

Do LVLMs Know When They Are Wrong? Trajectory-Based Correctness Estimation from a Single Forward Pass

Awal Ahmed Fime, Md Zarif Hossain and Ahmed Imteaj

Florida Atlantic University

{afime2025, mdzarifhossa2025, aimteaj}@fau.edu

Abstract

Large vision-language models (LVLMs) exhibit strong multimodal reasoning capabilities, yet reliably estimating whether a generated response is correct remains a fundamental challenge. Existing confidence estimation methods commonly rely on repeated querying, self-consistency sampling, or layer-wise probing, which introduce additional inference overhead, self-preference bias, or costly full-depth analysis. In this work, we propose a trajectory-based framework¹ for estimating LVLM response correctness from a single forward generation pass. Instead of treating internal representations as isolated layer-wise signals, we model response generation as a structured trajectory through the LVLM, capturing the evolution of visual evidence in the vision encoder, geometric transformations induced by cross-modal projection, and probability dynamics within the language decoder. Our key insight is that response correctness is encoded in the geometry and dynamics of these component-specific trajectories. Based on this observation, we derive lightweight trajectory descriptors for each major LVLM module and integrate them through a calibrated correctness estimator trained with as few as 500 labeled examples. At test time, the proposed estimator requires no additional inference passes or ground-truth labels. Extensive experiments across multiple LVLMs and benchmark datasets demonstrate consistent improvements over existing baselines, including up to 16.44% higher AUROC on Qwen3-VL-2B over VizWiz and 15.61% higher AUPR on LLaVA-Next-7B over ScienceQA compared to the strongest competing methods.

1 Introduction

Large vision-language models (LVLMs) achieve strong performance across multimodal reasoning tasks, yet their reliability remains a major challenge. LVLMs can generate incorrect or hallucinated responses, particularly in tasks requiring

fine-grained visual understanding or dense text interpretation (Zhang et al., 2025; Li et al., 2023c; Guan et al., 2024; Liu et al., 2024a). This has motivated growing interest in self-evaluation methods that enable LVLMs to assess the reliability of their own responses during inference. Existing LVLM self-evaluation methods primarily follow two directions, each with important limitations. The first re-prompts the model to assess the correctness of its own response (Park et al., 2026; Zhao et al., 2024; Wanyin Lin et al., 2025). While intuitive, these approaches suffer from self-preference bias, where the model tends to assign high confidence to its original prediction even when it is incorrect. They also require additional inference passes, increasing computational cost and limiting scalability in real-time settings. The second direction estimates correctness from internal representations, such as intermediate activations or final-layer logits (Li et al., 2023b; Burns et al., 2022; Wang et al., 2025b). Yet these methods are primarily developed for language-only models and do not naturally extend to LVLMs. It remains unclear which internal signals are most informative, how they should be integrated across modalities, and whether language-only confidence indicators remain reliable in visually grounded reasoning. Moreover, existing approaches treat correctness as a property of the model at a *single point* in the processing pipeline, rather than as something that *emerges progressively* throughout the model’s depth. We argue that this view overlooks the most informative signal in LVLMs: the evolution of internal representations along trajectories through the model’s latent space. These limitations raise a fundamental question: *can an LVLM recognize when its own response is incorrect without additional prompting or repeated inference?*

We propose a trajectory-based framework (*Trajectory-Lens*) for estimating generation correctness in LVLMs based on the observation that

¹Project website: speedlab-git.github.io/Trajectory-Lens

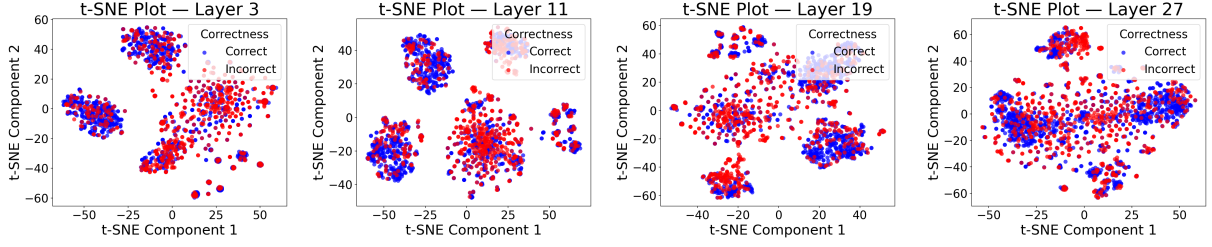


Figure 1: t-SNE visualizations of 1,247 correctly answered samples (blue) and 753 incorrectly answered samples (red) from VQA2. The representations are extracted from different layers of LLaVA-NeXT-7B: Layer 3, Layer 11, Layer 19, and Layer 27, shown from left to right.

transformer residual streams evolve progressively across normalized depth $\tau \in [0, 1]$. Instead of estimating correctness from a single layer, our approach captures discriminative signals across the full trajectory of the model. The language decoder score captures probability-gap evolution across depth, the vision encoder score accumulates visual evidence from discriminative trajectory regions, and the projection score measures geometric separation within low-variance cross-modal subspaces. These component-specific trajectory scores are combined via adaptive weighting to produce a unified correctness estimate over the multimodal residual stream. The entire framework operates within a *single forward pass* without additional prompting or repeated inference.

Contributions: ❶ To the best of our knowledge, this is the first work to study trajectory-level self-evaluation from generated token in LVLMs by jointly analyzing the vision encoder, projection layer, and language decoder within a unified framework. We demonstrate that different components of a LVLM encode distinct trajectory signals of response correctness. ❷ We introduce three trajectory-based correctness scores derived from depth-wise representation evolution of the residual stream, including a causal depth integral for the language model, a conditional depth integral for the vision encoder, and a discriminative subspace integral over cross-modal projection directions. ❸ Extensive experiments across diverse benchmarks and LVLMs show that our framework consistently improves correctness estimation from a single generation pass without additional inference, self-consistency sampling, or ground-truth labels at test time.

2 Related Works

Self-evaluation for LVLMs remains under-explored relative to rapid progress in multimodal benchmarking. Early pipelines relied on human annotations or external expert models to verify image-conditioned

responses (Wang et al., 2025c; Hu et al., 2023; Jing et al., 2023), while recent work shifts toward using LVLMs as judges in scoring and ranking settings (Chen et al., 2024b; Lee et al., 2024; Xiong et al., 2025). Such approaches risk inheriting self-preference biases, where models systematically favor outputs resembling their own generations (Wataoka et al., 2024; Chen et al., 2025). A complementary direction probes internal model states directly, showing that hidden representations and intermediate attention patterns encode reliable hallucination signals without requiring additional generation (Chen et al., 2024a; Wang et al., 2025b; Jiang et al., 2025; Wu et al., 2024). Other methods estimate reliability by aggregating multiple sampled responses via self-consistency or semantic equivalence (Chen et al., 2023; Gupta et al., 2025; Yu et al., 2026; Kuhn et al., 2023), though raw model probabilities often remain poorly calibrated, motivating dedicated calibration strategies for both LLMs and VLMs (Plaut et al., 2025; Xiao et al., 2026). These approaches either require multiple inference passes or auxiliary components, increasing cost and complexity. In contrast, our framework estimates trajectory-level distributional behavior from a single model invocation with only a small amount of data, offering a lightweight and scalable alternative for LVLM self-evaluation.

3 Empirical Findings on Trajectory Dynamics in Vision-Language Models

Recent advances in LLM self-evaluation have shown that internal representations and confidence dynamics contain meaningful signals about generation reliability (Wang et al., 2025b; Chen et al., 2024a; Plaut et al., 2025). However, whether similar trajectory-level self-evaluation signals emerge in LVLMs remains an open question. Unlike standard language models, LVLMs operate through a multimodal pipeline in which visual inputs are encoded, projected into the language embedding space, and progressively transformed before re-

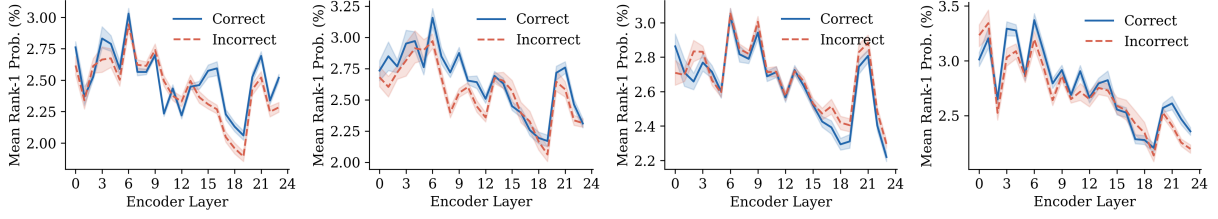


Figure 2: Top token probability trajectories across layers in the language model for ScienceQA, OCRBench, InfoVQA, and MME (left to right). Solid lines represent the average trajectory of the correct response, dashed lines represent the average trajectory of the incorrect response.

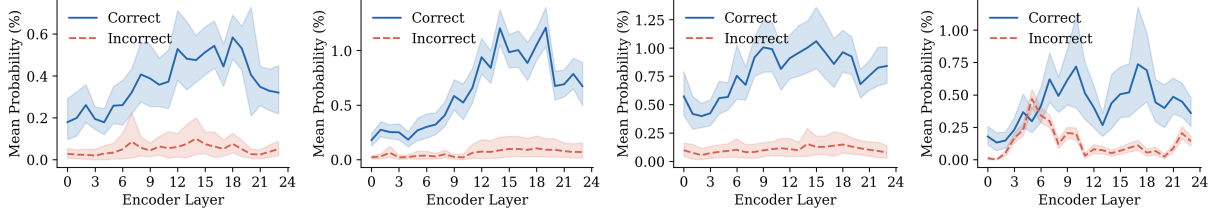


Figure 3: Token-probability trajectories from the vision encoder for InfoVQA, DocVQA, TextVQA, and VizWiz (left to right). Solid and dashed curves denote the mean trajectories of 250 correct samples $\mu_{c,ve}(\tau)$ and 250 incorrect samples $\mu_{i,ve}(\tau)$, respectively.

sponse generation. Existing studies primarily attribute hallucinations and incorrect responses to failures in the language generation stage (Li et al., 2026; Jiang et al., 2025; Wang et al., 2025a). Figure 1, generated from VQAv2 dataset (Goyal et al., 2017), suggests that inconsistencies can emerge much earlier in the multimodal pipeline and propagate through subsequent stages of the model. This observation motivates a finer-grained analysis of LVLM behavior across the vision encoder, projection layer, and language decoder. **We investigate whether incorrect or hallucinated responses can be identified directly from the internal trajectory dynamics of these components during a single forward pass.** To study this question, we analyze how internal representations evolve throughout response generation. Rather than treating hidden states at individual layers as isolated signals, we interpret each component as a continuous trajectory of the residual stream across normalized depth $\tau = (l - 1)/(L - 1) \in [0, 1]$. We then examine which trajectory patterns consistently distinguish correct generations from incorrect ones across different stages of the VLM pipeline. Unless otherwise specified, all analyses in Section 3 are conducted using LLaVA-NeXT-7B on a balanced set of 500 samples, comprising 250 correct and 250 incorrect responses.

3.1 Trajectory Dynamics Reveal Response Correctness in Language Models

We begin by analyzing the language decoder and examine how the top-token probability gap between correct and incorrect generations evolves

as a function of normalized depth $\tau \in [0, 1]$. Fig. 2 reveals a consistent pattern across all benchmarks: at shallow depths ($\tau \approx 0$) the gap $\Delta p(\tau) = \mu_{c,lm}(\tau) - \mu_{i,lm}(\tau)$ is negligible, indicating that the residual stream has not yet committed to a prediction. Here, $\mu_{c,lm}(\tau)$ and $\mu_{i,lm}(\tau)$ are the mean of the correct and incorrect samples from the language model. As τ increases through the middle layers, the trajectories of correct and incorrect generations diverge, which indicates that the model’s internal representation becomes more discriminative at those layers.

3.2 Vision Encoder Trajectories Reveal Progressive Visual Evidence Formation

We next investigate whether similar trajectory dynamics exist within the vision encoder. Unlike language decoders, vision encoders produce contextualized visual representations rather than token logits, making direct confidence estimation non-trivial. To probe these representations, we apply the image projection head to intermediate visual features and measure their similarity to text embeddings, producing a depth-dependent visual evidence signal at each trajectory position τ . Additional implementation details are provided in the Appendix. Fig. 3 reveals two important properties of vision encoder trajectories. First, the trajectories are inherently non-monotonic. Correct-generation evidence does not consistently dominate across all depths. Instead, discriminative trajectory regions emerge intermittently, interleaved with regions where the separation weakens or temporarily reverses. This behavior is particularly evident in the final exam-

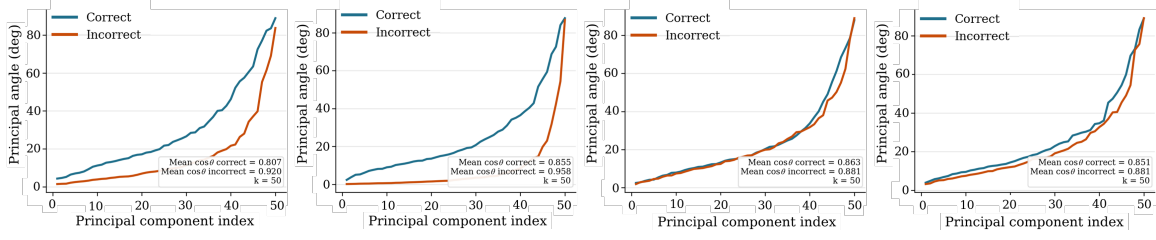


Figure 4: PCA analysis of cosine-angle changes between corresponding principal components before and after the projection layer for ScienceQA, OCRBench, ChartQA, and MME using Qwen3-VL-4B.

ple of Figure 3. Second, for visually challenging tasks involving dense text or fine-grained reasoning, correct-generation trajectories often begin with weak evidence and become strongly discriminative only in deeper layers. This suggests that correctness depends less on the initial strength of visual evidence and more on how visual evidence progressively accumulates throughout the trajectory. These findings rule out the alternative hypothesis that **H1**: higher visual evidence merely reflects easier examples. Instead, they support the hypothesis that **H2**: the progressive accumulation of evidence along the vision encoder trajectory is genuinely predictive of downstream correctness.

3.3 Projection Trajectories Reveal Hidden Geometric Instability

The projection layer forms the interface between visual perception and language generation by mapping visual representations into the language-model embedding space. Prior work (Verma et al., 2024; Tong et al., 2024; Cha et al., 2024) suggests that this transformation introduces measurable information loss. However, we observe that the overall transformation magnitude remains similar for both correct and incorrect generations, making global projection statistics insufficient for correctness estimation. In contrast, a finer-grained trajectory analysis reveals substantially different geometric structures between the two generation types. For Qwen-based models, we compute the cosine angle between token representations before and after projection and apply Principal Component Analysis (PCA) to the resulting angular-change patterns. Fig. 4 shows the mean angular differences across principal directions. Low-index components exhibit similar behavior for both correct and incorrect generations, suggesting that dominant projection directions mainly capture shared global transformations. Higher-index components reveal substantially larger separation between the two trajectories. Although these components explain less overall variance, they preserve finer-grained geometric differences introduced during cross-modal alignment.

These observations suggest that correctness signals are concentrated not in dominant shared projection structure, but in subtle trajectory variations emerging in low-variance projection directions.

4 Framework for Trajectory-Based Correctness Estimation

The empirical analysis in section 3 reveals three consistent patterns: *i*) the language model’s confidence gap between correct and incorrect generations evolves through distinct early, middle, and final phases across depth τ . *ii*) The vision encoder’s discriminative signal is intermittent and non-monotone along τ . *iii*) The projection layer’s geometric change is architecturally heterogeneous, concentrating its discriminative signal in low-variance subspace directions. These findings suggest that generation correctness is encoded in the evolution of multimodal representations across depth rather than in any isolated hidden state. Motivated by this observation, we introduce **Trajectory-Lens**, a trajectory-based correctness estimation framework that tracks and aggregates discriminative signals as they evolve throughout the multimodal pipeline.

4.1 Modeling Residual Stream as a Trajectory

Our central hypothesis is that response correctness is encoded in the evolution of the residual stream across depth, rather than in any isolated hidden state. We therefore interpret the sequence of transformer hidden states as a continuous trajectory through representation space. A transformer with L layers updates its residual stream through successive residual transformations:

$$\mathbf{h}_l = \mathbf{h}_{l-1} + \underbrace{\text{Attn}_l(\mathbf{h}_{l-1}) + \text{MLP}_l(\mathbf{h}_{l-1})}_{F_l(\mathbf{h}_{l-1}; \theta)} \quad (1)$$

We normalize layer depth using $\tau = \frac{l-1}{L-1} \in [0, 1]$ and interpret the residual updates as discrete steps of a continuous dynamical system:

$$\frac{d\mathbf{h}}{d\tau} = F(\mathbf{h}(\tau), \tau; \theta), \quad \mathbf{h}(0) = \mathbf{h}_{\text{in}} \quad (2)$$

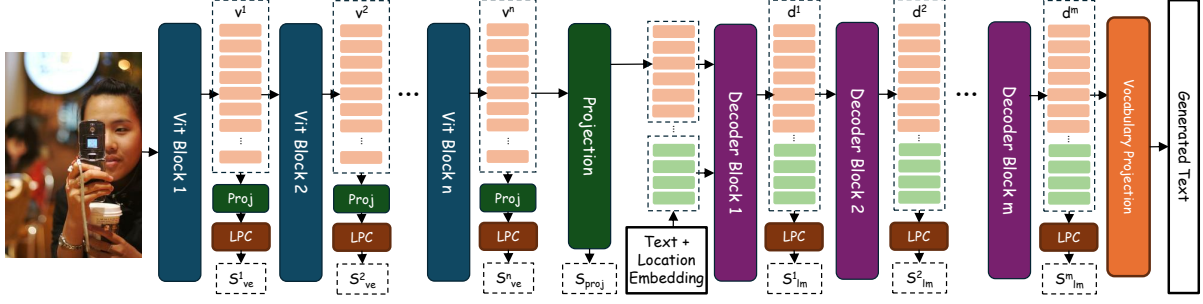


Figure 5: Overview of the multimodal trajectory-scoring pipeline. Projection modules (Proj) and logit probability calculators (LPC) estimate the vision-encoder (S_{ve}), projection-subspace (S_{proj}), and language-model (S_{lm}) trajectory scores from evolving multimodal representations.

This formulation follows the Neural-ODE interpretation of residual networks (Chen et al., 2018), where hidden representations evolve progressively across depth. Under this view, correctness depends on the evolution of representations across depth rather than any isolated hidden state $\mathbf{h}(\tau^*)$. Therefore, any layer-wise observable $f^{(l)}$ can be interpreted as a depth-dependent function $f(\tau)$ defined along the trajectory. This allows correctness-related signals to be accumulated across depth rather than extracted from a single layer.

4.2 Trajectory-Based Component Scores

Language-model trajectory score. The trajectory patterns in Figure 2 suggest that correctness is reflected in the evolution of token probabilities across depth rather than at any individual layer. This is calculated using finally generated answer token over full trajectory rather than the per-layer argmax. Early layers provide weak separation between correct and incorrect responses, while deeper layers become progressively more discriminative. Single-layer estimates may overlook how semantic confidence evolves across depth, while uniform averaging dilutes informative late-stage signals with weak early-layer representations. To capture this behavior, we define a depth-aware trajectory score:

$$S_{lm}[p] = \int_0^1 \tau (x_p(\tau) - w_p(\tau)) d\tau \quad (3)$$

where $x_p(\tau)$ and $w_p(\tau)$ denote the top-token probabilities of the current sample and a calibration-based weighted separator at normalized depth τ , respectively. Here, the weighted separator represents a calibration trajectory computed from labeled correct and incorrect samples, providing a stable reference direction for trajectory-level separation. The weighting kernel τ assigns greater importance to deeper layers, where semantic representations become more stable and discriminative

as the model commits to a response. The resulting score captures confidence evolution across depth rather than isolated layer-wise statistics.

Vision-encoder trajectory score. The trajectories in Fig. 3 exhibit strongly non-monotonic behavior. Correct-generation evidence does not consistently dominate incorrect-generation evidence across all depths. Instead, there exist intervals where the separation weakens, disappears, or reverses. Uniformly integrating evidence across the entire trajectory may dilute informative signals with ambiguous or misleading regions. To address this, we introduce a conditional trajectory integration that accumulates evidence only over discriminative depth intervals. We define a gating function:

$$\chi(\tau) = \mathbf{1}[\mu_{c,ve}(\tau) > \mu_{i,ve}(\tau)]$$

which selects trajectory regions where the vision encoder separates correct and incorrect generations. The resulting vision-encoder trajectory score is:

$$S_{ve}[\mathbf{h}] = \frac{\int_0^1 \chi(\tau) (x_{ve}(\tau) - w_{ve}(\tau)) d\tau}{\int_0^1 \chi(\tau) d\tau} \quad (4)$$

The denominator normalizes the score by the fraction of trajectory depth over which the encoder exhibits discriminative behavior. S_{ve} estimates correctness using only the informative portions of the visual trajectory rather than the full depth uniformly. This conditional structure also explains the distinction between **H1** and **H2** discussed in Section 3.2. Under **H1**, correctness is assumed to depend primarily on the immediate strength of visual evidence after feature extraction. **H2** suggests that correctness emerges progressively as the encoder refines semantically relevant visual information across depth. By selectively integrating only discriminative trajectory regions, $\chi(\tau)$ naturally

emphasizes the progressive evidence accumulation behavior described by *H2*.

Projection-layer subspace score. Our analysis shows that discriminative correctness signals are concentrated in low-variance projection directions that preserve subtle differences between correct and incorrect trajectory evolution. We parameterize the PCA component index using a normalized subspace coordinate $s = \text{idx}/N \in [0, 1]$ and define an indicator function that selects projection directions with meaningful angular separation:

$$\mathcal{I}(s) = \mathbf{1}[|\mu_c(s) - \mu_i(s)| > 1^\circ] \quad (5)$$

The resulting projection-subspace trajectory score:

$$S_{\text{proj}}[\phi] = \frac{\int_0^1 \mathcal{I}(s) (\cos \phi^x(s) - \cos \phi^w(s)) ds}{\int_0^1 \mathcal{I}(s) ds} \quad (6)$$

where $\phi^x(s)$ and $\phi^w(s)$ denote the projection angles of the current sample trajectory and weighted-separator hidden states along the s -th principal direction, respectively. Low-index components primarily capture dominant transformations shared across both correct and incorrect generations, resulting in limited discriminative separation. Higher-index components instead preserve finer-grained residual differences introduced during cross-modal alignment, where the two trajectory classes diverge more strongly. The indicator $\mathcal{I}(s)$ concentrates the score on these informative projection directions while suppressing subspaces with negligible discriminative value.

4.3 Unified Trajectory Correctness Score

The trajectory scores S_{ve} , S_{lm} , and S_{proj} capture complementary signals from visual separability, semantic confidence, and projection-space structure. We combine them into a unified correctness score:

$$\mathcal{S}[\mathbf{h}, p, \phi] = \alpha S_{ve}[\mathbf{h}] + \beta S_{lm}[p] + \gamma S_{\text{proj}}[\phi] \quad (7)$$

where α , β , and γ are component-adaptive calibration coefficients estimated from a small labeled validation set. Higher values of $\mathcal{S}$ indicate that the sample follows multimodal trajectories more consistent with correct generations.

5 Experimental Analysis

5.1 Experimental Setup

We evaluate the proposed framework, *Trajectory-Lens* using **LLaVA-NeXT-7B** and **LLaVA-NeXT-13B** from the LLaVA-NeXT (Liu et al., 2024b)

family, along with **Qwen3-VL-2B** and **Qwen3-VL-4B** from the Qwen3 family (Bai et al., 2025). We also evaluate ShareGPT4V (Chen et al., 2024c), InstructBLIP (Dai et al., 2023), and InternVL (Chen et al., 2024d) to study the generalizability of our framework. To calibrate α , β , and γ , we randomly sample 250 correct and 250 incorrect examples per task, following calibration strategies used in prior steering-based methods (Gadgil et al., 2026; Soo et al., 2025; Tan et al., 2024). The calibrated parameters are then used for full-task evaluation (see Appendix for details).

5.2 Baselines

We compare our method with representative state-of-the-art self-evaluation approaches spanning single-pass and multi-pass settings. Single-pass baselines include **CoE** (Wang et al., 2025b), **INSIDE** (Chen et al., 2024a), Devils in **Middle Layers** (Jiang et al., 2025), log-probability-based confidence estimation **Calib-Corr** (Plaut et al., 2025), and **VL-Calibration** (Xiao et al., 2026), which rely on internal signals such as hidden states, attention weights, or logits. Multi-pass baselines include **Semantic Uncertainty** (Kuhn et al., 2023) and **Logical Closed Loop** (Wu et al., 2024), which estimate reliability from output consistency across repeated inferences. These methods also differ in modality scope: CoE, INSIDE, and Semantic Uncertainty operate only on the language decoder, whereas Logical Closed Loop, Devils in Middle Layers and VL-Calibration explicitly incorporate visual information. For fair comparison we maintain same setting for all baselines.

5.3 Benchmarks and Metrics

We evaluate our framework across five categories: document VQA, chart reasoning, OCR and scene text understanding, science reasoning, and real-world VQA. Specifically, we use **DocVQA** (Mathew et al., 2021), **InfoVQA** (Mathew et al., 2022), **ChartQA** (Masry et al., 2022), **MME** (Fu et al., 2025), **TextVQA** (Singh et al., 2019), **OCRBench** (Liu et al., 2024c), **ScienceQA** (Lu et al., 2022), and **VizWiz** (Bigham et al., 2010). As our primary objective is detecting incorrect model responses, we report negative-class **AUROC**, **AUPR**, and **F1**. AUROC measures separability between positive and negative samples across thresholds, while AUPR and F1 better capture performance under class imbalance.

Table 1: **AUROC**, **AUPR**, and **F1** results (in percentage) across eight different tasks. $\uparrow$ indicates that higher values are better. $\dagger$ indicates that the method requires multiple inference passes to determine correctness. * indicates that the method uses both vision and text to determine correctness. **Bold** denotes the best performance among all methods.

	SQA $\uparrow$	InfoVQA $\uparrow$	ChartQA $\uparrow$	TextVQA $\uparrow$	DocVQA $\uparrow$	OCRBench $\uparrow$	MME $\uparrow$	VizWiz $\uparrow$
LLaVA-NeXT-7B (AUROC/AUPR/F1)								
COE_R	59.35/36.73/50.65	53.02/66.31/77.61	58.24/52.22/62.98	52.12/38.87/52.36	50.26/27.33/43.08	51.09/50.86/66.35	52.59/21.90/36.84	57.41/45.78/57.97
COE_C	59.04/34.99/48.93	54.28/65.84/77.62	61.62/56.28/63.57	52.76/38.53/52.35	57.83/33.17/43.66	63.08/64.21/66.31	55.03/24.08/37.14	56.10/44.32/57.99
INSIDE	70.80/52.63/56.88	59.07/68.69/ 77.96	60.43/54.34/64.07	59.48/53.13/59.06	60.10/46.05/50.13	59.25/67.33/70.77	66.92/34.13/41.51	60.94/49.50/60.44
Logical Closed Loop* $\dagger$	54.10/33.85/46.90	51.26/63.77/77.57	58.74/54.54/62.86	54.25/37.82/52.35	52.35/28.55/42.97	61.56/53.46/67.00	49.66/18.60/35.48	51.51/44.49/58.06
Middle Layers*	33.94/22.70/46.90	50.63/62.55/77.57	56.38/51.34/62.86	49.12/34.41/52.35	57.68/33.77/42.97	54.04/59.21/66.31	48.30/22.64/35.48	34.04/31.40/57.95
Semantic Uncertainty $\dagger$	50.74/32.87/46.90	53.74/68.08/77.57	56.07/56.65/62.86	60.52/59.25/52.76	62.20/40.46/44.36	57.09/60.29/66.31	64.41/34.52/44.06	50.45/42.64/57.96
Calib-Corr	73.73/54.08/82.06	56.54/72.66/56.54	54.43/51.42/70.73	61.53/60.30/78.05	62.67/50.64/83.57	62.34/ 74.91/75.78	66.82/33.91/87.15	60.28/46.53/74.38
VL-Calibration*	53.30/71.13/81.13	53.21/37.62/52.87	54.19/61.23/70.27	54.38/67.19/78.35	50.75/72.44/82.51	53.13/59.45/73.48	53.51/83.03/82.11	42.03/56.18/61.54
Trajectory-Lens (Ours)*	74.80/86.74/82.68	61.14/72.81/55.22	62.35/62.22/71.19	62.80/73.66/78.62	61.68/80.01/84.23	65.25/62.76/72.05	67.47/88.45/87.91	61.25/67.15/74.45
LLaVA-NeXT-13B (AUROC/AUPR/F1)								
COE_R	53.19/29.58/42.92	51.56/62.04/75.07	56.49/47.08/58.60	51.86/35.25/50.07	56.90/27.47/39.30	65.18/60.59/66.12	52.02/29.97/35.19	53.27/41.77/55.67
COE_C	65.66/41.69/47.02	55.94/64.43/75.03	57.97/49.36/58.99	55.06/38.39/50.03	61.41/32.53/40.65	70.38/71.43/66.66	54.03/31.99/37.28	58.45/44.77/55.98
INSIDE	68.17/44.74/53.36	57.57/68.54/ 77.88	58.52/52.52/65.11	58.64/45.93/54.71	66.04/40.55/47.04	63.23/63.38/69.16	57.48/27.16/37.07	57.10/47.73/58.12
Logical Closed Loop* $\dagger$	54.09/28.45/43.79	53.71/63.50/75.03	60.51/50.91/58.55	56.99/39.66/50.00	58.48/30.64/37.94	65.73/64.21/63.66	58.87/23.76/35.93	52.39/42.71/55.60
Middle Layers*	33.45/19.62/42.48	50.42/59.07/75.03	55.38/45.63/58.55	49.64/32.42/50.00	54.96/27.44/37.81	58.58/62.00/63.66	47.80/20.57/33.34	37.30/30.84/55.60
Semantic Uncertainty $\dagger$	50.19/27.33/42.57	50.62/62.88/75.03	54.78/45.00/58.59	56.04/44.02/50.00	55.15/28.42/37.81	54.39/57.27/63.66	53.03/21.91/33.34	51.03/41.19/55.83
Calib-Corr	70.31/50.92/84.52	58.21/70.42/59.95	64.43/54.39/74.14	54.18/60.87/82.49	66.12/48.24/ 87.89	70.85/72.68/76.56	55.99/37.93/86.92	57.94/42.56/76.16
VL-Calibration*	55.02/72.92/81.86	53.41/42.09/55.32	54.19/61.23/72.27	53.81/69.62/79.15	56.19/81.28/85.94	62.71/61.69/72.42	55.19/81.66/81.97	38.39/56.23/59.58
Trajectory-Lens (Ours)*	74.15/87.80/85.27	58.86/72.66/57.95	66.37/72.32/74.30	60.48/75.30/80.02	70.91/82.83/86.81	71.25/61.28/69.53	61.31/86.75/88.89	59.00/69.65/76.21
Qwen3-VL-2B (AUROC/AUPR/F1)								
COE_R	54.93/15.05/25.91	63.34/34.29/46.37	57.12/23.66/42.89	58.23/26.05/34.89	59.29/7.22/14.62	50.17/18.34/34.19	63.84/31.75/39.79	51.86/34.41/47.81
COE_C	51.92/13.71/25.91	56.09/30.30/39.86	51.43/24.24/34.38	52.07/19.80/34.72	53.37/6.25/12.44	50.37/18.73/33.02	61.85/30.37/38.71	52.19/33.66/47.81
INSIDE	53.67/16.01/24.24	62.85/35.39/47.68	60.22/34.02/47.75	64.51/29.71/38.70	63.73/9.11/15.51	52.08/22.33/34.30	59.22/27.80/36.70	53.78/32.83/48.33
Logical Closed Loop* $\dagger$	50.38/15.74/23.84	59.49/34.79/42.88	60.44/33.31/37.85	55.69/22.70/33.47	63.20/9.23/17.37	59.39/37.47/48.02	57.20/20.45/34.44	44.01/28.95/47.81
Middle Layers*	48.92/12.23/23.84	57.22/29.86/39.66	60.74/37.54/33.88	43.18/17.16/33.47	54.33/7.11/10.95	52.49/23.44/32.35	44.22/17.81/34.44	33.50/23.41/47.81
Semantic Uncertainty $\dagger$	50.44/13.04/23.84	53.61/29.94/39.66	54.28/31.94/33.88	54.26/27.59/33.47	51.44/21.88/32.35	50.45/5.94/10.95	52.35/23.23/34.44	50.08/31.22/47.84
Calib-Corr	61.38/36.03/72.74	53.75/27.62/81.93	58.35/79.29/78.66	58.37/30.97/88.82	62.17/8.65/ 97.01	53.00/23.85/89.36	63.92/39.37/78.50	50.80/29.57/71.37
VL-Calibration*	58.98/73.01/73.04	49.84/67.65/81.77	49.62/82.78/78.49	54.39/73.91/ 89.40	52.68/65.73/96.95	59.63/84.39/89.52	52.62/83.17/86.73	39.83/65.59/72.06
Trajectory-Lens (Ours)*	66.11/79.11/82.04	65.41/60.11/83.90	64.87/84.88/81.57	64.52/78.00/78.51	66.02/77.11/84.14	60.05/57.57/68.38	61.27/84.85/87.91	70.22/77.48/75.14
Qwen3-VL-4B (AUROC/AUPR/F1)								
COE_R	62.68/10.68/18.80	62.41/24.50/34.96	52.05/17.29/27.43	61.14/25.34/35.00	56.63/4.49/8.89	56.69/18.59/28.05	68.90/24.00/31.46	52.27/35.14/47.48
COE_C	62.49/11.38/19.61	63.62/28.07/36.99	53.73/19.28/27.44	64.54/27.71/37.99	59.12/5.77/10.87	58.54/20.13/29.29	67.39/22.67/30.47	51.78/33.94/48.17
INSIDE	51.66/9.50/14.07	53.80/21.39/30.12	51.52/17.11/27.54	58.68/24.99/32.86	51.55/4.61/8.34	51.54/16.98/27.18	61.96/18.62/27.47	53.95/32.86/46.82
Logical Closed Loop* $\dagger$	58.17/8.19/15.82	58.36/20.80/33.02	62.96/22.00/35.27	60.77/25.93/35.95	60.41/5.33/10.86	60.15/30.69/39.69	52.93/13.50/23.87	42.08/25.59/46.49
Middle Layers*	49.08/6.20/12.63	56.25/22.42/29.88	68.95/30.36/27.40	43.33/16.29/31.53	52.57/5.50/7.48	59.07/17.61/25.17	51.73/13.69/22.76	37.30/30.84/55.60
Semantic Uncertainty $\dagger$	50.87/8.32/12.87	50.03/17.73/29.88	50.79/11.39/27.75	54.93/26.93/31.53	51.67/4.51/7.48	52.06/18.83/25.17	56.19/19.05/23.72	50.39/32.57/46.49
Calib-Corr	63.62/21.40/96.37	56.72/21.94/ 90.37	63.20/19.80/91.37	63.50/25.83/88.67	60.38/5.48/ 98.01	58.13/19.78/92.14	67.48/24.01/91.26	54.32/31.81/72.15
VL-Calibration*	50.56/93.51/95.58	50.51/44.66/89.73	44.50/84.24/91.24	59.20/86.05/88.78	53.21/76.89/97.91	59.95/89.12/92.14	60.19/90.56/92.48	42.27/57.21/71.79
Trajectory-Lens (Ours)*	66.47/96.66/96.51	64.14/48.63/55.30	71.78/91.48/91.53	62.45/86.60/89.66	66.45/82.83/84.32	64.34/89.69/92.23	75.26/95.10/92.86	55.93/64.34/74.38

Table 2: Pipeline-stage timings on **OCRBench** for **LLaVA-NeXT-7B**. VE: Vision Encoder, RT: Rank Trajectory, PS: Projection Subspace, ENS: Ensemble.

Stage	Time (s)	Forward Passes	Throughput (samples/s)	Latency (ms/sample)	Description
VE	608.98	250	0.41	2435.92	α calib.
RT	146.38	250	1.70	585.52	β calib.
PS	56.23	250	4.44	224.92	γ calib.
ENS	4.60	0	217.39	4.60	Computes $\mathcal{S}$

6 Results and Discussion

6.1 Quantitative Analysis

Table 1 reports AUROC, AUPR, and F1 across eight benchmarks. For LLaVA-NeXT models-7B, our method achieves the best AUROC on **7 out of 8** benchmarks, with gains especially pronounced on SQA, where our AUPR exceeds the next best method by over **15%**. On VizWiz, our AUPR margin over the strongest competitor exceeds **10%**, showing robustness to noisy real-world inputs. On document and OCR-heavy tasks (DocVQA, InfoVQA, TextVQA), we consistently lead in AUROC. The sole exception is OCRBench, where Calib-Corr surpasses our AUPR on 7B (**74.91** vs. **62.76**), indicating sensitivity to model capacity on OCR-heavy tasks. Our framework generalizes effectively to the Qwen family as well. On Qwen3-VL-2B, we achieve the best AUROC on

7 out of 8 benchmarks, with notable AUPR gains on VizWiz (**+11.89%** over VL-Calibration) and DocVQA (**+11.38%** over VL-Calibration). Among baselines, Calib-Corr is the strongest competitor yet underperforms our framework in AUROC by an average of **2.8%** across tasks and model scales. Methods requiring multiple inference passes or additional visual inputs underperform our framework by over **5%** AUROC on average, demonstrating the effectiveness of our single-pass approach across diverse LVLMS and benchmarks.

6.2 Computational Efficiency

Table 2 reports the computational efficiency of our pipeline on OCRBench. We use 500 samples to calibrate the hyperparameters α , β , and γ . For LLaVA-NeXT-7B, calibrating α from the vision encoder takes 608.98 s, calibrating β from the language model takes 146.38 s, and calibrating γ takes 56.23 s. The total calibration cost is therefore 811.59 s, or approximately **13.5 minutes**. After calibration, the pipeline is highly efficient at inference, requiring only **4.60 ms** per sample.

6.3 Ablation Study

6.3.1 Generalization Across Architectures

Table 3 demonstrates that our framework generalizes consistently across diverse LVLMS and benchmarks.

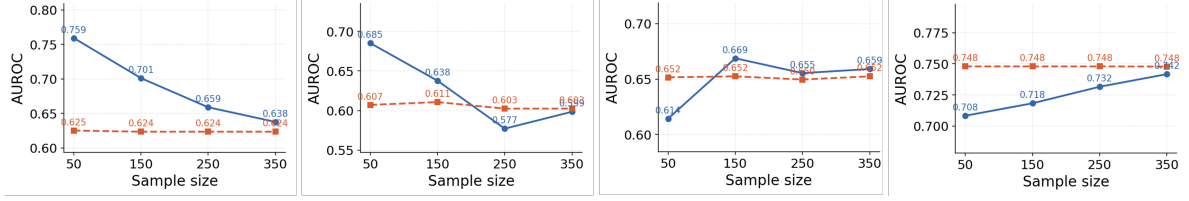


Figure 6: Performance vs. calibration sample size across ChartQA, InfoVQA, OCRBench, and ScienceQA.

Table 3: Performance evaluation of five LVLMS integrated with our Trajectory-Lens, measured by AUROC, AUPR, and F1 (in percentage) on **eight** benchmark datasets.

	SQA ↑	InfoVQA ↑	ChartQA ↑	TextVQA ↑	DocVQA ↑	OCRBench ↑	MME ↑	VizWiz ↑
(AUROC/AUPR/F1)								
InstructBLIP-XL	69.6/83.1/81.6	58.7/29.8/33.8	82.6/33.0/42.6	83.1/67.2/66.0	92.0/92.8/87.0	58.9/28.9/44.2	66.9/83.1/82.2	50.8/62.0/74.4
InstructBLIP-XXL	76.1/90.0/85.2	64.0/67.0/67.1	58.6/53.5/60.1	82.0/67.8/63.9	58.6/53.5/50.1	86.0/70.9/65.7	59.1/78.6/84.0	74.0/73.9/76.2
InternVL2-4B	52.9/96.4/97.9	55.0/68.5/78.7	60.2/86.0/89.6	62.8/82.8/85.9	61.8/93.2/94.6	67.3/85.9/86.0	57.0/88.5/91.4	52.9/63.7/76.5
InternVL2-8b	64.8/98.0/98.5	57.7/75.2/83.0	65.2/88.8/90.3	57.2/82.4/87.7	56.1/93.5/95.8	64.3/81.6/88.1	65.9/91.2/92.5	56.3/66.4/76.0
ShareGPT4V	53.0/76.4/84.4	54.9/44.1/57.4	55.8/63.1/73.9	55.8/71.0/80.0	58.9/81.7/86.8	60.6/59.2/73.9	54.8/54.8/88.9	50.8/60.0/74.5

chitectures. InternVL2-8B and InternVL2-4B achieve particularly strong AUPR and F1 scores on document-oriented tasks, while InstructBLIP-XXL obtains competitive AUROC on benchmarks such as OCRBench (85.99) and ScienceQA (76.08). InstructBLIP-XL also performs strongly on DocVQA (see [Appendix](#) for details).

6.3.2 Task-Specific Calibration Sample

To calibrate the task-specific weights, we use a small calibration subset following the intuition of steering-vector estimation. Figure 6 analyzes calibration sizes of **50, 150, 250, and 350** examples from each type. Performance improves with increasing calibration size, approaching the fully cal-

ibrated optimum. Using the elbow point of the performance-latency trade-off curve, we select **250** examples as the operating point, achieving near-optimal performance with low calibration overhead. Although the calibration weights (α, β, γ) are estimated independently for each task, they exhibit consistent structural patterns across datasets and architectures ([Appendix: Table 5](#)). Tasks with dense visual grounding, such as OCRBench and DocVQA, assign larger weights to visual trajectory components, whereas reasoning-oriented tasks such as ScienceQA and MME rely more heavily on the language-model trajectory score. This consistency suggests that the learned coefficients capture stable

Table 4: Task transferability analysis of Trajectory-Lens, where calibration is performed on one task (columns) and evaluated on other tasks (rows) using the same model.

	TextVQA	DocVQA	OCRBench	InfoVQA	ChartQA	MME	SQA	VizWiz
LLaVA-NeXT-7B (AUROC/AUPR/F1)								
TextVQA	62.80/73.66/78.62	60.88/79.60/84.19	61.02/59.09/69.05	60.98/71.95/75.15	61.08/61.58/71.05	61.88/86.09/87.91	63.88/76.01/82.60	56.66/64.25/74.42
DocVQA	62.13/73.06/78.59	61.68/80.01/84.23	64.62/61.91/71.69	60.43/45.42/55.25	62.24/62.12/71.03	66.76/88.20/87.91	72.91/84.53/82.65	60.53/66.68/74.44
OCRBench	61.69/72.68/78.55	61.63/79.96/84.17	65.25/62.76/72.05	60.25/45.27/55.03	62.35/62.22/71.19	65.25/62.76/72.05	74.78/86.75/82.65	61.25/67.15/74.45
InfoVQA	62.78/73.58/78.62	60.14/45.18/55.04	61.12/59.17/69.05	61.14/72.81/55.22	61.13/61.67/71.08	61.12/59.17/69.05	63.78/75.91/82.64	56.87/64.49/74.42
ChartQA	61.69/72.68/78.55	61.63/79.96/84.17	65.25/62.76/72.05	60.25/45.27/55.03	62.35/62.22/71.19	65.25/62.76/72.05	74.78/86.75/82.65	61.25/67.15/74.45
MME	61.72/72.76/78.53	61.63/79.99/84.18	65.16/62.68/72.05	67.47/88.45/87.91	62.25/62.13/71.14	67.47/88.45/87.91	74.79/86.73/82.71	60.90/66.87/74.45
SQA	61.71/72.75/78.52	61.63/79.98/84.17	65.19/62.70/72.05	60.11/45.16/55.06	62.29/62.16/71.14	65.19/62.70/72.05	74.80/86.74/82.68	60.90/66.87/74.45
VizWiz	61.69/72.68/78.55	61.63/79.96/84.17	65.25/62.76/72.05	60.25/45.27/55.03	62.35/62.22/71.19	65.25/62.76/72.05	74.78/86.75/82.65	61.25/67.15/74.45
LLaVA-NeXT-13B (AUROC/AUPR/F1)								
TextVQA	60.48/75.30/80.02	60.88/82.90/86.83	67.90/60.26/69.53	57.89/57.89/57.60	66.30/72.25/74.30	54.83/82.74/89.00	65.64/80.58/84.53	48.88/61.36/76.15
DocVQA	60.46/75.14/80.02	70.91/82.83/86.81	67.65/59.87/69.53	57.61/46.31/57.57	66.37/72.32/74.30	53.76/82.00/88.98	63.33/79.14/84.45	48.88/61.31/76.15
OCRBench	58.97/74.88/80.08	58.97/82.16/86.89	71.25/61.28/69.53	58.72/46.68/58.07	65.33/71.27/74.24	59.25/85.75/88.94	72.27/85.92/84.93	42.27/56.48/76.26
InfoVQA	57.37/74.05/80.02	57.26/81.17/86.85	68.06/61.35/69.65	58.86/72.66/57.95	64.62/70.31/74.16	61.06/86.67/88.92	73.96/87.52/85.12	51.48/62.35/76.24
ChartQA	60.46/75.14/80.02	70.91/82.83/86.81	67.65/59.87/69.53	57.61/46.31/57.57	66.37/72.32/74.30	53.76/82.00/88.98	63.33/79.14/84.45	48.88/61.31/76.15
MME	56.90/73.70/80.02	57.08/81.00/86.88	68.19/61.74/69.66	61.31/86.75/88.89	64.47/70.25/74.23	61.31/86.75/88.89	74.15/87.80/85.27	43.97/57.04/76.26
SQA	56.90/73.70/80.02	57.08/81.00/86.88	68.19/61.74/69.66	61.31/86.75/88.89	64.47/70.25/74.23	61.31/86.75/88.89	74.15/87.80/85.27	43.97/57.04/76.26
VizWiz	57.37/74.05/80.02	57.26/81.17/86.85	68.06/61.35/69.65	58.86/72.66/57.95	64.62/70.31/74.16	61.06/86.67/88.92	73.96/87.52/85.12	51.48/62.35/76.24
Qwen3-VL-2B (AUROC/AUPR/F1)								
TextVQA	64.52/78.00/78.51	51.31/73.14/84.14	60.01/57.30/68.48	53.89/39.51/53.88	57.54/59.93/71.05	61.04/84.67/87.91	66.09/79.08/82.09	70.13/77.71/75.14
DocVQA	54.12/67.86/78.50	66.02/77.11/84.14	59.77/56.20/68.43	54.64/39.78/54.21	64.87/84.88/81.57	61.27/84.85/87.91	66.11/79.11/82.04	69.97/78.55/75.10
OCRBench	51.62/64.66/78.50	49.93/71.89/84.14	60.05/57.57/68.38	51.53/37.72/53.76	61.21/62.49/71.24	60.27/83.66/87.91	65.83/78.87/82.06	70.16/77.06/75.13
InfoVQA	52.77/65.49/78.50	53.43/75.49/84.14	59.91/56.56/68.40	65.41/60.11/83.90	61.21/62.49/71.24	61.18/84.79/87.91	66.10/79.10/82.07	70.10/78.44/75.09
ChartQA	54.12/67.86/78.50	66.02/77.11/84.14	59.77/56.20/68.43	54.64/39.78/54.21	64.87/84.88/81.57	61.27/84.85/87.91	66.11/79.11/82.04	69.97/78.55/75.10
MME	54.12/67.86/78.50	66.02/77.11/84.14	59.77/56.20/68.43	54.64/39.78/54.21	64.87/84.88/81.57	61.27/84.85/87.91	66.11/79.11/82.04	69.97/78.55/75.10
SQA	54.12/67.86/78.50	66.02/77.11/84.14	59.77/56.20/68.43	54.64/39.78/54.21	64.87/84.88/81.57	61.27/84.85/87.91	66.11/79.11/82.04	69.97/78.55/75.10
VizWiz	54.12/67.86/78.50	66.02/77.11/84.14	59.77/56.20/68.43	54.64/39.78/54.21	64.87/84.88/81.57	61.27/84.85/87.91	66.11/79.11/82.04	69.97/78.55/75.10
Qwen3-VL-4B (AUROC/AUPR/F1)								
TextVQA	62.45/86.60/89.66	63.62/80.62/84.29	59.76/89.06/92.53	53.40/37.72/54.10	70.00/90.68/91.39	72.71/93.94/92.86	60.97/95.65/96.51	52.99/62.52/74.41
DocVQA	62.31/86.67/89.66	66.45/82.83/84.32	59.23/89.17/92.49	64.14/48.63/55.50	71.20/90.81/91.45	75.23/95.07/92.86	66.47/96.66/96.51	54.70/63.76/74.38
OCRBench	62.22/86.76/89.67	66.13/82.63/84.31	64.34/89.69/92.23	64.08/48.77/55.39	71.40/90.95/91.48	75.26/95.10/92.86	66.45/96.66/96.51	55.02/63.92/74.38
InfoVQA	62.31/86.67/89.66	66.45/82.83/84.32	59.23/89.17/92.49	64.14/48.63/55.50	71.20/90.81/91.45	75.23/95.07/92.86	66.47/96.66/96.51	54.70/63.76/74.38
ChartQA	61.13/86.43/89.70	66.13/82.63/84.31	59.83/89.34/92.55	62.53/47.13/55.70	71.78/91.48/91.53	75.12/95.05/92.86	66.42/96.66/96.51	55.47/64.18/74.38
MME	62.22/86.76/89.67	66.13/82.63/84.31	64.34/89.69/92.23	64.08/48.77/55.39	71.40/90.95/91.48	75.26/95.10/92.86	66.45/96.66/96.51	55.02/63.92/74.38
SQA	62.31/86.67/89.66	66.45/82.83/84.32	59.23/89.17/92.49	64.14/48.63/55.50	71.20/90.81/91.45	75.23/95.07/92.86	66.47/96.66/96.51	54.70/63.76/74.38
VizWiz	62.06/86.42/89.67	59.97/78.59/84.27	59.95/88.90/92.53	48.79/34.71/53.65	66.12/89.17/91.37	66.20/91.98/92.86	54.87/93.45/96.53	55.93/64.34/74.38

task-dependent multimodal reliability characteristics rather than dataset-specific memorization.

6.3.3 Contribution of Individual Components

Figure 7 analyzes the contributions of the vision encoder, language model, and projection layer. On datasets such as DocVQA and ChartQA, the language model alone achieves performance comparable to the full system. For more visually grounded tasks such as InfoVQA and TextVQA, the combined framework consistently outperforms individual components. These findings indicate that while the language model provides the dominant signal, the vision encoder and projection layer contribute complementary multimodal information.

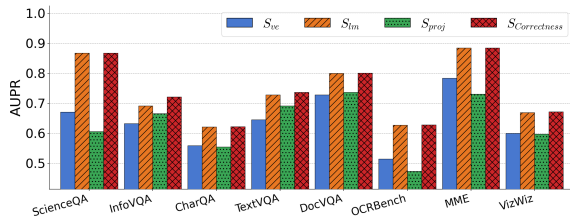


Figure 7: AUPR comparison of individual trajectory components and their combined score across eight benchmark datasets.

6.4 Cross-Task/Unknown distribution

Table 4 evaluates the transferability of the proposed estimator by fitting the calibrator on one source task and directly applying it to the remaining target tasks without re-estimation or task-specific adaptation. Across the four evaluated LVLMS, the average AUROC under cross-task calibration is 62.10, compared with 64.89 under matched-task calibration, resulting in an average decrease of only 2.79 AUROC. This relatively small gap indicates that the trajectory-based calibrator captures general correctness-related patterns that transfer across heterogeneous tasks rather than overfitting to a particular benchmark. Transferability is especially strong for LLaVA-NeXT-7B and Qwen3-VL-4B, with average AUROC decreases of only 1.44 and 2.10, respectively, and several transferred calibrators perform comparably to their task-specific counterparts. Transferability varies across model-task combinations, with larger drops for Qwen3-VL-2B on TextVQA and InfoVQA, suggesting that smaller models may be more sensitive to task-specific answer distributions and trajectory patterns.

7 Conclusions

We introduced a trajectory-based framework for estimating response correctness in LVLMS from a

single forward pass. Beyond the proposed estimator, this work presents the first systematic analysis of trajectory dynamics across the vision encoder, projection layer, and language decoder in LVLMS. Our findings show that correctness emerges progressively through the evolution of multimodal representations across depth rather than from isolated hidden states or final-layer logits. Motivated by this observation, we developed complementary trajectory-based scores for the vision encoder, projection layer, and language decoder, and integrated them into a unified correctness estimator without requiring repeated inference or auxiliary prompting. Extensive experiments across multiple LVLMS and benchmark datasets demonstrate consistent improvements over existing self-evaluation methods, including gains of up to 16.44% AUROC and 15.61% AUPR, while maintaining high efficiency with only 4.6 ms latency per sample after response. These results establish trajectory dynamics as an effective foundation for correctness estimation in LVLMS and suggest that reliable multimodal self-evaluation may require modeling how internal representations evolve across depth rather than relying solely on static confidence signals from individual layers or final outputs. A limitation of our framework is that it is primarily designed for short-form responses with localized semantic targets. Extending trajectory-based correctness estimation to abstractive generation, refusal behaviors, and long-form multimodal reasoning remains an important direction for future work.

8 Limitations

Our trajectory-based framework is designed and evaluated in the setting of short-answer visual question answering, where both correct and incorrect responses can be represented by a single token or a small set of semantically equivalent tokens. We outline several directions for future work below.

Abstractive and Descriptive Answers. The current formulation assumes that the target response can be mapped to a discrete answer token or token group. This assumption does not hold for open-ended outputs such as captions, descriptions, or free-form explanations.

Complex Spatial Reasoning. Our LogSumExp-based aggregation treats visual patches independently when estimating evidence for an answer. As a result, the method may be limited in cases where correctness depends on relational or compositional

reasoning across multiple image regions, e.g., comparing object positions, counting distributed entities, or interpreting interactions between objects.

Refused and Hedged Responses. The current framework assumes that the model produces a committed answer at generation time. Consequently, it does not assign a score when the model abstains, refuses, or provides a hedged response.

Long-form Reasoning Chains. Our analysis focuses on the trajectory of the final answer token. For responses that include intermediate reasoning before the final answer, the generated reasoning tokens may introduce trajectory dynamics that differ from those of the answer itself.

9 Ethical considerations

The authors used AI-assisted tool (OpenAI’s ChatGPT), for grammar refinement, formatting assistance, and readability improvement during the preparation of this manuscript. All research ideas, methodological developments, experiments, analyses, and conclusions were conceived, conducted, and verified solely by the authors.

Our work aims to improve the reliability by identifying potentially incorrect responses before they are trusted in downstream applications. However, correctness estimates are not guarantees of factual accuracy and may degrade under distribution shifts or unseen inputs. Therefore, they should be used as auxiliary reliability signals rather than as a replacement for human verification, especially in high-stakes domains such as healthcare, legal decision-making, and safety-critical systems.

Acknowledgement

This material is partly based upon work supported by the U.S. National Science Foundation (NSF) under Grant No. CRII-IIS-RI-2553868 and NAIRR Pilot under Allocation No. NAIRR250393. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF and NAIRR.

References

Marah Abidin, 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*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samuel White, and 1 others. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23rd annual ACM symposium on User interface software and technology*, pages 333–342.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, and 1 others. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. 2024. Honeybee: Locality-enhanced projector for multimodal llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13817–13827.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024a. Inside: LLMs’ internal states retain the power of hallucination detection. *arXiv preprint arXiv:2402.03744*.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024b. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Proceedings of the International Conference on Machine Learning*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024c. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. 2018. Neural ordinary differential equations. In *NeurIPS*.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. 2023. Universal self-consistency for large language model generation. *arXiv preprint arXiv:2311.17311*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024d. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

- Zhi-Yuan Chen, Hao Wang, Xinyu Zhang, Enrui Hu, and Yankai Lin. 2025. Beyond the surface: Measuring self-preference in llm judgments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1653–1672.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tjong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2025. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38.
- Soham Gadgil, Chris Lin, and Su-In Lee. 2026. Where to steer: Input-dependent layer selection for steering improves llm alignment. *arXiv preprint arXiv:2604.03867*.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14375–14385.
- Mihir Gupta, Abhay Mangla, Pratik Desai, and Ross Greer. 2025. Self-consistency in vision-language models for precision agriculture: multi-response consensus for crop disease management. In *2025 IEEE International Conference on Image Processing Workshops (ICIPW)*, pages 185–190. IEEE.
- Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417.
- Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. 2025. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25004–25014.
- Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2023. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. *arXiv preprint arXiv:2311.01477*.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.
- Seongyun Lee, Seungone Kim, Sue Hyun Park, Geewook Kim, and Minjoon Seo. 2024. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. *arXiv preprint arXiv:2401.06591*.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.
- Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. In *Proceedings of 2023 conference on empirical methods in natural language processing*, pages 292–305.
- Yifan Li, Kun Zhou, Xin Zhao, Lei Fang, and Jirong Wen. 2026. Analyzing and mitigating object hallucination: A training bias perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 6636–6643.
- Hanchao Liu, Wenyan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23*, Red Hook, NY, USA. Curran Associates Inc.
- Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024c. Ocr-bench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102.

- Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.
- Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.
- Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.
- Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.
- Seongheon Park, Changdae Oh, Hyeong Kyu Choi, Xuefeng Du, and Sharon Li. 2026. Vauq: Vision-aware uncertainty quantification for lvlm self-evaluation. *arXiv preprint arXiv:2602.21054*.
- Benjamin Plaut, Khanh Xuan Nguyen, and Tu Trinh. 2025. Probabilities of chat llms are miscalibrated but still predict correctness on multiple-choice q&a. *Transactions on Machine Learning Research*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763.
- Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.
- Samuel Soo, Wesley Teng, Chandrasekaran Balaganesh, Tan Guoxian, and Ming YAN. 2025. Interpretable steering of large language models with feature guided activation additions. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.
- Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. 2024. Analysing the generalisation and reliability of steering vectors. *Advances in Neural Information Processing Systems*, 37:139179–139212.
- Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9568–9578.
- Gaurav Verma, Minje Choi, Kartik Sharma, Jamelle Watson-Daniels, Sejoon Oh, and Srijan Kumar. 2024. Cross-modal projection in multimodal llms doesn’t really project visual attributes to textual space. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 657–664.
- Chao Wang, Xuancheng Zhou, Weiwei Fu, and Yang Zhou. 2025a. Mitigating hallucinations in large vision-language models with internal fact-based contrastive decoding. *arXiv preprint arXiv:2502.01056*.
- Yiming Wang, Pei Zhang, Baosong Yang, Derek Wong, and Rui Wang. 2025b. Latent space chain-of-embedding enables output-free llm self-evaluation. In *International Conference on Learning Representations*, volume 2025, pages 70938–70970.
- Ziyue Wang, Chi Chen, Fuwen Luo, Yurui Dong, Yuanchi Zhang, Yuzhuang Xu, Xiaolong Wang, Peng Li, and Yang Liu. 2025c. Actiview: Evaluating active perception ability for multimodal large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7605–7633.
- Inna Wanyin Lin, Yushi Hu, Shuyue Stella Li, Scott Geng, Pang Wei Koh, Luke Zettlemoyer, Tim Althoff, and Marjan Ghazvininejad. 2025. Self-improving vlm judges without human annotations. *arXiv e-prints*, pages arXiv–2512.
- Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*.
- Junfei Wu, Qiang Liu, Ding Wang, Jinghao Zhang, Shu Wu, Liang Wang, and Tieniu Tan. 2024. **Logical closed loop: Uncovering object hallucinations in large vision-language models**. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6944–6962, Bangkok, Thailand. Association for Computational Linguistics.
- Wenyi Xiao, Xinchu Xu, and Leilei Gan. 2026. Vl-calibration: Decoupled confidence calibration for large vision-language models reasoning. *arXiv preprint arXiv:2604.09529*.
- Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. 2025. Llava-critic: Learning to evaluate multimodal models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Chung-En Johnny Yu, Brian Jalaian, and Nathaniel D. Bastian. 2026. Scoop: Semantic consistent opinion pooling for uncertainty quantification in multiple vision-language model systems. *arXiv preprint arXiv:2603.23853*.
- Can Zhang, Ziheng Wu, Zhenghao Chen, Yufei Zhan, Yifan Li, Zhao Zhang, Xian Wang, Minghui Qiu, and 1 others. 2025. Seeing is believing? mitigating ocr hallucinations in multimodal large language models.

Advances in Neural Information Processing Systems, 38:74230–74248.

Henry Hengyuan Zhao, Pan Zhou, Difei Gao, Zechen Bai, and Mike Zheng Shou. 2024. Lova3: Learning to visual question answering, asking and assessment. *Advances in Neural Information Processing Systems*, 37:115146–115175.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Appendix

A Logit Lens to Diverse Answer Formats

In this section, we analyze different types of challenging question-answer pairs and present how each case is resolved. Figure 8 illustrates representative examples of these question types.

Yes/No Questions. For yes/no questions, we first trace back to the original question and identify the key entity or attribute being queried. To automate this step, we use Qwen3-VL-2B to extract the relevant keyword or phrase from the question. The logit lens of that word is then used in evaluating the confidence in the answer.

Multiple-choice Questions. For multiple-choice questions, we evaluate the semantic content of the correct answer option rather than only the option label. If the option consists of a single word, we directly compute its logit-lens score. If an option is a longer phrase or sentence, we use the Qwen3-VL-2B to extract the keyword or phrase representing the option, and then compute the logit-lens score for that extracted target.

Counted or Calculated Numerical Answers. For numerical answers, we apply the logit lens in a teacher-forced manner, similar to the multi-word answer setting. Since numerical values may be tokenized into multiple subword units and may have several valid textual forms, we score each candidate number by summing token-level log-probabilities. We further aggregate equivalent surface forms using log-sum-exp. Finally, we track the layer-wise rank or probability of the correct numerical answer within a fixed candidate set.

Multi-word Answers. For answers consisting of multiple tokens, we apply the logit lens in a teacher-forced manner. At each layer, we compute the probability of each target token conditioned on the prompt and the preceding ground-truth answer tokens. We then aggregate token-level log-probabilities across the answer sequence, using the average log-probability as the layer-wise confidence score for the complete answer.

B Methodology

B.1 Weighted Separator

We compute a variance-aware separator between correct and incorrect trajectories using the mean

Algorithm 1 Vision Encoder Logit Lens

Require: Image $\mathbf{I}$; vision encoder layers $\{V_{el}\}_{l=1}^L$; projection head $Proj$; LM embedding matrix $\mathbf{E} \in \mathbb{R}^{|\mathcal{V}| \times d}$; target token sequence $\mathbf{t} = (t_1, \dots, t_K)$; trajectory fractions $\tau \in (0, 1]$

Ensure: Logit-lens scores $x_{ve}(\tau)$

- 1: Encode image through vision encoder layers:
 $\mathbf{h}_l \leftarrow V_{el}(\mathbf{h}_{l-1})$
 - 2: Map trajectory fraction to layer index:
 $l_\tau \leftarrow \lfloor \tau \cdot L \rfloor$
 - 3: Project trajectory features to LM embedding space:
 $\mathbf{u}_\tau \leftarrow Proj(\mathbf{h}_{l_\tau}), \quad \mathbf{u}_\tau \in \mathbb{R}^{N \times d}$
 - 4: Normalize projected patch features:
 $\hat{\mathbf{u}}_\tau \leftarrow \text{NormalizeRows}(\mathbf{u}_\tau)$
 - 5: Vocabulary logits at trajectory point τ :
 $\mathbf{Z}_\tau \leftarrow \hat{\mathbf{u}}_\tau \hat{\mathbf{E}}^\top, \quad \mathbf{Z}_\tau \in \mathbb{R}^{N \times |\mathcal{V}|}$
 - 6: LSE over patches per token, sum over token sequence:

$$x_{ve}(\tau) \leftarrow \sum_{k=1}^K \log \sum_{n=1}^N \exp(\mathbf{Z}_\tau[n, t_k])$$
 - 7: Return trajectory logit-lens scores
return $x_{ve}(\tau)$
-

and standard deviation of their trajectory scores. For normalized layer τ layer and information source z , the separator is defined as

$$w_z(\tau) = \frac{(\sigma_{i,z}(\tau))^2 \mu_{c,z}(\tau) + (\sigma_{c,z}(\tau))^2 \mu_{i,z}(\tau)}{(\sigma_{c,z}(\tau))^2 + (\sigma_{i,z}(\tau))^2 + \varepsilon}, \quad (8)$$

where $\mu_{c,z}(\tau)$ and $\sigma_{c,z}(\tau)$ are the mean and standard deviation for correct trajectories, and $\mu_{i,z}(\tau)$ and $\sigma_{i,z}(\tau)$ are the corresponding statistics for incorrect trajectories. The constant ε ensures numerical stability. This inverse-variance weighting places the separator closer to the trajectory distribution with lower uncertainty. We consider three sources: language-model logits ($z = \text{lm}$), logits obtained after projecting internal representations from vision encoder into the language space ($z = \text{ve}$), and the angular deviation $\phi(s)$ induced by the projection ($z = \text{proj}$).

B.2 Logit-lens Calculation from Vision Encoder

Algorithm 1 outlines the logit-lens computation along the vision encoder trajectory. At each trajectory fraction $\tau \in (0, 1]$, the intermediate vision encoder representation $\mathbf{h}_{l_\tau}$ is projected into the lan-

Yes/No Questions:



Question: Is this photo taken in a place of baseball field? Please answer yes or no.
Answer: Yes
Keyword: Baseball field



Question: Is this a photo of Cotehele? Please answer yes or no.
Answer: Yes
Keyword: Cotehele

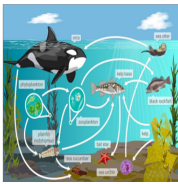


Question: Is this photo taken in a place of recreation room? Please answer yes or no
Answer: No
Keyword: Recreation room

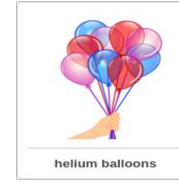


Question: Is this an image of Tanemakidaishi? Please answer yes or no.
Answer: No
Keyword: Tanemakidaishi

Multiple-choice Questions:



Question: Which of these organisms contains matter that was once part of the phytoplankton?
Choices: ["black rockfish", "sea otter"]
Answer: 0
Keyword: black rockfish



Question: Which property matches this object?
Choices: ["flexible", "sticky"]
Answer: 0
Keyword: flexible



Question: Select the organism in the same species as the North American beaver.
Choices: ["Ovis orientalis", "Alouatta caraya", "Castor canadensis"]
Answer: 2
Keyword: Castor canadensis



Question: Which is this organism's scientific name?
Choices: ["goldfish", "Carassius auratus"]
Answer: 1
Keyword: Carassius auratus

Counted or Calculated Numerical Answers:



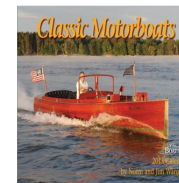
Question: how much is the sprayer?
Answers: 39.99



Question: what is the time shown?
Answer: 10:09



Question: What number is on the plane?
Answers: 41



Question: Which year's calendar is this?
Answer: 2013

Multi-word Answers.



Question: When was the privacy & security policy last updated?
Answers: August 8, 2016



Question: what is the name of the funeral home?
Answer: McMullen funeral home



Question: who is the author of dancers of arun?
Answers: Elizabeth a. Lynn



Question: Select the organism in the same species as the great blue heron?
Answer: Ardea Herodias

Figure 8: Examples of different challenging question-answer pairs used in logit lens analysis

guage model embedding space via the projection head $Proj$. Vocabulary logits $\mathbf{Z}_\tau \in \mathbb{R}^{N \times |\mathcal{V}|}$ are computed by projecting onto the embedding matrix $\mathbf{E}$. For a single target token t_i , the logit-lens score directly reflects how strongly the vision encoder encodes the target concept at depth τ . For a multi-token target sequence $\mathbf{t} = (t_1, \dots, t_K)$, the score is computed by applying Log-Sum-Exp (LSE) for each target token t_k , and summing the resulting per-token scores across the sequence. This formulation enables the trajectory score $x_{ve}(\tau)$ to capture the collective evidence across all image patches for the entire target token sequence at each depth of the vision encoder.

C Experimental Analysis

C.1 Detailed Experimental Setup

We evaluate the proposed frameworks on five vision-language model families: **LLaVA-NeXT-13B** (Liu et al., 2024b), **Qwen3-VL-4B** (Bai et al., 2025), **ShareGPT4V** (Chen et al., 2024c), **InstructBLIP** (Dai et al., 2023) and **InternVL** (Chen et al., 2024d). For InstructBLIP, we use two variants InstructBLIP Flan-T5-XL and InstructBLIP Flan-T5-XXL, both of which share a frozen ViT-g/14 image encoder (Radford et al., 2021) and an instruction-tuned Q-Former (Li et al., 2023a) while differing in their language decoder, using Flan-T5-XL (3B) and Flan-T5-XXL (11B) (Chung et al., 2024) respectively. For InternVL2, we use InternVL2-4B and InternVL2-8B, both built on the ViT-MLP-LLM architecture (Chen et al., 2024d) with a shared InternViT-300M-448px visual encoder and an MLP projector, but paired with different language decoders, namely Phi-3-mini-128k-instruct (Abdin et al., 2024) for the 4B variant and InternLM2.5-7B-Chat (Cai et al., 2024) for the 8B variant. We also evaluate ShareGPT4V-7B, which follows the LLaVA-style architecture (Liu et al., 2023) with a CLIP-Large visual encoder (Radford et al., 2021) and a Vicuna-v1.5 (Zheng et al., 2023) language decoder. We also considered **ShareGPT4V** (Chen et al., 2024c), 2 variant of **InstructBLIP** (Dai et al., 2023) and **InternVL** (Chen et al., 2024d)

D More Result Analysis

D.1 Analysis on Diverse Models

Table 3 reports the AUROC, AUPR, and F1 scores of our technique applied to five LVLMS across

eight benchmarks. Overall, our approach demonstrates consistent applicability across diverse model architectures and task types. Among the evaluated models, InternVL2-8b and InternVL2-4B tend to yield strong AUPR and F1 scores, particularly on document-oriented tasks such as DocVQA and OCRBench, suggesting that higher-capacity models with richer visual representations provide a more reliable basis for uncertainty estimation. InstructBLIP-XXL achieves competitive AUROC on several benchmarks, including OCRBench (85.99) and SQA (76.08), while InstructBLIP-XL performs well on DocVQA (91.99 AUROC). In contrast, ShareGPT4V consistently underperforms relative to the other models across most benchmarks, potentially reflecting limitations in its visual grounding capabilities. Performance also varies substantially across tasks: DocVQA and SQA generally yield higher scores, whereas InfoVQA and OCRBench remain more challenging, highlighting the difficulty of reliable uncertainty quantification in tasks requiring fine-grained visual-textual understanding and complex information extraction.

D.1.1 Contribution of Individual Components from AUROC

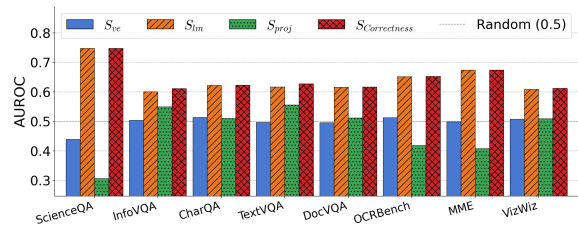


Figure 9: AUROC comparison of individual trajectory components and their combined score across eight benchmark datasets.

The AUROC trends in Figure 9 further show that S_{lm} is the strongest individual component across most benchmarks, often closely matching $S_{correctness}$. In contrast, S_{ve} demonstrate that the vision encoder alone has limited discriminative ability for correctness estimation. The projection layer S_{proj} shows more task-dependent behavior. It provides useful signal on OCR- and document-oriented datasets such as InfoVQA, TextVQA, DocVQA, and VizWiz, but performs poorly on ScienceQA, OCRBench, and MME. Overall, the results further suggest that correctness prediction is mainly driven by the language model, while the visual and projection components provide additional task-specific cues rather than serving as strong standalone predictors.

Table 5: Optimal weighting coefficients ($\alpha/\beta/\gamma$) calculated from the vision encoder, language model and projector, respectively, for each model across eight vision-language benchmarks.

	SQA $\uparrow$	InfoVQA $\uparrow$	ChartQA $\uparrow$	TextVQA $\uparrow$	DocVQA $\uparrow$	OCRBench $\uparrow$	MME $\uparrow$	VizWiz $\uparrow$
	$(\alpha/\beta/\gamma)$							
LLaVA-NeXT-7B	0.62/0.37/0.00	0.71/0.14/0.14	0.83/0.16/0.00	0.00/0.50/0.50	0.14/0.71/0.14	0.83/0.17/0.00	0.20/0.80/0.00	0.83/0.17/0.00
LLaVA-NeXT-13B	0.83/0.16/0.00	0.00/1.00/0.00	0.00/0.33/0.66	0.00/0.37/0.62	0.00/0.33/0.67	0.57/0.28/0.14	0.83/0.17/0.00	0.00/1.00/0.00
Qwen3-VL-2B	0.00/1.00/0.00	0.25/0.75/0.00	0.00/1.00/0.00	0.00/0.44/0.55	0.00/1.00/0.00	0.80/0.20/0.00	0.00/1.00/0.00	0.00/1.00/0.00
Qwen3-VL-4B	0.00/1.00/0.00	0.00/1.00/0.00	0.55/0.44/0.00	0.00/0.57/0.42	0.00/1.00/0.00	0.20/0.00/0.80	0.20/0.80/0.00	0.00/0.33/0.67
Instructblip-XL	0.00/0.80/0.20	0.83/0.17/0.00	0.00/1.00/0.00	0.16/0.83/0.00	0.00/1.00/0.00	0.00/0.83/0.17	0.50/0.17/0.33	0.83/0.17/0.00
Instructblip-XXL	0.00/1.00/0.00	0.00/1.00/0.00	0.00/1.00/0.00	0.00/1.00/0.00	0.83/0.17/0.00	0.83/0.17/0.00	0.56/0.33/0.11	0.00/0.75/0.25
InternV12-4B	0.00/0.75/0.25	0.55/0.11/0.33	0.00/1.00/0.00	0.50/0.30/0.20	0.00/0.50/0.50	0.00/0.17/0.83	0.80/0.20/0.00	0.00/1.00/0.00
InternV12-8b	0.00/0.17/0.83	0.28/0.25/0.57	0.37/0.62/0.00	0.00/0.52/0.57	0.40/0.60/0.00	0.71/0.14/0.14	0.00/1.00/0.00	0.00/0.33/0.67
ShareGPT4V	0.00/1.00/0.00	0.71/0.14/0.14	0.45/0.09/0.45	0.00/0.20/0.80	0.83/0.00/0.17	0.83/0.17/0.00	0.00/1.00/0.00	0.55/0.11/0.33

Table 6: Cross-model performance analysis on same model family

Calibrator	Model	TextVQA	DocVQA	OCRBench	InfoVQA	ChartQA	MME	SQA	VizWiz
LLaVA-NeXT-7B	LLaVA-NeXT-13B	60.12/75.39/80.05	58.13/74.48/80.11	58.19/61.74/69.66	58.37/46.40/57.83	64.47/70.25/74.23	61.07/86.67/88.92	74.08/87.65/85.13	43.97/57.04/76.26
LLaVA-NeXT-13B	LLaVA-NeXT-7B	61.69/72.68/78.55	59.09/78.49/84.16	63.54/60.89/70.34	60.04/45.11/55.05	59.08/60.39/70.83	67.46/88.45/87.91	74.78/86.75/82.65	60.85/66.85/74.45
Qwen3-VL-2B	Qwen3-VL-4B	62.39/86.59/89.66	66.45/82.83/84.32	60.69/89.73/92.59	64.04/48.76/55.56	71.20/90.81/91.45	75.23/95.07/92.86	66.47/96.66/96.51	54.70/63.76/74.38
Qwen3-VL-4B	Qwen3-VL-2B	54.47/68.02/78.51	56.02/77.11/84.14	52.42/49.77/67.06	54.64/39.78/54.21	55.74/58.66/71.02	61.20/84.81/87.91	66.11/79.11/82.04	57.30/66.37/74.52

Table 7: Standard deviation (AUROC/AUPR/F1) for different model and task.

Model	TextVQA	InfoVQA	DocVQA	OCRBench	ChartQA	MME	SQA	VizWiz
LLaVA-NeXT-7B	0.49/0.41/0.04	2.50/17.52/12.70	0.56/12.28/10.30	1.89/1.63/1.36	0.56/0.26/0.07	2.20/13.30/8.58	4.97/4.88/0.03	1.98/1.22/0.01
LLaVA-NeXT-13B	1.66/0.69/0.02	1.51/17.51/14.40	6.08/0.88/0.03	4.70/0.80/0.07	0.89/1.00/0.06	3.54/2.25/0.04	5.05/4.05/0.36	3.58/2.57/0.05
Qwen3-VL-2B	3.98/4.07/0.00	4.14/7.34/10.54	7.54/2.11/0.00	0.12/0.56/0.03	2.77/12.06/5.38	0.35/0.41/0.00	0.10/0.08/0.02	0.08/0.56/0.02
Qwen3-VL-4B	0.42/0.13/0.01	6.06/5.74/0.77	2.30/1.55/0.02	2.24/0.29/0.13	1.86/0.67/0.05	3.19/1.11/0.00	4.27/1.14/0.01	0.86/0.55/0.01

D.2 Calibrated Variable Range:

Table 5 reports the optimal weighting triplets (α, β, γ) for the composite scoring function in Eq. 7. The results reveal three distinct patterns of component dominance across models and tasks. First, several models, including Qwen3-VL-2B, LLaVA-NeXT-13B, and InstructBLIP-XXL, assign full weight to the language-model score $S_{lm}[p]$ on a majority of benchmarks, suggesting that token-level confidence serves as the most reliable uncertainty signal in general-purpose visual question answering. Second, models such as LLaVA-NeXT-7B and ShareGPT4V exhibit higher α values on OCR-intensive tasks (e.g., OCRBench, DocVQA), indicating that visual encoder hidden-state uncertainty $S_{ve}[\mathbf{h}]$ becomes increasingly informative when the task is dominated by text recognition. Third, γ -dominant configurations, though less frequent, emerge in tasks requiring fine-grained spatial or compositional reasoning (e.g., VizWiz, TextVQA), where the vision-language projection score $S_{proj}[\phi]$ captures uncertainty not reflected in either $S_{ve}[\mathbf{h}]$ or $S_{lm}[p]$ alone. Taken together, these findings suggest that no single component universally dominates, and that task-adaptive weighting is essential for reliable uncertainty estimation in multimodal models.

D.3 Cross-Task/Unknown distribution

Table 6 further evaluates whether a calibrator estimated from one model can be directly transferred to another model within the same family, without re-estimation or model-specific adaptation. Overall, the results demonstrate meaningful cross-model transferability across different model scales. For the LLaVA-NeXT family, transferring the calibrator from 7B to 13B achieves an average AUROC of 59.80, while the 13B-to-7B transfer achieves 63.32. Similarly, for the Qwen3-VL family, the calibrator estimated from Qwen3-VL-2B transfers particularly well to Qwen3-VL-4B, achieving an average AUROC of 65.15 across the eight tasks. The reverse Qwen3-VL-4B-to-2B transfer is comparatively weaker, with an average AUROC of 57.24, indicating some sensitivity to model capacity and scale. Nevertheless, the overall results suggest that the learned trajectory-based correctness patterns are not strictly model-specific and can generalize across different-sized models within the same architecture family, although the degree of transferability can vary with the source and target model.

D.4 Standard deviation

Table 7 reports the standard deviation of AUROC, AUPR, and F1 score for every model-task combination over the five independently sampled calibration sets. The results indicate that performance is

generally stable across calibration draws. Particularly for F1 score, whose median standard deviation across all model–task combinations is **0.04**. AUROC also shows moderate variability, with a median standard deviation of **2.22**. However, we observe larger variability for several AUPR results, specially on InfoVQA, DocVQA, ChartQA, and MME, reflecting the greater sensitivity of AUPR to the calibration composition and class-score distribution.