VLM-UQBench: A Benchmark for Modality-Specific and Cross-Modality Uncertainties in Vision Language Models

Anonymous ACL submission

Abstract

Uncertainty quantification (UQ) is important for the safe and reliable deployment of vision–language models (VLMs). While many UQ methods developed for LLMs have been applied to VLMs with promising results, uncertainty in VLMs can stem from the image, the text, or their misalignment, and it remains unclear how effectively existing methods capture these sources. To address this gap, we introduce **VLM-UQBench**, a benchmark for modality-specific and cross-modal uncertainty in VLMs. It includes 600 real-world VizWiz samples spanning clean, image-uncertain, text-uncertain, and cross-modal subsets, along with a scalable perturbation pipeline that extends to other datasets by converting original samples into counterfactual examples with targeted uncertainty. We also propose two simple metrics to measure the sensitivity of UQ scores to these uncertainty sources and their correlation with hallucinations. Extensive experiments across multiple UQ methods, VLMs, and datasets show that (i) UQ methods originally developed for LLMs lack robustness when transferred to VLMs, and their performance is highly sensitive to both the source of uncertainty and the underlying VLM architecture; (ii) Modality-specific uncertainty often induces hallucinations, yet those UQ methods are often not capable of providing reliable signals of such risk; and (iii) despite these limitations, current UQ methods remain practically useful and can rival GPT-4o with chain-of-thought prompting for ambiguity detection. In addition, simple calibrated ensembles of these UQ methods can provide a stronger and more practical uncertainty estimator. Code and dataset will be publicly available.

1 Introduction

Uncertainty quantification (UQ) is increasingly important for deploying vision–language models (VLMs) in applications such as assistive question answering, visual assistants, robotics, and medical or scientific image analysis (Venkataramanan et al., 2025; Huang et al., 2024a; Ren et al., 2023; Wang et al., 2025). In these settings, models should not only answer questions, but also know when to abstain, ask for clarification, or defer to humans, since hallucinated outputs can be harmful. For VLMs, this challenge is inherently multimodal: uncertainty may come from the image, the text, or their misalignment. A useful UQ method should therefore capture uncertainty arising from different sources and help indicate the risk of hallucination. However, although many UQ methods originally developed for LLMs have been increasingly applied to VLMs (Zhang et al., 2024b; Lau et al., 2025), it remains unclear whether they are robust to modality-specific uncertainty, generalize across model architectures, or produce scores that reliably reflect hallucination risk.

To address this gap, both a tailored benchmark and a tailored evaluation protocol are needed. Existing VQA benchmarks make partial progress: UNK-VQA (Guo et al., 2024) and FocusAmbiguity (Chen et al., 2025) provide annotations for answerability and grounded ambiguity, respectively; MM-UPD (Miyai et al., 2025) introduces finer-grained unsolvability categories for different types of unanswerable data; MuirBench (Wang et al., 2024a) adopts a contrastive pair design, with one answerable instance paired with one annotated unanswerable counterpart; and CertainlyUncertain (Chandu et al., 2024) provides instance-wise aleatoric–epistemic uncertainty annotations for free-form VQA. Existing UQ evaluation mainly focuses on calibration, selective prediction, and error detection, measuring whether confidence aligns with correctness (Naeini et al., 2015; Huang et al., 2024b), whether uncertainty separates reliable from unreliable outputs (Nadeem et al., 2009; Kapoor et al., 2024; Kuhn et al., 2023), and whether it can identify incorrect or hallucinated generations (Xiao

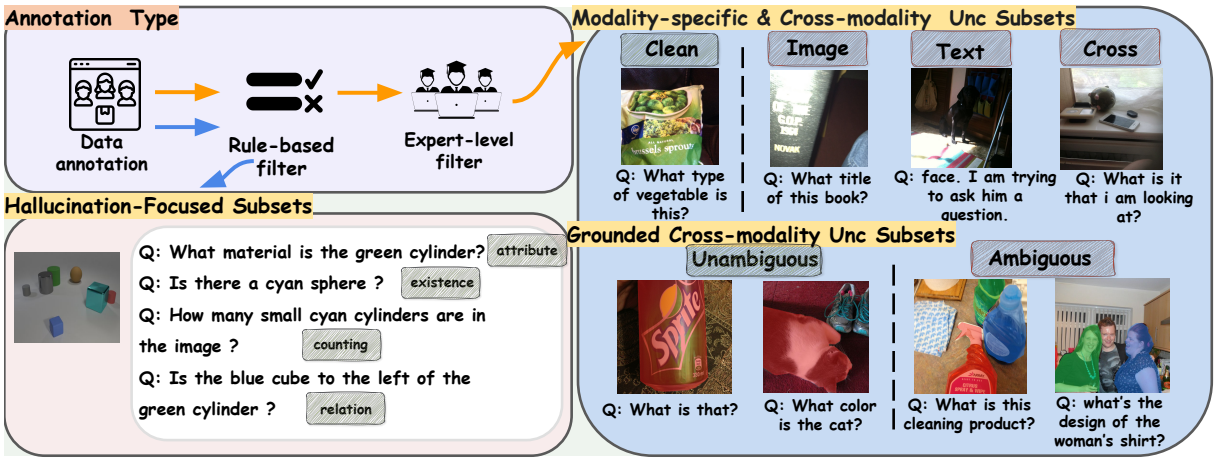

Figure 1: Benchmark composition and curation workflow. We integrate diverse data sources through a three-stage pipeline of *human annotation*, *rule-based filtering*, and *expert curation* (upper left), yielding four VizWiz-based subsets with modality-specific uncertainty labels (Clean, Image, Text, Cross; top right). VQ-FocusAmbiguity is leveraged as a grounded case of cross-modal ambiguity, providing unambiguous vs. ambiguous text–image alignment examples (middle right). Hallucination-focused subsets are built from CLEVR scene graphs using compositional templates and rule-based generation, enabling controlled evaluation of attribute, existence, counting, and relation hallucinations (bottom left).

and Wang, 2021). Some recent work also separates epistemic and aleatoric uncertainty (Hou et al., 2023; Yang et al., 2025). However, these benchmark and evaluation settings are still not sufficient for assessing the key properties discussed above. They do not explicitly separate image, text, and cross-modal uncertainty, and they do not directly evaluate whether UQ methods remain reliable across uncertainty sources, model architectures, and hallucination settings.

We therefore introduce VLM-UQBench, a benchmark for modality-specific uncertainty in free-form visual question answering. It combines curated real-world subsets with explicit labels for image, text, and cross-modal uncertainty, grounded cross-modal ambiguity from VQ-FocusAmbiguity (Chen et al., 2025), a scalable perturbation pipeline for controlled visual, textual, and cross-modal edits, and two paired metrics: Uncertainty Reflection Rate (URR) and Hallucination Consistency Coefficient (HCC). Together, these components provide a tailored framework for systematically evaluating how UQ methods behave in VLMs across uncertainty sources, model architectures, and downstream hallucination settings.

The contributions of this paper are fourfold:

- We introduce **VLM-UQBench**, a benchmark for modality-specific uncertainty in free-form VQA, covering image, text, and cross-modal uncertainty with curated real-world subsets and grounded ambiguity examples.
- We develop a scalable perturbation pipeline and propose two paired metrics, *Uncertainty Reflection Rate (URR)* and *Hallucination Consistency Coefficient (HCC)*, to evaluate UQ sensitivity to modality-specific uncertainty and its alignment with hallucination risk.
- We conduct a systematic benchmark of nine UQ methods originally designed for LLMs across four VLMs and three datasets, showing that these methods are highly sensitive to uncertainty source and model architecture, with weak alignment to hallucination risk.
- We show that UQ methods remain useful for practical VLM tasks such as ambiguity detection, and that simple calibrated ensembles can further improve performance by combining complementary UQ signals.

2 Related Works

2.1 UQ for LLM

Uncertainty estimation methods for LLMs are often divided into white-box methods, which access logits or hidden states, and black-box methods, which rely only on generated outputs (Shorinwa et al., 2025). Representative examples include probability-based scores such as Perplexity, Maximum Sequence Probability, Mean Token Entropy, and PMI (Fomicheva et al., 2020; Takayama and Arase, 2019), semantic measures such as Semantic Entropy (Kuhn et al., 2023), reflexive methods such as p(True) (Kadavath et al., 2022), and diversity-based methods such as DegMat and Lexical Sim-

ilarity (Lin et al., 2023; Fomicheva et al., 2020). Although these methods have shown promise in LLMs, applying them to VLMs is still not well studied, especially when uncertainty can come from the image, the text, or their interaction. We address this by systematically evaluating nine representative UQ methods on VLM-UQBench.

2.2 UQ Benchmark

Existing UQ benchmarks mainly evaluate whether uncertainty correlates with output quality, hallucination, or ambiguity. For LLMs, benchmarks are built on datasets such as CoQA, TriviaQA, MMLU, GSM8k, TruthfulQA, AmbigQA, and AmbigEnt (Reddy et al., 2019; Joshi et al., 2017; Hendrycks et al., 2020; Cobbe et al., 2021; Vashurin et al., 2025b; Lin et al., 2021; Kuhn et al., 2022; Hou et al., 2023). For VLMs, *Certainly Uncertain* provides epistemic and aleatoric labels in VQA (Chandu et al., 2024), but it does not distinguish whether uncertainty arises from the image, the text, or their interaction. VLM-UQBench addresses this gap by introducing explicit image, text, and cross-modal uncertainty, together with a scalable perturbation pipeline for controlled uncertainty injection.

2.3 UQ Evaluation

Uncertainty measures are commonly evaluated through calibration, selective prediction, and error detection. Calibration metrics assess alignment between confidence and correctness (Naeini et al., 2015; Huang et al., 2024b). Selective prediction uses metrics such as AUROC and AUARC to test whether uncertainty separates reliable from unreliable outputs (Nadeem et al., 2009; Kapoor et al., 2024; Kuhn et al., 2023). Error detection evaluates whether uncertainty can identify incorrect or hallucinated generations (Xiao and Wang, 2021). Recent work further decomposes uncertainty into epistemic and aleatoric components (Hou et al., 2023; Yang et al., 2025). However, these paradigms still treat uncertainty as a single scalar and ignore its modality source. In contrast, we use modality-specific subsets and contrastive perturbations, together with URR and HCC, to test sensitivity to visual, textual, and cross-modal uncertainty and their relationship to hallucinations.

3 VLM-UQBench

VLM-UQBench combines four components into a unified benchmark for uncertainty in VLMs: (i) human-annotated modality-specific subsets from VizWiz (clean, image, text, cross-modal), (ii) grounded cross-modal ambiguity cases from VQ-FocusAmbiguity, (iii) hallucination-focused subsets constructed from CLEVR scene graphs, and (iv) a synthetic perturbation pipeline that injects modality-specific uncertainty into standard VQA datasets. Figure 1 summarizes these components and the overall curation workflow. Table 1 summarizes how VLM-UQBench relates to existing uncertainty and ambiguity benchmarks.

3.1 Benchmark Overview

Human-Annotated Modality Subsets (VizWiz). We build the core modality-specific component of VLM-UQBench from the VizWiz dataset, which contains images captured by blind users, natural language questions, and ten crowd-sourced answers per question. Crucially, VizWiz provides annotations of plausible reasons for answer disagreement (Bhattacharya et al., 2019), including low-quality images, insufficient evidence, invalid or ambiguous questions, and subjective judgments.

For our benchmark, we focus on input-modality factors and group these labels into three categories: *visual uncertainty* (e.g., low-quality image artifacts such as blur, occlusion, or poor lighting), *textual uncertainty* (e.g., invalid or subjective questions that are vague or incoherent), and *cross-modality uncertainty* (e.g., ambiguous references or insufficient visual evidence given an otherwise clear question and image). Details of the full taxonomy and mapping are provided in Appendix A.

We then apply a filtering rule to construct non-overlapping subsets and have three domain-expert PhD annotators curate high-quality examples, yielding 150 samples for each modality-specific subset, plus a clean subset with no modality-induced uncertainty. These curated splits provide clear, human-labeled sources of visual, textual, and cross-modal uncertainty for evaluating UQ methods in VLMs.

Grounded Cross-Modality Subset (VQ-FocusAmbiguity). To complement the VizWiz subsets, we incorporate a grounded variant of cross-modal ambiguity from the *VQ-FocusAmbiguity* dataset (Chen et al., 2025). The dataset filters VQA pairs where a question may refer to multiple plausible image regions and uses segmentation-based grounding to annotate all valid referents, providing explicit supervision. We adopt its 60 ambiguous and 80 unambiguous examples as

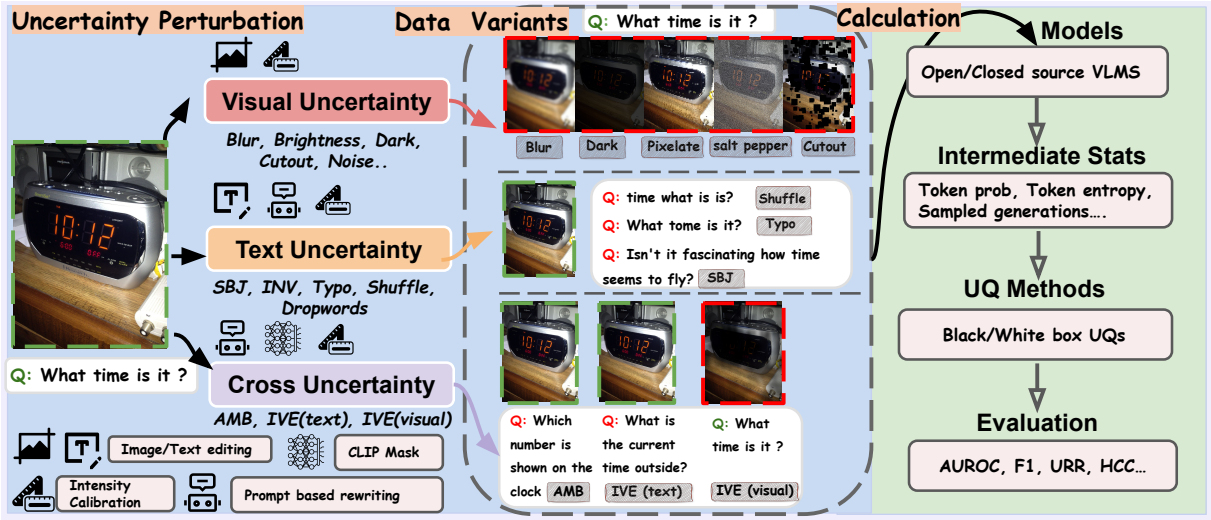

Figure 2: Overview of our pipeline. A clean image–question pair is processed through three stages: *uncertainty perturbation*, *variant generation*, and *uncertainty calculation*. In the perturbation stage (left), we inject *visual*, *textual*, and *cross-modal* uncertainty (e.g., blur, brightness, and occlusion for images; subjective or invalid rewrites, typos, and shuffles for text; AMB- and IVE-based edits for cross-modality). Perturbation intensity is calibrated on small validation subsets of the target datasets for VLMs to avoid being too weak (no effect) or too strong (catastrophic failure); we manually select these levels using a visualization tool, as detailed in Appendix B.4. In the variant-generation stage (middle), each original clean pair is expanded into a set of perturbed counterparts. In the calculation stage (right), all variants are fed into VLMs, from which we collect intermediate statistics (such as token probabilities, entropy, and sampled generations). Different UQ methods use these statistics to compute uncertainty scores, which are evaluated with AUROC, F1, URR, and HCC to study modality sensitivity and the relationship between uncertainty and hallucinations.

Table 1: Comparison of existing uncertainty benchmarks and VLM-UQBench on task modality, annotation type, and contrast granularity.

Benchmark	Task / Modality	Annotation Type	Contrast Granularity
AmbigQA (Min et al., 2020)	Open-domain QA	Ambiguity	Group
AbstentionBench (Kirichenko et al., 2025)	Open-domain QA	Abstention Scenarios	Group
UNK-VQA (Guo et al., 2024)	Open-ended VQA	Answerability	Group
VQ-FocusAmbiguity (Chen et al., 2025)	Free-form VQA	Ambiguity (Grounded)	Group
MM-UPD (Miyai et al., 2025)	Multiple-choice VQA	Unsolvability Categories	Group
MuirBench (Wang et al., 2024a)	Multiple-choice VQA	Answerability	Sample
CertainlyUncertain (Chandu et al., 2024)	Free-form VQA	Aleatoric–Epistemic Uncertainty	Sample
VLM-UQBench (Ours)	Free-form VQA	*Modality-specific & Cross-Modality* Uncertainty	Sample

grounded cross-modality samples, enabling precise evaluation of how UQ methods capture uncertainty from ambiguous text–image alignment.

Hallucination-Focused Subset (CLEVR-Hallucination). To analyze how modality-specific uncertainty contributes to different types of hallucination in VQA, we construct controlled evaluation sets using the CLEVR dataset (Johnson et al., 2017). Scene graphs and compositional templates are leveraged to generate four hallucination categories: attribute, existence, counting, and relation. Each question is grounded in object-level metadata, enabling reliable assessment. Further dataset generation details are provided in the Appendix. A.

3.2 Generalizing to Datasets Without Annotations: Synthetic Perturbation Pipeline

While the modality-specific and cross-modality subsets provide strong annotated uncertainty signals, they are costly to scale. To extend evaluation to larger and more diverse VQA datasets, we introduce a synthetic perturbation pipeline that automatically converts standard datasets into uncertainty-aware benchmarks without additional labels. Figure 2 illustrates the overall pipeline, from modality-specific perturbation injection to variant generation and uncertainty computation. The pipeline applies aligned perturbations to images, text, or both, corresponding to visual (U_v), textual (U_t), and cross-modal (U_x) uncertainty, with tunable intensity for

controlled evaluation.

Pipeline formulation. Given a clean image–question pair (x_v, x_t), the pipeline generates perturbed instances through a joint transformation T, i.e., $(\tilde{x}_v, \tilde{x}_t) = T(x_v, x_t; \boldsymbol{\lambda})$, where $\boldsymbol{\lambda} = [\lambda_v, \lambda_t, \lambda_x]$ controls the perturbation strength. We decompose $T(x_v, x_t; \boldsymbol{\lambda}) = \big(T_v^{(k_v)}(x_v; \lambda_v),\, T_t^{(k_t)}(x_v, x_t; \lambda_t),\, T_x^{(k_x)}(x_v, x_t; \lambda_x)\big)$, where $T_v^{(k_v)}$, $T_t^{(k_t)}$, and $T_x^{(k_x)}$ denote the k-th perturbation type for visual, textual, and cross-modal modalities, respectively. Each modality includes K_v, K_t, and K_x perturbation types, and varying $\lambda_i \in [0, 1]$ yields multiple uncertainty levels per sample while preserving the original semantic intent.

Visual modality. Visual uncertainty (U_v) simulates degradation of perceptual evidence through pixel-level and geometric distortions. For each visual perturbation type $k_v \in \{1, \ldots, K_v\}$, we define $\tilde{x}_v^{(k_v)} = T_v^{(k_v)}(x_v; \lambda_v)$, where $T_v^{(k_v)}$ applies operations such as blur, noise, compression, brightness or contrast shift, motion blur, or cutout, reproducing LOW-QUALITY IMAGE (LQI) conditions common in real-world VQA.

Textual modality. Textual uncertainty (U_t) captures ambiguity or inconsistency in the language input. For each textual perturbation type $k_t \in \{1, \ldots, K_t\}$, we generate $\tilde{x}_t^{(k_t)} = T_t^{(k_t)}(x_v, x_t; \lambda_t)$, where $T_t^{(k_t)}$ may condition on the image (e.g., to decide how to underspecify the question) but only modifies the text, using LLM-based paraphrases (INV/SBJ), typos, word drops, and phrase shuffles.

Cross-modal modality. Cross-modal uncertainty (U_x) reduces semantic alignment between the image and question. For each cross-modal perturbation type $k_x \in \{1, \ldots, K_x\}$, we obtain $(\tilde{x}_v^{(k_x)}, \tilde{x}_t^{(k_x)}) = T_x^{(k_x)}(x_v, x_t; \lambda_x)$, where $T_x^{(k_x)}$ introduces insufficient evidence (IVE) or ambiguous reference (AMB) conditions.

Outcome. Each modality-specific transformation independently injects uncertainty into the visual, textual, and cross-modal inputs:

$$\begin{aligned}\mathcal{D}_{\text{perturbed}} = \{\, &(T_v^{(k_v)}(x_v; \lambda_v),\ T_t^{(k_t)}(x_t; \lambda_t), \\ &T_x^{(k_x)}(x_v, x_t; \lambda_x)) \mid (x_v, x_t) \in \mathcal{D}, \\ &k_i \in \{1, \ldots, K_i\},\ \lambda_i \in [0, 1]\,\}.\end{aligned} \tag{1}$$

This yields a scalable, dataset-agnostic benchmark for evaluating UQ under controlled, modality-specific perturbations. Implementation details and augmentation parameters are provided in Appendix B.

3.3 Evaluation

We employ both conventional classification metrics and tailored correlation measures to evaluate UQ performance across modality-specific groups (image, text, and cross-modality) in VLMs. **AUROC** and **F1** assess how well each UQ method distinguishes uncertain samples from clean ones within each subset, with AUROC capturing discriminative power and F1 the precision–recall trade-off. Beyond these aggregate metrics, we introduce two measures that explicitly link UQ to modality-targeted perturbations and hallucination behavior.

Uncertainty Reflection Rate (URR). To test whether a UQ method responds appropriately to injected uncertainty, we define the *Uncertainty Reflection Rate* for a given UQ method u and a fixed perturbation operator $T^{(k)}$ with intensity λ (as introduced in Sec. 3.2). Given a clean input x_i and its perturbed version $x_i^{\text{pert}} = T^{(k)}(x_i; \lambda)$, let $U_i^{\text{clean}} = u(x_i)$ and $U_i^{\text{pert}} = u(x_i^{\text{pert}})$ be the corresponding uncertainty scores. We set $\delta_i = 1$ if $U_i^{\text{pert}} > U_i^{\text{clean}}$ and $\delta_i = 0$ otherwise, and define $\mathrm{URR}(u, T^{(k)}, \lambda) = \frac{1}{n}\sum_{i=1}^{n} \delta_i$.

A higher URR indicates that method u more consistently increases its uncertainty under this perturbation, showing stronger modality-specific sensitivity.

Hallucination Consistency Coefficient (HCC). To evaluate whether *changes* in uncertainty track hallucination behavior, we use the *Hallucination Consistency Coefficient* for a UQ method u under a fixed perturbation setting. For each sample i, let $U_i^{\text{clean}} = u(x_i)$, $U_i^{\text{pert}} = u(x_i^{\text{pert}})$, and $\Delta U_i = U_i^{\text{pert}} - U_i^{\text{clean}}$, and let $H_i \in \{0, 1\}$ be the hallucination label (1 = hallucinated). We compute the point-biserial correlation between ΔU and H, which in closed form is

$$\mathrm{HCC}(\Delta u) = \frac{\overline{\Delta U}_1 - \overline{\Delta U}_0}{s_{\Delta U}} \sqrt{\frac{n_1 n_0}{n^2}}, \tag{2}$$

where $\overline{\Delta U}_1$ and $\overline{\Delta U}_0$ are the mean uncertainty changes for hallucinated and non-hallucinated samples, $s_{\Delta U}$ is the standard deviation of all ΔU_i values, n_1 and n_0 are their counts, and $n = n_1 + n_0$.

A positive $\mathrm{HCC}(\Delta u)$ indicates that hallucinated cases tend to exhibit larger uncertainty increases under perturbation, suggesting the UQ reliably reflects hallucination sensitivity.

4 Experiments

We evaluate nine UQ methods originally designed for LLM, spanning both white-box and black-box settings. Across three datasets and four VLMs, our study is guided by three research questions:

- **RQ1:** How robust are these UQ methods when applied to VLMs across different uncertainty sources and model architectures?
- **RQ2:** Does modality-specific uncertainty induce hallucinations, and do these UQ methods reflect the resulting risk?
- **RQ3:** To what extent are these UQ methods remain useful for practical VLM tasks?
- **RQ4:** Can multiple complementary UQ signals be synergized to form a stronger and more practical uncertainty baseline?

4.1 Experiment Setup

Datasets. We evaluate across datasets spanning real-world, grounded, and synthetic settings. The human-annotated subsets from **VizWiz** and **VQ-FocusAmbiguity** provide modality-specific and cross-modality uncertainty labels for real and grounded images. The **VizWiz-Clean**, **VQAv2**, and **CLEVR** datasets serve as bases for our perturbation pipeline, enabling controlled injection of fine-grained visual, textual, and cross-modal uncertainty for large-scale evaluation. Finally, **CLEVR-Hallucination** offers scene-graph–based control for analyzing how modality uncertainty contributes to different hallucination types.

Models. We benchmark four vision–language models spanning different architectures and access regimes: three open-source models, **Kosmos-2** (Peng et al., 2023), **Qwen-VL** (Wang et al., 2024b), and **LLaVA** (Liu et al., 2023), and one closed-source API model, **GPT-4o-mini** (Hurst et al., 2024). All models are evaluated under two decoding regimes: (1) *greedy decoding* with temperature $= 0$ for deterministic inference, and (2) *sampling-based decoding* with temperature $= 0.7$ for diversity-driven UQ, where $N{=}10$ samples are drawn per input. All other decoding hyperparameters follow the LM-Polygraph protocol (Vashurin et al., 2025a).

Figure 3: HCC(ΔU) on CLEVR-Existence for Qwen-VL. Rows list UQ estimators; columns denote modality-specific perturbations (visual in blue, textual in orange). Cells show the point-biserial correlation between hallucination flips and uncertainty change—higher (warmer) is better; negative values share the zero color. Corresponding results for LLaVA and GPT-4o-mini are provided in the appendix.

Uncertainty Quantification Methods. We evaluate nine UQ methods spanning white-box and black-box approaches. **White-box** methods (*Mean Token Entropy*, *Perplexity*, *Maximum Sequence Probability*, *PMI*, *Semantic Entropy*, *p(True)*) compute uncertainty directly from model logits or token probabilities, with Semantic Entropy additionally clustering semantically similar outputs and p(True) eliciting reflexive confidence scores. **Black-box** methods (*Lexical Similarity*, *DegMat*, *LUQ*) estimate uncertainty from the variability and structure of $N{=}10$ sampled responses.

4.2 Experiment Results

RQ1. UQ methods show limited robustness across uncertainty sources and VLM architectures. Table 2 reveals (i) strong architecture dependence: Qwen-VL and GPT-4o-mini generally enable stronger UQ discrimination. Averaged across methods, Qwen-VL attains AUROC values of 0.73 for image uncertainty, 0.74 for text uncertainty, and 0.71 for cross-modal uncertainty, while GPT-4o-mini reaches 0.62, 0.71, and 0.66, respectively. By contrast, Kosmos-2 and LLaVA stay below 0.60 on most averaged settings, suggesting that UQ performance depends strongly on the underlying VLM. Table 2 further reveals (ii) strong sensitivity to modality-specific uncertainty: the same estimator does not perform equally well across image, text, and cross-modal settings. For example, Semantic Entropy is among the strongest methods

Table 2: UQ discrimination results on modality-specific uncertainty subsets. Each cell reports AUROC for discriminating uncertain vs. clean samples under visual (Image), textual (Text), and cross-modal (Cross) conditions. Higher values indicate better discrimination. "–" denotes results that are unavailable due to missing token-level statistics required by the corresponding UQ methods. The rightmost column reports the mean AUROC of each method across all available settings, while the bottom row reports the mean AUROC of each setting across all available methods.

Estimator	**Kosmos-2**			**LLaVA**			**Qwen-VL**			**GPT-4o-mini**			**Avg.**
	Image	Text	Cross	Image	Text	Cross	Image	Text	Cross	Image	Text	Cross	
MSP	0.40	0.58	0.74	0.51	0.54	0.55	0.84	0.88	0.83	0.70	0.78	0.77	0.68
Perplexity	0.48	0.61	0.58	0.50	0.44	0.48	0.80	0.81	0.79	0.73	0.76	0.75	0.64
SemanticEntropy	0.60	0.70	0.77	0.63	0.73	0.62	0.85	0.89	0.81	0.56	0.71	0.62	0.71
MeanTokenEntropy	0.52	0.62	0.52	0.56	0.52	0.52	0.81	0.81	0.79	–	–	–	0.63
DegMat	0.58	0.60	0.48	0.69	0.61	0.64	0.73	0.70	0.60	0.56	0.70	0.64	0.63
LexSim	0.62	0.59	0.38	0.72	0.69	0.64	0.85	0.87	0.76	0.70	0.74	0.72	0.69
LUQ	0.52	0.55	0.46	0.60	0.42	0.51	0.61	0.63	0.54	0.46	0.57	0.47	0.53
PMI	0.60	0.48	0.36	0.61	0.58	0.51	0.79	0.82	0.79	–	–	–	0.62
PTrue	0.65	0.50	0.68	0.49	0.55	0.46	0.27	0.26	0.44	–	–	–	0.48
Avg.	0.55	0.58	0.55	0.59	0.56	0.55	0.73	0.74	0.71	0.62	0.71	0.66	0.63

Table 3: UQ discrimination results on the Grounded Cross-Modality Ambiguity (Cross-Modality–AMB) subset. Higher values indicate better discrimination. "–" denotes closed-source models that cannot be evaluated due to lack of access to model-internal information.

Estimator	**Kosmos-2**		**LLaVA**		**Qwen-VL**		**GPT-4o-mini**	
	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1
MSP	**0.64**	0.63	0.58	0.62	0.69	0.68	0.62	0.63
Perplexity	0.60	0.62	0.57	**0.64**	**0.70**	0.69	0.64	0.64
SE	0.57	0.62	0.54	0.60	0.68	0.66	**0.66**	0.66
MeanTokenEntropy	0.56	0.63	0.58	0.64	0.70	0.69	–	–
DegMat	0.57	0.62	0.58	0.62	0.63	0.62	0.64	0.63
LexSim	0.52	0.62	0.56	0.62	0.61	0.64	**0.66**	**0.66**
LUQ	0.55	0.63	**0.61**	0.63	0.59	0.62	0.61	0.64
PMI	0.46	0.36	0.52	0.60	0.62	0.63	–	–
PTrue	0.57	0.62	0.46	0.61	0.45	0.60	–	–

Table 4: F1 for FocusAmbiguity (Cross-Modality–AMB) detection. Columns compare the strongest uncertainty estimator from Table 3 ("UQ-best") with the zero-shot prompting strategies used in (Chen et al., 2025).

Model	UQ-best	ZS	ZS-CoT	ZS-ECoT
Qwen-VL	**0.69**	0.63	0.63	0.62
GPT-4o-mini	**0.66**	0.62	0.63	0.64

overall, with an average AUROC of 0.71. Still, its performance varies notably across models and uncertainty types. Overall, different UQ methods exhibit distinct performance profiles across models and modality-specific uncertainty settings, and no single method remains consistently robust across all settings.

RQ2. Modality-specific uncertainty induces hallucinations, but UQ methods do not reliably reflect the resulting risk. By applying our synthetic perturbation pipeline to the hallucination-focused subsets to generate counterfactual samples, we observe two key findings. First, modality-specific uncertainty can induce substantial hallucinations, with attribute and existence questions being the most consistently affected. As shown in Table 5, visual uncertainty is particularly damaging; for example, visual blur alone causes a 64.7% hallucination rate on attribute tasks for Qwen-VL. Second, these UQ methods do not reliably capture the resulting hallucination risk. Figure 3 shows that HCC(ΔU) is generally weak for Qwen-VL. Notably, UQ methods exhibit low HCC for visual blur despite its strong hallucination effect, while sometimes responding more strongly to lower-impact perturbations such as typos. Results for other models are provided in the appendix. Overall, this indicates a weak alignment between uncertainty estimates and uncertainty-induced hallucinations.

RQ3. UQ methods retain utility for VLM tasks, and ensemble aggregation can further enhance performance. On the grounded ambiguity benchmark VQ-FocusAmbiguity, these UQ methods remain practically useful for ambiguity detection: the best UQ estimator achieves F1 of 0.69 on Qwen-VL and 0.66 on GPT-4o-mini, outperforming zero-shot, zero-shot CoT, and zero-shot ECoT prompting methods in both cases (Table 4). This shows that, despite their limited robustness under modality-specific and architecture-dependent uncertainty, LLM-derived UQ methods can still provide strong signals for real VLM tasks such as ambiguity detection. Moreover, since single-estimator performance is strongly method- and model-dependent, it is natural to ask whether different estimators can be combined to provide a stronger practical baseline.

Uncertainty Type	LLaVA (Hallucination Rate %)				Qwen (Hallucination Rate %)				GPT-4o-mini (Hallucination Rate %)			
	Attr.	Count.	Rel.	Exist.	Attr.	Count.	Rel.	Exist.	Attr.	Count.	Rel.	Exist.
Visual Uncertainty												
Blur	12.7	4.0	8.7	26.0	64.7	12.7	37.3	50.7	25.3	1.3	28.0	34.7
Brightness (Dark)	4.0	4.0	6.0	6.7	9.3	6.0	6.0	14.7	16.0	1.3	26.0	12.0
Brightness (Bright)	8.0	5.3	11.3	12.7	26.0	10.0	11.3	22.7	17.3	0.0	20.0	20.7
Cutout	7.3	2.7	12.0	6.7	24.7	12.0	12.0	18.0	20.0	1.3	27.3	27.3
Noise	10.0	6.0	1.3	14.0	33.3	12.7	14.0	22.0	16.7	1.3	24.0	30.0
Pixelate	8.0	5.3	1.3	12.0	11.3	9.3	8.0	6.0	12.0	1.3	18.7	16.7
Salt & Pepper	8.7	6.0	1.3	12.7	46.7	12.7	16.7	32.7	16.7	1.3	20.7	26.7
Solarize	13.3	6.7	5.3	23.3	52.0	10.0	15.3	45.3	30.7	0.7	14.0	38.7
Textual Uncertainty												
Dropwords	11.3	4.7	5.3	14.7	8.7	6.0	11.3	6.7	11.3	1.3	15.3	17.3
INV	24.0	9.3	8.0	24.0	44.0	12.0	29.3	30.0	23.3	1.3	22.7	30.7
SBJ	22.0	10.0	22.0	58.0	20.7	12.7	30.0	58.0	27.3	1.3	28.0	60.0
Shuffle	16.0	5.3	12.7	18.0	4.0	3.3	6.0	0.7	6.7	1.3	13.3	7.3
Typos	22.7	8.7	21.3	37.3	7.3	8.0	15.3	27.3	7.3	1.3	21.3	22.0

Table 5: Hallucination Rate (%) across CLEVR tasks with representative perturbations. Perturbations are calibrated to induce modality-specific uncertainty via selected levels. Hallucination rate is the percentage of items that are correct on the clean sample but become incorrect on its perturbed counterpart. We do not evaluate *cross-modality uncertainty* here, as such perturbations would alter the original VQA ground truth and invalidate direct comparison.

Method	Avg. AUC	# Wins
Avg. Single	0.63	–
Best Single	0.75	–
Simple Calibrated Ensemble	**0.76**	**12/16**
Oracle Perfect	0.99	16/16

Table 6: Summary of ensemble results across 16 model–attack scenarios. # Wins counts the number of scenarios where a method outperforms Best Single. Oracle Perfect is an optimistic upper bound. Detailed ensemble definitions are provided in the Appendix.

RQ4. Complementary UQ signals can be synergized into a stronger and more practical uncertainty baseline. Following RQ3, we ask whether the useful uncertainty information captured by individual estimators can be combined more effectively. We find that the answer is yes, but only when estimators are selected and aligned carefully. To quantify the available headroom, we first define **Oracle Perfect**, an optimistic sample-wise upper bound that has access to labels and, for each sample, effectively selects the estimator output that best matches the true label. Its near-perfect AUC across scenarios shows that substantial complementary information is already present across estimators, even though no single method is consistently optimal.

Table 6 summarizes the main practical comparison. Here, **Avg. Single** reports the mean performance across individual UQ estimators, while **Best Single** denotes the strongest individual estimator in each scenario and therefore serves as an optimistic single-method reference. Our best practical method, a **simple calibrated ensemble**, uses a small labeled calibration split to align estimator directions and combine their normalized scores. It improves over both the average single-estimator level and the optimistic best-single baseline, reaching an average AUC of 0.76 and outperforming Best Single in 12 of the 16 model–attack scenarios. Overall, these results suggest that current UQ estimators contain useful but distributed information: they are informative individually, strongly complementary in principle, and can yield a stronger practical uncertainty baseline when combined appropriately. Detailed ensemble definitions and full results are deferred to the Appendix D.

However, perturbation strength must be carefully calibrated as shown in Appendix B.4, edits that are too subtle yield no usable signal, while overly strong edits can collapse output diversity.

5 Conclusion

We introduced **VLM-UQBench**, a benchmark for modality-specific and cross-modal uncertainty in vision–language models. Our experiments show that UQ methods originally developed for LLMs transfer imperfectly to VLMs: they are highly sensitive to both uncertainty source and model architecture, and often fail to reliably reflect hallucination risk. Still, these methods remain useful for practical tasks such as ambiguity detection, and simple calibrated ensembles can further strengthen performance by combining complementary signals. We hope **VLM-UQBench** provides a useful testbed for developing more robust uncertainty estimation methods for multimodal models.

Limitations

VLM-UQBench is intended as a diagnostic benchmark for source-aware uncertainty in VLMs, but it has several limitations. First, the human-curated core benchmark is deliberately small. The VizWiz-based modality-specific subsets provide clean and interpretable supervision, but their scale remains limited, and the final labels still depend on expert judgment despite cross-checking and manual review. As a result, the benchmark is better suited for controlled analysis than for drawing population-level conclusions about all real-world multimodal uncertainty.

Second, part of the benchmark relies on controlled or synthetic constructions. Our perturbation pipeline is designed to inject visual, textual, and cross-modal uncertainty while preserving the underlying task semantics as much as possible, and the CLEVR-based hallucination subsets provide precise oracle answers. However, such settings cannot fully capture the richness and variability of naturally occurring uncertainty in open-world multimodal data. The benchmark should therefore be viewed as a structured stress test rather than a complete proxy for deployment conditions.

Third, perturbation strengths are calibrated heuristically. Although we manually tune intensities to avoid perturbations that are either too weak to matter or so strong that they destroy the task, there is no universal notion of matched difficulty across perturbation families. We therefore do not interpret differences across perturbation types as a strict ranking of which uncertainty source is intrinsically harder.

Finally, our experiments mainly study UQ methods originally developed for LLMs, since these are the methods most commonly transferred to VLMs in current practice. While this focus matches the main motivation of our benchmark, VLM-UQBench itself is not limited to LLM-based UQ and can also be used to evaluate uncertainty methods designed specifically for VLMs. Exploring such methods, as well as broader VLM tasks beyond open-ended VQA, remains important future work.

Ethical Considerations

This work is motivated by safety and reliability. By studying whether uncertainty estimates in VLMs are sensitive to source-specific ambiguity and aligned with hallucination risk, VLM-UQBench is intended to support the development of models that can better abstain, defer, or request clarification in high-stakes settings.

At the same time, the benchmark and perturbation pipeline could be misused to stress-test models in ways that resemble adversarial manipulation. We therefore position the pipeline as a diagnostic evaluation tool rather than an attack framework, and we encourage future use in settings that improve robustness, calibration, and safeguards rather than degrade deployed systems.

Our benchmark also inherits limitations and potential biases from its source datasets. VizWiz reflects assistive, real-world image capture conditions, while CLEVR provides synthetic scenes with strong structural control; neither dataset is fully representative of all users, domains, or deployment environments. As a result, benchmark performance should not be interpreted as a complete measure of fairness or real-world safety.

Finally, uncertainty estimates can influence downstream decisions such as abstention or escalation. Poorly calibrated uncertainty may create false reassurance or excessive conservatism. We therefore do not recommend using the benchmark as the sole basis for deployment claims; instead, it should be combined with domain-specific evaluation, human oversight, and broader robustness and safety testing.

References

Nilavra Bhattacharya, Qing Li, and Danna Gurari. 2019. Why does a visual question have different answers? In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4271–4280.

Khyathi Raghavi Chandu, Linjie Li, Anas Awadalla, Ximing Lu, Jae Sung Park, Jack Hessel, Lijuan Wang, and Yejin Choi. 2024. Certainly uncertain: A benchmark and metric for multimodal epistemic and aleatoric awareness. *arXiv preprint arXiv:2407.01942*.

Chongyan Chen, Yu-Yun Tseng, Zhuoheng Li, Anush Venkatesh, and Danna Gurari. 2025. Acknowledging focus ambiguity in visual questions. *arXiv preprint arXiv:2501.02201*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. 2019. Autoaugment:

Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123.

Terrance DeVries and Graham W Taylor. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Yangyang Guo, Fangkai Jiao, Zhiqi Shen, Liqiang Nie, and Mohan Kankanhalli. 2024. Unk-vqa: A dataset and a probe into the abstention ability of multi-modal large models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. 2023. Decomposing uncertainty for large language models through input clarification ensembling. *arXiv preprint arXiv:2311.08718*.

Andrew G Howard. 2013. Some improvements on deep convolutional neural network based image classification. *arXiv preprint arXiv:1312.5402*.

Ling Huang, Su Ruan, Yucheng Xing, and Mengling Feng. 2024a. A review of uncertainty quantification in medical image analysis: Probabilistic and non-probabilistic methods. *Medical Image Analysis*, 97:103223.

Xinmeng Huang, Shuo Li, Mengxin Yu, Matteo Sesia, Hamed Hassani, Insup Lee, Osbert Bastani, and Edgar Dobriban. 2024b. Uncertainty in language models: Assessment through rank-calibration. *arXiv preprint arXiv:2404.03163*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Katie Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew G Wilson. 2024. Large language models must be taught to know what they don't know. *Advances in Neural Information Processing Systems*, 37:85932–85972.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J Bell. 2025. Abstentionbench: Reasoning llms fail on unanswerable questions. *arXiv preprint arXiv:2506.09038*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2022. Clam: Selective clarification for ambiguous questions with generative language models. *arXiv preprint arXiv:2212.07769*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Gregory Kang Ruey Lau, Hieu Dao, and Bryan Kian Hsiang Low. 2025. Uncertainty quantification for mllms. In *ICLR Workshop: Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambigqa: Answering ambiguous open-domain questions. *arXiv preprint arXiv:2004.10645*.

Atsuyuki Miyai, Jingkang Yang, Jingyang Zhang, Yifei Ming, Qing Yu, Go Irie, Yixuan Li, Hai Li, Ziwei Liu, and Kiyoharu Aizawa. 2025. Unsolvable problem detection: Robust understanding evaluation for large multimodal models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Malik Sajjad Ahmed Nadeem, Jean-Daniel Zucker, and Blaise Hanczar. 2009. Accuracy-rejection curves (arcs) for comparing classification methods with a reject option. In *Machine Learning in Systems Biology*, pages 65–81. PMLR.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Allen Z Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, and 1 others. 2023. Robots that ask for help: Uncertainty alignment for large language model planners. *arXiv preprint arXiv:2307.01928*.

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

Junya Takayama and Yuki Arase. 2019. Relevant and informative response generation using pointwise mutual information. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 133–138.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. 2025a. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, and 1 others. 2025b. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248.

Aishwarya Venkataramanan, Paul Bodesheim, and Joachim Denzler. 2025. Probabilistic embeddings for frozen vision-language models: Uncertainty quantification with gaussian process latent variable models. *arXiv preprint arXiv:2505.05163*.

Chenyu Wang, Weichao Zhou, Shantanu Ghosh, Kayhan Batmanghelich, and Wenchao Li. 2025. Semantic consistency-based uncertainty quantification for factuality in radiology report generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, and 1 others. 2024a. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. *arXiv preprint arXiv:2103.15025*.

Ziran Yang, Shibo Hao, Hao Sun, Lai Jiang, Qiyue Gao, Yian Ma, and Zhiting Hu. 2025. Understanding the sources of uncertainty for large language and multimodal models. In *ICLR Workshop: Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI*.

Runpeng Yu, Weihao Yu, and Xinchao Wang. 2024. Api: Attention prompting on image for large vision-language models.

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. 2024a. Luq: Long-text uncertainty quantification for llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5244–5262, Singapore. Association for Computational Linguistics.

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

Appendix Table of Contents

We provide all additional details for our paper in the following sections.

- Appendix A details the data construction process, describing how modality-specific subsets are derived from VizWiz disagreement annotations and how the hallucination-focused subset is generated from CLEVR.
- Appendix B provides comprehensive details on the uncertainty perturbation pipeline, covering the specific algorithms for text, visual, and cross-modality perturbations, as well as the strength calibration protocol.
- Appendix C presents formal definitions for all uncertainty quantification estimators used in the benchmark and provides additional experimental results, including heatmaps for Uncertainty Reflection Rate (URR) and Hallucination Consistency Coefficient (HCC).
- Appendix D provides additional details on ensemble baselines, including oracle upper bounds, practical few-shot ensembles, and comparisons to unsupervised and fully supervised baselines.

A Data Creation Details

In this section, we detail how we construct the datasets used in our benchmark. First, we build modality-specific subsets from VizWiz using human disagreement annotations. Then, we construct a hallucination-focused subset from CLEVR (Johnson et al., 2017) using its scene graphs and question templates.

A.1 Data Creation for Modality-Specific Subsets From VizWiz

We leverage the crowdsourced disagreement annotations from VizWiz (Bhattacharya et al., 2019), where for each question–image pair five annotators indicate one or more reasons for answer disagreement from a fixed taxonomy: Low Quality Image (LQI), Answer Not Present in Image (IVE), Invalid Question (INV), Difficult Question (DFF), Ambiguous Question (AMB), Subjective Question (SBJ), Synonymous Answers (SYN), Granularity Mismatch (GRN), Spam (SPM), and Other (OTH). We follow the original formulation and category definitions (Bhattacharya et al., 2019). For each pair, we aggregate these annotations into a count vector (0–5 for each reason) and use the resulting per-reason counts as the basis for deriving modality-specific subsets.

Clean subset. Pairs satisfying at most one annotator selects LQI and, for all question-related reasons (IVE, INV, DFF, AMB, SBJ, SYN), no more than two annotators select the reason.

Visual-uncertainty subset. Pairs with strong evidence of image degradation but clear questions: at least four annotators select LQI, while all question-related reasons (IVE, INV, DFF, AMB, SBJ, SYN) are selected by at most two annotators.

Textual-uncertainty subset. Pairs with reliable visual input but noisy questions: at most one annotator selects LQI, at least three annotators select either INV or SBJ (invalid or subjective question), and neither AMB nor IVE receives unanimous support.

Cross-modality uncertainty subset. Pairs characterized by strong question–image mismatch: all annotators select either AMB or IVE, indicating ambiguity or missing visual evidence.

These thresholds yield substantial subsets for each modality, providing reliable coverage of visual, textual, and cross-modality uncertainty. To further ensure dataset quality, three expert annotators with research experience in vision-and-language manually review the automatically selected examples and remove cases with annotation errors or unclear categorization.

A.2 Data Creation for Hallucination-Focused Subset From CLEVR

Question Categories. We define four question categories, each aligned with a different hallucination type and grounded in CLEVR's scene graphs. **Attribute** questions ask for an intrinsic property of a specific object (color, size, material, or shape). **Existence** questions ask whether an object with certain attributes is present in the image. **Relation** questions ask about spatial relationships between objects (left/right/front/behind) or about objects located in a given direction. **Counting** questions ask for the number of objects satisfying given attribute constraints. For all categories, the CLEVR scene graphs provide exact ground truth, and any discrepancy between the model's answer and this oracle is interpreted as the corresponding type of hallucination.

Algorithm 1 CLEVR hallucination subset construction

Require: CLEVR scenes $\mathcal{S}$ with object attributes and relations;
question types $\mathcal{T} = \{\text{Attr}, \text{Exist}, \text{Rel}, \text{Count}\}$;
templates $\{\tau_t\}_{t \in \mathcal{T}}$; target per type N

Ensure: Hallucination subset $\mathcal{D}$ of QA pairs

```
D ← ∅
count[t] ← 0 for all t ∈ T
while ∃ t ∈ T such that count[t] < N do
    s ~ S {sample scene}
    for all t ∈ T with count[t] < N do
        e ← SelectSceneElements(s, t)
        q ← FillTemplate(τ_t, e)
        a ← OracleAnswer(s, q, t)
        if Valid(q, a) then
            D ← D ∪ {(image_id(s), q, a, t)}
            count[t] ← count[t] + 1
        end if
    end for
end while
return D
```

Subset Creation. We construct a hallucination-focused subset from CLEVR by using its scene annotations as an oracle. For each split (train/val/test), we load the corresponding CLEVR scene file, which lists all objects (with color, size, material, shape) and all pairwise spatial relations (left, right, front, behind). For each question type t (attribute, existence, relation, counting), we define a small set of templates τ_t and a type-specific routine $\textsc{SelectSceneElements}(s, t)$ that selects from the CLEVR scene graph of a scene s the objects, attributes, and/or relations required for that type. Given these elements e, we instantiate a natural language question via $\textsc{FillTemplate}(\tau_t, e)$ and compute the ground-truth answer with $\textsc{OracleAnswer}(s, q, t)$, again using only the scene graph. A validity check $\textsc{Valid}(q, a)$ ensures that queries are well-formed and unambiguous (e.g., "negative" existence questions truly refer to absent object types). We repeat this procedure across scenes until a fixed target number of questions per type is reached, yielding a balanced subset where every question–answer pair is fully determined by the CLEVR oracle. The overall generation pipeline is summarized in Algorithm 1.

B Uncertainty Perturbation Details

In this section, we describe how we inject controlled perturbations into the text, image, and cross-modal components of VLM-UQBench. Text perturbations use prompt-based rewrites (INV, SBJ) and local lexical changes (typos, dropwords, shuffle) to add textual uncertainty, while visual perturbations apply pixel-level and geometric transformations to add image uncertainty. Cross-modality perturbations either rewrite the question conditioned on the image (AMB, IVE) or use CLIP-based attention (Yu et al., 2024) masks to perturb localized image–text alignments. We also describe a heuristic calibration procedure for choosing perturbation strengths that introduce meaningful uncertainty while preserving core semantics. Implementation details for each perturbation are provided below for reproducibility.

B.1 Text Perturbation

Text perturbations operate on the question component of a VQA input to inject controlled textual uncertainty. The specific perturbation types we use are detailed below.

Prompt-based Semantic Perturbations. We leverage `gpt-4o` (Hurst et al., 2024)) to introduce subtle uncertainty while preserving the semantic content of the question. Specifically:

- **INV**: rewrites introduce slight incoherence or underspecification.
- **SBJ**: rewrites inject subjective phrasing (e.g., "Do you think...").

These rewrites are generated with temperature sampling ($T = 0.7$) and retry logic to ensure deterministic coverage.

Prompt for INV

```
You are an expert at stress-testing Visual
Question Answering tasks.
Rewrite Visual Question Answering questions to
add text uncertainty while keeping them natural
and close to the original meaning.
Type of uncertainty:
INV: rewrite the question into a natural
but non-visual or semantically invalid form.
Goal: rewrite each question into INV form with
minimal, natural change.
```

Prompt for SBJ

```
You are an expert at stress-testing Visual
Question Answering tasks.
Rewrite Visual Question Answering questions to
add text uncertainty while keeping them natural
and close to the original meaning.

Type of uncertainty:
SBJ: make the question subjective — add mild
opinion or emotion while keeping the same
visual topic.

Keep rewrites fluent, minimal, and human-like.
Do not add typos or random noise.

SBJ examples

Image: (photo of a flower)
Original: What color is the flower?
Rewritten (SBJ): Don't you think the color of
flower looks beautiful?

Image: (photo of dessert)
Original: What is on top of the cake?
Rewritten (SBJ): Would you enjoy eating this
cake?

Goal: rewrite each question into SBJ form with
minimal, natural change.
```

Lexical Perturbations. We implement lightweight stochastic functions to distort the surface form of text:

- **Typos (minor/strong)**: characters are randomly replaced with neighboring keyboard keys at probability $p \in \{0.03, 0.15\}$.
- **Dropwords (minor/strong)**: tokens are randomly deleted with probability $p \in \{0.05, 0.25\}$.
- **Shuffle (minor/strong)**: short phrases (split by punctuation or conjunctions) are swapped $k \in \{1, 2\}$ times.

Each augmentation ensures syntactic validity (ending with a question mark) and retains a copy of the original question for reference.

B.2 Visual Perturbation

Visual perturbations operate on the image component of a VQA input to inject controlled image uncertainty. For each image, we apply pixel-level and geometric perturbations. The specific perturbation types we use are detailed below.

- **Blur**: Applies a Gaussian blur with varying radius to reduce spatial high-frequency details, mimicking out-of-focus or motion-induced blur (Hendrycks and Dietterich, 2019). Implemented via convolution with a Gaussian kernel of standard deviation σ (strength controlled by the blur radius).
- **Additive Gaussian noise**: Adds zero-mean Gaussian noise to each pixel, controlled by standard deviation σ. This simulates sensor noise or low-light imaging conditions (Hendrycks and Dietterich, 2019). The noise level σ acts as the perturbation strength.
- **Brightness/Contrast/Saturation/Sharpness**: Scales image attributes multiplicatively using coefficients sampled within predefined ranges, following photometric distortions from (Howard, 2013). Adjustments mimic illumination or camera exposure variations, with the scaling coefficient serving as the perturbation strength.
- **Pixelate**: Downsamples the image to a coarse grid (e.g., $k \times k$) and resizes it back to the original resolution using nearest-neighbor interpolation. Mimics low-resolution capture or privacy-preserving filtering. The grid resolution k determines the perturbation strength.
- **Solarization**: Inverts pixel intensities above a fixed threshold T, i.e., $I' = 255 - I$ if $I > T$, creating non-linear tonal inversions (Cubuk et al., 2019). The threshold parameter T controls the strength level.
- **Cutout**: Randomly masks a square region of the image, replacing it with a constant value (typically gray). Encourages spatial invariance and robustness to occlusion (DeVries and Taylor, 2017). The cutout ratio is used as the perturbation strength.
- **Salt-and-pepper noise**: Randomly replaces a small fraction p of pixels with 0 (black) or 255 (white), modeling impulse noise from degraded transmission or sensor dropout (DeVries and Taylor, 2017). The corruption probability p controls the perturbation strength.

B.3 Cross-modality Perturbation

Cross-modality perturbations operate on the interaction between the image and the question to inject controlled cross-modal uncertainty. We use two prompt-based transformations (AMB and IVE) that rewrite the question given the image, and a CLIP-based masking scheme that perturbs localized image–text alignments. The specific procedures are detailed below.

System Prompt for AMB/IVE Generation

```
You are an expert at stress-testing visual QA tasks. Every input includes a clean Base Question plus an accompanying image.
Follow this procedure silently (no scratch work in the response):

1. Analysis: Carefully inspect the image and the Base Question. Summarize only the relevant visible objects (with counts if helpful), key attributes, and salient relations directly tied to the Base Question.

2. AMB (Ambiguity injection):
- Make only minimal semantic tweaks to the Base Question so it can now be grounded to multiple distinct regions/objects seen in this image (e.g., drop a distinguishing attribute, broaden a relation, merge similar entities).
- Keep the question natural, image-specific, and requiring a non-yes/no answer.
- Provide short plausible answer candidates strictly derived from what is visibly present; DO NOT list franchise/options not visible. Only produce AMB if at least two distinct visible candidates exist after minimal generalization; otherwise set AMB to null.

3. IVE (Insufficient Visual Evidence):
- Ask about a closely related object/attribute that would normally resolve the Base Question but is hidden, cropped out, or otherwise unverifiable in the current view.
- The reason must be exactly one of: Missing Interaction Target, Implied Missing Associate, Out of Frame. Return null if none fit.

4. Output ONLY strict JSON in this schema (no additional keys, comments, or prose):
{
"analysis": "Concise scene + question analysis that motivates the variants",
"AMB": {"variant_question": "...", "plausible_answers": ["candidate A", "candidate B", "candidate C"]} OR null,
"IVE": {"variant_question": "...", "reason_unanswerable": "Missing Interaction Target / Implied Missing Associate / Out of Frame"} OR null
}
```

Cross-modality masking via CLIP attention. To capture localized correspondences between textual and visual inputs, we employ a CLIP-based cross-attention mechanism (Yu et al., 2024). Given an input image and its paired text (e.g., caption or question), we extract per-patch attention weights from the CLIP image encoder conditioned on the text embedding.

Formally, for each image patch feature v_i and text token embedding t_j, CLIP computes the scaled dot-product attention:

$$a_{ij} = \text{softmax}\left(\frac{v_i \cdot t_j}{\sqrt{d}}\right),$$

where d denotes the feature dimension. We then aggregate attention across all text tokens to derive a per-patch relevance score:

$$m_i = \frac{1}{|T|}\sum_{j \in T} a_{ij}.$$

This yields an attention map that highlights spatial regions most semantically related to the text prompt. The map is normalized to $[0, 1]$ and smoothed using a Gaussian kernel, producing a continuous relevance mask M. Patches with higher attention values indicate stronger visual-text alignment. To use this mechanism as a perturbation, we selectively suppress regions according to the relevance mask: highly aligned patches are attenuated, altering the model's access to semantically critical spatial evidence. A scalar hyperparameter $\lambda \in [0, 1]$ controls the perturbation strength, and the modified image is computed as

$$\tilde{I} = (1 - \lambda M) \odot I,$$

providing a smooth continuum from mild masking (λ small) to aggressive removal of text-relevant regions (λ close to 1). Thus, CLIP-derived masks offer a principled, alignment-aware way to generate controllable cross-modality perturbations.

B.4 Perturbation Strength Calibration

As discussed in the main paper, selecting an appropriate intensity λ for synthetic perturbations is critical for valid uncertainty evaluation. If the perturbation is too weak, the model remains confident, yielding no useful uncertainty signal. Conversely, if the perturbation is too strong, the semantic content is destroyed, causing the model to correctly refuse to answer (e.g., "I cannot see anything") rather than exhibiting epistemic uncertainty.

To address this trade-off, we developed a custom *Augmentation Visualizer* interface to manually calibrate noise levels for each dataset and perturbation type.

Calibration Interface. As shown in Figure 4, 5, the interface provides a real-time, human-in-the-loop workflow for selecting λ. The tool features:

- **Batch Visualization:** Users can load random batches (e.g., $N = 150$) from the target

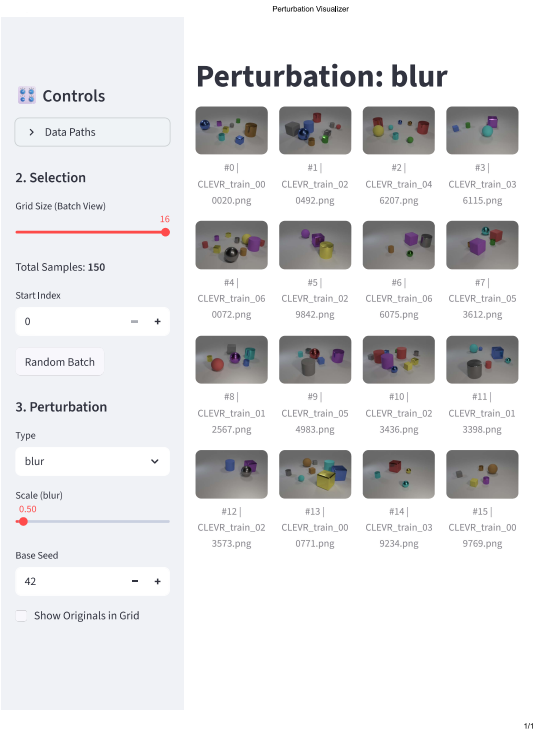

Figure 4: Calibration Interface (Too Weak). Our interactive tool allows for batch visualization of perturbation intensities. In this example, the selected intensity is insufficient to induce meaningful uncertainty, representing a trivial case where the perturbation has no effect.

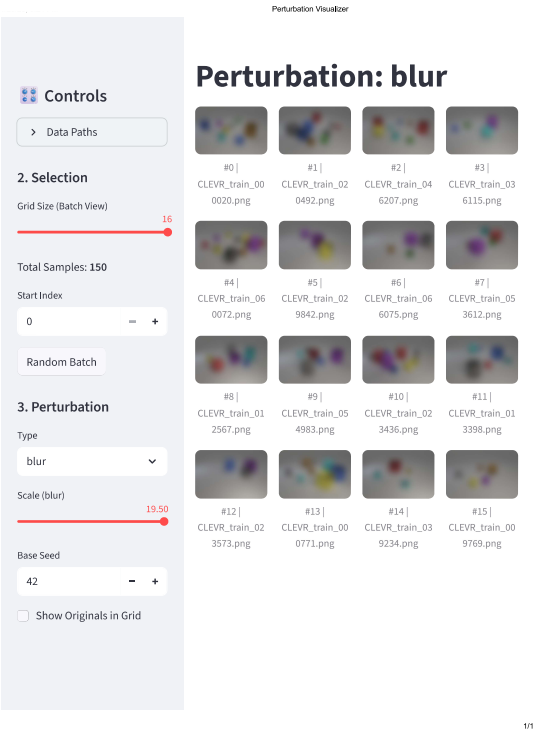

Figure 5: Calibration Interface (Too Strong). An example of catastrophic semantic destruction. At extreme intensities (e.g., blur scale 19.00), the image content becomes unrecognizable, leading the model to refuse to answer and rendering UQ metrics uninformative.

datasets (VizWiz, CLEVR, etc.) to ensure the selected intensity generalizes across different images.

- **Dynamic Control:** A slider allows for granular adjustment of the perturbation scale (e.g., blur radius, noise variance) with real-time visual feedback.
- **Grid Inspection:** Original and augmented images are displayed in a grid, allowing immediate comparison of semantic preservation.

Selection Protocol. We employed domain experts to select parameters within the "uncertainty Goldilocks zone." The criteria were defined as follows: the perturbation must be strong enough to introduce visible artifacts that could plausibly cause disagreement among annotators or models (avoiding the scenario in Figure 4), but weak enough that the core object or attribute remains identifiable by a human observer (avoiding the catastrophic failure shown in Figure 5). The final intensities reported reflect the calibrated midpoints derived from this process.

B.5 Perturbation Strength

In the main paper, we use a fixed set of calibrated visual perturbations and raw strengths: 10.00 for blur, 0.20 for brightness (dark), 4.00 for brightness (light), 0.20 for cutout, 50.00 for Gaussian noise, 5 for pixelation, 0.200 for salt-and-pepper noise, and 1 for solarization. For text perturbations, we use fixed probability-based strengths: typos at probability $p = 0.15$, drop-word perturbations at $p = 0.15$, and phrase shuffling with $k = 1$ swap per question. For cross-modality perturbations based on CLIP attention masks, we use a fixed masking coefficient $\lambda = 1.0$ as the default perturbation strength. This corresponds to applying the raw, unscaled attention map M from CLIP.

These values are selected to lie in a regime where the perturbation is neither trivial nor catastrophically strong. In the appendix, we additionally include nearby intensity levels that also fall within this calibrated range.

For prompt-based perturbations (INV, SBJ, AMB, IVE), we do not have a continuous raw strength parameter; instead, we calibrate their effective strength by manually tweaking the prompts

(and sampling temperature), treating them as discrete perturbation types rather than numeric intensity levels.

C Modality-Specific Uncertainty Analysis

In this section, we provide the exact definitions of all uncertainty quantification estimators evaluated in the main paper, together with additional results for Uncertainty Reflection Rate (URR) and Hallucination Consistency Coefficient (HCC). Below, we detail each UQ method used in our evaluation.

Mean Token Entropy (Fomicheva et al., 2020). Given a generated response $y = (y_1, \ldots, y_T)$, let $p_\theta(\cdot \mid y_{<t}, x)$ denote the softmax distribution over the vocabulary at step t. We compute the Mean Token Entropy as follows:

$$\mathcal{U}_{\text{MTE}}(x) = \frac{1}{T} \sum_{t=1}^{T} H(p_\theta(\cdot \mid y_{<t}, x)),$$

where

$$H(p_\theta(\cdot \mid y_{<t}, x)) = -\sum_{v} p_\theta(v \mid y_{<t}, x) \log p_\theta(v \mid y_{<t}, x)$$

is the entropy of the model's next-token predictive distribution over the vocabulary at step t. Intuitively, $\mathcal{U}_{\text{MTE}}(x)$ is large when these next-token distributions are diffuse (non-peaked) across positions in the sequence, indicating higher token-level predictive uncertainty.

Perplexity (Fomicheva et al., 2020). Perplexity quantifies uncertainty as the negative average log-probability of generated tokens:

$$U_{\text{PPL}}(\mathbf{y}, x) = -\frac{1}{T} \sum_{i=1}^{T} \log p(y_i \mid y_{<i}, x),$$

where x is the input, $\mathbf{y} = (y_1, y_2, \ldots, y_T)$ is the model-generated sequence of T tokens. Higher values indicate lower model confidence in the generated output.

Lexical Similarity (LeXSim) (Fomicheva et al., 2020). quantifies uncertainty as the negative mean pairwise similarity between sampled generations:

$$U_{\text{LexSim}}(\mathbf{x}) = -\frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} s(\mathbf{y}^i, \mathbf{y}^j),$$

where $\mathbf{x}$ is the input, $\{\mathbf{y}^1, \ldots, \mathbf{y}^N\}$ are N sampled generations from the model, and $s(\cdot, \cdot)$ is a similarity metric (e.g., ROUGE-L or BLEU). Higher values indicate greater diversity among samples, suggesting higher uncertainty.

Maximum Sequence Probability (MSP). Quantifies uncertainty using the joint probability of the full generated response $y = (y_1, \ldots, y_T)$. We compute

$$\mathcal{U}_{\text{MSP}}(x) = -\log p_\theta(y \mid x) = -\sum_{t=1}^{T} \log p_\theta\big(y_t \mid y_{<t}, x\big),$$

so larger values indicate lower model confidence thus high uncertainty.

Degree Matrix (DegMat) (Lin et al., 2024). Given an input x, we draw M responses $\{y^{(1)}, \ldots, y^{(M)}\}$, each a full sequence. We first build a similarity matrix $W \in [0,1]^{M \times M}$ with

$$W_{m,m'} = \text{sim}\big(y^{(m)}, y^{(m')}\big),$$

where $\text{sim}(\cdot, \cdot)$ is typically an NLI-based entailment score. We define the diagonal degree matrix $D \in R^{M \times M}$ by

$$D_{m,m'} = \begin{cases} \sum_{j=1}^{M} W_{m,j}, & m = m', \\ 0, & m \neq m'. \end{cases}$$

The DegMat uncertainty estimate is then

$$\mathcal{U}_{\text{DegMat}}(x) = \frac{\text{tr}(MI - D)}{M^2},$$

where higher values indicate lower agreement among sampled responses, suggesting higher uncertainty.

Long-text Uncertainty Quantification(LUQ) (Zhang et al., 2024a). Given an input x, we draw M stochastic responses $\{y^{(0)}, \ldots, y^{(M)}\}$. Each response $y^{(m)}$ is split into sentences $y^{(m)} = \{s_1^{(m)}, \ldots, s_{T_m}^{(m)}\}$. For any two responses $y^{(m)}$ and $y^{(m')}$, we compute the entailment-based similarity

$$S\Big(y^{(m)}, y^{(m')}\Big) = \frac{1}{T_m} \sum_{t=1}^{T_m} P_{\text{entail}}\Big(s_t^{(m)} \mid y^{(m')}\Big),$$

where P_{entail} is the NLI-based entailment probability. The confidence of response $y^{(m)}$ is

$$C(x, y^{(m)}) = \frac{1}{|R'| - 1} \sum_{\substack{m'=0 \\ m' \neq m}}^{M} S\Big(y^{(m)}, y^{(m')}\Big).$$

The LUQ uncertainty score is then

$$\mathcal{U}_{\mathrm{LUQ}}(x) = 1 - \frac{1}{|R'|} \sum_{m=0}^{M} C(x, y^{(m)}),$$

so higher values indicate greater semantic disagreement across sampled responses.

P(true) Score (Kadavath et al., 2022). Given an input x the model is asked to choose from a binary set {True, False}. The uncertainty score is computed as the negative log-probability assigned to the token corresponding to "True." Formally, if $p_\theta(\text{"True"} \mid q, a)$ is the model's predicted probability, the metric is:

$$\mathcal{U}_{\mathrm{PTrue}}(x, a) = -\log p_\theta(\text{"True"} \mid x).$$

Pointwise Mutual Information (PMI) (Takayama and Arase, 2019). Given input x and generated sequence $\mathbf{y} = (y_1, \ldots, y_T)$, PMI quantifies how much the input influences each token. For token t, the pointwise mutual information is $\mathrm{PMI}(y_t; x \mid y_{<t}) = \log p(y_t \mid y_{<t}, x) - \log p(y_t \mid y_{<t})$, comparing the conditional probability with input to the language model prior. The uncertainty is:

$$\mathcal{U}_{\mathrm{PMI}}(\mathbf{y}, x) = -\frac{1}{T} \sum_{t=1}^{T} \mathrm{PMI}(y_t; x \mid y_{<t}),$$

where higher values indicate lower input dependence, suggesting higher uncertainty.

C.1 Uncertainty Reflection Rate Results

Figures 6–13 present URR heatmaps on real-world VizWiz and VQAv2 data for Qwen-VL, LLaVA (Liu et al., 2023), Kosmos-2 (Peng et al., 2023), and GPT-4o-mini (Hurst et al., 2024) using our perturbation pipeline, showing how strongly each UQ method responds to image-, text-, and cross-modality perturbations; higher values (warmer colors) indicate better behavior, meaning the corresponding UQ method is more reflective of the injected modality-specific uncertainty.

C.2 Hallucination Consistency Coefficient Results

Figures 14–25 present HCC heatmaps on the hallucination-focused CLEVR subsets (Attribute, Counting, Existence, Relation) for GPT-4o-mini, LLaVA, and Qwen-VL under fine-grained image- and text-only perturbations. Higher values (warmer colors) indicate better behavior, meaning that the corresponding UQ method more reliably captures hallucinations induced by the given modality-specific perturbation.

C.3 Additional Cross-Model HCC Results for CLEVR-Existence

The main paper shows the CLEVR-Existence HCC(ΔU) results for Qwen-VL as a representative example. Here, we provide the corresponding results for LLaVA and GPT-4o-mini for completeness. As shown in Figures 26–28, similar to the Qwen-VL results discussed in the main paper, HCC(ΔU) remains generally weak across models. Notably, Figures 26 and 27 show that several UQ methods exhibit low HCC for visual blur despite its strong hallucination effect, while Figure 28 shows that GPT-4o-mini follows a similar overall trend, although the available estimators are more limited due to API constraints.

Overall, Figures 26–28 further support a weak alignment between uncertainty estimates and uncertainty-induced hallucinations across models. Although the exact magnitudes vary by architecture and perturbation type, the overall conclusion remains consistent: strong hallucination-inducing perturbations are not reliably accompanied by strong uncertainty increases.

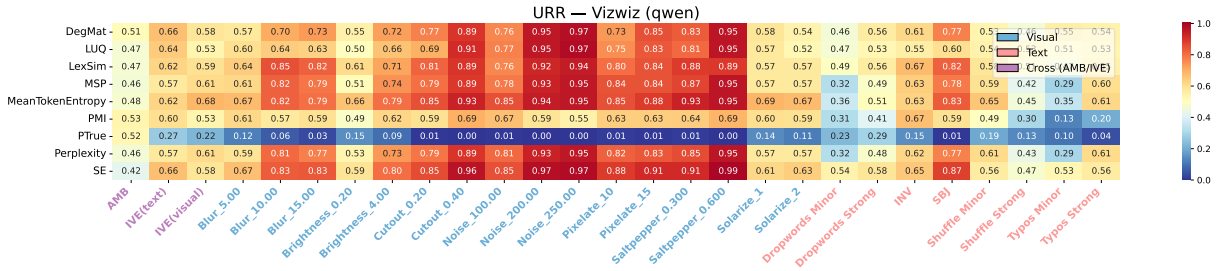

Figure 6: Qwen-VL, VizWiz. Uncertainty reflection rate under fine-grained image-, text-, and cross-modality perturbations on the VizWiz dataset using Qwen-VL.

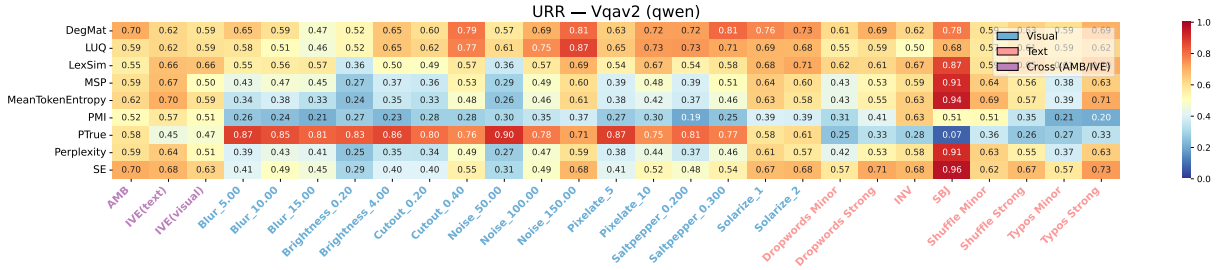

Figure 7: Qwen-VL, VQAv2. Uncertainty reflection rate under fine-grained image-, text-, and cross-modality perturbations on the VQAv2 dataset using Qwen-VL.

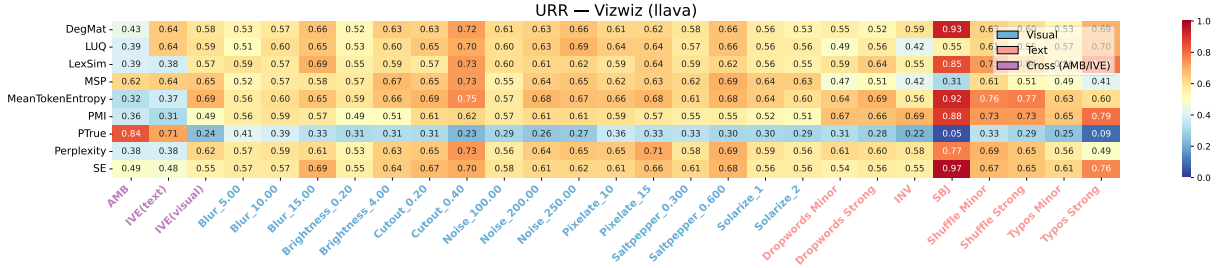

Figure 8: LLaVA, VizWiz. Uncertainty reflection rate under fine-grained image-, text-, and cross-modality perturbations on the VizWiz dataset using LLaVA.

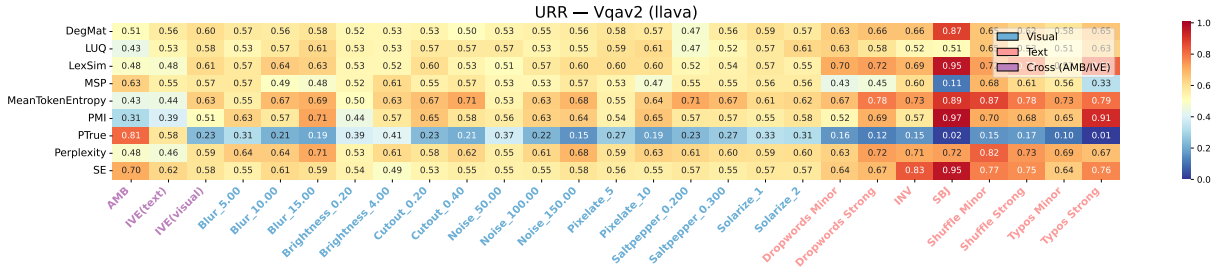

Figure 9: LLaVA, VQAv2. Uncertainty reflection rate under fine-grained image-, text-, and cross-modality perturbations on the VQAv2 dataset using LLaVA.

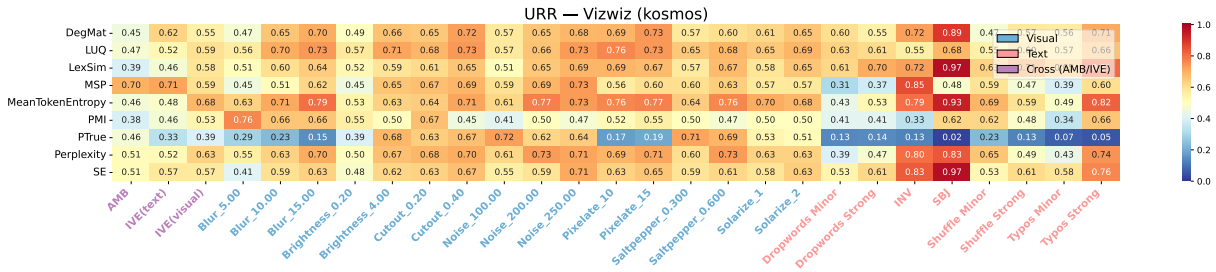

Figure 10: Kosmos, VizWiz. Uncertainty reflection rate under fine-grained image-, text-, and cross-modality perturbations on the VizWiz dataset using Kosmos.

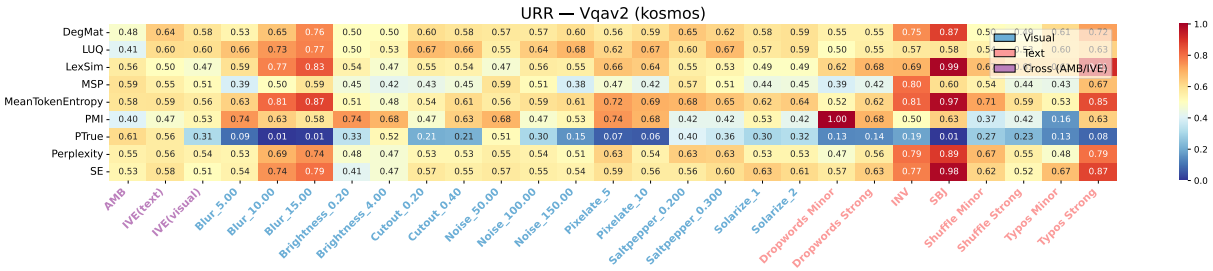

Figure 11: Kosmos, VQAv2. Uncertainty reflection rate under fine-grained image-, text-, and cross-modality perturbations on the VQAv2 dataset using Kosmos.

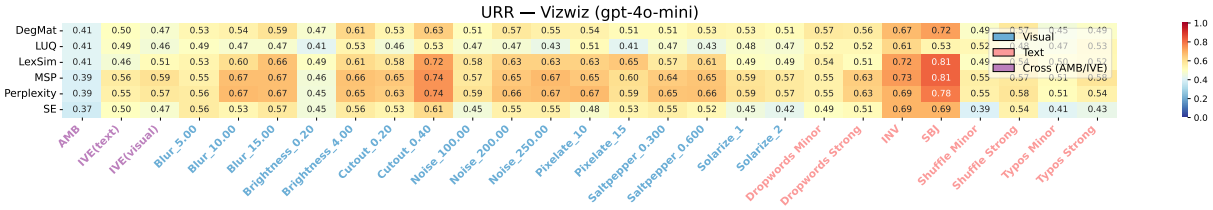

Figure 12: GPT-4o-mini, VizWiz. Uncertainty reflection rate under fine-grained image-, text-, and cross-modality perturbations on the VizWiz dataset using GPT-4o-mini.

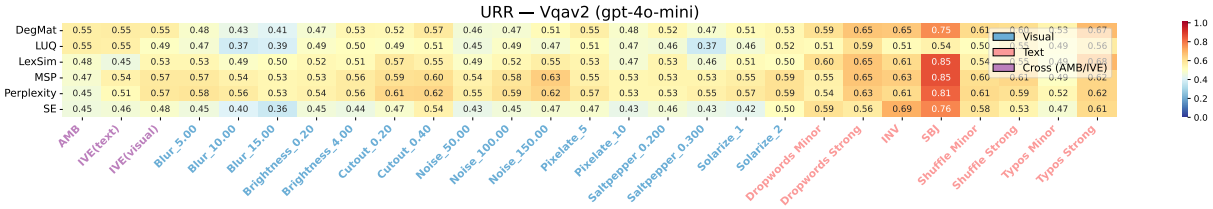

Figure 13: GPT-4o-mini, VQAv2. Uncertainty reflection rate under fine-grained image-, text-, and cross-modality perturbations on the VQAv2 dataset using GPT-4o-mini.

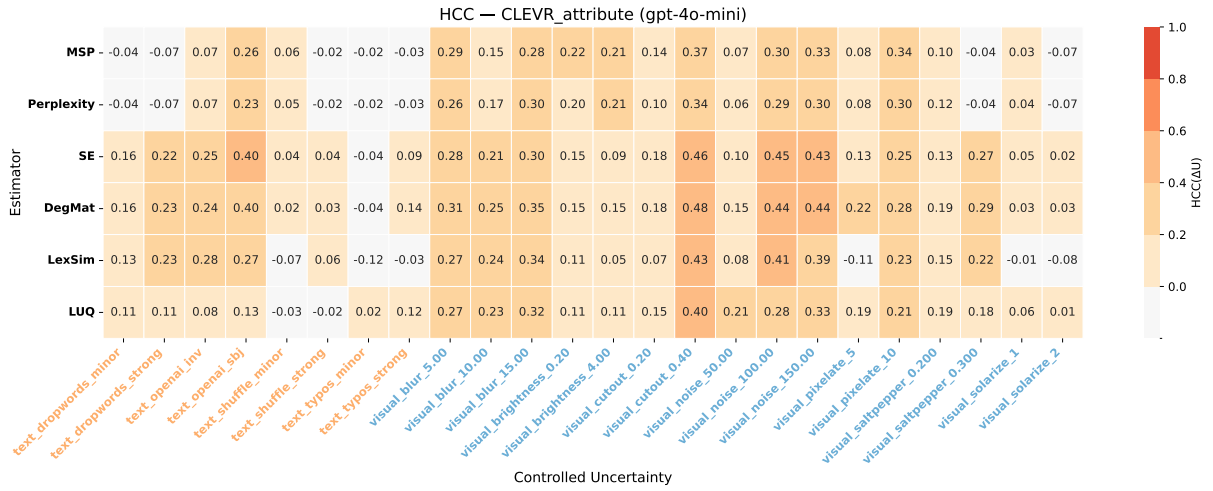

Figure 14: GPT-4o-mini, CLEVR-Attribute. Uncertainty reflection rate under fine-grained image-, text-, and cross-modality perturbations on CLEVR-Attribute using GPT-4o-mini.

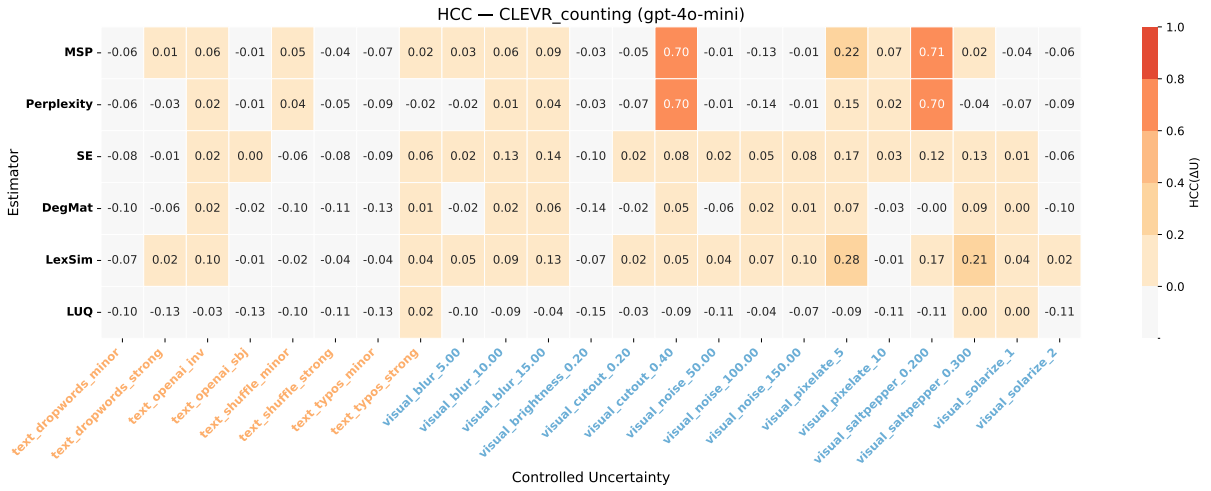

Figure 15: GPT-4o-mini, CLEVR-Counting. Uncertainty reflection rate under fine-grained image-, text-, and cross-modality perturbations on CLEVR-Counting using GPT-4o-mini.

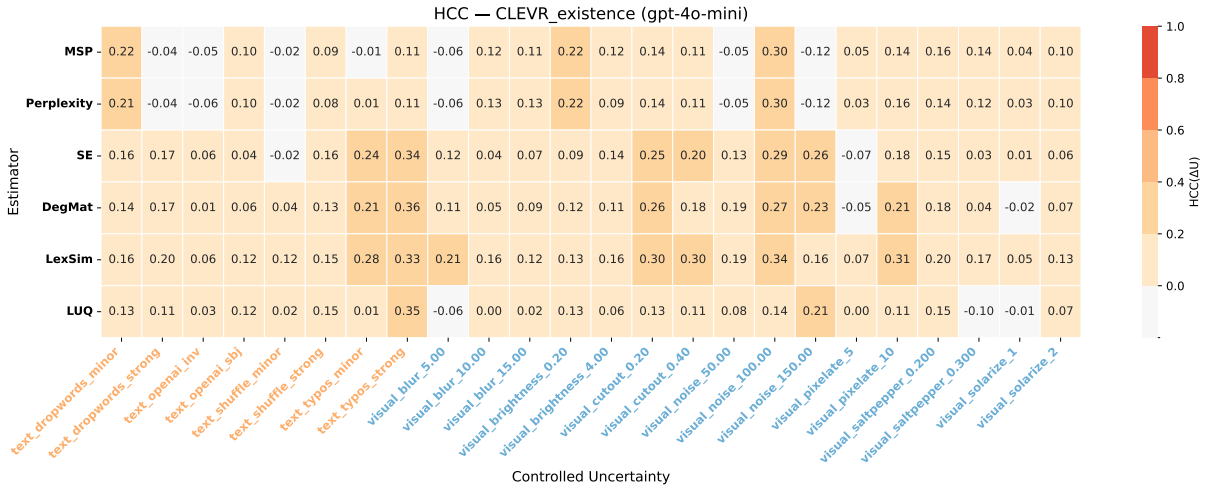

Figure 16: GPT-4o-mini, CLEVR-Existence. Uncertainty reflection rate under fine-grained image-, text-, and cross-modality perturbations on CLEVR-Existence using GPT-4o-mini.

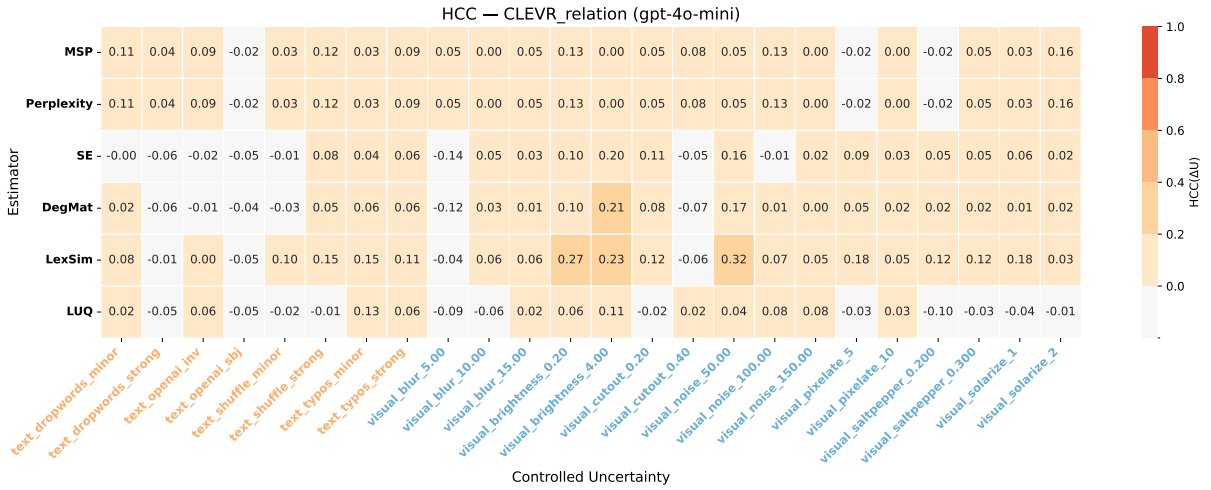

Figure 17: GPT-4o-mini, CLEVR-Relation. Uncertainty reflection rate under fine-grained image-, text-, and cross-modality perturbations on CLEVR-Relation using GPT-4o-mini.

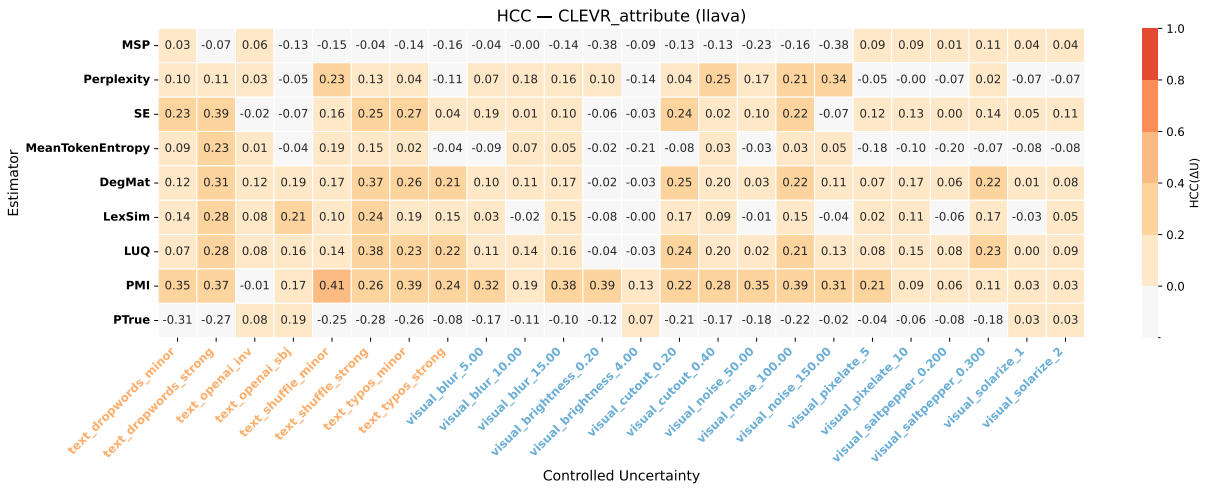

Figure 18: LLaVA, CLEVR-Attribute. Uncertainty reflection rate under fine-grained image-, text-, and cross-modality perturbations on CLEVR-Attribute using LLaVA.

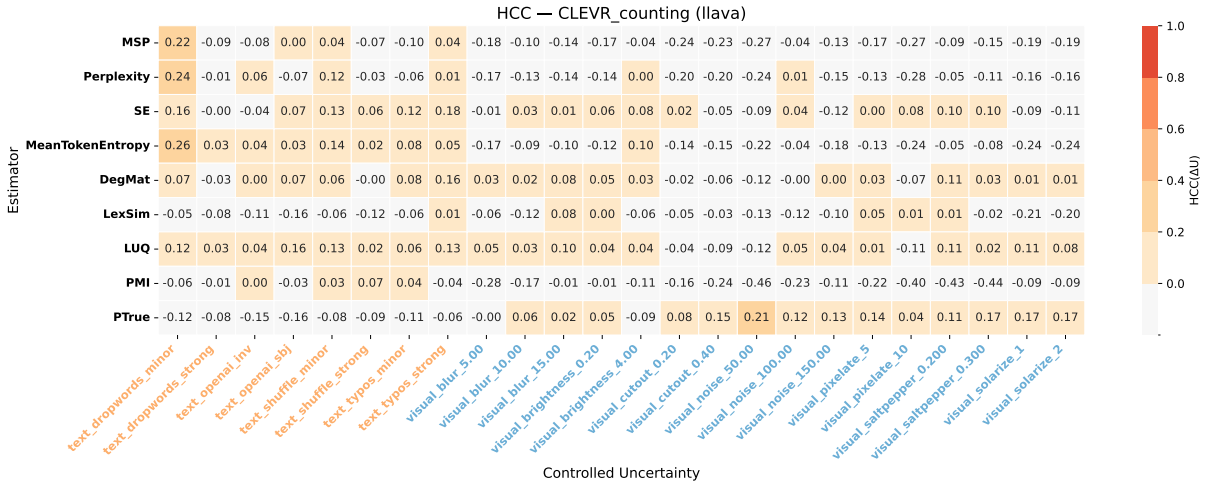

Figure 19: LLaVA, CLEVR-Counting. Uncertainty reflection rate under fine-grained image-, text-, and cross-modality perturbations on CLEVR-Counting using LLaVA.

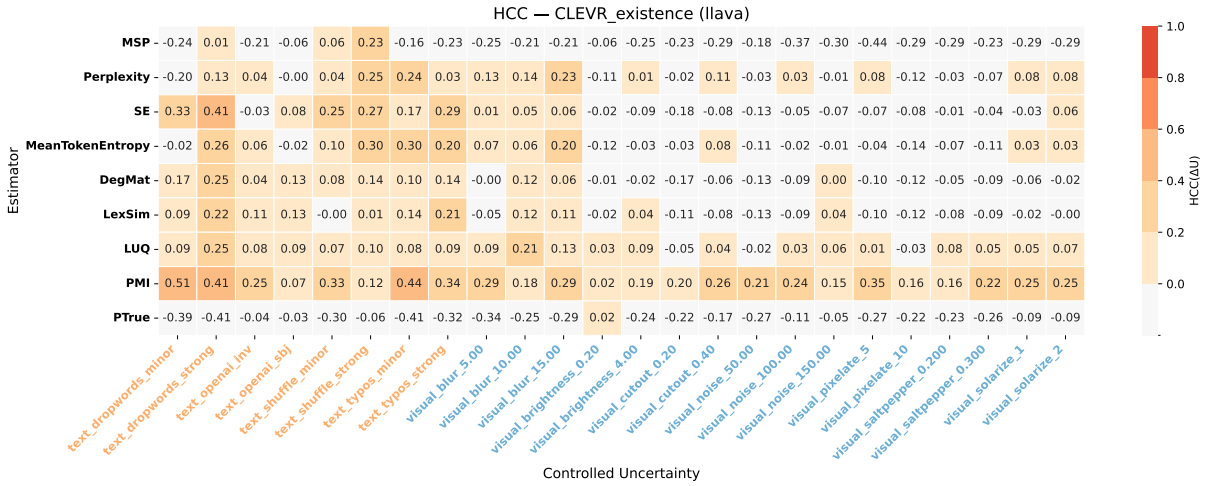

Figure 20: LLaVA, CLEVR-Existence. Uncertainty reflection rate under fine-grained image-, text-, and cross-modality perturbations on CLEVR-Existence using LLaVA.

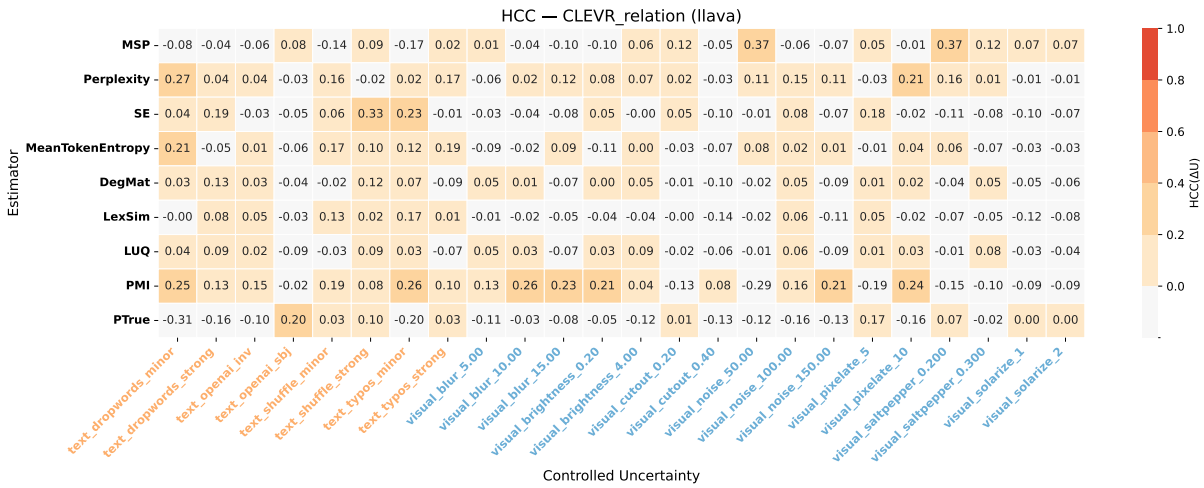

Figure 21: LLaVA, CLEVR-Relation. Uncertainty reflection rate under fine-grained image-, text-, and cross-modality perturbations on CLEVR-Relation using LLaVA.

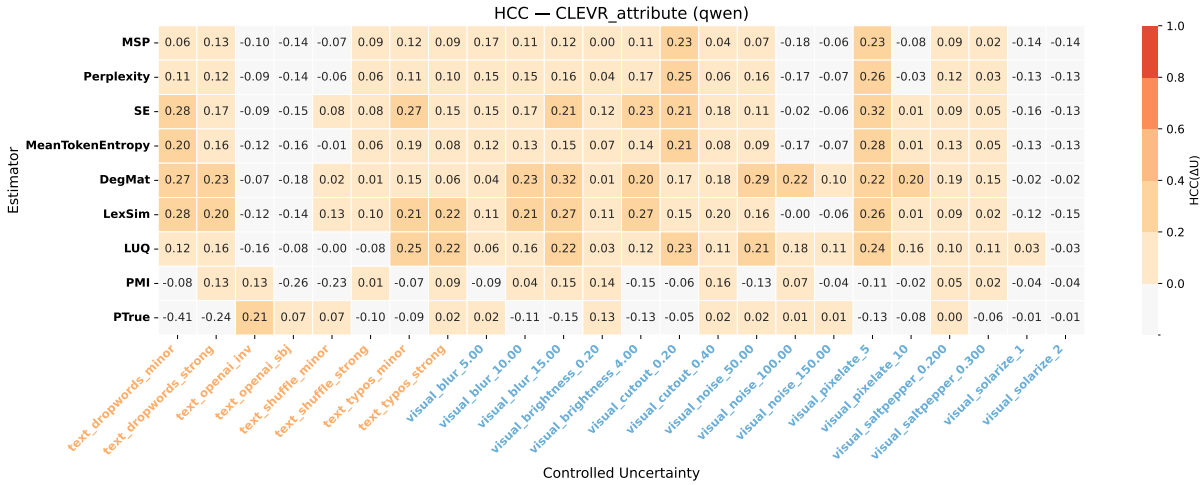

Figure 22: Qwen-VL, CLEVR-Attribute. Uncertainty reflection rate under fine-grained image-, text-, and cross-modality perturbations on CLEVR-Attribute using Qwen-VL.

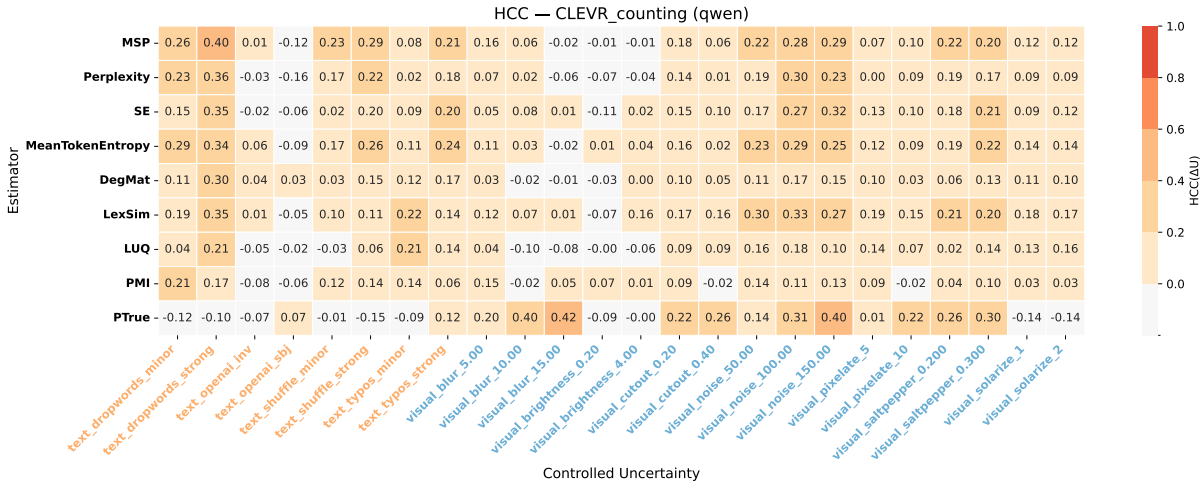

Figure 23: Qwen-VL, CLEVR-Counting. Uncertainty reflection rate under fine-grained image-, text-, and cross-modality perturbations on CLEVR-Counting using Qwen-VL.

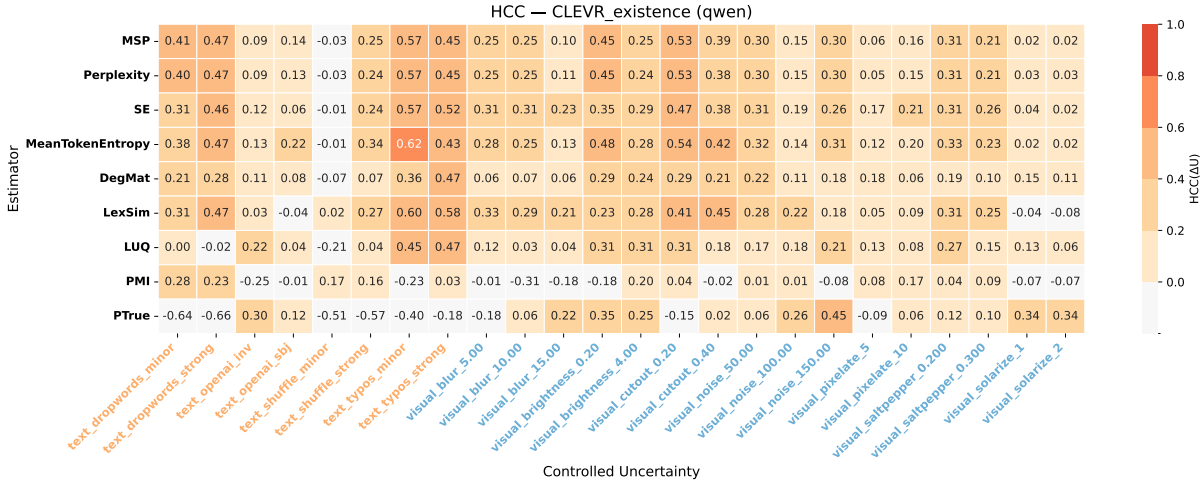

Figure 24: Qwen-VL, CLEVR-Existence. Uncertainty reflection rate under fine-grained image-, text-, and cross-modality perturbations on CLEVR-Existence using Qwen-VL.

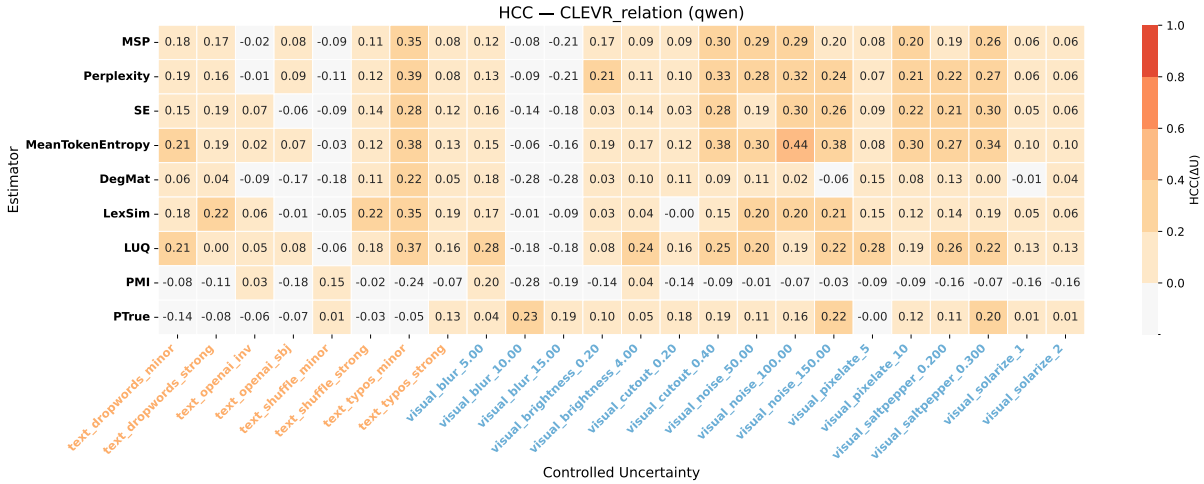

Figure 25: Qwen-VL, CLEVR-Relation. Uncertainty reflection rate under fine-grained image-, text-, and cross-modality perturbations on CLEVR-Relation using Qwen-VL.

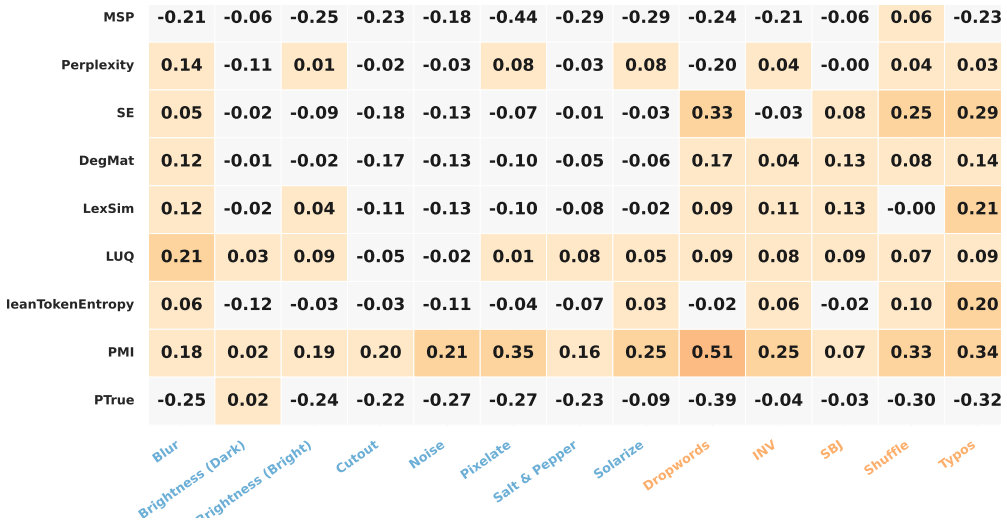

Figure 26: LLaVA on CLEVR-Existence. Rows list UQ estimators; columns denote modality-specific perturbations (visual in blue, textual in orange). Cells show the point-biserial correlation between hallucination flips and uncertainty change, where higher (warmer) values indicate better alignment between uncertainty change and hallucination flips; negative values share the zero color.

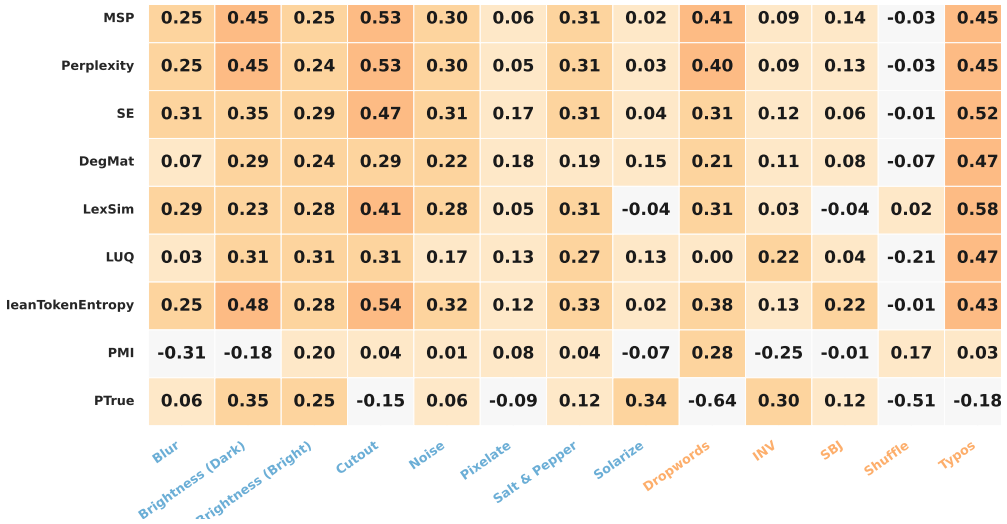

Figure 27: Qwen-VL on CLEVR-Existence. Rows list UQ estimators; columns denote modality-specific perturbations (visual in blue, textual in orange). Cells show the point-biserial correlation between hallucination flips and uncertainty change, where higher (warmer) values indicate better alignment between uncertainty change and hallucination flips; negative values share the zero color. This is the representative result discussed in the main paper.

D Additional Details on Ensemble Baselines

This appendix provides additional details for the ensemble results summarized in **RQ4** of the main paper. Our goal is not to introduce a new standalone method, but to better understand whether complementary UQ signals can be combined into a stronger uncertainty baseline, and under what level of supervision this becomes effective. We therefore report three groups of results: (i) **oracle upper bounds** that quantify the available headroom, (ii) **practical few-shot ensemble baselines** that correspond to the main paper results, and (iii) **additional unsupervised and fully supervised baselines** for completeness.

D.1 Experimental Setup

Evaluation scenarios. We evaluate ensemble baselines on the same hallucination-focused setting used in the main paper: four VLMs—`gpt-4o-mini`, `kosmos`, `llava`, and `qwen`—across three attack types (*image perturbation*, *image–text mismatch*, and *text manipulation*) plus an *all-combined* setting, yielding 16 scenarios in total. Each per-attack scenario contains 150 clean and 150 attacked samples, while the combined setting contains 150 clean and 450 attacked samples.

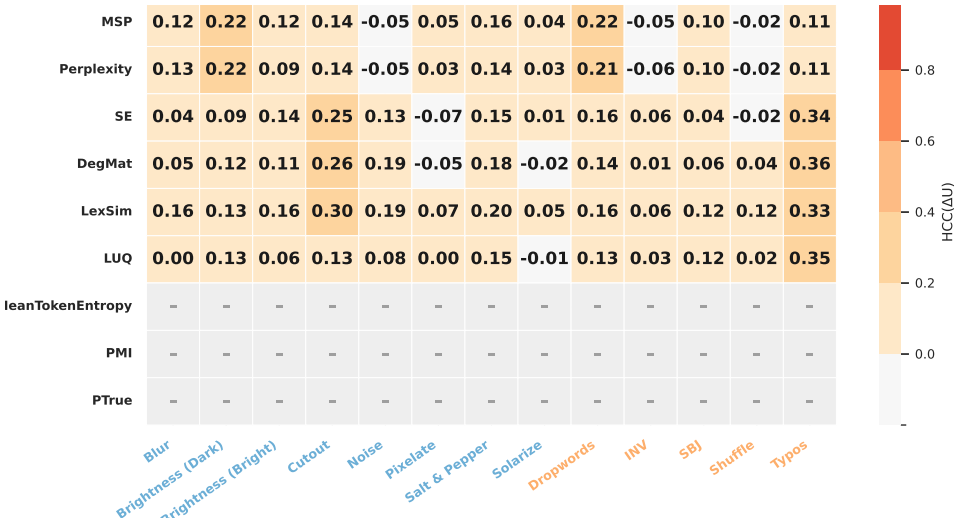

Figure 28: GPT-4o-mini on CLEVR-Existence. Rows list UQ estimators; columns denote modality-specific perturbations (visual in blue, textual in orange). Cells show the point-biserial correlation between hallucination flips and uncertainty change, where higher (warmer) values indicate better alignment between uncertainty change and hallucination flips; negative values share the zero color. PMI, PTrue, and MeanTokenEntropy are omitted because the closed-source API does not expose full token-level distributions.

Candidate UQ estimators. The number of available UQ estimators differs by model. `gpt-4o-mini` has 6 estimators: MAXIMUMSEQUENCEPROBABILITY, PERPLEXITY, LEXICALSIMILARITY, SEMANTICENTROPY, DEGMAT, and LUQ. `kosmos` has 8 estimators (the above plus PTRUE and MEANTOKENENTROPY), and `llava` and `qwen` each have 9 estimators (further including MEANPOINTWISEMUTUALINFORMATION).

Task and metric. All ensemble methods are evaluated as binary classifiers for *clean vs. attacked* samples using **AUC**. Higher AUC indicates better discrimination between clean and hallucination-prone samples.

Direction alignment. Different UQ estimators use opposite uncertainty conventions: some assign larger values to *more certain* outputs (e.g., MAXIMUMSEQUENCEPROBABILITY), while others assign larger values to *more uncertain* outputs (e.g., SEMANTICENTROPY). Before combining estimators, we therefore perform **direction alignment**. Given labeled calibration data, we compare the AUC of each estimator under its original sign and its negation, and retain the direction that yields higher train AUC.

Rank normalization. Unless otherwise noted, ensemble methods operate on **rank-normalized** estimator scores. For each estimator, scores are converted to normalized ranks in $[0, 1]$ within the evaluation split. This reduces scale mismatch across heterogeneous estimators and makes aggregation or classification more stable.

Few-shot protocol. For the practical few-shot experiments corresponding to the main paper, we use 5-fold stratified cross-validation with **reversed splits**: the smaller fold ($\sim$20%, around 60 samples) serves as training/calibration data, while the larger fold ($\sim$80%, around 240 samples) serves as evaluation data. This setup stresses the low-label regime and tests whether a small calibration split is sufficient to unlock useful complementarity across estimators.

D.2 Oracle Upper Bound

We first ask whether the candidate UQ estimators contain complementary information at all. To quantify the best-case potential of estimator combination, we define three oracle-style references using ground-truth labels:

- **Oracle-Max:** for each sample, take the maximum rank-normalized score across estimators.
- **Oracle-Min:** for each sample, take the minimum rank-normalized score across estimators.
- **Oracle-Perfect:** for clean samples, take the max; for attacked samples, take the min.

Oracle-Perfect is not a deployable method. Instead, it is an optimistic sample-wise upper bound

Table 7: Oracle upper-bound analysis. Oracle-Perfect is a sample-wise optimistic upper bound.

Model	Attack	Best Single	Oracle-Perfect	Gap
gpt-4o-mini	image	0.726	0.999	+0.273
	mismatch	0.767	0.999	+0.232
	text	0.778	0.995	+0.217
	combined	0.748	0.979	+0.231
kosmos	image	0.651	1.000	+0.349
	mismatch	0.768	1.000	+0.232
	text	0.696	1.000	+0.304
	combined	0.689	1.000	+0.311
llava	image	0.724	1.000	+0.276
	mismatch	0.640	1.000	+0.360
	text	0.726	1.000	+0.274
	combined	0.685	1.000	+0.315
qwen	image	0.854	0.999	+0.145
	mismatch	0.829	0.996	+0.167
	text	0.893	0.999	+0.106
	combined	0.851	0.998	+0.147

that shows how much useful complementary information is already distributed across the estimator pool.

Table 7 reports the oracle analysis. Across all 16 scenarios, Oracle-Perfect reaches AUC between 0.98 and 1.00, indicating that the candidate UQ estimators collectively contain enough information to almost perfectly separate clean from attacked samples. The gap between the best single estimator and Oracle-Perfect ranges from $+0.106$ to $+0.360$, confirming that no single estimator exhausts the available signal and motivating ensemble baselines as a meaningful extension of the benchmark.

D.3 Practical Few-Shot Ensemble Baselines

The main paper reports the performance of a simple calibrated ensemble as the strongest practical baseline under the few-shot setting. Here we provide additional implementation details and broader comparisons within that same low-label regime.

We first considered two simple calibration settings. In a **zero-shot domain-heuristic** setting, estimator directions are assigned using heuristic semantic priors (e.g., entropy-like estimators should increase with uncertainty, probability-like estimators should decrease). These heuristics fail to beat the strongest single estimator in any scenario, showing that estimator semantics do not transfer cleanly across models and attack types. We then consider **small-sample calibration**, where a modest labeled subset is used only for direction alignment and estimator selection. Using $K = 50$ labeled examples (25 clean, 25 attacked), a simple **Calibrate-Top-3** strategy already beats the strongest single estimator in 11/16 scenarios, suggesting that a small calibration budget is sufficient to recover a substantial portion of the available ensemble benefit.

D.3.1 Representative Few-Shot Methods

Below we describe the representative few-shot methods used in the broader comparison.

Calibrate-Top-K Weighted (Cal-Top$K$-W). On the training fold, each estimator is direction-aligned and assigned a train AUC. The top-K estimators are selected and their rank-normalized outputs on the evaluation fold are combined using weights proportional to $(\mathrm{AUC}_j - 0.5)$. We evaluate $K \in \{2, 3, 4, 5\}$ and an adaptive-K variant.

Cal-Top-3 Median. This method uses the same direction alignment and top-3 selection as Cal-Top-3-W, but aggregates the selected estimators using the median rather than a weighted mean. It is more robust to occasional outlier scores.

Leave-Worst-Out (LWO). Starting from all estimators, we iteratively remove the estimator with the lowest train AUC. At each subset size, we evaluate a weighted rank-average on the train split and select the best-performing subset size. The selected subset is then applied to the evaluation split.

LR on Ranks. All estimators are direction-aligned using train labels, transformed to normalized ranks in $[0, 1]$, and then combined using logistic regression (`liblinear`, $C = 1.0$). This method is deliberately simple; its main advantage comes from preprocessing and feature comparability rather than classifier complexity.

Ensemble of Ensembles. This method averages the normalized predictions of Cal-Top-3-W, Cal-Top-3 Median, and LWO, providing a simple hybrid that can be more robust to the failure mode of any one constituent method.

Nested Selection. An inner 3-fold CV on the small training fold is used to choose among Cal-Top-3-W, Cal-Top-3 Median, and LWO; the selected method is then applied to the outer evaluation fold.

D.3.2 Main Few-Shot Results

We next evaluate a broader family of 53 few-shot ensemble baselines spanning calibration-based aggregation, low-capacity trained classifiers, normalization variants, feature-selection strategies, and hybrid methods.

For reference, we include **Oracle Best Single**, an *optimistic single-method reference* that selects the strongest individual estimator using evaluation labels. Since this baseline uses test labels, it is not deployable; rather, it answers the narrow question of whether an ensemble can outperform even the best single estimator one could have selected in hindsight.

Table 8 reports the strongest few-shot methods across all 16 scenarios. **LR on Ranks** achieves the highest average AUC (0.761) and beats the optimistic single-estimator reference in 12/16 scenarios. **LWO** and **Ensemble of Ensembles** are close behind, while the simpler **Cal-Top-3 Weighted** baseline remains highly competitive despite requiring almost no model training. These results show that a small calibration split is sufficient to exploit useful complementarity across UQ estimators.

For completeness, Table 9 groups the broader 53-method few-shot search space by category, and Table 10 reports the strongest overall few-shot methods ranked by average AUC across the 16 scenarios.

D.4 Additional Baselines

Unsupervised baselines. A natural baseline class is **unsupervised ensemble aggregation**, which does not use labels at all. This setting is attractive in principle, but substantially harder in practice because both direction alignment and estimator selection must be inferred indirectly. We evaluate standard label-free strategies such as mean-rank, median-rank, agreement-weighted averaging, PCA projection, and cluster-representative selection, as well as stronger variants based on exhaustive sign search, variance-based filtering, and iterative pruning. Table 11 shows that unsupervised ensembles remain weak overall: the main bottleneck is unstable direction alignment, while weak or near-random estimators often dilute stronger ones. Even the best unsupervised method remains clearly below the few-shot setting.

Fully supervised baselines. We also evaluate strong **fully supervised** ensemble baselines that use all labels and therefore serve as upper-reference baselines rather than realistic deployment-time methods. We test 23 methods spanning linear models, tree ensembles, shallow neural baselines, soft voting, stacking, and exhaustive subset search. Table 12 summarizes the results. A consistent pattern is that **feature selection matters more than model complexity**: best-subset search with simple classifiers (LR or GBT) outperforms more expressive architectures, while MLP-based baselines tend to overfit in this low-dimensional regime.

D.5 Additional Analysis

Feature selection dominates model complexity. Across both the fully supervised and few-shot settings, the most consistent pattern is that *which estimators are included* matters more than *how they are combined*. Methods such as best-subset search and Leave-Worst-Out are strong largely because they remove weak or noisy estimators that would otherwise dilute the signal.

Rank normalization is critical. The strongest practical method, LR on Ranks, differs from standard logistic regression mainly in its preprocessing. By converting each estimator to normalized ranks before fitting the classifier, the method removes scale mismatch, reduces the influence of outliers, and produces a better-conditioned feature space for linear combination.

Simple calibrated baselines are already strong. Cal-Top-3 Weighted performs only slightly below the best trained few-shot method while requiring almost no model fitting. This suggests that, in practice, a small calibration split may be sufficient to identify the most reliable estimators and align their directions.

Ensemble benefit is model-dependent. The value of estimator combination varies substantially by backbone. For `qwen`, many individual estimators are already strong, so ensembles provide modest but consistent gains. For `llava`, estimator quality is more heterogeneous, and ensembles yield larger improvements when they successfully identify complementary subsets. For `kosmos`, improvements are limited by the fact that only a small number of estimators rise meaningfully above random.

Failure cases reveal a genuine limit. A few scenarios remain resistant to ensemble improvement, notably `kosmos`/text, `kosmos`/combined, and `llava`/mismatch. In these cases, the estimator pool appears to contain one moderately useful signal surrounded by near-random ones. When the remaining estimators do not add complementary information, even strong combination strategies cannot outperform the best single estimator.

Table 8: Few-shot ensemble baselines under 5-fold CV with approximately 20% train / 80% eval splits. "Oracle Best Single" is an optimistic reference that selects the strongest individual estimator with access to evaluation labels. Gains are measured relative to that reference. Bold indicates the best non-oracle method in each row.

Model	Attack	Oracle Best Single	Cal-T3 Wt.	LWO	LR on Ranks	Best Practical	Best Gain
gpt-4o-mini	image	0.726	0.736	0.735	0.743	**0.746**	+0.021
	mismatch	0.767	0.779	0.776	**0.792**	**0.792**	+0.025
	text	0.779	0.782	0.780	**0.787**	**0.787**	+0.008
	combined	0.751	0.769	0.767	0.765	**0.775**	+0.024
kosmos	image	0.654	0.642	0.663	**0.673**	**0.673**	+0.018
	mismatch	0.769	**0.792**	0.784	0.786	**0.792**	+0.023
	text	0.696	0.682	0.680	**0.686**	**0.686**	−0.010
	combined	0.689	0.672	0.677	**0.679**	**0.679**	−0.010
llava	image	0.724	0.755	0.744	0.746	**0.756**	+0.032
	mismatch	0.642	0.621	0.632	**0.637**	**0.637**	−0.006
	text	0.726	0.725	**0.739**	0.736	**0.739**	+0.013
	combined	0.685	0.720	0.717	**0.720**	**0.720**	+0.036
qwen	image	0.856	0.854	0.864	**0.868**	**0.868**	+0.012
	mismatch	0.829	0.817	0.819	0.816	**0.819**	−0.010
	text	0.894	0.897	0.894	0.896	**0.900**	+0.007
	combined	0.854	0.852	0.852	0.851	**0.856**	+0.002
Average		0.753	0.756	0.758	**0.761**	**0.761**	+0.012
# Beats Oracle Best Single		—	10/16	10/16	**12/16**	**12/16**	—

Table 9: Taxonomy of the 53 few-shot ensemble baselines evaluated under 5-fold cross-validation with approximately 20% train / 80% eval splits.

Category	Representative Methods	#Methods
Calibration-based	Top-K weighted/unweighted, adaptive-K, inner-CV K	12
Direction alignment	Domain prior, train-calibrated, Bayesian weighting	4
Aggregation rules	Mean, median, Borda, Copeland, harmonic/geometric	7
Feature selection	Leave-worst-out, best-pair, best-subset, pruning	6
Normalization variants	CDF-weighted, Platt, isotonic, z-score raw values	4
Trained classifiers	LR, LR+Poly2, GBT, RF, ET, SVM, Ridge, LDA, KNN	12
Meta / hybrid	Nested selection, ensemble-of-ensembles, meta-switch	5
Baselines	Best single, domain-mean-rank, optimistic references	3

Table 10: Top few-shot ensemble baselines ranked by average AUC across all 16 scenarios. #Best counts the number of scenarios in which the method attains the highest non-oracle AUC.

Rank	Method	Avg. AUC	#Best	Category
1	LR on Ranks	**0.7613**	**5**	Trained classifier
2	Leave-Worst-Out	0.7576	4	Feature selection
3	Ensemble of Ensembles	0.7575	1	Meta / hybrid
4	CDF-Weighted Top-3	0.7563	1	Normalization variant
5	Cal-Top-3 Weighted	0.7560	1	Calibration
6	Platt Scaling Top-3	0.7558	0	Normalization variant
7	Borda Count Top-3	0.7554	0	Aggregation rule
8	Bayesian Weighted	0.7553	1	Direction alignment
9	Nested Selection	0.7553	2	Meta / hybrid
10	Cal-Top-3 Wt. Exp.	0.7537	0	Calibration
11	Inner-CV Best-K	0.7537	1	Calibration
12	Cal-Top-5 Weighted	0.7532	1	Calibration
13	Cal-Top-3 Median	0.7461	4	Aggregation rule
14	Stacking (LR+ET)	0.7454	3	Trained classifier
15	Ridge Classifier	0.7447	1	Trained classifier
–	Oracle Best Single	0.7525	–	Optimistic reference

Table 11: Unsupervised ensemble baselines. Without labels, direction alignment and estimator filtering become the dominant failure modes.

Method	#Wins	Avg. Gain	Description
Standard unsupervised methods			
Cluster-representative	3/16	−0.039	One estimator per correlation cluster
Agreement-weighted	1/16	−0.039	Weighted by average correlation
Median-rank	0/16	−0.044	Median of aligned ranks
Mean-rank	0/16	−0.049	Mean of aligned ranks
Improved unsupervised methods			
Top-3 by Variance	5/16	−0.024	Variance filtering + direction search
Exhaustive Direction	2/16	−0.028	Search over sign assignments
LOO-Consistency Wt.	0/16	−0.032	Correlation-based pruning/weighting

Table 12: Fully supervised ensemble baselines using all labels. The strongest methods are simple models combined with explicit subset selection, suggesting that estimator choice is more important than model complexity.

Method	Avg. AUC	Avg. Gain	#Best	Note
Best-subset GBT	0.774	+0.040	4	Nonlinear + subset selection
Best-subset LR	0.773	+0.041	5	Simple linear + subset selection
Soft Voting (5)	0.772	+0.034	1	Diverse base models
ExtraTrees	0.771	+0.033	2	Robust tree ensemble
Random Forest	0.769	+0.031	1	Stable supervised baseline
LR + Poly2	0.763	+0.027	2	Captures pairwise interactions
AdaBoost	0.740	+0.016	1	Moderate gains
SVM-RBF	0.735	+0.011	0	Nonlinear boundary only
Logistic Reg. (L2)	0.729	+0.005	0	Plain linear baseline
MLP (32,16)	0.641	−0.086	0	Overfits at this scale
Stacking (MLP meta)	0.590	−0.137	0	Severe overfitting