OTU: PROCESS-AWARE UNCERTAINTY ESTIMATION FOR MULTIMODAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) have demonstrated strong capabilities in visual understanding and multi-step reasoning, yet their predictions often lack reliable uncertainty estimates. Existing uncertainty quantification (UQ) methods predominantly operate at the answer level, treating the final response as a whole rather than modeling its internal reasoning process. As a result, they often fail to capture uncertainties caused by ambiguous evidence, imprecise grounding, or errors in intermediate reasoning steps. This limitation is particularly critical in multimodal settings, where uncertainty may stem from both perception and reasoning. In this work, we propose **OTU**, a process-aware uncertainty estimation framework that models reasoning trajectories as chain graphs and quantifies uncertainty via cross-trajectory consistency. In a fully black-box setting, OTU employs optimal transport to align multiple sampled trajectories and compute a consensus barycenter that captures dominant semantic and structural patterns. Uncertainty is then derived from the deviation of individual trajectories from this consensus, enabling both answer-level confidence estimation and fine-grained step-level reliability analysis. Extensive experiments on multiple multimodal benchmarks show that OTU consistently improves calibration over existing black-box baselines. Moreover, it provides interpretable insights into where and how reasoning fails, highlighting that uncertainty in MLLMs is inherently a process-level phenomenon rather than solely an outcome-level signal.

1 INTRODUCTION

A key aspect of human intelligence lies in the ability to assess uncertainty throughout the reasoning process (Cosmides & Tooby, 1996), rather than only evaluating final outcomes. Recent Multimodal Large Language Models (MLLMs) (Wang et al., 2025; Bai et al., 2025; Team et al., 2023) have made rapid progress in understanding complex visual scenes and performing multi-step reasoning grounded in visual evidence (Su et al., 2025; Huang et al., 2025). However, as these models become more capable, a critical question remains: *when and where do they fail*?

Current MLLMs often produce answers that appear confident yet are incorrect (Groot & Valdenegro-Toro, 2024). In language models, most existing uncertainty estimation methods operate at the level of final answers, relying on signals such as token likelihoods (Malinin & Gales, 2020; Fu et al., 2025), verbalized confidence elicited through prompting strategies (Lin et al., 2022; Tian et al., 2023), or agreement across multiple sampled outputs with various aggregation schemes (Xiong et al., 2023; Lin et al., 2023). While these approaches provide coarse estimates of answer correctness for relatively simple problems, they struggle to capture how uncertainty evolves throughout the reasoning process. This limitation becomes especially pronounced in complex reasoning tasks (Jiang et al., 2025; Lu et al., 2023), where errors often arise in intermediate steps, and models may exhibit hesitation, revision, or inconsistent reasoning trajectories due to underlying uncertainty (Hao et al., 2024). Even when the final answer is correct, these unstable intermediate reasoning processes suggest that the model's confidence should be calibrated more carefully. Therefore, capturing this process-level uncertainty is crucial for revealing the source of errors and improving the reliability of model predictions.

Multimodal reasoning adds another source of uncertainty because perception and inference are tightly coupled (Chen et al., 2025a; Zhou et al., 2025). Errors may arise from visual misinterpretation (Chen

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

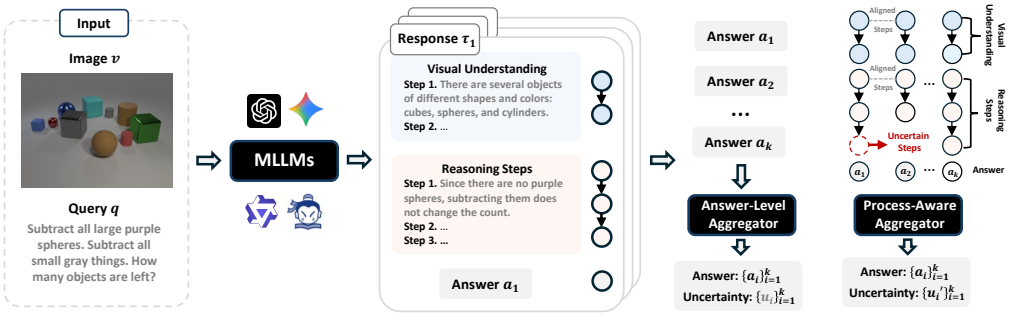

Figure 1: Answer-level aggregation summarizes sampled MLLM trajectories by their final predictions, whereas OTU compares visual-understanding and reasoning steps to estimate process-aware confidence and localize inconsistency.

et al., 2025b; Li et al., 2023), flawed intermediate reasoning (Yan et al., 2025), or their interaction. Locating the source matters because visual and reasoning failures require different mitigation strategies (Zou et al., 2024).

To address these challenges, we propose **OTU**, an optimal transport-based framework for uncertainty quantification in MLLMs that models internal reasoning trajectories in a fully black-box setting (Team et al., 2023; Achiam et al., 2023), requiring only input-output interactions and no labeled reasoning supervision. As illustrated in Figure 1, instead of treating each response as a whole, we represent it as a structured chain that captures both visual grounding and intermediate reasoning steps. We then quantify uncertainty by measuring consistency across multiple sampled trajectories, using optimal transport (Cuturi, 2013; Titouan et al., 2019) to jointly align their semantic content and structural dependencies, thereby enabling soft matching across trajectories with varying lengths and structures, without requiring explicit one-to-one alignment between reasoning steps. This formulation enables us to assess not only whether a prediction is unreliable, but also where and how the reasoning process deviates from a consensus.

In summary, our contributions are as follows:

- We identify a key limitation of existing black-box UQ methods for multimodal large language models: uncertainty is typically estimated at the final answer level, without modeling how it emerges across reasoning trajectories and intermediate reasoning steps.
- We propose OTU, a process-aware uncertainty estimation framework that models multimodal reasoning as structured trajectories. Our method represents each reasoning path as a chain graph and quantifies uncertainty by measuring cross-trajectory consistency via optimal transport, capturing both semantic and structural variations across modalities.
- We conduct extensive experiments on multiple multimodal benchmarks, showing that our approach achieves the strongest aggregate performance across calibration and failure-detection metrics, while providing interpretable insights into reasoning uncertainties.

2 Preliminaries

This section establishes the black-box uncertainty quantification setting for MLLMs, including the sampled trajectory set, candidate answers, and the process-aware uncertainty assigned to each answer (Section 2.1). It then introduces optimal transport and the FGW barycenter as the mathematical tools for aligning structured reasoning trajectories and extracting their shared consensus (Section 2.2).

2.1 Uncertainty Quantification in Black-Box MLLMs

Uncertainty in MLLMs (Groot & Valdenegro-Toro, 2024) arises from visual ambiguities, prompt variations, and the stochastic nature of autoregressive generation. Given an input image $\mathbf{v}$ and a text query $\mathbf{q}$, an MLLM generates a reasoning trajectory τ, consisting of intermediate steps and a final answer a. In black-box settings, only input-output interactions are accessible; model parameters,

internal states, and token-level probabilities are unavailable. Uncertainty is therefore typically estimated by sampling multiple reasoning trajectories under stochastic decoding. Concretely, we obtain a set of K trajectories $\mathcal{T} = \{\tau_k\}_{k=1}^{K}$, where K denotes the number of sampled outputs. Standard approaches, such as Self-Consistency (Xiong et al., 2023; Lyu et al., 2025), quantify uncertainty by measuring the agreement among the final answers $\{a_k\}_{k=1}^{K}$, e.g., through frequency or semantic consistency (Lin et al., 2023). However, these methods focus solely on the final outcome and ignore how uncertainty develops and propagates throughout the reasoning process.

To incorporate process-level variability without conflating trajectory-specific diagnostics with the final prediction, we estimate the uncertainty of each candidate answer from the full set of sampled trajectories. The problem can be formalized as follows:

Problem 1 (Black-Box Uncertainty Quantification for MLLMs). *Given a black-box MLLM and an input pair* $(\mathbf{v}, \mathbf{q})$*, stochastic decoding produces* K *reasoning trajectories* $\mathcal{T} = \{\tau_k\}_{k=1}^{K}$ *with candidate answers* $\mathcal{A} = \{a_k : k = 1, \ldots, K\}$*. The goal is to define an uncertainty function* $\mathcal{U}$ *that assigns a process-aware uncertainty score to each candidate answer using the sampled trajectory set,*

$$u(a \mid \mathcal{T}) = \mathcal{U}(a \mid \mathcal{T}), \quad a \in \mathcal{A}, \qquad u_k := u(a_k \mid \mathcal{T}). \tag{1}$$

Here, each trajectory inherits the uncertainty of the answer it supports, and trajectories sharing the same answer have the same u_k. Their reasoning processes determine this answer-level uncertainty through trajectory deviations and provide separate step-level diagnostic signals.

2.2 Optimal Transport and FGW Barycenter

In our framework, each reasoning trajectory τ is represented as a structured chain graph $G = (\mathbf{X}, \mathbf{C}, \mathbf{p})$, capturing both visual grounding and reasoning steps, as illustrated in Figure 1. Such trajectories often involve complex structures, including sequential dependencies as well as parallel reasoning patterns. To compare them, we require a distance that can align their steps while respecting both semantic content and structural roles. Optimal Transport (OT) (Peyré & Cuturi, 2019; Villani et al., 2009; Cuturi, 2013) provides a principled way to align elements between two sets by learning a transport plan. In our setting, OT can be viewed as matching reasoning steps across trajectories with minimal semantic cost.

To capture both semantic and structural alignment, we adopt the Fused Gromov-Wasserstein (FGW) distance (Titouan et al., 2019), which jointly aligns node features and structural dependencies. Formally, let a trajectory be represented as a graph $G = (\mathbf{X}, \mathbf{C}, \mathbf{p})$, where $\mathbf{X} \in \mathbb{R}^{n \times d}$ denotes node features (e.g., embeddings of steps), $\mathbf{C}^{n \times n}$ encodes pairwise structural relations within trajectories, and $\mathbf{p}$ is a pre-defined distribution over nodes. Given two trajectories $G_1 = (\mathbf{X}_1, \mathbf{C}_1, \mathbf{p})$ and $G_2 = (\mathbf{X}_2, \mathbf{C}_2, \mathbf{q})$, the FGW distance is defined as:

$$d_{FGW,\alpha}(G_1, G_2) = \min_{\mathbf{T} \in \Pi(\mathbf{p}, \mathbf{q})} (1 - \alpha)\langle \mathbf{M}, \mathbf{T} \rangle + \alpha \sum_{i,j,k,l} L\left(C_{1,i,k}, C_{2,j,l}\right) T_{i,j} T_{k,l}, \tag{2}$$

where $\Pi(\mathbf{p}, \mathbf{q})$ denotes the set of nonnegative transport plans whose row and column marginals are $\mathbf{p}$ and $\mathbf{q}$, respectively. Here, $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2}$ is the transport plan, where $T_{i,j}$ indicates how much mass from step i in G_1 is matched to step j in G_2, and $\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}$ is the inter-graph feature cost matrix, with $M_{i,j}$ measuring semantic dissimilarity between steps across trajectories (e.g., embedding distance). The inner product $\langle \mathbf{M}, \mathbf{T} \rangle = \sum_{i,j} M_{i,j} T_{i,j}$ represents the linear sum of semantic alignment costs, weighting pairwise dissimilarities by how strongly steps are matched. The function L measures discrepancies between structural relations, typically comparing pairwise relations of step pairs across trajectories. The parameter $\alpha \in [0, 1]$ balances semantic alignment and structural consistency, trading off feature similarity and relational structure. Intuitively, FGW aligns two reasoning trajectories by matching steps that are both semantically similar and occupy analogous positions in the reasoning structure.

Beyond pairwise comparison, we summarize a set of trajectories into a shared reference via the FGW barycenter (Brogat-Motte et al., 2022). Intuitively, the barycenter can be viewed as a consensus trajectory that captures the dominant semantic and structural patterns shared across multiple reasoning trajectories. In our framework, it serves as a reference against which individual trajectories are compared to assess their consistency and uncertainty.

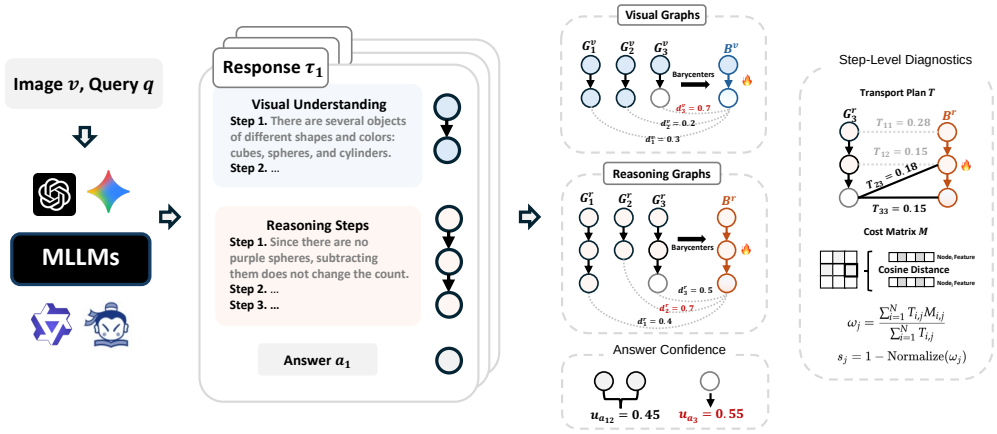

Figure 2: OTU decomposes sampled trajectories into visual and reasoning graphs, learns one FGW barycenter per modality, and uses graph-to-consensus deviations for answer confidence and transport-weighted costs for step-level diagnostics.

Definition 1 (FGW Barycenter). *Given a set of trajectory graphs* $\{G_k = (\mathbf{X}_k, \mathbf{C}_k, \mathbf{p}_k)\}_{k=1}^K$ *and their corresponding weights* $\{w_k\}_{k=1}^K$ *satisfying* $\sum_{k=1}^K w_k = 1$*, the FGW barycenter* $G^* = (\mathbf{X}^*, \mathbf{C}^*, \mathbf{p}^*)$ *is defined as the minimizer of the weighted sum of FGW distances:*

$$(\mathbf{X}^*, \mathbf{C}^*) = \arg\min_{\mathbf{X},\mathbf{C}} \sum_{k=1}^{K} w_k \cdot d_{FGW,\alpha}\big((\mathbf{X}, \mathbf{C}, \mathbf{p}^*), G_k\big). \tag{3}$$

Here, $\mathbf{X}^*$ and $\mathbf{C}^*$ represent the optimized node features and structural matrix of the barycenter, respectively, while $\mathbf{p}^*$ is a fixed reference node distribution. The hyperparameter $\alpha \in [0, 1]$ controls the trade-off between the feature information and the graph structure.

3 Methodology

This section presents OTU in four stages, as summarized in Figure 2. We first represent visual understanding and reasoning as separate chain graphs (Section 3.1) and learn a modality-specific FGW barycenter for each graph family (Section 3.2). We then aggregate graph-to-consensus deviations into process-aware answer confidence (Section 3.3) and use transport alignment to derive relative step-level diagnostics (Section 3.4).

3.1 Visual and Reasoning Chain Graphs

Following the notation in Section 2, we represent trajectory τ_k with two chain graphs: a visual graph G_k^v and a reasoning graph G_k^r. Generated trajectories already impose a natural step order, so a chain captures their observed structure without additional inference. In contrast, a non-chain topology would require an external parser or auxiliary model to infer dependencies between steps, introducing extra assumptions and potential parsing errors into the black-box pipeline. Using $m \in \{v, r\}$ to index the two modalities, both graphs take the form

$$G_k^m = (\mathbf{X}_k^m, \mathbf{C}_k^m, \mathbf{p}_k^m), \tag{4}$$

where $\mathbf{X}_k^m$, $\mathbf{C}_k^m$, and $\mathbf{p}_k^m$ are the feature matrix, structure matrix, and mass vector for modality m.

Feature matrix. $\mathbf{X}_k^m \in \mathbb{R}^{n_k^m \times d}$ encodes the n_k^m steps as d-dimensional text representations that capture their semantic content.

Structure matrix. $\mathbf{C}_k^m \in \mathbb{R}^{n_k^m \times n_k^m}$ models relationships between steps:

$$\mathbf{C}_k^m = (1 - \beta)\mathbf{C}_{\text{pos},k}^m + \beta \mathbf{C}_{\text{sem},k}^m. \tag{5}$$

Here $\mathbf{C}^m_{\text{pos},k}[i,j] \propto |i-j|$ captures the relative positions of steps along the reasoning chain, while $\mathbf{C}^m_{\text{sem},k}$ contains pairwise cosine distances between their embeddings. The parameter $\beta \in [0,1]$ balances positional and semantic structure.

Mass vector. We use the uniform distribution $\mathbf{p}^m_k = \mathbf{1}/n^m_k$ as a neutral prior. Since black-box access provides neither token probabilities nor ground-truth step importance, learned masses require additional assumptions. Uniform mass sets only the prior node weights; transport and alignment costs still determine each step's agreement with the consensus.

This formulation treats each trajectory as a structured process, aligning steps by both semantic similarity and their roles in the reasoning sequence. The semantic component further distinguishes structurally similar but semantically inconsistent steps. Although we use chains as the default topology, FGW can accommodate richer structures such as directed acyclic graphs by changing $\mathbf{C}^m_k$ without altering the remaining pipeline. We evaluate this extension in Section 5.4.

3.2 Consensus Learning via FGW Barycenters

Assessing trajectory reliability without ground-truth reasoning requires comparing graphs whose steps differ in meaning, position, or granularity. FGW supplies soft semantic–structural correspondences across these graphs, while its barycenter summarizes each modality's recurring pattern (Section 2.2).

For each $m \in \{v, r\}$, we fix a reference mass $\mathbf{p}^*_m$ and compute $\mathcal{B}^m = (\mathbf{X}^*_m, \mathbf{C}^*_m, \mathbf{p}^*_m)$ by solving:

$$(\mathbf{X}^*_m, \mathbf{C}^*_m) = \arg\min_{\mathbf{X},\mathbf{C}} \frac{1}{K} \sum_{k=1}^{K} d_{\text{FGW},\alpha}\big((\mathbf{X}, \mathbf{C}, \mathbf{p}^*_m), G^m_k\big). \tag{6}$$

Here, $\mathbf{X}^*_m$ and $\mathbf{C}^*_m$ are the consensus feature and structure matrices, $\mathbf{p}^*_m$ is its fixed mass, and $1/K$ weights all graphs equally. Each term measures the best alignment with G^m_k, so minimizing their average yields the shared pattern. The learned correspondences expose fine-grained deviations at aligned roles, while distances quantify overall semantic–structural differences. Separate modality-specific barycenters distinguish grounding from reasoning failures. The many-to-many transport also handles node decomposition: if one trajectory splits a logical step and another merges it, mass is distributed across related nodes instead of enforcing one-to-one matches.

3.3 Process-aware Answer Confidence

The learned barycenters provide process-level references, but uncertainty must ultimately be assigned to candidate answers. Under black-box access, we use sampled behavior as an observable proxy for confidence: predictions supported across multiple trajectories often exhibit more consistent visual grounding and reasoning. We therefore estimate answer confidence from how closely each supporting trajectory follows the two modality-specific consensuses, rather than from answer frequency alone. For trajectory τ_k, we first compute the deviation

$$d_k = \lambda_v d_{\text{FGW},\alpha}(\mathcal{B}^v, G^v_k) + \lambda_r d_{\text{FGW},\alpha}(\mathcal{B}^r, G^r_k). \tag{7}$$

Here, λ_v and λ_r balance visual grounding and reasoning consistency. A smaller d_k indicates that the trajectory follows the consensus semantic–structural pattern, whereas a larger value marks atypical evidence or reasoning.

Because different reasoning paths may reach the same answer, we aggregate their evidence at the answer level. Let $\mathcal{A}$ be the set of unique answers and $\mathcal{T}_a = \{\tau_k \in \mathcal{T} \mid a_k = a\}$ the trajectories supporting answer a. We compute their average deviation:

$$\bar{d}_a = \frac{1}{|\mathcal{T}_a|} \sum_{\tau_k \in \mathcal{T}_a} d_k. \tag{8}$$

This aggregation captures the overall consistency of trajectories supporting answer a. Answers whose supporting trajectories are more aligned with the consensus will have smaller $\bar{d}_a$. We convert these deviations into probabilities through a softmax normalization:

$$P(a) = \frac{\exp(-\gamma \bar{d}_a)}{\sum_{a' \in \mathcal{A}} \exp(-\gamma \bar{d}_{a'})}. \tag{9}$$

As in other candidate-normalized consistency estimators (Xiong et al., 2023; Zuo et al., 2025; Zhang et al., 2025), $P(a)$ measures relative support among sampled answers rather than absolute correctness; unanimous samples therefore yield $P(a) = 1$. Each trajectory inherits $u_k = 1 - P(a_k)$, while its deviation d_k still records disagreement with the visual and reasoning consensuses. Consequently, trajectories that reach the same answer through inconsistent evidence or reasoning can remain distinguishable, and the step-level diagnostics below localize these differences.

3.4 Relative Step-level Inconsistency via Transport Alignment

Beyond answer confidence, we estimate the relative inconsistency of individual steps. Let N_m be the number of nodes in $\mathcal{B}^m$. FGW alignment between $\mathcal{B}^m$ and G_k^m yields a transport plan $\mathbf{T}_k^m \in \mathbb{R}^{N_m \times n_k^m}$ and a feature-cost matrix $\mathbf{M}_k^m$ of the same shape, where $M_{k,i,j}^m$ is the cosine distance between barycenter node i and trajectory step j. We summarize these pairwise matches for step j using the transport-weighted cost

$$\omega_{k,j}^m = \frac{\sum_{i=1}^{N_m} T_{k,i,j}^m M_{k,i,j}^m}{\sum_{i=1}^{N_m} T_{k,i,j}^m}. \tag{10}$$

The cost $\omega_{k,j}^m$ aggregates the semantic mismatch of step j over its transported mass: consistently aligned steps have lower values, whereas misaligned steps have higher values. We normalize it into a relative step-level score:

$$s_{k,j}^m = 1 - \frac{\omega_{k,j}^m - \min(\boldsymbol{\omega}_k^m)}{\max(\boldsymbol{\omega}_k^m) - \min(\boldsymbol{\omega}_k^m) + \epsilon}, \tag{11}$$

where $\boldsymbol{\omega}_k^m = \{\omega_{k,j}^m\}_{j=1}^{n_k^m}$ and $\epsilon > 0$ ensures numerical stability. The resulting scores rank steps within a trajectory according to their relative agreement with the sampled consensus. They are diagnostic indicators rather than calibrated probabilities of step correctness.

4 Experiments

This section establishes the empirical performance of OTU. We first describe the evaluation setup (Section 4.1), then report its discrimination and calibration performance across datasets and backbones (Section 4.2), and finally examine how its confidence supports step-wise reliability analysis (Section 4.3).

4.1 Experimental Setup

Datasets and models. We evaluate on MathVista (Lu et al., 2023), MMMU-Pro (Yue et al., 2025), and MME-CoT (Jiang et al., 2025). Together, they cover visual information extraction, knowledge-intensive question answering, and explicit multi-step reasoning, allowing us to test whether calibration improvements persist across different sources of difficulty rather than on one task format. We use InternVL3.5-8B (Wang et al., 2025), Qwen3-VL-8B (Bai et al., 2025), and GPT-5 (Singh et al., 2025). The two open-weight models and one API model also test whether OTU's output-only signal transfers across model families and access conditions.

Baselines. Our main evaluation follows a fully *black-box* protocol: all methods use only input–output interactions and require neither token likelihoods nor model internals. The baselines represent four common confidence signals: verbalized confidence from CoT (Xiong et al., 2023), final-answer frequency from Self-Consistency (Xiong et al., 2023; Zuo et al., 2025), embedding agreement from Degree Matrix (Lin et al., 2023), and heuristic graph structure from CenConf, PathConf, and PathWeight (Zhang et al., 2025). This grouping tests whether aligning the internal process adds information beyond self-report, answer agreement, and graph statistics. When output likelihoods are accessible, we additionally compare with Semantic Entropy (Kuhn et al., 2023) and Kernel Language Entropy (Nikitin et al., 2024). Appendix C gives the complete comparison, Appendix H describes the baselines, and Appendix I provides the full protocol and implementation details.

Table 1: Uncertainty quantification across datasets and backbones; Avg. Imp. is the mean gain over the strongest baseline.

Backbone	Method	MathVista			MMMU-Pro			MME-COT			Avg. Imp.(%)
		AUROC ↑	ECE ↓	BS ↓	AUROC ↑	ECE ↓	BS ↓	AUROC ↑	ECE ↓	BS ↓	
InternVL3.5-8B	CoT	**0.6322**	0.6405	0.6189	0.6002	0.4463	0.4387	0.5523	0.4875	0.4823	**+8.57**
	Self-Con	0.6160	0.5442	0.5101	0.6431	0.2939	0.3239	0.6640	0.2854	0.3132	
	Degree	0.6100	0.6163	0.5810	0.6428	0.3643	0.3652	0.6675	0.3868	0.3764	
	CenConf	0.6178	0.4952	0.4706	0.6375	0.2165	0.2974	0.6675	0.2854	0.3117	
	PathConf	0.6165	0.5491	0.5137	0.6514	0.3045	0.3280	0.6640	0.2849	0.3132	
	PathWeight	0.6190	0.5609	0.5260	0.6362	0.3071	0.3353	0.6640	0.2854	0.3132	
	OTU	0.6267	**0.4287**	**0.4234**	**0.6522**	**0.2062**	**0.2911**	**0.6694**	**0.1788**	**0.2804**	
Qwen3-VL-8B	CoT	**0.6186**	0.5758	0.5760	0.6251	0.3017	0.3079	0.5545	0.4074	0.4105	**+2.33**
	Self-Con	0.5973	0.5148	0.4969	0.6564	0.2901	0.3138	0.5991	0.3660	0.3809	
	Degree	0.5990	0.5700	0.5449	0.6557	0.3467	0.3469	0.6012	0.4090	0.4069	
	CenConf	0.5780	0.4828	0.4820	0.6326	0.2093	0.2859	0.5955	0.3023	0.3505	
	PathConf	0.5926	0.5276	0.5134	0.6555	0.2641	0.3008	**0.6199**	0.3525	0.3642	
	PathWeight	0.5905	0.5226	0.5088	0.6492	0.2946	0.3154	0.6197	0.3822	0.3801	
	OTU	0.5845	**0.4499**	**0.4674**	**0.6708**	**0.1895**	**0.2664**	0.6052	**0.3017**	**0.3494**	
GPT-5	CoT	0.5372	0.4521	0.4525	**0.7290**	0.2809	0.2916	0.5484	0.3956	0.3794	**+1.02**
	Self-Con	0.6175	0.4250	0.4196	0.4884	0.2675	0.3446	0.5768	0.3286	0.3716	
	Degree	0.6408	0.4764	0.4581	0.5000	0.2721	0.2721	**0.6431**	0.3607	0.3589	
	CenConf	0.5942	**0.3597**	0.3904	0.4518	0.2967	0.3349	0.5466	0.2503	0.3514	
	PathConf	0.6049	0.3933	0.4018	0.5116	0.2523	0.3217	0.5850	0.3054	0.3574	
	PathWeight	0.6222	0.4068	0.4026	0.4921	0.2637	0.3321	0.5690	0.3065	0.3673	
	OTU	**0.6588**	0.3703	**0.3731**	0.5622	**0.2465**	**0.2404**	0.5887	**0.2212**	**0.3140**	

Metrics. Following standard practice (Xiong et al., 2023; Zhang et al., 2025), we report AUROC (Bradley, 1997), Expected Calibration Error (ECE) (Naeini et al., 2015), and Brier Score (BS) (Glenn et al., 1950). AUROC measures whether erroneous predictions receive higher uncertainty than correct ones, whereas ECE and BS assess whether the confidence values themselves are meaningful. Reporting both views matters because a score may rank failures correctly while remaining systematically overconfident.

4.2 Main Result

Across the nine settings in Table 1, OTU obtains the lowest BS in all nine and the lowest ECE in eight, while achieving the highest AUROC in four. CoT and Degree sometimes rank errors competitively but remain poorly calibrated, showing that discrimination does not necessarily translate into reliable confidence. The largest calibration gain occurs on MME-CoT with InternVL3.5-8B, where OTU also maintains strong discrimination. GPT-5 shows a different trade-off: OTU does not attain the best AUROC on MMMU-Pro or MME-CoT but yields the lowest ECE and BS on both. Compared with structure-based baselines, OTU replaces heuristic graph statistics with optimal-transport alignment of semantic content and relational structure, yielding better calibration overall across the three backbones. Its weaker AUROC in several settings may partly arise because a single barycenter can treat valid minority reasoning paths as atypical. Section 5.3 examines this limitation and shows that multiple barycenters provide further calibration gains. Overall, these results demonstrate that OTU is an effective *black-box* calibration method across datasets and model families, with stronger gains in calibration than in error ranking. Appendix C further compares OTU with likelihood-aware methods and reports the complete results.

4.3 Step-wise Reliability

We next examine whether step-wise consistency helps identify unreliable predictions. At threshold t, predictions with confidence below t are flagged. Coverage is the fraction of predictions flagged, wrong-case precision is the fraction of flagged predictions that are incorrect, and wrong-case recall is the fraction of all incorrect predictions recovered. We sweep t separately for each method and report the operating point(s) with the strongest overall trade-off across these metrics: the best Step-CoT threshold and the two best OTU thresholds. Step-CoT elicits confidence at each reasoning step and achieves high precision but extremely low coverage. In contrast, OTU recovers substantially

Table 2: Process-aware reliability analysis. Cov, Prec_w, and Rec_w denote coverage, wrong-case precision, and wrong-case recall.

Method	Cov ↑	Prec_w ↑	Rec_w ↑
Step-CoT ($t = 0.9$)	0.113	**0.920**	0.143
OTU ($t = 0.55$)	0.576	0.699	0.562
OTU ($t = 0.60$)	**0.750**	0.710	**0.743**

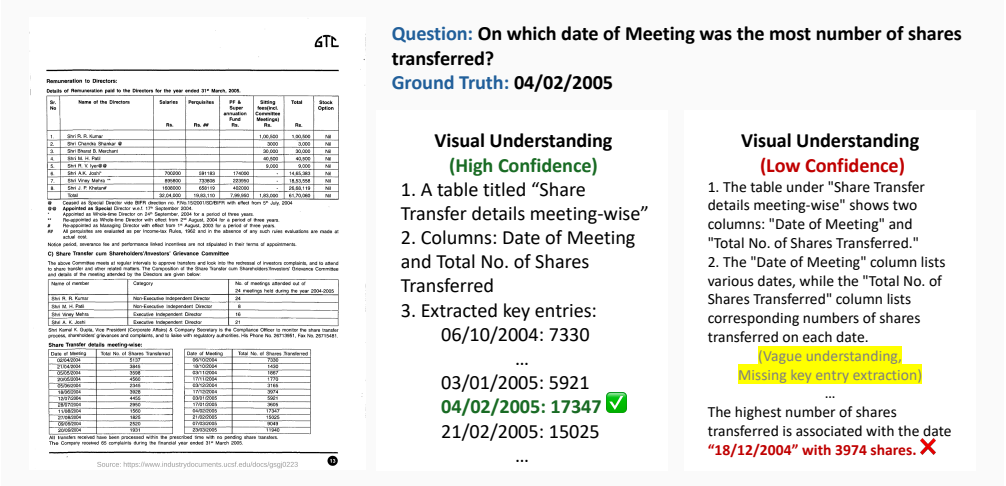

Figure 3: Visual-error diagnosis. The low-confidence trajectory omits key table entries and is assigned inconsistency primarily at visual-grounding steps.

more wrong predictions and broadens coverage without reducing precision, showing that cross-step consistency is more informative than elicited step confidence.

5 Analysis

Beyond aggregate performance, we analyze failure localization (Section 5.1), trajectory scaling (Section 5.2), multiple reasoning modes (Section 5.3), and robustness and efficiency (Section 5.4), connecting confidence scores to their failure modes and computational trade-offs.

5.1 Qualitative Analysis

Figure 3 illustrates visual-error localization on a task that requires identifying the date with the largest number of transferred shares. The high-confidence trajectory extracts the relevant table entries and reaches the correct answer, whereas the low-confidence trajectory omits key visual evidence and produces an incorrect prediction. By measuring cross-trajectory consistency, OTU assigns lower confidence to the latter answer, and its relative step scores associate the mismatch primarily with visual grounding rather than subsequent reasoning. A complementary reasoning-error case in Figure 6 of Appendix E shows the opposite pattern: trajectories share similar visual understanding, but an omitted category causes the low-confidence trajectory to fail during the median calculation. Together, these cases show that modality-specific graphs can distinguish failures originating in visual evidence from those arising in later reasoning. The diagnosis identifies both the unreliable trajectory and where its deviation begins.

5.2 Calibration Scaling

Figure 4 shows that increasing the trajectory budget K initially raises AUROC and lowers ECE and Brier score because repeated trajectories make stable visual evidence and reasoning transitions easier to identify. The curves separate after $K \approx 20$: AUROC remains nearly stable, whereas ECE and Brier score reach their lowest values and then increase slightly. Once common patterns are covered, additional samples tend to duplicate existing paths or introduce rare and noisy modes. A single barycenter must then summarize a broader mixture, which can weaken probability calibration without changing error ranking. A moderate K therefore balances pattern coverage against redundancy and motivates the mode-aware extension below.

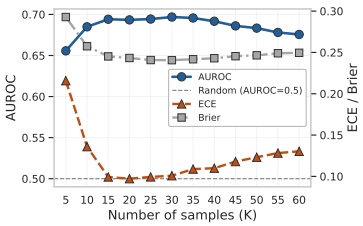

Figure 4: Calibration scaling.

5.3 Multiple Reasoning Modes

Appendix D defines D_{ij} as a weighted combination of visual and reasoning FGW distances. We cluster this matrix into C groups, learn modality-specific barycenters for each group, and score a trajectory by its distance to the nearest group without changing answer aggregation. This lets coherent minority strategies form distinct modes. Table 3 shows better calibration than answer- and graph-level baselines with comparable discrimination, supporting multiple barycenters when distinct reasoning modes occur.

Table 3: Mode-aware results (MathVista; InternVL3.5-8B).

Method	AUROC ↑	ECE ↓	BS ↓
Self-Con	0.6230	0.5060	0.4928
Degree	0.6249	0.6024	0.5632
CenConf	0.6284	0.5092	0.4771
PathWeight	<u>0.6348</u>	0.5249	0.4934
OTU	**0.6350**	<u>0.4021</u>	<u>0.4022</u>
OTU-Multi	0.6311	**0.3946**	**0.3997**

5.4 Robustness and Efficiency

We test robustness to hyperparameters and graph topology. Sweeping α, β, and λ_v over 0.1–0.9 yields stable AUROC, ECE, and BS (Appendix A.1, Figure 5). A DAG variant uses an auxiliary model to parse step dependencies and improves discrimination and calibration over chains (Appendix A.2, Table 4). On MME-CoT with InternVL3.5-8B, aggregation remains below one second at $K = 60$ and takes about 0.2% of generation time, leaving sampling as the runtime bottleneck (Appendix B, Table 5).

6 Related Work

Uncertainty Quantification for LLMs. UQ methods for LLMs and MLLMs are broadly white-box or black-box. *White-box* approaches estimate uncertainty from token probabilities or logits, often with supervised calibration or fine-tuning (Kadavath et al., 2022; Lin et al., 2022; Tian et al., 2023). MMBoundary further calibrates confidence at individual reasoning steps using token-level uncertainty and cross-modal self-rewarding signals, followed by supervised fine-tuning and reinforcement learning (He et al., 2025). This makes confidence expression part of the trained MLLM, whereas OTU requires neither internal model signals nor parameter updates. *Black-box* approaches instead use generated outputs: semantic entropy measures answer variability (Kuhn et al., 2023), while agreement and confidence-elicitation methods compare repeated samples (Xiong et al., 2023; Lin et al., 2023; Groot & Valdenegro-Toro, 2024). Their comparison unit is usually the completed answer or response, so they can miss trajectories that reach similar answers through incompatible visual evidence or intermediate reasoning.

Graph-Based Reasoning and Confidence Estimation. Prior work constructs semantic or entailment graphs over generated outputs and derives confidence from centrality or agreement (Lin et al., 2023; Jiang et al., 2024). Topo-UQ elicits subquestion–answer nodes and their logical dependencies through additional prompting, then measures uncertainty from pairwise semantic graph-edit distances between the resulting reasoning topologies (Da et al., 2025). All Roads Lead to Rome instead merges sampled reasoning paths into one directed graph, adds cross-path equivalence edges identified by an auxiliary model, and scores answer nodes using centrality, path convergence, or weighted path counts (Zhang et al., 2025). OTU neither elicits dependency graphs nor requires an auxiliary equivalence parser. It retains separate visual and reasoning trajectories and learns a modality-specific FGW barycenter, whose soft many-to-many transport aligns variable-length steps to a shared consensus. This formulation produces both answer confidence and localized process diagnostics rather than a graph-level or answer-node score alone.

7 Conclusion

We propose OTU, a black-box framework that estimates uncertainty by aligning separate visual and reasoning graphs to modality-specific optimal-transport consensuses. This process-level view captures whether sampled trajectories agree on both evidence and reasoning, rather than only on the final answer. Across three benchmarks and three backbones, OTU achieves the strongest overall balance between calibration and failure detection while providing step-level signals that distinguish visual-grounding failures from later reasoning errors. These results show that process consistency supports both reliable answer confidence and fine-grained diagnosis without access to model internals.

AI USE STATEMENT

During the preparation of this paper, we used ChatGPT solely as an auxiliary writing tool to improve the grammar, fluency, and readability of text written by the authors. The authors developed and validated all scientific ideas, as well as the methodology, experimental design, analyses, and conclusions. They reviewed the entire manuscript and take full responsibility for its content. This use of ChatGPT was limited to writing assistance and does not constitute authorship.

ETHICS STATEMENT

Our work studies black-box uncertainty estimation for MLLMs on established research benchmarks. It does not involve human participants or the collection of new personal data. The framework is intended for methodological research on model reliability rather than deployment-specific decision making. Because confidence estimates may inform downstream use, they should be validated in the target setting rather than interpreted as guarantees of correctness. All authors have reviewed the ICLR Code of Ethics and affirm adherence to its principles throughout this work.

REPRODUCIBILITY STATEMENT

We describe the trajectory-sampling protocol, graph construction, FGW objective, confidence aggregation, datasets, models, baselines, and evaluation metrics in the main paper. Appendix I provides the complete implementation settings and software environment, while the remaining appendix reports the prompts, hyperparameters, sensitivity analyses, ablations, and runtime measurements needed to reproduce the evaluation. The complete code is included in the supplementary material and is also available through the anonymous repository at https://anonymous.4open.science/r/OTU_Open-2F84.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

Luc Brogat-Motte, Rémi Flamary, Céline Brouard, Juho Rousu, and Florence D'Alché-Buc. Learning to predict graphs with fused gromov-wasserstein barycenters. In *International Conference on Machine Learning*, pp. 2321–2335. PMLR, 2022.

Shiqi Chen, Jinghan Zhang, Tongyao Zhu, Wei Liu, Siyang Gao, Miao Xiong, Manling Li, and Junxian He. Bring reason to vision: Understanding perception and reasoning through model merging. *arXiv preprint arXiv:2505.05464*, 2025a.

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas. *arXiv preprint arXiv:2503.01773*, 2025b.

Leda Cosmides and John Tooby. Are humans good intuitive statisticians after all? rethinking some conclusions from the literature on judgment under uncertainty. *cognition*, 58(1):1–73, 1996.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Longchao Da, Xiaoou Liu, Jiaxin Dai, Lu Cheng, Yaqing Wang, and Hua Wei. Understanding the uncertainty of llm explanations: A perspective based on reasoning topology. *arXiv preprint arXiv:2502.17026*, 2025.

Yichao Fu, Xuewei Wang, Yuandong Tian, and Jiawei Zhao. Deep think with confidence. *arXiv preprint arXiv:2508.15260*, 2025.

W Brier Glenn et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Tobias Groot and Matias Valdenegro-Toro. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pp. 145–171, 2024.

Shibo Hao, Yi Gu, Haotian Luo, Tianyang Liu, Xiyan Shao, Xinyuan Wang, Shuhua Xie, Haodi Ma, Adithya Samavedhi, Qiyue Gao, et al. Llm reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models. *arXiv preprint arXiv:2404.05221*, 2024.

Zhitao He, Sandeep Polisetty, Zhiyuan Fan, Yuchen Huang, Shujin Wu, and Yi R Fung. Mmboundary: Advancing mllm knowledge boundary awareness through reasoning step confidence calibration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16427–16444, 2025.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Xu Tang, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, et al. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*, 2025.

Mingjian Jiang, Yangjun Ruan, Prasanna Sattigeri, Salim Roukos, and Tatsunori Hashimoto. Graph-based uncertainty metrics for long-form language model generations. *Advances in Neural Information Processing Systems*, 37:32980–33006, 2024.

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

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

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*, 2022.

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

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. Calibrating large language models with sample consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19260–19268, 2025.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*, 2020.

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

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

Gabriel Peyré and Marco Cuturi. *Computational optimal transport: With applications to data science*. Now Foundations and Trends, 2019.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025.

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

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 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*, pp. 5433–5442, 2023.

Vayer Titouan, Nicolas Courty, Romain Tavenard, and Rémi Flamary. Optimal transport for structured data with application on graphs. In *International Conference on Machine Learning*, pp. 6275–6284. PMLR, 2019.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023.

Qianqi Yan, Yue Fan, Hongquan Li, Shan Jiang, Yang Zhao, Xinze Guan, Ching-Chen Kuo, and Xin Eric Wang. Multimodal inconsistency reasoning (mmir): A new benchmark for multimodal reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18829–18845, 2025.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, 2025.

Caiqi Zhang, Chang Shu, Ehsan Shareghi, and Nigel Collier. All roads lead to rome: Graph-based confidence estimation for large language model reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 31802–31812, 2025.

Yuhao Zhou, Yiheng Wang, Xuming He, Ao Shen, Ruoyao Xiao, Zhiwei Li, Qiantai Feng, Zijie Guo, Yuejin Yang, Hao Wu, et al. Scientists' first exam: Probing cognitive abilities of mllm via perception, understanding, and reasoning. *arXiv preprint arXiv:2506.10521*, 2025.

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, et al. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2410.03577*, 2024.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, et al. Ttrl: Test-time reinforcement learning. *arXiv preprint arXiv:2504.16084*, 2025.

A Robustness Analysis

We analyze how OTU behaves under changes to two core design choices. The hyperparameter study varies the coefficients governing semantic–structural alignment, graph construction, and visual–reasoning aggregation to characterize their trade-offs between discrimination and calibration. The graph-structure study then replaces the parser-free chain with an auxiliary-parser DAG, testing whether richer dependencies improve alignment without changing the remaining pipeline.

A.1 Hyperparameter Sensitivity

We vary one factor at a time around $\alpha = 0.3$, $\beta = 0.3$, and $\lambda_v = \lambda_r = 0.5$, using identical samples within each sweep. These sweeps test how much alignment should depend on node content, how the graph should combine semantic and sequential relations, and how visual evidence should be weighted against reasoning. Figure 5 plots all three metrics over a shared range of 0.1–0.9; for the modality sweep, $\lambda_r = 1 - \lambda_v$. Performance remains stable across broad neighborhoods of the defaults, while the curves reveal meaningful trade-offs among the three metrics. For α, a moderate semantic–structural balance around the default yields the strongest AUROC, whereas larger values slightly improve ECE and Brier score at the expense of discrimination. Varying β produces only modest changes, showing that the semantic and sequential graph relations can be combined over a relatively broad range. The modality sweep exhibits a clearer balance: small λ_v values weaken calibration, while strongly emphasizing visual evidence reduces AUROC. Assigning comparable weight to visual and reasoning deviations therefore provides the best overall compromise across discrimination and calibration.

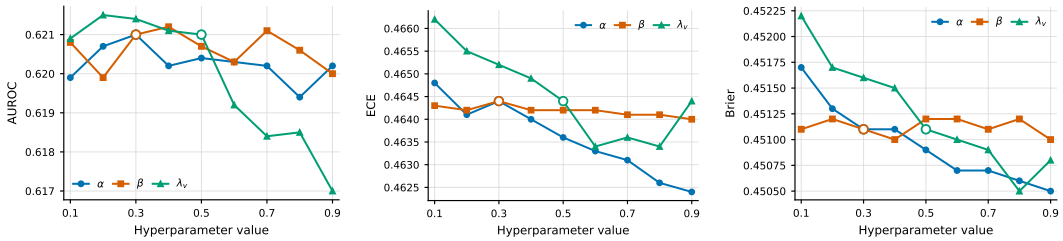

Figure 5: Hyperparameter sensitivity on MathVista with InternVL3.5-8B. Each panel reports one metric while varying α, β, or λ_v independently; for the modality sweep, $\lambda_r = 1 - \lambda_v$. Larger markers indicate the default settings.

A.2 Beyond Chain Graph Representation

The default chain connects adjacent steps and can be constructed directly from the generated sequence. To test whether it misses useful non-local dependencies, we build a directed acyclic graph from auxiliary dependency annotations while keeping the features, transport objective, and confidence aggregation fixed. Table 4 shows that the DAG improves discrimination and both calibration metrics, indicating that FGW can use richer structure when the edges are reliable. The chain remains the default because it requires no auxiliary parser; the DAG identifies a concrete route for improving graph construction.

Table 4: Graph-structure ablation on MathVista with InternVL3.5-8B.

Graph structure	AUROC ↑	ECE ↓	BS ↓
Chain	0.6267	0.4287	0.4234
DAG	**0.6472**	**0.4018**	**0.3972**

B Efficiency Analysis

We measure per-example runtime on MME-CoT with InternVL3.5-8B and separate trajectory generation from uncertainty aggregation. Generating one trajectory takes 4.861 seconds on average, excluding model and dataset loading. Table 5 shows that OTU's aggregation remains below one second up to $K = 60$ and accounts for approximately 0.2% of estimated sampling time. Although it is slower than PathWeight for small K, trajectory generation remains the dominant cost.

Table 5: Runtime on MME-CoT in seconds per example.

K	Est. sampling	PathWeight	**OTU**
10	48.613	0.018	0.100
20	97.226	0.063	0.197
30	145.840	0.136	0.300
40	194.453	0.238	0.398
50	243.066	0.375	0.531
60	291.679	0.540	0.548

Inference was measured on an NVIDIA A100 GPU and aggregation on an NVIDIA V100 GPU with CUDA 12.2, on a server with dual Intel Xeon Gold 6240R CPUs. Computation is independent across examples, modalities, and trajectory-to-barycenter alignments, so these small OT problems can also be batched or parallelized. For substantially longer trajectories, approximate or low-rank FGW solvers provide a natural extension.

C Additional Likelihood-Aware Baselines

The main evaluation follows a fully black-box protocol and assumes no access to token likelihoods or model internals. To test whether process-level consistency remains useful when likelihoods are available, we additionally compare against Semantic Entropy with answer log-probabilities (Kuhn et al., 2023) and Kernel Language Entropy (Nikitin et al., 2024) on InternVL3.5-8B. Table 6 shows that these methods improve calibration over several output-only baselines, while OTU achieves the lowest ECE and Brier score. This comparison is restricted to the open-weight setting and is not available for GPT-5 under our API access.

Table 6: Likelihood-aware comparison on MathVista with InternVL3.5-8B. OTU remains fully output-based.

Method	AUROC ↑	ECE ↓	BS ↓
CoT	**0.6322**	0.6405	0.6189
Self-Con	0.6160	0.5442	0.5101
Degree	0.6100	0.6163	0.5810
CenConf	0.6178	0.4952	0.4706
PathConf	0.6165	0.5491	0.5137
PathWeight	0.6190	0.5609	0.5260
Semantic Entropy + log-prob.	0.6070	<u>0.4470</u>	0.4390
Kernel Language Entropy	0.6020	0.4630	<u>0.4360</u>
OTU	<u>0.6267</u>	**0.4287**	**0.4234**

D Multiple Reasoning Modes

A single barycenter assumes that sampled trajectories concentrate around one dominant reasoning pattern. To accommodate multiple valid modes, we first form a pairwise process-distance matrix

$$D_{ij} = \lambda_v d_{\mathrm{FGW},\alpha}(G_i^v, G_j^v) + \lambda_r d_{\mathrm{FGW},\alpha}(G_i^r, G_j^r), \tag{12}$$

and cluster the trajectories into C groups $\{\mathcal{T}^{(c)}\}_{c=1}^{C}$. For each cluster and modality, we compute a mode-specific barycenter $\mathcal{B}_c^m$ and use

$$\widetilde{d}_k = \min_c \left[\lambda_v d_{\mathrm{FGW},\alpha}(\mathcal{B}_c^v, G_k^v) + \lambda_r d_{\mathrm{FGW},\alpha}(\mathcal{B}_c^r, G_k^r)\right] \tag{13}$$

in place of d_k, leaving the rest of OTU unchanged. A trajectory is therefore considered reliable when it is consistent with any discovered mode, rather than only with a global average.

Table 3 reports the full comparison with $K = 10$. The mode-aware variant slightly improves calibration while preserving comparable discrimination. This pattern is useful because the extension is intended to distinguish coherent alternative strategies, not to change the ranking signal supplied by the underlying trajectory distances.

E Case Study: Reasoning Error Detection

Figure 6 provides a detailed example of reasoning error detection under similar visual understanding. The task asks whether *Orange Red* is the high median of a set of bar values. All sampled trajectories correctly interpret the visual content, identifying the categories and their corresponding values from the bar chart.

However, the trajectories diverge in the reasoning process. The high-confidence trajectory correctly lists all eight values, sorts them, and computes the median as the average of the 4th and 5th elements, leading to the correct answer. In contrast, the low-confidence trajectory omits one category, incorrectly assumes an odd number of elements, and applies an incorrect median rule, resulting in a wrong prediction.

This example highlights that even with consistent visual understanding, errors can arise purely from reasoning inconsistencies. By analyzing step-wise alignment across trajectories, OTU assigns lower confidence to such incomplete or inconsistent reasoning paths, effectively identifying failures that are not detectable from final answers alone.

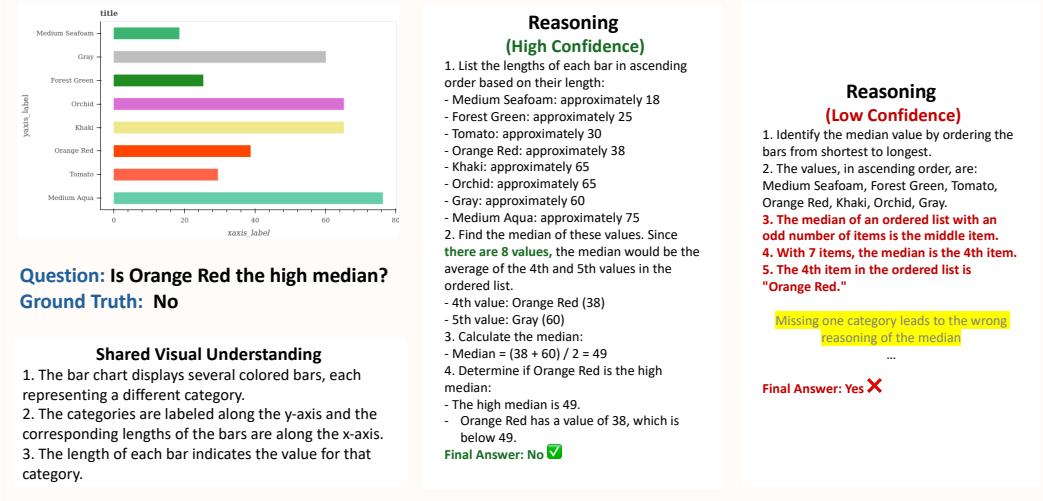

Figure 6: Case study of reasoning uncertainty. Low-confidence trajectories exhibit inconsistent reasoning despite similar visual understanding.

F Alternative Node Masses

Uniform mass is a neutral prior for the fully black-box setting: it requires neither token probabilities nor an auxiliary importance estimator. It also does not imply that every node produces the same alignment cost, since the learned transport plan and semantic-structural mismatch determine the

final deviation. When step-importance scores $w_j \geq 0$ are available, non-uniform masses can be incorporated directly as

$$p_j = \frac{w_j}{\sum_{\ell=1}^{n} w_\ell}. \tag{14}$$

Possible signals include step-level token entropy for open-weight models, verbalized confidence for black-box models, position-aware weights, or graph centrality. We retain uniform mass to avoid coupling OTU to an additional confidence model.

G Additional Analysis: Reliability Diagrams

Figure 7 presents reliability diagrams comparing confidence calibration between CoT and OTU. The left panel shows CoT confidence, which exhibits significant overconfidence: predictions with high confidence are often incorrect, leading to large gaps between predicted confidence and empirical accuracy (ECE = 0.4513). In contrast, the right panel shows OTU (with $K = 35$), where predictions are much better aligned with the ideal diagonal, resulting in substantially improved calibration (ECE = 0.1088). This improvement demonstrates that OTU produces more reliable uncertainty estimates by aligning confidence with actual correctness. By leveraging cross-trajectory consistency, our method reduces systematic overconfidence and yields confidence scores that more faithfully reflect prediction reliability. This suggests that process-level consistency provides a stronger signal for calibration than answer-level confidence alone.

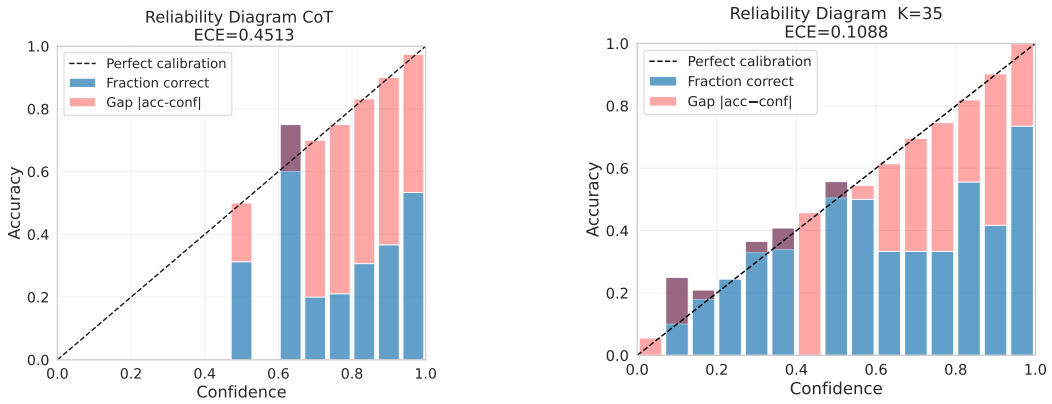

Figure 7: Reliability diagrams on MMMU-Pro with InternVL3.5-8B. Left: CoT confidence (ECE $= 0.4513$). Right: OTU at $K = 35$ (ECE $= 0.1088$).

H Baseline Implementations

We provide brief descriptions of the output-only baselines used in our main experiments. All main baselines are black-box and rely only on model inputs and outputs. For baselines with publicly available implementations, we adopt their official codebases. For those without released code, we implement them following the descriptions provided in the original papers. The separate likelihood-aware comparison is described in Appendix C.

- **CoT Confidence (CoT Conf).** The model is prompted to generate a reasoning chain together with a self-reported confidence score. The confidence is directly extracted from the model's verbalized output, reflecting its subjective belief about the correctness of the final answer.
- **Self-Consistency (Self-Con).** Multiple reasoning trajectories are sampled under stochastic decoding. Confidence is estimated based on the agreement among final answers, where higher agreement indicates higher confidence.
- **Degree Matrix (Degree).** An embedding-based aggregation method that represents sampled outputs in a semantic space and constructs a similarity graph. Confidence is derived from the degree (connectivity) of each answer node, reflecting how well it is supported by semantically similar outputs.

- **CenConf.** A graph-based method that merges reasoning paths into a unified graph and computes confidence using node centrality (e.g., Katz centrality). Answers that are more central in the graph are assigned higher confidence.
- **PathConf.** Confidence is estimated based on the number of distinct reasoning paths that lead to each candidate answer. Answers supported by more paths are considered more reliable.
- **PathWeight.** An extension of path-based aggregation that assigns higher weights to paths containing commonly shared reasoning steps. Confidence is computed by aggregating weighted path scores, emphasizing consistent reasoning patterns across trajectories.

I Implementation Details

Inference and Sampling. For open-source models, we perform stochastic decoding with temperature set to 0.6 during inference scaling. For each query, we sample $K = 5$ reasoning trajectories unless otherwise specified. Every trajectory follows the same structured response format, separating visual observations, reasoning steps, and the final answer. We reuse the exact sampled set across aggregation-based baselines and OTU, which removes decoding variation from method comparisons. The sample count changes only in the scaling study.

Optimal Transport and Hyperparameters. We implement optimal transport using the POT (Python Optimal Transport) library. Our default configuration is $\alpha = 0.3$, $\beta = 0.3$, $\lambda_v = \lambda_r = 0.5$, and $\gamma = 3.0$, and we keep it fixed across datasets and backbones. Appendix A.1 varies one factor at a time while holding the others fixed, isolating which component could affect calibration.

Software Environment. Experiments use Python 3.11.9, PyTorch 2.8.0, torchvision 0.23.0, Transformers 4.56.1, vLLM 0.10.2, LMDeploy 0.9.2, Accelerate 1.8.1, and qwen-vl-utils 0.0.11. Optimal transport and evaluation use POT 0.9.6.post1, NumPy 1.26.4, SciPy 1.16.1, scikit-learn 1.3.2, and netcal 1.3.6. Data and representation utilities use datasets 4.0.0, sentence-transformers 5.2.0, and Pillow 10.4.0.

DAG Ablation. For the optional DAG variant, we use GPT-4o-mini to extract directed dependencies among visual-understanding and reasoning steps, then encode these dependencies in the structure matrix. The feature representation, FGW solver, barycenter construction, and uncertainty aggregation remain unchanged. This variant tests the flexibility of the framework, while the default chain avoids dependence on an auxiliary parser.

Embedding Representation. OTUis agnostic to the choice of text representation, and contextual embedding models can be substituted without changing the graph construction or FGW objective. In our experiments, wherever a method requires step or output embeddings, we use the same lightweight hashing-based representation. This common backend avoids loading an external embedding encoder and its associated inference latency, while ensuring that differences between methods do not arise from encoder capacity. Specifically, each step is tokenized into alphanumeric tokens, mapped independently into a fixed-dimensional vector through a hashing function, and then ℓ_2-normalized. The resulting representation requires no additional training or labeled step supervision and keeps representation overhead small relative to trajectory generation.

Compute Resources. All experiments are conducted on servers with two NVIDIA A100 or V100 GPUs, using mixed-precision inference where applicable. Generation and uncertainty aggregation are timed separately because they have different computational profiles; this separation makes clear how much overhead is introduced by uncertainty estimation rather than by the underlying MLLM.

J PROMPT TEMPLATE

Multimodal VQA Reasoning Prompt

System. You are a multimodal VQA assistant. Given an image and a question, first describe the key visual facts you will use, then reason step-by-step, and finally answer.

Task. Given an image–question pair, provide a response that follows the system requirements and strictly adheres to the format below.

Expected response format.

```
## Visual Understanding
<visual facts, bullet/numbered steps>
## Reasoning
<reasoning steps>
## Final
Answer:  <final answer>
```

Rules.

1. Keep the final answer as short as possible (e.g., a number, choice letter, or brief phrase).
2. Use English only.
3. Do not output any text outside the required format.

Output. {Placeholder}