CONFIDENCE NEEDS EVIDENCE: RELIABLE CONFIDENCE ESTIMATION FOR MULTIMODAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable confidence estimation is essential for identifying trustworthy predictions and enabling reliable model deployment. However, we find that in video-based tasks, existing methods based on output probabilities or response consistency can assign unreliable confidence driven by language priors. This motivates us to reweight confidence by the evidence support obtained from multimodal inputs. We propose EVICONF, a confidence estimator that accounts for both content and temporal information in video. It measures content support through the selected answer's probability lift from a question-only input to the full input, and temporal support through the fraction of temporal shuffles that lower the same answer's probability. The resulting evidence score reweights normalized predictive-entropy confidence through a geometric mean to distinguish correct high-confidence predictions from confident errors. Extensive experiments across eight models and four benchmarks covering video QA and audio-visual QA show that EVICONF achieves state-of-the-art correctness AUROC. Ablation studies further show that evidence support reduces the confident-error rate by up to 92% relative to internal confidence alone, demonstrating the robustness and effectiveness of EVICONF in mitigating false confidence in multimodal large language models (MLLMs).

1 INTRODUCTION

Large language models play an increasingly important role in understanding and generating information, making reliable assessment of their predictions essential for practical deployment. Confidence estimation serves this purpose by helping systems decide which predictions to trust and which to defer for review or reject (Guo et al., 2017; Jiang et al., 2021). However, false confidence, defined as high confidence in incorrect predictions, remains a persistent challenge. Existing estimators based on output probabilities or response consistency can assign high scores even when a model confidently repeats an error (Geng et al., 2024). This problem is particularly acute in multimodal video question answering, where rich content and temporal information require predictions to depend on the supplied input: reliable answers must reflect both what occurs and how events unfold, yet high confidence does not necessarily indicate that the model has used this evidence.

To understand why false confidence persists in these tasks, we conduct three connected diagnostic analyses (Section 3). First, we compare probability-based and sampling-based estimators and find strong correlations, indicating that they largely capture a shared internal-confidence signal. We then examine whether this signal becomes more reliable as models scale: high-confidence predictions become more prevalent, but their error rates do not decrease correspondingly. Finally, we test whether support from the input distinguishes reliable from unreliable confidence. Holding the full-input prediction fixed, we compare its probability with and without multimodal evidence and find that, among predictions with similarly high internal confidence, stronger evidence support corresponds to fewer errors. Together, these findings indicate that language priors can sustain high confidence without sufficient support from the input, leading to confident errors. Confidence unsupported by multimodal evidence is therefore unreliable, motivating us to measure the support contributed by that evidence and use it to reweight internal confidence.

We therefore propose EVICONF (*Evidence-Supported Confidence*), a training-free estimator that weights internal confidence by support from multimodal inputs. Because video evidence contains both content and temporal information, we assess the contribution of each while holding the full-input

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.

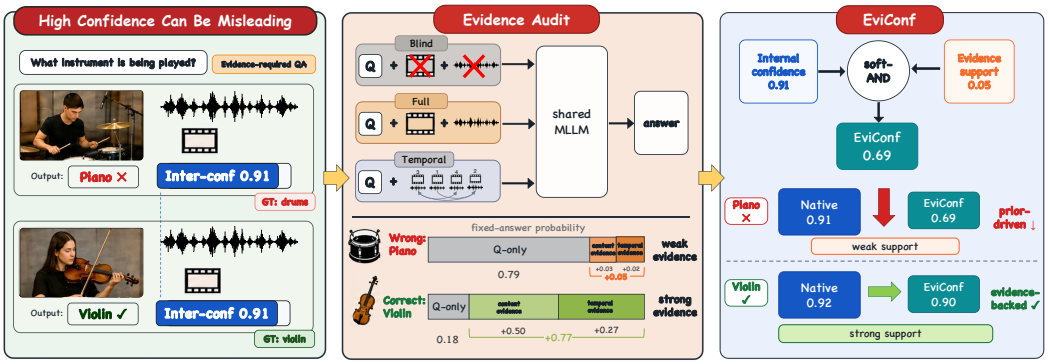


Figure 1: Overview of EVICONF. Existing estimators can assign high confidence to incorrect predictions driven by language priors. Our method uses an evidence score derived from content and temporal information to reweight internal confidence to better distinguish correct high-confidence predictions from confident errors.

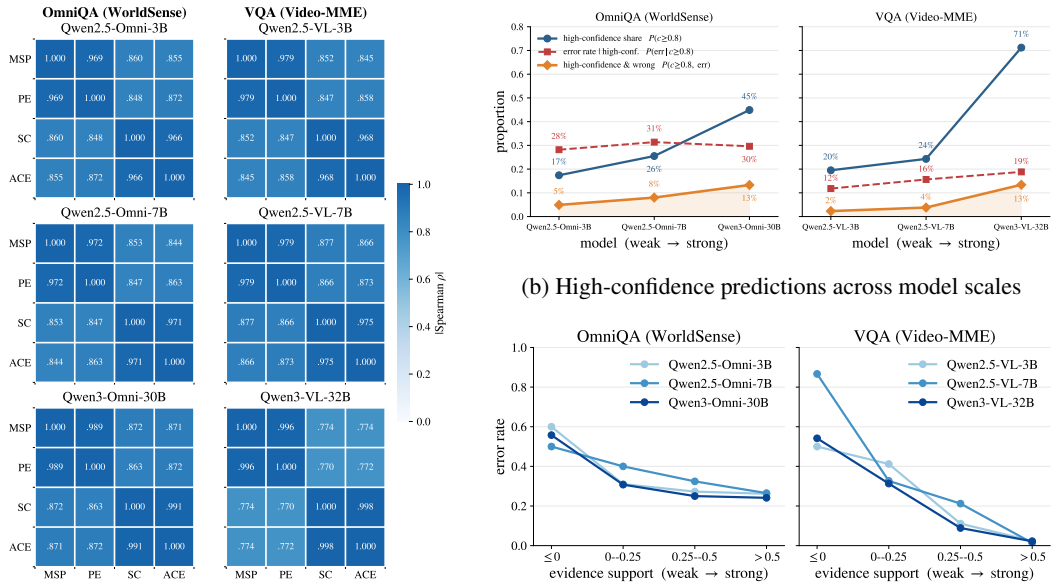
prediction fixed. Content support is measured by the selected answer's probability lift from a question-only input to the full input; temporal support is measured by the fraction of four temporal shuffles that lower the same answer's probability. For audio-visual inputs, frames and aligned audio segments are shuffled jointly to preserve local correspondence. The temporal factor modulates the content lift to form an evidence score, which reweights normalized predictive-entropy confidence through a parameter-free geometric mean. This evidence weighting helps distinguish correct high-confidence predictions from confident errors without sampled answer rollouts. Figure 1 illustrates EVICONF.

We conduct extensive experiments across eight models from multiple families and scales on four benchmarks: Video-MME and MVBench for video QA, and WorldSense and DailyOmni for audio-visual QA. EVICONF achieves state-of-the-art correctness AUROC across the evaluated settings, while ablation studies validate the contributions of evidence support and its integration with internal confidence. In particular, at a fixed prediction coverage, adding evidence support reduces the confident-error rate by up to 92% relative to internal confidence alone. These results demonstrate the effectiveness of EVICONF in mitigating false confidence across MLLMs and support using multimodal evidence to select more reliable predictions. Our contributions are as follows:

- We reveal that existing confidence estimators can assign unreliable confidence to multimodal predictions driven by the model's language priors. To address this, we measure how much evidence the selected answer receives from multimodal inputs and use this support to reweight internal confidence, thereby mitigating false confidence.
- We propose EVICONF, a training-free confidence estimator that incorporates both content and temporal evidence. It measures content support through the selected answer's full-versus-blind probability lift, modulates this support by the fraction of temporal shuffles that lower the same answer's probability, and combines the resulting evidence score with normalized predictive-entropy confidence through a parameter-free geometric mean.
- We present extensive experiments across multiple model families and scales on four benchmarks (Video-MME, MVBench, WorldSense, and DailyOmni), covering both video QA and audio-visual QA. EVICONF achieves state-of-the-art correctness AUROC in all 16 benchmark–model settings, while ablation studies validate the effectiveness of our evidence-based confidence estimation in reducing confident errors.

2 RELATED WORK

Confidence estimation. Confidence estimation seeks a score that predicts whether a model output is correct (Guo et al., 2017; Jiang et al., 2021). Existing methods draw on three broad signals. *Probability-based methods* use the concentration of the predictive distribution (Hendrycks & Gimpel, 2017). *Sampling-based methods* use agreement or semantic variation across repeated generations (Wang et al., 2023; Manakul et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024; Nikitin



(a) Pairwise estimator correlations

(b) High-confidence predictions across model scales

(c) Error rate by signed evidence lift

Figure 2: **Confidence needs evidence. (a)** Operationally different estimators largely recover the same internal-confidence signal. **(b)** Larger models become more confident without reducing high-confidence errors. **(c)** Stronger full-versus-blind content support corresponds to lower error rates.

et al., 2024; Nguyen et al., 2025). Multimodal extensions such as VL-Uncertainty and DOUBT introduce input or reasoning perturbations to expose false confidence that may remain hidden under direct sampling, but ultimately infer confidence from response variation or concentration (Zhang et al., 2024; Chen et al., 2026). *Representation-based methods* instead use hidden activations or the stability of response representations, often through an external correctness estimator trained on these internal signals (Chen et al., 2024; Radharapu et al., 2026; Khanmohammadi et al., 2025). Across these families, confidence estimation primarily relies on output behavior or internal state without accounting for the directional support contributed by task evidence, which can lead to unreliable confidence estimates.

Multimodal evidence contrasts. Prior work extracts information about modality contributions by contrasting full inputs with degraded or missing-modality references. In VQA, RUBi and Counterfactual VQA use such contrasts to mitigate language biases (Goyal et al., 2017; Cadene et al., 2019; Niu et al., 2021), while VAUQ quantifies visual evidence through its contribution to uncertainty reduction in LVLM self-evaluation (Park et al., 2026). For generation, visual contrastive decoding and related methods suppress hallucinations through visual contrasts (Leng et al., 2024; Favero et al., 2024), and PMI-style rescoring uses a content-free reference to improve answer selection (Holtzman et al., 2021). Motivated by these approaches, we measure content support through the full-versus-blind probability lift of a fixed selected answer and temporal support through the consistency with which shuffling lowers that answer's probability. Together, these signals provide content and temporal evidence for reweighting internal confidence.

3 Confidence Needs Evidence

To investigate the causes of false confidence, we conduct three diagnostic analyses on 1,000-example evaluation sets from Video-MME and WorldSense, covering three Qwen model scales per benchmark. We examine estimator agreement, confidence reliability across scales, and the contribution of input evidence (Figure 2). Appendix A provides the complete protocol.

We first compare probability-based estimators (MSP and PE) with sampling-based estimators (SC and ACE). Despite using different mechanisms, their scores are strongly correlated across benchmarks and models (Figure 2a), indicating that they largely capture a shared internal-confidence signal. We

then use normalized predictive-entropy confidence c_{int} as a representative signal to examine whether internal confidence becomes more reliable as models scale. Larger models assign high confidence to more predictions, yet the error rate among these predictions increases on Video-MME and remains substantial on WorldSense (Figure 2b). Increasing model scale therefore strengthens confidence without resolving confident errors. Finally, we test whether input evidence helps distinguish reliable from unreliable confidence. Holding the full-input prediction fixed, we measure its probability lift relative to a blind, question-only input. Among predictions with similarly high internal confidence, stronger evidence support corresponds to lower error rates on both benchmarks (Figure 2c). Thus, the contribution of multimodal evidence provides information about correctness beyond internal confidence alone. Together, these analyses indicate that high confidence can persist through language priors without sufficient input evidence, making it unreliable in multimodal tasks. Evidence support helps distinguish correct high-confidence predictions from confident errors and is therefore important for reliable confidence estimation.

4 METHOD

EVICONF combines internal confidence with evidence support that accounts for both content and temporal order. The content term measures how much the supplied evidence supports the model's selected answer relative to a blind input. The temporal factor measures how consistently disrupting the evidence order reduces support for that same answer. Both are model-relative measurements; neither assumes that the supported answer is factually correct.

Problem formulation and internal confidence. We consider a multiple-choice question with textual query q, answer choices $\mathcal{Y} = \{1, \ldots, C\}$, and task evidence $e_{\mathcal{M}}$. In our experiments, $\mathcal{M} = V$ for video QA and $\mathcal{M} = AV$ for joint audio-visual QA. The full-input option distribution, prediction, and internal confidence are defined as:

$$\boldsymbol{p}^{\mathcal{M}} = p(\cdot \mid q, e_{\mathcal{M}}), \qquad \hat{y} = \arg\max_{y \in \mathcal{Y}} p_y^{\mathcal{M}}, \qquad c_{\text{int}} = 1 - \frac{H(\boldsymbol{p}^{\mathcal{M}})}{\log C}. \tag{1}$$

Here, c_{int} is the normalized complement of predictive entropy. We hold $\hat{y}$ fixed under all subsequent interventions. Our primary objective is to rank correct predictions above incorrect ones. Absolute agreement between confidence and empirical accuracy is a secondary calibration diagnostic.

Content evidence support. We construct a blind distribution $\boldsymbol{p}^{\emptyset} = p(\cdot \mid q)$ by retaining the question, answer choices, and prompting format while removing the audio and video evidence. The content term is the signed probability lift $\ell^{\mathcal{M}} = p_{\hat{y}}^{\mathcal{M}} - p_{\hat{y}}^{\emptyset} \in [-1, 1]$. A positive lift indicates that adding the task evidence increases support for the selected answer relative to the language-only reference. This comparison measures the contribution of the available content, but does not separately determine whether its temporal order matters.

Temporal evidence support. We generate $K = 4$ shuffled versions of the task evidence. Each intervention randomly permutes frame content while keeping the timestamp labels in ascending order, so that the labels do not reveal the original frame positions. For audio-visual inputs, the same permutation is applied jointly to frames and their time-aligned audio segments, preserving their local correspondence while disrupting the original sequence. Let $p_{k,\hat{y}}^{\mathcal{M}}$ denote the probability of the fixed answer under shuffled input k. We define:

$$\tau^{\mathcal{M}} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}\left[p_{k,\hat{y}}^{\mathcal{M}} < p_{\hat{y}}^{\mathcal{M}}\right] \in [0, 1]. \tag{2}$$

This score measures the fraction of shuffles that lower the answer's probability. It is large when the original order consistently supports that answer relative to the shuffled alternatives. It measures the frequency of a decrease, rather than its magnitude, and does not by itself establish correct temporal reasoning. Strict inequality means that unchanged probabilities do not count as decreases.

Evidence-weighted confidence. We normalize the product of content and temporal support to $[0, 1]$ and combine it with internal confidence through a geometric mean:

$$c_{\text{evid}}^{\mathcal{M}} = \frac{1 + \ell^{\mathcal{M}} \tau^{\mathcal{M}}}{2}, \qquad c_{\text{EviConf}}^{\mathcal{M}} = \sqrt{c_{\text{int}}\, c_{\text{evid}}^{\mathcal{M}}}. \tag{3}$$

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Benchmarks. We evaluate video QA on fixed random samples of 1,000 examples from Video-MME and 1,000 from MVBench (Fu et al., 2025; Li et al., 2024), and audio-visual QA on a fixed random sample of 1,000 WorldSense examples and the complete 1,197-example DailyOmni evaluation set (Hong et al., 2025; Zhou et al., 2025). We use these subsets to reduce computational cost. Each subset is sampled once and shared across all models and confidence estimators.

Models. For video QA, we use Qwen2.5-VL-3B/7B, Qwen3-VL-32B, and the cross-family InternVL3-8B model (Zhu et al., 2025). For audio-visual QA, we use Qwen2.5-Omni-3B/7B, Qwen3-Omni-30B-A3B, and the cross-family Ola-7B model. This design tests both scaling within the Qwen family and transfer of the confidence principle across model families.

Baselines. We compare with probability-based MSP and predictive entropy (PE), rollout-based self-consistency (SC) and answer-choice entropy (ACE), the perturbation-based VL-Uncertainty estimator, and the object-aware rollout method DOUBT. We also include a supervised white-box Probe reference that predicts correctness from the last-token hidden state. One probe is trained per model using only non-test data: 3,700 pooled Video-MME/MVBench examples for vision-language models and 2,000 WorldSense examples for omni-modal models, with the latter transferred unchanged to DailyOmni. All remaining methods are training-free.

Evaluation metrics. AUROC is the primary metric because it measures whether a confidence score ranks correct predictions above errors, the discrimination required for threshold-based acceptance or deferral. We also report 10-bin expected calibration error (ECE) as a secondary diagnostic of absolute score–accuracy agreement. See Appendix D for the complete protocols.

5.2 MAIN RESULTS

Video QA. As shown in Table 1, EVICONF achieves the highest AUROC in all eight video-QA settings, including the cross-family InternVL model, with a maximum gain of 10.0 points. The gains are especially pronounced for the largest Qwen model, showing that evidence support remains useful as models scale and helps demote confident but unsupported predictions rather than merely reproducing output concentration.It also achieves the lowest ECE in three settings. **Audio-Visual QA.** eviconf is not limited to video QA: for any set of task evidence, it measures the corresponding evidence support beyond internal confidence. Table 2 applies the same estimator to joint audio-visual evidence, where EVICONF again ranks first in all eight settings and transfers to the cross-family Ola model. The AUROC gains follow the same direction as in video QA, supporting the generality of EVICONF. EVICONF also obtains the lowest ECE in three settings.

Enhanced confidence discrimination. Figure 3 provides a visual illustration of EVICONF's improved separation of correct and incorrect predictions across both task families and all three Qwen scales. The horizontal axis shows internal confidence, the vertical axis shows evidence support, and the blue curves denote EVICONF iso-confidence contours. Correct and incorrect predictions overlap substantially along the internal-confidence axis, but many incorrect predictions exhibit weaker evidence support and thus receive lower joint scores. By down-weighting these confident yet weakly supported errors, EVICONF yields a clearer separation between correct and incorrect predictions than internal confidence alone. This pattern is consistent across all settings.

5.3 DEEP ANALYSIS

We test whether the full-versus-blind lift reflects support from task evidence rather than arbitrary multimodal input, reference artifacts, or language answerability. Holding the full-input prediction $\hat{y}$ fixed, we compare matched and mismatched task evidence, text-only and format-matched references, and examples with strong and weak blind text priors. Appendix D.3 provides the complete protocol.

Content specificity. Matched task evidence produces a substantial positive lift, whereas mismatched evidence has a near-zero mean effect (Figure 4a). Although mismatched inputs can shift individual probabilities in either direction, the matched lift is more consistent: the matched-minus-mismatch gap is positive for 82–91% of examples and is more predictive of correctness in every evaluated cell.

Table 1: Results for vision-language models. Avg. denotes the macro-average AUROC over four models within each benchmark. # Inf. counts model calls per example, with each rollout or perturbation counted once. †Probe requires supervised training.

		Qwen2.5-VL-3B		Qwen2.5-VL-7B		Qwen3-VL-32B		InternVL3-8B		
Method	# Inf.	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	Avg.
Video-MME (random $n = 1{,}000$)										
MSP	1	74.7	17.1	71.5	13.0	75.8	19.6	75.0	15.8	74.3
PE	1	76.3	7.1	72.9	**4.2**	76.2	16.9	74.9	10.3	75.1
SC	10	70.8	19.8	70.5	15.6	64.7	20.2	69.8	18.8	69.0
ACE	10	71.9	9.9	70.1	13.4	64.7	17.6	70.0	13.6	69.2
VL-Uncertainty	5	64.5	15.4	65.5	13.4	58.2	19.3	61.0	11.4	62.3
Probe	1†	65.7	**6.6**	64.3	6.1	64.9	13.2	63.5	**9.5**	64.6
DOUBT	21	68.4	22.8	65.4	22.0	62.7	22.9	65.5	23.3	65.5
EVICONF (Ours)	6	**80.2**	15.6	**79.3**	9.6	**85.0**	**12.4**	**78.1**	10.8	**80.7**
MVBench (random $n = 1{,}000$)										
MSP	1	74.7	8.7	71.5	9.4	75.8	17.0	84.8	10.2	76.7
PE	1	77.6	**4.1**	74.6	5.9	76.5	13.2	83.7	**4.6**	78.1
SC	10	72.8	13.1	70.7	11.1	67.7	17.8	72.2	13.5	70.9
ACE	10	74.3	13.7	72.5	15.4	68.2	15.2	72.3	12.5	71.8
VL-Uncertainty	5	65.8	11.1	63.9	12.7	62.3	14.7	64.1	15.7	64.0
Probe	1†	64.5	8.2	59.2	12.7	67.9	16.5	74.7	9.5	66.6
DOUBT	21	72.1	16.8	67.6	19.4	64.9	19.2	74.1	15.8	69.7
EVICONF (Ours)	6	**79.1**	6.6	**78.9**	**5.5**	**86.5**	**9.6**	**85.9**	7.7	**82.6**

Table 2: Results for omni-modal models. Avg. denotes the macro-average AUROC over four models within each benchmark. # Inf. counts model calls per example, with each rollout or perturbation counted once. †Probe requires supervised training.

		Qwen2.5-Omni-3B		Qwen2.5-Omni-7B		Qwen3-Omni-30B		Ola-7B		
Method	# Inf.	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	Avg.
WorldSense (random $n = 1{,}000$)										
MSP	1	63.6	22.2	66.6	25.9	68.5	28.4	63.5	17.8	65.6
PE	1	64.4	14.5	67.0	14.3	67.9	19.1	63.3	**10.8**	65.7
SC	10	60.7	26.3	63.6	28.3	66.0	31.1	61.5	22.1	63.0
ACE	10	61.6	20.1	63.5	20.1	66.2	24.5	62.1	17.0	63.4
VL-Uncertainty	5	59.3	22.2	61.0	21.4	58.3	27.7	59.2	20.9	59.5
Probe	1†	57.5	**6.1**	57.3	**11.1**	57.6	**8.7**	57.4	12.8	57.5
DOUBT	21	61.0	30.1	62.6	31.5	62.1	31.9	59.4	26.5	61.3
EVICONF (Ours)	6	**66.3**	15.5	**68.8**	18.0	**68.7**	17.4	**64.6**	14.5	**67.1**
DailyOmni (full set, $n = 1{,}197$)										
MSP	1	73.3	10.0	70.2	15.4	74.6	17.1	72.6	14.1	72.7
PE	1	73.4	14.4	70.2	11.2	74.8	11.4	72.8	**6.1**	72.8
SC	10	70.4	13.9	67.0	18.0	68.8	18.4	69.1	18.2	68.8
ACE	10	70.6	11.7	67.5	13.4	69.0	14.5	70.1	12.5	69.3
VL-Uncertainty	5	62.6	13.8	59.8	18.1	60.3	18.2	64.7	16.8	61.9
Probe	1†	49.4	14.1	55.1	13.2	52.4	25.4	64.9	8.2	55.5
DOUBT	21	67.2	18.7	63.1	21.9	63.2	18.7	69.2	22.3	65.7
EVICONF (Ours)	6	**75.9**	**6.1**	**73.0**	**6.0**	**77.1**	**5.5**	**73.0**	10.4	**74.8**

Appendix D.4 reports the sample-level analysis. **Reference validity.** Black and noise references yield lifts similar to the text-only reference, ruling out artifacts caused by removing the multimodal input interface (Figure 4b). **Language priors.** The lift remains informative under both strong and weak text priors (Figure 4c), showing that its usefulness is not explained solely by question-only answerability. This result supports its complementarity with internal confidence. These controls suggest that EVICONF mitigates unsupported false confidence. However, it cannot in principle fully resolve false confidence caused by misinterpreted evidence.

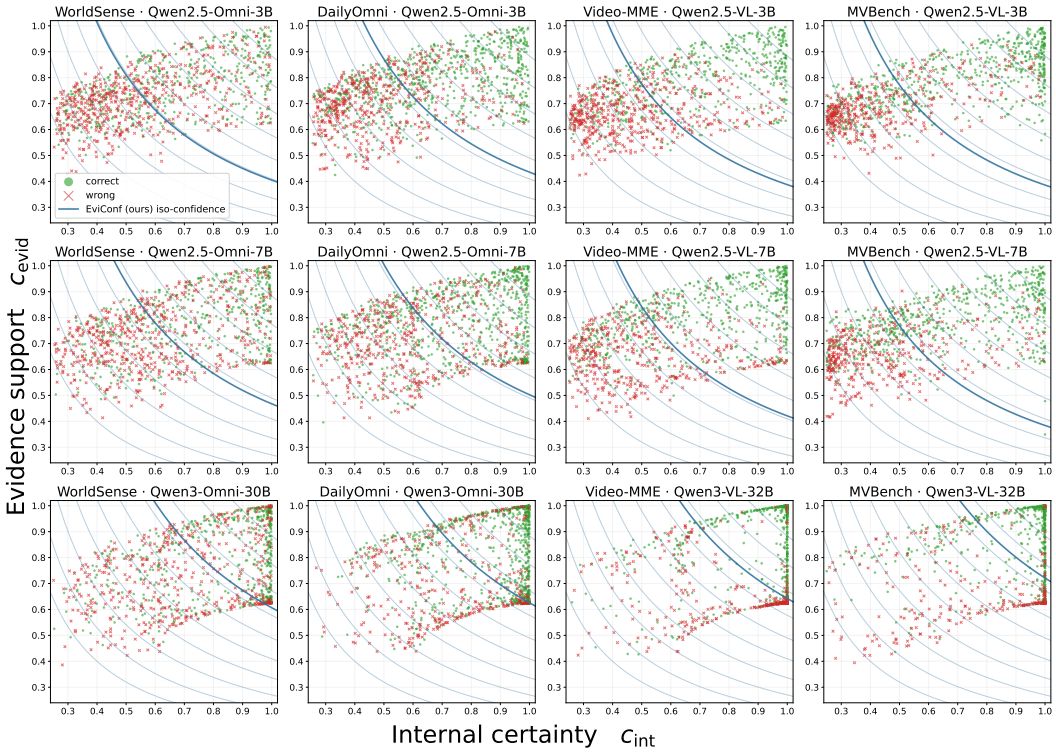
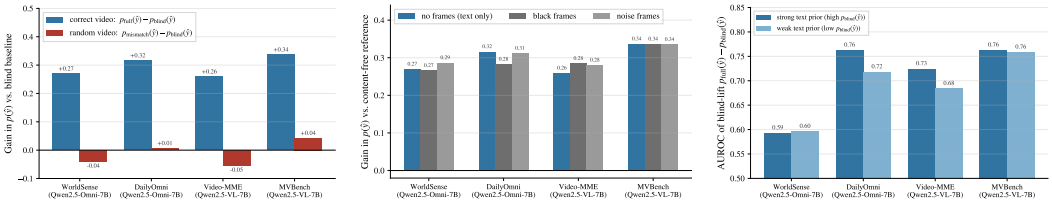


Figure 3: EVICONF combines internal confidence with evidence support to better separate correct from incorrect predictions. Blue curves show EVICONF iso-confidence contours.



(a) Evidence specificity (b) Reference-condition controls (c) Strong vs. weak text priors

Figure 4: Evidence lift captures evidence support rather than artifacts or language priors.

6 ABLATION STUDY

Our ablation study examines whether evidence support reduces confident errors, joint audio-visual evidence outperforms either source alone, temporal support adds value beyond content support, and performance depends on geometric-mean fusion. We follow the main experimental settings but use smaller random subsets of $n = 200$ examples to reduce computational cost. Detailed results and protocols are reported in Appendix B.

6.1 COMPONENT ABLATIONS

Video QA. Table 3 examines two components in video QA: internal confidence and evidence support. Combining both yields higher AUROC than either component alone in all six settings, indicating that evidence support complements internal confidence in distinguishing correct from incorrect predictions. The combined score also achieves the lowest or tied-lowest ECE in three settings, while its ECE falls between those of the individual components in the remaining settings. This suggests a calibration trade-off, with model-specific recalibration on held-out data left to future work.

Table 3: Component ablations for vision-language models, evaluated on a fixed random $n = 200$ diagnostic subset for each benchmark–model cell.

			Video-MME		MVBench	
Model	Ablation	Variant	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓
Qwen2.5-VL-3B	Full	EVICONF	82.9	8.1	77.9	5.4
	(a) Component	Internal Confidence c_{int}	78.9	18.6	76.5	8.1
		Evidence Support c^{V}_{evid}	82.2	24.9	74.9	9.8
Qwen2.5-VL-7B	Full	EVICONF	82.0	11.2	79.9	7.2
	(a) Component	Internal Confidence c_{int}	75.2	4.4	76.3	7.2
		Evidence Support c^{V}_{evid}	81.1	16.3	79.7	13.3
Qwen3-VL-32B	Full	EVICONF	86.3	16.4	85.3	10.3
	(a) Component	Internal Confidence c_{int}	77.7	20.5	71.8	15.1
		Evidence Support c^{V}_{evid}	78.4	9.7	85.1	7.9

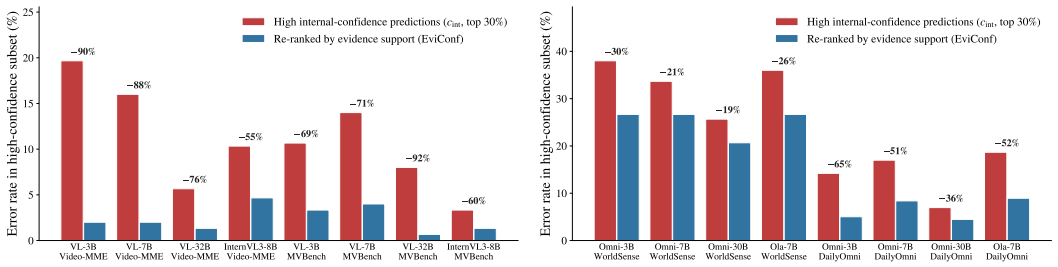
Table 4: Component and evidence-source ablations for omni-modal models, evaluated on a fixed random $n = 200$ diagnostic subset for each benchmark–model cell.

			WorldSense		DailyOmni	
Model	Ablation	Variant	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓
Qwen2.5-Omni-3B	Full	EVICONF	64.4	17.0	74.3	9.1
	(a) Component	Internal Confidence c_{int}	56.4	16.6	73.6	11.5
		Evidence Support c^{AV}_{evid}	60.4	29.4	67.9	7.1
	(b) Source	EVICONF (c^{A}_{evid})	59.6	20.6	74.9	5.1
		EVICONF (c^{V}_{evid})	57.0	19.0	72.7	8.8
Qwen2.5-Omni-7B	Full	EVICONF	63.5	17.8	74.3	6.9
	(a) Component	Internal Confidence c_{int}	60.7	17.4	70.1	7.0
		Evidence Support c^{AV}_{evid}	62.4	25.9	68.1	7.8
	(b) Source	EVICONF (c^{A}_{evid})	63.2	21.1	72.5	7.8
		EVICONF (c^{V}_{evid})	59.6	19.2	68.2	2.1
Qwen3-Omni-30B	Full	EVICONF	67.2	18.6	83.7	8.6
	(a) Component	Internal Confidence c_{int}	66.0	24.2	75.9	9.0
		Evidence Support c^{AV}_{evid}	63.9	24.1	68.5	9.6
	(b) Source	EVICONF (c^{A}_{evid})	69.0	24.4	75.0	6.9
		EVICONF (c^{V}_{evid})	65.0	21.3	69.7	5.2

Audio-Visual QA. *Component ablation.* We compare the full score with internal confidence alone and evidence support alone under joint audio-visual inputs. Combining the two components consistently improves AUROC, confirming that their complementarity extends to audio-visual QA. ECE follows the same pattern as in video QA, with the combined score either matching or improving on both components or falling between them. *Evidence intervention.* Both single-source interventions yield informative scores based on evidence support, but the audio-conditioned variant has higher AUROC than the video-conditioned variant in all six settings. This suggests stronger correctness discrimination from audio evidence on these benchmarks, where many questions concern acoustic events. Nevertheless, the variant conditioned on joint audio-visual evidence is strongest in four settings and remains close to the best single-source result in the other two. Conditioning on both modalities provides a balanced default that does not commit the estimator to one evidence source.

6.2 CONFIDENT-ERROR REDUCTION

Figure 5 further isolates the effect of adding evidence support to internal confidence at a fixed 30% coverage. For each benchmark–model cell, we compare the error rate among the top 30% of predictions ranked by internal confidence with that among the top 30% ranked by the complete



(a) Video QA (b) Audio-visual QA

Figure 5: **Adding evidence support reduces confident errors.** Each pair compares error rates at 30% coverage under rankings by internal confidence c_{int} and the complete EVICONF score. Annotations report the relative reductions.

Table 5: Temporal-module contribution: $\Delta\text{AUROC} = \text{Full} - \text{w/o Temporal}$, in percentage points. Each benchmark–model cell reuses the preceding 200-example ablation subset; temporal/static counts depend on its question-type composition. Positive values favor Full; Avg. is the unweighted mean of the 12 cell-wise differences. Absolute AUROC and model identities are given in Table 7.

	Video-MME			MVBench			WorldSense			DailyOmni			
Subset	3B	7B	32B	3B	7B	32B	3B	7B	30B	3B	7B	30B	Avg.
Overall ($n = 200$)	+0.8	+1.3	+3.3	+0.4	+1.3	+5.4	+0.3	-0.5	+0.8	+0.4	+0.5	-2.7	+0.9
Temporal	+1.2	+1.7	+7.4	+0.7	+1.0	+5.6	+0.9	+0.3	+1.0	+0.6	-0.2	-0.3	+1.7
Static	+0.6	+0.9	+1.4	-6.7	+2.2	+3.0	-0.9	-0.6	+0.8	+0.6	+0.4	-6.1	-0.4

EVICONF score. Adding evidence support reduces the confident-error rate in all 16 settings, by 55–92% in video QA and 19–65% in audio-visual QA. These results show that evidence weighting in EVICONF effectively mitigates false confidence.

6.3 TEMPORAL MODULE ABLATION

Table 5 isolates the contribution of temporal support by removing it while keeping all other components unchanged. Temporal questions benefit most, with AUROC improving in 10 of 12 settings by an average of 1.7 percentage points. This suggests that temporal support provides more reliable confidence weighting for questions requiring temporal reasoning. For static questions that do not require temporal information, the average change is small (-0.4 points), although individual settings vary. This question-type breakdown shows that the gains are concentrated on temporal questions, with limited average impact on static questions. Overall AUROC improves in 10 of 12 settings, with a mean gain of 0.9 points, indicating a modest overall benefit beyond content support. Absolute scores are reported in Appendix B.3, Table 7.

7 CONCLUSION

We showed that confidence estimators based on output probabilities or response consistency can assign unreliable confidence driven by language priors in video-based tasks. To address this problem, we proposed EVICONF, which reweights internal confidence using content and temporal evidence support from multimodal inputs. Content support is measured by the selected answer's probability lift from question-only to full input, while temporal support is measured by the fraction of temporal shuffles that lower the same answer's probability. These signals form an evidence score that is combined with normalized predictive-entropy confidence through a geometric mean. Experiments across eight models and four benchmarks covering video QA and audio-visual QA demonstrate state-of-the-art correctness AUROC, while ablations show confident-error reductions of up to 92% relative to internal confidence alone. These results highlight the value of grounding confidence in multimodal evidence to distinguish reliable predictions from confident errors.

AI USE STATEMENT

Generative AI tools were used to assist with paper organization, language polishing, and literature retrieval and discovery. The authors reviewed and verified the resulting text, citations, and technical claims and take full responsibility for the paper's content.

REPRODUCIBILITY STATEMENT

Section 5 summarizes the benchmarks, model families, baselines, and evaluation metrics. The appendix provides the complete diagnostic, method, experiment, and ablation protocols.

REFERENCES

Remi Cadene, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. RUBi: Reducing unimodal biases for visual question answering. In *Advances in Neural Information Processing Systems*, 2019.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' internal states retain the power of hallucination detection. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0d1986a61e30e5fa408c81216a616e20-Abstract-Conference.html.

Kaiqi Chen, Yang Qin, Changhao He, Xi Peng, and Peng Hu. DOUBT: Decoupled object-level understanding and bridging via vMF-based trustworthiness for hallucination detection in MLLMs. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=QMf9CBpEIf.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. doi: 10.1038/s41586-024-07421-0. URL https://www.nature.com/articles/s41586-024-07421-0.

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

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24108–24118, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.html.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. *arXiv preprint arXiv:2311.08298*, 2024. URL https://arxiv.org/abs/2311.08298.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 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*, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. WorldSense: Evaluating real-world omnimodal understanding for multimodal LLMs. *arXiv preprint arXiv:2502.04326*, 2025. URL https://arxiv.org/abs/2502.04326.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021.

Reza Khanmohammadi, Erfan Miahi, Mehrsa Mardikoraem, Simerjot Kaur, Ivan Brugere, Charese H. Smiley, Kundan Thind, and Mohammad M. Ghassemi. Calibrating LLM confidence by probing perturbed representation stability. *arXiv preprint arXiv:2505.21772*, 2025. URL https://arxiv.org/abs/2505.21772.

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

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Li_MVBench_A_Comprehensive_Multi-modal_Video_Understanding_Benchmark_CVPR_2024_paper.html.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Dang Nguyen, Ali Payani, and Baharan Mirzasoleiman. Beyond semantic entropy: Boosting LLM uncertainty quantification with pairwise semantic similarity. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4530–4540. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.234. URL https://aclanthology.org/2025.findings-acl.234/.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0283. URL https://papers.nips.cc/paper_files/paper/2024/hash/10c456d2160517581a234dfde15a7505-Abstract-Conference.html.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual VQA: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Seongheon Park, Changdae Oh, Hyeong Kyu Choi, Sean Du, and Sharon Li. VAUQ: Vision-aware uncertainty quantification for LVLM self-evaluation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 26534–26550, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.1321. URL https://aclanthology.org/2026.findings-acl.1321/.

Bhaktipriya Radharapu, Eshika Saxena, Kenneth Li, Chenxi Whitehouse, Adina Williams, and Nicola Cancedda. Calibrating LLM judges: Linear probes for fast and reliable uncertainty estimation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 193–216, San Diego, California, USA, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-industry.14. URL https://aclanthology.org/2026.acl-industry.14/.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Ruiyang Zhang, Hu Zhang, and Zhedong Zheng. VL-Uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. *arXiv preprint arXiv:2411.11919*, 2024. URL https://arxiv.org/abs/2411.11919.

Ziwei Zhou, Rui Wang, and Zuxuan Wu. Daily-Omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv preprint arXiv:2505.17862*, 2025. URL https://arxiv.org/abs/2505.17862.

Jinguo Zhu et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. URL https://arxiv.org/abs/2504.10479.

A Additional Details for Confidence Needs Evidence

A.1 Diagnostic Experimental Setup

The motivation study uses fixed 1,000-example evaluation pools from Video-MME and WorldSense, representing video QA and audio-visual QA, respectively. We evaluate Qwen2.5-VL-3B, Qwen2.5-VL-7B, and Qwen3-VL-32B on Video-MME, and Qwen2.5-Omni-3B, Qwen2.5-Omni-7B, and Qwen3-Omni-30B-A3B on WorldSense. Within each benchmark, every diagnostic uses the same examples and model predictions; no samples are selected after observing correctness or confidence.

Figure 2a computes the absolute Spearman rank correlation among MSP, PE-derived confidence, SC, and ACE-derived confidence separately within each benchmark–model cell, without pooling models or benchmarks. Figure 2b uses the prespecified threshold $c_{\text{int}} \geq 0.8$ and reports both the fraction of all predictions above the threshold and the correct–incorrect composition within that subset. Figure 2c retains the same high-confidence subset and stratifies it by the raw signed lift $\ell^{\mathcal{M}} \leq 0$, $\ell^{\mathcal{M}} \in (0, 0.25]$, $\ell^{\mathcal{M}} \in (0.25, 0.5]$, and $\ell^{\mathcal{M}} > 0.5$. This is the content lift defined in Section 4, before temporal modulation and the fixed mapping to $c^{\mathcal{M}}_{\text{evid}}$; the diagnostic threshold and bin boundaries are fixed across model scales.

MSP and PE represent probability-based confidence, while SC and ACE are canonical rollout-consistency scores computed from the same stochastic generations. Perturbation-augmented rollout estimators such as VL-Uncertainty and object-aware rollout estimators such as DOUBT are evaluated in the main benchmark comparison; the redundancy diagnostic isolates the underlying probability and rollout-consistency signal families without introducing estimator-specific intervention procedures.

A.2 Additional Diagnostic Analyses

Redundancy among output-only estimators. Table 6 summarizes the six unique off-diagonal correlations in each benchmark–model matrix from Figure 2a. The standard deviation is computed across estimator pairs within each cell and describes pairwise variation rather than bootstrap uncertainty. The mean absolute rank correlation ranges from 0.847 to 0.910 across all six cells, confirming that output-signal redundancy persists across model scales and task families. The larger variation for Video-MME with Qwen3-VL-32B arises because the within-family correlations for the probability and rollout pairs remain nearly perfect, whereas the correlations across these two families are weaker.

B Additional Details for Ablation Study

B.1 Evaluation Protocol

The component, evidence-source, fusion-rule, and additional-signal ablations reuse a fixed 200-example subset from each corresponding main-evaluation pool. Both task families use fixed random subsets: Video-MME and MVBench for video QA, and WorldSense and DailyOmni for audio-visual QA, as specified in Tables 3 and 4. The fixed-coverage analysis instead uses the complete main-evaluation pools and therefore covers all 16 benchmark–model settings. Within each setting, predictions are ranked separately by c_{int} and the complete EviConf score; the top 30% under each ranking are retained, and their error rates are compared. The reported relative reduction is computed with the internal-confidence error rate as the reference. Correctness labels, score orientation, AUROC, and 10-bin ECE follow the main evaluation protocol.

B.2 Ablation Variants

The component ablation compares the full score $\sqrt{c_{\text{int}} c^{\mathcal{M}}_{\text{evid}}}$ with c_{int} and $c^{\mathcal{M}}_{\text{evid}}$ alone. The omni-modal evidence ablation replaces c^{AV}_{evid} with c^{A}_{evid} or c^{V}_{evid} while retaining the same internal-confidence factor. The fusion-rule analysis replaces the geometric mean with the harmonic mean, minimum, or arithmetic mean without tuning any parameters. The additional-signal ablation extends the two-factor geometric mean with MSP or SC:

$$c_{3\text{way}}(u) = \left(c_{\text{int}} c^{\mathcal{M}}_{\text{evid}} u\right)^{1/3}, \qquad u \in \{c_{\text{MSP}}, c_{\text{SC}}\}. \tag{4}$$

Table 6: Summary of redundancy among output-derived confidence estimators. Each entry is the mean ± sample standard deviation across the six unique pairwise absolute Spearman correlations among MSP, PE, SC, and ACE within one benchmark–model cell.

Benchmark	Model	Pairwise $\lvert\rho\rvert$ (mean ± std.)
Video-MME	Qwen2.5-VL-3B	0.891 ± 0.064
Video-MME	Qwen2.5-VL-7B	0.906 ± 0.055
Video-MME	Qwen3-VL-32B	0.847 ± 0.116
WorldSense	Qwen2.5-Omni-3B	0.895 ± 0.057
WorldSense	Qwen2.5-Omni-7B	0.892 ± 0.062
WorldSense	Qwen3-Omni-30B	0.910 ± 0.062

Table 7: Overall AUROC for the temporal-module ablation summarized in Table 5. Each benchmark–model cell reuses the same 200 examples as the preceding ablations, shared across both variants; temporal/static counts depend on the subset composition. Full includes temporal support; w/o Temporal removes only that module, retaining internal confidence and the full-versus-blind content support. Entries report AUROC in percent. Avg. is the unweighted mean over the 12 benchmark–model cells for each variant. Video-MME/MVBench use Qwen2.5-VL-3B/7B and Qwen3-VL-32B; WorldSense/DailyOmni use Qwen2.5-Omni-3B/7B and Qwen3-Omni-30B.

	Video-MME			MVBench			WorldSense			DailyOmni			
Variant	3B	7B	32B	3B	7B	32B	3B	7B	30B	3B	7B	30B	Avg.
Overall ($n = 200$)													
w/o Temporal	82.1	80.7	83.0	77.5	78.6	79.9	64.1	64.0	66.4	73.9	73.8	86.4	75.9
Full	82.9	82.0	86.3	77.9	79.9	85.3	64.4	63.5	67.2	74.3	74.3	83.7	76.8

The video QA study omits the modality-replacement family because video evidence is its only non-textual source.

B.3 Temporal Module Ablation

Table 7 provides the overall absolute scores corresponding to the overall differences in Table 5. This analysis reuses the same fixed random 200-example subset for each benchmark–model cell as the preceding ablations, with identical examples across the two temporal variants. Temporal and static groups are formed within these subsets; their sizes depend on the question-type composition rather than a balanced resampling. The ablated variant retains internal confidence and the full-versus-blind content support; only temporal support is removed by setting $\tau^{\mathcal{M}} = 1$. The temporal/static counts are 66/134 for Video-MME, 164/36 for MVBench, 69/131 for WorldSense, and 67/133 for DailyOmni, shared across the three models within each benchmark. Overall Full scores match the original component-ablation results in Tables 3 and 4; overall w/o Temporal scores are computed by subtracting the corresponding differences in Table 5 from Full.

B.4 Fusion and Additional-Signal Sensitivity

Tables 8 and 9 test whether the gains in video QA and audio-visual QA depend on the fusion operator, thereby separating the contribution of evidence support from the choice of geometric-mean fusion. Module (a) compares the geometric mean used by EviConf with three parameter-free fusion rules. Geometric fusion is within 0.6 AUROC points of the best rule in every video-QA cell and within 0.3 points in every audio-visual-QA cell. The leading rule varies across cells, and ECE shows no consistent preference. We therefore retain the geometric mean as a simple, symmetric fusion rule that was not selected by tuning on these diagnostic results.

Module (b) adds MSP or self-consistency as a third factor to the two-factor EviConf reference shown in module (a). Neither signal yields a consistent AUROC improvement. The additions reduce AUROC in five of six video QA cells and all six audio-visual QA cells; the only improvement is the 0.3-point gain from adding self-consistency on MVBench with Qwen2.5-VL-7B, accompanied by higher ECE.

Table 8: Fusion-rule and additional-signal ablations for vision-language models on fixed $n = 200$ diagnostic subsets. Module (a) compares parameter-free fusion rules; module (b) adds a third output-derived signal to the EVICONF reference in module (a).

			Video-MME		MVBench	
Model	Module	Variant	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓
Qwen2.5-VL-3B	(a) Fusion	EVICONF (geometric mean)	82.9	8.1	77.9	5.4
		Harmonic mean	82.0	17.3	77.3	4.9
		Minimum	79.6	8.8	76.1	8.5
		Arithmetic mean	83.5	19.5	78.2	4.3
	(b) Additional	3-way (+MSP)	81.2	17.3	76.1	5.1
		3-way (+SC)	81.2	17.2	76.8	4.3
Qwen2.5-VL-7B	(a) Fusion	EVICONF (geometric mean)	82.0	11.2	79.9	7.2
		Harmonic mean	80.0	6.9	79.4	5.6
		Minimum	76.2	5.9	78.0	8.7
		Arithmetic mean	81.4	12.8	80.5	9.1
	(b) Additional	3-way (+MSP)	78.5	7.4	78.3	6.6
		3-way (+SC)	80.9	9.0	80.2	8.7
Qwen3-VL-32B	(a) Fusion	EVICONF (geometric mean)	86.3	16.4	85.3	10.3
		Harmonic mean	85.8	14.6	85.7	9.3
		Minimum	83.3	9.0	85.6	8.8
		Arithmetic mean	86.1	15.1	85.2	11.2
	(b) Additional	3-way (+MSP)	84.2	17.0	83.6	12.4
		3-way (+SC)	84.1	17.4	83.7	12.7

Table 9: Fusion-rule and additional-signal ablations for omni-modal models on fixed $n = 200$ diagnostic subsets. Module (a) compares parameter-free fusion rules; module (b) adds a third output-derived signal to the EVICONF reference in module (a).

			WorldSense		DailyOmni	
Model	Module	Variant	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓
Qwen2.5-Omni-3B	(a) Fusion	EVICONF (geometric mean)	64.4	17.0	74.3	9.1
		Harmonic mean	60.1	16.0	74.3	9.1
		Minimum	58.8	11.4	73.2	15.0
		Arithmetic mean	60.3	18.0	74.0	7.5
	(b) Additional	3-way (+MSP)	59.0	19.1	74.2	7.5
		3-way (+SC)	57.7	20.0	73.5	4.5
Qwen2.5-Omni-7B	(a) Fusion	EVICONF (geometric mean)	63.5	17.8	74.3	6.9
		Harmonic mean	63.6	15.6	73.8	5.8
		Minimum	63.5	8.5	71.4	9.5
		Arithmetic mean	63.4	18.8	74.6	6.1
	(b) Additional	3-way (+MSP)	63.0	19.6	72.7	5.4
		3-way (+SC)	62.0	19.6	72.3	3.8
Qwen3-Omni-30B	(a) Fusion	EVICONF (geometric mean)	67.2	18.6	83.7	8.6
		Harmonic mean	65.3	17.5	83.2	9.2
		Minimum	62.9	13.0	78.5	15.8
		Arithmetic mean	66.7	19.8	84.0	9.0
	(b) Additional	3-way (+MSP)	66.8	21.9	81.8	5.1
		3-way (+SC)	67.1	22.8	82.0	5.4

This behavior is consistent with the signal redundancy observed in Section 3: another output-derived confidence signal changes the score scale but generally contributes little new information for ranking correctness.

C Additional Details for Method

C.1 Input Construction and Inference

The full condition contains the question, answer choices, and all task evidence. For video QA, the task evidence is the video without an audio input; for audio-visual QA, it is the native video with its audio track enabled. The blind condition retains only the textual question and answer choices. Prompts and all non-target textual inputs are held fixed across the two conditions.

The primary EviConf score uses six forward passes per example: one with all task evidence, one with the blind text-only input, and $K = 4$ with temporally shuffled evidence as defined in Section 4. Additional audio-only and video-only passes are required only for the omni-modal diagnostic ablations and are not part of the deployed score. Option probabilities are read from the first decoded token after a fixed JSON scaffold ({`"answer": "`) that teacher-forces the next token onto the answer letter. For each answer letter we sum the probability mass of its case- and whitespace-variant token ids and renormalize over the option set, giving a proper distribution $\boldsymbol{p}$ over $\mathcal{Y}$; the predicted letter $\hat{y}$ is the argmax of the full-input option distribution, as in Equation 1, and is held fixed when reading its blind-input and shuffled-input probabilities. The blind pass preserves the question, options, and prompting format while removing all frames and audio. All passes use the same tokenizer, option letters, and numerical precision, so $\ell^{\mathcal{M}} = p_{\hat{y}}^{\mathcal{M}} - p_{\hat{y}}^{\emptyset}$ compares two probabilities of the same answer under one changed input condition.

C.2 Modality-Specific Diagnostic

To analyze the contribution of each omni-modal channel, we define $\boldsymbol{p}^A = p(\cdot \mid q, a)$ and $\boldsymbol{p}^V = p(\cdot \mid q, v)$ and compute

$$\ell^A = p_{\hat{y}}^{AV} - p_{\hat{y}}^{V}, \qquad \ell^V = p_{\hat{y}}^{AV} - p_{\hat{y}}^{A}, \qquad \ell^{AV} = p_{\hat{y}}^{AV} - p_{\hat{y}}^{\emptyset}. \tag{5}$$

The superscript denotes the evidence source being measured: A for audio, V for video evidence, and AV for their joint support. Audio is evaluated relative to the video-only condition and video evidence relative to the audio-only condition, so the non-target source is held fixed. For evidence-source ablations, each content lift replaces the joint content term while retaining the full-input temporal factor: $c_{\text{evid}}^{S} = (1 + \ell^{S}\tau^{AV})/2$, where $S \in \{A, V, AV\}$. These quantities diagnose which source changes support for the fixed full-input prediction; they do not alter Equation 3.

C.3 Statistical Protocol

All EviConf scores are computed independently per example. Test-set ranks, z-scores, min–max statistics, fitted combinations, and correctness labels are excluded. The fixed mapping for evidence support uses only the theoretical range of the signed lift and requires no dataset-level statistics or learned weights.

D Additional Details for Experiments

D.1 Benchmarks

For video QA, we use independent fixed random samples of 1,000 examples from Video-MME and 1,000 examples from MVBench. For audio-visual QA, we use a fixed random sample of 1,000 examples from WorldSense, which covers 26 real-world task types, and the complete 1,197-example DailyOmni evaluation set. The three 1,000-example subsets are drawn once by random sampling with seed 42 and held fixed across all models and confidence estimators; DailyOmni uses the complete evaluation set. Results are computed separately for each benchmark–model cell. Each question is presented in its native multiple-choice form (C ranges up to eight options on MVBench and is four on the other benchmarks), and the gold letter is the option whose text matches the released answer. All methods share the same prompt, option ordering, output constraint, and answer parser within a cell.

D.2 Models and Inference Configuration

The vision-language evaluation uses Qwen2.5-VL-3B/7B-Instruct, Qwen3-VL-32B-Instruct, and InternVL3-8B. The omni-modal evaluation uses Qwen2.5-Omni-3B, Qwen2.5-Omni-7B, Qwen3-Omni-30B-A3B-Instruct, and Ola-7B. Qwen2.5-VL-3B/7B run under vLLM; Qwen3-VL-32B and InternVL3-8B run under Hugging Face Transformers with automatic device mapping; all omni-modal models run under Transformers in `bfloat16` with SDPA attention and the speech decoder disabled (text-only output). The 32B model is sharded across two GPUs and each smaller model occupies a single GPU; Ola-7B runs in an isolated environment. All models are used at their released checkpoints without fine-tuning.

Frame sampling follows two fixed regimes. For video QA, we uniformly sample 32 frames per video at evenly spaced bin midpoints, with a per-frame budget in $[3{,}136,\ 50{,}176]$ pixels and no subtitles (video-only). For audio-visual QA, we sample video at 2 frames per second capped at 120 frames with a per-frame budget in $[12{,}544,\ 50{,}176]$ pixels, preserve the native audio track at 16 kHz mono, and jointly provide temporally aligned audio and visual inputs to preserve synchronization. For each example, the full pass reads the option distribution at the answer position and determines $\hat{y}$; the blind pass reads the probability of that same answer with all frames and audio removed. Four additional shuffled passes estimate $\tau^{\mathcal{M}}$ using the same fixed answer, for six passes in total; the score requires no sampled answer generations. The rollout-based baselines use the generation settings specified in the baseline implementation subsection. The random seed is fixed to 42 for subset construction and all sampling-based baselines.

D.3 Blind-Control Validation Protocol

The controlled validation in Figure 4 uses Qwen2.5-VL-7B for Video-MME and MVBench and Qwen2.5-Omni-7B for WorldSense and DailyOmni. Each benchmark contributes a fixed stratified subset of $n = 200$ examples selected with seed 42, using the corresponding full-input predictions from the main evaluation. For every example, we fix $\hat{y}$ to the full-input prediction from the main evaluation and preserve the question, answer choices, option ordering, frame count, spatial resolution, and timestamp tags across applicable conditions. Consequently, every comparison reads the probability of the same answer under an input intervention rather than allowing the prediction target to change.

All reference conditions are constructed from the same decoded input. One forward pass per condition yields the option distribution and the probability $p_{\mathrm{ref}}(\hat{y})$ of the fixed answer. Probabilities are the renormalized first-token probabilities over the canonical option letters. For omni-modal examples, frame interventions are paired with the corresponding audio intervention: black frames use silence, noise frames use RMS-matched Gaussian noise, and mismatched inputs use the other example's audio track. The independently reconstructed full condition uses the same task evidence and inference route as the reconstructed reference conditions. Its numerical agreement with the full-input probabilities from the main evaluation is quantified in Appendix D.4.

The reference conditions are: (i) *blind*, containing only the question and options; (ii) *format null*, containing all-zero frames with the same count and size; (iii) *noise null*, containing uniform-noise frames and, for omni-modal inputs, loudness-matched Gaussian audio; and (iv) *mismatch*, containing another example's real multimodal input. Mismatches are fixed-point-free derangements within task type and option count, repeated with three seeds and averaged, so the control remains structurally valid while being answer-irrelevant.

Panels (a) and (b) report sample means of probability differences. Panel (a) compares $p_{\mathrm{full}}(\hat{y}) - p_{\mathrm{blind}}(\hat{y})$ with $p_{\mathrm{mismatch}}(\hat{y}) - p_{\mathrm{blind}}(\hat{y})$. Panel (b) reports $p_{\mathrm{full}}(\hat{y}) - p_{\mathrm{ref}}(\hat{y})$ for the blind, format-null, and noise-null references. Panel (c) uses the deployed blind lift $p_{\mathrm{full}}(\hat{y}) - p_{\mathrm{blind}}(\hat{y})$ as a correctness score after splitting each benchmark–model cell at the median $p_{\mathrm{blind}}(\hat{y})$ into strong- and weak-text-prior groups of approximately 100 examples each. AUROC is computed separately within each group; the split is diagnostic and is not used by the deployed estimator.

D.4 Sample-Level Mismatch-Lift Analysis

The near-zero mean mismatch lift in Figure 4a may conceal substantial probability shifts that cancel across examples. We therefore examine whether matched task evidence supports the fixed full-input prediction $\hat{y}$ more consistently than mismatched evidence at the sample level. The analysis uses

Table 10: Sample-level mismatch analysis on four blind-control cells ($n = 200$ each), reporting mismatch-lift magnitude, paired specificity, and correctness AUROC with 95% confidence intervals.

		Mismatch lift $\bar{r}_i$			Specificity gap d_i		Correctness AUROC	
Benchmark	Model	mean	mean|·|	5–95%	$P(d_i{>}0)$	mean [95% CI]	matched / mism.	Δ [95% CI]
Video-MME	VL-7B	-0.05	0.17	$[-0.47, 0.30]$	0.84	0.31 [0.27, 0.36]	0.68/0.47	$+0.21$ [0.13, 0.30]
MVBench	VL-7B	$+0.04$	0.23	$[-0.39, 0.67]$	0.91	0.29 [0.25, 0.34]	0.63/0.54	$+0.09$ [0.01, 0.17]
WorldSense	Omni-7B	-0.04	0.19	$[-0.43, 0.38]$	0.82	0.31 [0.26, 0.36]	0.61/0.47	$+0.14$ [0.05, 0.23]
DailyOmni	Omni-7B	$+0.01$	0.15	$[-0.34, 0.33]$	0.82	0.31 [0.26, 0.36]	0.70/0.60	$+0.10$ [0.02, 0.18]

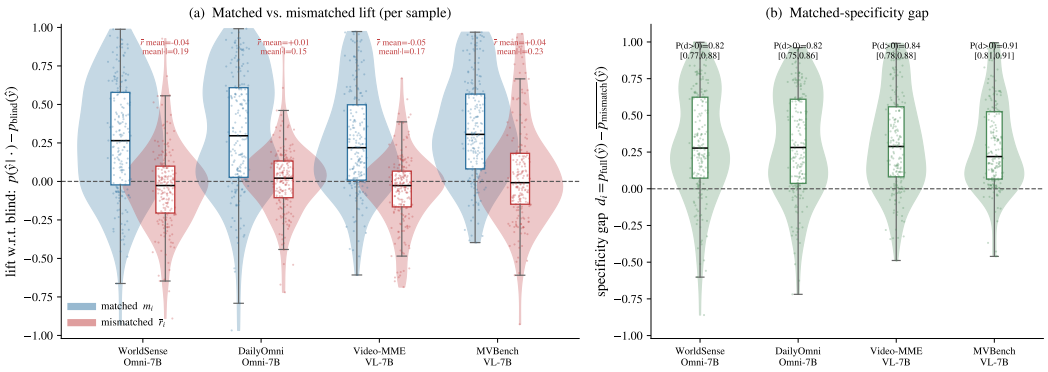


Figure 6: **Sample-level mismatch analysis. (a)** Distributions of matched lift m_i and mean mismatch lift $\bar{r}_i$. **(b)** Distribution of the paired specificity gap $d_i = m_i - \bar{r}_i$. Annotations report mismatch magnitude and $P(d_i > 0)$ with 95% confidence intervals.

the probabilities from the four blind-control settings in Appendix D.3 and requires no additional inference. For each example, we read $p_{\text{full}}(\hat{y})$, $p_{\text{blind}}(\hat{y})$, and the probabilities from three fixed derangements, $p_{\text{mismatch}(s)}(\hat{y})$. We compute the matched lift $m_i = p_{\text{full}}(\hat{y}) - p_{\text{blind}}(\hat{y})$, the mean mismatch lift $\bar{r}_i = \frac{1}{3}\sum_s p_{\text{mismatch}(s)}(\hat{y}) - p_{\text{blind}}(\hat{y})$, and the matched-specificity gap $d_i = m_i - \bar{r}_i$. For this analysis, $p_{\text{full}}(\hat{y})$ is read from the reconstructed full condition, and reference probabilities are obtained through the same reconstruction route; $\hat{y}$ remains the fixed full-input prediction from the main evaluation. The mean absolute difference between the reconstructed full-input probabilities and those from the main evaluation is 0.007 for video QA and 0.12–0.13 for audio-visual QA. Confidence intervals for the gap use a hierarchical paired bootstrap with 10,000 replicates, resampling examples in the outer loop and the three mismatch seeds within each example in the inner loop; AUROC intervals use paired resampling of examples.

Table 10 and Figure 6 show that the mismatch lift has a small mean (-0.05 to $+0.04$) but a substantially larger mean absolute value (0.15–0.23) and wide 5–95% ranges. Its near-zero average therefore reflects cancellation between positive and negative shifts rather than an absence of sample-level effects. Matched task evidence nevertheless provides substantially more consistent support: d_i is positive for 82–91% of examples, its mean ranges from 0.29 to 0.31, and its 95% confidence interval excludes zero in every cell. The matched lift also predicts correctness better than the mismatch lift in all four cells, with paired ΔAUROC values from $+0.09$ to $+0.21$ and confidence intervals above zero. In contrast, mismatch-lift AUROC ranges from 0.47 to 0.60. Thus, mismatched evidence has directionally inconsistent and weakly predictive effects, whereas matched evidence provides consistent support for the selected answer.

D.5 Baseline Implementations

Let $\boldsymbol{p}^{\mathcal{M}}$ be the full-input distribution over C answer choices. The output-based baseline scores are defined as

$$c_{\text{MSP}} = \max_y p_y^{\mathcal{M}}, \qquad c_{\text{PE}} = 1 - \frac{H(\boldsymbol{p}^{\mathcal{M}})}{\log C}, \qquad c_{\text{SC}} = \frac{1}{R}\sum_{r=1}^{R} \mathbb{I}[y_r = \tilde{y}], \qquad c_{\text{ACE}} = 1 - \frac{H(\boldsymbol{q})}{\log C}, \tag{6}$$

where $\tilde{y}$ is the modal answer among $R = 10$ stochastic generations and $q_k = R^{-1} \sum_{r=1}^{R} \mathbb{I}[y_r = k]$ is the empirical distribution over canonical option identifiers. SC and ACE use the same generations, decoding parameters, and answer parser. ACE does not perform free-form semantic clustering because the multiple-choice task already supplies fixed answer classes. The $R = 10$ generations are sampled at temperature 0.8 (top-p 0.95) with at most 32 new tokens for vision-language models and at temperature 1.0 (top-p 0.95) for omni-modal models, using the same fixed seed as the other experiments.

VL-Uncertainty perturbs the visual input and semantically preserves the textual query, clusters the resulting answers by semantic content, and uses the entropy of the cluster distribution as uncertainty; we report its normalized confidence orientation. Following the original method it uses 5 paired perturbations—Gaussian blur on the frames at radii $\{0.6, 0.8, 1.0, 1.2, 1.4\}$ paired with five temperature-varied question rephrasings—and samples one answer per perturbed pair, clustering by option letter for the multiple-choice task. Probe is a supervised white-box reference adapted from the LLM-judge setting and is reproduced as detailed below. DOUBT follows its object-aware response elicitation and directional-concentration confidence score: a greedy pass first lists up to five salient objects/events, then 10 direct and 10 object-bridged answers are sampled at temperature 0.5 (top-k 10, top-p 0.99), embedded with an `nli-roberta-large` sentence encoder, and scored by the debiased von Mises–Fisher mean-resultant length averaged over the direct and bridged sets. The main tables report total per-example model calls: 1 for MSP and PE, 10 for SC and ACE, 5 for VL-Uncertainty, 1 for Probe, 21 for DOUBT (1 object pass + 20 sampled answers), and 6 for EVICONF. The single inference call reported for Probe excludes supervised offline training.

Supervised Probe reproduction. For each evaluated model, we train a separate affine probe from its hidden dimension to one scalar, followed by a sigmoid: $\hat{p} = \sigma(\boldsymbol{w}^\top \boldsymbol{h} + b)$. The feature $\boldsymbol{h}$ is the last-token hidden state from the middle transformer layer, indexed by integer division of the number of layers by two. The layer indices are 18, 14, 32, and 14 for Qwen2.5-VL-3B, Qwen2.5-VL-7B, Qwen3-VL-32B, and InternVL3-8B, respectively; and 18, 14, 24, and 14 for Qwen2.5-Omni-3B, Qwen2.5-Omni-7B, Qwen3-Omni-30B, and Ola-7B, respectively. Features are standardized coordinate-wise using the training-set mean and standard deviation, which are then reused unchanged at evaluation time. The target is answer correctness $y \in \{0, 1\}$, and the probe minimizes Brier loss $(\hat{p} - y)^2$ using Adam with learning rate 10^{-4}, weight decay 0.01, batch size 4, and 10 epochs. We apply no post-hoc calibration.

For each vision-language model, we pool all non-test examples from the two benchmarks: 1,700 from Video-MME and 2,000 from MVBench, for 3,700 training examples total. One model-specific probe is trained on this pool and evaluated separately on the fixed Video-MME-1,000 and MVBench-1,000 test sets. For each omni-modal model, the training set contains the same fixed 2,000 WorldSense examples excluded from the random 1,000-example WorldSense test set. The resulting model-specific probe is evaluated in-domain on WorldSense-1,000 and transferred without refitting to the complete DailyOmni-1,197 set, which has no probe-training pool and is therefore an out-of-domain evaluation. No reported test example or test correctness label is used for feature standardization, probe fitting, model selection, or calibration.

D.6 EVALUATION METRICS

For example i, let $z_i = \mathbb{I}[\hat{y}_i = y_i^*]$ denote correctness and $s_i \in [0, 1]$ its confidence score, oriented so that larger values predict correctness. AUROC is the probability that a randomly selected correct example receives a higher score than a randomly selected incorrect example, with half credit for ties:

$$\mathrm{AUROC} = \Pr(s^+ > s^-) + \tfrac{1}{2} \Pr(s^+ = s^-). \tag{7}$$

For calibration, we partition $[0, 1]$ into $B = 10$ equal-width bins and let I_b denote the set of example indices whose confidence scores fall in bin b. We compute ECE over the nonempty bins:

$$\mathrm{ECE} = \sum_{b:\, |I_b| > 0} \frac{|I_b|}{N} \left| \frac{1}{|I_b|} \sum_{i \in I_b} z_i - \frac{1}{|I_b|} \sum_{i \in I_b} s_i \right|. \tag{8}$$

AUROC and ECE are multiplied by 100 in the main-result and ablation tables. Table 10 instead reports AUROC, AUROC differences, and probability statistics on the 0–1 scale (signed for differences). Higher AUROC and lower ECE are better. AUROC is the primary evaluation criterion because

the estimator is designed to discriminate correct from incorrect predictions; ECE is a secondary diagnostic, and the score is not assumed to be a calibrated probability by construction. Scores are evaluated without test-set normalization, fitted combinations, or test-label tuning; boldface marks the best populated value within each benchmark–model–metric cell.