Do Vision Language Models Recognize Visual Ambiguity?

Ta Duc Huy[1], Trang Nguyen[1]
Townim Chowdhury[1], Ankit Yadav[1], Minh-Son To[2]
Zhibin Liao[1], Johan W.Verjans[1], and Vu Minh Hieu Phan[1]

[1] Australian Institute for Machine Learning, Adelaide University, Australia
[2] Flinders University, Australia
tdh512194@gmail.com

Abstract. Vision-language models can produce confident answers on visually ambiguous inputs, resulting in biased predictions. Common entropy-based methods, such as Semantic Entropy (SE), rely on output diversity. Yet our analysis shows that *overconfident visual embeddings suppress output diversity* under stochastic decoding, causing SE to underestimate uncertainty in such cases. Recent methods instead probe output diversity through input perturbations, including textual paraphrasing or joint text-image perturbations, and show improved performance. We study these approaches and reveals that the resulting variability is often dominated by textual changes rather than visual evidence, causing uncertainty estimates to reflect *prompt sensitivity* rather than visual ambiguity. We therefore propose **Visual Semantic Entropy (VSE)**, which perturbs *only the image* to probe nearby visual variations while keeping the text query fixed. VSE measures uncertainty by clustering generated answers into semantic prototypes and computing the mass-weighted dispersion among them. Extensive evaluation across five modern vision-language models and five diverse VQA benchmarks demonstrates that VSE effectively captures visual ambiguity, establishing a new state-of-the-art for VLM uncertainty estimation. Code is available: https://github.com/tadeephuy/visual-semantic-entropy

Keywords: vision-language model · visual semantic entropy

1 Introduction

Vision-language models (VLMs) excel at visual question answering (VQA), a setting in which uncertainty can arise from ambiguity in the visual evidence. In this work, we study uncertainty estimation for VLMs in VQA. Reliable uncertainty estimates are essential for detecting unreliable predictions and enabling trustworthy deployment of VLMs. Current approaches to uncertainty estimation (UE) in VLMs include *verbalized* methods that prompt models to report confidence [5, 19, 21, 24], *logit-based* methods derived from output probabilities or token entropy [7, 10], and *consistency-based* methods that measure agreement

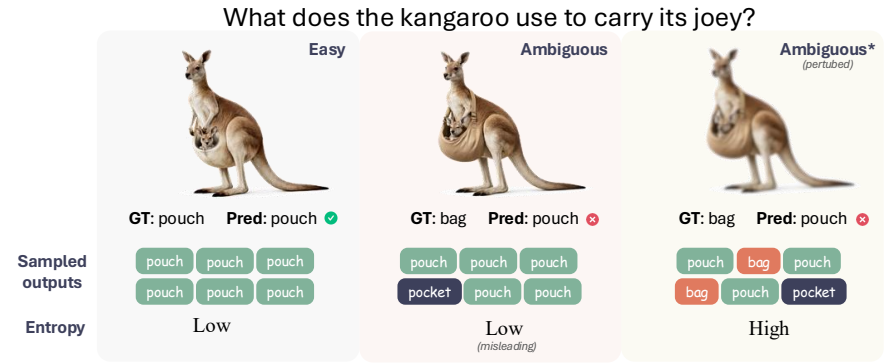

Fig. 1: VLMs can be confidently wrong on ambiguous images. Sampling-based methods such as Semantic Entropy estimate uncertainty by repeatedly sampling model outputs and computing the entropy of the resulting answer distribution, where low entropy indicates high certainty. *Left (Easy):* For a visually clear image, the model correctly predicts "pouch" and repeated sampling consistently outputs "pouch", producing low entropy that appropriately reflects high certainty. *Middle (Ambiguous):* When the image is visually biased, the ground truth is "bag" but the model predicts "pouch". Despite this ambiguity, repeated sampling still yields only "pouch" resulting in low entropy, which is **misleading**. *Right (Ambiguous, perturbed):* After perturbing the image to probe alternative local views, sampling produces diverse answers such as "pouch", "bag", and "pocket", increasing entropy and exposing visual uncertainty.

across sampled generations [8, 13, 27]. Another widely used class is *entropy-based* methods, which estimate uncertainty from the semantic distribution of sampled answers, such as Semantic Entropy (SE) and its variants [3, 15, 17, 26].

In this paper, we set out to analyze existing UE methods and identify 3 limitations that motivate our approach: **(1)** Overconfident visual representations can produce highly consistent outputs under sampling, causing entropy-based methods such as SE [3] to underestimate uncertainty. **(2)** Wording variations among semantically equivalent answers can inflate pairwise semantic distances, leading methods such as SNNE [15] to overestimate uncertainty. **(3)** Textual perturbations can dominate output variability, causing methods that rely on text paraphrasing or joint text-image perturbations (e.g., VL-Uncertainty [26] and C&U [8]) to reflect prompt sensitivity rather than visual ambiguity.

First, Semantic Entropy [3] estimates uncertainty by sampling multiple outputs under stochastic decoding and computing the entropy of the answer distribution, where diverse answers correspond to high entropy and consistent answers to low entropy. While effective, this approach assumes that epistemic uncertainty manifests through decoding randomness. In VLMs, however, uncertainty can originate from the visual input which SE does not effectively capture (Fig. 1). We show that visually biased images can induce visual embeddings that lead to highly confident predictions (Fig. 3). Under such conditions, repeated stochastic decoding produces little output variation, rendering SE ineffective (Sec. 3.2).

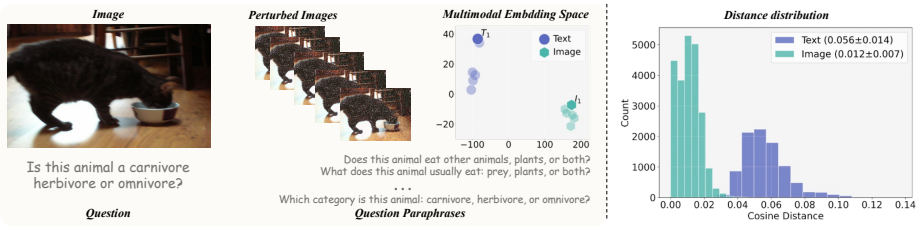

Fig. 2: Text perturbations induce large semantic shifts. *Left*: Given an input image and question, we generate perturbed images and textual paraphrases. In the multimodal embedding space, image perturbations form a tight local cluster around the original representation (gray dashed circle), while text paraphrases produce larger semantic shifts. *Right*: Cosine distance distributions between perturbed samples and the original representation, measured on AOKVQA dataset [20] using Qwen3-VL-Embedding-8B [9]. Text perturbations lead to substantially larger deviations than image perturbations (mean distance 0.056 ± 0.014 vs. 0.012 ± 0.007).

Second, given the resulting answer variants, SE treats answers as categorical labels and estimates uncertainty by counting their frequencies, ignoring semantic distances between categories. Semantic Nearest Neighbor Entropy (SNNE) [15] addresses this by incorporating semantic distances through pairwise aggregation. However, SNNE operates directly on text outputs. Because *language is discrete*, answers that share the same meaning but differ in wording may appear separated in embedding space. As a result, distance-based aggregation can *mistake wording variation for semantic disagreement* and inflate uncertainty (Sec. 3.4).

Third, recent work [8,26] explores input-space perturbations through textual paraphrasing or joint text-image perturbations. As illustrated in Fig. 2, due to *prompt sensitivity*, text paraphrases can induce non-local, larger semantic shifts than visual perturbations, which typically produce only small, local changes. While such perturbations increase output diversity, the *resulting variability is often dominated by textual changes* rather than visual evidence, with clustering is largely text-driven (Fig. 5, Sec. 3.3).

To address these limitations, we propose **Visual Semantic Entropy (VSE)**, which perturbs the image to induce alternative answers and measures uncertainty through dispersion among semantic prototypes of the generated responses. *First*, VSE only perturbs the input image, leaving the text query unchanged, to probe prediction instability that sampling alone fails to capture. *Second*, VSE clusters semantically similar answers into *prototypes* so that wording variations do not inflate uncertainty. Uncertainty is measured as the frequency-weighted pairwise distance among prototypes. *Finally*, by perturbing only the image, VSE avoids modality confounding from textual perturbations and yields uncertainty that better reflects visual ambiguity. Extensive experiments show that VSE is particularly effective on visually adversarial datasets, outperforming methods based on text perturbation and joint image-text perturbations. In summary, our contributions are:

1. Our analysis reveals three failure modes of current VLMs uncertainty estimators: **(1)** suppressed output variation from overconfident visual embeddings; **(2)** inflated distances from wording variation; and **(3)** textual perturbations dominate output variability, causing uncertainty to reflect prompt sensitivity rather than visual ambiguity.
2. We introduce **Visual Semantic Entropy (VSE)**, which perturbs the image, clusters semantically similar answers, and quantifies uncertainty using weighted pairwise distances between prototype embeddings.
3. We present an extensive benchmark across 5 VLMs and 5 datasets, evaluating multiple uncertainty estimation methods and showing that VSE consistently yields more reliable uncertainty estimates.

2 Related Work

Uncertainty Estimation in VLMs aims to quantify the reliability of model predictions, which is particularly important for visual question answering (VQA), where ambiguous visual evidence can lead to unreliable answers.

Verbalized methods prompt models to report confidence [5,19,21,24], while **logit-based** methods derive uncertainty from output probabilities [7, 10]. Both do not explicitly account for visual ambiguity.

Consistency-based approaches, including SelfCheckGPT [13] and C&U [8], estimate uncertainty by measuring agreement between sampled responses and the original answer. C&U additionally generates question paraphrases before probing agreement. However, when visual embeddings are confident, stochastic decoding can yield highly consistent outputs, causing these methods to underestimate uncertainty.

Entropy-based approaches estimate uncertainty from the semantic distribution of sampled answers. Semantic Entropy (SE) [3] computes entropy over outputs sampled under stochastic decoding. Later methods improve semantic representation, including SNNE [15] and KLE [17], which aggregates pairwise distances across sampled answers. However, similar to consistency-based methods, these approaches rely on diversity in generated text under stochastic decoding and may fail to capture visual uncertainty when visual embeddings are confident, as shown in our analysis (Sec. 3.2). Moreover, pairwise distance aggregation can inflate uncertainty due to wording variations across answers(Sec. 3.4).

Input Perturbation. Recent work probes uncertainty through input perturbations. C&U [8] perturbs the question, while VL-Uncertainty [26] perturbs both the image and the text. Our analysis shows that textual perturbations often dominate output variability due to prompt sensitivity, causing uncertainty estimates to reflect linguistic variation rather than visual ambiguity (Sec. 3.3).

Motivated by these limitations, we focus on uncertainty arising from visual ambiguity. Our design probes uncertainty through visual perturbations while avoiding text perturbations that introduce prompt sensitivity, and aggregates predictions at the semantic level to mitigate wording variations.

3 Problem Analysis

We analyze limitations of existing uncertainty estimators in VLMs and formalize two hypotheses and one proposition that guide our method.

3.1 Problem Formulation

A visual question instance consists of a question q and an image v. Given (q, v), a VLM produces an answer $a = \text{VLM}(q, v)$. Our goal is to estimate the reliability of this prediction. Specifically, we seek an uncertainty estimator U such that $\tilde{u} = U(q, v; \text{VLM})$, where larger uncertainty score $\tilde{u}$ indicates a higher likelihood that the generated answer a is incorrect. This is of high interests in scenarios where unreliable predictions must be filtered.

3.2 Confident visual embeddings induce low semantic entropy

Semantic Entropy estimates uncertainty by sampling multiple outputs under stochastic decoding and computing the entropy of the resulting categorical answer distribution. This approach assumes that epistemic uncertainty manifests as variability induced by decoding randomness. In VLMs, however, predictions are also conditioned on visual embeddings. When these embeddings are highly confident, stochastic decoding produces nearly identical outputs, leading to low SE even if the underlying image is ambiguous.

Hypothesis 1*. Confident visual embeddings induce low semantic entropy.* If the visual embedding yields a sharply peaked conditional answer distribution, decoding randomness alone cannot generate diverse outputs. As a result, SE may underestimate visual uncertainty.

To test this hypothesis, we utilize the *Adversarial* split of the VILP dataset [12], which consists of visually ambiguous images. We then filter samples that the VLM *answers incorrectly* and quantify their visual confidence as follows: Let $\mathbf{v}_i \in \mathbb{R}^D$ denote the embedding of the i-th visual token at the final layer and $W \in \mathbb{R}^{\mathcal{V} \times D}$ the language modeling head with vocabulary size of $\mathcal{V}$ of the VLM. Each visual token is projected into the vocabulary space using LogitLens [18]:

$$\mathbf{z}_i = W\mathbf{v}_i, \qquad \tilde{\mathbf{z}}_i = \text{softmax}(\mathbf{z}_i). \tag{1}$$

We compute the entropy of each visual token distribution and average across visual tokens to obtain the visual entropy H_{vis}:

$$H_{\text{vis}}(x) = \operatorname*{avg}_i \left(-\sum_{k=1}^{|\mathcal{V}|} \tilde{\mathbf{z}}_i(k) \log \tilde{\mathbf{z}}_i(k) \right). \tag{2}$$

Lower visual entropy H_{vis} indicates a more confident visual representation. Based on this metric, we partition samples into two groups: Visually confident (low

visual entropy) and Visually uncertain (high visual entropy). Importantly, this metric only reflects the model certainty *in the visual input*, not the predicted answer. We expect high SE (uncertainty) in both groups, since the model answers these samples incorrectly. However, as shown in Fig. 3-*Left*, standard SE yields high uncertainty only for visually uncertain samples. Visually confident samples exhibit low SE despite the visual ambiguity, indicating that confident visual embeddings suppress output variability, yielding low semantic entropy.

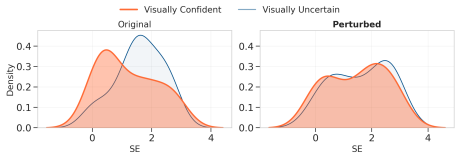

Fig. 3: Decoding-based Semantic Entropy underestimates visual ambiguity. On the samples with incorrect answer, models are expected to express high uncertainty (SE). Using LogitLens [18] on the original image, we partition samples into Visually Confident (low visual entropy) and Visually Uncertain (high visual entropy). *Left*: Under stochastic decoding, high SE appears only for Visually Uncertain samples. Visually Confident samples remain low despite the ambiguous input. *Right*: After perturbing the image and aggregating predictions across perturbed views, we can induce high SE for Visually Confident samples. Results are shown for Qwen2.5-VL on VILP dataset.

We then perturb the input image to generate alternative local views while preserving semantic content. As shown in Fig. 3-*Right*, perturbation induces higher semantic entropy for visually confident samples. We note that the grouping is determined by calculating visual entropy H_{vis} on the original image. More details on experimental settings are included in the Supplementary.

Discussion. These results demonstrate that decoding stochasticity alone fails under confident visual embeddings, while image perturbation induces high semantic entropy for visually ambiguous inputs. This motivates probing uncertainty in the visual input space.

3.3 Text perturbation dominates output semantic shifts

Beyond perturbing visual inputs, recent work also perturbs the text query [8,26]. Unlike visual perturbations, which introduce small local changes around the same image representation, textual paraphrases can induce larger semantic shifts. Semantically equivalent sentences may occupy distant regions in embedding space due to the discrete nature of language. As a result, paraphrasing can substantially alter output semantics. This motivates the following hypothesis.

Hypothesis 2. *Text perturbations dominate output diversity under joint perturbation.* We hypothesize that output diversity is driven primarily by textual

Table 1: Text (P_T) vs. Image (P_I) perturbations Purity on VILP dataset for Qwen2.5-VL, Gemma3, Intern3.5-VL and LlaVA-NeXT.

	Qwen2.5-VL			**Gemma3**			**Intern3.5-VL**			**LlaVA-NeXT**		
	Easy	*Adv.*	*All*	*Easy*	*Adv.*	*All*	*Easy*	*Adv.*	*All*	*Easy*	*Adv.*	*All*
P_T(%)	51.9	55.6	54.3	35.6	38.3	37.4	46.7	48.4	47.8	54.5	53.9	54.1
P_I(%)	36.4	40.1	38.8	30.4	36.8	34.6	37.8	41.4	40.2	41.0	44.8	43.4

variation rather than visual ambiguity. Then uncertainty measured under joint perturbation would mostly reflect *prompt sensitivity*.

To test this, we analyze several VLMs on the *Easy* and *Adversarial* split of the VILP dataset. For each image, we construct an $L \times M$ perturbation grid, where L denotes the number of textual paraphrases and M denotes the number of image perturbations. For each (l, m) pair, we generate a model output and cluster the resulting $L \times M$ responses based on semantic similarity. This allows us to analyze how cluster membership depends on paraphrase identity or image perturbation.

Purity analysis. To quantify the alignment between clusters and perturbation types, we compute cluster purity. For each cluster c, we define *text purity* $P_T(c)$ and *image purity* $P_I(c)$ as the fraction of samples in c originating from the most frequent textual paraphrase and image perturbation. Let n_c denote the size of cluster c, $n_c(l)$ the number of samples in c generated from paraphrase l, and $n_c(m)$ the number generated from image perturbation m. We define

$$P_T(c) = \frac{\max_l n_c(l)}{n_c}, \quad P_I(c) = \frac{\max_m n_c(m)}{n_c}. \tag{3}$$

If $P_T(c) > P_I(c)$, most samples in cluster c originate from a single paraphrase $l_{\max}$, indicating that cluster membership is primarily determined by the textual input rather than visual perturbations, and vice versa. To obtain an overall text and image purity for each image across clusters, we compute the size-weighted purity: $P_T = \sum_c \frac{n_c}{L \times M} P_T(c)$ and $P_I = \sum_c \frac{n_c}{L \times M} P_I(c)$.

Overall, text purity consistently exceeds image purity (Tab. 1), and the gap widens on the *Easy* split. Furthermore, we estimate the joint density of (P_T, P_I) across images using kernel smoothing and visualize it as a filled contour plot (Fig. 4). The diagonal $P_T = P_I$ separates text-dominant (above) from image-dominant regions (below). We observe that the density mass concentrates predominantly above this line, meaning that for most images, cluster assignments align more strongly with paraphrase identity than with visual perturbations. This pattern shows that variability under joint perturbation is largely text-driven rather than visual ambiguity. Additional visualizations and details are in Supp.

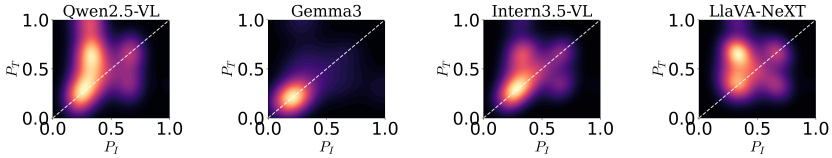

Fig. 4: Kernel density of text P_T and image purity P_I across models.. The diagonal $P_T = P_I$ separates text- from image-dominant regions. Density mass lying predominantly above the diagonal indicates that clustering is primarily text-driven.

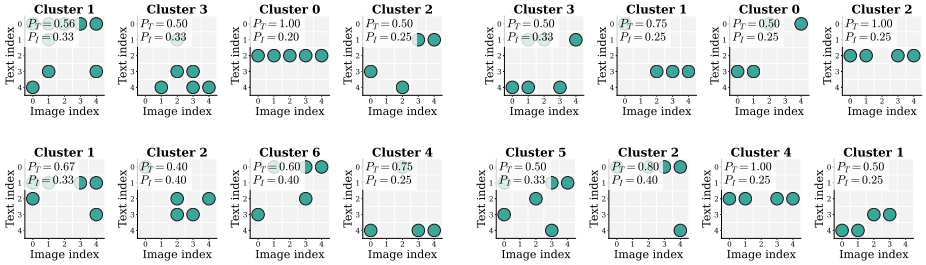

Fig. 5: Cluster occupancy map. Each panel shows cluster assignments on the $L \times M$ perturbation grid of a random sample (rows: textual paraphrases; columns: image perturbations). Horizontal stripes indicate invariance across image perturbations, showing that clustering is dominant by textual paraphrase. We provide more results in Supp.

To qualitatively examine this effect, we visualize cluster occupancy on the $L \times M$ perturbation grid. Rows correspond to textual paraphrases and columns to image perturbations. Under joint perturbation, we observe *horizontal stripe patterns* in the occupancy maps (Fig. 5), indicating that cluster membership is invariant across image perturbations but consistent within a given paraphrase. This pattern confirms that output are primarily conditioned on the textual input.

Discussion. The quantitative purity analysis and qualitative examples support Hypothesis 2: under joint image-text perturbation, output diversity is largely driven by textual variation rather than visual ambiguity. Our finding aligns with Image-DPO [12], where the authors perturb only the image while keeping the question fixed to dismiss model over-reliance on language. In uncertainty estimation, prompt sensitivity leads paraphrasing to induce semantic shifts that inflate uncertainty without reflecting ambiguity in visual information. Hence, restricting perturbations to the visual domain avoids artificial prompt sensitivity and yields more faithful visual uncertainty.

3.4 Linguistic variations inflates distance-based uncertainty

Other approaches estimate uncertainty by measuring semantic distances among sampled outputs [15, 17]. However, because answers are discrete text, semantically equivalent responses with different wording can be mapped to distant points in embedding space. As a result, distance-based aggregation may yield inflated uncertainty even when there is no true semantic disagreement.

Formally, given n sampling iterations, the model generates answer variants $\{a_i\}^n$. A general distance-based estimator computes uncertainty as the average pairwise distance:

$$U(q) = \frac{1}{n(n-1)} \sum \mathrm{d}(a, a'), \tag{4}$$

where $\mathrm{d}(\cdot, \cdot)$ is a semantic distance function.

Proposition 1. *Distance inflation under linguistic variations.* Suppose all sampled answers express the same semantic meaning. If wording differences induce non-zero embedding distances, $\mathrm{d}(a, a') > 0$ for some $a \neq a'$, then $U(q) > 0$ despite the absence of semantic disagreement.

Intra- vs. Inter-cluster effect. To separate meaning-level disagreement from wording variation, consider partitioning the answers into semantic clusters, where each cluster groups answers that convey the same meaning. Let c denote the cluster assignment of a. For a distance metric $\mathrm{d}(\cdot, \cdot)$, pairwise aggregation can be decomposed as:

$$U(q) = \sum_{k=1}^{K} \mathbb{E}[\mathrm{d}(a, a') \mid c = c' = k] + \sum_{k \neq k'} \mathbb{E}[\mathrm{d}(a, a') \mid c = k, c' = k'] \,. \tag{5}$$

The first term captures variation within clusters, which mainly reflects wording differences. The second term captures variation across clusters, corresponding to genuine semantic disagreement. Because distance-based aggregation sums both terms, uncertainty is influenced not only by disagreement between meanings but also by wording variation within each meaning. Therefore, uncertainty estimation should first group semantically equivalent answers and then measure distances across groups rather than between individual text outputs.

4 Method

The analysis above shows three limitations of current uncertainty estimators: (1) decoding stochasticity underestimates uncertainty when visual embeddings are overly confident, (2) text perturbations can dominate output variability under joint image-text perturbation, and (3) linguistic variations inflate distance-based estimators. These observations suggest that VQA uncertainty estimation should probe ambiguity in the visual input space, avoid textual perturbations that introduce large semantic shifts, and aggregate outputs at the semantic level.

Guided by these principles, we propose **Visual Semantic Entropy (VSE)**, a method that perturbs the image to elicit alternative visual interpretations (Sec. 4.1), clusters semantically equivalent answers into prototypes, and quantifies uncertainty as the weighted dispersion among prototype embeddings (Sec. 4.2). An overview of VSE is shown in Fig. 6.

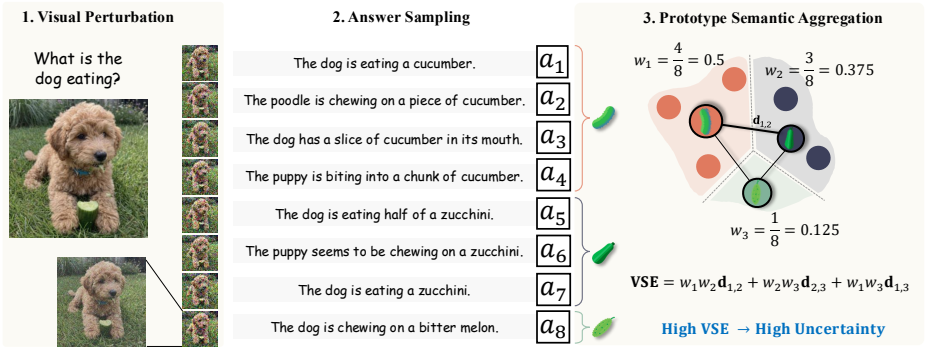

Fig. 6: Visual Semantic Entropy for VQA Uncertainty Estimation. (1) We perturb the input image to generate local visual variants. (2) The VLM produces multiple answer samples across perturbed views. (3) We cluster semantically equivalent answers, select a representative prototype for each cluster, and compute uncertainty as the mass-weighted dispersion among prototype embeddings.

4.1 Visual Perturbation

To probe visual epistemic uncertainty, we perturb the input image while keeping the question fixed. Unlike textual paraphrasing, which may induce non-local semantic shifts, visual perturbations introduce controlled variations around the original image, enabling us to explore local visual variants.

Let $\mathcal{T}$ denote a visual perturbation operator. Given an input image v, we generate M perturbed views:

$$v_m = \mathcal{T}(v; \xi_m), \quad m = 1, \ldots, M, \tag{6}$$

where ξ_m parameterizes the perturbation. The operator $\mathcal{T}$ is designed to preserve high-level semantic content while inducing local variations in pixel space. Each perturbed image v_m is paired with the same question q and fed into the VLM to obtain answer samples. Variability across these perturbed views v_m serves as a probe of visual uncertainty.

Discussion. Prior work [26] progressively increases perturbation magnitude to generate samples at multiple noise levels. In contrast, we fix a small perturbation scale for two reasons. First, uncertainty should quantify ambiguity conditioned on the original input (q, v). Progressively increasing perturbations magnitude may move samples outside the local neighborhood of v, so answer variation reflects changes to the input rather than visual ambiguity. Second, sampling-based estimators aggregate samples uniformly. Mixing perturbation magnitudes inappropriately assigns equal weight to local and non-local, large deviations. By using a single fixed scale, the measured dispersion consistently reflects variability within a local neighborhood.

4.2 Prototype Semantic Aggregation

Given M perturbed image variants $\{v_m\}^M$ and the question q, the VLM produces a corresponding set of answer samples $\{a_m\}^M$. These samples may differ in wording or semantic meaning. We then perform Prototype Semantic Aggregation (**ProtoSem**) as follows:

Clustering. We cluster the answer variants $\{a_m\}$ into K groups $\{c_k\}_{k=1}^K$ based on semantic distance base on a function $\mathrm{d}(\cdot,\cdot)$.

Prototype selection. For each cluster c_k, we select a prototype answer p_k that minimizes the average distance to other members in the same cluster:

$$p_k = \arg\min_{a \in c_k} \sum_{a' \in c_k} \mathrm{d}(a, a'). \tag{7}$$

This prototype serves as the representative semantic meaning of cluster k.

Prototype dispersion. Let $w_k = \frac{|c_k|}{M}$ denote the empirical probability of cluster k. We define uncertainty as the expected semantic distance between two independently drawn cluster prototypes:

$$\tilde{u} = U(q, v; \mathrm{VLM}) = \mathbb{E}_{k \sim w,\, k' \sim w}[\mathrm{d}(p_k, p_{k'})] = \sum_{k \neq k'} w_k w_{k'}\, \mathrm{d}(p_k, p_{k'}). \tag{8}$$

This measures disagreement among supported semantic meanings: semantic variations are consolidated within each cluster and represented by a single prototype p_k, and each prototype is weighted by its mass w_k. Uncertainty increases only when *multiple high-mass prototypes are separated.*

Discussion. SE [3] treats answers as discrete categories and ignores semantic proximity, whereas SNNE [15] measures pairwise distances and is sensitive to linguistic variation. VSE instead groups semantically equivalent outputs and compute weighted dispersion over cluster prototypes, eliminating intra-cluster wording inflation while preserving semantic disagreement across clusters.

5 Experiments

5.1 Setup

Metrics. Following prior work [3,15], we evaluate uncertainty estimation using the Area Under the ROC Curve (AUC). Each prediction is labeled as correct or incorrect, and the predicted uncertainty score is used to distinguish between them. AUC measures how well uncertainty scores rank incorrect predictions above correct ones across different thresholds. A higher AUC indicates better alignment between predicted uncertainty and model reliability.

Datasets. We evaluate our method on five benchmarks covering knowledge-based VQA, multimodal reasoning: OKVQA [14], AOKVQA [20],MMVet [25]; and adversarial datasets to expose visual bias: VILP [12], VLM-are-biased [22]. We detail on the dataset and their splits in the Supplementary material.

Other Baselines. We conduct an extensive benchmark to compare **VSE** against these uncertainty estimation approaches: *Verbalized*: Verbalized Uncertainty [5] . *Logit-based*: Confidence measures derived from output probabilities, including AvgEnt and MaxEnt (token entropy), and AvgProb and MaxProb (token probability) [10]. *Consistency-based*: SelfCheckGPT [13] and C&U [8]. *Entropy-based*: Semantic Entropy (SE) [3], Semantic Nearest Neighbor Entropy (SNNE) [15], Kernel Language Entropy (KLE) [17], and VL-Uncertainty [26]. Our method, **VSE**, also belongs to this category.
Models. We evaluate all approaches on five vision-language models that cover a range of architecture: Qwen2.5-VL-7B [2], Gemma3-4B [4], Intern3.5-VL-8B [23], LLaVA-NeXt-8B [11], and Qwen3-VL-8B [1].
Implementation Details. To perturb the image, we add Gaussian noise with standard deviation $\sigma = 20$, introducing small visual variations while preserving semantic content. The initial answer a is generated using greedy decoding ($T = 0.0$). After obtaining this initial answer, we sample $M = 10$ additional responses with decoding temperature $T = 1.0$, following the setup of SE [3]. For answer clustering, we use the DeBERTa-v2-xlarge-mnli [6] as the semantic distance function $\mathrm{d}(\cdot, \cdot)$ and apply hierarchical clustering [16]. Additional hyperparameter analysis can be found in supplementary material.

5.2 Results

Qualitative Results. We report results on AOKVQA in Tab. 2. Entropy-based methods perform best overall, while logit-based approaches perform poorly across all models. Verbalized uncertainty performs competitively only on Qwen3-VL. Consistency-based methods also show strong performance, particularly C&U with text perturbations. Among entropy-based methods, **VSE** consistently achieves the best results across all models, reaching AUC scores of 0.783, 0.778, 0.724, 0.798, and 0.792 on Qwen2.5-VL, Gemma3, LLaVA-NeXT, Qwen3-VL, and Intern3.5-VL, and outperforming VL-Uncertainty by up to 9.2% AUC. The same trend holds on MMVet and OKVQA (Tabs. 3 and 4). VSE achieves the best performance across all models, improving over the strongest baselines such as SE (e.g., 0.767 vs. 0.758 on Qwen2.5-VL in OKVQA) and reaching AUC scores of 0.781 and 0.778 on MMVet for Qwen2.5-VL and Gemma3, respectively. We observe that SE and SNNE perform well on MMVet, suggesting that repeated sampling better captures uncertainty in this benchmark, which primarily evaluates reasoning ability rather than visual ambiguity.
Qualitative Results. We provide qualitative result in Fig. 7. More results are provided in Supplementary.

5.3 Ablation Study

VSE reflects visual ambiguity more faithfully. We evaluate uncertainty estimation on the visually adversarial datasets VILP and VLM-are-biased (Tab. 5), where images are intentionally designed to be highly ambiguous to probe model biases and test reliance on visual evidence. VSE consistently achieves the best

Fig. 7: Qualitative Result of Visual Semantic Entropy. *Top*: VSE is high, showing high uncertainty for wrong prediction. *Bottom*: VSE is low, showing more certainty for correct prediction. Results are shown for Qwen2.5-VL on AOKVQA.

Table 2: AOKVQA results. AUC scores of uncertainty estimation methods across Qwen2.5-VL, Gemma3, LLaVA-NeXT, Qwen3-VL, and Intern3.5-VL. Higher is better.

Category	Method	Venue	Qwen2.5	Gemma3	LLaVA	Qwen3	Intern3.5
Verbalized	Verb-U [5]	NAACL'24	0.635	0.654	0.580	0.700	0.630
Logit	AvgEnt [10]	EMNLP'24	0.625	0.530	0.635	0.581	0.613
	MaxEnt [10]	EMNLP'24	0.596	0.529	0.638	0.578	0.632
	AvgProb [10]	EMNLP'24	0.612	0.508	0.634	0.556	0.593
	MaxProb [10]	EMNLP'24	0.600	0.561	0.522	0.541	0.512
Consistency	SCG-NLI [13]	EMNLP'23	0.575	0.576	0.579	0.591	0.618
	SCG-Pr [13]	EMNLP'23	0.734	0.701	0.608	0.708	0.730
	C&U [8]	CVPR'24	0.731	0.712	0.602	0.720	0.734
Entropy	SE [3]	ICLR'23	0.702	<u>0.732</u>	0.622	0.746	0.775
	SNNE [15]	ACL'25	0.744	0.700	<u>0.651</u>	0.747	0.745
	KLE [17]	NIPS'25	<u>0.774</u>	0.721	0.541	<u>0.772</u>	0.764
	VL-U [26]	arxiv	0.756	0.715	0.616	0.706	<u>0.781</u>
	VSE	**Ours**	**0.783**	**0.778**	**0.724**	**0.798**	**0.792**

Table 3: MMVet results. AUC scores of uncertainty estimation methods across Qwen2.5-VL, Gemma3. Higher is better.

Model	Verb-U [5]	A-E [10]	A-P [10]	SCG [13]	C&U [8]	SE [3]	SNNE [15]	VL-U [26]	**VSE**
Qwen2.5-VL	0.595	0.687	0.678	0.718	0.587	0.716	0.758	0.659	**0.781**
Gemma3	0.598	0.607	0.601	0.625	0.646	0.722	0.740	0.693	**0.778**

Table 4: OKVQA results. AUC scores of uncertainty estimation methods across Qwen2.5-VL, Gemma3, LLaVA-NeXT, Qwen3-VL. Higher is better.

Category	**Method**	**Venue**	**Qwen2.5**	**Gemma3**	**LLaVA**	**Qwen3**
Verbalized	Verb-U [5]	NAACL'24	0.663	0.647	0.532	0.593
Logit	AvgEnt [10]	EMNLP'24	0.690	0.577	0.729	0.703
	MaxEnt [10]	EMNLP'24	0.714	0.567	0.724	0.715
	AvgProb [10]	EMNLP'24	0.653	0.570	0.705	0.679
	MaxProb [10]	EMNLP'24	0.519	0.555	0.542	0.565
Consistency	SCG-NLI [13]	EMNLP'23	0.714	0.584	0.658	0.597
	SCG-Pr [13]	EMNLP'23	**0.767**	0.657	0.675	0.657
	C&U [8]	CVPR'24	0.742	0.678	0.718	0.693
Entropy	SE [3]	ICLR'23	0.758	0.703	0.685	0.790
	SNNE [15]	ACL'25	0.751	0.667	0.698	0.711
	KLE [17]	NIPS'25	0.744	0.673	0.672	0.724
	VL-U [26]	arxiv	0.731	0.702	0.695	0.731
	VSE	**Ours**	**0.767**	**0.715**	**0.749**	**0.799**

performance across all settings, improving over the strongest baseline by large margins (e.g., 0.650 vs. 0.535 on VILP with Qwen2.5-VL and 0.826 vs. 0.783 on VLM-are-biased). In contrast, several existing methods degrade substantially on these datasets, suggesting that they are sensitive to language biases or decoding artifacts. These results highlight that explicitly probing visual variations help to capture uncertainty arising from ambiguous visual inputs.

Table 5: AUC scores of uncertainty estimation methods on the VILP and VLM-are-biased benchmarks.

Dataset	**Model**	Verb-U [5]	A-E [10]	SE [3]	C&U [8]	VL-U [26]	**VSE**
VILP	**Qwen2.5-VL**	0.521	0.507	0.535	0.515	0.489	**0.650**
	Gemma3	0.503	0.528	0.660	0.540	0.501	**0.665**
VLM-are-biased	**Qwen2.5-VL**	0.722	0.537	0.728	0.697	0.783	**0.826**
	Gemma3	0.600	0.694	0.700	0.737	0.758	**0.776**

Effect of Visual Perturbation and Prototype Aggregation. Tab. 6 evaluates two components of our approach. For SE and SNNE, we report the original

estimator (*Base*) and a variant where we add visual perturbation ($+\mathcal{T}$). We also evaluate **ProtoSem**, our prototype semantic aggregation strategy, both with and without visual perturbation.

(1) Comparing *Base* and $+\mathcal{T}$ shows consistent improvements for SE and SNNE across all models, with gains up to +9.5% AUC, demonstrating that visual perturbations are beneficial for uncertainty estimation.

(2) Comparing methods across $+\mathcal{T}$ columns (with visual perturbation), **ProtoSem** consistently achieves the best performance, demonstrating that prototype-based semantic aggregation better captures uncertainty than entropy computed over sampled answers, as used in SE and SNNE.

These results justify the core design of VSE, combining visual perturbation with prototype-based semantic aggregation to improve uncertainty estimation.

Table 6: Effect of visual perturbation ($+\mathcal{T}$) and Prototype Semantic Aggregation (ProtoSem) on entropy-based uncertainty estimation methods on the AOKVQA dataset. The ProtoSem + $\mathcal{T}$ combination is our proposed **VSE**.

	SE [3]			SNNE [15]			**ProtoSem**		
Model	Base	$+\mathcal{T}$	Δ	Base	$+\mathcal{T}$	Δ	Base	$+\mathcal{T}_{(\mathbf{VSE})}$	Δ
Qwen2.5-VL	0.702	0.761	+0.059	0.700	0.720	+0.020	0.711	0.783	+0.072
Gemma3	0.732	0.775	+0.043	0.651	0.746	+0.095	0.745	0.778	+0.033
Qwen3-VL	0.747	0.788	+0.041	0.747	0.757	+0.010	0.762	0.798	+0.036

6 Conclusion

We study uncertainty estimation in VLMs for VQA, where ambiguity often arises from the visual input. Our analysis shows that decoding-based estimators can underestimate uncertainty when confident visual embeddings suppress output variation, while text perturbations and wording variations can inflate uncertainty estimates. Motivated by these observations, we propose **Visual Semantic Entropy**, which probes uncertainty through visual perturbations and aggregates predictions at the semantic prototype level. Experiments across multiple VLMs and VQA benchmarks demonstrate that VSE provides more reliable uncertainty estimates than existing approaches. A limitation of our method is the additional computational cost from perturbation-based sampling, which is shared with other sampling-based uncertainty estimators, as well as sensitivity to the decoding temperature used during generation. Our approach also relies on semantic similarity models to cluster answers, and its performance may depend on the quality of the sematic distance function. While we evaluate VSE across several representative VLMs and benchmarks, extending the analysis to larger model families and additional multimodal tasks remains an important direction for future work. We hope this work encourages future research on uncertainty estimation that explicitly accounts for visual ambiguity.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025) 12
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025), https://arxiv.org/abs/2502.13923, computer Vision and Pattern Recognition (cs.CV) 12
3. Farquhar, S., Kossen, J., Kuhn, L., Gal, Y.: Detecting hallucinations in large language models using semantic entropy. Nature **630**(8017), 625–630 (2024) 2, 4, 11, 12, 13, 14, 15
4. Gemma Team: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (2025), https://arxiv.org/abs/2503.19786 12
5. Groot, T., Valdenegro-Toro, M.: Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. In: Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024). pp. 145–171 (2024) 1, 4, 12, 13, 14
6. He, P., Liu, X., Gao, J., Chen, W.: Deberta: Decoding-enhanced bert with disentangled attention. arXiv preprint arXiv:2006.03654 (2020) 12
7. Huang, J., Xu, J., Shi, X., Hu, P., Feng, L., Zhu, X.: Revisiting confidence calibration for misclassification detection in vlms. In: The Fourteenth International Conference on Learning Representations 1, 4
8. Khan, Z., Fu, Y.: Consistency and uncertainty: Identifying unreliable responses from black-box vision-language models for selective visual question answering. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition. pp. 10854–10863 (2024) 2, 3, 4, 6, 12, 13, 14
9. Li, M., Zhang, Y., Long, D., Chen, K., Song, S., Bai, S., Yang, Z., Xie, P., Yang, A., Liu, D., et al.: Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. arXiv preprint arXiv:2601.04720 (2026) 3
10. Li, Q., Geng, J., Lyu, C., Zhu, D., Panov, M., Karray, F.: Reference-free hallucination detection for large vision-language models. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 4542–4551 (2024) 1, 4, 12, 13, 14
11. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/ 12
12. Luo, T., Cao, A., Lee, G., Johnson, J., Lee, H.: Probing visual language priors in vlms. arXiv preprint arXiv:2501.00569 (2024) 5, 8, 11
13. Manakul, P., Liusie, A., Gales, M.: Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In: Proceedings of the 2023 conference on empirical methods in natural language processing. pp. 9004–9017 (2023) 2, 4, 12, 13, 14
14. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. pp. 3195–3204 (2019) 11

15. Nguyen, D., Payani, A., Mirzasoleiman, B.: Beyond semantic entropy: Boosting llm uncertainty quantification with pairwise semantic similarity. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 4530–4540 (2025) 2, 3, 4, 8, 11, 12, 13, 14, 15
16. Nielsen, F.: Hierarchical clustering. In: Introduction to HPC with MPI for Data Science, pp. 195–211. Springer (2016) 12
17. Nikitin, A., Kossen, J., Gal, Y., Marttinen, P.: Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. Advances in Neural Information Processing Systems **37**, 8901–8929 (2024) 2, 4, 8, 12, 13, 14
18. nostalgebraist: interpreting gpt: the logit lens. LessWrong post (Aug 31 2020), https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens 5, 6
19. Pelucchi, M.: Exploring ChatGPT's accuracy and confidence in high-resource languages. Ph.D. thesis (2023) 1, 4
20. Schwenk, D., Khandelwal, A., Clark, C., Marino, K., Mottaghi, R.: A-okvqa: A benchmark for visual question answering using world knowledge. In: European conference on computer vision. pp. 146–162. Springer (2022) 3, 11
21. Valdenegro-Toro, M.: I find your lack of uncertainty in computer vision disturbing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1263–1272 (2021) 1, 4
22. Vo, A., Nguyen, K.N., Taesiri, M.R., Dang, V.T., Nguyen, A.T., Kim, D.: Vision language models are biased. arXiv preprint arXiv:2505.23941 (2025) 11
23. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025) 12
24. Xuan, W., Zeng, Q., Qi, H., Wang, J., Yokoya, N.: Seeing is believing, but how much? a comprehensive analysis of verbalized calibration in vision-language models. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 1408–1450 (2025) 1, 4
25. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. In: International conference on machine learning. PMLR (2024) 11
26. Zhang, R., Zhang, H., Zheng, Z.: Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. arXiv preprint arXiv:2411.11919 (2024) 2, 3, 4, 6, 10, 12, 13, 14
27. Zhang, S., Sambara, S., Banerjee, O., Acosta, J., Fahrner, L.J., Rajpurkar, P.: Radflag: A black-box hallucination detection method for medical vision language models. arXiv preprint arXiv:2411.00299 (2024) 2