NEURAL INFORMATION PROCESSING SYSTEMS

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



The Hallucination Problem in MLLMs

Multimodal LLMs connect visual perception with textual semantics, and have achieved remarkable success in tasks like VQA and complex reasoning.
The core challenge: MLLMs suffer from **hallucinations** — they confidently generate factually incorrect but grammatically plausible content.
Examples include non-existent objects, incorrect attributes, and fabricated events in responses.

Critical need: MLLMs do not internally realize their visual capability boundary
→ robust **Uncertainty Quantification (UQ)** is essential for trustworthy deployment.

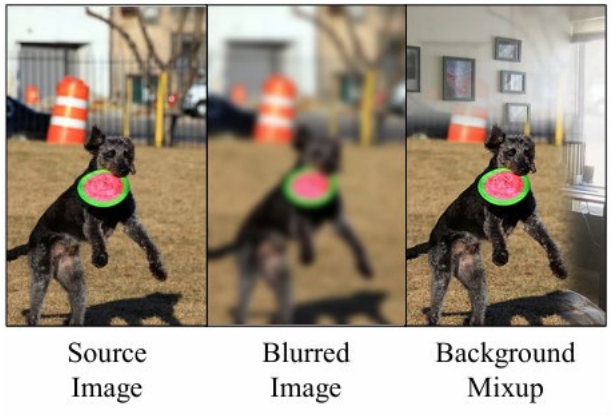


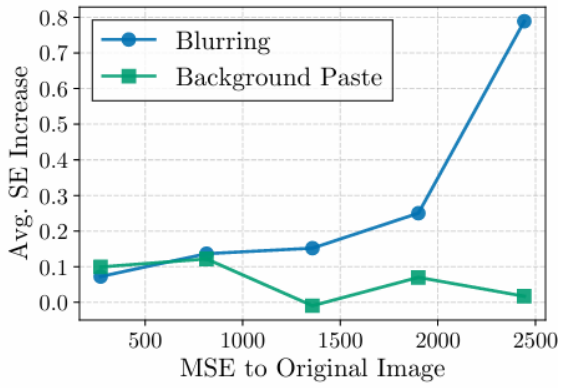
The Key Limitation: Aleatoric Bias

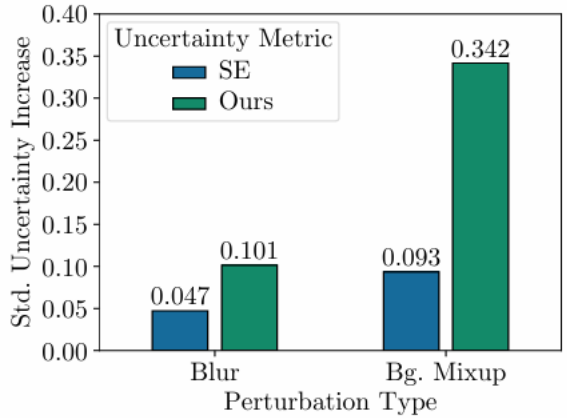
Two types of uncertainty (Hüllermeier & Waegeman, 2021):

- **Aleatoric uncertainty** — arises from data ambiguity (noise, incomplete information)
- **Epistemic uncertainty** — arises from model limitations (insufficient perception or comprehension)

The problem: Existing UQ methods implicitly assume all uncertainty can be estimated equally.
We show their paradigm is **biased toward aleatoric uncertainty**, leaving epistemic uncertainty underestimated.
Consequence: They fail to detect hallucinations caused by **superficial/spurious correlations** — e.g., a background object misleading the model about the foreground.
Empirical evidence: Semantic Entropy rises under both visual blur and background mixup. Our Semantic Divergence stays stable under blur but spikes under hallucinative perturbations — capturing epistemic uncertainty.







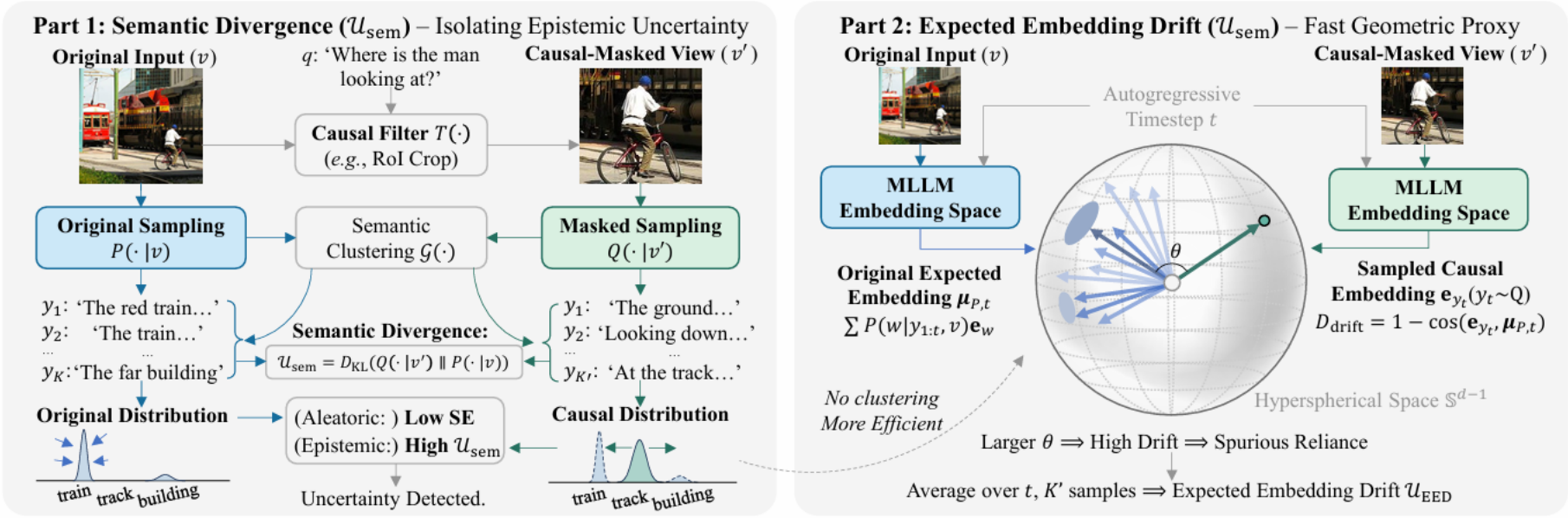
Causal-Invariant Masking (CIM)

- **Core Insight:** Robust reasoning must remain **invariant** when task-irrelevant features are removed.

Causal-Invariant Masking (CIM): A framework that **disentangles epistemic instability from data ambiguity** by probing semantic consistency under causally invariant transforms.

Key idea:

1. Construct a **Region-of-Interest (RoI) view** that retains only causal (query-relevant) features.
2. Measure the **shift in the model's belief distribution** between the original and focused inputs.
3. Large shift → model relies on spurious correlations → **epistemic uncertainty**



Semantic Divergence

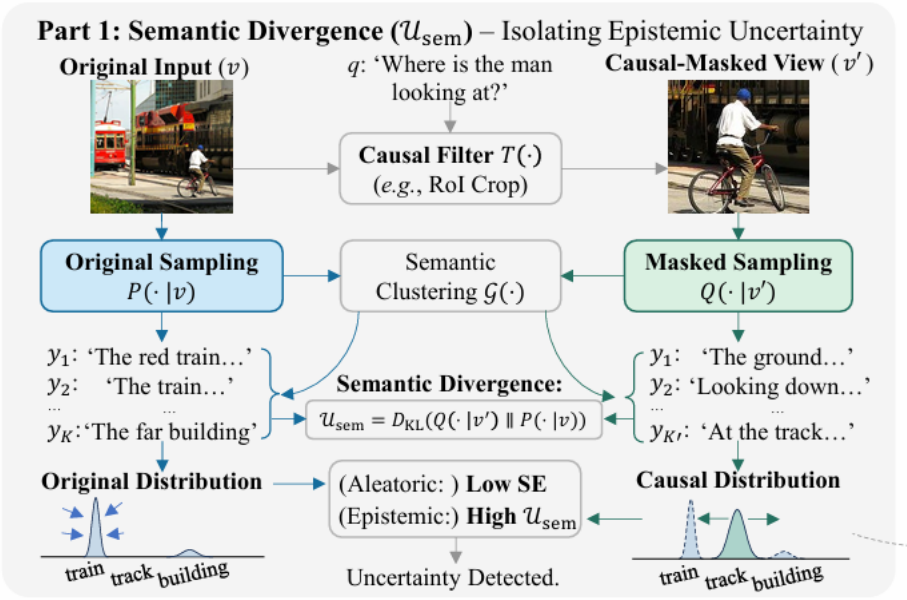
Definition: KL divergence between semantic distributions of the original input v and the causally-filtered view v':

$$\mathcal{U}_{\text{sem}}(v, q) = D_{KL}(P(s|v') \parallel P(s|v)) = \sum_{s \in \mathcal{S}} P(s|v') \log \frac{P(s|v')}{P(s|v)}$$

Interpretation: *Information gain* from removing non-causal features.
Low $\mathcal{U}_{\text{sem}}$: Model is robust (or data is ambiguous but model is honest)
High $\mathcal{U}_{\text{sem}}$: Model relies on background context → epistemic failure
Computation: Sample K' sequences → cluster via entailment $\mathcal{G}(\cdot)$ → aggregate meaning-level probabilities (similar to SE pipeline).

Theoretical Justification
Under an energy-based model of the VLM, the causal (query-relevant) energy cancels in the likelihood ratio, making SD invariant to aleatoric ambiguity.
It estimates half the variance of the model's sensitivity to non-causal features, thereby isolating epistemic uncertainty.

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



Expected Embedding Drift (EED)

Problem: Semantic Divergence requires expensive discrete clustering ($\mathcal{O}((K + K')^2)$ entailment judgments).

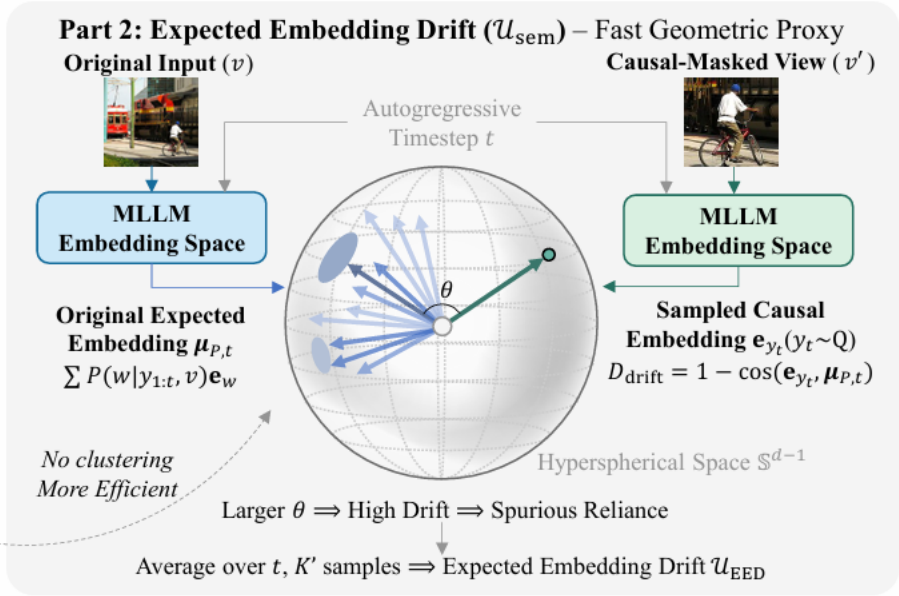
Key observation: Semantic meaning is already embedded in the model's high-dimensional geometry — tokens with similar meanings lie close together on the hypersphere.

EED approach:

1. Compute **expected embedding vector** $\boldsymbol{\mu}_{P,t}$ — the probability-weighted semantic center
2. Measure **cosine drift** of each generated token from this center:

$$D_{\text{drift}}(y_t) = 1 - \cos(\boldsymbol{e}_{y_t}, \boldsymbol{\mu}_{P,t})$$

3. Aggregate across K' samples → scalar $\mathcal{U}_{\text{EED}}$



Theoretical Justification

Under a von Mises-Fisher assumption on the embedding hypersphere, the KL divergence between two semantic distributions is strictly proportional to the cosine distance between their mean directions

$$D_{KL}(Q \parallel P) = \kappa A_d(\kappa) \cdot (1 - \boldsymbol{\mu}_Q^\top \boldsymbol{\mu}_P)$$

EED is an unbiased, clustering-free geometric estimator of Semantic Divergence.

Results — VQA Benchmarks

Key findings (VQAv2 / OKVQA / AdVQA):
CIM (SD) achieves SOTA across all benchmarks and both models
On the challenging **AdVQA** (adversarial VQA):

- **76.8% AUROC** (InternVL), surpassing SE (75.5%) and EigenScore (67.1%)

Key observations (on POPE):

- SD dominates across all splits (~75% AUROC vs. SE's ~68–72%)

EED performs competitively — it does not degrade SD much, while being ~50% faster

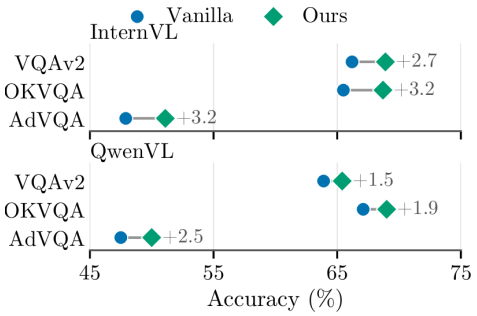
Method	VQAv2	OKVQA	AdVQA	POPE Rand.	POPE Pop.	POPE Adv.
SE	78.7	80.6	75.5	70.6	68.7	72.3
EigenScore	72.6	72.2	67.1	65.5	64.4	63.8
UMPIRE	74.3	72.6	68.2	68.7	68.7	68.2
Ours (SD)	**81.3**	**81.8**	**76.8**	**75.9**	**74.8**	**74.9**
Ours (EED)	80.3	80.9	75.8	73.1	71.0	73.0

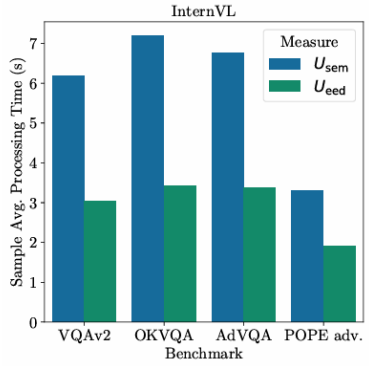
Efficiency & Applications

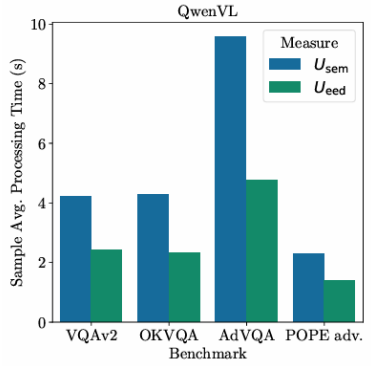
Computational Efficiency (EED):
- **~50% reduction** in processing time vs. SD
- AdVQA with QwenVL: 9.6s → 4.8s per sample
- Eliminates $\mathcal{O}((K + K')^2)$ entailment-based clustering

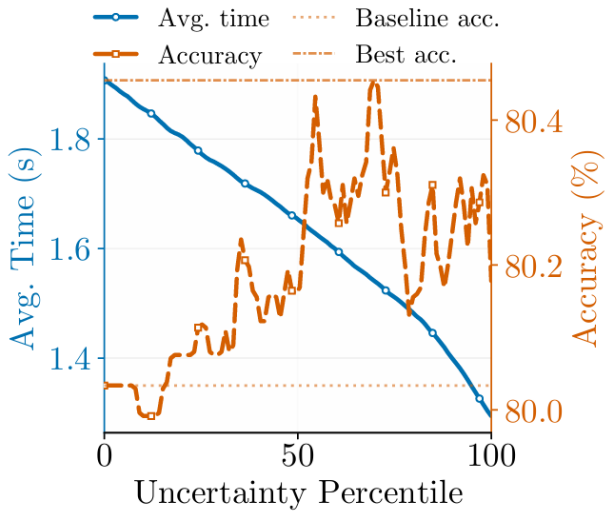
Application 1 — Hallucination Mitigation:
- **Reject-and-refine strategy:** if $\mathcal{U}_{\text{sem}}$ exceeds threshold → re-generate answer
- AdVQA: InternVL improves from 47.9% → **51.1%**; QwenVL from 47.5% → **50.0%**

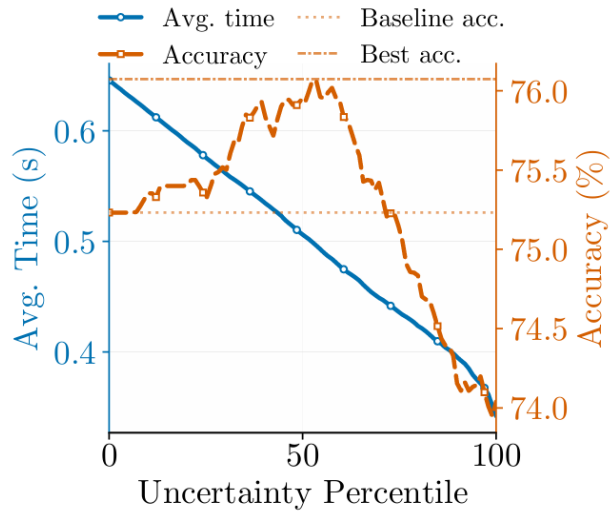
Application 2 — Adaptive-Resolution Inference:
- Low uncertainty → process at low resolution (256 px max)
- Reduces latency (QwenVL: 0.65s → 0.50s) while **improving accuracy** (75.2% → 76.1%)











Conclusion

Problem addressed: Existing MLLM UQ methods are biased toward aleatoric uncertainty, overlooking epistemic uncertainty from spurious correlations.

Our contributions:

1. **Causal-Invariant Masking (CIM)** — a framework that isolates epistemic uncertainty via causal invariance.
2. **Semantic Divergence (SD)** — theoretically grounded metric (converges to variance of sensitivity to non-causal features).
3. **Expected Embedding Drift (EED)** — fast geometric proxy (~50% faster, clustering-free, theoretically equivalent).
4. **State-of-the-art hallucination detection** across diverse benchmarks (76.8% AUROC on AdVQA).
5. **Practical applications** in hallucination mitigation and adaptive inference.

Outcome: CIM enables a prominent path toward trustworthy multimodal reasoning.

Thank You for Your Attention!