Revealing Epistemic Uncertainty in MLLMs via Causal-Invariant Masking

Haoyang Luo[1] **Linwei Tao**[2] **Jie Gui**[3] **Xinghao Chen**[4]
Chang Xu[2] **Jianyuan Guo**[1] **Minjing Dong**[1]*

[1] City University of Hong Kong [2] University of Sydney
[3] Southeast University [4] Huawei Noah's Ark Lab

Abstract

Multimodal Large Language Models (MLLMs) suffer from hallucinations, creating a critical need for Uncertainty Quantification (UQ) to ensure reliable deployment. However, existing approaches struggle to detect uncertainty caused by superficial associations, especially when the query-relevant signal is weak. We mainly attribute this issue to their bias toward aleatoric uncertainty arising from data ambiguity, overlooking epistemic uncertainty stemming from model limitations. To better capture epistemic uncertainty for a comprehensive UQ, we propose Causal-Invariant Masking (**CIM**), which measures the semantic shift between the original predictions and those conditioned on a causally-focused view. Based on this framework, we introduce Semantic Divergence as our core metric for UQ and provide theoretical evidence that, under evidence-preservation and small-shift assumptions, it approximates the variance of the model's sensitivity to non-causal correlations. To accelerate UQ in MLLMs, we further propose Expected Embedding Drift (EED), a fast geometric proxy metric that estimates semantic shift directly within the hyperspherical embedding space. Experiments show that combining our signal with Semantic Entropy achieves state-of-the-art performance on various benchmarks, while the proposed EED accelerates by nearly 50% with comparable performance. Our code is available at `https://github.com/E-Galois/Causal-Invariant-Masking`.

1 Introduction

Large Language Models (LLMs) [51, 32] and Multimodal Large Language Models (MLLMs) [47, 4] have significantly altered the landscape of artificial intelligence. By connecting visual perception with textual semantics, MLLMs are able to demonstrate consistent success in tasks ranging from visual question answering to complex reasoning [28, 37, 50]. However, their reliable deployment in real-world applications requires solving the severe problem of hallucination [5, 17], where MLLMs confidently generate factually incorrect but grammatically plausible content. Especially for VLMs, non-existent objects, attributes, and events may unreasonably occur in the responses when misleading factors, either in vision or language, dominate the query. It suggests that MLLMs do not internally realize their visual capability boundary and highlights a critical need for robust safety measures targeting generative trustworthiness.

With this aim, numerous Uncertainty Quantification (UQ) methods have been developed. By estimating the confidence level of a model, systems can abstain from answering or flag potential hallucinations when encountering high uncertainty. Therefore, precise UQ becomes a proxy for hallucination detection. Traditional UQ approaches typically measure the entropy of the predictive

*Correspondence to Minjing Dong (minjdong@cityu.edu.hk).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

distribution, often aggregated at the token level [46, 42]. Recent advancements [21, 20, 34] have shifted towards measuring *semantic uncertainty* using sampling-based methods, such as Semantic Entropy (SE) [13], which aggregates probabilities over clusters of similar generations to calculate the response variety.

However, these approaches exhibit limited discrimination under hallucinative visual content, especially when the query-irrelevant elements generate spurious hints. For instance, background objects with prominent colors can dominate the response to queries on the foreground object, making consistent mistakes in the sampled output. Thus, a more comprehensive UQ is necessary for MLLMs. In fact, both *epistemic* and *aleatoric* uncertainties play important roles in UQ, where epistemic uncertainty reveals insufficiency in the model's perception or comprehension and aleatoric uncertainty reflects the poor data quality or incomplete information [19]. Quantifying both of them is crucial in UQ [41, 1], however, existing approaches fail to do so because they typically assume that all forms of uncertainty can be equally estimated. In this paper, we show that they are biased towards quantifying mostly aleatoric uncertainty in MLLMs, leaving epistemic uncertainty underestimated.

We introduce Causal-Invariant Masking (**CIM**), a framework that disentangles intrinsic epistemic instability from data ambiguity by probing semantic consistency under causally invariant transforms. Our core insight is that robust reasoning must remain invariant when task-irrelevant features are removed. By constructing a Region-of-Interest (RoI) view that aims to preserve query-relevant causal features, we measure the shift in the model's belief distribution. We quantify this using Semantic Divergence, defined as the shift from the semantic distributions of the original to the focused inputs. Supported by theoretical derivation, this metric targets the model stability against irrelevant distractions to capture context-related epistemic uncertainty, thereby alleviating the biased quantification in previous works. To address the computational costs of discrete semantic clustering, we then propose a fast proxy called Expected Embedding Drift (EED). Leveraging the hyperspherical geometry of the embedding space, EED estimates semantic shift directly from token embeddings, eliminating costly discrete operations. Experiments demonstrate that combining our signal with SE outperforms baselines, achieving state-of-the-art hallucination detection of **76.8%** AUROC on the challenging AdVQA and superior results on POPE. Furthermore, EED reduces computational overhead by **50%** with competitive detection performance, facilitating applications such as hallucination mitigation and adaptive-resolution inference.

2 Related Work

2.1 Uncertainty Estimation in LLMs

UQ in LLMs is pivotal for reliability and hallucination detection, generally categorized into intrinsic, verbalized, and consistency-based methods. Intrinsic metrics like perplexity [41] measure token-level confidence, while verbalized methods [45, 43] prompt models to explicitly express uncertainty, often yielding better calibration than raw probabilities. Consistency-based works are currently dominant approaches. Semantic Entropy [13] and Semantic Dispersion [27] detect hallucinations by aggregating multiple sampled responses based on meaning rather than exact phrasing. Recent works further exploit internal representations as uncertainty indicators by utilizing activation covariance and propose EigenScore [10] or geometric convex hull analysis [9] to quantify dispersion in high-dimensional embedding spaces. However, these methods target general UQ for LLMs and can be challenged in vision-language tasks.

2.2 Uncertainty Estimation in MLLMs

Extending UQ to MLLMs involves challenges from combined visual and textual modalities, often focusing on hallucination detection in vision-language tasks. Similar to those for LLMs, consistency-based sampling techniques are explored, *e.g.*, [23] develops UMPIRE, a training-free UQ framework that estimates consistency in multiple sampled responses by a coherence-adjusted semantic volume. Perturbation-based methods introduce variations in inputs to probe uncertainty. VL-Uncertainty [48] and Uncertainty-o [49] measure hallucination in LVLMs by computing entropy over clustered responses, but using semantic-equivalently perturbed visual and textual queries, supporting tasks of hallucination mitigation and uncertainty-aware reasoning. However, these approaches typically rely on *global* perturbations under the full context, without explicitly isolating the model's reliance on critical and irrelevant features. CIM is also a perturbation-style test in its implementation, but its

perturbation is query-conditioned and RoI-preserving. The goal is to compare the original belief with a view that keeps the answer evidence while suppressing irrelevant context, rather than purely testing variance under general information loss.

3 Method

3.1 Preliminary

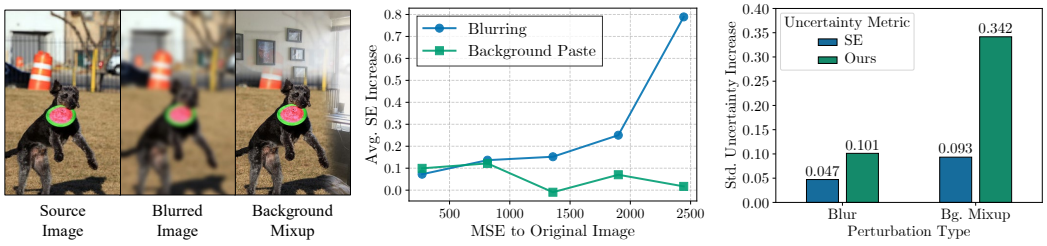

Figure 1: **Left:** A case of blurred and background-mixup image pair. **Middle:** Sensitivity of Semantic Entropy to visual perturbations in MLLMs. Similar levels of error to the original image are binned into the same point. **Right:** Comparison of uncertainty increase under blurring and background mixup perturbations. While Semantic Entropy increases slightly across both, it fails to distinguish hallucination. In contrast, our proposed Semantic Divergence remains robust to visual blur but exhibits a distinct spike under hallucinative perturbations. The values are normalized by the standard deviation of each metric.

We consider an MLLM parameterized by θ over visual space $\mathcal{V}$ and vocabulary $\mathcal{W}$. Given an image v and query q, it generates a token sequence $y = (y_1, \ldots, y_T)$ autoregressively via $P_\theta(y|v,q) = \prod_{t=1}^{T} P_\theta(y_t|y_{<t}, v, q)$. For simplicity, we henceforth omit q and θ, denoting the distribution as $P(y|v)$. Token-level entropy, *i.e.*, $\mathcal{H}(y) = -\sum P(y)\log P(y)$, or perplexity values, *i.e.*, $\text{PPL}(y) = \exp\left(-\frac{1}{T}\sum_{t=1}^{T}\log P(y_t|y_{<t}, v)\right)$, often misestimate uncertainty because semantically equivalent outputs are treated as distinct. Semantic Entropy (SE) [13] samples K sequences, maps them to a semantic space $\mathcal{S}$ with a clustering operation $\mathcal{G} : \mathcal{W}^* \to \mathcal{S}$, and aggregates meaning-level probabilities: $\mathcal{H}_{\text{sem}}(v) = -\sum_{s\in\mathcal{S}} P(s|v)\log P(s|v)$, where $\mathcal{W}^*$ is the set of $\mathcal{W}$-strings and $P(s|v) = \sum_{y:\mathcal{G}(y)=s} P(y|v)$ is the aggregated probability. While SE handles linguistic variance, it still fails to distinguish epistemic from aleatoric uncertainty.

3.2 Epistemic Uncertainty Isolation via Causal Masking

Reliable MLLM deployment requires separating aleatoric uncertainty, which reflects data ambiguity, from epistemic uncertainty, which reflects model limitation. Existing metrics often blend the two and therefore skew toward aleatoric noise. As illustrated in Fig. 1 middle, Semantic Entropy rises quickly under blurring for aleatoric uncertainty, but is much less sensitive to background mixup for epistemic uncertainty, suggesting that it treats causal and non-causal features alike. Robust reasoning instead should satisfy causal invariance: predictions should stay stable when task-irrelevant context changes, whereas uncertain responses should not.

Methodology. We propose **CIM** to isolate the epistemic component through **causal invariance**. Given a query q, we decompose the visual input v into causal features v_c and non-causal features v_s, where v_c is query-relevant, while v_s may carry spurious hints of correlations. This split is defined with respect to an oracle predictor satisfying $P(s|v_c, v_s) = P(s|v_c)$, rather than assuming that the evaluated MLLM already knows it. Accordingly, reliable reasoning should remain stable when non-causal variables are removed, and any large shift after filtering v_s exposes reliance on spurious correlations. We therefore introduce a v_c-preserving masking operator $T(\cdot) : \mathcal{V} \to \mathcal{V}$, mainly implemented as query-conditioned Region-of-Interest (RoI) cropping. In practice, we detect key objects with GroundingDINO [29], filter overly large boxes, and crop the union bounding box so

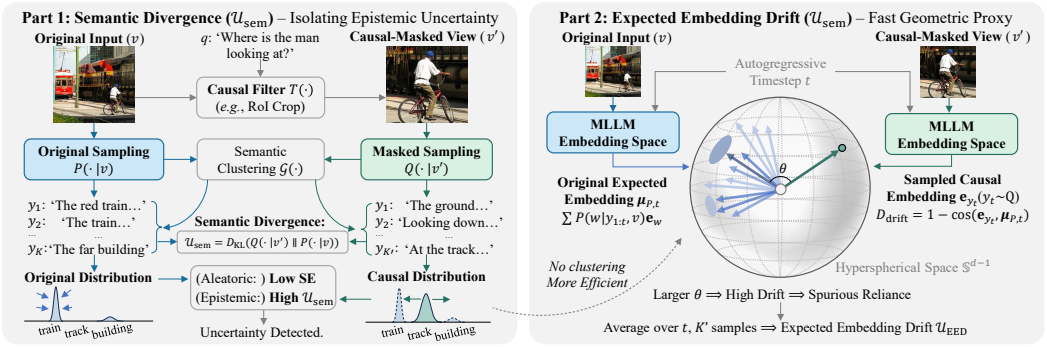

Figure 2: **The detailed framework of CIM.** Left: We isolate epistemic uncertainty via Semantic Divergence, which measures the KL divergence between the semantic distribution of the original input and a causally-filtered view v', revealing the model's reliance on query-irrelevant information. Right: Expected Embedding Drift geometrically estimates semantic shift via cosine distance in the hyperspherical embedding space, eliminating the computational overhead of discrete clustering.

inter-object relations and local context are preserved. This acts as an empirical proxy for the oracle v_c filter: it preserves query-relevant evidence while removing as much v_s as possible, yielding a focused view $v' = T(v)$. Because only a v_c-preserving transform is required, the detector choice is not essential, as validated in Appendix C.6. Here, "causal" refers to the invariance of an answer when query-relevant evidence is preserved and distracting context is removed. RoI cropping approximates this property without implementing a formal causal intervention.

To quantify the shift caused by causal masking, we define **Semantic Divergence** (**SD**) as the Kullback-Leibler divergence between the semantic probability distributions of the focused input $P(s|v', q)$ and the original model $P(s|v, q)$. In practice, we sample K' sequences from the model, group them into clusters $\{s_i\}_{j=1}^{M}$ via entailment $\mathcal{G}(\cdot)$, and aggregate their likelihoods similarly to SE. After grouping, the Semantic Divergence is thus calculated as:

$$\begin{aligned}\mathcal{U}_{\text{sem}}(v, q) &= D_{KL}(P(s|v')\|P(s|v)) \\ &= \sum_{s\in\mathcal{S}} P(s|v') \log \frac{P(s|v')}{P(s|v)}.\end{aligned} \tag{1}$$

This KL divergence measures the *information gain* from removing non-causal features. If the model is robust, or the data is ambiguous while the model is honest, the distribution over meanings remains stable, yielding low $\mathcal{U}_{\text{sem}}$. In contrast, hallucinations that rely on background context cause a semantic shift and high $\mathcal{U}_{\text{sem}}$. Unlike global perturbations that flatten both distributions, this measure targets context-related epistemic failures. As illustrated in Fig. 1 right, $\mathcal{U}_{\text{sem}}$ responds more strongly than SE to misleading background changes. This provides an epistemic-oriented signal complementary to SE.

3.3 Theoretical Analysis

We justify why $\mathcal{U}_{\text{sem}}$ isolates epistemic uncertainty by framing the MLLM's semantic output as an Energy-Based Model (EBM) [8, 24, 12]. This is compatible with autoregressive decoding: prior work interprets pre-softmax logits as token energies [33], connects left-to-right language models and EBMs through a function-space equivalence [6], and uses energy functions for constrained text decoding [39]. Because MLLM decoding remains autoregressive after conditioning on images, the same energy view applies. Under standard causal assumptions, the divergence in Eq. (1) converges to the variance of sensitivity to spurious features.

Standard causal theory [36, 35] models the input v as causal variables v_c and non-causal variables v_s. An oracle predictor would satisfy $P(s|v_c, v_s) = P(s|v_c)$, but real models often fail to be invariant. We therefore introduce a unified assumption on the masking operation and the energy landscape:

Assumption 3.1 (Energy Decomposition)**.** The MLLM's semantic output over $\mathcal{S}$ is modeled as a Boltzmann distribution $P_\theta(s|v) = \frac{1}{Z(v)} e^{-E_\theta(s,v)}$, where the total energy decomposes additively

into a causal term $E_c(s, v_c)$ and a non-causal term $E_s(s, v_s)$. We assume that the causal energy E_c remains invariant under the causal-preserving $T(\cdot)$, *i.e.*, $E_c(s, v_c') = E_c(s, v_c)$, while the non-causal energy shifts by a quantity $\Delta_s(s) \triangleq E_s(s, v_s) - E_s(s, v_s')$.

This invariant/variant split follows the broader causal-invariance view that causal mechanisms should remain stable across interventions or environments [38, 2, 40]. Related decomposition work also separates shared invariant factors from domain-specific variants [7]. Rather than requiring pixel-level separation, Assumption 3.1 only asks that $T(\cdot)$ preserve query-relevant evidence while suppressing non-causal context. The empirical validations in Appendix C.8, Table 3, and Appendix C.4 show that causal-energy shifts are far below the model's logit margins, while detected RoIs rarely fail and closely match ground truth RoIs.

We first show that the shared causal-energy term or the aleatoric component ϕ_c cancels in the likelihood ratio. This interpretation is conditional on Assumption 3.1: the empirical validations support its approximation in the evaluated settings.

Lemma 3.2 (Aleatoric Cancellation)**.** *Given Assumption 3.1, the log-likelihood ratio between the perturbed and original distributions depends solely on the non-causal energy shift $\Delta_s(s)$ and the partition function ratio, i.e.,*

$$\log \frac{Q(s)}{P(s)} = \Delta_s(s) + \log \frac{Z(v)}{Z(v')}. \tag{2}$$

The proof is provided in Appendix A.1. Lemma 3.2 shows that the shared $E_c(s, v_c)$ term cancels in the likelihood ratio under Assumption 3.1, providing the basis for the epistemic interpretation. Next, we establish the exact form of the divergence.

Theorem 3.3 (Epistemic Jensen Gap)**.** *The Semantic Divergence $\mathcal{U}_{sem} = D_{KL}(Q\|P)$ is equivalent to the Jensen Gap of the exponential function acting on $\Delta_s(s)$ with respect to distribution Q.*

Furthermore, if we assume the shift magnitude is bounded such that $\sup|\Delta_s(s)| \leq \epsilon$ for a small $\epsilon > 0$. The Semantic Divergence is approximated by half the variance of the energy shift $\Delta_s(s)$ with a third-order error bound:

$$\mathcal{U}_{sem} = \frac{1}{2}\mathrm{Var}_{s\sim Q}[\Delta_s(s)] + \mathcal{O}(\epsilon^3). \tag{3}$$

We provide the detailed proof in Appendix A.2. This result shows that $\mathcal{U}_{\text{sem}}$ quantifies epistemic instability caused by irrelevant features under the stated approximation. When the v_c-focused operation $T(\cdot)$ removes most of v_s, we have $\Delta_s(s) = E_s(s, v_s)$, so $\mathcal{U}_{\text{sem}}$ measures variance on query-irrelevant signals. Appendix C.8 reports empirical Δ_s magnitudes relative to the model's decision margins.

3.4 Fast Quantification via Expected Embedding Drift

Despite its theoretical validity, Semantic Divergence $\mathcal{U}_{\text{sem}}$ is expensive to compute. Equation (1) requires estimating $P(s|v) = \sum_{y:\mathcal{G}(y)=s} P(y|v)$, which in turn needs samples from P and Q, followed by expensive LLM clustering $\mathcal{G}(\cdot)$.

However, we observe that semantic meaning is already embedded in the model's high-dimensional geometry, where tokens with similar meanings lie close together. We exploit this and propose **Expected Embedding Drift** (EED), a clustering-free proxy that bypasses discrete semantic grouping.

Specifically, it measures the geometric divergence between the masked-input distribution Q and the original semantic distribution P by mapping token probabilities to single vectors on a hypersphere.

Expected Semantic Vector. For a response $y \sim Q$ at time step t conditioned on $y_{<t}$ and v, rather than treating the output solely as a discrete categorical distribution, we view the token distribution as a distribution over *semantic vectors* and compute the **expected embedding vector**, $\boldsymbol{\mu}_{P,t}$ as

$$\boldsymbol{\mu}_{P,t} = \mathbb{E}_{w\sim P(\cdot|y_{<t},v)}[\mathbf{e}_w] = \sum_{w\in\mathcal{W}} P(w|y_{<t}) \cdot \mathbf{e}_w, \tag{4}$$

where $\mathbf{e}_w \in \mathbb{R}^d$ is the L_2-normalized embedding vector for token w in $\mathcal{W}$. This expected embedding is the probability-weighted semantic center in hyperspherical space, filtering low-probability noise and pointing to the aggregate intent.

Drift Calculation. To measure epistemic inconsistency, for each generated response sequence $y = (y_1, \ldots, y_T) \sim Q$, we measure the cosine drift of y_t from the source semantic expectation $\boldsymbol{\mu}_{P,t}$ at each autoregressive timestep t:

$$D_{\text{drift}}(y_t) = 1 - \cos(\mathbf{e}_{y_t}, \boldsymbol{\mu}_{P,t}) = 1 - \frac{\mathbf{e}_{y_t} \cdot \boldsymbol{\mu}_{P,t}}{\|\mathbf{e}_{y_t}\| \|\boldsymbol{\mu}_{P,t}\|}, \tag{5}$$

where $\mathbf{e}_{y_t}$ is the embedding vector of token y_t. Lower drift means the perturbed token still aligns with the original average belief, while higher drift indicates a stronger rotation after irrelevant features are removed. We average this drift over K' stochastic sampling passes and their tokens to obtain the sequence-level EED:

$$\mathcal{U}_{\text{EED}}(v, q) = \frac{1}{K'} \sum_{k=1}^{K'} \sum_{t=1}^{T_k} D_{\text{drift}}(y_t^{(K+k)}), \tag{6}$$

This aggregates the step-wise deviations into a scalar score, where $\mathbf{e}_{y_t^{(k)}}$ is the embedding of the token generated by the perturbed model in the k-th sample. Notably, EED scores require far fewer samples than SD to stabilize, thanks to its non-discrete semantic aggregation. It also avoids a semantic clustering model and $\mathcal{O}((K + K')^2)$ entailment judgments, reducing the overhead to just $\mathcal{O}(K')$. We present the framework of the proposed uncertainty metrics in Fig. 2

Theoretical Justification To justify EED, we show that under a geometric distribution assumption, minimizing $\mathcal{U}_{\text{EED}}$ is equivalent to minimizing the KL divergence between the semantic distributions.

Assumption 3.4 (Spherical Semantic Distribution)**.** We model the continuous semantic distribution of the MLLM output as a von Mises-Fisher (vMF) distribution on the unit hypersphere $\mathbb{S}^{d-1}$ [30, 22]. The probability density of a semantic vector $\mathbf{z}$ given mean direction $\boldsymbol{\mu} \in \mathbb{S}^{d-1}$ and concentration $\kappa > 0$ is:

$$p(\mathbf{z}|\boldsymbol{\mu}, \kappa) = C_d(\kappa) \exp(\kappa \boldsymbol{\mu}^\top \mathbf{z}), \tag{7}$$

where $C_d(\kappa) = \kappa^{d/2-1} / ((2\pi)^{d/2} I_{d/2-1}(\kappa))$ is the normalization constant.

This connects the discrete token space to the continuous geometry, where $\boldsymbol{\mu}$ corresponds to the normalized expected embedding $\boldsymbol{\mu}_{P,t}$.

Lemma 3.5 (KL Divergence of vMF Distributions)**.** *Let $Q \sim vMF(\boldsymbol{\mu}_Q, \kappa)$ and $P \sim vMF(\boldsymbol{\mu}_P, \kappa)$ be two von Mises-Fisher distributions with identical concentration κ. The Kullback-Leibler divergence between them is strictly proportional to the cosine distance between their mean directions:*

$$D_{KL}(Q\|P) = \kappa A_d(\kappa)(1 - \boldsymbol{\mu}_Q^\top \boldsymbol{\mu}_P), \tag{8}$$

where $A_d(\kappa) = \|\mathbb{E}_{\mathbf{z}\sim Q}[\mathbf{z}]\| = \frac{I_{d/2}(\kappa)}{I_{d/2-1}(\kappa)}$ is the ratio of Bessel functions.

The derivation is detailed in Appendix A.3.

Theorem 3.6 (Geometric Proxy Equivalence)**.** *Under the vMF model with identical concentration in Lemma 3.5, the Expected Embedding Drift $\mathcal{U}_{EED}$ is an unbiased estimator of the KL divergence between the perturbed semantic distribution Q and the original belief P, up to a scaling factor and additive constant.*

The full proof is available in Appendix A.4, which connects the proxy to our core objective. The point-wise drift in Eq. (5) is aggregated in Eq. (6) to estimate the expected cosine distance. As established in Eq. (8) and Thm. 3.6, this distance effectively estimates the Semantic Divergence in Eq. (1), giving an efficient proxy for epistemic uncertainty. This relationship reflects the expected behavior under the stated geometric assumptions, providing theoretical insights into the practical application of EED.

4 Experiment

We adopt InternVL2-8B [11] and Qwen2.5-VL-7B-Instruct [3] as our evaluated models. We test the performance of $\mathcal{U}_{\text{sem}}$ and $\mathcal{U}_{\text{EED}}$ on various benchmarks, including common visual question answering (VQAv2 [15], OKVQA [31], AdVQA [25]) and hallucination-targeted benchmarks (POPE [26],

Table 1: **Uncertainty estimation performance** (%) on visual question answering benchmarks. AUC is short for AUROC. SE+SD/EED are blended scores; pure rows omit SE. – denotes unreported AURAC.

Benchmark	VQAv2				OKVQA				AdVQA			
MLLM	InternVL2-8B		Qwen2.5-VL-7B		InternVL2-8B		Qwen2.5-VL-7B		InternVL2-8B		Qwen2.5-VL-7B	
Metric	AUC	AURAC	AUC	AURAC	AUC	AURAC	AUC	AURAC	AUC	AURAC	AUC	AURAC
Perplexity	70.3	80.7	77.6	81.4	65.4	76.8	75.1	82.3	65.7	60.3	66.4	61.2
Self Evaluation [45]	60.9	75.1	62.4	71.9	66.4	77.0	63.1	75.5	60.4	57.1	55.3	49.7
Verbalized Confidence [43]	54.3	68.3	57.2	69.2	55.3	67.8	58.2	72.1	50.4	48.6	57.1	51.9
Semantic Entropy [13]	78.7	82.5	76.5	79.5	80.6	84.5	78.6	83.8	75.5	66.6	73.6	63.1
EigenScore [10]	72.6	80.7	75.8	79.9	72.2	78.5	76.8	83.0	67.1	60.2	64.6	59.1
UMPIRE [23]	74.3	81.7	72.9	78.4	72.6	79.0	76.4	83.0	68.2	61.9	65.5	59.6
Uncertainty-o [49]	72.1	80.1	64.9	74.2	67.7	77.8	61.8	74.2	60.1	56.1	57.4	53.7
Pure SD (ours)	76.8	83.3	73.7	79.7	77.6	84.0	74.4	82.6	70.6	64.3	68.3	62.2
Pure EED (ours)	69.0	79.7	64.2	73.4	68.1	77.0	65.7	77.1	62.7	59.2	60.1	54.7
SE+SD (ours)	**81.3**	**85.2**	**78.4**	**81.5**	**81.8**	**85.7**	**79.6**	**84.2**	**76.8**	**66.9**	74.5	**64.8**
SE+EED (ours)	80.3	84.5	77.9	81.1	80.9	84.5	79.3	84.0	75.8	66.5	**74.6**	64.6

Table 2: **Uncertainty estimation performance** (%) on all 3 sub-tasks of the POPE benchmark. SE+SD/EED denote the blended scores.

Benchmark	POPE random				POPE popular				POPE adversarial			
MLLM	InternVL		QwenVL		InternVL		QwenVL		InternVL		QwenVL	
Metric	AUC	AURAC	AUC	AURAC	AUC	AURAC	AUC	AURAC	AUC	AURAC	AUC	AURAC
Perplexity	68.7	91.2	68.5	91.5	68.1	90.4	67.3	90.2	68.9	90.8	65.2	89.2
Self Evaluation [45]	52.0	84.6	61.3	91.4	53.6	84.7	51.5	88.1	53.5	85.2	57.4	89.5
Verbalized Confidence [43]	50.9	84.2	60.4	88.2	51.8	82.9	60.5	87.2	50.6	82.8	60.6	87.5
Semantic Entropy [13]	70.6	89.6	62.9	88.7	68.7	88.5	64.8	88.0	72.3	90.2	63.0	86.3
EigenScore [10]	65.5	89.4	56.0	87.2	64.4	87.8	56.6	86.5	63.8	88.5	54.6	85.4
UMPIRE [23]	68.7	91.2	80.5	94.4	68.7	90.5	79.1	93.4	68.2	90.5	77.0	92.7
Uncertainty-o [49]	64.6	88.8	51.5	85.4	66.4	88.6	55.1	85.9	62.8	88.2	56.6	86.1
SE+SD (ours)	**75.9**	**92.4**	66.2	89.3	**74.8**	**91.3**	66.1	88.7	**74.9**	**91.7**	64.5	88.1
SE+EED (ours)	73.1	91.6	**83.2**	**95.1**	71.0	90.5	**81.5**	**94.3**	73.0	91.3	**78.6**	**93.6**

MME [14]). For simplicity, we adopt the lightweight GroundingDINO detector [29] as our RoI generator. In practice, specialized or query-conditioned detectors can also instantiate $T(\cdot)$, as verified in Appendix C.6. We fix the evaluation subset and total generation budget to $K_{\text{total}} = 32$: sampling-based baselines use 32 original-view samples, while CIM uses $K = 16$ original-view and $K' = 16$ cropped-view samples. The original-view samples also yield SE without extra decoding. For total uncertainty quantification, we use $\mathcal{U}_{\text{final}} = (1 - \lambda)\mathcal{U}_{\text{SE},16} + \lambda\mathcal{U}_{\text{SD/EED}}$ with $\lambda = 0.05$. We denote these blends by SE+SD and SE+EED, and report the pure signals separately to distinguish their epistemic-oriented behavior from total-UQ performance. Appendix B provides the detailed experimental setup.

4.1 Results on Hallucination Detection

General VQA Benchmarks. Table 1 presents the results on VQAv2, OKVQA, and AdVQA. The blended CIM scores consistently improve AUROC over SE across all evaluated VQA settings. Pure SD/EED are weaker than SE on general detection, supporting their role as complementary signals rather than replacements for SE. Notably, on the AdVQA benchmark, which is specifically designed to elicit model errors, our method achieves an AUROC of **76.8%** on InternVL, surpassing Semantic Entropy (75.5%) and EigenScore (67.1%). This suggests that our method's specific reliance on *causal invariance* by checking if the answer holds after removing background context is more robust to adversarial visual perturbations than methods that solely attend to the output distribution. Furthermore, SE+EED performs competitively with SE+SD, often within one AUROC point, supporting the practical use of the efficient proxy.

We further assess the gains using 10,000 paired bootstrap resamples on evaluation sets (Appendix C.12). Eleven of the twelve VQA gain intervals are above zero at the reported 90% confidence level. All POPE/MME gain intervals are above zero. Appendix C.2 provides qualitative examples of low-SE hallucinations.

Hallucination-targeted Benchmarks. Table 2 details performance on the POPE benchmark. This benchmark is crucial as it directly evaluates on the topic of object hallucinations. On the **InternVL**

Table 3: **Ablation study of the x_c-focused image transform.** AUROC and AURAC use the final mixed score, while AUROC^{EPI} uses the SE-free pure SD score. GT RoI is available only on GQA; – denotes not applicable. IVL / QVL denotes InternVL / QwenVL, respectively.

METRIC	AUROC						AUROC^{EPI}						AURAC					
BENCHMARK	VQAv2		AdVQA		GQA		VQAv2		AdVQA		GQA		VQAv2		AdVQA		GQA	
MLLM	IVL	QVL	IVL	QVL	IVL	QVL	IVL	QVL	IVL	QVL	IVL	QVL	IVL	QVL	IVL	QVL	IVL	QVL
RANDOM CROP	80.2	77.9	75.7	71.9	77.9	68.8	70.3	62.1	61.4	54.3	61.0	57.5	83.6	81.6	66.7	61.5	81.5	61.3
REVERSED CROP	77.6	76.1	74.8	71.9	76.8	69.5	53.7	56.5	54.2	50.7	58.7	54.5	81.5	79.2	65.0	61.3	80.1	61.5
GT ROI CROP	–	–	–	–	**80.3**	**71.4**	–	–	–	–	72.1	**66.3**	–	–	–	–	82.4	**63.9**
OURS (ROI CROP)	**81.3**	**78.5**	**76.8**	**74.5**	79.9	70.2	**76.8**	**73.7**	**70.6**	**68.3**	**73.3**	64.6	**85.2**	**81.9**	**66.9**	**64.8**	**82.5**	63.1

model, our explicit method dominates the performance across all splits, achieving ∼**75% AUROC**, significantly outperforming Semantic Entropy (∼68-72%). Interestingly, on **QwenVL**, SE+EED outperforms SE+SD (e.g., 83.2% vs 66.2% on Random). QwenVL's average token entropy on POPE is only 0.06–0.08, compared with 0.51–0.54 for InternVL (Appendix C.9). In this sharp, mostly binary regime, both views can remain in the same hard answer cluster despite changes in the underlying token distributions. EED retains continuous token-level changes that finite-sample SD can miss, which helps explain their different rankings without requiring identical finite-sample behavior. A significant gap is observed between our CIM and other baselines on the challenging **Adversarial** split, confirming CIM's identification of whether the context priors are relied on rather than visual evidence. We further evaluate on the comprehensive MME benchmark in Appendix C.1, where Fig. 7 shows that CIM remains the best-performing method. This demonstrates CIM's generalization beyond simple object existence to complex reasoning in MME, and again verifies its consistency in uncertainty estimation.

4.2 Discussions

Causal Masking. Our theoretical framework proposes that the visual input operator $T(\cdot)$ must focus on isolating causal features x_c to separate epistemic from aleatoric uncertainty. To validate such a design choice of x_c-focused RoI Cropping, we compare it against: 1) **Random Crop**, which crops an arbitrary region, 2) **Reversed Crop**, which masks the detected RoIs and keeps only the background, and 3) **Ground-Truth RoIs**, which are from the box-annotated GQA benchmark. The ablation results are presented in Table 3. We also compare with the commonly used gaussian noise perturbation in Appendix C.3. From the results, **Random Crop** yield inferior performance compared to our targeted approach. This aligns with our hypothesis that random perturbations degrade the aleatoric and epistemic signals indiscriminately. Crucially, **Reversed Crop** performs significantly worse than Random Crop in most tested cases. This is because feeding only the background forces the model to hallucinate or guess, but does not provide an anchor point for the correct reasoning. In contrast, our method of RoI Crop achieves significantly larger AUROC across all settings (e.g., **81.3%** on VQAv2 for InternVL). By preserving necessary evidence while maximally removing spurious context, our measurement effectively grounds the MLLM's confidence in *causal* signals, supporting the intended epistemic interpretation.

While the SE component is included to retain aleatoric sensitivity for hallucination detection, it can dilute the random-vs-RoI gap on epistemic UQ. To evaluate the proposed signal separately, AUROC^{EPI} reports hallucination-detection AUROC using pure SD without SE. Under this setting, the separation becomes much larger: Our RoI Crop improves over Random Crop by 6.5/9.2 AUROC on InternVL VQAv2/AdVQA or by 11.6/14.0 AUROC on QwenVL VQAv2/AdVQA, and is even more superior to the Reversed Crop. We additionally include GQA [18] because it provides ground-truth object annotations, enabling a controlled comparison with GT RoIs. On GQA, our operation is comparable to **GT RoIs Crop** and both clearly outperform the others, validating that CIM relies on preserving query-relevant causal evidence rather than arbitrary perturbation, and exact crops are less critical than retaining causal evidence.

We also show the flexibility of our pipeline using question-only RoIs generated with Florence-2, keeping evaluation examples and sampling settings fixed (Appendix C.7). Pure and blended SD show mixed changes across settings, indicating that CIM can operate without conditioning the crop on the original answer.

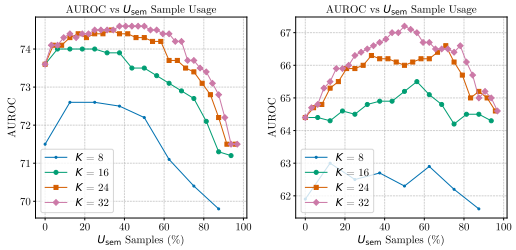

Figure 3: **Hallucination detection performance with fixed numbers of total samples**, *w.r.t.* the proportion of samples for $\mathcal{U}_{\text{sem}}$ calculation besides SE calculation, showing QwenVL results using AdVQA (left) and MME (right).

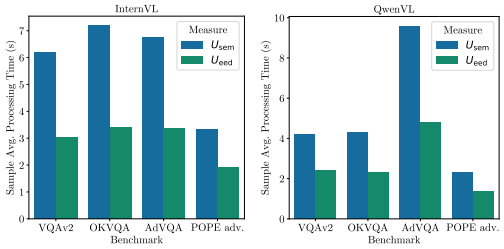

Figure 4: **Sample-wise efficiency comparison for $\mathcal{U}_{\text{sem}}$ and $\mathcal{U}_{\text{EED}}$.** It takes only half time for $\mathcal{U}_{\text{EED}}$ to compute the uncertainty under saturated performance.

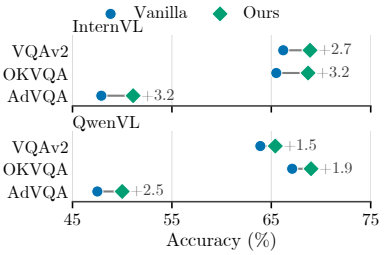

Figure 5: **Hallucination mitigation performance** (Acc%) using CIM to threshold and re-answer the query.

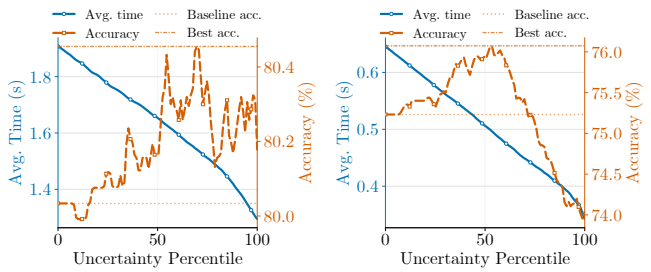

Figure 6: **Adaptive-resolution VQA on MME.** Confident samples use low resolutions. Left: InternVL; Right: QwenVL.

Number of Samples. In Fig. 3, we show the optimal sampling allocation under a fixed total number $K + K'$ between estimating the original $P(s|v)$, *i.e.*, for SE and $\mathcal{U}_{\text{sem}}$, and measuring the masked $P(s|v')$ only for $\mathcal{U}_{\text{sem}}$. We observe a consistent middle-peaking trend across benchmarks. The maximal performance is achieved at a balanced sampling budget of 40%-60% each. Biased allocations degrade performance, as insufficient samples for either distribution lead to mis-estimation of one type of uncertainty. Furthermore, while increasing the total budget universally improves AUROC, the optimal allocation proportion remains stable, suggesting its robustness in such scaling.

Efficiency of EED. Fig. 4 compares the average processing time per sample between the Semantic Divergence $\mathcal{U}_{\text{sem}}$ and our proposed proxy $\mathcal{U}_{\text{EED}}$. Across both MLLM architectures and all four benchmarks, $\mathcal{U}_{\text{EED}}$ consistently reduces the computational cost by approximately 50%. For instance, on the computationally demanding AdVQA benchmark with QwenVL, the processing time drops from $\sim$ 9.6s to $\sim$ 4.8s per sample. Compared with sampling-and-clustering methods under the same generation budget, EED removes the NLI-based $\mathcal{O}((K + K')^2)$-time clustering for the perturbed input and operates directly in the embedding space. Appendix C.10 further shows that SD is not strongly tied to one entailment model, but EED avoids that clustering dependency altogether.

Applications. Beyond detection, a practical uncertainty metric should enable the model to correct its own errors. We study whether **CIM** can act as a filter to improve factual quality. In this setting, we perform a reject-and-refine strategy: if the estimated uncertainty $\mathcal{U}_{\text{sem}}$ exceeds a threshold, we instruct the model to re-generate the answer. Accuracy improvements on VQAv2, OKVQA, and AdVQA are reported in Fig. 5, which demonstrates the effectiveness of such intervention. By using our uncertainty estimates to trigger re-answering, we observe consistent accuracy gains across all benchmarks. Importantly, on the difficult **AdVQA** benchmark, the performance of InternVL improves from 47.9% to **51.1%**, and QwenVL from 47.5% to **50.0%**. This significant increase indicates that our method successfully flags epistemic failures where the model was previously hallucinating due to adversarial visual contexts.

We further apply CIM to adaptive-resolution inference, where samples with uncertainty below a percentile threshold are processed at a low maximum resolution of 256, reserving high-resolution

inference for uncertain samples. As shown in Fig. 6, thresholding confident samples reduces inference time and also boosts accuracy beyond the high-resolution baseline. For QwenVL, downscaling the most confident 50% samples reduces average latency from 0.65s to 0.50s and raises accuracy from $\sim 75.2\%$ to $\sim 76.1\%$; InternVL shows similar gains around the 50th–70th percentile thresholds. This suggests that high resolution can distract the inference on already sufficient samples, while CIM identifies when low-resolution evidence is adequate.

Blending Weight. Table 4 varies λ from 0 to 1 to show its effective range. Both signals improve AUROC over ranges containing the default $\lambda = 0.05$, so the benefit is not confined to this weight. Larger weights eventually lose the improvement as the complementary entropy signal is reduced.

Table 4: **Blending-weight sensitivity.** Best λ and evaluated ranges with AUROC above SE-16. IVL/QVL denote InternVL/QwenVL.

		SD		EED	
MLLM	Benchmark	Best λ	Range above SE	Best λ	Range above SE
IVL	VQAv2	0.105	0.001–0.771	0.312	0.001–0.626
IVL	OKVQA	0.167	0.001–0.475	0.298	0.001–0.514
IVL	AdVQA	0.088	0.001–0.234	0.159	0.001–0.304
QVL	VQAv2	0.039	0.001–0.378	0.122	0.001–0.435
QVL	OKVQA	0.031	0.001–0.264	0.167	0.001–0.410
QVL	AdVQA	0.035	0.001–0.157	0.026	0.001–0.163

5 Conclusion

This work addresses the critical bias in existing MLLM uncertainty quantification methods toward aleatoric uncertainty, which often ignores hallucinations driven by spurious correlations. We proposed CIM, a framework that probes epistemic uncertainty by measuring the semantic shift under causally invariant perturbations, supported by theory under explicit assumptions. Our proposed Semantic Divergence targets reliance on non-causal features and complements Semantic Entropy, achieving state-of-the-art performance in combination. Furthermore, the Expected Embedding Drift (EED) reduces computational latency by 50% with competitive detection performance. By enabling robust hallucination detection and efficient adaptive inference, CIM enables a rigorous path toward trustworthy multimodal reasoning.

Limitations. CIM depends on RoI quality and semantic equivalence estimation. Cropping can remove necessary evidence or change scale and spatial context, while answer-conditioned grounding may propagate hallucinated objects. Our controls support the approach in the evaluated settings, but do not eliminate crop-induced distribution shift or establish an exact decomposition of uncertainty. The theoretical interpretation remains conditional on the stated energy and geometric assumptions. Connecting CIM more tightly with causal inference theory and establishing stronger guarantees for the uncertainty decomposition constitutes a promising direction for future work.

Acknowledgments and Disclosure of Funding

This work was supported in part by Young Scientist Fund (No. 62406265) of NSFC, Early Career Scheme (No. 21212526) of RGC, REG (No. 7020206) of the City University of Hong Kong, and the Australian Research Council under Projects DP240101848 and FT230100549.

References

[1] Gustaf Ahdritz, Tian Qin, Nikhil Vyas, Boaz Barak, and Benjamin L Edelman. Distinguishing the knowable from the unknowable with language models. In Proceedings of the 41st International Conference on Machine Learning, pages 503–549, 2024.

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. arXiv preprint arXiv:1907.02893, 2019.

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

[4] Tianyi Bai, Hao Liang, Binwang Wan, Yanran Xu, Xi Li, Shiyu Li, Ling Yang, Bozhou Li, Yifan Wang, Bin Cui, et al. A survey of multimodal large language model from a data-centric perspective. arXiv preprint arXiv:2405.16640, 2024.

[5] 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.

[6] Mathieu Blondel, Michael E Sander, Germain Vivier-Ardisson, Tianlin Liu, and Vincent Roulet. Autoregressive language models are secretly energy-based models: Insights into the lookahead capabilities of next-token prediction. arXiv preprint arXiv:2512.15605, 2025.

[7] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. Advances in Neural Information Processing Systems, 29, 2016.

[8] Miguel A Carreira-Perpinan and Geoffrey Hinton. On contrastive divergence learning. In International workshop on artificial intelligence and statistics, pages 33–40. PMLR, 2005.

[9] Ferhat Ozgur Catak and Murat Kuzlu. Uncertainty quantification in large language models through convex hull analysis. Discover Artificial Intelligence, 4(1):90, 2024.

[10] Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms' internal states retain the power of hallucination detection. In Proceedings of the Twelfth International Conference on Learning Representations, 2024.

[11] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. Science China Information Sciences, 67(12):220101, 2024.

[12] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. Advances in neural information processing systems, 32, 2019.

[13] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. Nature, 630(8017):625–630, 2024.

[14] 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. In The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2025.

[15] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6904–6913, 2017.

[16] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In International Conference on Learning Representations, 2021.

[17] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43(2):1–55, 2025.

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

[19] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Machine learning, 110(3):457–506, 2021.

[20] Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in llms. arXiv preprint arXiv:2406.15927, 2024.

[21] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In The Eleventh International Conference on Learning Representations, 2023.

[22] Sachin Kumar and Yulia Tsvetkov. Von mises-fisher loss for training sequence to sequence models with continuous outputs. In International Conference on Learning Representations, 2019.

[23] Gregory Kang Ruey Lau, Hieu Dao, Nicole Kan Hui Lin, and Bryan Kian Hsiang Low. Uncertainty quantification for mllms. In ICML 2025 Workshop on Reliable and Responsible Foundation Models, 2025.

[24] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, Fujie Huang, et al. A tutorial on energy-based learning. Predicting structured data, 1(0), 2006.

[25] Linjie Li, Jie Lei, Zhe Gan, and Jingjing Liu. Adversarial vqa: A new benchmark for evaluating the robustness of vqa models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 2042–2051, 2021.

[26] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne 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, pages 292–305, 2023.

[27] Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. Transactions on Machine Learning Research, 2024, 2024.

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. Advances in Neural Information Processing Systems, 36:34892–34916, 2023.

[29] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In European conference on computer vision, pages 38–55. Springer, 2024.

[30] Kanti V Mardia and Peter E Jupp. Directional statistics. John Wiley & Sons, 2009.

[31] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3195–3204, 2019.

[32] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. arXiv preprint arXiv:2402.06196, 2024.

[33] Robert Adrian Minut, Hazem Dewidar, Iacopo Masi, et al. Spilled energy in large language models. In Proceedings of the International Conference on Learning Representations, 2026.

[34] Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. Advances in Neural Information Processing Systems, 37:8901–8929, 2024.

[35] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. Causal inference in statistics: A primer. John Wiley & Sons, 2016.

[36] Judea Pearl et al. Models, reasoning and inference. Cambridge, UK: CambridgeUniversityPress, 19(2):3, 2000.

[37] Jingwei Peng, Jiehao Chen, Mateo Alejandro Rojas, and Meilin Zhang. Mv-core: Multimodal visual-conceptual reasoning for complex visual question answering. arXiv preprint arXiv:2508.07023, 2025.

[38] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. Journal of the Royal Statistical Society Series B: Statistical Methodology, 78(5):947–1012, 2016.

[39] Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. Cold decoding: Energy-based constrained text generation with langevin dynamics. Advances in Neural Information Processing Systems, 35:9538–9551, 2022.

[40] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. Proceedings of the IEEE, 109(5):612–634, 2021.

[41] Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. ACM Computing Surveys, 2025.

[42] Linwei Tao, Yi-Fan Yeh, Minjing Dong, Tao Huang, Philip Torr, and Chang Xu. Revisiting uncertainty estimation and calibration of large language models. arXiv preprint arXiv:2505.23854, 2025.

[43] Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 5433–5442, 2023.

[44] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4818–4829, 2024.

[45] Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. arXiv preprint arXiv:2306.13063, 2023.

[46] Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In The Twelfth International Conference on Learning Representations, 2024.

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

[48] Ruiyang Zhang, Hu Zhang, and Zhedong Zheng. Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. arXiv preprint arXiv:2411.11919, 2024.

[49] Ruiyang Zhang, Hu Zhang, Hao Fei, and Zhedong Zheng. Uncertainty-o: One model-agnostic framework for unveiling uncertainty in large multimodal models. arXiv preprint arXiv:2506.07575, 2025.

[50] Xiaofeng Zhang, Fanshuo Zeng, Yihao Quan, Zheng Hui, and Jiawei Yao. Enhancing multimodal large language models complex reason via similarity computation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pages 10203–10211, 2025.

[51] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. arXiv preprint arXiv:2303.18223, 1(2), 2023.

Appendix

A Proofs of Equations and Propositions

A.1 Proof of Lemma 3.2 (Aleatoric Cancellation)

Proof. Let $P(s) \triangleq P_\theta(s|v)$ and $Q(s) \triangleq P_\theta(s|v')$ be the distribution given the original input and the masked input $v' = T(v)$, respectively. The log-probabilities for these inputs are

$$\log Q(s) = -(E_c(s, v_c) + E_s(s, v'_s)) - \log Z(v')$$
$$\log P(s) = -(E_c(s, v_c) + E_s(s, v_s)) - \log Z(v).$$

Subtracting the two equations yields:

$$\begin{aligned}\log \frac{Q(s)}{P(s)} &= (E_s(s, v_s) - E_s(s, v'_s)) + \log \frac{Z(v)}{Z(v')} \\ &= \Delta_s(s) + \log \frac{Z(v)}{Z(v')}. \end{aligned} \tag{9}$$

□

A.2 Proof of Theorem 3.3 (Epistemic Jensen Gap)

Part 1: Equivalence to Jensen Gap

Proof. We first expand the partition function ratio. By definition, $Z(v) = \sum_y \exp(-(E_c(s, v_c) + E_s(s, v_s)))$. The exponent can be rewritten as $-E_c - E_s(v'_s) - (E_s(v_s) - E_s(v'_s)) = -E(s, v') - \Delta_s(s)$. Therefore, we can derive that

$$Z(v) = \sum_s e^{-E_\theta(s,v')} e^{-\Delta_s(s)}. \tag{10}$$

By definition of $Z(v')$, we have $e^{-E_\theta(s,v')} = Z(v')Q(s)$ which simplifies Eq. (10) as

$$\begin{aligned} Z(v) &= \sum_s Z(v')Q(s)e^{-\Delta_s(s)} \\ &= Z(v') \cdot \mathbb{E}_{s\sim Q}[e^{-\Delta_s(s)}] \end{aligned} \tag{11}$$

Thus, the constant term from Lemma 3.2 is $\log \frac{Z(v)}{Z(v')} = \log \mathbb{E}_{y\sim Q}[e^{-\Delta_s(s)}]$. Now, substituting Lemma 3.2 into the definition in Eq. (1) yields

$$\begin{aligned} \mathcal{U}_{\text{sem}} &= \mathbb{E}_{s\sim Q}\left[\log \frac{Q(s)}{P(y)}\right] \\ &= \mathbb{E}_{s\sim Q}\left[\Delta_s(s) + \log \mathbb{E}_{s'\sim Q}[e^{-\Delta_s(s')}]\right] \\ &= \log \mathbb{E}_{s\sim Q}[e^{-\Delta_s(s)}] - \mathbb{E}_{s\sim Q}[-\Delta_s(s)], \end{aligned} \tag{12}$$

which is precisely the Jensen Gap $\mathcal{J}(e^x; -\Delta_s) \geq 0$. The divergence is strictly zero if and only if $\Delta_s(s)$ is constant almost surely over $\mathcal{S}$, implying the model is insensitive to spurious features. While Eq. (12) provides an exact measure, it can be interpreted best at a low hallucination scenario, *i.e.*, the model is trained well enough to produce relatively smaller $|\Delta_s(s)|$, which results in Eq. 3. □

Part 2: Convergence to Variance (Proof of Eq. (3))

Proof. Let $K_{-\Delta}(t) = \log \mathbb{E}_{s\sim Q}[e^{t(-\Delta_s(s))}]$ be the Cumulant Generating Function (CGF) of the random variable $-\Delta_s(s)$. From Theorem 1, we identify that:

$$D_{KL}(Q\|P) = K_{-\Delta}(1) - \mathbb{E}_{s\sim Q}[-\Delta_s(s)] = K_{-\Delta}(1) + \mathbb{E}_{s\sim Q}[\Delta_s(s)] \tag{13}$$

We apply Taylor's Theorem with the Lagrange form of the remainder to $K_{-\Delta}(t)$ on the interval $[0, 1]$. There exists some $\xi \in (0, 1)$ such that:

$$K_{-\Delta}(1) = K_{-\Delta}(0) + K'_{-\Delta}(0)(1) + \frac{K''_{-\Delta}(0)}{2!}(1)^2 + \frac{K'''_{-\Delta}(\xi)}{3!}(1)^3 \tag{14}$$

Using the properties of cumulants, we have $K_{-\Delta}(0) = 0$, $K'_{-\Delta}(0) = \mathbb{E}_Q[-\Delta_s(s)]$, and $K''_{-\Delta}(0) = \mathrm{Var}_Q(-\Delta_s(s))$. Given the assumption that the shift is bounded by $\sup|\Delta_s(s)| \leq \epsilon$, the third derivative satisfies $|K^{(3)}_{-\Delta}(\xi)| \leq C\epsilon^3$ for some constant C, effectively bounding the remainder term by $\mathcal{O}(\epsilon^3)$. Substituting these values:

$$K_{-\Delta}(1) = -\mathbb{E}_Q[\Delta_s(s)] + \frac{1}{2}\mathrm{Var}_Q(\Delta_s(s)) + \mathcal{O}(\epsilon^3) \tag{15}$$

Substituting this expansion back into the KL equation, the linear expectation terms cancel:

$$\begin{aligned} D_{KL}(Q\|P) &= \left(-\mathbb{E}_Q[\Delta_s(s)] + \frac{1}{2}\mathrm{Var}_Q(\Delta_s(s)) + \mathcal{O}(\epsilon^3)\right) + \mathbb{E}_Q[\Delta_s(s)] \\ &= \frac{1}{2}\mathrm{Var}_{s\sim Q}[\Delta_s(s)] + \mathcal{O}(\epsilon^3) \end{aligned} \tag{16}$$

□

A.3 Proof of Lemma 3.5 (KL Divergence of vMF Distributions)

Proof. By definition, the KL divergence is:

$$\begin{aligned} D_{KL}(Q\|P) &= \int_{\mathbb{S}^{d-1}} q(\mathbf{z}) \log \frac{q(\mathbf{z})}{p(\mathbf{z})} d\mathbf{z} \\ &= \int_{\mathbb{S}^{d-1}} q(\mathbf{z}) \left(\log \frac{C_d(\kappa)e^{\kappa \boldsymbol{\mu}_Q^\top \mathbf{z}}}{C_d(\kappa)e^{\kappa \boldsymbol{\mu}_P^\top \mathbf{z}}}\right) d\mathbf{z} \\ &= \int_{\mathbb{S}^{d-1}} q(\mathbf{z}) \left(\kappa \boldsymbol{\mu}_Q^\top \mathbf{z} - \kappa \boldsymbol{\mu}_P^\top \mathbf{z}\right) d\mathbf{z} \\ &= \kappa(\boldsymbol{\mu}_Q - \boldsymbol{\mu}_P)^\top \int_{\mathbb{S}^{d-1}} \mathbf{z} q(\mathbf{z}) d\mathbf{z}. \end{aligned} \tag{17}$$

A property of the vMF distribution is that the expected vector is $\mathbb{E}_{\mathbf{z}\sim Q}[\mathbf{z}] = A_d(\kappa)\boldsymbol{\mu}_Q$. Substituting this into the integral:

$$\begin{aligned} D_{KL}(Q\|P) &= \kappa(\boldsymbol{\mu}_Q - \boldsymbol{\mu}_P)^\top (A_d(\kappa)\boldsymbol{\mu}_Q) \\ &= \kappa A_d(\kappa)(\boldsymbol{\mu}_Q^\top \boldsymbol{\mu}_Q - \boldsymbol{\mu}_P^\top \boldsymbol{\mu}_Q). \end{aligned} \tag{18}$$

Since $\boldsymbol{\mu}_Q$ is a unit vector, $\boldsymbol{\mu}_Q^\top \boldsymbol{\mu}_Q = 1$. Thus, $D_{KL}(Q\|P) = \kappa A_d(\kappa)(1 - \boldsymbol{\mu}_P^\top \boldsymbol{\mu}_Q)$. □

A.4 Proof of Theorem 3.6 (Geometric Proxy Equivalence)

Proof. The EED metric computes the expectation of the cosine distance over the sampled trajectory from Q. Taking the expectation of Eq. (6) with respect to $y_t \sim Q$:

$$\begin{aligned} \mathbb{E}_{y_t\sim Q}[\mathcal{U}_{\text{EED}}] &= 1 - \mathbb{E}_{y_t\sim Q}[\mathbf{e}_{y_t}]^\top \boldsymbol{\mu}_P \\ &= 1 - (A_d(\kappa)\boldsymbol{\mu}_Q)^\top \boldsymbol{\mu}_P. \end{aligned} \tag{19}$$

Rearranging terms for $\boldsymbol{\mu}_Q^\top \boldsymbol{\mu}_P$ and substituting into Lemma 3.5:

$$\begin{aligned} D_{KL}(Q\|P) &= \kappa A_d(\kappa)\left(1 - \frac{1 - \mathbb{E}[\mathcal{U}_{\text{EED}}]}{A_d(\kappa)}\right) \\ &= \kappa \mathbb{E}_{y_t\sim Q}[\mathcal{U}_{\text{EED}}] - \kappa(1 - A_d(\kappa)). \end{aligned} \tag{20}$$

The divergence relates to EED via an affine transformation. The term $\kappa(1 - A_d(\kappa))$ is a constant dependent only on concentration κ and dimension d. Under high concentration ($\kappa \to \infty$), $A_d(\kappa) \to 1$, and the relation simplifies to exact linearity. Thus, minimizing EED strictly minimizes the divergence between the semantic distributions. □

B Detailed Experiment Setup

Computational Resources We carried out all experiments on 4 NVIDIA RTX A6000 GPUs.

Model Parameters The implementation uses an inference temperature of 0.1, while the sampling temperature is 0.5. The RoI transform uses GroundingDINO with detection threshold 0.3, which filters boxes labeled as the whole image when they cover more than 95% of the image area and uses a central crop fallback when no valid detection remains.

Prompts The dataset question is passed to the evaluated MLLM with this prompt:

```
[Question] Answer this question shortly.
```

The prompt used to extract key objects for grounding is:

```
What are the key objects in this question-answer pair?
Q: [Question]
A: [Answer]
Output in json format: {"key_objects": ["key object 1", "key object 2", ...]}
```

where the placeholders are filled with the original question and the model's original answer, and the judge LLM is queried with temperature 0.1. The prompt used for pairwise semantic clustering of sampled answers is:

```
Question: [Question] True Answer: [Answer A] Is the following answer correct?
Answer: [Answer B] Respond with 'no' if this is totally different from the
True Answer. Respond with 'yes' when they are of the same meaning.
```

This clustering prompt is filled with the text query and the answer pairs, using a temperature of 0.1.

Causal-Invariant Transform. We use GroundingDINO as a lightweight RoI generator for the causal-invariant transform $T(\cdot)$. When no valid RoI is found, we use a central crop as the fallback view. This choice is not essential to the framework: task-specific or query-conditioned detectors can be substituted when they are available.

Dataset Configuration. To keep evaluation consistent and computationally feasible, we fix the evaluation subsets across all methods. For hallucination-targeted benchmarks, we use the full test sets: 3,000 POPE examples across the random, popular, and adversarial splits, and 2,374 MME examples. For VQAv2, OKVQA, AdVQA, and GQA, we use the first 2,000 validation examples from each dataset.

Sampling Budget and Fair Comparison. For all sampling-based methods, we fix the total decoding budget to $K_{\text{total}} = 32$ sequences per query.

- **Baselines:** SE, EigenScore, and UMPIRE sample only from the original distribution $P(y|v)$, so we set $K = 32$.
- **Ours (CIM):** CIM samples from both the original distribution $P(y|v)$ and the causally-filtered distribution $P(y|v')$. We use $K = 16$ original-view samples and $K' = 16$ filtered-view samples, keeping $K + K' = 32$.

Metric Interpolation. The $K = 16$ original-view samples used by CIM can also estimate standard Semantic Entropy without extra decoding. We therefore report a mixed score $\mathcal{U}_{\text{final}} = \lambda \mathcal{U}_{\text{ours},K'=16} + (1 - \lambda) \cdot \mathcal{U}_{\text{SE},K=16}$, where $\mathcal{U}_{\text{ours}}$ is either $\mathcal{U}_{\text{sem}}$ or $\mathcal{U}_{\text{EED}}$. Here, λ is set to 0.05. This score retains sensitivity to aleatoric uncertainty while adding the epistemic signal measured by CIM.

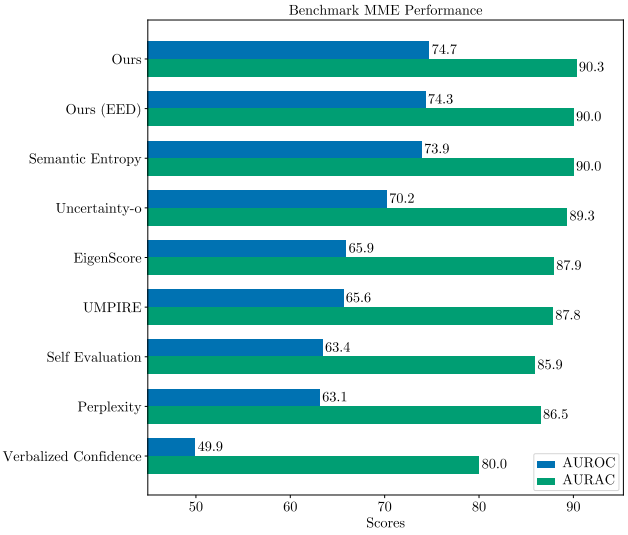

Figure 7: **Uncertainty estimation performance** (%) on the MME benchmark.

C More Experiments and Results

C.1 Additional MME Results

Figure 7 reports hallucination detection on MME. CIM achieves the best performance (74.7%), showing that the uncertainty signal generalizes beyond the VQA datasets in the main tables.

C.2 Case Study



Figure 8: **Representative instances of SE vs. $\mathcal{U}_{\text{sem}}$ on hallucinated samples.** In all instances, our proposed CIM uncertainty is high despite the standard SE remains low, indicating epistemic uncertainty that common sampling-based uncertainty metric fails to identify.

To better understand the behavior of CIM in practice, we present a case analysis of SE's failure cases in Fig. 8. As illustrated, the MLLM frequently generates incorrect answers with high confidence, leading to extremely low SE scores. For instance, in the left example, the model consistently mispredicts 'Japanese dish,' likely misled by background Japanese text. Similarly, in the right example, the model hallucinates the color possibly by looking at an incorrect object. Empirically, counting complex objects can often hallucinate but result in a stable number or low semantic entropy, as shown in the middle instance. In these scenarios, standard metrics fail but the proposed method consistently peaks. Such behavior confirms the successful detection of epistemic failure with our CIM.

C.3 Gaussian Noise Perturbation

We compare RoI masking with the commonly used gaussian noise perturbation under the same VQAv2 and AdVQA settings as the main masking ablation. Table 5 reports the final mixed uncertainty score. Gaussian noise perturbs the entire image and therefore cannot selectively preserve x_c while suppressing x_s. Its performance stays close to random masking and below RoI masking across benchmarks. This supports the main conclusion that the improvement is not caused by simply changing the image, but by constructing an evidence-preserving view that removes non-causal context.

Table 5: **Gaussian noise perturbation ablation.** Results are AUROC/AURAC on VQAv2 and AdVQA.

Benchmark	VQAv2				AdVQA			
MLLM	InternVL		QwenVL		InternVL		QwenVL	
Metric	AUROC	AURAC	AUROC	AURAC	AUROC	AURAC	AUROC	AURAC
Random Masking	80.2	83.6	77.9	81.6	75.7	66.7	71.9	61.5
Gaussian Noise	80.3	84.0	77.8	80.7	75.5	66.5	73.4	63.0
Reversed Masking	77.6	81.5	76.1	79.2	74.8	65.0	71.9	61.3
Ours (RoI Masking)	**81.3**	**85.2**	**78.5**	**81.9**	**76.8**	**66.9**	**74.5**	**64.8**

C.4 Detector Failure Sensitivity

We next test how performance changes when RoI quality is deliberately degraded. Starting from GT boxes on GQA, we replace an increasing fraction of boxes with random crops. We evaluate and show the uncertainty detection performance in Table 6.

Table 6: **GQA detector-failure sensitivity.** A higher failure ratio means that more GT RoIs are replaced by random crops.

Failure Ratio (%)	InternVL AUROC	QwenVL AUROC
0	72.1	66.3
20	69.4	62.8
40	67.1	61.4
60	65.1	59.7
80	63.3	58.7
100	61.0	57.5

The AUROC drops steadily as valid RoIs are replaced, confirming that CIM relies on preserving causal evidence rather than arbitrary perturbations such as random cropping. Meanwhile, In the actual pipeline, the detector fallback rate is low across benchmarks, as shown in Table 7.

Table 7: **GroundingDINO fallback rate.** Fallback means no valid RoI is detected and the central crop is used.

Benchmark	VQAv2	OKVQA	AdVQA	MME
Failure (%)	1.10	0.62	1.52	1.54

C.5 Relational and Global GQA Questions

A concern with RoI cropping is that it may remove relationships or scene context. Our implementation uses the union box over detected query-relevant objects, so the spatial region between objects is preserved. We evaluate this directly on GQA by comparing common questions with relational/global questions.

From the results in Table 8, we can see the gain is still maintained when testing relational/global questions, suggesting that the union box in our implementation keeps enough contextual structure for CIM to remain effective.

Table 8: **GQA relational/global analysis.** AUROC on common questions and complex questions involving relational or global reasoning.

SETTING	GQA-COMMON		GQA-REL-GLOBAL	
METRIC	INTERNVL	QWENVL	INTERNVL	QWENVL
SE	78.4	69.2	79.2	69.4
OURS ($K + K'$)	**79.9**	**70.2**	**80.2**	**71.5**

C.6 Alternative Grounding with Florence-2

CIM does not require GroundingDINO or a union box specifically. To test whether the framework is sensitive to the detector choice, we replace GroundingDINO with the query-conditioned Florence-2 detector [44].

Table 9: **Query-conditioned detector analysis.** Results are AUROC/AURAC. GD denotes GroundingDINO and FL denotes Florence-2.

BENCHMARK	INTERNVL (GD)	INTERNVL (FL)	QWENVL (GD)	QWENVL (FL)
VQAV2	81.3/85.2	**82.2/85.9**	**78.4**/81.5	77.2/**82.6**
ADVQA	76.8/66.9	**77.7/68.3**	**74.5**/64.8	73.5/**65.0**

In the results shown in Table 9, the changes are small, indicating that CIM mainly depends on preserving query-relevant evidence and is not tied to a particular RoI detector.

C.7 Question-Only Grounding and Box Disagreement

The default grounding prompt uses the model's original answer to identify key objects, which can propagate a hallucinated object into the crop. We therefore construct question-only RoIs with Florence-2 while keeping the evaluation examples and sampling settings fixed. Table 10 compares this pipeline with answer-conditioned grounding for pure SD and the SE+SD blend.

The two grounding choices show mixed changes: question-only grounding improves both scores on InternVL, while QwenVL shows smaller blended-score decreases and mixed pure-score changes. This indicates that conditioning on the original answer is not essential to CIM. Answer-conditioned grounding can supply localization cues for underspecified questions, whereas question-only grounding avoids using a potentially hallucinated answer.

Table 10: **Question-only grounding for SD.** AUROC (%) for pure SD and SE+SD with answer-conditioned (AC) or question-only (Q-only) crops. Differences are in AUROC points.

MLLM / Benchmark	Score	AC	Q-only	Difference
IVL / VQAv2	Pure	77.69	78.08	+0.39
IVL / VQAv2	Blended	81.78	82.66	+0.88
IVL / AdVQA	Pure	70.99	74.10	+3.11
IVL / AdVQA	Blended	77.45	78.10	+0.65
QVL / VQAv2	Pure	74.07	72.26	-1.81
QVL / VQAv2	Blended	78.51	78.33	-0.17
QVL / AdVQA	Pure	69.47	69.77	+0.30
QVL / AdVQA	Blended	75.87	75.12	-0.75

Table 11 measures overlap between the two sets of boxes. The frequency of IoU below 0.5 is higher on hallucinated than correct examples in all four settings. This disagreement motivates caution about answer-conditioned grounding, but is not itself a ground-truth crop-error rate: either grounding choice may fail to preserve the required evidence.

Table 11: **Grounding disagreement.** Box IoU compares answer-conditioned and question-only crops. Counts and mean/median IoU refer to hallucinated examples; the last two columns give disagreement rates by answer correctness.

				IoU < 0.5	
MLLM	Benchmark	Hallucinated n	Mean/median IoU	Hallucinated	Correct
IVL	VQAv2	639	0.663 / 0.862	34.6%	28.2%
IVL	AdVQA	1,002	0.592 / 0.678	42.4%	34.4%
QVL	VQAv2	674	0.649 / 0.806	35.5%	31.4%
QVL	AdVQA	996	0.594 / 0.682	41.6%	33.9%

C.8 Empirical Validations for Energy Assumptions

Practical causal energy could shift and spurious elements might remain. To further verify the influence of the decomposition assumption empirically, we evaluated logit causal energy by comparing our crops against vacant-background views whose area outside the crop is masked to black. We estimate causal and non-causal logit-energy shifts and compare their median magnitudes with the median top-1/top-2 logit gap. Table 12 reports retained-evidence shifts of 9.0–20.7% of the corresponding decision margins. These measurements show that the shift is relatively small, validating the decomposition accuracy.

Table 12: **Retained-evidence energy shift scale.** Median absolute shift and median top-1/top-2 logit gap; relative shift is their ratio.

Benchmark	MLLM	$\|\Delta_c\|$	Logit gap	Relative shift
OKVQA	InternVL	0.75	3.62	20.7%
OKVQA	QwenVL	0.38	4.21	9.0%
AdVQA	InternVL	0.50	4.57	10.9%
AdVQA	QwenVL	0.50	4.81	10.4%

To examine the energy-shift scale relevant to Theorem 3.3, we empirically evaluated its magnitude, i.e., the logit difference between original and masked views, on hallucinated samples. The results are shown in Table 13. Note that this value is accurate under our causal-invariance assumption. For scale anchoring, we compared $|\Delta_s|$ against the median top-1/top-2 logit gap. The median $|\Delta_s|$ (0.50-1.00) is consistently much smaller than the median decision margin (3.62-4.81). The relative shifts are 11.9–21.9% of the decision margins. This value is sufficiently small relative to the logit scale, confirming our theoretical approximation.

Table 13: **Context-related energy shift scale on hallucinated samples.** Median absolute shift and median top-1/top-2 logit gap; relative shift is their ratio.

Benchmark	MLLM	$\|\Delta_s\|$	Logit gap	Relative shift
OKVQA	InternVL	0.75	3.62	20.7%
OKVQA	QwenVL	0.50	4.21	11.9%
AdVQA	InternVL	1.00	4.57	21.9%
AdVQA	QwenVL	0.62	4.81	12.9%

Both $\|\Delta_c\|$ and $\|\Delta_s\|$ are substantially smaller than the corresponding decision margins. These estimates provide empirical context for the approximation in Theorem 3.3 and suggest limited retained-evidence disturbance in the evaluated settings.

C.9 POPE Entropy Distributions

Table 14 reports the average token-level entropy during decoding on POPE. QwenVL has a much sharper output distribution than InternVL. In this regime, a token-distribution change may leave the sampled answers in the same semantic cluster, whereas EED can retain continuous token-level changes. This helps explain the POPE performance reversal.

Table 14: **Average token-level entropy on POPE.**

MLLM	RANDOM	POPULAR	ADVERSARIAL
INTERNVL-8B	0.51	0.51	0.54
QWEN2.5-VL-7B	0.06	0.07	0.08

C.10 NLI Model Ablation

Table 15: **NLI model ablation.** Δ(SD Qwen, SD DeBERTa) represents the metric value difference between "SD w/ Qwen-7B" and "SD w/ DeBERTa-small".

METRIC	VQAV2		OKVQA	
	AUROC	AURAC	AUROC	AURAC
SE W/ DEBERTA-SMALL	74.5	78.0	77.5	83.5
SD W/ DEBERTA-SMALL	76.2	80.4	78.6	84.4
SE W/ QWEN-7B	76.5	79.5	78.6	83.8
SD W/ QWEN-7B	**78.4**	**81.5**	**79.6**	84.2
Δ(SD QWEN, SD DEBERTA)	2.2	1.1	1.0	-0.2

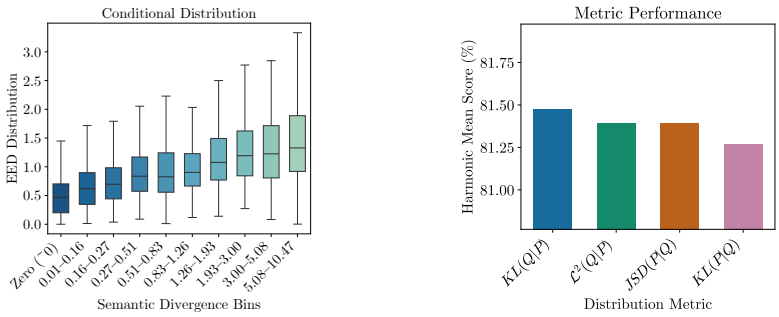

Figure 9: **Left:** Expected Embedding Drift (EED) as a proxy for Semantic Divergence. **Right:** Harmonic mean of AUROC and AURAC for different distribution distances, averaged across 2 MLLMs and 7 benchmarks.

Semantic clustering depends on an entailment model, so we compare the default Qwen-7B entailment model with a much smaller DeBERTa-small model [16]. Table 15 shows that the weaker entailment model slightly changes performance, but SD remains competitive and continues to outperform SE under the same NLI model.

C.11 EED Proxy and Distance Metrics

Figure 9 left validates the connection between EED and Semantic Divergence by grouping samples according to $\mathcal{U}_{\text{sem}}$ and plotting the corresponding EED values. The mean EED increases with Semantic Divergence, supporting EED as an efficient proxy for semantic distribution shift.

Figure 9 right compares squared L_2 distance, Jensen-Shannon Divergence, reverse KL $D_{KL}(P\|Q)$, and forward KL $D_{KL}(Q\|P)$. The method is not highly sensitive to the exact distance, but forward KL performs best in our experiments, consistent with its role as an information-gain measure from the causally-filtered view to the original view.

C.12 Statistical Uncertainty of Detection Gains

We evaluate statistical uncertainty using a paired bootstrap over shared evaluation examples. For each of 10,000 resamples, we resample examples jointly for SE and the blended score and recompute their AUROC difference. Tables 16 and 17 report the point estimates, gains, 90% confidence intervals, and paired bootstrap p-values. The bootstrap analyses use evaluation sets of different sizes from those

used for the main performance tables, explaining the differences in point estimates. Each comparison within the bootstrap analysis uses the same examples for both methods. These intervals quantify variability over evaluation examples, rather than variation across generation seeds.

On VQA, eleven of the twelve gain intervals are above zero. The exception is SE+EED on InternVL–AdVQA, whose interval includes zero; we therefore do not claim a statistically supported improvement in that setting. All POPE and MME intervals are above zero. The reported gains and interval endpoints are retained at their original precision, so gains need not equal differences between rounded table entries exactly.

Table 16: **Paired-bootstrap results on VQA.** AUROC (%) and gains in AUROC points, with 90% confidence intervals from 10,000 paired resamples. These results use a different evaluation size from the main performance tables.

MLLM	Benchmark	Blend	SE	Blended	Gain (90% CI)	p
IVL	VQAv2	SE+SD	78.73	81.32	+2.59 [1.84, 3.34]	< 0.001
IVL	VQAv2	SE+EED	78.73	80.53	+1.79 [1.22, 2.37]	< 0.001
IVL	OKVQA	SE+SD	80.55	81.82	+1.26 [0.67, 1.86]	< 0.001
IVL	OKVQA	SE+EED	80.55	81.22	+0.67 [0.26, 1.07]	0.003
IVL	AdVQA	SE+SD	75.51	76.85	+1.33 [0.76, 1.89]	< 0.001
IVL	AdVQA	SE+EED	75.51	75.75	+0.24 [-0.08, 0.55]	0.109
QVL	VQAv2	SE+SD	76.53	78.52	+1.98 [1.45, 2.54]	< 0.001
QVL	VQAv2	SE+EED	76.53	78.13	+1.59 [0.95, 2.25]	< 0.001
QVL	OKVQA	SE+SD	78.59	79.61	+1.02 [0.63, 1.44]	< 0.001
QVL	OKVQA	SE+EED	78.59	79.67	+1.08 [0.53, 1.64]	< 0.001
QVL	AdVQA	SE+SD	73.57	74.61	+1.04 [0.62, 1.47]	< 0.001
QVL	AdVQA	SE+EED	73.57	74.61	+1.05 [0.57, 1.52]	< 0.001

Table 17: **Paired-bootstrap results on POPE and MME.** AUROC (%) and gains in AUROC points, with 90% confidence intervals from 10,000 paired resamples. These results use a different evaluation size from the main performance tables.

MLLM	Benchmark	Blend	SE	Blended	Gain (90% CI)	p
IVL	POPE rand.	SE+SD	70.60	75.93	+5.33 [4.22, 6.48]	< 0.001
IVL	POPE rand.	SE+EED	70.60	73.10	+2.50 [1.26, 3.77]	< 0.001
IVL	POPE pop.	SE+SD	68.72	74.81	+6.09 [4.91, 7.35]	< 0.001
IVL	POPE pop.	SE+EED	68.72	70.97	+2.25 [1.09, 3.43]	0.001
IVL	POPE adv.	SE+SD	71.59	75.05	+3.46 [2.54, 4.41]	< 0.001
IVL	POPE adv.	SE+EED	71.59	73.00	+1.42 [0.38, 2.48]	0.013
QVL	POPE rand.	SE+SD	62.94	66.44	+3.50 [2.29, 4.74]	< 0.001
QVL	POPE rand.	SE+EED	62.94	83.27	+20.33 [18.62, 22.00]	< 0.001
QVL	POPE pop.	SE+SD	64.79	66.31	+1.53 [0.51, 2.55]	0.006
QVL	POPE pop.	SE+EED	64.79	81.51	+16.72 [15.13, 18.32]	< 0.001
QVL	POPE adv.	SE+SD	62.19	64.66	+2.47 [1.37, 3.60]	< 0.001
QVL	POPE adv.	SE+EED	62.19	78.68	+16.49 [14.93, 18.05]	< 0.001
IVL	MME	SE+SD	73.87	74.67	+0.79 [0.28, 1.33]	0.004
IVL	MME	SE+EED	73.87	74.86	+0.98 [0.29, 1.68]	0.009
QVL	MME	SE+SD	64.36	67.19	+2.83 [2.04, 3.65]	< 0.001
QVL	MME	SE+EED	64.36	70.77	+6.41 [5.51, 7.33]	< 0.001

C.13 Sensitivity to Uncertainty-Source

To further examine whether SD captures a signal different from SE, we compare Gaussian blur with background replacement. Blur degrades visual evidence and serves as a proxy for increased data ambiguity. Background replacement preserves the RoI while introducing misleading surrounding context, probing sensitivity to contextual distractions. Table 18 extends the comparison in Fig. 1 with mean QwenVL scores under both perturbations.

Pure SD is higher under background replacement than blur across all seven benchmarks. SE exhibits the opposite ordering on six benchmarks, with OKVQA as the exception. Thus, the metrics

respond differently to the two perturbations, supporting their complementary roles. Although the two perturbations are not perfectly isolated, such results can already provide insight into the different uncertainty sources captured by SE and SD.

Table 18: **Sensitivity to different uncertainty sources.** Mean SE and pure SD under Gaussian blur and background replacement (paste). Differences retain the precision reported in the evaluation.

Benchmark	SE: Blur / Paste	SE: Blur − Paste	SD: Blur / Paste	SD: Paste − Blur
VQAv2	0.613 / 0.565	+0.047	1.468 / 1.636	+0.168
OKVQA	0.719 / 0.744	-0.025	1.379 / 1.454	+0.075
AdVQA	0.846 / 0.680	+0.166	1.581 / 2.021	+0.440
POPE rand.	0.127 / 0.112	+0.015	0.395 / 0.709	+0.314
POPE pop.	0.156 / 0.139	+0.017	0.631 / 0.958	+0.327
POPE adv.	0.138 / 0.114	+0.025	0.513 / 0.788	+0.275
MME	0.196 / 0.169	+0.027	0.186 / 0.376	+0.190

C.14 Risk-Coverage Performance

We additionally evaluate selective prediction through the area under the risk-coverage curve (AURC), where lower values indicate lower retained-set risk as uncertain examples are rejected. Table 19 reports AURC for SE and both blends on the VQA benchmarks. Both SE+SD and SE+EED reduce AURC in all six model–benchmark settings, with reported reductions ranging from 0.28 to 2.25 points. This complements the AURAC results in the main tables and shows that the additional context-sensitivity signal also improves selective prediction.

Table 19: **Risk-coverage performance on VQA.** Lower AURC is better; reductions relative to SE are in points.

MLLM	Benchmark	SE ↓	SE+SD ↓	Reduction	SE+EED ↓	Reduction
InternVL	VQAv2	17.49	15.24	2.25	15.48	2.01
InternVL	OKVQA	15.76	14.73	1.04	15.48	0.28
InternVL	AdVQA	33.86	32.68	1.18	33.44	0.42
QwenVL	VQAv2	20.68	18.62	2.05	18.70	1.97
QwenVL	OKVQA	16.64	15.54	1.10	15.90	0.75
QwenVL	AdVQA	36.71	35.20	1.51	35.10	1.62

D Licenses for Existing Assets

We use existing public datasets, models, and detector assets for research evaluation only and do not redistribute their data or checkpoints. We cite the original creators in the main paper and follow the licenses or terms listed by their official project pages:

- **VQAv2** [15]: the VQA Consortium states that VQA annotations are licensed under Creative Commons Attribution 4.0 International, and its software is distributed under BSD-style redistribution terms; see `https://visualqa.org/terms.html`. The associated images are COCO images and remain subject to the original COCO/image-source terms.
- **OK-VQA** [31]: the official download page provides OK-VQA annotations/questions and COCO images, but does not state a separate OK-VQA annotation license; see `https://okvqa.allenai.org/download.html`. We therefore use the benchmark under the official access terms and the underlying COCO image terms.
- **AdVQA** [25]: the official release page lists the AdVQA/AVQA data and image sources, but does not state a separate dataset license; see `https://adversarialvqa.org/download.html`. We use it under the official access terms and the source-image terms for COCO, Conceptual Captions, Fakeddit, and VCR as applicable.
- **GQA** [18]: the official GQA website provides images that come from COCO and Flickr and scene graphs based on Visual Genome; see `https://cs.stanford.edu/people/`

dorarad/gqa/index.html and https://cs.stanford.edu/people/dorarad/gqa/download.html. We could not find a separate license statement; the LMMs-Lab Hugging Face mirror lists MIT, but we treat the original COCO, Flickr, and Visual Genome terms as applicable to them.

- **POPE** [26]: the official RUCAIBox/POPE repository is released under the MIT license; see https://github.com/RUCAIBox/POPE. Its COCO-derived components also inherit the relevant COCO image/annotation terms.
- **MME** [14]: the project page links to the Hugging Face dataset release, which lists Apache-2.0 as the dataset license; see https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models/tree/Evaluation and https://huggingface.co/datasets/darkyarding/MME.
- **InternVL2-8B** [11]: the official Hugging Face model card lists an MIT license and notes that its InternLM2.5-7B-Chat component is licensed under Apache-2.0; see https://huggingface.co/OpenGVLab/InternVL2-8B.
- **Qwen2.5-VL-7B-Instruct** and **Qwen2.5-7B-Instruct** [3]: their official Hugging Face model cards list Apache-2.0; see https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct and https://huggingface.co/Qwen/Qwen2.5-7B-Instruct.
- **GroundingDINO** [29]: the official repository license file is Apache-2.0; see https://github.com/IDEA-Research/GroundingDINO/blob/main/LICENSE.
- **Florence-2** [44]: the official Microsoft Hugging Face model card lists an MIT license; see https://huggingface.co/microsoft/Florence-2-large.
- **DeBERTa-small** [16]: the Microsoft DeBERTa repository and the microsoft/deberta-v3-small Hugging Face checkpoint list the MIT license; see https://github.com/microsoft/DeBERTa/blob/master/LICENSE and https://huggingface.co/microsoft/deberta-v3-small.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction match the paper's actual contributions in Sections 3 and 4: CIM/SD probe context-related epistemic uncertainty under the stated assumptions, EED is the faster geometric proxy, and the reported gains on AdVQA/POPE and the 50% runtime reduction are supported by the experiments and appendix.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper does discuss limitations, although not in a standalone limitations section: Section 3 and the appendix note the approximations behind SD/EED, the cost of semantic clustering, and the sensitivity to detector quality and sampling allocation.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The main theoretical statements are numbered in Section 3, their assumptions are stated there, and complete proofs are given in Appendix A.1–A.4.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper specifies the models, benchmarks, detector, subset sizes, and total sampling budget in Section 4 and Appendix B. It also discloses run-critical settings such as the decoding parameters, the prompts, and the interpolation weight λ.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We have provided them in the supplementary material, and will make them public after the final revision.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper gives the main data splits, models, detector, and sampling budgets in Section 4 and Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: In the Appendix, We report AUROC gains with 90% confidence intervals and paired bootstrap p-values from 10,000 resamples over shared evaluation examples. These quantify example-level variability, and the evaluation setup and the setting whose interval includes zero are explicitly described.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The appendix reports computing environment, sample budgets, and per-sample timing, including the roughly 50% speedup of the proposed EED. However, including the amount of compute required for each of the individual experimental runs can be too cumbersome given the large number of experiments conducted.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is an evaluation/method paper on public benchmarks and pretrained models, with no human-subject study or other obvious ethics issue, and we do not see any conflict with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: The paper emphasizes positive uses such as hallucination detection, mitigation, and adaptive inference, which generally exhibit little societal impact.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release a new high-risk model or dataset; it only evaluates existing models on public benchmarks, so there are no release safeguards to describe.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper cites the datasets and pretrained models it uses, and states their licenses and terms of use in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The contribution is a method rather than a new released dataset, code package, or model asset, so there is no new asset documentation to provide.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper uses public benchmarks and model outputs only; it does not report crowdsourcing or any human-subject data collection.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No human-subject study is reported, so IRB approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Section 3 and Appendix C.10 describe the use of MLLMs as the evaluated systems and the Qwen-7B/DeBERTa entailment models inside semantic clustering, so LLM usage is part of the core method and is explicitly described.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.