Quantifying and Evaluating Continuity Properties of Multi-modal LLMs

Anonymous ACL submission

Abstract

The recent advances in multimodal large language models (MLLMs), while extending the skills and capabilities of text-only LLMs, have also made the model responses vulnerable to increased hallucinations, reduced contextual awareness and inconsistency in complex reasoning. Most existing works on benchmarking of MLLMs focus on datasets consisting of isolated samples that do not allow the evaluation of continuity and monotonicity properties of these models. In this paper, we develop a synthetic benchmark for evaluating MLLM performance and uncertainty on continually varying dimensions of complexity. The benchmark relies on the core real-world principle that *inputs of increasing/decreasing ambiguity should ideally lead to higher/lower model uncertainty*. We experiment with 5 large vision language models (LVLMs) and 4 large audio language models (LALMs) across various image question-answering and audio question-answering tasks. Our findings show that most MLLMs lack the real-world continuity and monotonicity that are human-like.

1 Introduction

With the recent rise of multimodal LLMs (MLLMs) (Team et al., 2023; OpenAI, 2023), interactions and understanding of audio and visual content have become increasingly plausible with remarkable performances. However, across tasks, multimodal LLMs tend to exhibit hallucinatory behaviors and inconsistent responses, for example in visual and audio reasoning tasks (Fu et al., 2024; Bhattacharya et al., 2025a), due to the lack of grounding (Favero et al., 2024). In such a setting, the trustworthiness of the MLLMs can be enhanced through uncertainty estimators (Bhattacharya et al., 2025b; Khan and Fu, 2024; Kuhn et al., 2023; Nguyen et al., 2025). The use of uncertainty estimators to enhance the utility and reliability of MLLMs has been a rapidly evolving area of research (Zhi et al., 2025).

In terms of benchmarking and uncertainty evaluations, common dataset choices like visual spatial reasoning (Liu et al., 2023a; Fu et al., 2024) and audio reasoning (Sakshi et al., 2025; Bhattacharya et al., 2025a) allow the comparison and ranking of various MLLMs with respect to existing metrics of performance and uncertainty. However, most of these benchmark datasets consist of isolated and independent data samples that evaluate model performance with respect to specific proxy applications, like abstention. They do not allow the evaluation of fundamental properties of real-world deployment of models, such as the continuity property of model performance and uncertainty on specific dimensions of the input complexity. A handful of exceptions are the studies that evaluate the potential of model uncertainty to detect data shifts, data corruptions and OOD samples (Ovadia et al., 2019; Hendrycks and Dietterich, 2019).

In this paper, we propose a benchmarking approach that involves gradually modulating the input complexity for various reasoning-based tasks. For a given input, the reasoning ambiguity is continually modified for a question-answering (QA) task. The model uncertainty is tracked with the increase/decrease in input complexity. Specifically, given an audio/image input with a textual prompt, first we dissect the possible axes of varying the input complexity. Subsequently, new data points are generated by varying its complexity dimensions. We call the original input as the source scene/audio, and the ones synthesized by variation as sweeping scenes/audios. Using the sweeping samples, we monitor the model predictions and uncertainty estimates using established methods. A set of 5 LVLMs and 4 LALMs is used for the benchmarking. Among visual QA tasks, spatial reasoning, shape reasoning and object counting are used. The audio QA based tasks involve temporal reasoning, reasoning with

distractors, and volume-based reasoning. By allowing a continuous change in the input ambiguity, we attempt to answer the fundamental question - `which class of models exhibit monotonic patterns of performance, similar to human expectations.`
Our core contributions are the following:

1. Given an image/audio QA pair, we propose a synthetic dataset and quantitative metrics to benchmark and evaluate the MLLMs with tasks of varying levels of complexity.

2. The study is quite unique compared to other benchmarks, as various scenes/audios are generated that transform the input along dimensions of interest based on the task involved.

3. We find that multiple state-of-the-art MLLMs offer perfect task accuracy; however, they fail to offer monotonic uncertainty behavior.

2 Prior works

Uncertainty Estimation : The two most common approaches used for model uncertainty evaluation are proxy tasks like *abstention performance* (Kuhn et al., 2023; Nguyen et al., 2025; Nikitin et al., 2024) and *calibration* (Guo et al., 2017; Tian et al., 2023). The *abstention performance* relies on the hypothesis that a model has more uncertainty associated with mispredictions than with correct predictions. The metrics used for this are AUC and AUCPR (Kuhn et al., 2023; Nguyen et al., 2025). On the other hand, *calibration* measures whether the confidence (reciprocal of uncertainty) values are aligned with respect to the model accuracies. Recent studies (Simhi et al., 2025) have shown that the LLMs and MLLMs may often mis-predict with low uncertainty, and the number of such cases can be significant.
Input Data Manipulation: Some works evaluate whether model uncertainty becomes higher under data shifts, data corruption, out-of-distribution (OOD) samples (Ovadia et al., 2019; Hendrycks and Dietterich, 2019). The work by (Ovadia et al., 2019) studies model uncertainty under data shifts, corruptions and OOD input datasets. They find that the uncertainty increases and calibration gets upset with increasing data shift. (Hendrycks and Dietterich, 2019) performs benchmarking of the models under real-world, non-adversarial data shifts like blur, brightness, and noise. They report that architectural improvements for accuracy improvement do not always translate to better robustness. (Lu et al., 2025) shows theoretically, and on real-world datasets, that data drift can be detected using prediction uncertainty rather than error rate more efficiently. However, the effect of varying data discriminability without distribution shift remains to be studied. The closest work to this paper, based on data shift-based failure modes, is the work by (Li et al., 2023). It uses synthetic datasets generated using the CLEVR framework to probe the VLM failures along dimensions of visual complexity, question redundancy, concept distribution and concept redundancy.

3 Problem Statement - Continuity Evaluation of MLLMs

In this paper, we propose benchmarks derived from existing image and audio datasets for continuity-based evaluation of model performance and uncertainty. In order to measure the monotonic patterns of uncertainty under continuous variation of discriminability, we synthesize sampled data-points that define a continuum of spatial/temporal discriminability in images/audio samples respectively.

Let x denote a multimodal (audio/image) sample $\in \mathcal{R}^{m,n}$. We denote $\mathbb{I}_+ = \{\hat{x}_1, ...\hat{x}_{n_+}\}$ as a sequence of synthetic data samples that should ideally evoke a series of non-decreasing performance and uncertainty estimates. Similarly, we denote $\mathbb{I}_- = \{\tilde{x}_1, ...\tilde{x}_{n_-}\}$ as a sequence of synthetic data samples that should ideally evoke a series of non-increasing performance and uncertainty estimates. Here, $\hat{x}_i, \tilde{x}_j \in \mathcal{R}^{m,n}$ indicate transformations of the original sample x with varied discriminability. We call them sweeping samples of the original sample x. In the example in Figure 1, an object's ("ball") location is perturbed while keeping the rest of the entities intact.

Let y denote a black-box MLLM response to the input x and $U(x, y)$ denote the uncertainty estimator for this input-output pair. We define, the following two parameters,

- $s_i = -1$ for $i \in \{1, .., n_+\}$ and $s_j = +1$ for $j \in \{1, .., n_-\}$, denote monotonicity direction.

- $d_i := |i - n_+|$ denotes step-distance from the most "complex" input step for samples in $\mathbb{I}_+$ and $d_j := |j - 1|$ denotes the similar measure for samples in $\mathbb{I}_-$.

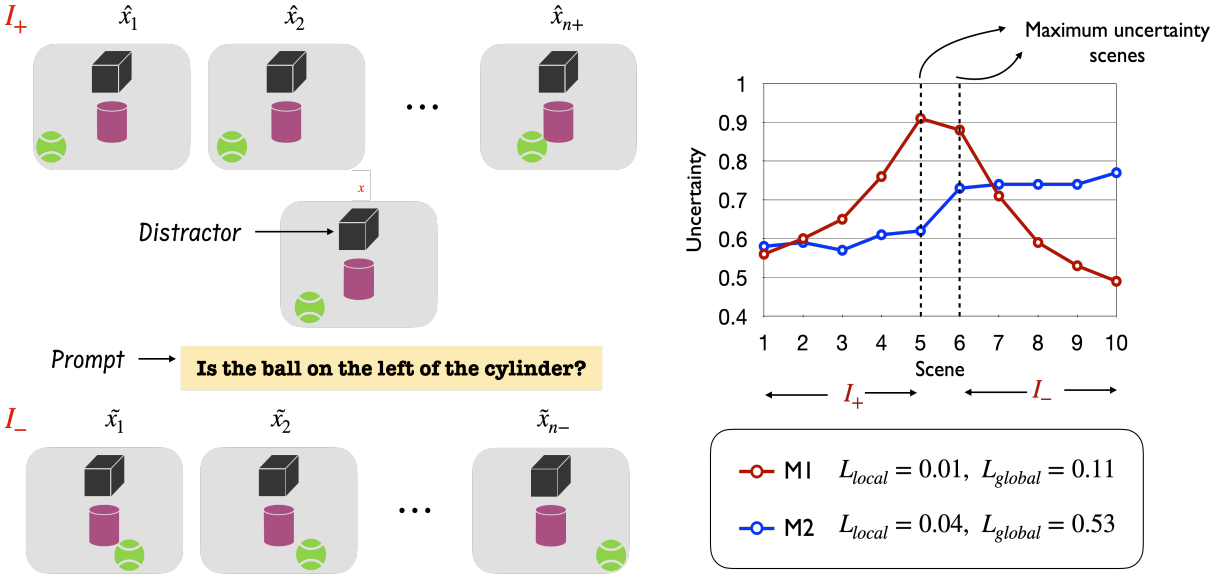


Figure 1: Left panel - Input x at the center along with the textual prompt for probing the spatial reasoning of an MLLM. The synthetic examples $\{\hat{x}_1, .., \hat{x}_{n_+}\} \in \mathbb{I}_+$, which should ideally evoke a non-decreasing uncertainty sequence are illustrated in top-panel, while $\{\tilde{x}_1, .., \tilde{x}_{n_-}\} \in \mathbb{I}_-$, which should ideally evoke a non-increasing uncertainty sequence are illustrated in bottom-panel. Right panel - An example uncertainty estimate from two MLLMs (M_1, M_2). M_1 has better continuous uncertainty behavior over other (M_2) as it exhibits lower monotonicity violation measured using losses $\mathcal{L}_{local}$ and $\mathcal{L}_{global}$.

Further, denoting $\hat{u}_i = U(\hat{x}_i, \hat{y}_i)$ and $\tilde{u}_j = U(\tilde{x}_j, \tilde{y}_j)$, we define two monotonicity violation losses to quantify the monotonicity of the uncertainty estimators.

$$\mathcal{L}_{\text{local}} := \sum_{i=1}^{n_+-1} \frac{\max\{0,\ s_i\,(\hat{u}_{i+1} - \hat{u}_i)\}}{|\hat{u}_{i+1} - \hat{u}_i|} + \sum_{j=2}^{n_-} \frac{\max\{0,\ s_j\,(\tilde{u}_j - \tilde{u}_{j-1})\}}{|\tilde{u}_j - \tilde{u}_{j-1}|} \quad \text{(1a)}$$

$$\mathcal{L}_{\text{global}} := \frac{1 + \rho_S}{2}, \text{ where} \quad \text{(1b)}$$

$$\rho_S := \text{SC}\big(\{[\hat{u}_i]_{i=1}^{n_+}, [\tilde{u}_j]_{j=1}^{n_-}\}; \{[-d_i]_{i=1}^{n_+}, [d_j]_{j=1}^{n_-}\}\big) \quad \text{(1c)}$$

Here $SC(\cdot, \cdot)$ denotes the Spearman rank correlation coefficient between two arrays. The $\mathcal{L}_{\text{local}}$ measures the total magnitude of local monotonicity violation between consecutive steps. $\mathcal{L}_{\text{local}}$ increases only in case of monotonicity loss between consecutive steps, as ideally, $s_i\,(\hat{u}_{i+1} - \hat{u}_i) \geq 0, i \in \mathbb{I}_+$ and $s_j\,(\tilde{u}_j - \tilde{u}_{j-1}) \geq 0, j \in \mathbb{I}_-$. $\mathcal{L}_{\text{global}}$ is a rank loss that globally measures the number of rank monotonicity violations. For $i \in I_+$, $-d_i$ is an increasing sequence which positively correlates with the monotonically increasing behavior of $\hat{u}_i$. Also, for $j \in \mathbb{I}_-$, d_j is a decreasing sequence which correlates with the expected monotonically decreasing behavior of $\tilde{u}_j$. Both the losses $\mathcal{L}_{\text{local}}$ and $\mathcal{L}_{\text{global}}$ have lower value for better monotonic behavior of uncertainty. Also, they both range from 0 to 1.

In the illustrative example in the right-panel of Figure 1, two uncertainty trajectories along the discriminability dimension are studied, obtained from two models M_1 and M_2. The greater number of monotonicity violations and the lack of peak uncertainty at the scenes of maximum ambiguity in uncertainty sequence of M_2 result in higher losses as compared to M_1.

4 Proposed benchmark

We develop 3 variants of the vision-language tasks and 3 variants of the audio-language tasks for continuous uncertainty evaluation.

4.1 Vision-QA tasks

We create 3 main tasks based on 3 different dimensions of varying data discriminability.

4.1.1 Spatial location dimension

In the spatial reasoning-based task, one object is moved gradually from left to right (*primary ob-*

ject), and another object is placed at the center (*anchor object*). Hence, the spatial separation of the primary object and the anchor object keeps on reducing gradually in $\mathbb{I}_+$ and vice-versa in $\mathbb{I}_-$. An illustrative example is shown in Figure 1, where the separation between the ball and the cylinder keeps varying while the textual prompt - "*Is the ball on the left of the cylinder?*" attempts to probe the continuity properties of the MLLM under consideration using an uncertainty estimator. In total, 30 various source scenes are created to create diverse set of images by - (a) keeping only primary and anchor objects, (b) changing colors of the objects, (c) changing the size of primary object, (d) changing the size of anchor object, (e) change of primary object, (f) change of anchor object, (g) changing both objects, (h) one distractor object, (i) two distractor objects, (j) shifting one of the objects diagonally instead of horizontal movements.

In these cases, we keep $n_+ = 5$ and $n_- = 5$. Further, we alter object shapes chosen among basic (cube, sphere cylinder, prism etc.) and complex shapes (cycle, airplane, bus etc.), and divide it into 3 sub-categories - (a) scenes involving both objects as basic (B-B), (b) one complex object and another basic one (C-B), (c) both objects being complex (C-C). A few examples are shown in Figure 2. Illustrative examples of these visual variations are shown in Appendix C (Figures 4–6).

4.1.2 Scale dimension

In this setting, the object sizes are smoothly varied. The primary object is initially considered smaller than the anchor object. Gradually, it increases to match the size of the anchor object ($\mathbb{I}_+$, and is increased further to make it even larger than the anchor ($\mathbb{I}_-$). The textual prompt probes the relative size between the objects. For this task, source scene variations include, (a) keeping only primary and anchor objects, (b) changing colors of the objects, (c) changing the shape of primary object, (d) changing the size of anchor object, (e) changing size of both objects, (f) one distractor, (g) two or more distractor objects, (h) swapping positions of the objects, (i) changing color of the primary object, (j) varying anchor object size. Here, $n_+ = 7$ and $n_- = 12$. Similar to the spatial reasoning, B-B, C-B and C-C variants are generated as well. An example is shown in Figure 2.

4.1.3 Count dimension

In this setting, the VLMs are probed in a counting task, where a certain type of object, is present with other distractors. The scenes contain a varying number of distractor objects, unrelated to the type of object that needs to be counted. The distractor objects are cluttered spatially. The number of such distractor objects is increased from 5 to 10 in images from 1 to $n_+ = 5$, and then decreased gradually from 10 to 6 in $n_- = 5$ images. Only basic object shapes are used in this task. An example is shown in Figure 2.

4.2 Audio-QA tasks

Similar to visual stimuli, several audio clips are also generated to probe large audio language models (LALMs).

4.2.1 Temporal location dimension

In this setting, two different audio events (from different audio sources) are placed at different temporal locations. One of them is designated as the primary event, which is moved along the temporal axis. The other source is an anchor source, which remains unchanged temporally. We use $n_+ = 5$, and $n_- = 5$. The textual prompt probes the LALMs about the temporal positional reasoning of the audio events. The audio events are randomly selected from the ESC-50 environmental sound events corpus (Piczak, 2015). Various such source scenes are created with, (a) only the primary and anchor events, (b) adding noise to the primary sound source, (c) noise added to the anchor source, (d) noise added to both sources, (e) changing source of primary sound event, (f) changing anchor sound source, (g) changing both primary and anchor sound sources, (h) distractor sound events before the primary and anchor events, (i) distractor sound events after the primary and anchor events, (j) distractor sound events both before and after the primary and anchor events.

4.2.2 Number of Distractors

In this setting, apart from the primary and anchor audio events of interest, several distractor audio events unrelated to the task are added to the audio recording without overlap. In this task as well, the temporal location of the primary source with respect to the anchor source is probed. The number and position of distractor events is changed to introduce variations and the effect on the uncertainty of the LALMs is studied. The variations

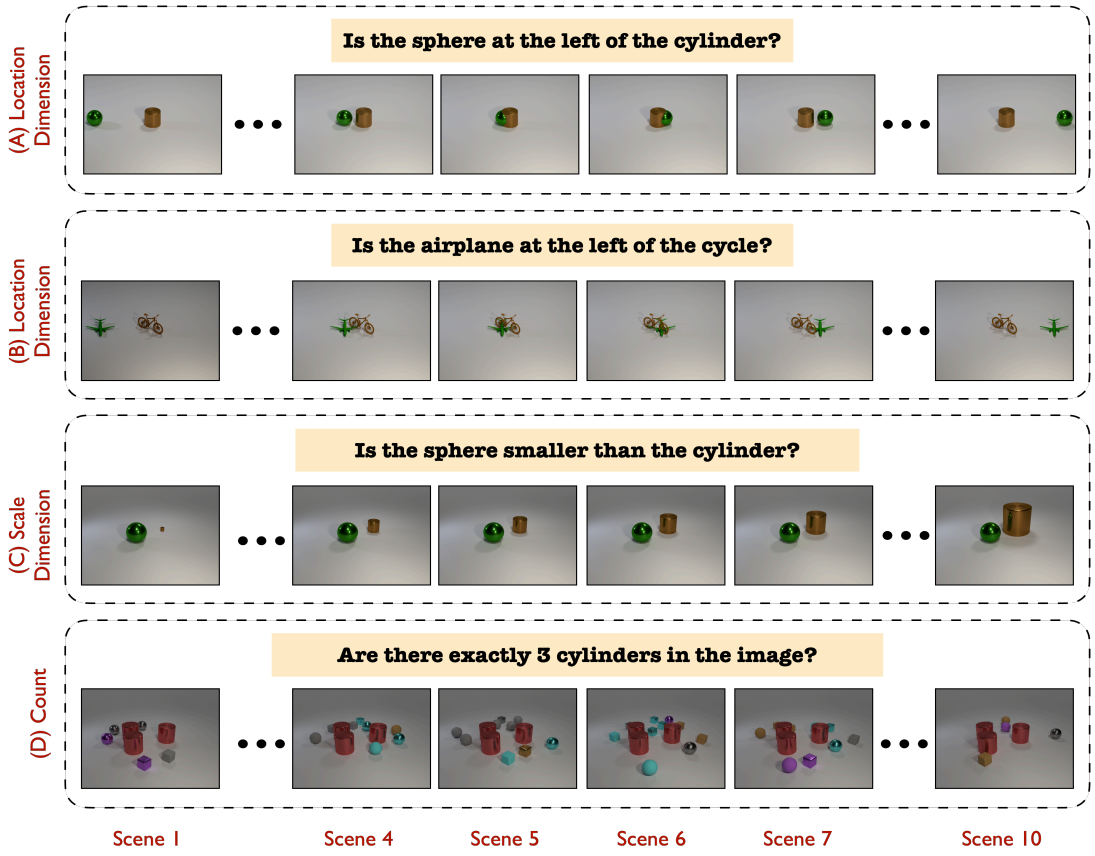


Figure 2: Illustrative examples of visual QA tasks based on spatial location, scale and count. (A) Spatial location dimension task with basic-basic (B-B) object shapes, (B) Spatial location dimension task with complex-complex (C-C) object shapes, (C) Scale dimension task, (D) Count task.

are, (a) only the primary and anchor events present, (b) one distractor event before both events, (c) one distractor event after both events, (d) one distractor before and one distractor after both events, (e) two distractors before and one distractor after both events, (f) one distractor before and two distractors after both events, (g) two distractors before and two distractors after both events.

4.2.3 Volume dimension

In this setting, two different audio events (from different audio sources) are placed at different temporal locations. The volume of the primary event is initially significantly lesser than the anchor event. Then, the volume of the primary event is increased gradually ($\mathbb{I}_+$) and subsequently, made more and more louder than the anchor event ($\mathbb{I}_-$). Textual prompt probes the model to check if the primary event is louder than the anchor event.

5 Experimental setup

Apart from creating datasets based on Section 4, the continuous uncertainty evaluation involves selecting a sound black-box uncertainty estimation method and an MLLM. We experiment with 2 black-box uncertainty estimation methods and 9 MLLMs (5 LVLMs and 4 LALMs).

5.1 Image QA dataset

The synthetic vision-language datasets are generated using the CLEVR (Johnson et al., 2017) framework, which provides procedural 3D scene generation with Blender (Blender Online Community, 2025) as the rendering backend. To extend beyond simple geometric shapes in CLEVR, we integrate complex 3D object models from the Super-CLEVR (Li et al., 2023) dataset, enabling experiments with realistic shapes. All images are rendered using Blender's Cycles path-tracing engine at 640×480 resolution with 512 samples per pixel. For spatial reasoning tasks, we employ an orthographic camera projection with scale 6.0 to eliminate perspective distortions. A fixed metal material is applied to all objects for visual consistency. Scene reproducibility is ensured through seed-controlled random number generators for object placement, orientation ($\theta \in [0°, 360°]$), and color selection.

Object complexity levels: The vision QA tasks support three types of object shape combinations: (a) **Basic-Basic (B-B)**: both objects are basic CLEVR shapes (sphere, cube, cylinder); (b) **Basic-Complex (B-C)**: one among the primary and anchor obejcts is basic shape, another is a complex shape from Super-CLEVR; (c) **Complex-Complex (C-C)**: both objects are complex shapes.

Dataset statistics: For each of the *spatial location dimension* and *scale dimension* tasks, 30 images are created and for the *counting* task, 10 images are created. With $n_+ = 5$ and $n_- = 5$, this leads to a corpus of 700 image-text pairs in the image QA category.

5.2 Audio QA dataset

We synthesize all audio samples using PyDub (Robert, 2014) by composing short event clips sourced from the ESC-50 environmental sound dataset (Piczak, 2015). Each sample is constructed on a 20 sec. stereo timeline (44.1 kHz, 16-bit), with global loudness normalization to -20 dBFS and 500 ms fade-in/out at event boundaries to reduce transition artifacts. We place two 5 sec. events on the timeline: a fixed audio event at start time 7.5 s, and a second audio event swept across 10 temporal positions from 0.0 sec. to 15.0 sec.. When the events overlap, we mix the overlap segment using equal-power crossfading with cosine/sine weighting to avoid perceived loudness dips.

Dataset statistics: Several variations are also introduced using different primary and anchor audio sources, adding noise, and distractor events and perturbations discussed in Section 4. In total, 20 such audio clips are created for the *temporal location* dimension, 30 clips for the *volume dimension* task, and 14 clips are created for the *distractor* task. Hence, in total, we get a corpus 640 audio clips in the audio QA experiments.

5.3 MLLMs considered

- The LVLMs explored are `LLaVA-1.6` (Liu et al., 2023b), `Phi-4` (Abdin et al., 2024), `Qwen-2.5-VL` (Yang et al., 2025), `Pixtral` (Agrawal et al., 2024) and `Gemma-3` (Team et al., 2025) are used.

- The LALMs used are `Qwen2.5-Omni` (Xu et al., 2025), `Kimi-Audio` (Ding et al., 2025), `Audio Flamingo 3` (Ghosh and Duraiswami, 2025) and `Qwen-Omni-3` (Xu et al., 2025).

5.4 Uncertainty estimation methods

We use two different uncertainty estimation methods. Firstly, we utilize semantic entropy (SE) (Kuhn et al., 2023), which calculates the entropy in the semantic space of open-text predictions from multimodal LLMs, instead of the lexical space. In the black-box settings, SE is computed assuming all sampling responses from the MLLM are equally likely, as the log-likelihoods are not available. This is referred to as discrete semantic entropy (DSE). Another approach to measuring uncertainty is to find the deviation of equivalent/complimentary sampling responses from the greedy decoded response (Bhattacharya et al., 2025b) (FESTA). For the object counting task, where complementary data generation is challenging, the FESTA approach uses only equivalent samples. The rest of the experiments use both equivalent and complementary samples. Table 1 and Table 2 use 30 sampling responses for both methods. The effect of the number of sampling responses used for uncertainty estimation is studied in Appendix Section 6.7.

6 Results

We evaluate continuous uncertainty across both audio-QA and vision-QA tasks, as discussed in Section 4. We report results on vision-QA tasks in Table 1 and audio-QA tasks in Table 2. Across all combinations of tasks and MLLMs, we report the monotonicity violation losses $\mathcal{L}_{local}$ and $\mathcal{L}_{global}$ (Equation 1) for two different uncertainty estimators DSE and FESTA. We report the abstention AUCs obtained using them. The task accuracies of different MLLMs are also included.

6.1 Vision-QA: spatial location

For the task based on spatial location dimension in Table 1, near-perfect task accuracy is offered by most LVLMs except `Gemma-3` (81.33%). Although `Gemma-3` offers much worse accuracy as compared to others, it offers second-best $\mathcal{L}_{global}$ using DSE. Also, `Phi-4` and `Qwen-2.5-VL`, despite having near-perfect accuracy, have significant monotonicity violation losses observed using FESTA. These observations highlight why monotonicity studies are necessary, even among high-performance models, $\mathcal{L}_{\text{global}}$ varies substantially, indicating different degrees of transition localization. `LlaVa-1.6` exhibits the best monotonicity behavior as evident from the minimum $\mathcal{L}_{\text{global}}$ loss

Table 1: Evaluation of monotonicity of FESTA and DSE based uncertainty scores for vision-QA tasks.

MLLMs	Acc.	DSE			FESTA		
		AUC (↑)	L_{local} (↓)	L_{global} (↓)	AUC (↑)	L_{local} (↓)	L_{global} (↓)
Spatial location dimension							
`LlaVa-1.6`	95.67	61.1	0.0097	0.4261	98.92	0.0359	0.1882
`Phi-4`	99.67	50.0	0.0000	0.5000	98.99	0.0304	0.3139
`Qwen-2.5-VL`	99.33	50.0	0.0000	0.5000	98.98	0.0334	0.3388
`Pixtral`	97.00	50.0	0.0000	0.5000	99.49	0.0285	0.2700
`Gemma-3`	81.33	68.6	0.0396	0.4831	96.70	0.0328	0.3047
Avg.	94.60	55.9	0.0099	0.4818	98.62	0.0322	0.2831
Scale dimension							
`LlaVa-1.6`	54.00	68.6	0.0907	0.6622	87.83	0.0559	0.5258
`Phi-4`	84.33	77.5	0.0631	0.3472	83.91	0.0297	0.3004
`Qwen-2.5-VL`	84.33	80.1	0.0368	0.3409	89.46	0.0306	0.1954
`Pixtral`	88.67	70.7	0.0323	0.3846	81.65	0.0370	0.2793
`Gemma-3`	73.33	56.4	0.0535	0.5028	89.79	0.0375	0.2530
Avg.	76.93	70.7	0.0553	0.4475	86.53	0.0381	0.3108
Count							
`LlaVa-1.6`	100.00	–	0.0855	0.3575	–	0.1000	0.4110
`Phi-4`	100.00	–	0.0882	0.4690	–	0.1013	0.3337
`Qwen-2.5-VL`	99.00	99.5	0.1911	0.6724	46.46	0.1319	0.6537
`Pixtral`	99.00	100.0	0.0459	0.4926	98.99	0.0615	0.3267
`Gemma-3`	70.00	59.1	0.0782	0.4128	81.76	0.1343	0.5816
Avg.	93.60	86.2*	0.0978	0.4809	75.74*	0.1058	0.4613

Table 2: Evaluation of monotonicity of FESTA and DSE based uncertainty scores for Audio-QA tasks.

MLLMs	Acc.	DSE			FESTA		
		AUC (↑)	L_{local} (↓)	L_{global} (↓)	AUC (↑)	L_{local} (↓)	L_{global} (↓)
Temporal location dimension							
`Audio-Flam.3`	89.50	90.35	0.0492	0.2707	88.30	0.0354	0.2185
`Qwen2.5-Omni`	87.00	78.51	0.0957	0.5190	99.56	0.0285	0.2075
`Qwen-Omni-3`	96.00	77.22	0.0255	0.4199	98.50	0.0261	0.3152
`Kimi-Audio`	68.50	72.05	0.0982	0.3599	84.10	0.0474	0.2939
Avg.	85.25	79.53	0.0671	0.3924	92.61	0.0343	0.2588
Number of distractors							
`Audio-Flam.3`	87.14	62.80	0.0525	0.3483	70.00	0.0411	0.4277
`Qwen2.5-Omni`	70.00	66.25	0.0916	0.6227	93.26	0.0467	0.4126
`Qwen-Omni-3`	92.86	88.49	0.0739	0.5472	85.83	0.0315	0.3725
`Kimi-Audio`	56.43	58.58	0.0878	0.4145	78.36	0.0628	0.4317
Avg.	76.61	69.03	0.0765	0.4832	81.86	0.0455	0.4111
Volume dimension							
`Audio-Flam.3`	57.50	50.45	0.0758	0.4559	52.08	0.0664	0.4865
`Qwen2.5-Omni`	59.00	60.74	0.0657	0.4223	65.20	0.0550	0.3788
`Qwen-Omni-3`	54.50	67.44	0.0589	0.4656	83.51	0.0526	0.3182
`Kimi-Audio`	53.50	60.14	0.0700	0.5089	70.38	0.0671	0.5049
Avg.	56.13	59.69	0.0676	0.4637	67.79	0.0603	0.4221

using both FESTA and DSE.

6.2 Vision-QA: scale dimension

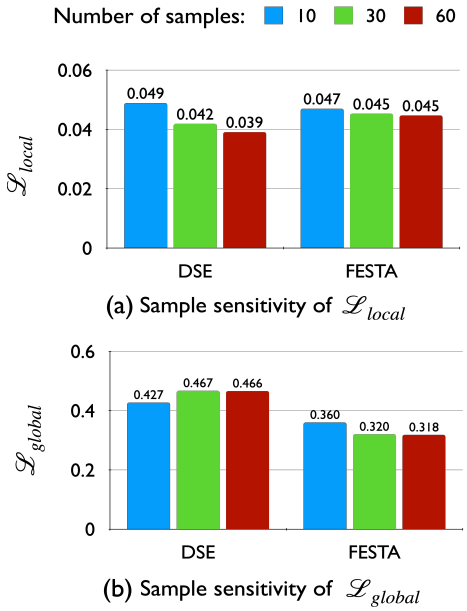
For the task based on the scale dimension in Table 1, the LVLMs show more spread in task accuracy than spatial location, with `Pixtral` achieving the best accuracy and `LlaVa-1.6` the lowest. However, accuracy does not directly translate to monotonicity: `Qwen-2.5-VL` exhibits the best trajectory-faithfulness under FESTA, achieving the minimum $\mathcal{L}_{\text{global}}$ (0.1954). In contrast, `LlaVa-1.6` shows the largest global deviation under both DSE and FESTA ($\mathcal{L}_{\text{global}} = 0.6622$ and 0.5258), indicating that its uncertainty peak is poorly localized around the equal-size transition. These observations again highlight why monotonicity studies are necessary: even among models with similar accuracy (e.g., `Phi-4` and `Qwen-2.5-VL`), the global violation losses can differ substantially, reflecting different degrees of transition localization.

6.3 Vision-QA: count

For the count task in Table 1, several LVLMs reach saturated task accuracy (e.g., `LlaVa-1.6` and `Phi-4`), making the AUC undefined and reducing the diagnostic value of ranking-based metrics in those cases. In this regime, monotonicity losses become the primary evidence of whether uncertainty meaningfully tracks increasing cluttering of distractors. Under FESTA, `Pixtral` exhibits the best monotonicity behavior with the minimum $\mathcal{L}_{\text{global}}$ (0.3267), whereas `Qwen-2.5-VL` shows the largest global deviation (0.6537) despite near-perfect accuracy, indicating a diffuse or mislocalized uncertainty profile at maximum number of distractors. Moreover, `Gemma-3` achieves lower accuracy but still incurs high $\mathcal{L}_{\text{global}}$ under FESTA (0.5816), reinforcing that both correctness and monotonicity can degrade independently. Overall, these results motivate reporting $\mathcal{L}_{\text{global}}$ alongside accuracy, since it remains informative even when accuracy saturates, and AUC is undefined.

6.4 Audio-QA: Temporal location dimension

For temporal ordering tasks (Table 2), most audio LALMs achieve strong accuracy, with `Qwen-Omni-3` performing best. However, accuracy alone does not fully reflect the reliability of evaluations. The monotonicity metrics in Table 2 reveal clear differences in how uncertainty evolves along temporal trajectories. With FESTA, `Audio-Flam.3` and `Qwen2.5-Omni` exhibit more localized and smoother transitions, whereas even high-accuracy models show larger global deviations, indicating less faithful temporal uncertainty localization. This gap between correctness and trajectory behavior motivates monotonicity analysis for audio reasoning. Notably, `Kimi-Audio` exhibits lower accuracy and reduced continuity, suggesting limited temporal reasoning capability across all the



(a) Sample sensitivity of $\mathscr{L}_{local}$

(b) Sample sensitivity of $\mathscr{L}_{global}$

Figure 3: Sample sensitivity for uncertainty estimators across vision QA tasks.

metrics measured.

6.5 Audio-QA: Distractor dimension

For audio tasks with increasing numbers of distractors (Table 2), accuracy drops and varies widely across models, indicating that presence of more distractors significantly increases task difficulty. Under FESTA, models exhibit substantially different uncertainty behaviors, with `Qwen2.5-Omni` showing stronger abstention skills despite reduced moderate accuracy, while other high-accuracy models display weaker localization of uncertainty.

6.6 Audio-QA: Volume dimension

For loudness-based reasoning tasks (Table 2), overall accuracy is lower than in temporal and distractor-based settings, indicating increased difficulty in modeling volume transitions. The monotonicity metrics reveal substantial variation in uncertainty behavior across models, even where accuracy is comparable. Under FESTA, some models exhibit smoother and more localized uncertainty evolution, while others show reduced continuity, suggesting weaker alignment with gradual loudness changes. These results further demonstrate that performance accuracy alone is insufficient to characterize reasoning quality for volume-based audio tasks. Overall, the consistently low accuracy and AUC in this setting suggest that current audio LALMs have limited capability to reason reliably about loudness differences, making volume-based reasoning a particularly challenging dimension.

6.7 Effect of number of samples

The uncertainty estimators either sample the output distribution (DSE) or both input and output distributions (FESTA). However, both these estimators are sensitive to the number of samples that allow a stable estimation of the uncertainty measure. We perform a sensitivity analysis using different number of output or input/output samples for both these estimators. The analysis is reported in Figure 3 for vision-QA tasks. In particular, both the local and the global measures are seen to be relatively stable with varying number of samples. A higher number of samples gives a slight improvement in most cases except for the DSE approach with $\mathcal{L}_{global}$. Further analysis is included in Section B.

7 Discussion

The probes on multimodal LLMs for their uncertainty evaluation across axes of complexity show that the uncertainty patterns often suffer from lack of monotonicity, evident from high monotonicity violation losses. Moreover, we also highlight that high accuracy and near-perfect model performance may often mask real-world human-like behavioral profiles that are expected from the MLLMs. While our analysis probed only a subset of dimensions, it lays out an open call for modelling efforts to analyze MLLMs on diverse axes of complexity. Further, smoothness and continuity are often desired properties as models make transitions from static images and audios to multi-input and video based scene understanding. Smooth uncertainty calibration along various axes of discriminability is highly sought after because of its potential to work as a trigger mechanism in various applications, such as robotic applications (Hsiao et al., 2011; Kim et al., 2012), and autonomous vehicle functionality (Basavaraj et al., 2023), among others.

8 Conclusion

We propose a synthetic benchmark for evaluating uncertainty in multimodal LLMs for inputs with varying discriminability. Our experiments with vision-LLMs and audio-LLMs have shown that many of them fail to demonstrate monotonic uncertainty behaviors with a gradual increase/decrease in input ambiguity, which violates the fundamental notion of uncertainty. Our evaluation protocol points to a critical limitation in uncertainty behavior of multimodal LLMs and calls for modelling efforts to address this.

Limitations

Although the proposed evaluation and benchmarking of continuous evaluations of MLLM uncertainty finds intriguing observations, it has the following limitations and opens scope for future work.

1. The proposed study explores a limited number of dimensions of discriminability/ambiguity across vision-language and audio-language tasks. The benchmarking can further be enhanced to more sophisticated tasks and associated ambiguity.

2. As compared to simple performance metrics like accuracy, computation of monotonicity violation losses involves much larger computation because for a single data point, multiple sweeping scenes/audios are created along the dimension of discriminability.

3. The proposed benchmark has potential applications in robotics, such as object handling using a robotic arm, and in autonomous driving, including path following, among others. This work has not covered such applications.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, and 1 others. 2024. Pixtral 12b. *arXiv preprint arXiv:2410.07073*.

Meghana Basavaraj, Upendra Suddamalla, and Shenxin Xu. 2023. Lanenet++: Uncertainty-aware lane detection for autonomous vehicle. In *International Symposium on Visual Computing*, pages 245–258. Springer.

Debarpan Bhattacharya, Apoorva Kulkarni, and Sriram Ganapathy. 2025a. Benchmarking and Confidence Evaluation of LALMs For Temporal Reasoning. In *Interspeech 2025*, pages 2068–2072.

Debarpan Bhattacharya, Apoorva Kulkarni, and Sriram Ganapathy. 2025b. Festa: Functionally equivalent sampling for trust assessment of multimodal llms. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12277–12295.

Blender Online Community. 2025. Blender - a 3d modelling and rendering package. https://www.blender.org/. Blender Foundation. Accessed: 2025-07-21.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, and 1 others. 2025. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer.

Sreyan Ghosh and Ramani Duraiswami. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. In *TTIC Summer Workshop on Foundations of Speech and Audio Foundation Models 2025*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*.

Kaijen Hsiao, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. 2011. Robust grasping under object pose uncertainty. *Autonomous Robots*, 31(2):253–268.

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.

Zaid Khan and Yun Fu. 2024. Consistency and uncertainty: Identifying unreliable responses from black-box vision-language models for selective visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10854–10863.

Junggon Kim, Kunihiro Iwamoto, James J Kuffner, Yasuhiro Ota, and Nancy S Pollard. 2012. Physically-based grasp quality evaluation under uncertainty. In *2012 IEEE International Conference on Robotics and Automation*, pages 3258–3263. IEEE.

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

Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. 2023. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14963–14973.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2023a. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651.

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

Pengqian Lu, Jie Lu, Anjin Liu, and Guangquan Zhang. 2025. Early concept drift detection via prediction uncertainty. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19124–19132.

Dang Nguyen, Ali Payani, and Baharan Mirzasoleiman. 2025. Beyond semantic entropy: Boosting llm uncertainty quantification with pairwise semantic similarity. *arXiv preprint arXiv:2506.00245*.

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

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Sebastian Nowozin, Joshua V Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems*, 32:13991–14002.

Karol J. Piczak. 2015. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press.

James Robert. 2014. pydub: Manipulate audio with a simple and easy high level interface. https://github.com/jiaaro/pydub. Accessed: 2025-07-21.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2025. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *The Thirteenth International Conference on Learning Representations*.

Adi Simhi, Itay Itzhak, Fazl Barez, Gabriel Stanovsky, and Yonatan Belinkov. 2025. Trust me, i'm wrong: High-certainty hallucinations in llms. *arXiv preprint arXiv:2502.12964*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

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

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhuo Zhi, Chen Feng, Adam Daneshmend, Mine Orlu, Andreas Demosthenous, Lu Yin, Da Li, Ziquan Liu, and Miguel RD Rodrigues. 2025. Seeing and reasoning with confidence: Supercharging multimodal llms with an uncertainty-aware agentic framework. *arXiv preprint arXiv:2503.08308*.

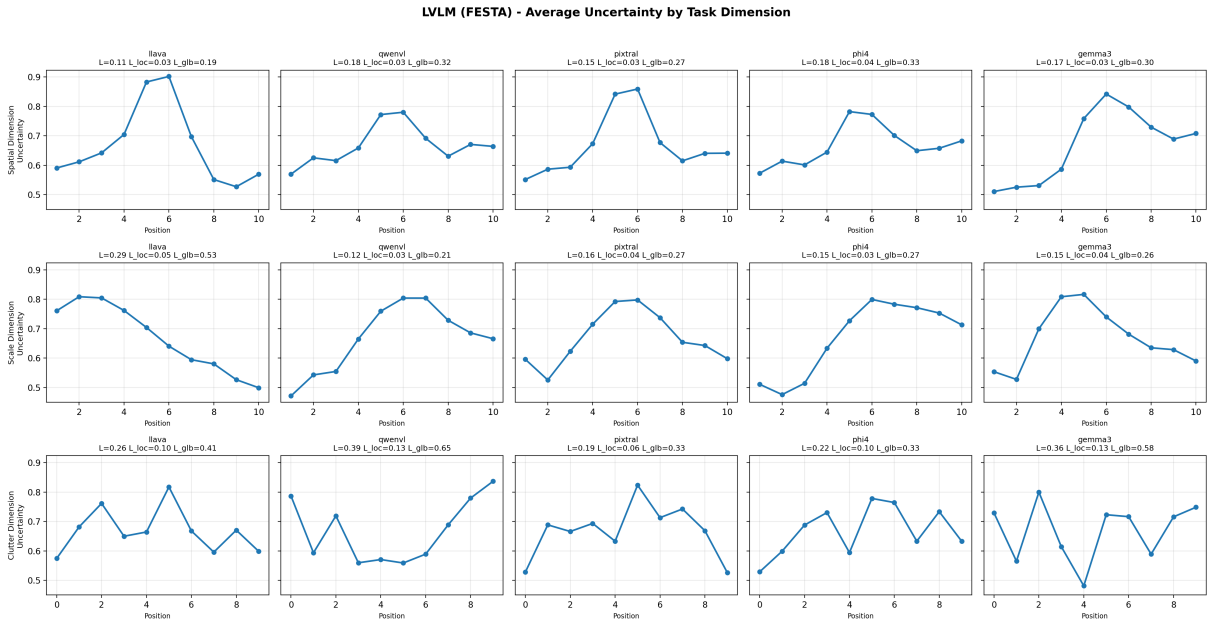


Figure 4: LVLMs (FESTA): average uncertainty trajectories across task dimensions (spatial translation, scale, clutter).

A Appendix

A.1 Average Uncertainty Plots

Figures 4–6 visualize the *average* uncertainty values along each continuous trajectory step (x-axis), aggregated across all scenes/questions within a task variant. These plots provide a qualitative complement to the quantitative monotonicity losses: a trajectory-faithful estimator should exhibit a smooth "approach–transition–depart" profile with uncertainty peaking near the designed transition point. For LVLMs, we report a single grid figure covering all three image axes (spatial location, scale, count dimensions) for both FESTA and DSE. Similarly for Audio-LLMs, we present (DSE vs. FESTA) aligned per-task pairs for a direct visual comparison of peak localization and smoothness across estimators.

B Sampling Ablation

We study the effect of sample count used by sampling-based uncertainty estimation by sweeping the number of samples from 10 to 60. Increasing the number of samples reduces the *local* monotonicity violations for both estimators (DSE: 0.0489→0.0391; FESTA: 0.0470→0.0447), suggesting improved step-to-step consistency of the uncertainty trajectory. For the *global* loss, FESTA exhibits a clear improvement (0.3596→0.3182), while DSE shows a modest increase from 10 to 30 samples (0.4268→0.4670) and then stabilizes (0.4657 at 60). Overall, FESTA appears more stable under increased sampling, whereas entropy-based scoring shows sensitivity in global ordering despite improved local smoothness.
The Tables 3- 20 reports aggregated performance under three sampling regimes ($n \in \{10, 30, 60\}$) for both image and audio settings.

Audio tasks. We report two audio tasks: *Temporal Reasoning* and *Count Reasoning*.

Image tasks. We report three image tasks as discussed before: *Spatial Reasoning*, *Scale Reasoning*, and *Count Reasoning*. Each reported entry corresponds to a *model-level average across task complexities* available in the runs for that task and n.

Metrics. We report Accuracy (Acc) and AUC for the two uncertainty measures: *DSE* and *FESTA*. Additionally, we report local and global loss terms L_{local} and L_{global} for both variants.

C Variants

C.1 Visual variations used in the synthetic scenes

Figures 7–9 illustrate the image variations used by the three sweeps. Each figure is a 10×10 grid of example scenes. Across all grids, **columns** (left→right) correspond to increasing values of the

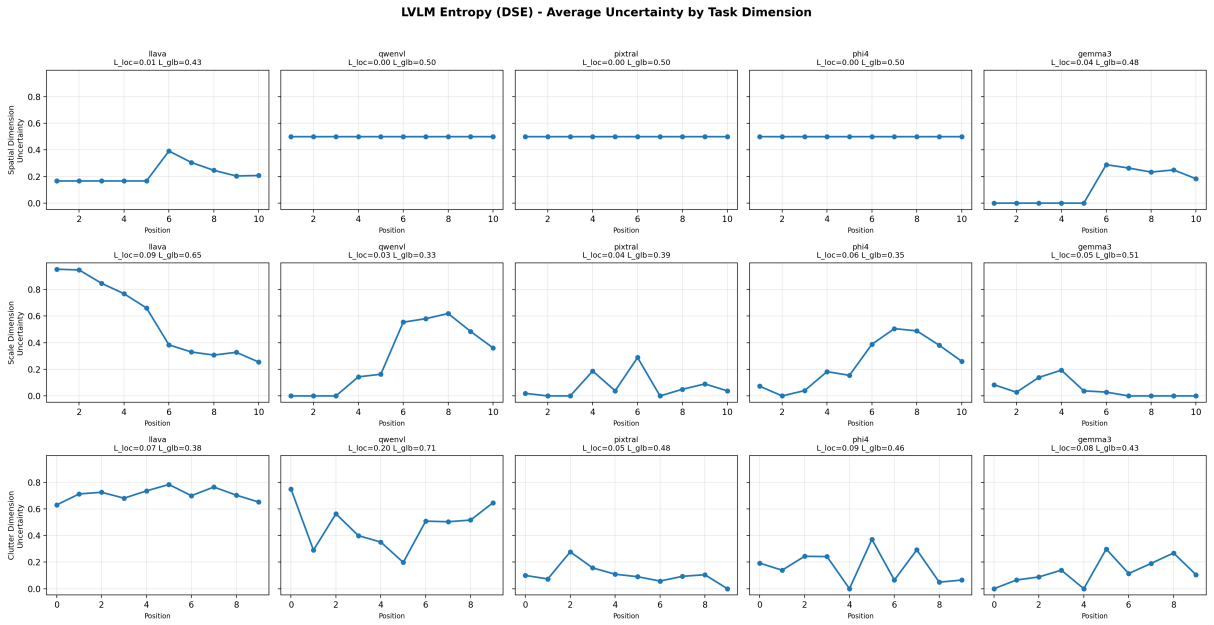


Figure 5: LVLMs (DSE/Entropy): average uncertainty trajectories across task dimensions (spatial translation, scale, clutter).

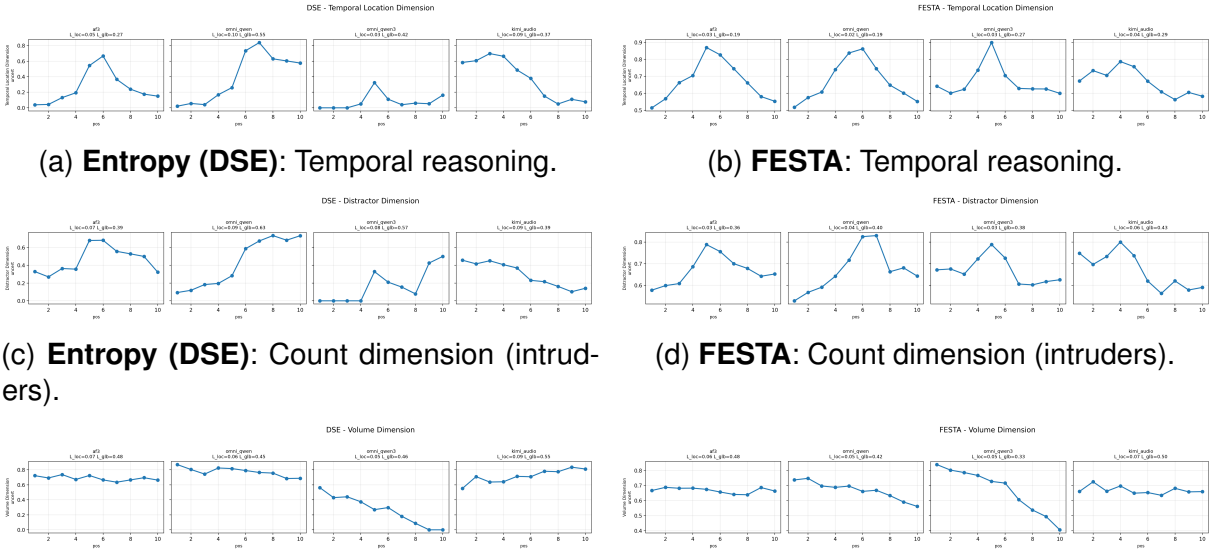
Table 3: Sample count $n = 10$; Temporal Reasoning.

Model	Acc	AUC_{DSE}	$\text{AUC}_{\text{FESTA}}$	$L_{\text{local,DSE}}$	$L_{\text{local,FESTA}}$	$L_{\text{global,DSE}}$	$L_{\text{global,FESTA}}$
`Audio-Flam.3`	0.8950	0.8889	0.8033	0.0700	0.0457	0.3071	0.2316
`Qwen2.5-Omni`	0.8700	0.7701	0.9841	0.1271	0.0518	0.5172	0.2931
`Qwen-Omni-3`	0.9600	0.8709	0.8786	0.0294	0.0213	0.3810	0.3108
`Kimi-Audio`	0.6850	0.7092	0.7337	0.0933	0.0576	0.3808	0.3781

sweep parameter, while **rows** show different underlying templates/complexities (e.g., different object type/color combinations and, in some cases, additional non-target objects).

Row legend for Fig. 7 (Spatial Reasoning, `basic_basic`). Rows correspond to the 10 sub-variations `Im01`–`Im10`.
Row 1 (Im01, `original`): clean baseline with default pair/colors (e.g., green sphere as source; brown cylinder as destination).
Row 2 (Im02, `color_both`): both source and destination colors are changed (randomly selected).
Row 3 (Im03, `size_up_src`): source object scaled up (2× larger) while destination unchanged.
Row 4 (Im04, `size_down_dst`): destination object scaled down (0.5× smaller) while source unchanged.
Row 5 (Im05, `shape_src`): source shape changed: (sphere→cube); destination unchanged.
Row 6 (Im06, `shape_dst`): destination shape changed: (cylinder→cube); source unchanged.
Row 7 (Im07, `shape_both`): both shapes changed: sphere→cube and cylinder→sphere).
Row 8 (Im08, `intruder_one`): add one cube intruder (different color), placed off the sweep axis.
Row 9 (Im09, `intruder_two`): add two cube intruders (different colors), placed off the sweep axis.
Row 10 (Im10, `move_dst_up`): destination object shifted diagonally (upward), so the comparison is not purely horizontal.

Row legend for Fig. 8 (Scale Reasoning, `basic_basic`). Rows correspond to the 10 sub-variations `Im01`–`Im10`.
Row 1 (Im01, `original`): baseline size-comparison setup (left object fixed size; right object varies across columns).
Row 2 (Im02, `color_both`): both objects recolored (different colors).
Row 3 (Im03, `shape_src`): left object shape changed: (sphere→cube).
Row 4 (Im04, `shape_dst`): right object shape changed: (cylinder→cube).



(a) **Entropy (DSE)**: Temporal reasoning.

(b) **FESTA**: Temporal reasoning.

(c) **Entropy (DSE)**: Count dimension (intruders).

(d) **FESTA**: Count dimension (intruders).

(e) **Entropy (DSE)**: Amplitude/volume reasoning.

(f) **FESTA**: Amplitude/volume reasoning.

Figure 6: Audio-LLMs: average uncertainty trajectories for each audio task dimension, shown as aligned pairs (Entropy/DSE vs. FESTA) to facilitate direct visual comparison of trajectory shape and peak localization.

Table 4: Sample count $n = 10$; Count Reasoning.

Model	Acc	AUC_{DSE}	$\text{AUC}_{\text{FESTA}}$	$L_{\text{local,DSE}}$	$L_{\text{local,FESTA}}$	$L_{\text{global,DSE}}$	$L_{\text{global,FESTA}}$
`Audio-Flam.3`	0.8714	0.6755	0.7206	0.1123	0.0620	0.3900	0.4220
`Qwen2.5-Omni`	0.7000	0.6890	0.9140	0.1190	0.0474	0.6237	0.3863
`Qwen-Omni-3`	0.9286	0.8056	0.8594	0.0752	0.0353	0.5636	0.3845
`Kimi-Audio`	0.5643	0.7145	0.6384	0.0889	0.0658	0.3203	0.3786

Row 5 (Im05, `shape_both`): both shapes changed (sphere→cube, cylinder→sphere).
Row 6 (Im06, `intruder_one`): add one cube intruder in the foreground.
Row 7 (Im07, `intruder_two`): add two cube intruders in the foreground.
Row 8 (Im08, `swap_positions`): swap left/right object positions (cylinder on the left, sphere on the right), inverting the comparison logic.
Row 9 (Im09, `color_src`): only the left object color is changed; right object remains as in the baseline.
Row 10 (Im10, `size_both_change`): both objects change size in opposite directions across columns (one increases while the other decreases).

Row legend for Fig. 9 (Clutter Reasoning, `basic_basic`). Rows correspond to the 10 sub-variations `01`–`10`
Row 1 (01, `original`): baseline (3 blue cylinders as focus; distractors are random spheres/cubes with colors from a fixed pool).
Row 2 (02, `color_both`): recolor both focus objects (e.g., blue→red) and the distractor color palette.
Row 3 (03, `color_focus`): recolor focus objects only; distractors unchanged from baseline.
Row 4 (04, `color_distractor`): change distractor colors only; focus unchanged from baseline.
Row 5 (05, `shape_focus`): change focus shape (cylinder→cube); distractor shapes exclude the focus shape to keep counting unambiguous.
Row 6 (06, `shape_distractor`): restrict distractors to a single shape class (e.g., spheres only).
Row 7 (07, `size_up_focus`): scale up focus objects (1.3× larger).
Row 8 (08, `size_down_focus`): scale down focus objects (0.7× smaller).
Row 9 (09, `size_up_distractor`): scale up distractor objects (1.3× larger).
Row 10 (10, `material_change`): change material globally (e.g., metal→rubber).

Table 5: Sample count $n = 10$; Volume Reasoning.

Model	Acc	$\mathrm{AUC}_{\mathrm{DSE}}$	$\mathrm{AUC}_{\mathrm{FESTA}}$	$L_{\mathrm{local,DSE}}$	$L_{\mathrm{local,FESTA}}$	$L_{\mathrm{global,DSE}}$	$L_{\mathrm{global,FESTA}}$
Audio-Flam.3	0.5750	0.4878	0.5735	0.1218	0.0707	0.5251	0.4699
Qwen2.5-Omni	0.5900	0.5760	0.5471	0.1077	0.0756	0.4551	0.4617
Qwen-Omni-3	0.5450	0.6406	0.7725	0.0534	0.0602	0.4361	0.3597
Kimi-Audio	0.5350	0.5322	0.6466	0.1208	0.0760	0.5473	0.5281

Table 6: Sample count $n = 30$; Temporal Reasoning.

Model	Acc	$\mathrm{AUC}_{\mathrm{DSE}}$	$\mathrm{AUC}_{\mathrm{FESTA}}$	$L_{\mathrm{local,DSE}}$	$L_{\mathrm{local,FESTA}}$	$L_{\mathrm{global,DSE}}$	$L_{\mathrm{global,FESTA}}$
Audio-Flam.3	0.8950	0.9035	0.8830	0.0492	0.0354	0.2707	0.2185
Qwen2.5-Omni	0.8700	0.7851	0.9956	0.0957	0.0285	0.5190	0.2075
Qwen-Omni-3	0.9600	0.7722	0.9850	0.0255	0.0261	0.4199	0.3152
Kimi-Audio	0.6850	0.7205	0.8410	0.0982	0.0474	0.3599	0.2939

D Computation hardware details

A cluster with 8 X Nvidia RTX A6000 cards is used for the experiments.

Table 7: Sample count $n = 30$; Count Reasoning.

Model	Acc	$\mathrm{AUC}_{\mathrm{DSE}}$	$\mathrm{AUC}_{\mathrm{FESTA}}$	$L_{\mathrm{local,DSE}}$	$L_{\mathrm{local,FESTA}}$	$L_{\mathrm{global,DSE}}$	$L_{\mathrm{global,FESTA}}$
Audio-Flam.3	0.8714	0.6280	0.7000	0.0525	0.0411	0.3483	0.4277
Qwen2.5-Omni	0.7000	0.6625	0.9326	0.0916	0.0467	0.6227	0.4126
Qwen-Omni-3	0.9286	0.8849	0.8583	0.0739	0.0315	0.5472	0.3725
Kimi-Audio	0.5643	0.5858	0.7836	0.0878	0.0628	0.4145	0.4317

Table 8: Sample count $n = 30$; Volume Reasoning.

Model	Acc	$\mathrm{AUC}_{\mathrm{DSE}}$	$\mathrm{AUC}_{\mathrm{FESTA}}$	$L_{\mathrm{local,DSE}}$	$L_{\mathrm{local,FESTA}}$	$L_{\mathrm{global,DSE}}$	$L_{\mathrm{global,FESTA}}$
Audio-Flam.3	0.5750	0.5045	0.5208	0.0758	0.0664	0.4559	0.4865
Qwen2.5-Omni	0.5900	0.6074	0.6520	0.0657	0.0550	0.4223	0.3788
Qwen-Omni-3	0.5450	0.6744	0.8351	0.0589	0.0526	0.4656	0.3182
Kimi-Audio	0.5350	0.6014	0.7038	0.0700	0.0671	0.5089	0.5049

Table 9: Sample count $n = 60$; Temporal Reasoning.

Model	Acc	$\mathrm{AUC}_{\mathrm{DSE}}$	$\mathrm{AUC}_{\mathrm{FESTA}}$	$L_{\mathrm{local,DSE}}$	$L_{\mathrm{local,FESTA}}$	$L_{\mathrm{global,DSE}}$	$L_{\mathrm{global,FESTA}}$
Audio-Flam.3	0.8950	0.9135	0.9055	0.0475	0.0291	0.2536	0.1852
Qwen2.5-Omni	0.8700	0.7991	0.9971	0.0940	0.0236	0.5353	0.1898
Qwen-Omni-3	0.9600	0.8351	0.9888	0.0246	0.0264	0.4178	0.2687
Kimi-Audio	0.6850	0.7207	0.8598	0.0802	0.0421	0.3508	0.2925

Table 10: Sample count $n = 60$; Count Reasoning.

Model	Acc	$\mathrm{AUC}_{\mathrm{DSE}}$	$\mathrm{AUC}_{\mathrm{FESTA}}$	$L_{\mathrm{local,DSE}}$	$L_{\mathrm{local,FESTA}}$	$L_{\mathrm{global,DSE}}$	$L_{\mathrm{global,FESTA}}$
Audio-Flam.3	0.8714	0.6451	0.7875	0.0575	0.0336	0.3853	0.3640
Qwen2.5-Omni	0.7000	0.6656	0.9679	0.0806	0.0399	0.6319	0.3992
Qwen-Omni-3	0.9286	0.8901	0.8821	0.0760	0.0329	0.5574	0.3815
Kimi-Audio	0.5643	0.6212	0.7965	0.0681	0.0633	0.3919	0.4334

Table 11: Sample count $n = 60$; Volume Reasoning.

Model	Acc	$\mathrm{AUC}_{\mathrm{DSE}}$	$\mathrm{AUC}_{\mathrm{FESTA}}$	$L_{\mathrm{local,DSE}}$	$L_{\mathrm{local,FESTA}}$	$L_{\mathrm{global,DSE}}$	$L_{\mathrm{global,FESTA}}$
Audio-Flam.3	0.5750	0.4981	0.5282	0.0590	0.0611	0.4635	0.4831
Qwen2.5-Omni	0.5900	0.6295	0.6581	0.0531	0.0457	0.4022	0.4248
Qwen-Omni-3	0.5450	0.6891	0.8509	0.0527	0.0472	0.4774	0.3294
Kimi-Audio	0.5350	0.5702	0.7201	0.0664	0.0708	0.5363	0.4986

Table 12: Sample count $n = 10$; Spatial Reasoning. Metrics are averaged across available complexities per model.

Model	Acc	$\mathrm{AUC}_{\mathrm{DSE}}$	$\mathrm{AUC}_{\mathrm{FESTA}}$	$L_{\mathrm{local,DSE}}$	$L_{\mathrm{local,FESTA}}$	$L_{\mathrm{global,DSE}}$	$L_{\mathrm{global,FESTA}}$
LlaVa-1.6	0.96	0.9571	0.9694	0.0081	0.0381	0.2179	0.2674
Phi-4	1.00	0.4798	1.0000	0.0000	0.0223	0.4371	0.3634
Qwen-2.5-VL	0.99	0.9745	1.0000	0.0000	0.0282	0.4607	0.3978
Pixtral	0.97	0.9420	0.9849	0.0035	0.0295	0.3018	0.3296
Gemma-3	0.81	0.6968	0.9560	0.0768	0.0367	0.4208	0.3215

Table 13: Sample count $n = 10$; Scale Reasoning. Metrics are averaged across available complexities per model.

Model	Acc	$\mathrm{AUC}_{\mathrm{DSE}}$	$\mathrm{AUC}_{\mathrm{FESTA}}$	$L_{\mathrm{local,DSE}}$	$L_{\mathrm{local,FESTA}}$	$L_{\mathrm{global,DSE}}$	$L_{\mathrm{global,FESTA}}$
LlaVa-1.6	0.54	0.6512	0.8259	0.1053	0.0724	0.6413	0.5601
Phi-4	0.84	0.7582	0.8482	0.0649	0.0379	0.3563	0.3147
Qwen-2.5-VL	0.84	0.7984	0.8873	0.0441	0.0345	0.3258	0.2704
Pixtral	0.89	0.6669	0.7919	0.0346	0.0385	0.3965	0.3052
Gemma-3	0.73	0.5205	0.8631	0.0456	0.0367	0.5217	0.2518

Table 14: Sample count $n = 10$; Count Reasoning. Metrics are averaged across available complexities per model.

Model	Acc	$\mathrm{AUC}_{\mathrm{DSE}}$	$\mathrm{AUC}_{\mathrm{FESTA}}$	$L_{\mathrm{local,DSE}}$	$L_{\mathrm{local,FESTA}}$	$L_{\mathrm{global,DSE}}$	$L_{\mathrm{global,FESTA}}$
LlaVa-1.6	1.00	—	—	0.1597	0.1116	0.6119	0.4528
Phi-4	1.00	—	—	0.0716	0.0843	0.4987	0.4116
Qwen-2.5-VL	0.99	0.8990	0.7879	0.1947	0.1398	0.5933	0.6242
Pixtral	0.99	1.0000	0.9697	0.0526	0.0636	0.4906	0.4175
Gemma-3	0.70	0.5593	0.8086	0.0829	0.1213	0.5024	0.5344

Table 15: Sample count $n = 30$; Spatial Reasoning. Metrics are averaged across available complexities per model.

Model	Acc	$\mathrm{AUC}_{\mathrm{DSE}}$	$\mathrm{AUC}_{\mathrm{FESTA}}$	$L_{\mathrm{local,DSE}}$	$L_{\mathrm{local,FESTA}}$	$L_{\mathrm{global,DSE}}$	$L_{\mathrm{global,FESTA}}$
LlaVa-1.6	0.96	0.6115	0.9892	0.0097	0.0359	0.4261	0.1882
Phi-4	1.00	0.5000	0.9899	0.0000	0.0304	0.5000	0.3139
Qwen-2.5-VL	0.99	0.5000	0.9898	0.0000	0.0334	0.5000	0.3388
Pixtral	0.97	0.5000	0.9949	0.0000	0.0285	0.5000	0.2700
Gemma-3	0.81	0.6858	0.9670	0.0396	0.0328	0.4831	0.3047

Table 16: Sample count $n = 30$; Scale Reasoning. Metrics are averaged across available complexities per model.

Model	Acc	$\mathrm{AUC}_{\mathrm{DSE}}$	$\mathrm{AUC}_{\mathrm{FESTA}}$	$L_{\mathrm{local,DSE}}$	$L_{\mathrm{local,FESTA}}$	$L_{\mathrm{global,DSE}}$	$L_{\mathrm{global,FESTA}}$
LlaVa-1.6	0.54	0.6863	0.8783	0.0907	0.0559	0.6622	0.5258
Phi-4	0.84	0.7745	0.8390	0.0631	0.0297	0.3472	0.3004
Qwen-2.5-VL	0.84	0.8006	0.8946	0.0368	0.0306	0.3409	0.1954
Pixtral	0.89	0.7070	0.8165	0.0323	0.0370	0.3846	0.2793
Gemma-3	0.73	0.5636	0.8979	0.0535	0.0375	0.5028	0.2530

Table 17: Sample count $n = 30$; Count Reasoning. Metrics are averaged across available complexities per model.

Model	Acc	$\mathrm{AUC}_{\mathrm{DSE}}$	$\mathrm{AUC}_{\mathrm{FESTA}}$	$L_{\mathrm{local,DSE}}$	$L_{\mathrm{local,FESTA}}$	$L_{\mathrm{global,DSE}}$	$L_{\mathrm{global,FESTA}}$
LlaVa-1.6	1.00	—	—	0.0855	0.1000	0.3575	0.4110
Phi-4	1.00	—	—	0.0882	0.1013	0.4690	0.3337
Qwen-2.5-VL	0.99	0.9949	0.4646	0.1911	0.1319	0.6724	0.6537
Pixtral	0.99	1.0000	0.9899	0.0459	0.0615	0.4926	0.3267
Gemma-3	0.70	0.5907	0.8176	0.0782	0.1343	0.4128	0.5816

Table 18: Sample count $n = 60$; Spatial Reasoning. Metrics are averaged across available complexities per model.

Model	Acc	$\mathrm{AUC}_{\mathrm{DSE}}$	$\mathrm{AUC}_{\mathrm{FESTA}}$	$L_{\mathrm{local,DSE}}$	$L_{\mathrm{local,FESTA}}$	$L_{\mathrm{global,DSE}}$	$L_{\mathrm{global,FESTA}}$
LlaVa-1.6	0.96	0.6098	0.9966	0.0072	0.0344	0.4249	0.1878
Phi-4	1.00	0.5000	1.0000	0.0000	0.0357	0.5000	0.3252
Qwen-2.5-VL	0.99	0.5000	1.0000	0.0000	0.0319	0.5000	0.3248
Pixtral	0.97	0.5000	0.9935	0.0000	0.0266	0.5000	0.2684
Gemma-3	0.81	0.6865	0.9731	0.0396	0.0319	0.4848	0.2994

Table 19: Sample count $n = 60$; Scale Reasoning. Metrics are averaged across available complexities per model.

Model	Acc	$\mathrm{AUC}_{\mathrm{DSE}}$	$\mathrm{AUC}_{\mathrm{FESTA}}$	$L_{\mathrm{local,DSE}}$	$L_{\mathrm{local,FESTA}}$	$L_{\mathrm{global,DSE}}$	$L_{\mathrm{global,FESTA}}$
LlaVa-1.6	0.54	0.6983	0.8867	0.0749	0.0499	0.6496	0.5306
Phi-4	0.84	0.7822	0.8396	0.0601	0.0289	0.3486	0.2701
Qwen-2.5-VL	0.84	0.8104	0.8885	0.0356	0.0299	0.3378	0.2070
Pixtral	0.89	0.7077	0.8149	0.0332	0.0399	0.3828	0.2728
Gemma-3	0.73	0.5631	0.9074	0.0547	0.0365	0.5028	0.2571

Table 20: Sample count $n = 60$; Count Reasoning. Metrics are averaged across available complexities per model.

Model	Acc	$\mathrm{AUC}_{\mathrm{DSE}}$	$\mathrm{AUC}_{\mathrm{FESTA}}$	$L_{\mathrm{local,DSE}}$	$L_{\mathrm{local,FESTA}}$	$L_{\mathrm{global,DSE}}$	$L_{\mathrm{global,FESTA}}$
LlaVa-1.6	1.00	—	—	0.0751	0.1000	0.3423	0.4110
Phi-4	1.00	—	—	0.0813	0.1013	0.4609	0.3337
Qwen-2.5-VL	0.99	0.9899	0.4646	0.1713	0.1319	0.6905	0.6537
Pixtral	0.99	1.0000	0.9899	0.0509	0.0615	0.5010	0.3267
Gemma-3	0.70	0.5883	0.8176	0.0746	0.1343	0.4096	0.5816

Figure 7: **Spatial Reasoning** The queried object is translated horizontally relative to a reference object (left→right across columns), spanning clearly-left, near-overlap/ambiguous midpoints, and clearly-right configurations.

Figure 8: **Scale Reasoning** The relative object scale is varied monotonically across columns (small→large), producing progressively larger/smaller target instances while keeping the overall scene layout consistent within each row. Rows show different base templates/complexities (different object pairs and, in some rows, additional objects).

Figure 9: **Count Reasoning** The number and composition of non-target (distractor) objects is varied across columns, increasing scene count and potential confounds while keeping the core target configuration consistent within each row. Rows correspond to different target templates (e.g., different target colors/shapes).