

From Token Correction to Claim Admission: Training-Free Hallucination Mitigation for Large Vision-Language Models

Hyeonggeun Han
Independent Researcher
hygnhan@gmail.com

Taehyun Cho
Vector Institute
Fields Institute
Seoul National University
talium@cml.snu.ac.kr

Sangwoo Hong
Konkuk University
swhong06@konkuk.ac.kr

Abstract

Large vision-language models often exhibit object hallucination, generating claims that are not supported by the input image. To mitigate this issue, we propose Evidence-Guided Visual Claim Admission (EGVCA), a training-free framework that admits a semantic claim only when it receives sufficient image-dependent support. EGVCA measures this support by comparing the claim likelihood under the observed image and a visual-null input. For captioning, it controls semantic expansion without altering token generation. For binary visual question answering, it combines atomic visual support with the image-induced change in the model’s native Yes–No margin. Experiments on CHAIR, POPE, and MME across three LVLM backbones show that EGVCA substantially reduces object hallucination while preserving caption quality. It achieves the highest POPE F1 in eight of nine backbone–sampling settings and improves the native MME total score on all three backbones. These results demonstrate the effectiveness of claim-level admission as a simple alternative to token-level distribution correction.

1 Introduction

Large Vision-Language Models (LVLMs) have demonstrated remarkable capabilities in multi-modal understanding and generation, enabling a broad range of applications that require interaction with visual environments. Despite this progress, their reliability remains fundamentally challenged by object hallucination (OH), where generated responses contain object-centric information that is inconsistent with the visual input (Li et al., 2023b; Wang et al., 2023, 2026). Such hallucinations can manifest in various forms, including mentioning objects that are not present, describing incorrect attributes or properties of visible objects, or asserting relationships between objects that are not supported by the image. These errors can remain

linguistically plausible even when they are visually ungrounded, making them particularly difficult to detect from the generated response alone. As LVLMs are increasingly expected to provide faithful interpretations of visual scenes, mitigating OH is therefore essential for improving their reliability and enabling trustworthy real-world deployment.

Numerous approaches have been proposed to reduce hallucinations through additional training, external supervision, representation intervention, and decoding-time correction. Among them, training-free decoding methods have received particular attention because they can be directly applied to pre-trained LVLMs without additional training data or parameter updates. A major line of work improves visual grounding by contrasting predictions obtained under different visual or conditional contexts (Leng et al., 2024; Favero et al., 2024), while another exploits internal decoding signals, such as attention patterns, intermediate-layer predictions, or visual-token and layer-wise information, to suppress hallucination-prone outputs (Huang et al., 2024; Wang et al., 2025; Huo et al., 2025; Liang et al., 2025). More recent approaches further suggest that correcting generation selectively according to the visual support of emerging content can avoid unnecessary intervention (Yang et al., 2026). Despite their effectiveness, existing methods primarily mitigate hallucination by modifying token distributions or intervening on particular internal signals during decoding. Such correction can suppress hallucination-prone generation, but modifying the generation process itself may also alter visual information that is already well grounded. As a result, the model may favor safer but less semantically specific responses. We therefore take a different perspective. Rather than directly correcting how a claim is generated, we first ask whether the semantic claim itself has sufficient visual evidence to be admitted into the response. This separation allows hallucination mitigation to focus on

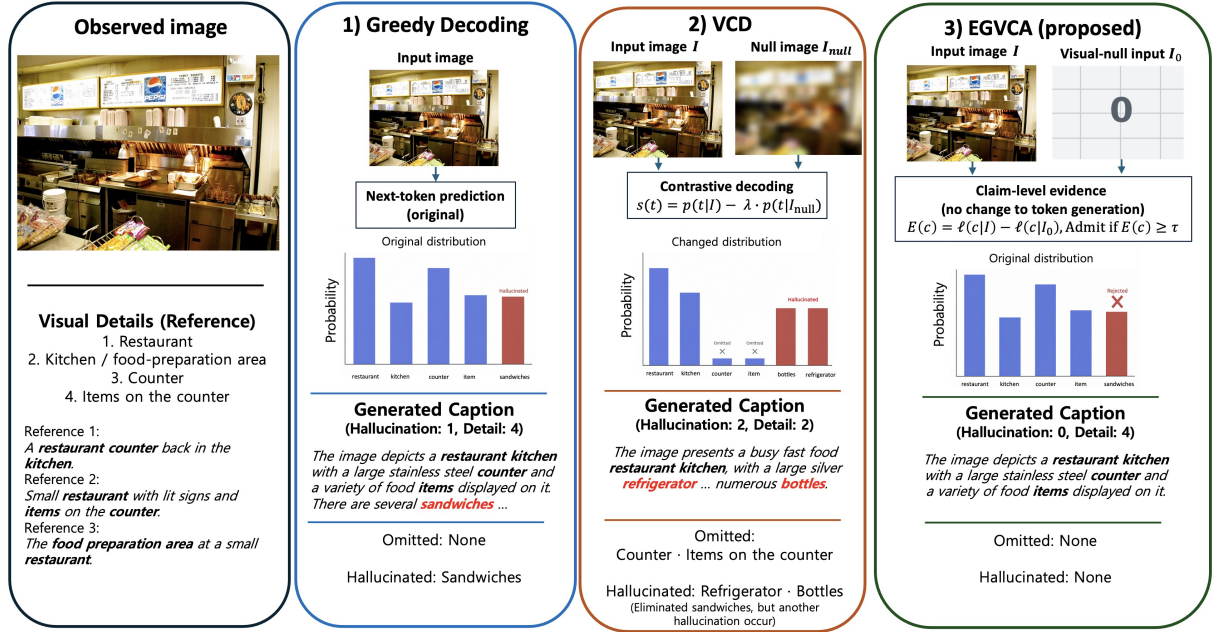


Figure 1: Comparison of greedy decoding, VCD, and EGVCA. Unlike VCD, which modifies next-token predictions, EGVCA uses original-versus-visual-null evidence to reject unsupported claims without altering candidate generation.

unsupported semantic commitments while leaving grounded generation otherwise unchanged.

Motivated by this perspective, we propose **Evidence-Guided Visual Claim Admission (EGVCA)**, a training-free framework that formulates hallucination mitigation as visual claim admission. As illustrated in Figure 1, EGVCA is based on a simple principle: a visual claim should be incorporated into the response only when its plausibility can be meaningfully attributable to the observed image. We estimate this support by comparing the normalized likelihood of the same claim under the observed image and a visual-null input, implemented as an all-zero preprocessed visual tensor. This counterfactual comparison isolates image-dependent evidence from plausibility induced by linguistic context alone.

Unlike contrastive decoding methods that directly modify the next-token distribution, EGVCA interprets the resulting contrast as evidence for a semantic claim. It then uses this evidence to regulate whether the model commits to that claim, while leaving the underlying candidate generation unchanged. This separation between claim-level evidence estimation and claim admission provides a unified mechanism for suppressing visually unsupported content while retaining supported semantic information. Experiments across three LVLMM backbones on CHAIR, POPE, and MME show that EGVCA consistently reduces object hallucination

while maintaining competitive generation quality.

2 Related Work

Large Vision-Language Models. LVLMMs have rapidly advanced by integrating visual encoders with large language models, achieving strong performance across a wide range of multimodal understanding and generation tasks (Li et al., 2023a; Liu et al., 2023; Zhu et al., 2024; Ye et al., 2024; Bai et al., 2023). Recent progress in large-scale multimodal pretraining and visual instruction tuning has further improved their ability to follow natural-language instructions and reason over complex visual content. As a result, LVLMMs have become a central foundation for general-purpose multimodal systems. Despite these advances, imperfect visual grounding remains a persistent limitation, causing LVLMMs to generate semantically plausible content that is inconsistent with the visual input.

Object Hallucination in LVLMMs. Object hallucination (OH) refers to the generation of object-centric content that is not faithfully grounded in the given image (Rohrbach et al., 2018; Li et al., 2023b). It can appear as incorrect object identification as well as erroneous attributes, quantities, or relations associated with visual entities (Wang et al., 2023). Prior studies have linked OH to biased object co-occurrences, strong language priors, insufficient vision-language alignment, and

the diminishing influence of visual information during autoregressive generation (Li et al., 2023b; Leng et al., 2024; Favero et al., 2024). To address these issues, various mitigation strategies have been explored, with training-free inference-time methods receiving particular attention for improving visual faithfulness without modifying the underlying LVLM.

Training-Free Decoding for Hallucination Mitigation. Training-free decoding provides an efficient means of mitigating hallucinations without retraining LVLMs. Many approaches derive corrective signals from the predictions or internal computations of a frozen model. VCD contrasts predictions under the original and distorted visual inputs to suppress language priors, while OPERA penalizes over-trusted attention patterns during decoding (Leng et al., 2024; Huang et al., 2024). DeCo uses visual information preserved in intermediate layers to correct next-token predictions, and MoLE combines predictions from multiple layer-wise experts to improve visual grounding (Wang et al., 2025; Liang et al., 2025). Recent analyses further show that reliance on linguistic context tends to increase as generation proceeds, while visual dependency can weaken during later decoding stages (Min et al., 2025; Anand et al., 2026). This tendency motivates explicit control over the visual support of later generated content. The use of such internal signals is also connected to broader research that seeks to understand and control model behavior through interpretable internal features, mechanistic circuits, role-aware information flow, confidence-aware fusion, and evidence-grounded structured representations (Huben et al., 2024; Lee et al., 2026b,a; Yoon et al., 2026; Lee et al., 2025). EGVCA shares the broader goal of deriving informative signals from a frozen model, but differs in how these signals are used. Rather than directly modifying next-token predictions, EGVCA interprets the original-versus-null likelihood contrast as evidence for a semantic claim and uses it to determine whether that claim should be admitted.

3 Claim-Level View of Object Hallucination

Given an input image I and a textual instruction x , a pretrained LVLM parameterized by θ generates a response y according to $p_\theta(y \mid I, x)$. Rather than viewing hallucination solely as an error at the token level, we consider the semantic commitments

expressed in the response. Specifically, we define y as a sequence of semantic claims $y = (c_1, \dots, c_T)$, where each c_t may describe the existence of an object, an attribute of a visible object, or a relation among visual entities. Letting $h_t = (c_1, \dots, c_{t-1})$ denote the preceding claims, with h_1 being empty, we express the response likelihood as

$$p_\theta(y \mid I, x) = \prod_{t=1}^T p_\theta(c_t \mid I, x, h_t). \quad (1)$$

Under this formulation, object hallucination arises when the model introduces a semantic commitment that is not sufficiently grounded in the observed image.

A central difficulty is that the likelihood of a claim does not by itself indicate whether the claim is visually grounded. A claim may receive high probability because it is linguistically plausible or favored by the model’s learned priors even when the image provides little evidence for it. We therefore distinguish the overall plausibility of a claim from the extent to which that plausibility increases in the presence of the observed image. To this end, we consider a visual-null counterpart I_0 of the input image I , under which task-relevant visual information is removed. This motivates comparing the likelihood of the same claim under I and I_0 through the log-likelihood contrast

$$\log p_\theta(c_t \mid I, x, h_t) - \log p_\theta(c_t \mid I_0, x, h_t). \quad (2)$$

A positive contrast indicates that the observed image increases the model-assigned likelihood of the claim relative to the visual-null input. This comparison provides a basis for assessing image-dependent evidence beyond the claim’s overall plausibility.

Crucially, we do not formulate this image-dependent likelihood difference as a corrective signal for directly modifying next-token probabilities. Instead, we treat it as evidence for deciding whether a semantic commitment should be admitted into the model response. This shifts the target of intervention from token generation itself to the model’s commitment to semantic content: generation remains unchanged for content that is sufficiently grounded, while visually unsupported commitments can be regulated at the admission stage. We consider this principle in both open-ended generation and binary visual question answering. Although the form of the response and

the corresponding admission mechanism differ between the two settings, both instantiate the same underlying problem of determining whether a semantic commitment is sufficiently attributable to the observed image. Section 4 describes how this likelihood comparison is operationalized as evidence and how claim admission is implemented in each setting.

4 Evidence-Guided Visual Claim Admission

Building on the distinction between claim plausibility and image-dependent evidence in Section 3, we propose Evidence-Guided Visual Claim Admission (EGVCA), a training-free framework for regulating semantic commitments in LVLM responses. EGVCA uses the likelihood contrast in Eq. 2 to estimate visual evidence and guide claim admission. We first define a claim-level visual evidence score and specify the visual-null input used to compute it. We then describe how this evidence determines the retained response prefix in open-ended generation and adjusts the final Yes–No decision in binary visual question answering.

4.1 Image-Attributable Visual Evidence

EGVCA normalizes the claim-level likelihood contrast by token count to reduce its dependence on claim length. Using the preceding context h_t defined in Section 3, we compute the average token log-likelihood of a claim as

$$\ell_\theta(c_t | I, x, h_t) = \frac{\sum_{j=1}^{|c_t|} \log p_\theta(c_{t,j} | I, x, h_t, c_{t,<j})}{|c_t|}, \quad (3)$$

where $|c_t|$ is the number of tokens in c_t , $c_{t,j}$ denotes its j -th token, and $c_{t,<j}$ denotes its preceding tokens. Averaging places claims of different lengths on a per-token scale. We then define the image-attributable evidence as

$$E(c_t) = \ell_\theta(c_t | I, x, h_t) - \ell_\theta(c_t | I_0, x, h_t). \quad (4)$$

Thus, $E(c_t)$ is the log-likelihood contrast in Eq. 2 divided by $|c_t|$.

The visual-null input I_0 is constructed by replacing the backbone-specific preprocessed visual input with a zero-valued tensor of the same shape. This construction removes the image’s scene content without substituting another scene or introducing a randomly generated visual pattern. It therefore provides a deterministic reference for measuring

the change in claim likelihood induced by the observed image. Accordingly, we interpret $E(c_t)$ as evidence relative to this reference, rather than as an absolute measure of visual correctness.

A positive $E(c_t)$ indicates that the observed image increases the claim’s average token log-likelihood relative to I_0 . A value near zero indicates little change, whereas a negative value indicates that the claim is less likely under the observed image than under the reference. EGVCA uses this score to evaluate claim admission, without feeding it back into the next-token distribution.

4.2 Claim Admission for Open-Ended Generation

For open-ended generation, EGVCA applies evidence-based admission to a candidate response produced by the LVLM’s native decoding strategy. We partition this response into the claim sequence described in Section 3 and evaluate the claims in their original order. Each claim is defined as a sentence-like span separated by terminal punctuation. Admission retains a prefix of this sequence without rewriting individual claims.

The first claim c_1 is retained without an evidence check to prevent an empty response. Each subsequent claim is evaluated using $E(c_t)$ conditioned on its preceding context h_t and admitted according to

$$\text{Admit}(c_t) = \mathbb{I}[E(c_t) \geq \tau], \quad t \geq 2, \quad (5)$$

where $\mathbb{I}[\cdot]$ is the indicator function and τ is the minimum evidence required to retain an additional claim. This rule uses image-attributable evidence to determine whether the response should extend beyond the claims already retained.

Admission stops at the first claim that fails the criterion. If c_t is the first rejected claim, EGVCA discards it together with all subsequent claims and returns

$$\hat{y} = (c_1, \dots, c_{t-1}). \quad (6)$$

If no claim is rejected, the full candidate response is retained. Stopping at the first rejection preserves the preceding context under which each retained claim was generated and evaluated. It also avoids retaining later claims that may depend on rejected content. The threshold τ controls how much of the response is admitted beyond the initial claim. Increasing τ makes admission stricter, while decreasing it allows a longer prefix to be retained.

4.3 Evidence-Guided Claim Admission for Binary VQA

In binary visual question answering, claim admission concerns the proposition expressed by a question q . Answering Yes affirms this proposition, whereas answering No rejects it. EGVCA combines evidence about the question’s visual components with evidence about the complete proposition, then uses the combined score to adjust the model’s final decision.

Component-level evidence measures how strongly the image increases the likelihood of the visual content mentioned in the question. We decompose q into K visual components $a_1, \dots, a_K$, corresponding to objects, attributes, or relations involved in the proposition. Visual components are extracted using a rule-based parser with predefined lexical variants, without requiring a trained semantic parser. Each component is scored under a fixed neutral visual-description context x_{neu} , which requests visual description without posing a Yes–No question. This context allows us to evaluate the component’s textual content outside the original binary question framing. For each component, we compute

$$e(a_k) = \ell_{\theta}(a_k | I, x_{\text{neu}}) - \ell_{\theta}(a_k | I_0, x_{\text{neu}}), \quad (7)$$

using the token-averaged log-likelihood in Eq. 3, with no preceding claim context. When multiple equivalent textual forms are considered for a component, we use the form with the largest evidence score. We aggregate the component scores as

$$V(q) = \frac{1}{K} \sum_{k=1}^K e(a_k). \quad (8)$$

Averaging reduces the dependence of the aggregate score on the number of components.

Component-level evidence alone does not establish whether the complete proposition is supported. For example, evidence for a dog, a couch, and a sitting action does not necessarily establish that the dog is sitting on the couch. EGVCA therefore includes a second signal that measures how the image changes the model’s preference for affirming the question as a whole. We compute the native Yes–No margin under the observed image as

$$m_I(q) = \log p_{\theta}(\text{Yes} | I, q) - \log p_{\theta}(\text{No} | I, q), \quad (9)$$

and the corresponding margin under the visual-null input as

$$m_0(q) = \log p_{\theta}(\text{Yes} | I_0, q) - \log p_{\theta}(\text{No} | I_0, q). \quad (10)$$

The proposition-level evidence is their difference,

$$M(q) = m_I(q) - m_0(q). \quad (11)$$

A positive $M(q)$ indicates that the observed image shifts the model’s relative preference toward Yes, whereas a negative value indicates a shift toward No. Unlike $V(q)$, this score evaluates the image’s effect on the decision for the complete question.

EGVCA combines these complementary signals through an equally weighted joint evidence score,

$$J(q) = \frac{1}{2}V(q) + \frac{1}{2}M(q). \quad (12)$$

Here, $V(q)$ contributes evidence for the visual components, while $M(q)$ contributes evidence for affirming the complete proposition. The joint score adjusts the native decision margin through

$$m'(q) = m_I(q) + \lambda (J(q) - \tau), \quad (13)$$

where $\lambda \geq 0$ controls the strength of the adjustment and τ is the evidence reference level for binary VQA. Evidence above τ increases the margin toward Yes, while evidence below τ decreases it. Unlike the hard threshold used for open-ended generation, τ is not a standalone admission condition. The final decision depends jointly on the native margin and the evidence adjustment.

The final answer is

$$\hat{a} = \begin{cases} \text{Yes}, & m'(q) > 0, \\ \text{No}, & m'(q) \leq 0. \end{cases} \quad (14)$$

This rule applies the evidence adjustment at the final answer-selection stage, leaving the LVLM probabilities used to compute the scores unchanged. Both response settings therefore use visual evidence to regulate semantic commitment, through prefix admission in open-ended generation and decision-level adjustment in binary VQA.

5 Experiments

5.1 Experimental Setup

Models. We evaluate EGVCA on three pre-trained LVLM backbones, LLaVA-1.5-7B (Liu et al., 2024), MiniGPT-4 with Vicuna-13B (Zhu et al., 2024), and mPLUG-Owl2-7B (Ye et al., 2024). All model parameters remain frozen throughout evaluation.

Table 1: CHAIR-500 comparison across three LVLM backbones.

Method	LLaVA-1.5		MiniGPT-4		mPLUG-Owl2	
	CHAIR _S ↓	CHAIR _I ↓	CHAIR _S ↓	CHAIR _I ↓	CHAIR _S ↓	CHAIR _I ↓
Greedy	21.93 ± 1.79	7.07 ± 0.59	20.33 ± 0.52	7.26 ± 0.39	21.47 ± 1.00	7.12 ± 0.39
Beam	18.73 ± 2.49	6.23 ± 0.37	20.00 ± 1.88	7.20 ± 0.48	19.27 ± 1.59	6.61 ± 0.39
VCD	23.87 ± 0.25	7.91 ± 0.27	23.73 ± 0.90	9.02 ± 0.07	25.40 ± 2.32	8.77 ± 0.73
OPERA	18.00 ± 1.02	6.02 ± 0.19	17.80 ± 1.57	7.08 ± 0.35	17.73 ± 0.74	6.21 ± 0.16
DoLa	17.27 ± 1.51	8.72 ± 2.08	15.07 ± 1.89	7.08 ± 0.56	18.87 ± 1.23	6.44 ± 0.23
DeCo	16.20 ± 1.34	6.16 ± 0.65	17.80 ± 1.31	6.75 ± 0.46	17.33 ± 1.79	7.33 ± 1.07
SID	22.13 ± 0.98	6.50 ± 0.31	22.93 ± 2.66	8.65 ± 0.89	25.40 ± 1.85	8.68 ± 0.46
MoLE	18.13 ± 1.31	6.51 ± 0.59	17.80 ± 1.28	6.50 ± 0.28	18.53 ± 0.25	6.21 ± 0.18
HALC	23.07 ± 1.75	8.52 ± 0.62	17.93 ± 1.15	6.31 ± 0.61	20.00 ± 1.84	6.83 ± 0.33
EGVCA	10.87 ± 1.43	5.77 ± 0.61	12.73 ± 0.66	6.19 ± 0.11	12.60 ± 1.61	5.63 ± 0.70

Benchmarks and Evaluation Metrics. We evaluate object hallucination in open-ended generation using CHAIR (Rohrbach et al., 2018) on 500 images sampled from the MS COCO 2014 validation set (Lin et al., 2014) per run, with three independent sampling seeds for all three backbones. We report CHAIR_S and CHAIR_I to measure hallucination at the caption and object-mention levels, respectively. Caption quality is assessed using BLEU-4 (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), and SPICE (Anderson et al., 2016) to characterize retained content. For binary VQA, we evaluate the Random, Popular, and Adversarial settings of MS COCO POPE (Li et al., 2023b), reporting accuracy and F1 for object-existence judgments. We additionally report perception, cognition, and total scores on the full MME benchmark (Fu et al., 2025) to assess broader multimodal capabilities.

Baselines. We compare EGVCA with standard greedy decoding, beam search, and the native Yes–No decision, together with seven inference-time methods, VCD (Leng et al., 2024), OPERA (Huang et al., 2024), DoLa (Chuang et al., 2024), DeCo (Wang et al., 2025), SID (Huo et al., 2025), MoLE (Liang et al., 2025), and HALC (Chen et al., 2024). These baselines allow us to compare claim admission with alternative inference-time interventions for improving factuality or mitigating hallucination without updating the evaluated LVLM.

Implementation Details. For open-ended generation, EGVCA applies claim admission to greedy responses with $\tau = 0.25$ for LLaVA-1.5, $\tau = 0.30$ for MiniGPT-4, and $\tau = 0.125$ for mPLUG-Owl2. For binary VQA, a single calibration parameter pair is selected for each backbone–benchmark set-

ting using a predefined parameter sweep and is applied uniformly to all examples, using a held-out validation set. Appendix A.3 reports the local sensitivity around the selected operating points. CHAIR results are reported as the mean and standard deviation over three seeds, while caption-quality results are reported as three-seed means. All experiments were conducted on a single NVIDIA GeForce RTX 5090 GPU with 32 GB of memory.

5.2 Experimental Results

Open-ended Object Hallucination. Table 1 demonstrates the effectiveness of EGVCA in reducing object hallucination in open-ended generation. EGVCA achieves the lowest CHAIR_S in each backbone comparison. On LLaVA-1.5, it obtains a score of 10.87, outperforming DeCo, the strongest baseline on this metric, by 5.33 percentage points. On MiniGPT-4 and mPLUG-Owl2, EGVCA reduces CHAIR_S from 20.33 to 12.73 and from 21.47 to 12.60, respectively, while also improving CHAIR_I. This result implies that a larger proportion of the resulting captions contain no hallucinated object mentions. This outcome aligns with the purpose of claim admission, which determines whether additional semantic content should enter the response. The improvements across all three backbones support the applicability of this principle to different LVLMs without modifying their underlying token-generation rules.

Binary Visual Question Answering. Table 2 shows that EGVCA improves object-presence prediction across different backbones and negative-sampling settings. EGVCA improves F1 over native predictions and achieves the highest F1 among the evaluated methods in eight of the nine backbone–setting combinations. Applied to

Table 2: POPE results under Random, Popular, and Adversarial sampling across three LVLM backbones.¹

Method	LLaVA-1.5						MiniGPT-4						mPLUG-Owl2					
	Random		Popular		Adversarial		Random		Popular		Adversarial		Random		Popular		Adversarial	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Native	87.20	85.69	85.83	84.40	83.47	82.25	74.83	72.56	71.40	69.94	69.50	68.70	88.67	88.59	84.77	85.24	80.90	82.12
VCD	84.23	82.50	82.37	80.62	79.33	77.97	58.70	58.74	57.00	57.26	57.60	58.94	83.97	83.39	81.17	81.30	79.00	79.48
OPERA	86.23	84.42	85.17	83.41	82.87	81.27	74.50	68.71	72.03	66.69	69.67	65.00	86.83	85.80	84.40	83.60	80.80	80.51
DoLa	88.43	87.44	86.37	85.52	83.57	83.03	-	-	-	-	-	-	86.67	85.55	84.47	83.56	81.07	80.61
DeCo	88.53	87.60	86.33	85.56	83.37	82.95	52.57	67.78	51.77	67.42	51.83	67.45	86.47	86.68	82.30	83.27	78.03	80.04
SID	89.40	89.05	86.10	86.11	80.47	81.49	59.07	57.74	58.13	57.07	56.57	55.39	84.03	83.60	80.70	80.87	77.10	77.99
MoLE	88.87	88.11	86.87	86.27	83.20	83.05	55.67	21.12	55.23	20.95	55.03	21.25	86.93	85.89	84.50	83.69	81.00	80.68
HALC	88.50	87.52	86.43	85.60	83.57	83.05	-	-	-	-	-	-	86.53	85.85	83.87	83.51	80.20	80.46
EGVCA	90.43	90.05	87.60	87.47	80.97	81.94	76.80	78.45	76.07	77.92	69.40	73.41	90.37	90.11	87.37	87.42	82.40	83.27

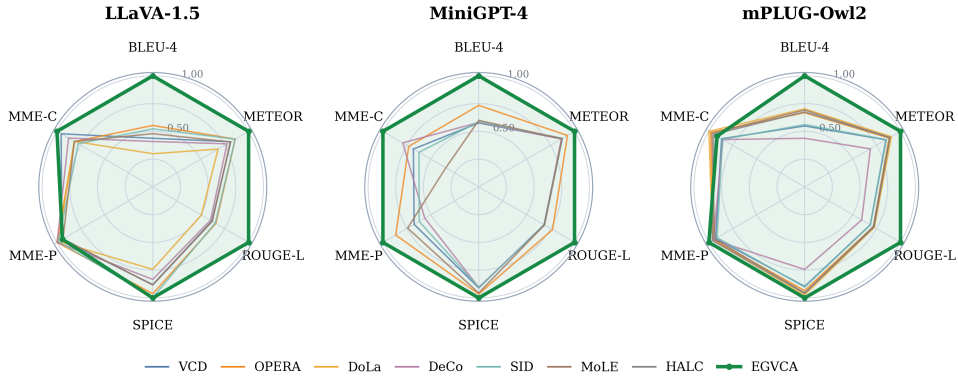


Figure 2: General capability profiles across three LVLM backbones. BLEU-4, METEOR, ROUGE-L, and SPICE are evaluated on COCO captions, whereas MME-P and MME-C are evaluated on the separate MME benchmark. Each metric is normalized by the best matched result for the corresponding backbone and metric. Greedy and Beam are omitted for clarity.

LLaVA-1.5, EGVCA obtains the best accuracy and F1 in both the Random and Popular settings. On MiniGPT-4, it achieves the highest F1 across all three settings, with Popular F1 increasing from 69.94 to 77.92. On mPLUG-Owl2, EGVCA leads in both accuracy and F1 under every sampling setting. These results extend the evidence for claim admission beyond open-ended descriptions. Because POPE requires a binary response, its F1 improvements demonstrate that the benefit of EGVCA also applies to deciding whether an object-existence claim should be affirmed, beyond controlling how much descriptive content is generated.

General Capability Preservation. Figure 2 jointly summarizes caption quality and MME performance. For captioning, EGVCA achieves the best matched BLEU-4, METEOR, ROUGE-L, and SPICE scores on all three backbones. On LLaVA-1.5, SPICE increases from 21.44 with greedy decoding to 22.19, while EGVCA obtains SPICE scores of 20.45 and 22.55 on MiniGPT-4 and

mPLUG-Owl2, respectively. The consistent improvements across all four reference-based metrics indicate that hallucination reduction does not come at the expense of semantic caption quality.

EGVCA also improves MME Total over the native model on all three backbones, with the largest gain of 249.21 points on MiniGPT-4. It obtains the highest cognition score on LLaVA-1.5, the best perception and cognition scores on MiniGPT-4, and the highest perception and total scores on mPLUG-Owl2. Together, these results show that EGVCA preserves broader visual understanding while reducing hallucination.

5.3 Ablation and Analysis.

We analyze the key design choices of EGVCA on LLaVA-1.5. Specifically, we examine how the caption admission threshold affects the trade-off be-

¹DoLa and HALC yielded unstable binary outputs with MiniGPT-4 on both POPE and MME under the matched protocol, indicating sensitivity to the backbone-decoding interaction.

Table 3: Sensitivity of caption admission to τ on LLaVA-1.5. Results are the mean and standard deviation over three CHAIR-500 runs.

Setting	CHAIR _S ↓	CHAIR _I ↓	SPICE ↑
Greedy	21.93±1.79	7.07±0.59	21.44±0.30
$\tau = 0$	20.47±1.48	6.96±0.56	21.67±0.25
$\tau = 0.125$	14.73±1.47	6.47±0.57	22.23±0.14
$\tau = 0.25$ (Ours)	10.87±1.43	5.77±0.61	<u>22.19±0.24</u>
$\tau = 0.5$	<u>7.47±1.37</u>	<u>4.56±0.80</u>	21.73±0.25
$\tau = 1.0$	6.47±1.11	4.24±0.61	21.27±0.25

Table 4: Binary component ablation on full POPE and MME using LLaVA-1.5.

Variant	POPE F1 ↑	MME-P ↑	MME-C ↑	MME Total ↑
Native LVLM	84.11	1360.14	321.79	1681.92
w/o $M(q)$	81.93	<u>1371.71</u>	361.79	<u>1733.50</u>
w/o $V(q)$	<u>85.33</u>	1360.32	321.79	1682.11
w/o $m_I(q)$	84.95	498.33	204.29	702.62
Full EGVCA	86.48	1373.34	366.79	1740.13

tween hallucination reduction and semantic quality, assess the contributions of the visual-content and native-margin evidence components to binary decisions, and report the inference efficiency of the resulting framework.

Caption Admission Threshold. Table 3 reports the sensitivity to τ on CHAIR. Increasing τ consistently reduces both CHAIR_S and CHAIR_I, confirming that the admission threshold directly controls the strictness of semantic expansion. The best SPICE score is obtained at $\tau = 0.125$, while $\tau = 0.25$ substantially reduces hallucination with only a negligible SPICE difference (22.19 versus 22.23). We therefore select $\tau = 0.25$ as the final operating point. Relative to greedy generation, it reduces CHAIR_S from 21.93 to 10.87 and CHAIR_I from 7.07 to 5.77, while increasing SPICE from 21.44 to 22.19. More aggressive thresholds further reduce hallucination, but also decrease SPICE, demonstrating a trade-off between hallucination suppression and semantic quality.

Binary Evidence Components. Table 4 separates atomic visual support $V(q)$ from native margin gain $M(q)$. Using $V(q)$ alone substantially improves MME performance but degrades POPE, whereas $M(q)$ alone improves POPE while leaving MME nearly unchanged. This indicates that the two components capture complementary aspects of visual evidence. Using the joint evidence as an independent classifier performs poorly, particularly on MME, showing that counterfactual evidence

Table 5: Inference efficiency on LLaVA-1.5 over CHAIR images.

Method	Tokens/sec. ↑	Peak VRAM (GB) ↓
Greedy	66.61	15.27
Beam	42.78	22.35
VCD	33.56	16.37
OPERA	5.43	25.45
DoLa	19.69	26.31
DeCo	<u>59.10</u>	<u>15.75</u>
SID	33.49	16.36
MoLE	31.00	29.26
HALC	1.45	19.54
EGVCA	21.90	15.27

should calibrate rather than replace the LVLM’s native decision. Combining the native margin with joint evidence in the full admission head achieves the strongest joint result, improving POPE F1 from 84.11 to 86.48 and MME Total from 1681.92 to 1740.13.

Inference Efficiency. Table 5 reports end-to-end throughput and peak GPU memory over 200 CHAIR images. EGVCA processes 21.90 final-output tokens per second while using 15.27 GB of peak memory. Its throughput exceeds DoLa, OPERA, and HALC, while its memory usage matches greedy decoding. Detailed latency measurements and the EGVCA runtime breakdown are provided in the appendix.

6 Conclusion

We presented EGVCA, a training-free framework that admits visual claims according to their evidence under the observed image relative to a visual-null input. Across three LVLM backbones, EGVCA reduces CHAIR hallucination, improves POPE performance, and preserves or improves caption quality and MME performance. These results demonstrate the effectiveness of claim-level admission for controlling visual hallucination.

Limitations

EGVCA requires additional original and visual-null scoring, resulting in higher inference cost than standard greedy decoding. Our experiments focus on three LVLM backbones and widely used English-language benchmarks. Moreover, because EGVCA mitigates rather than verifies model outputs, residual errors exist in the resulting responses. It should therefore not be treated as a factuality guarantee in high-stakes settings.

References

- Neeraj Anand, Samyak Jha, Udbhav Bamba, and Rahul Rahaman. 2026. Crops: A training-free hallucination mitigation framework for vision-language models. *arXiv preprint arXiv:2601.00659*.
- Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. [Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond](#). *Preprint*, arXiv:2308.12966.
- Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.
- Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024. HALC: Object hallucination reduction via adaptive focal-contrast decoding. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 7824–7846.
- Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. [Dola: Decoding by contrasting layers improves factuality in large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14303–14312.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2025. MME: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38.
- Hyeonggeun Han, Sehwan Kim, Hyungjun Joo, Sangwoo Hong, and Jungwoo Lee. 2025. Adjusting initial noise to mitigate memorization in text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 38:684–711.
- Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. 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 (CVPR)*, pages 13418–13427.
- Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, volume 2024, pages 7827–7845.
- Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. 2025. Self-introspective decoding: Alleviating hallucinations for large vision-language models. In *International Conference on Learning Representations*, volume 2025, pages 24272–24295.
- Jaewook Lee, Woojin Lee, Oh-Woog Kwon, and Harksoo Kim. 2025. [Do large language models have “emotion neurons”? investigating the existence and role](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15617–15639. Association for Computational Linguistics.
- Jungseob Lee, Jaehyung Seo, and Heuiseok Lim. 2026a. The hallucination signal is a mean shift: Why simple probes suffice. *arXiv preprint arXiv:2608.28930*.
- Nakyung Lee, Sangwoo Hong, and Jungwoo Lee. 2026b. Mechanistic circuit identification for controllable data generation. *arXiv preprint arXiv:2608.24065*.
- Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13872–13882.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 19730–19742.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 292–305.
- Tian Liang, Yuetian Du, Jing Huang, Ming Kong, Luyuan Chen, Yadong Li, Siye Chen, and Qiang Zhu. 2025. Mole:decoding by mixture of layer experts alleviates hallucination in large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18684–18692.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.
- Kyungmin Min, Minbeom Kim, Kang-il Lee, Dongryeol Lee, and Kyomin Jung. 2025. Mitigating hallucinations in large vision-language models via summary-guided decoding. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4183–4198.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4035–4045.
- Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. 2025. [MLLM can see? dynamic correction decoding for hallucination mitigation](#). In *The Thirteenth International Conference on Learning Representations*.
- Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, et al. 2023. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.
- Tianbo Wang, Yuqing Ma, Kewei Liao, Zhange Zhang, Simin Li, Jinyang Guo, and Xianglong Liu. 2026. [AFTER: Mitigating the object hallucination of LVLm via adaptive factual-guided activation editing](#). In *The Fourteenth International Conference on Learning Representations*.
- Lei Yang, Xinze Liu, Dayan Wu, Ding Wang, Hengjie Zhu, Zihao Zhang, Tianzhu Hu, Hanqi Wu, Peng Fu, and Zheng Lin. 2026. Hallucinations leave a grounding signature: Verifier-guided decoding for selective object correction. *arXiv preprint arXiv:2607.27823*.
- Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13040–13051.
- Yeo-Chan Yoon, Sookyun Kim, and Gyeongmin Kim. 2026. Adaptive orthogonal role-aware transformer for robust instruction alignment. *ETRI Journal*.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. [MiniGPT-4: Enhancing vision-language understanding with advanced large language models](#). In *The Twelfth International Conference on Learning Representations*.

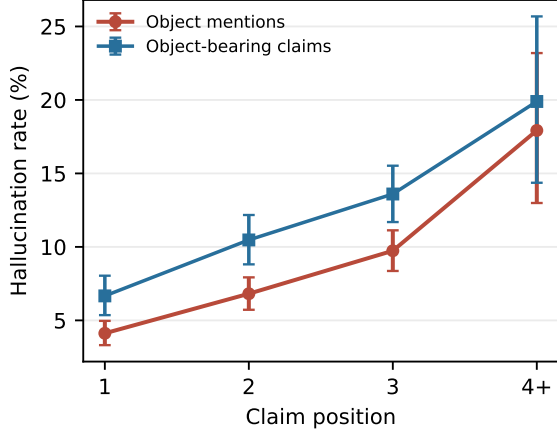


Figure 3: Object hallucination across claim positions in LLaVA-1.5 Greedy captions. Error bars indicate 95% confidence intervals obtained using image-grouped bootstrap resampling. Positions four and later are pooled.

A Additional Analyses

A.1 Hallucination across Claim Positions

Previous studies report that LVLMs increasingly rely on linguistic context as generation proceeds, while their dependence on visual information tends to weaken in later decoding stages (Min et al., 2025; Anand et al., 2026). We examine whether the same tendency appears across semantic claims in our setting. The analysis uses the frozen LLaVA-1.5 Greedy captions from three CHAIR-500 runs, comprising 1,500 captions and 4,989 claims. We include all candidate claims, including those occurring after the first operational rejection.

We report two complementary rates. The object-mention rate is the proportion of recognized COCO object mentions labeled as hallucinated by the CHAIR evaluator. The claim-level rate is the proportion of object-bearing claims that contain at least one hallucinated object. Claims at position four and later are pooled because their individual sample counts are smaller.

As shown in Figure 3, the object-mention hallucination rate increases from 4.12% in the first claim to 6.81% in the second, 9.74% in the third, and 17.92% at position four or later. The proportion of object-bearing claims containing a hallucinated object similarly increases from 6.67% to 10.47%, 13.60%, and 19.90%. These results confirm that later claims are increasingly susceptible to object hallucination in our generated captions, supporting the motivation for controlling the admission of additional semantic content.

A.2 Visual-Null Construction

We compare EGVCA’s zero tensor with VCD-style diffusion noise, a CLIP-normalized black image, and a randomly shuffled COCO image. Captions are always generated from the original image, and only the reference used for evidence scoring is changed. Since diffusion noise can be sensitive to its configuration (Han et al., 2025), we report it as a separate reference condition. As shown in Table 6, the normalized-black reference behaves similarly to the zero tensor, with slightly lower CHAIR scores but slightly lower SPICE. In contrast, the diffusion-noise reference is worse on both CHAIR metrics and SPICE and the shuffled-image reference produces the highest hallucination rates despite maintaining a comparable SPICE score. We use the zero tensor because it retains the highest semantic quality and does not require a stochastic corruption process.

Table 6: Visual-null ablation on LLaVA-1.5 CHAIR-500. Results are mean $\pm$ standard deviation over three seeds.

Null reference	CHAIR _S ↓	CHAIR _I ↓	SPICE ↑
Zero tensor (EGVCA)	10.87 $\pm$ 1.43	5.77 $\pm$ 0.61	22.19 $\pm$ 0.24
Normalized black	10.53 $\pm$ 1.20	5.59 $\pm$ 0.45	22.15 $\pm$ 0.24
VCD diffusion noise	11.13 $\pm$ 2.07	5.98 $\pm$ 0.87	22.04 $\pm$ 0.24
Shuffled image	13.00 $\pm$ 0.85	6.63 $\pm$ 0.38	22.11 $\pm$ 0.27

A.3 Binary Parameter Sensitivity

Tables 7 and 8 present local 3×3 sensitivity grids over τ and λ . The same decision equation, $m'(q) = m_I(q) + \lambda[J(q) - \tau]$, is used throughout. On POPE, all displayed settings outperform the native F1 score of 84.11, indicating stable performance around the selected operating point. MME is more sensitive to calibration, but eight of the nine settings remain above the native Total score of 1681.92. In both benchmarks, the selected setting achieves the best result within the reported local neighborhood.

A.4 Caption Admission Statistics

We additionally inspect the decisions made by the caption admission gate on the LLaVA-1.5 CHAIR-500 runs. Statistics are pooled over seeds 42, 777, and 2024, yielding 1,500 generated captions. The first claim is admitted deterministically. Therefore, the admitted-evidence statistic in Table 9 is computed only over later claims. Admitted claims receive substantially stronger counterfactual evidence than rejected claims.

Table 7: Parameter sensitivity on POPE using LLaVA-1.5. Rows vary τ and columns vary λ . Values are Avg. F1.

$\tau \backslash \lambda$	1.0	2.0	4.0
0.5	86.20	86.42	85.22
1.0	86.14	86.48	86.18
1.5	85.89	85.65	86.07

Table 8: Parameter sensitivity on MME using LLaVA-1.5. Rows vary τ and columns vary λ . Values are MME Total.

$\tau \backslash \lambda$	0.010	0.025	0.050
-2.0	1685.71	1737.87	1706.80
-1.5	1684.83	1740.13	1724.06
-1.0	1678.08	1728.12	1737.34

Table 9: Claim-admission statistics on LLaVA-1.5 CHAIR-500, pooled over three seeds.

Statistic	Value
Number of captions	1,500
Candidate claims per caption	3.326
Admitted claims per caption	1.462
Captions stopped at a rejected claim	92.27%
Mean first-rejection position	2.320
Evidence of admitted post-first claims	0.484 ± 0.235
Evidence of rejected claims	0.088 ± 0.093

A.5 Inference Efficiency

We measure end-to-end inference efficiency on the same 200 CHAIR images using LLaVA-1.5, batch size one, FP16 inference, and max_new_tokens=64. We use three warm-up images followed by one synchronized timing pass for each method. Model-loading time is excluded. Because the methods generate outputs of different lengths, we report both latency per image and final-output token throughput. For EGVCA, end-to-end latency includes candidate generation, claim segmentation, original and visual-null claim scoring, and claim admission.

EGVCA requires $1.49\times$ the end-to-end latency of greedy decoding. Its throughput exceeds DoLa, OPERA, and HALC, while its peak memory usage matches that of greedy decoding. The lower final-output token count reflects the shorter captions produced by claim admission. Table 11 further separates the cost of candidate generation from the additional claim-scoring stage.

A.6 Complete General-Capability Results

Tables 12 and 13 provide the complete numerical results summarized by Figure 2. The dashes in Table 12 denote the omitted MiniGPT-4 results for DoLa and HALC, whose binary outputs were unstable under the matched protocol.

Table 10: Inference efficiency on LLaVA-1.5 over 200 CHAIR images after three warm-up images. Each method is measured in one synchronized pass.

Method	Sec./image↓	Tokens/sec.↑	Output tokens	Peak VRAM (GB)↓	Relative overhead
Greedy	0.953	66.61	63.48	15.27	1.00×
Beam	1.493	42.78	63.88	22.35	1.57×
VCD	1.905	33.56	63.91	16.37	2.00×
OPERA	11.760	5.43	63.84	25.45	12.34×
DoLa	3.244	19.69	63.85	26.31	3.40×
DeCo	<u>1.080</u>	<u>59.10</u>	63.85	<u>15.75</u>	<u>1.13×</u>
SID	1.905	33.49	63.80	16.36	2.00×
MoLE	2.058	31.00	63.81	29.26	2.16×
HALC	43.068	1.45	62.30	19.54	45.19×
EGVCA	1.424	21.90	31.18	15.27	1.49×

Table 11: End-to-end latency breakdown of EGVCA on LLaVA-1.5.

Component	Sec./image↓
Candidate generation	0.953060
Claim segmentation	0.000018
Original/null claim scoring	0.470771
Admission decision	0.000006
End-to-end	1.423854

Table 12: General visual understanding on the full MME benchmark. Perception and cognition scores are reported alongside their total.

Method	LLaVA-1.5			MiniGPT-4			mPLUG-Owl2		
	Perception	Cognition	Total	Perception	Cognition	Total	Perception	Cognition	Total
Native	1360.14	321.79	1681.92	693.94	196.79	890.73	<u>1398.51</u>	329.64	1728.15
VCD	1381.62	<u>350.71</u>	1732.33	599.77	170.71	770.48	1310.22	331.43	1641.65
OPERA	1456.40	299.29	<u>1755.68</u>	<u>771.00</u>	182.86	<u>953.86</u>	1344.54	<u>375.00</u>	1719.54
DoLa	1434.32	300.71	1735.04	-	-	-	1354.54	382.14	<u>1736.68</u>
DeCo	<u>1453.74</u>	322.50	1776.24	501.13	<u>198.57</u>	699.70	1343.98	325.71	1669.69
SID	<u>1433.76</u>	285.36	1719.11	552.97	<u>156.07</u>	709.04	1287.26	327.86	1615.11
MoLE	1364.30	300.36	1664.65	660.29	86.07	746.36	1324.62	369.29	1693.91
HALC	1440.07	295.36	1735.43	-	-	-	1358.87	362.14	1721.01
EGVCA	1373.34	366.79	1740.13	889.58	250.36	1139.94	1409.34	349.29	1758.62

Table 13: Reference-based caption quality across three LVLM backbones. Results are averaged over three seeds.

Method	LLaVA-1.5				MiniGPT-4				mPLUG-Owl2			
	B-4↑	MET.↑	R-L↑	SPICE↑	B-4↑	MET.↑	R-L↑	SPICE↑	B-4↑	MET.↑	R-L↑	SPICE↑
Greedy	7.89	23.01	27.01	<u>21.44</u>	7.94	22.52	26.37	19.07	8.05	23.26	27.25	21.72
Beam	<u>8.04</u>	23.11	27.55	<u>21.26</u>	8.30	23.09	26.89	<u>20.02</u>	8.48	<u>23.82</u>	28.00	21.90
VCD	6.36	21.92	25.88	19.53	7.17	22.02	26.03	18.61	6.65	22.31	26.37	20.21
OPERA	<u>8.04</u>	<u>23.19</u>	<u>27.63</u>	21.35	<u>9.14</u>	<u>23.54</u>	<u>29.40</u>	19.74	8.09	23.17	27.59	21.06
DoLa	4.34	18.35	21.29	16.48	5.08	17.93	21.62	14.99	<u>8.51</u>	<u>23.82</u>	<u>28.03</u>	<u>21.94</u>
DeCo	5.94	20.88	25.27	18.49	7.22	21.93	26.23	18.50	5.30	18.04	22.93	16.77
SID	6.33	21.81	25.78	19.54	7.29	22.21	26.21	18.53	6.79	22.43	26.40	20.35
MoLE	6.96	21.81	26.06	19.61	7.46	22.07	26.17	19.54	8.39	23.41	27.87	21.52
HALC	4.47	20.06	23.64	17.92	6.49	21.05	25.41	17.86	8.18	23.39	27.61	21.60
EGVCA	14.53	26.93	41.96	22.19	12.45	25.41	38.18	20.45	12.11	26.33	38.53	22.55