Not All Inconsistency is Equal: Decomposing LVLM Uncertainty into Belief Divergence and Belief Conflict

Jie Shi[1], Xiaodong Yue[2,3*], Wei Liu[4], Yufei Chen[4], Feifan Dong[3]
[1] School of Computer Engineering and Science, Shanghai University, Shanghai, China
[2] Institute of Artificial Intelligence, Shanghai University, Shanghai, China
[3] School of Future Technology, Shanghai University, Shanghai, China
[4] School of Computer Science and Technology, Tongji University, Shanghai, China
jieshi@shu.edu.cn, yswantfly@shu.edu.cn, ldachuan@outlook.com, yufeichen@tongji.edu.cn, dongfeifan@shu.edu.cn

Abstract

Uncertainty Quantification (UQ) is critical for detecting hallucinations in black-box Large Vision-Language Models (LVLMs). However, prevailing methods like Discrete Semantic Entropy (DSE) are unreliable, as their scores are primarily dominated by the number of semantic clusters. This renders them incapable of distinguishing between benign semantic ambiguity (varied but coherent responses) and severe belief conflict (contradictory responses). We address this limitation by proposing a novel framework rooted in Dempster-Shafer theory of evidence, built on the premise that not all inconsistency is equal. Our method decomposes uncertainty into two complementary metrics: Belief Divergence, which quantifies ambiguity by measuring the separation between viewpoints, and Belief Conflict, which captures direct logical contradictions. Extensive experiments demonstrate that our framework provides a more reliable measure of uncertainty.

Introduction

Large Vision-Language Models (LVLMs) have demonstrated powerful capabilities across a spectrum of vision and language tasks, becoming foundational tools in numerous real-world applications (Hu et al. 2024; Peng et al. 2023; Cui et al. 2024). A significant portion of these cutting-edge models are deployed as black-box systems, accessible only through APIs without exposing internal parameters. While this broadens access, it presents a critical challenge: ensuring the reliability of models whose inner workings are opaque (Woo et al. 2025). The phenomenon of *hallucination*, defined as the generation of plausible but factually incorrect information, is particularly concerning in this context (Bai et al. 2024; Liu et al. 2024b). Therefore, developing robust methods for Uncertainty Quantification (UQ) for these black-box models is a crucial necessity for responsible AI deployment.

A prevailing strategy for black-box UQ involves generating a diverse answer set via semantic-preserving perturbations and quantifying its *inconsistency* as a proxy for uncertainty (Zhang, Zhang, and Zheng 2024). This is often realized as Discrete Semantic Entropy (DSE) (Farquhar et al. 2024), which clusters responses by meaning, typically using bidirectional entailment predictions from a Natural Language Inference (NLI) model, before calculating entropy from the cluster distribution. However, **as its evaluation is primarily dominated by the number of semantic clusters**, this approach is incapable of distinguishing between two fundamentally different types of inconsistency: a mild form arising from *semantic ambiguity* and a severe form arising from *semantic conflict*.

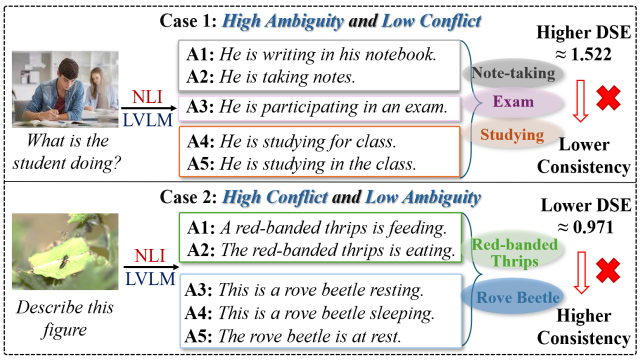

Figure 1: An illustration of the two primary failure modes of Discrete Semantic Entropy (DSE). In the high-ambiguity case (Top), DSE yields a high score (≈ 1.522), incorrectly suggesting high uncertainty. In the high-conflict case (Bottom), it yields a deceptively lower score (≈ 0.971), failing to capture the severe contradiction.

This dual-limitation is vividly illustrated in Figure 1. The benign ambiguity in Case 1 results in a high DSE score (≈ 1.522), while the severe conflict in Case 2 yields a deceptively lower score (≈ 0.971). These counter-intuitive results reveal a fundamental principle that existing metrics overlook: **not all inconsistency is created equal.** A mild form like ambiguity signals a stable exploration of a topic, whereas a severe form like conflict signals a critical failure in reasoning.

To achieve this differentiation, we propose a novel UQ framework rooted in Dempster-Shafer theory (DST) of evidence (Shafer 1992), chosen for its inherent ability to explicitly model ignorance. Crucially, our framework operates in a fully black-box setting. It moves beyond a single, coarse-grained score by decomposing uncertainty

*Corresponding author

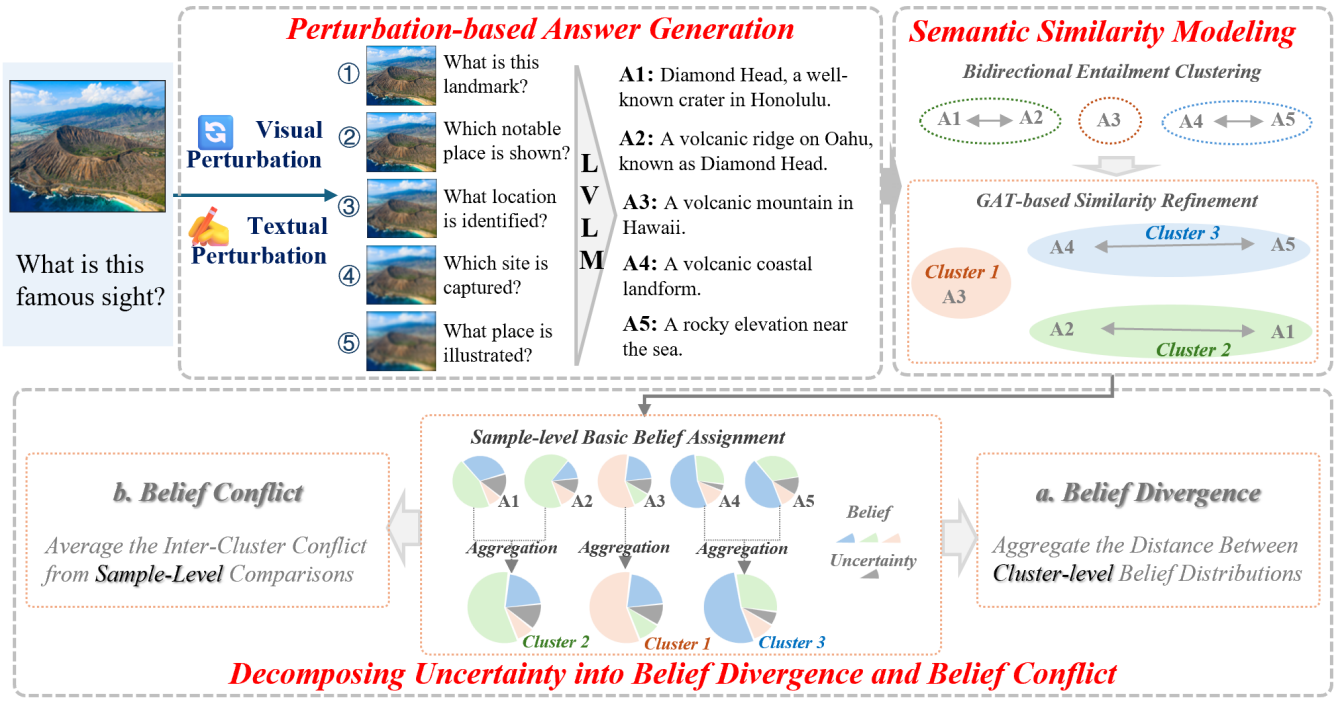

Figure 2: An overview of our proposed framework for decomposing LVLM uncertainty. **(1)** First, we generate a diverse answer set by applying semantic-preserving perturbations to the initial vision-language input. **(2)** Next, the answer set's semantic structure is modeled through NLI-based hard clustering and then refined with a Graph Attention Network (GAT). **(3)** Finally, within our DST-based module, each answer is converted into a Basic Belief Assignment (BBA), which is used to quantify uncertainty via two complementary metrics: Belief Divergence and Belief Conflict.

into two complementary components: *Belief Divergence*, which quantifies semantic ambiguity, and *Belief Conflict*, which captures direct logical contradictions. By integrating these two dimensions, our framework provides a more robust, fine-grained, and interpretable measure of LVLM uncertainty. In summary, our contributions are as follows:

- We are the first to conceptualize and decompose LVLM inconsistency into two distinct forms: benign semantic ambiguity and critical semantic conflict.
- We develop a novel, black-box framework using Dempster-Shafer evidence theory to quantify uncertainty via two complementary metrics: Belief Divergence and Belief Conflict.
- Empirical analyses validate the effectiveness of the proposed method in providing a more reliable measure of uncertainty.

Related Work

A prevailing strategy in UQ of black-box LVLMs is to generate a diverse answer set via semantic-preserving perturbations and then quantify its inconsistency as a proxy for uncertainty. Existing methods primarily operate on these answer sets (Kadavath et al. 2022; Zhang et al. 2024). A foundational approach is to first perform *Semantic Clustering*, typically using a pre-trained NLI model to group semantically equivalent answers. Methods like DSE (Zhang, Zhang, and Zheng 2024) and NumSet (Kuhn, Gal, and Farquhar 2023; Lin, Trivedi, and Sun 2024) then quantify uncertainty based on the distribution or simply the number of these clusters. However, the primary limitation of these methods is that their scores are dominated by the cluster count, ignoring the finer-grained relationships between them. To address this, other methods incorporate pairwise similarity. These include lexical-based approaches like LexSim (Fomicheva et al. 2020), nearest-neighbor inspired techniques like SNNE (Nguyen, Payani, and Mirzasoleiman 2025), and a family of graph-based methods that derive uncertainty from the properties of a semantic similarity graph, such as Deg, SumEigv, and Eccen, proposed by (Lin, Trivedi, and Sun 2024). While these approaches provide a more nuanced view than simple cluster counting, they still treat inconsistency as a monolithic concept, lacking a formal mechanism to distinguish between benign semantic ambiguity and severe semantic conflict. In contrast, Evidence Theory, also known as Dempster-Shafer Theory (DST) (Shafer 1992), offers a dedicated framework for reasoning under uncertainty. It has been successfully applied to diverse tasks, including trustworthy multi-view learning (Han et al. 2022; Liu, Chen, and Yue 2025a,b; Liang et al. 2025a,b; Xu et al. 2024; Liu et al. 2023c, 2022), ensemble learning (Fu et al. 2022; Lv et al. 2021), fine-grained image classification (Xu et al. 2023), patent classification (Zhang et al. 2022), and medical diagnosis (Liu et al.

2025; Liu, Chen, and Yue 2024; Fu et al. 2023).

Method

In this section, we introduce a novel framework rooted in DST that decomposes inconsistency into its complementary parts. The framework unfolds in three stages, as illustrated in Figure 2.

Modeling Answer Set Semantics

Perturbation-based Answer Generation: To assess a model's consistency, we first generate a diverse answer set $\mathcal{A} = \{a_1, a_2, \ldots, a_N\}$ by querying the LVLM with N semantically equivalent but perturbed input pairs, a strategy inspired by prior work (Zhang, Zhang, and Zheng 2024).

For the visual modality, we perturb the original image I into a series of images I_i using a 2D Gaussian Blur, controlling the intensity via the blur radius r_i. Concurrently, for the textual modality, we employ a pre-trained Large Language Model (LLM) to generate semantically equivalent paraphrases T_i of the original text T, controlling diversity by the temperature parameter τ_i. These perturbations are synchronized by intensity to form a set of input pairs, $\Lambda = \{\langle I_i, T_i \rangle\}$, which are then fed to the LVLM to produce the answer set $\mathcal{A}$ for our analysis.

Structured Semantic Modeling: To model the internal semantic structure of the answer set $\mathcal{A}$, we perform an initial hard clustering followed by a graph-based refinement. First, we use a pre-trained NLI model to apply a strict bidirectional entailment test to every answer pair, grouping semantically equivalent answers into a set of N_c clusters, $\{C_k\}_{k=1}^{N_c}$.

While this provides a coarse overview, we refine it using a Graph Attention Network (GAT (Velickovic et al. 2017)) to create a more nuanced similarity measure. We first obtain initial embeddings, $\mathbf{X}$, for all answers using a standard sentence-embedding model. We then encode the clustering result into a graph where edges connect answers belonging to the same cluster. This graph and the embeddings $\mathbf{X}$ are processed by a pre-trained GAT to produce refined, "structure-aware" embeddings. From these, we compute the final similarity matrix $\mathbf{S}$ using a Gaussian kernel function, where each element S_{ij} captures the deep semantic consistency between answers a_i and a_j.

Belief Representation of Answers

After modeling the semantic structure, we move to the formal framework of DST, chosen for its inherent ability to explicitly represent ignorance. Each answer a_i is represented as a Basic Belief Assignment (BBA), denoted m_i.

Our frame of discernment (FOD), $\Theta = \{C_1, \ldots, C_{N_c}\}$, is the set of mutually exclusive semantic clusters. The focal sets for our BBA, m_i, are restricted to the *singleton sets* $\{C_k\}$ and the *universal set* Θ. The belief masses are calculated as follows. First, we calculate a semantic affinity distribution $\mathbf{p}^i = \left[p_1^i, \ldots, p_{N_c}^i\right]$ for answer a_i across all clusters based on the refined similarity matrix $\mathbf{S}$:

$$p_k^i = \tilde{S}_i\left(C_k\right) \Big/ \sum_{l=1}^{N_c} \tilde{S}_i\left(C_l\right), \tag{1}$$

where

$$\tilde{S}_i\left(C_k\right) = \frac{1}{|C_k|} \sum_{a_j \in C_k} S_{ij} \tag{2}$$

is the support score for each answer a_i with respect to each cluster C_k by averaging its similarity to the cluster's members. The ignorance term $m_i\left(\Theta\right)$ is then quantified using the normalized Shannon entropy of this distribution:

$$m_i\left(\Theta\right) = \frac{-\sum_{k=1}^{N_c} p_k^i \log_2 p_k^i}{\log_2 N_c}. \tag{3}$$

The remaining belief mass is then distributed proportionally among the singleton sets:

$$m_i\left(\{C_k\}\right) = \left(1 - m_i\left(\Theta\right)\right) \cdot p_k^i. \tag{4}$$

Quantifying Consistency via Two Metrics

The BBA representation allows us to decompose inconsistency into two complementary components.

Belief Divergence (Measuring Ambiguity): This metric quantifies inconsistency from *semantic ambiguity*. A higher divergence corresponds to a higher degree of ambiguity.

To calculate this, we first establish a centroid belief distribution $\bar{m}_k = \frac{\sum_{i \in C_k} m_i}{|C_k|}$ for each semantic cluster by averaging the BBAs of its members. Our raw inconsistency score for divergence, d, is then the total sum of the pairwise Jousselme distances between every unique cluster centroid:

$$d = \sum_{k=1}^{N_c - 1} \sum_{l=k+1}^{N_c} d_J\left(\bar{m}_k, \bar{m}_l\right), \tag{5}$$

where

$$d_J\left(\bar{m}_k, \bar{m}_l\right) = \sqrt{\frac{1}{2}\left(\bar{m}_k - \bar{m}_l\right)^T \mathbf{D}\left(\bar{m}_k - \bar{m}_l\right)}. \tag{6}$$

Here, $\mathbf{D}$ is the matrix of pairwise Jaccard similarities between focal sets. This summation-based design, rather than an average, ensures that, *similar to DSE, the metric is sensitive to the number of distinct semantic clusters, while the Jousselme distance provides a nuanced measure of their actual separation.* This total distance d is then mapped to a normalized consistency score, $S_{div} = \exp\left(-d\right)$.

Belief Conflict (Measuring Contradiction): This metric quantifies inconsistency from *semantic conflict*. To obtain a global measure of conflict, we employ a hierarchical averaging strategy. The process begins at the sample-level by calculating the Dempster conflict coefficient, κ, between any two answers, a_i and a_j, from different clusters. Since the focal sets in our framework are restricted to singletons, we have a simplified form:

$$\kappa\left(m_i, m_j\right) = \sum_{k \neq l} m_i\left(\{C_k\}\right) m_j\left(\{C_l\}\right). \tag{7}$$

Next, these sample-level values are averaged to find the representative conflict between any two clusters, C_k and C_l:

$$\kappa\left(C_k, C_l\right) = \frac{1}{|C_k|\,|C_l|} \sum_{a_i \in C_k} \sum_{a_j \in C_l} \kappa\left(m_i, m_j\right). \tag{8}$$

Finally, our raw inconsistency score for conflict, $\bar{\kappa}$, is:

$$\bar{\kappa} = \frac{1}{\binom{N_c}{2}} \sum_{k=1}^{N_c-1} \sum_{l=k+1}^{N_c} \kappa\left(C_k, C_l\right). \tag{9}$$

This conflict level is then mapped to a normalized consistency score, $S_{con} = \exp\left(-\bar{\kappa}\right)$.

Final Integrated Consistency Score: While the diagnostic power lies in the individual metrics, a single score is necessary for benchmarking. We combine the two consistency scores into a final score, $S_{overall}$, using a weighted average:

$$S_{overall} = \alpha S_{div} + (1 - \alpha)\, S_{con}, \tag{10}$$

where $\alpha \in [0, 1]$ is a weighting parameter. This score represents the model's overall consistency and is inversely proportional to its uncertainty.

Complexity Analysis

The overall time complexity of our framework is $O\left(N^2\right)$, where N is the number of generated answers. This complexity is comparable to standard methods like DSE-based method ($O\left(N^2\right)$) and is primarily dominated by the pairwise semantic clustering step. The cost of each major module is as follows. 1) The initial **Answer Generation** module has a complexity of $O\left(N\right)$, as it requires N forward passes through the LVLM. The **Structured Semantic Modeling** Module contains the main computational bottleneck with the NLI-based hard clustering requiring $O\left(N^2\right)$ pairwise inferences. The subsequent GAT refinement is an efficient, single forward pass over the constructed graph. Finally, the **Belief Framework Quantification** stage, including the BBA construction and the calculation of our two metrics, also has a complexity of approximately $O\left(N^2\right)$, as its operations depend on pairwise comparisons between the N answers and N_c clusters (where $N_c \leq N$).

Experiments

Experimental Setup

Datasets: We evaluate our method on three diverse benchmarks with various LVLMs. First, **LLaVA-Bench** (Liu et al. 2023a) provides 60 challenging questions designed to test convention, detail, and complex reasoning. Second, **MM-Vet-V2** (Yu et al. 2024), a comprehensive benchmark with 218 questions, systematically evaluates 6 core and 16 combined capabilities, from OCR to chart understanding. Finally, **VQA-RAD** (Lau et al. 2018), a medical visual question answering dataset with over 3500 questions on 315 radiological images, tests domain-specific abilities like identifying abnormalities.

Evaluated LVLMs: We conduct experiments on 13 LVLMs from four distinct model families: **Qwen** (Wang et al. 2024; Bai et al. 2025), **InternVL** (Chen et al. 2024; Zhu et al. 2025), **LLaVA** (Liu et al. 2023b, 2024a), and **DeepSeek** (Lu et al. 2024). While many of these models are open-source, we treat all of them as **black-box** in our experiments, as our proposed method only requires access to their final text outputs.

Implementation Details: For each sample, we first generate a reference answer at a low temperature (0.1) for correctness evaluation. To create the answer set for our uncertainty analysis, we generate N perturbed inputs. **Unless otherwise specified, we set $N = 10$ and the weighting parameter $\alpha = 0.5$ for all main experiments.** Visual perturbations are 2D Gaussian Blurs with radii r_i varying from 0.1 to 0.55 (step 0.05), while textual perturbations are paraphrases from the **Qwen2.5-7B-Instruct** model (Team 2024) with temperatures τ_i in the same range and step size. The LVLM under evaluation then generates an answer for each perturbed input. We deliberately use a low temperature (0.1) for this step, as we empirically found that for our VQA tasks, the input perturbations alone were sufficient to elicit a rich and semantically diverse answer set. Moreover, we use the **jina-embeddings-v3** (Sturua et al. 2024) to obtain initial answer embeddings. The same **Qwen2.5-7B-Instruct** model also serves as the NLI model for the hard clustering step. The GAT-based refinement is implemented using the *GATConv* layer from the PyTorch Geometric library with a single-layer network. To ensure a fair comparison, all baseline methods were evaluated under the exact same experimental settings as our proposed framework. All experiments were conducted on a single NVIDIA A100 40GB GPU.

Evaluation metrics: Following prior work (Nguyen, Payani, and Mirzasoleiman 2025), we adopt two standard metrics: **AUROC**, which measures the ability of an uncertainty score to discriminate between correct and incorrect answers, and **AUARC**, which reflects the practical utility of using the score to reject low-quality answers.

Baselines: We compare our method against eight existing black-box UQ baselines. These include methods based on lexical similarity (**LexSim** (Fomicheva et al. 2020)), semantic clustering and entropy (**DSE** (Zhang, Zhang, and Zheng 2024), **NumSet** (Kuhn, Gal, and Farquhar 2023)), three graph-based methods proposed by (Lin, Trivedi, and Sun 2024) (**SumEigv**, **Deg**, **Eccen**), and other recent approaches (**LUQ** (Zhang et al. 2024), **SNNE** (Nguyen, Payani, and Mirzasoleiman 2025)). For SNNE, we use the variant with the ROUGE similarity function (**SNNE-ROUGE**), as it was demonstrated to be the most effective in the original work.

Experimental Results

Comparison with Baselines: We evaluate our proposed framework against a suite of strong black-box UQ baselines, with the AUROC and AUARC results presented in Table 1 and Table 2, respectively. The results demonstrate the robust performance of our framework, which achieves state-of-the-art results on the vast majority of configurations. Our framework's advantage is particularly pronounced on the more complex and nuanced benchmarks, MM-Vet-V2 and VQA-RAD. On these datasets, our method achieves the highest AUROC and AUARC scores for nearly every model, highlighting its robustness in handling challenging questions that require deep comprehension. On the simpler LLaVA-Bench dataset, our method remains highly competitive, though other specialized methods also perform well, particularly on the AUARC benchmark. We attribute this to

	Qwen				LLaVA				InternVL				DeepSeek
	Qwen2-VL	Qwen2.5-VL			LLaVA-1.5		LLaVA-NeXT		InternVL2	InternVL3			DeepSeek-VL
	7B	7B	32B	72B	7B	13B	7B	13B	8B	8B	9B	14B	7B
LLaVA-Bench													
LUQ	0.6203	0.5520	0.4710	0.5156	0.5882	0.5925	0.5496	0.6605	0.5847	0.6099	0.4746	0.5752	0.6098
NumSet	0.7033	0.5349	0.5831	0.5368	0.5675	0.5854	0.5917	0.6960	**0.6687**	0.5556	0.6154	0.6343	0.6098
LexSim	**0.7692**	**0.7006**	0.6094	**0.6116**	0.6906	**0.6925**	0.6596	0.6662	0.6442	0.5521	0.5843	0.5660	0.5539
SumEigv	0.4249	0.5211	0.4922	0.5033	0.6645	0.6724	0.4714	**0.7017**	0.4458	0.4016	0.4435	0.4502	0.5931
Deg	0.4200	0.5246	0.4978	0.4944	0.6688	0.6724	0.4501	0.6932	0.4392	0.3924	0.4419	0.4549	0.5815
Eccen	0.4811	0.5749	0.4604	0.5346	0.6961	0.6126	0.5172	0.6506	0.4881	0.4520	0.4059	0.4699	0.6014
DSE	0.7039	0.5800	0.6166	0.5926	0.5817	0.6149	0.6156	0.6733	0.6541	0.6389	0.6383	**0.6829**	0.6444
SNNE-ROUGE	0.6947	0.6594	0.5022	0.5458	**0.6993**	0.6460	0.5565	0.6165	0.5622	0.5185	0.5368	0.5706	0.5308
Ours ($S_{overall}$)	0.7106	0.6160	**0.6272**	0.6049	0.6449	0.6553	**0.6622**	0.6989	0.6620	**0.6620**	**0.6604**	0.6736	**0.6605**
MM-Vet-V2													
LUQ	0.5428	0.5734	0.5923	0.5647	0.6758	0.5956	0.6366	0.5686	0.5730	0.5714	0.6147	0.5687	0.6091
NumSet	0.6538	0.6614	0.6316	0.5894	0.7109	0.6812	0.6620	0.6420	0.6640	0.6410	0.6466	0.6195	0.6394
LexSim	0.6491	0.6659	0.5967	0.6094	0.6558	0.6626	0.6417	0.6479	0.6332	0.6178	0.6273	0.6478	0.5452
SumEigv	0.6334	0.6224	0.6077	0.6150	0.6949	0.6310	0.5210	0.5278	0.5511	0.5298	0.5189	0.5612	0.6134
Deg	0.6324	0.6215	0.6026	0.6188	0.6929	0.6350	0.5180	0.5322	0.5539	0.5305	0.5201	0.5582	0.6139
Eccen	0.6430	0.6317	0.6084	0.6344	0.6775	0.6432	0.5362	0.5663	0.5764	0.5550	0.5535	0.5985	0.5941
DSE	0.6708	0.6949	0.6753	0.6345	0.7462	0.7229	0.6807	0.6629	0.6890	0.6876	0.6826	0.6640	0.6589
SNNE-ROUGE	0.6300	0.6747	0.6208	0.6294	0.6575	0.6635	0.6406	0.6206	0.6381	0.6239	0.6483	0.6494	0.5626
Ours ($S_{overall}$)	**0.6727**	**0.7019**	**0.6851**	**0.6453**	**0.7703**	**0.7393**	**0.7033**	**0.6753**	**0.7019**	**0.6954**	**0.6940**	**0.6748**	**0.6627**
VQA-RAD													
LUQ	0.5478	0.5299	0.6354	0.5400	0.5400	0.5144	0.5236	0.5178	0.5404	0.5834	0.5655	0.5605	0.5665
NumSet	0.5898	0.6231	0.6288	0.6277	0.6222	0.6153	0.5725	0.5774	0.6316	0.6561	0.6413	0.6526	0.5888
LexSim	0.5547	0.5493	0.6018	0.5706	0.5814	0.6061	0.5299	0.5252	0.5044	0.5764	0.5043	0.5616	0.5607
SumEigv	0.5395	0.5276	0.5933	0.5643	0.5910	0.5981	0.4791	0.4989	0.5001	0.5482	0.5411	0.5669	0.5708
Deg	0.5396	0.5222	0.5918	0.5645	0.5871	0.5993	0.4754	0.5014	0.4926	0.5506	0.5372	0.5659	0.5697
Eccen	0.5468	0.5245	0.6004	0.5604	0.5785	0.6061	0.4827	0.5100	0.4815	0.5601	0.5416	0.5600	0.5823
DSE	0.6033	0.6431	0.6540	**0.6681**	0.6398	0.6394	0.5811	0.5919	0.6717	0.6896	0.6825	0.6895	0.5963
SNNE-ROUGE	0.5405	0.5495	0.6111	0.5709	0.5778	0.5926	0.5230	0.5218	0.5075	0.5663	0.5111	0.5594	0.5601
Ours ($S_{overall}$)	**0.6171**	**0.6504**	**0.6651**	0.6661	**0.6484**	**0.6450**	**0.5943**	**0.6171**	**0.6756**	**0.6986**	**0.6946**	**0.6986**	**0.5991**

Table 1: AUROC results on all datasets across 13 LVLMs. All methods are evaluated in a black-box setting using $N = 10$ generated answers. Our method uses a balanced weight of $\alpha = 0.5$. Bold text indicates the best performance in each column.

	Qwen				LLaVA				InternVL				DeepSeek
	Qwen2-VL	Qwen2.5-VL			LLaVA-1.5		LLaVA-NeXT		InternVL2	InternVL3			DeepSeek-VL
	7B	7B	32B	72B	7B	13B	7B	13B	8B	8B	9B	14B	7B
LLaVA-Bench													
LUQ	0.4022	0.4606	0.4742	0.4521	0.1828	0.2613	0.3035	0.3001	0.3137	0.4675	0.2337	0.4185	0.3804
NumSet	0.5082	0.4782	0.6227	0.5626	0.1986	0.2999	0.2639	**0.4246**	0.3806	0.4257	0.2131	0.4551	0.4185
LexSim	**0.5716**	**0.5673**	0.6056	0.5243	0.2624	**0.3630**	0.3247	0.4041	**0.4187**	0.4532	**0.2646**	0.4084	0.3496
SumEigv	0.3658	0.4856	0.5528	0.4741	0.2374	0.3576	0.2087	0.4100	0.2995	0.3572	0.2085	0.3958	0.3203
Deg	0.3616	0.4879	0.5569	0.4650	0.2374	0.3572	0.2013	0.4028	0.2969	0.3313	0.2059	0.3979	0.3150
Eccen	0.3942	0.5106	0.5471	0.4817	**0.2689**	0.3278	0.2345	0.4007	0.3099	0.3995	0.1789	0.3966	0.3814
DSE	0.4937	0.5048	0.6443	0.6362	0.2050	0.3066	0.3114	0.4035	0.3738	0.4873	0.2267	**0.4870**	0.4282
SNNE-ROUGE	0.5229	0.5424	0.5746	0.4905	0.2599	0.3442	0.2736	0.3843	0.3602	0.4296	0.2478	0.4190	0.3492
Ours ($S_{overall}$)	0.4970	0.5176	**0.6473**	**0.6396**	0.2214	0.3187	**0.3319**	0.4133	0.3781	**0.4982**	0.2330	0.4840	**0.4318**
MM-Vet-V2													
LUQ	0.3560	0.4145	0.5293	0.5046	0.3016	0.2294	0.3012	0.2734	0.3730	0.4875	0.4417	0.4427	0.3543
NumSet	0.4779	0.5330	0.5506	0.6316	0.3064	0.2919	0.3013	0.3159	0.4276	0.5714	0.4960	**0.5712**	**0.3928**
LexSim	0.4766	0.5364	0.5466	0.7388	0.3104	0.2974	0.3023	0.3265	0.4232	0.5574	0.4987	0.5385	0.3210
SumEigv	0.4700	0.5046	0.5528	0.7603	0.3271	0.2831	0.2429	0.2619	0.3840	0.5018	0.4403	0.4971	0.3640
Deg	0.4692	0.5048	0.5492	0.7621	0.3265	0.2845	0.2413	0.2625	0.3852	0.5009	0.4394	0.4935	0.3642
Eccen	0.4681	0.5034	0.5527	**0.7825**	0.3074	0.2825	0.2534	0.2851	0.4042	0.5138	0.4595	0.5098	0.3539
DSE	**0.4781**	**0.5832**	**0.6271**	0.5396	0.3477	0.3192	0.3423	0.3317	0.4539	0.5562	**0.5498**	0.5493	0.3591
SNNE-ROUGE	0.4727	0.5403	0.5658	0.7620	0.2953	0.2863	0.3091	0.3174	0.4346	0.5687	0.5167	0.5509	0.3347
Ours ($S_{overall}$)	0.4227	0.5369	0.5623	0.5576	**0.3495**	**0.3295**	**0.3658**	**0.3574**	**0.4850**	**0.5860**	0.5353	0.5446	**0.3928**
VQA-RAD													
LUQ	0.5019	0.4743	0.7004	0.6005	0.4034	0.3468	0.3072	0.3102	0.4604	0.5933	0.5582	0.5512	0.4031
NumSet	0.5229	0.5740	0.6911	0.6532	0.4666	0.4346	0.3407	0.3589	0.5803	0.6611	0.6416	0.6551	0.3978
LexSim	0.5279	0.5159	0.6739	0.6522	0.4629	0.4272	0.2959	0.3229	0.4754	0.6165	0.5521	0.5961	0.4020
SumEigv	0.5105	0.5115	0.6842	0.6492	0.4469	0.4191	0.2701	0.3190	0.4809	0.6082	0.5781	0.6036	0.4048
Deg	0.5107	0.5091	0.6830	0.6487	0.4456	0.4198	0.2671	0.3192	0.4756	0.6084	0.5752	0.6012	0.4045
Eccen	0.5393	0.5060	0.6849	0.6515	0.4497	0.4425	0.2751	0.3235	0.4676	0.6185	0.5840	0.5961	0.4031
DSE	0.5575	0.5644	0.7006	0.6737	**0.4873**	**0.4667**	0.3510	0.3735	0.5976	0.6677	0.5839	0.6563	0.4151
SNNE-ROUGE	0.5433	0.5208	0.6795	0.6523	0.4549	0.4161	0.2945	0.3238	0.4832	0.6092	0.5545	0.5923	0.3994
Ours ($S_{overall}$)	**0.5584**	**0.5872**	**0.7103**	**0.6919**	0.4865	0.4529	**0.3554**	**0.3915**	**0.6021**	**0.6707**	**0.6606**	**0.6695**	**0.4227**

Table 2: AUARC results on all datasets across 13 LVLMs. All experimental settings are identical to those in Table 1. Bold text indicates the best performance in each column.

GAT	S_{con}	S_{div}	Qwen2-VL 7B	Qwen2.5-VL 7B	Qwen2.5-VL 32B	Qwen2.5-VL 72B	LLaVA-1.5 7B	LLaVA-1.5 13B	LLaVA-NeXT 7B	LLaVA-NeXT 13B	InternVL2 8B	InternVL3 8B	InternVL3 9B	InternVL3 14B	DeepSeek-VL 7B
LLaVA-Bench															
	✓		0.6899	0.5989	0.6261	0.6004	0.5839	0.5916	0.6504	0.6875	0.6382	0.6574	0.6358	0.6748	0.6438
		✓	0.6618	0.5897	0.6250	0.6015	0.6557	0.5885	0.6370	0.6704	0.6224	0.6574	0.6358	0.6655	0.6284
	✓	✓	0.6716	0.6034	0.6261	0.5993	0.6340	0.5978	0.6370	0.6804	0.6276	0.6609	0.6375	0.6690	0.6322
✓	✓		0.6752	0.6000	0.6217	**0.6093**	0.5904	0.6294	**0.6711**	0.6558	0.6422	0.6605	**0.6817**	**0.6840**	0.6476
✓		✓	0.7069	0.5989	0.6261	0.6004	**0.6840**	0.6397	0.6607	0.6804	0.6580	0.6609	0.6473	0.6735	0.6463
✓	✓	✓	**0.7106**	**0.6160**	**0.6272**	0.6049	0.6449	**0.6553**	0.6622	**0.6989**	**0.6620**	**0.6620**	0.6604	0.6736	**0.6605**
MM-Vet-V2															
	✓		0.6621	0.6947	0.6798	0.6394	0.7430	0.7149	0.6872	0.6622	0.6859	0.6891	0.6863	0.6669	0.6558
		✓	0.6517	0.6984	0.6805	0.6414	0.7391	0.6961	0.6958	0.6607	0.6816	0.6934	0.6873	0.6743	0.6622
	✓	✓	0.6572	0.6956	0.6813	0.6410	0.7412	0.7024	0.6961	0.6623	0.6842	0.6898	0.6876	0.6703	0.6626
✓	✓		0.6397	0.7007	0.6804	0.6426	0.6971	0.6607	0.6591	0.6205	0.6474	0.6879	0.6702	0.6689	0.6261
✓		✓	0.6683	0.6987	0.6842	0.6441	0.7682	0.7365	0.6903	0.6695	0.6942	0.6950	0.6884	0.6731	0.6596
✓	✓	✓	**0.6727**	**0.7019**	**0.6851**	**0.6453**	**0.7703**	**0.7393**	**0.7033**	**0.6753**	**0.7019**	**0.6954**	**0.6940**	**0.6748**	**0.6627**
VQA-RAD															
	✓		0.6130	0.6469	**0.6651**	0.6629	0.6403	0.6324	0.5862	0.5974	**0.6811**	0.6955	0.6910	0.6912	0.5970
		✓	0.5984	0.6424	0.6627	0.6574	0.6267	0.6000	0.5904	0.6118	0.6701	0.6941	0.6888	0.6867	0.5864
	✓	✓	0.6047	0.6460	0.6643	0.6613	0.6322	0.6086	0.5886	0.6081	0.6633	0.6960	0.6909	0.6908	0.5912
✓	✓		0.5725	0.6310	0.6461	0.6479	0.6221	0.5922	0.5853	0.5951	0.6286	0.6728	0.6621	0.6797	0.5732
✓		✓	0.6170	**0.6505**	0.6648	0.6658	0.6446	0.6399	0.5896	0.6160	0.6763	0.6972	0.6916	0.6964	0.5990
✓	✓	✓	**0.6171**	0.6504	**0.6651**	**0.6661**	**0.6484**	**0.6450**	**0.5943**	**0.6171**	0.6756	**0.6986**	**0.6946**	**0.6986**	**0.5991**

Table 3: Ablation study of our framework's components, reported in AUROC across all datasets. Bold text indicates the best-performing configuration for each LVLM. Experimental settings are identical to those in Table 1.

the fact that the answer sets generated from LLaVA-Bench's simpler questions are less semantically diverse, making the full power of our decomposition less critical. **Most importantly, the primary contribution of our framework is not just the superior performance of the final integrated score, but its diagnostic power.** While the overall score is necessary for benchmarking, the true value lies in the decomposition of uncertainty. Our framework, by providing two distinct metrics for ambiguity and contradiction, offers a level of interpretability that single-score methods cannot.

Ablation Study: We conduct an ablation study to validate the contribution of each component, with AUROC results shown in Table 3. The analysis provides strong empirical evidence for our core thesis that decomposing inconsistency is a more effective approach. The full framework, which integrates S_{div} and S_{con} and utilizes our GAT-based refinement, almost consistently achieves the best performance.

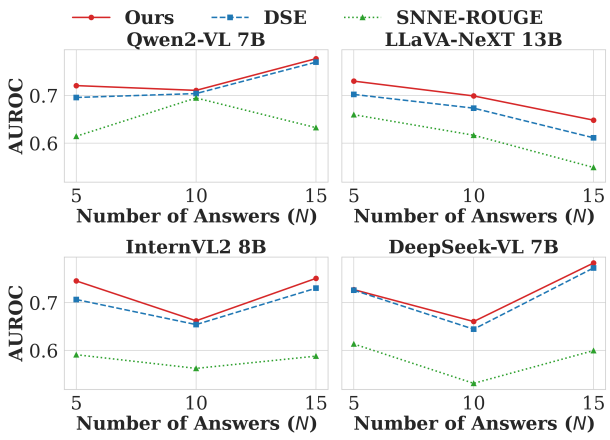

Figure 3: The effect of N on AUROC for our method and key baselines on the LLaVA-Bench dataset.

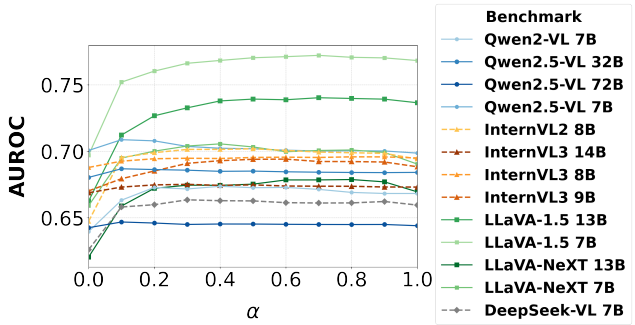

Figure 4: The effect of α on AUROC for all 13 LVLMs on the MM-Vet-V2 dataset, using $N = 10$ answers.

Parameter Analysis of N: Here, we analyze the sensitivity of our framework to the number of generated answers on the LLaVA-Bench dataset across $N \in \{5, 10, 15\}$. We select one model from each of the four families under evaluation (Qwen2-VL, LLaVA-NeXT, InternVL2, and DeepSeek-VL) and compare our method against two key baselines, DSE and SNNE-ROUGE.

The results, shown in Figure 3, reveal a complex and non-monotonic relationship between performance and sample size. Our method's AUROC on several models shows a notable dip at $N = 10$ before surging to its highest point at $N = 15$. This suggests an intricate interplay between the diversity of the answer set and our decomposition-based approach. Despite the strong performance at $N = 15$ on this dataset, we selected $N = 10$ as the default setting for our main experiments to maintain a pragmatic balance between performance and the significant computational cost required across all models and datasets.

Parameter Analysis of α: To analyze the relative contributions of two metrics, we vary the weighting parameter α from 0 (purely Conflict-based) to 1 (purely Divergence-

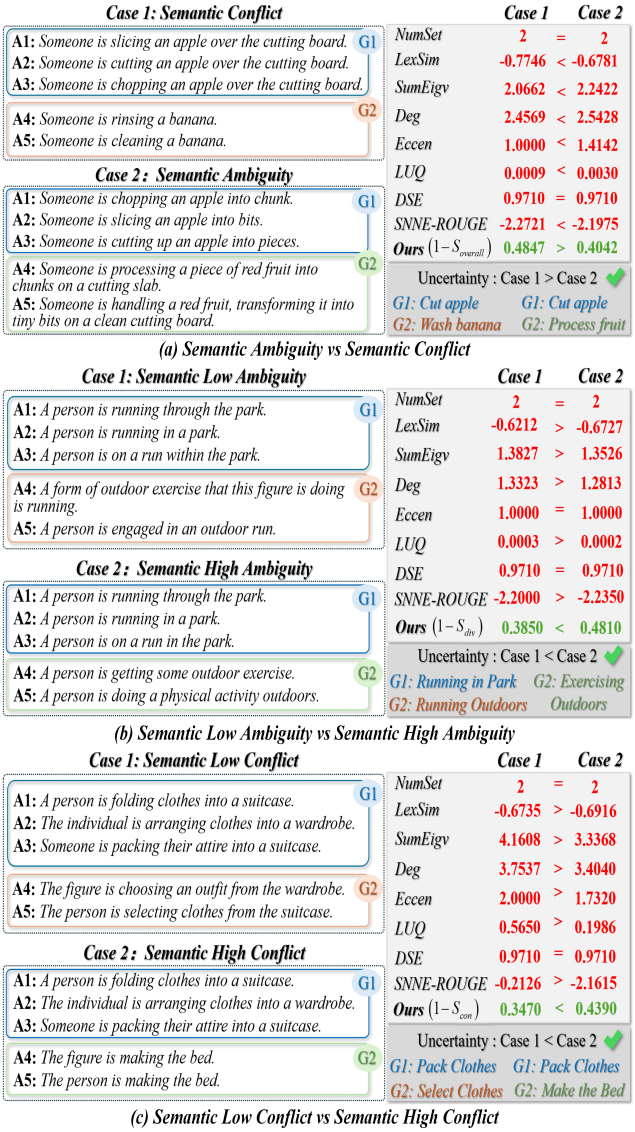

Figure 5: Illustrative experiments validating our decomposition-based approach. **(a)** A targeted experiment demonstrating that baseline methods fail to distinguish between semantic ambiguity and conflict, proving the necessity of decomposition. **(b)** Validation of our Belief Divergence metric $(1 - S_{div})$ on a controlled ambiguity gradient. **(c)** Validation of our Belief Conflict metric $(1 - S_{con})$ on a controlled conflict gradient.

based) on MM-Vet-V2 benchmark. The result, shown in Figure 4, provides strong empirical evidence for our core thesis that a decomposed approach is superior. For the vast majority of models, performance peaks when $0 < \alpha < 1$, demonstrating a clear synergistic effect where the combination of both metrics consistently outperforms using either in isolation. Interestingly, we also observe that Belief Divergence ($\alpha = 1$) often serves as a stronger individual baseline than Belief Conflict ($\alpha = 0$), suggesting our perturbation strategy primarily elicits semantic ambiguity. Given the robust performance of a balanced combination, we select $\alpha = 0.5$ as a general-purpose setting for all our main experiments.

Discussion

The Necessity of Decomposing Uncertainty

To demonstrate why decomposing inconsistency is essential, we constructed two targeted scenarios shown in Figure 5(a): a Semantic Conflict case containing a direct factual contradiction (e.g., "Cut apple" vs. "Wash banana"), and a Semantic Ambiguity case containing a benign difference in descriptive granularity (e.g., "Cut apple" vs. "Process fruit"). The results of this illustrative experiment are stark: **every baseline method incorrectly assigns a higher uncertainty score to the benign ambiguity case than to the severe conflict case.** This proves their fundamental inability to distinguish between different types of inconsistency and validates our core thesis that *not all inconsistency is equal*. In stark contrast, our framework's uncertainty score $(1 - S_{overall})$, even with a balanced $\alpha = 0.5$, correctly identifies the semantic conflict as the more severe and uncertain scenario. This is possible only because our Belief Conflict and Belief Divergence metrics can assess these two phenomena independently.

Independent Validation of Decomposed Metrics

To validate that our two metrics independently measure the phenomena they are designed for, we constructed two pairs of targeted, illustrative cases. First, to test the ability of our Belief Divergence metric to measure semantic ambiguity, we created a Low Ambiguity case and a High Ambiguity case. As shown in Figure 5(b), our uncertainty score $(1 - S_{div})$ correctly and intuitively assigns a lower score to the Low Ambiguity case (0.3850) than to the High Ambiguity case (0.4810), demonstrating its sensitivity to the degree of semantic separation. Second, to test the ability of our Belief Conflict metric to measure semantic conflict, we created a Low Conflict case and a High Conflict case with mutually exclusive answers. As shown in Figure 5(c), our uncertainty score $(1 - S_{con})$ behaves as expected, assigning a significantly lower score to the Low Conflict case (0.3470) compared to the High Conflict case (0.4390), confirming its ability to effectively detect direct logical contradictions.

Conclusion

In this work, we address a fundamental limitation of existing black-box UQ methods for LVLMs: their inability to distinguish between benign semantic ambiguity and severe semantic conflict. We propose a novel framework, rooted in DST, that operationalizes the thesis that *not all inconsistency is equal*. Our approach successfully decomposes uncertainty into two distinct and complementary metrics: *Belief Divergence*, to quantify ambiguity, and *Belief Conflict*, to capture direct logical contradictions. Extensive experiments, including targeted illustrative cases and evaluations on 13 LVLMs across three diverse benchmarks, demonstrate that our framework provides a more robust, fine-grained, and reliable measure of uncertainty than existing methods.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 62476165, 62472315, 62406182), and Fundamental Research Program of Shanxi Province (Serial No. 202403021212176), and Science and Technology Innovation Project for Higher Education Institutions of Shanxi Province (Serial No. 2024L004).

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Bai, Z.; Wang, P.; Xiao, T.; He, T.; Han, Z.; Zhang, Z.; and Shou, M. Z. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24185–24198.

Cui, C.; Ma, Y.; Cao, X.; Ye, W.; Zhou, Y.; Liang, K.; Chen, J.; Lu, J.; Yang, Z.; Liao, K.-D.; et al. 2024. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 958–979.

Farquhar, S.; Kossen, J.; Kuhn, L.; and Gal, Y. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017): 625–630.

Fomicheva, M.; Sun, S.; Yankovskaya, L.; Blain, F.; Guzmán, F.; Fishel, M.; Aletras, N.; Chaudhary, V.; and Specia, L. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8: 539–555.

Fu, H.; Yue, X.; Liu, W.; and Denoeux, T. 2022. Stable clustering ensemble based on evidence theory. In *2022 IEEE International Conference on Image Processing (ICIP)*, 2046–2050. IEEE.

Fu, W.; Chen, Y.; Liu, W.; Yue, X.; and Ma, C. 2023. Evidence reconciled neural network for out-of-distribution detection in medical images. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, 305–315. Springer.

Han, Z.; Zhang, C.; Fu, H.; and Zhou, J. T. 2022. Trusted multi-view classification with dynamic evidential fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(2): 2551–2566.

Hu, Y.; Li, T.; Lu, Q.; Shao, W.; He, J.; Qiao, Y.; and Luo, P. 2024. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22170–22183.

Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Hatfield-Dodds, Z.; DasSarma, N.; Tran-Johnson, E.; et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Kuhn, L.; Gal, Y.; and Farquhar, S. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Lau, J. J.; Gayen, S.; Ben Abacha, A.; and Demner-Fushman, D. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1): 1–10.

Liang, X.; Fu, P.; Qian, Y.; Guo, Q.; and Liu, G. 2025a. Trusted multi-view classification via evolutionary multi-view fusion. In *Proceedings of the International Conference on Learning Representations*, 1–14.

Liang, X.; Wang, S.; Qian, Y.; Guo, Q.; Du, L.; Jiang, B.; Luo, T.; and Li, F. 2025b. Trusted multi-view classification with expert knowledge constraints. In *Proceedings of the International Conference on Machine Learning*, volume 267, 37409–37426.

Lin, Z.; Trivedi, S.; and Sun, J. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual instruction tuning. In *Advances in Neural Information Processing Systems*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36: 34892–34916.

Liu, H.; Xue, W.; Chen, Y.; Chen, D.; Zhao, X.; Wang, K.; Hou, L.; Li, R.; and Peng, W. 2024b. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Liu, W.; Chen, Y.; and Yue, X. 2024. Building trust in decision with conformalized multi-view deep classification. In *Proceedings of the ACM International Conference on Multimedia*, 7278–7287.

Liu, W.; Chen, Y.; and Yue, X. 2025a. Enhancing multi-view classification reliability with adaptive rejection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 18969–18977.

Liu, W.; Chen, Y.; and Yue, X. 2025b. Enhancing testing-time robustness for trusted multi-view classification in the wild. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 15508–15517.

Liu, W.; Chen, Y.; Yue, X.; Zhang, C.; and Xie, S. 2023c. Safe multi-view deep classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 8870–8878.

Liu, W.; Chen, Y.; Yue, X.; Zhang, C.; and Xie, S. 2025. Enhancing reliability in medical image classification of imperfect views. *IEEE Transactions on Circuits and Systems for Video Technology*.

Liu, W.; Yue, X.; Chen, Y.; and Denoeux, T. 2022. Trusted multi-view deep learning with opinion aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 7585–7593.

Lu, H.; Liu, W.; Zhang, B.; Wang, B.; Dong, K.; Liu, B.; Sun, J.; Ren, T.; Li, Z.; Yang, H.; et al. 2024. Deepseek-vl: Towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Lv, Y.; Zhang, B.; Yue, X.; Xu, Z.; and Liu, W. 2021. Ensemble of adapters for transfer learning based on evidence theory. In *International Conference on Belief Functions*, 66–75. Springer.

Nguyen, D.; Payani, A.; and Mirzasoleiman, B. 2025. Beyond semantic entropy: Boosting LLM uncertainty quantification with pairwise semantic similarity. *arXiv preprint arXiv:2506.00245*.

Peng, Z.; Wang, W.; Dong, L.; Hao, Y.; Huang, S.; Ma, S.; and Wei, F. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Shafer, G. 1992. Dempster-shafer theory. *Encyclopedia of Artificial Intelligence*, 1: 330–331.

Sturua, S.; Mohr, I.; Akram, M. K.; Günther, M.; Wang, B.; Krimmel, M.; Wang, F.; Mastrapas, G.; Koukounas, A.; Koukounas, A.; Wang, N.; and Xiao, H. 2024. jina-embeddings-v3: Multilingual embeddings with task LoRA. arXiv:2409.10173.

Team, Q. 2024. Qwen2.5: A Party of Foundation Models.

Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y.; et al. 2017. Graph attention networks. *stat*, 1050(20): 10–48550.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Woo, S.; Zhou, K.; Zhou, Y.; Wang, S.; Guan, S.; Ding, H.; and Cheong, L. L. 2025. Black-Box visual prompt engineering for mitigating object hallucination in large vision language models. In *Proceedings of the Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, 529–538.

Xu, C.; Si, J.; Guan, Z.; Zhao, W.; Wu, Y.; and Gao, X. 2024. Reliable conflictive multi-view learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 16129–16137.

Xu, Z.; Yue, X.; Lv, Y.; Liu, W.; and Li, Z. 2023. Trusted fine-grained image classification through hierarchical evidence fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 10657–10665.

Yu, W.; Yang, Z.; Ren, L.; Li, L.; Wang, J.; Lin, K.; Lin, C.-C.; Liu, Z.; Wang, L.; and Wang, X. 2024. Mm-vet v2: A challenging benchmark to evaluate large multimodal models for integrated capabilities. *arXiv preprint arXiv:2408.00765*.

Zhang, C.; Liu, F.; Basaldella, M.; and Collier, N. 2024. Luq: Long-text uncertainty quantification for llms. *arXiv preprint arXiv:2403.20279*.

Zhang, L.; Liu, W.; Chen, Y.; and Yue, X. 2022. Reliable multi-view deep patent classification. *Mathematics*, 10(23): 4545.

Zhang, R.; Zhang, H.; and Zheng, Z. 2024. Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. *arXiv preprint arXiv:2411.11919*.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.