reduces hallucination without compromising general multimodal reasoning on the evaluated tasks.

Methods	MME-Hall	Object-Level		Attribute-Level		Perception	Cognition
	Total↑	Existence↑	Count↑	Position↑	Color↑	Total↑	Total↑
LLaVA-1.5-7B	643.3 ↑0.0	190.0 ↑0.0	155.0 ↑0.0	128.3 ↑0.0	170.0 ↑0.0	1508.9 ↑0.0	335.7 ↑0.0
+VCD	648.3 ↑5.0	**190.0** ↑0.0	**155.0** ↑0.0	133.3 ↑5.0	170.0 ↑0.0	1515.0 ↑6.0	337.9 ↑2.1
+DeCo	608.3 ↓35.	**190.0** ↑0.0	135.0 ↓20.	118.3 ↓10.	165.0 ↓5.0	1419.9 ↓89.	304.6 ↓31.
+MemVR	648.3 ↑5.0	**190.0** ↑0.0	**155.0** ↑0.0	133.3 ↑5.0	170.0 ↑0.0	1512.8 ↑3.8	335.7 ↑0.0
+AEDG	**658.3** ↑15.	**190.0** ↑0.0	**155.0** ↑0.0	**138.3** ↑10.	**175.0** ↑5.0	**1521.6** ↑13.	**361.4** ↑26.
Qwen-VL-Chat	600.0 ↑0.0	180.0 ↑0.0	125.0 ↑0.0	130.0 ↑0.0	165.0 ↑0.0	1442.8 ↑0.0	342.1 ↑0.0
+VCD	613.3 ↑13.	180.0 ↑0.0	125.0 ↑0.0	133.3 ↑3.3	175.0 ↑10.	1403.2 ↓40.	317.9 ↓24.
+DeCo	613.3 ↑13.	180.0 ↑0.0	125.0 ↑0.0	133.3 ↑3.3	175.0 ↑10.	1489.5 ↑47.	345.4 ↑3.3
+MemVR	610.0 ↑10.	180.0 ↑0.0	130.0 ↑5.0	130.0 ↑0.0	170.0 ↑5.0	1473.4 ↑31.	347.5 ↑5.4
+AEDG	**648.3** ↑48.	**185.0** ↑5.0	**145.0** ↑20.	**138.3** ↑8.3	**180.0** ↑15.	**1491.2** ↑48.	**355.4** ↑13.
GLM-4V-9B	703.3 ↑0.0	200.0 ↑0.0	168.3 ↑0.0	156.7 ↑0.0	178.3 ↑0.0	1680.9 ↑0.0	479.6 ↑0.0
+VCD	675.0 ↓28.	**200.0** ↑0.0	166.7 ↓1.6	138.3 ↓18.	170.0 ↓8.3	1624.1 ↓57.	481.4 ↑1.8
+DeCo	610.5 ↓93.	**200.0** ↑0.0	160.0 ↓8.3	105.5 ↓51.	145.0 ↓33.	1605.9 ↓75.	443.6 ↓36.
+MemVR	703.3 ↑0.0	**200.0** ↑0.0	**173.3** ↑5.0	151.7 ↓5.0	**178.3** ↑0.0	1683.0 ↑2.1	487.1 ↑7.5
+AEDG	**708.3** ↑5.0	**200.0** ↑0.0	**173.3** ↑5.0	**156.7** ↑0.0	**178.3** ↑0.0	**1690.7** ↑9.8	**497.1** ↑18.

Table 3: Performance comparison on MME with LLaVA-1.5-7B, Qwen-VL-Chat, and GLM-4V-9B. MME-Hall is the hallucination-sensitive subset score aggregated over Existence, Count, Position, and Color; Perception and Cognition are the standard MME category totals. The best value among the compared methods in each column is bolded.

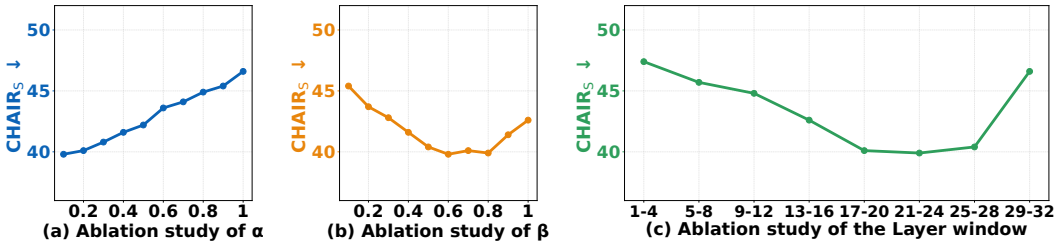


Figure 5: Hyperparameter study on CHAIR with LLaVA-1.5-7B: (a) the negative-anchor strength α, (b) the positive-anchor strength β, and (c) the candidate layer window.

4.4 Ablation Studies

Impact of the layer-selection signal. To evaluate how the choice of layer-selection signal affects performance, we conduct an ablation study on CHAIR with LLaVA-1.5-7B, as reported in Table 4. We keep the candidate set, layer window, fusion rule, and hyper-parameters fixed, and vary only the signal used to select the intermediate layers. Attention mass reduces CHAIR_S from 50.0 to 46.2, while logit entropy and full attention entropy further reduce it to 44.3 and 41.9, respectively. The full AEDG configuration achieves the lowest scores, with CHAIR_S and CHAIR_I of 39.8 and 10.6. These results show that the distribution of attention provides a more informative layer-selection cue than attention magnitude or vocabulary-level uncertainty. Complementing this ablation, Figure 6 shows that lower visual attention entropy is generally associated with higher answer accuracy on MME-Hall questions with ground-truth answers of *No*, with the largest gap of 37.3 percentage points.

Effectiveness of the anchor components. The second block of Table 4 evaluates the two anchors. Removing the positive anchor increases CHAIR_S from 39.8 to 46.7, while removing the negative anchor increases it to 43.5. This indicates that both anchors contribute to hallucination reduction, with the positive anchor providing visual evidence and the negative anchor suppressing unreliable decoding states. The confidence gate is kept fixed throughout and is not ablated.

Table 4: Ablation study on CHAIR with LLaVA-1.5-7B.

Selection signal	$CHAIR_S\downarrow$	$CHAIR_I\downarrow$	Rec.↑
Layer-selection signal			
Vanilla	50.0	15.4	77.1
Attention mass	46.2	13.9	76.6
Logit entropy	44.3	13.1	76.5
Laplacian energy	43.0	12.4	77.3
Full attention entropy	41.9	12.0	76.3
Anchor components			
w/o pos. anchor	46.7	13.8	76.7
w/o neg. anchor	43.5	12.1	76.4
Layer pair and anchor direction			
Fixed pair (22)	46.4	13.0	76.8
Random pair	49.6	14.5	77.4
Reversed anchors	52.0	14.6	77.9
AEDG	**39.8**	**10.6**	77.2

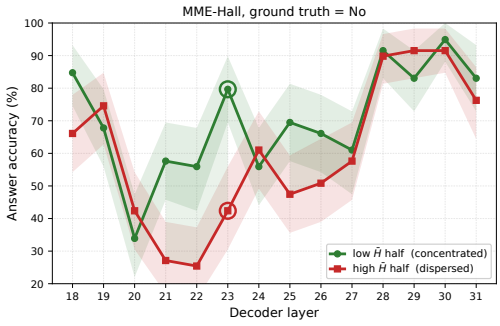


Figure 6: Visual attention entropy and answer accuracy on MME-Hall.The largest gap occurs at intermediate layer 23, where the low-entropy group achieves 79.7% accuracy versus 42.4% for the high-entropy group.Shaded bands indicate 95% bootstrap confidence intervals.

Method	Memory (MB)↓	Latency (ms/token)↓	$CHAIR_S\downarrow$	$CHAIR_I\downarrow$	Average.↓
LLaVA-1.5-7B	14,356	31.0	50.0	15.4	32.7
+VCD	15,237	62.8	48.6	14.9	31.8
+DeCo	14,383	34.8	42.8	12.6	27.7
+MemVR	14,449	46.9	47.6	13.8	30.7
+AEDG	14,486	31.8	**39.8**	**10.6**	**25.2**

Table 5: Inference efficiency on LLaVA-1.5-7B, Comparison of inference speed, GPU memory usage, and hallucination performance on the CHAIR using a single A100 GPU.

Effect of the anchor strengths. Figure 5(a)(b) studies the two decoding strengths over $\{0.1, \ldots, 1.0\}$. Increasing the negative-anchor strength α steadily worsens hallucination performance, indicating that overly strong suppression can distort the final prediction. The positive-anchor strength β performs best at a moderate value, whereas both very small and very large values are less effective. We therefore use $\alpha = 0.1$ and $\beta = 0.6$ in all experiments.

Impact of the candidate layer window. Figure 5(c) shows that mid-to-deep layers provide the most effective operating range. Very early layers contain insufficiently developed visual representations, whereas very late layers are more affected by language priors. The selected window is therefore kept in the mid-to-deep range and adjusted empirically for different backbones.

4.5 Efficiency Analysis

Table 5 shows that AEDG achieves the lowest $CHAIR_S$ and $CHAIR_I$ among the compared methods with a small inference overhead. Relative to vanilla LLaVA-1.5-7B, latency increases from 31.0 to 31.8 ms/token and GPU memory from $14{,}356$ to $14{,}486$ MB. AEDG has lower latency than VCD and MemVR. These results indicate that AEDG reduces hallucination while remaining close to the baseline in inference cost.

5 Conclusion

In this work, we introduced AEDG, a training-free decoding framework that uses visual attention entropy to select reliable intermediate layers and mitigate hallucinations in MLLMs. AEDG combines high- and low-entropy anchors to calibrate the final prediction without parameter updates or an additional forward pass. Experiments on LLaVA-1.5-7B, Qwen-VL-Chat, and GLM-4V-9B show that AEDG reduces hallucination while maintaining strong multimodal performance. Overall, AEDG provides an efficient and effective solution for building more reliable MLLMs.

AI USE STATEMENT

We used generative AI tools to aid and polish the writing of this paper, including drafting and revising the theoretical analysis in Appendix A. Separately, GPT-4o-mini served as an automated judge for open-ended HallusionBench responses, as described in Section 4.1; its judgments are subject to errors.

ETHICS STATEMENT

AEDG aims to reduce visually unsupported outputs from multimodal language models through inference-time calibration of released checkpoints. Our experiments use established vision–language benchmarks. Although reducing hallucination can improve reliability, AEDG does not guarantee factual correctness or eliminate biases inherited from the underlying models and evaluation data. In particular, concentrated visual attention and confident predictions can still be incorrect, and benchmark improvements do not establish safety in high-stakes applications. Automated evaluation also has limitations, including potential errors and biases in model-based judgments. Downstream use should respect the applicable model and dataset licenses, protect sensitive information, and include task-specific validation and appropriate human oversight rather than treating calibrated outputs as verified facts.

REPRODUCIBILITY STATEMENT

Section 3 specifies the visual attention normalization, entropy-based layer selection, confidence gate, dual-anchor logit correction, and candidate filtering that define AEDG. Section 4.1 identifies the evaluated backbones, benchmarks, metrics, baseline configurations, default decoding hyperparameters, and hardware used in our experiments. AEDG operates on released checkpoints without parameter updates or an auxiliary trained model. Section 4.4 examines the effects of the selection signal, anchor components, and hyperparameters, while Section 4.5 reports inference latency and memory usage. Appendix A provides the assumptions and derivations underlying the theoretical analysis and distinguishes exact algebraic identities from conditional guarantees and their limitations. Appendix B describes the benchmarks, evaluation metrics, backbones, and baselines.

REFERENCES

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

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

Fanpu Cao, Xin Zou, Xuming Hu, and Hui Xiong. When looking is not enough: Visual attention structure reveals hallucination in mllms, 2026. URL https://arxiv.org/abs/2605.11559.

Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Linfeng Zhang, Lijie Wen, and Xuming Hu. Ict: Image-object cross-level trusted intervention for mitigating object hallucination in large vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4209–4221, 2025.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations*, volume 2024, pp. 54158–54183, 2024.

Xin Dong, Shichao Dong, Jin Wang, Jing Huang, Li Zhou, Zenghui Sun, Lihua Jing, Jinsong Lan, Xiaoyong Zhu, and Bo Zheng. Inter: Mitigating hallucination in large vision-language models by interaction guidance sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2534–2544, 2025.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Team Glm, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14375–14385, 2024.

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

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. *arXiv preprint arXiv:2408.02032*, 2024.

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

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024.

Xiaoye Qu, Jiashuo Sun, Wei Wei, Daizong Liu, Jianfeng Dong, and Yu Cheng. Look, compare, decide: Alleviating hallucination in large vision-language models via multi-view multi-path reasoning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 4428–4441, 2025.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, 2018.

Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, et al. Vigc: Visual instruction generation and correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5309–5317, 2024a.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. Mllm can see? dynamic correction decoding for hallucination mitigation. In *International Conference on Learning Representations*, volume 2025, pp. 13712–13736, 2025.

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

Yujun Wang, Jinhe Bi, Soren Pirk, Yunpu Ma, et al. Ascd: Attention-steerable contrastive decoding for reducing hallucination in mllm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10306–10314, 2026.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, et al. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19985–19995, 2025.

Yongchang Zhang, Oliver Ma, Tianyi Liu, Guangquan Zhou, and Yang Chen. See it, say it, sorted: An iterative training-free framework for visually-grounded multimodal reasoning in lvlms. *arXiv preprint arXiv:2602.21497*, 2026.

Zefeng Zhang, Hengzhu Tang, Jiawei Sheng, Zhenyu Zhang, Yiming Ren, Zhenyang Li, Dawei Yin, Duohe Ma, and Tingwen Liu. Debiasing multimodal large language models via noise-aware preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9423–9433, 2025.

Xingyu Zhu, Kesen Zhao, Liang Yi, Shuo Wang, Zhicai Wang, Beier Zhu, Hanwang Zhang, and Xiangnan He. Look carefully: Adaptive visual reinforcements in multimodal large language models for hallucination mitigation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=KhEjFE1QvQ.

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, et al. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2410.03577*, 2024.

A Theoretical Analysis

This appendix analyzes three aspects of AEDG: what visual attention entropy measures, when the dual-anchor correction favors grounded continuations, and how gating and candidate filtering constrain the resulting distribution. The algebraic identities and perturbation bounds below hold without a statistical model of the backbone. In contrast, interpreting concentration as relevant visual evidence requires an explicit alignment assumption. We keep these two types of statements separate; low entropy alone is not a certificate of factual correctness.

A.1 Notation and Scope

We fix an image, a query, a generated prefix, and a decoding step t, and suppress t when unambiguous. Let $q_l(i) = p_i^{(l)}$ in Eq. 3 denote the distribution over the K visual tokens, and let $P_l(u) = p_t^{(l)}(u)$ in Eq. 2 denote the distribution over a vocabulary $\mathcal{W}$ of size D. Thus, visual attention entropy and vocabulary uncertainty refer to different distributions. We assume $K > 1$, $M_l > 0$ for all candidate layers, a nonempty layer window $\mathcal{C}$, finite vocabulary logits, $\alpha, \beta \geq 0$, $1 \leq k \leq D$, and a cutoff $p \in (0, 1]$. Write $f = l_{\text{focus}}$, $d = l_{\text{disp}}$, and $\lambda = \beta_{\text{eff}} = \beta \max_u P_f(u)$. The candidate set $\mathcal{T}$ is exactly the set in Eq. 9. Define

$$c = P_N(\mathcal{T}), \qquad \pi(u) = \frac{P_N(u)}{c} \quad (u \in \mathcal{T}), \qquad \Delta(u) = \alpha\big(z^{(N)}(u) - z^d(u)\big) + \lambda z^f(u). \tag{11}$$

Here π is the final-layer distribution conditioned on the same candidates, not the unfiltered vanilla distribution. All logarithms are natural, with $0 \log 0 = 0$. Probability statements concern the unit-temperature softmax in Eq. 10; we state separately the implications for greedy decoding, which is used in our experiments.

A.2 What Visual Attention Entropy Measures

Concentration rather than attention magnitude. Let $U_{\mathbb{V}}$ be uniform on the visual span. Directly from the definition,

$$D_{\text{KL}}(q_l \| U_{\mathbb{V}}) = \log K - H_l = (1 - \bar{H}_l) \log K, \qquad n_{\text{eff}}(q_l) := \exp(H_l) = K^{\bar{H}_l}. \tag{12}$$

Thus, minimizing normalized entropy maximizes divergence from uniform visual attention and minimizes its entropy-based effective support. Moreover,

$$\max_{i \in \mathbb{V}} q_l(i) \geq \exp(-H_l) = K^{-\bar{H}_l}, \tag{13}$$

since $H_l = \sum_i q_l(i) \log(1/q_l(i)) \geq \log(1/\max_i q_l(i))$. These identities do not identify which visual token receives the mass.

The distinction from full attention entropy is also exact. Let r_l be the attention distribution conditioned on nonvisual positions and suppose $0 < M_l < 1$. The entropy chain rule gives

$$H(\boldsymbol{a}^{(l)}) = h_{\mathrm{b}}(M_l) + M_l H_l + (1 - M_l) H(r_l), \qquad h_{\mathrm{b}}(x) = -x \log x - (1-x) \log(1-x). \tag{14}$$

Indeed, substituting $a_i^{(l)} = M_l q_l(i)$ on visual positions and $a_i^{(l)} = (1-M_l) r_l(i)$ elsewhere yields the decomposition. Full attention entropy therefore also depends on the visual mass and the organization of nonvisual attention. Conditioning on the visual span removes those terms. For any fixed q_l and any $M_l \in (0,1)$, these factorizations define a valid attention row, so visual mass alone does not determine visual concentration.

An explicit relevance-alignment model. To describe when concentration tracks useful evidence, consider a nonempty proper subset $\mathcal{S} \subset \mathbb{V}$ of relevant visual tokens, with $m = |\mathcal{S}| < K$. Let $U_{\mathcal{S}}$ denote its uniform distribution, extended by zero outside $\mathcal{S}$. The following is an illustrative assumption on candidate layers, not a property implied by Transformer attention:

$$q_l = (1 - \rho_l) U_{\mathcal{S}} + \rho_l U_{\mathbb{V}}, \qquad 0 \leq \rho_l \leq 1. \tag{15}$$

The same relevant set is shared across layers at the fixed decoding step; ρ_l measures contamination by spatially uniform attention.

Proposition A.1 (Entropy ordering under aligned contamination) *Under Eq. 15, H_l is strictly increasing in ρ_l, while the attention mass on relevant tokens,*

$$q_l(\mathcal{S}) = 1 - \rho_l \left(1 - \frac{m}{K}\right), \tag{16}$$

is strictly decreasing. Consequently, the focused anchor maximizes relevant attention mass and the dispersed anchor minimizes it within $\mathcal{C}$.

Proof. At contamination level ρ, each token in $\mathcal{S}$ has probability $A(\rho) = (1-\rho)/m + \rho/K$, and each token outside has probability $B(\rho) = \rho/K$. Hence $H(\rho) = -mA(\rho)\log A(\rho) - (K-m)B(\rho)\log B(\rho)$. For $0 < \rho < 1$,

$$\frac{\mathrm{d}H(\rho)}{\mathrm{d}\rho} = \frac{K-m}{K} \log \frac{A(\rho)}{B(\rho)} > 0. \tag{17}$$

Continuity extends the strict ordering to $[0,1]$. Summing the probabilities on $\mathcal{S}$ gives Eq. 16. Division by the common positive constant $\log K$ leaves the ordering unchanged. □

This proposition supplies a sufficient model for the anchor directions in Eq. 6. It does not establish that every low-entropy layer is well grounded: a point mass on an irrelevant token has zero entropy but no relevant mass. Likewise, evidence distributed over several objects can have high entropy. Even under Eq. 15, converting relevant attention into a correct vocabulary prediction requires an aligned readout. The token-level conditions needed by the calibration are made explicit in Section A.5 rather than inferred from entropy alone.

A.3 Stability of Entropy-Based Layer Selection

Proposition A.2 (Selection stability with an entropy gap) *Suppose perturbed visual distributions $\widehat{q}_l$ satisfy $\frac{1}{2}\|q_l - \widehat{q}_l\|_1 \leq \delta$ for every $l \in \mathcal{C}$, where $0 \leq \delta \leq 1 - 1/K$. Set*

$$\varepsilon_K(\delta) = \frac{h_{\mathrm{b}}(\delta) + \delta \log(K-1)}{\log K}. \tag{18}$$

Then $|\bar{H}_l - \widehat{\bar{H}}_l| \leq \varepsilon_K(\delta)$. If the focused anchor is unique and

$$\gamma_f := \min_{l \in \mathcal{C} \setminus \{f\}} (\bar{H}_l - \bar{H}_f) > 2\varepsilon_K(\delta), \tag{19}$$

its index is unchanged by the perturbations. The same holds for the unique dispersed anchor when $\gamma_d := \min_{l \in \mathcal{C}\setminus\{d\}} (\bar{H}_d - \bar{H}_l) > 2\varepsilon_K(\delta)$.

Proof. For two distributions with total variation η, couple random variables X, Y by matching their common probability mass, so that $\Pr(X \neq Y) = \eta$. Conditional on a mismatch and on Y, X takes at most $K-1$ values. Thus $H(X \mid Y) \leq h_{\mathrm{b}}(\eta) + \eta \log(K-1)$, and the same bound holds for $H(Y \mid X)$. Since $H(X) - H(Y) = H(X \mid Y) - H(Y \mid X)$, their entropy difference in absolute value is at most this bound. The bound is nondecreasing for $\eta \in [0, 1-1/K]$, so $\eta \leq \delta$ gives Eq. 18. For any $l \neq f$,

$$\hat{\bar{H}}_l - \hat{\bar{H}}_f \geq \bar{H}_l - \bar{H}_f - 2\varepsilon_K(\delta) > 0. \tag{20}$$

The maximum-entropy case follows by reversing the differences. A singleton candidate window is trivially stable. □

Why nonnegligible visual mass matters. The preceding bound is on normalized visual attention. If unnormalized visual rows $a, \hat{a}$ obey $\|a - \hat{a}\|_1 \leq \eta$ and their masses $M, \widehat{M}$ are at least $m_0 > 0$, then

$$\left\| \frac{a}{M} - \frac{\hat{a}}{\widehat{M}} \right\|_1 \leq \frac{\|a - \hat{a}\|_1 + |M - \widehat{M}|}{M} \leq \frac{2\eta}{m_0}. \tag{21}$$

This follows by adding and subtracting $\hat{a}/M$ and using $|M - \widehat{M}| \leq \|a - \hat{a}\|_1$. Consequently, very small visual mass can amplify numerical or attention perturbations after normalization. Neither a positive lower bound on this mass nor a large entropy gap is enforced by the selection rule itself. The result is therefore a conditional stability guarantee, not a claim that selected layers cannot switch near ties.

A.4 Dual-Anchor Calibration as Exponential Reweighting

Proposition A.3 (Exact distribution and variational characterization) *On the fixed candidate set $\mathcal{T}$, AEDG satisfies*

$$\tilde{p}(u) = \frac{\pi(u) e^{\Delta(u)}}{\sum_{w \in \mathcal{T}} \pi(w) e^{\Delta(w)}} \;\propto\; P_N(u) \left(\frac{P_N(u)}{P_d(u)} \right)^{\alpha} P_f(u)^{\lambda}. \tag{22}$$

It is also the unique solution to

$$\tilde{p} = \arg\max_{r \in \mathcal{P}(\mathcal{T})} \left\{ \mathbb{E}_{u \sim r}[\Delta(u)] - D_{\mathrm{KL}}(r \| \pi) \right\}, \tag{23}$$

where $\mathcal{P}(\mathcal{T})$ is the probability simplex on $\mathcal{T}$.

Proof. Equation 7 gives $\tilde{z}(u) = z^{(N)}(u) + \Delta(u)$. Exponentiating and normalizing on $\mathcal{T}$ proves the first identity. Replacing each logit by its log-probability plus its layer-specific log-normalizer yields the product expression, since all log-normalizers are constant in u. Write $Z_\Delta = \sum_{u \in \mathcal{T}} \pi(u) e^{\Delta(u)}$. For any $r \in \mathcal{P}(\mathcal{T})$,

$$\mathbb{E}_r[\Delta] - D_{\mathrm{KL}}(r \| \pi) = \log Z_\Delta - D_{\mathrm{KL}}(r \| \tilde{p}). \tag{24}$$

Nonnegativity of KL divergence, with equality only at $r = \tilde{p}$, proves the variational statement. □

The negative-anchor contribution is a likelihood-ratio correction $(P_N/P_d)^{\alpha}$, while the positive anchor contributes the factor P_f^{λ}. The KL term in Eq. 23 penalizes deviation from the filtered final-layer prediction. This is an exact optimization interpretation of the implemented logits, not an assumption that the three layer predictions are independent experts. In particular, subtracting z^d need not remove all language priors, and adding z^f can also reinforce a confident but unsupported prediction.

A.5 Conditions for Favoring Grounded Continuations

For two retained tokens $u, v \in \mathcal{T}$, define the layer-wise margin $m_l(u, v) = z^{(l)}(u) - z^{(l)}(v)$, using m_f, m_d for the selected anchors.

Proposition A.4 (Pairwise odds and greedy recovery) *The change in log-odds between any two retained tokens is exactly*

$$\log \frac{\tilde{p}(u)}{\tilde{p}(v)} - \log \frac{\pi(u)}{\pi(v)} = \alpha\big(m_N(u, v) - m_d(u, v)\big) + \lambda m_f(u, v) = \Delta(u) - \Delta(v). \tag{25}$$

Thus, for a grounded token u and an unsupported competitor v, the conditions $m_N(u,v) \geq m_d(u,v)$ and $m_f(u,v) \geq 0$ are sufficient for nondecreasing grounded-token odds. The improvement is strict whenever one of the two weighted terms on the right is strictly positive. More generally, u defeats v under greedy decoding if and only if

$$\Delta(u) - \Delta(v) > -m_N(u,v), \tag{26}$$

where strict inequality excludes ties. If this holds for every $v \in \mathcal{T} \setminus \{u\}$, AEDG greedily selects u.

Proof. The normalizers cancel in each pairwise log-odds. Subtracting the two calibrated logits gives $\tilde{z}(u) - \tilde{z}(v) = m_N(u,v) + \alpha(m_N(u,v) - m_d(u,v)) + \lambda m_f(u,v)$. Subtracting the baseline margin establishes Eq. 25; the remaining claims follow from the signs of the terms and the definition of a greedy maximizer. □

The negative anchor is useful for this pair only when its relative preference for the unsupported token is stronger than that of the final layer, or when the positive term compensates for an unfavorable contrastive increment. The focused anchor must favor the relevant continuation at the vocabulary level, not merely concentrate attention. These conditions explain why anchor direction matters and why increasing α need not improve decoding.

A corresponding set-level statement avoids relying on a single competitor. At the fixed prefix, suppose the candidates can be partitioned into a nonempty grounded set $\mathcal{G}$ and a nonempty unsupported set $\mathcal{B}$. This semantic partition is supplied for analysis; AEDG does not observe it.

Corollary A.5 (Conditional suppression of unsupported probability) *Let $r = \pi(\mathcal{B})$ and $\delta_* = \min_{u\in\mathcal{G}} \Delta(u) - \max_{v\in\mathcal{B}} \Delta(v)$. Then*

$$\tilde{p}(\mathcal{B}) \leq \frac{r}{r + e^{\delta_*}(1-r)}. \tag{27}$$

In particular, if $\delta_* > 0$, the unsupported probability is strictly smaller than under π.

Proof. Let $A = \sum_{u\in\mathcal{G}} \pi(u) e^{\Delta(u)}$ and $B = \sum_{v\in\mathcal{B}} \pi(v) e^{\Delta(v)}$. By bounding each exponent with its respective extremum, $A/B \geq e^{\delta_*}(1-r)/r$. Substituting this into $\tilde{p}(\mathcal{B}) = 1/(1 + A/B)$ proves the result. □

The corollary bounds the next-token probability under sampling at a fixed prefix. It does not by itself establish a reduction in greedy sequence-level hallucination, CHAIR, or POPE error: decoding changes future prefixes, and semantic hallucination is not generally a fixed token-level partition.

A.6 Confidence Gating and Distributional Distortion

For a function b on $\mathcal{T}$, define its oscillation $\mathrm{osc}_{\mathcal{T}}(b) = \max_{u\in\mathcal{T}} b(u) - \min_{u\in\mathcal{T}} b(u)$. Oscillation, rather than absolute logit magnitude, is relevant because adding a constant to all logits does not change softmax. Set

$$R = \mathrm{osc}_{\mathcal{T}}(\Delta), \qquad R_- = \mathrm{osc}_{\mathcal{T}}(z^{(N)} - z^d), \qquad R_+ = \mathrm{osc}_{\mathcal{T}}(z^f). \tag{28}$$

Proposition A.6 (A gate-dependent distortion bound) *For the fixed candidates and selected logits,*

$$R \leq \alpha R_- + \beta g_{\mathrm{focus}} R_+, \qquad e^{-R} \leq \frac{\tilde{p}(u)}{\pi(u)} \leq e^R \quad (u \in \mathcal{T}). \tag{29}$$

Both directions of KL divergence satisfy

$$D_{\mathrm{KL}}(\tilde{p}\|\pi) \leq \frac{R^2}{8}, \qquad D_{\mathrm{KL}}(\pi\|\tilde{p}) \leq \frac{R^2}{8}. \tag{30}$$

Proof. Subadditivity of oscillation and nonnegativity of the coefficients give the first bound. Writing $a = \min\Delta$ and $b = \max\Delta$, we have $e^a \leq Z_\Delta \leq e^b$. Equation 22 then gives the ratio bounds because $b - a = R$. For the KL bounds, let $F(s) = \log \sum_{u\in\mathcal{T}} \pi(u) e^{s\Delta(u)}$ for $s \in [0,1]$. Its

second derivative is the variance of Δ under the distribution proportional to $\pi e^{s\Delta}$, and therefore $F''(s) \le R^2/4$: any random variable in $[a, b]$ has variance at most $(b-a)^2/4$. Since $F(0) = 0$,

$$D_{\mathrm{KL}}(\pi\|\tilde{p}) = F(1) - F'(0) = \int_0^1 (1-s)F''(s)\,\mathrm{d}s \le R^2/8,$$
$$D_{\mathrm{KL}}(\tilde{p}\|\pi) = F'(1) - F(1) = \int_0^1 sF''(s)\,\mathrm{d}s \le R^2/8. \tag{31}$$

This proves both inequalities. □

The gate obeys $1/D \le g_{\mathrm{focus}} \le 1$, so it weakens the positive contribution to this worst-case bound relative to an ungated coefficient β. This need not monotonically reduce the actual KL divergence, because the positive and negative corrections can cancel. If P_f is exactly uniform over the vocabulary, its logits are constant and $R_+ = 0$, so its addition has no effect on the normalized distribution. However, a small maximum probability does not by itself bound all logit differences, and a confidently wrong anchor can have $g_{\mathrm{focus}} \approx 1$. The gate is consequently a magnitude-control mechanism, not a calibrated probability of visual correctness.

A.7 Candidate Filtering and the Limits of the Guarantees

Let $c_k = \sum_{u \in \mathrm{TopK}(P_N, k)} P_N(u)$. The precise filtering rule in Eq. 9 implies

$$|\mathcal{T}| \le k, \qquad c = P_N(\mathcal{T}) \ge \min(p, c_k) > 0. \tag{32}$$

If $c_k \ge p$, the shortest retained prefix has mass at least p; otherwise all top-k tokens are retained and $c = c_k < p$. Thus, the cutoff alone does not guarantee retained mass p when the top-k cap is active. The final-layer highest-probability token is always retained, but a grounded alternative outside $\mathcal{T}$ cannot be recovered by either anchor.

Filtering and calibration have an exact decomposition relative to the unfiltered final-layer distribution:

$$D_{\mathrm{KL}}(\tilde{p}\|P_N) = D_{\mathrm{KL}}(\tilde{p}\|\pi) + \log\frac{1}{c} \le \frac{R^2}{8} + \log\frac{1}{c}. \tag{33}$$

To see this, substitute $\pi(u) = P_N(u)/c$ into the KL sum over $\mathcal{T}$; $\tilde{p}$ has zero mass elsewhere. The two terms respectively capture within-candidate reweighting and the cost of removing baseline probability mass. For $\alpha = \beta = 0$, AEDG reduces to π, not generally to P_N. It nevertheless preserves the vanilla greedy choice when that choice is unique, since filtering retains the final-layer maximizer.

Implications and limitations. Taken together, the results explain how AEDG can improve decoding when concentrated attention aligns with relevant regions, the selected readouts provide favorable token margins, and the candidate set retains grounded alternatives. Entropy gaps control selection stability, while gate-weighted logit ranges and retained mass control distributional distortion. None of these results guarantees factuality from attention alone, determines optimal α or β, or establishes that smaller distortion implies lower hallucination. In particular, near-tied entropies, negligible visual mass, confidently incorrect focused anchors, and exclusion of the correct token are explicit failure cases. The empirical evaluations in Section 4 assess the net effect beyond these conditional, per-step guarantees.

B Experimental Details

B.1 Benchmarks and Metrics

Datasets. To evaluate the effectiveness of AEDG, we conduct experiments on three visual hallucination benchmarks and one general-purpose multimodal benchmark. The hallucination benchmarks include Polling-based Object Probing Evaluation (POPE) (Li et al., 2023), Caption Hallucination Assessment with Image Relevance (CHAIR) (Rohrbach et al., 2018), and HallusionBench (Guan et al., 2024). We additionally use MME (Fu et al., 2026) to assess whether AEDG preserves broad multimodal perception and reasoning capabilities. We describe these benchmarks and their evaluation metrics below.

CHAIR (Rohrbach et al., 2018) evaluates whether objects mentioned in a generated caption are supported by the input image. It comprises two complementary metrics: CHAIR_S, the fraction of captions containing at least one hallucinated object, and CHAIR_I, the fraction of hallucinated object mentions among all normalized object mentions. We evaluate on a 500-image subset of MSCOCO Val2014 (Lin et al., 2014), whose instance annotations cover 80 object categories, using the prompt "Please describe this image in detail." The metrics are defined as

$$\text{CHAIR}_\text{i} = \frac{|\{\text{hallucinated object mentions}\}|}{|\{\text{all object mentions}\}|},$$

$$\text{CHAIR}_\text{s} = \frac{|\{\text{captions containing hallucinated objects}\}|}{|\{\text{all captions}\}|}.$$

We report both metrics as percentages, with lower values indicating fewer hallucinations. We additionally report object recall to assess whether hallucination reduction preserves coverage of the image content.

POPE (Li et al., 2023) formulates object hallucination evaluation as a binary visual question-answering task, asking whether a specified object is present in an image. We evaluate on the MSCOCO-based Random, Popular, and Adversarial settings. These settings construct negative queries using randomly selected absent objects, frequently occurring absent objects, and absent objects that commonly co-occur with those present in the image, respectively. We use the benchmark question text and report accuracy and F1 score for each setting, together with their averages over the three settings.

HallusionBench (Guan et al., 2024) complements object-centric benchmarks by evaluating hallucinations involving visual perception and cross-modal reasoning. Its structured examples test whether models answer consistently across related questions and changes in visual context. Following the benchmark's scoring conventions, we report figure-level accuracy (fACC), question-pair accuracy (qACC), easy- and hard-subset group accuracies (easyA and hardA), and overall individual-question accuracy (aACC). Figure and question groups receive credit only when all their included answers are correct. We use GPT-4o-mini to judge open-ended responses, as described in Section 4.1.

MME (Fu et al., 2026) evaluates multimodal models across perception and cognition. Its perception tasks cover object existence, counting, spatial position, color, posters, celebrities, scenes, landmarks, artwork, and optical character recognition. The cognition tasks include commonsense reasoning, numerical calculation, text translation, and code reasoning. Questions are formulated as binary yes/no queries. We follow the official scoring protocol, which combines question-level accuracy and the accuracy of answering both questions for an image correctly. We report Perception and Cognition totals, as well as the MME-Hall subset, which sums the scores for Existence, Count, Position, and Color.

B.2 Backbones and Baselines

Backbones. We evaluate AEDG on three representative multimodal large language models: LLaVA-1.5-7B (Liu et al., 2024), Qwen-VL-Chat (Bai et al., 2023), and GLM-4V-9B (Glm et al., 2024). These backbones differ in their visual encoders, multimodal connectors, and language-model architectures. All experiments start from the released checkpoints; AEDG introduces neither fine-tuning nor parameter updates.

Baselines and decoding. We compare against vanilla decoding and three training-free hallucination-mitigation methods: VCD (Leng et al., 2024), DeCo (Wang et al., 2025), and MemVR (Zou et al., 2024). These methods represent visual contrastive decoding, intermediate-layer logit correction, and memory-space visual reinforcement, respectively. As described in Section 4.1, we use their official implementations and recommended method-specific hyperparameters, keeping the backbone, prompt, and decoding configuration identical across methods. AEDG uses greedy decoding with `do_sample=False` and temperature 0; its default calibration and candidate-filtering parameters are given in Section 4.1.

C LIMITATIONS

AEDG currently requires access to per-layer attention weights and identifiable visual-token spans, which limits its applicability to models that do not expose these quantities. Our experiments cover three open-source MLLMs and object-centric hallucination benchmarks. Extending AEDG to larger models, video and multi-image inputs, and more complex reasoning tasks, as well as exploring finer-grained or head-selective entropy signals, remains future work.