VISUAL CONTRAST FOR RELIABLE UNCERTAINTY IN LABEL-FREE VLM RANKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Label-free vision-language model (VLM) ranking commonly uses uncertainty as a proxy for model capability, yet this signal can become unreliable under overconfidence. We observe that predictions with strong dependence on either linguistic or visual information exhibit poorer calibration than those with more balanced source dependence. Our analysis further shows that strong support from a single information source can induce high confidence without guaranteeing prediction reliability, while visual degradation provides a tractable signal of relative visual support. Motivated by these findings, we propose Visual Contrast for Reliable Uncertainty (VCRU), a label-free method for ranking VLMs on multiple-choice question-answering tasks. VCRU compares first-token NLL and output entropy between original and visually degraded inputs. Opposite shifts in these signals define confidence conflicts, which serve as diagnostic patterns of potentially unreliable confidence. A drift gate further emphasizes high-confidence conflicts with substantial confidence changes, yielding a model-level penalty that adjusts the baseline uncertainty score. Experiments on 36 heterogeneous VLMs across four benchmarks show that VCRU consistently improves its uncertainty baselines across model families. These results suggest that visual contrast provides a useful complementary signal for confidence-based, label-free VLM ranking.

1 INTRODUCTION

Selecting an appropriate vision-language model (VLM) for a target task typically requires ground-truth labels to evaluate candidate models (Kembhavi et al., 2016; Lu et al., 2022; Liu et al., 2024b; Chen et al., 2024b). However, model performance can vary across tasks, domains, and deployment environments. Even when labels are available for one setting, a change in the target distribution may require new annotations and re-evaluation. This repeated labeling process can be costly as models and deployment conditions evolve. These challenges motivate *label-free model ranking*, which aims to recover the relative performance ordering of candidate VLMs using only unlabeled target data.

A common direction is to estimate model capability from uncertainty, mainly through consistency-based and logit-based signals. Consistency-based methods repeatedly sample responses to the same input and measure their agreement (Wang et al., 2023; Xiong et al., 2024), but require multiple stochastic generations and can incur substantial computational or API cost. Logit-based methods instead derive uncertainty directly from model outputs using token probability (Malinin & Gales, 2021), entropy (Kadavath et al., 2022; Kuhn et al., 2023), maximum softmax probability (Manakul et al., 2023), or negative log-likelihood (NLL) (Manakul et al., 2023; Tu et al., 2025; Aichberger et al., 2026; Buffa & Corro, 2026). These methods generally assume that lower uncertainty indicates stronger model capability. However, VLMs can be overconfident under poorly calibrated settings (Wang et al., 2024a; Liu et al., 2026; Li et al., 2026), making uncertainty less reliable for ranking. This raises a natural question: **when does confidence become unreliable for label-free VLM ranking, and how can such cases be identified without target labels?**

In this work, we investigate this question from the perspective of *information-source dependence*. We observe that predictions with strong reliance on either linguistic priors or visual information exhibit higher calibration error than those in intermediate reliance groups. This suggests that confidence can become less informative when a prediction is dominated by a single information source. Our theoretical analysis further shows that sufficiently strong support from one source can produce

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.

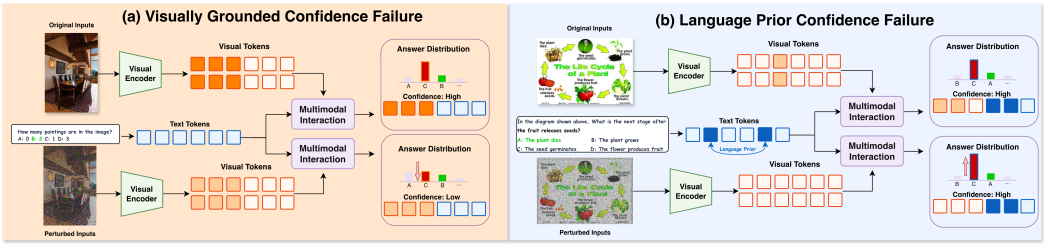

Figure 1: Confidence failure modes caused by excessive reliance on a single information source in VLM ranking. (a) Visually Grounded Confidence Failure: confidence is dominated by visual information, leading to overconfident predictions that do not reliably reflect multimodal capability. (b) Language Prior Confidence Failure: confidence is dominated by linguistic priors, leading to overconfident predictions without sufficient visual support.

high confidence without guaranteeing higher prediction reliability. Importantly, we do not assume that reliance on one source is inherently incorrect; rather, we study when such reliance makes confidence less reliable as a proxy for model capability. Based on this observation, we characterize two confidence-failure patterns, illustrated in Figure 1. **Visually Grounded Confidence Failure** describes cases in which the original prediction is strongly supported by visual information, such that weakening the image substantially reduces its confidence. **Language-Prior Confidence Failure** describes the complementary case in which weakening visual information increases support for the original prediction, suggesting stronger dependence on non-visual, including linguistic, information. These patterns motivate us to examine not only the magnitude of confidence, but also how confidence responds to controlled changes in visual evidence.

We therefore propose **Visual Contrast for Reliable Uncertainty (VCRU)**, a label-free approach for ranking VLMs on multiple-choice question-answering tasks. VCRU compares the NLL of the original first-token prediction and the entropy of the complete first-token distribution between the original and visually degraded inputs. When these two uncertainty signals move in opposite directions, we identify a *confidence conflict*. Specifically, an NLL increase accompanied by an entropy decrease defines **Conflict-V**, indicating that the original prediction loses support while competing predictions become more concentrated as visual evidence is weakened. Conversely, an NLL decrease accompanied by an entropy increase defines **Conflict-L**, indicating that the original prediction gains relative support despite increasing distribution-level uncertainty. These conflict patterns serve as diagnostic signals of potentially unreliable confidence rather than direct indicators of prediction error. To further distinguish meaningful confidence changes from negligible fluctuations, VCRU introduces a drift gate that emphasizes conflict samples with high original confidence and substantial confidence changes after visual degradation. These gated conflicts are aggregated into a model-level penalty that adjusts the baseline uncertainty score.

Taken together, our study highlights two confidence failure modes in label-free VLM ranking, Visually Grounded Confidence Failure and Language Prior Confidence Failure, and shows how visual contrast can be used to diagnose and correct potentially unreliable confidence. We evaluate VCRU on 36 heterogeneous VLMs across four benchmarks: MMStar (Chen et al., 2024b), MMBench (Liu et al., 2024b), AI2D (Kembhavi et al., 2016), and ScienceQA (Lu et al., 2022). Across diverse VLM architectures and benchmarks, VCRU consistently improves its underlying uncertainty baselines and outperforms existing consistency-based and logit-based ranking methods.

2 Reliable VLM Ranking via Visual Contrast

2.1 Problem Formulation

Target MCQA Dataset and Candidate VLMs. Let $\mathcal{D} = \{x_i\}_{i=1}^{N}$ denote a target multiple-choice question-answering (MCQA) dataset, where each sample is $x_i = (I_i, q_i, \mathcal{A}_i)$. Here, I_i denotes the input image, q_i the question, and $\mathcal{A}_i = \{a_{i,c}\}_{c=1}^{C_i}$ the set of candidate answers. We consider a pool of M candidate VLMs, denoted by $\mathcal{F} = \{f_m\}_{m=1}^{M}$.

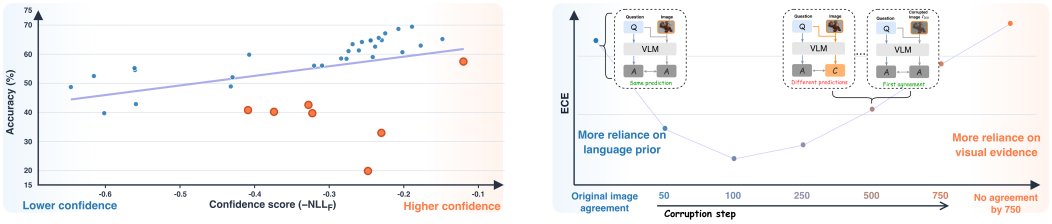

Figure 2: (a) Relationship between confidence measured by $-$NLL and model performance across VLMs. (b) Expected Calibration Error (ECE) across prediction groups with different degrees of reliance on visual and linguistic information. The groups are ordered by the visual-noise intensity required for the prediction to align with its text-only answer.

For a sample x_i, model f_m generates a token sequence $z_{m,i,1:K_{m,i}} = (z_{m,i,1}, \ldots, z_{m,i,K_{m,i}})$, where $K_{m,i}$ is the response length. At generation step j, the model defines the conditional next-token distribution $p_m(\cdot \mid x_i, z_{m,i,<j})$, where $z_{m,i,<j} = (z_{m,i,1}, \ldots, z_{m,i,j-1})$. The generated sequence is decoded into a textual response $r_{m,i} = \mathrm{Decode}(z_{m,i,1:K_{m,i}})$. Following VLMEvalKit (Duan et al., 2024), the response $r_{m,i}$ is further mapped to an option prediction $\hat{y}_{m,i} \in \mathcal{A}_i$ whenever a valid option can be identified. In our experiments, both model inference and answer evaluation are conducted using the standard VLMEvalKit pipeline.

Model Performance. Let $y_i \in \mathcal{A}_i$ denote the ground truth for sample x_i. The performance of candidate VLM f_m is measured by its accuracy $g_m = \frac{1}{N}\sum_{i=1}^{N} \phi(\hat{y}_{m,i}, y_i)$, where $\phi(\hat{y}_{m,i}, y_i) = \mathbb{I}[\hat{y}_{m,i} = y_i]$. A response that cannot be mapped to any candidate option is counted as incorrect. The resulting scores $\{g_m\}_{m=1}^{M}$ define the ground-truth ranking of VLMs on the target dataset.

Label-Free Ranking Objective. Our goal is to derive a label-free score $s_m = S(f_m, \mathcal{D})$ for each candidate VLM f_m, without access to the target labels $\{y_i\}_{i=1}^{N}$, such that the ranking induced by $\{s_m\}_{m=1}^{M}$ closely agrees with the ground-truth ranking induced by $\{g_m\}_{m=1}^{M}$. This allows candidate VLMs to be ranked and selected for the target MCQA dataset without using target annotations.

2.2 Motivation: When VLM Confidence Becomes Unreliable

Potential limitation of confidence-based ranking. Uncertainty-based VLM ranking commonly assumes that lower uncertainty indicates stronger model capability. While this trend generally holds, it can fail under overconfidence. As shown in Figure 2(a), confidence measured by $-$NLL is overall positively correlated with model accuracy, yet several VLMs remain highly confident despite lower accuracy. This suggests that confidence alone may not always reliably reflect model capability.

We hypothesize that one source of this mismatch is *single-source reliance*, where a prediction is supported predominantly by either linguistic information or visual evidence. Such reliance is not necessarily incorrect, but extreme dependence on one source may make confidence less reliable.

Empirical analysis of single-source reliance. To examine this hypothesis, we compare predictions under the original vision-language input and the corresponding text-only input. Samples with identical predictions form the *Original Image Agreement* group. For the remaining samples, we progressively degrade the image and record the minimum perturbation intensity required for the prediction to match the text-only prediction. Earlier agreement indicates weaker visual dependence, whereas later agreement indicates stronger visual influence. Samples that never match within the maximum perturbation level form the *No Agreement by 750* group.

As shown in Figure 2(b), the two extreme groups exhibit higher expected calibration error (ECE) than the intermediate groups, suggesting that extreme source dependence is associated with less reliable confidence. We next provide a theoretical explanation for how strong support from a single information source can produce high confidence without necessarily implying higher accuracy.

2.3 Why Single-Source Dominance Can Mislead Confidence

Notations. Let $|\mathcal{V}|$ denotes the possible tokens at the first decoding position, and $z = z_{m,i,1}$ refers to the first predicted token generated under the original language-image-pair input. For each token

k, let a_k and v_k denote the contributions from linguistic and visual information, respectively. Given the strength ρ of the available visual information under the original input, the corresponding logit is $s_k(\rho) = a_k + \rho v_k$.

Theorem 1. *(The connection between confidence and single-source reliance.) Suppose that one information source supports z over every competing token by at least γ, while any opposing contribution from the other source is bounded by B. Specifically, visual predominance corresponds to $\rho(v_z - v_k) \geq \gamma$ and $a_k - a_z \leq B$, whereas linguistic predominance corresponds to $a_z - a_k \geq \gamma$ and $\rho(v_k - v_z) \leq B$, for every $k \neq z$, where $\gamma > B$. Then, the confidence assigned to z satisfies*

$$p_z(\rho) \geq \frac{1}{1 + (|\mathcal{V}| - 1)\exp\big(-(\gamma - B)\big)}. \tag{1}$$

Theorem 1 shows that when evidence from one modality sufficiently outweighs competing evidence from the other modality, the confidence assigned to z increases as the margin $(\gamma - B)$ widens. However, higher confidence does not necessarily indicates higher accuracy, since it may reflect strong support from a predominant source rather than actual VLM capability. To further explain our observation that single-source reliance exacerbates VLM overconfidence, Theorem 2 establishes a lower bound on the confidence-accuracy gap for predictions supported by single-sources and reveals the resulting overconfidence.

Theorem 2. *(The connection between overconfidence and single-source reliance.) Let $\mathcal{G}$ denote a group of predictions satisfying these conditions, and $y^\star$ denotes the correct answer. If their accuracy is at most $r < 1$, then their average confidence-accuracy gap satisfies*

$$\begin{aligned} &\mathbb{E}[p_z(\rho) \mid \mathcal{G}] - \Pr(\hat{y} = y^\star \mid \mathcal{G}) \\ &\quad \geq \frac{1}{1 + (|\mathcal{V}| - 1)\exp\big(-(\gamma - B)\big)} - r, \end{aligned} \tag{2}$$

Theorem 2 shows that predictions supported predominantly by single information source are overconfident on average when their confidence lower bound exceeds their accuracy upper bound r due to a larger margin $(\gamma - B)$. This illustrates how strong support from a single source can produce confidence that exceeds prediction reliability, rather than indicating greater VLM capability. The proof of Theorem. 1 and Theorem. 2 defer to Appendix B.1.

Having established that high confidence driven by reliance on a single information source can result in overconfidence, we next seek approaches to identify whether a VLM prediction exhibits such single-source reliance. To this end, Theorem 3 examines whether confidence changes induced by visual degradation can reveal the relative contributions of visual evidence and linguistic information to the model's confidence.

Theorem 3. *(Connection between visual degradation and NLL changes.) Let ρ and $\rho' < \rho$ denote the strengths of the available visual information under the two inputs, respectively. Given the original input x_i and its visually distorted counterpart x_i', for $u \in [\rho', \rho]$, the token logits and corresponding output probabilities are $s_k(u) = a_k + uv_k$ and $p_k(u) = \frac{\exp(a_k + uv_k)}{\sum_{j \in \mathcal{V}} \exp(a_j + uv_j)}$. We keep the first token z fixed under visual distortion and define $\ell_z(u) = -\log p_z(u)$ and $\mathbb{E}_{p(u)}[v] = \sum_{j \in \mathcal{V}} p_j(u) v_j$. Its NLL response to visual distortion satisfies*

$$\Delta\ell_z = \ell_z(\rho') - \ell_z(\rho) = \int_{\rho'}^{\rho} \big(v_z - \mathbb{E}_{p(u)}[v]\big)\, \mathrm{d}u, \tag{3}$$

Theorem. 3 relates the original token's response to visual degradation to its relative visual support. The $v_z - \mathbb{E}_{p(u)}[v]$ compares the visual contribution to the original token z with the average visual contribution across all tokens, weighted by their probabilities. A positive $\Delta\ell_z$ indicates that this difference is positive on balance over the perturbation interval, so weakening visual information decreases the probability of z. Conversely, a negative $\Delta\ell_z$ indicates that the difference is negative on balance, so weakening visual information increases the probability of z. Thus, the NLL change of the original token z measures the accumulated relative visual support over the interval, motivating visual contrast as a signal for investigating the source reliance underlying VLM confidence. The proof of Theorem. 3 is provided in Appendix B.2.

Table 1: Sign taxonomy of confidence responses under visual degradation. The four regimes are determined by the sign patterns of $(\Delta\ell^F_{m,i}, \Delta h^F_{m,i})$.

	$\Delta h^F_{m,i} < 0$ (Entropy ↓)	$\Delta h^F_{m,i} > 0$ (Entropy ↑)
$\Delta\ell^F_{m,i} > 0$ (NLL ↑)	**Conflict-V** ($E^- N^+$)	Agreement
$\Delta\ell^F_{m,i} < 0$ (NLL ↓)	Agreement	**Conflict-L** ($E^+ N^-$)

2.4 Approach: Visual Contrast for Reliable Uncertainty (VCRU)

Based on the above analysis, we introduce **Visual Contrast for Reliable Uncertainty (VCRU)** for label-free VLM ranking. Theorem 1 and Theorem 2 show that strong support from a single information source can produce high confidence without guaranteeing prediction reliability. Theorem 3 further shows that likelihood changes under visual degradation reflect the relative visual support of the original prediction. Motivated by these results, VCRU probes confidence using fixed visual distortion and identifies responses that may be unreliable for model ranking.

Image perturbation. Inspired by contrastive approaches that compare model outputs under original and distorted visual inputs (Leng et al., 2024; Jung et al., 2025), we perturb images with Gaussian noise and examine the resulting changes in uncertainty. For each input $x_i = (I_i, q_i)$, we construct a distorted counterpart $x'_i = (I'_i, q_i)$, where I'_i denotes the perturbed image. Let $\mathbf{v}_{i,0}$ denote the original visual input corresponding to I_i. Following the Gaussian forward process in diffusion models (Ho et al., 2020), we progressively perturb $\mathbf{v}_{i,0}$ as follows: $q_{\text{diff}}\left(\mathbf{v}_{i,t} \mid \mathbf{v}_{i,t-1}\right) = \mathcal{N}\left(\mathbf{v}_{i,t}; \sqrt{1-\beta_t}\mathbf{v}_{i,t-1}, \beta_t \mathbf{I}_{\text{id}}\right)$. Here, $t \in \{1, \ldots, T\}$ denotes the perturbation step, β_t controls the noise strength, $\mathbf{I}_{\text{id}}$ is the identity matrix, and $\mathcal{N}(\cdot; \mu, \Sigma)$ denotes a Gaussian distribution with mean μ and covariance Σ. Starting from $\mathbf{v}_{i,0}$, the process progressively removes visual information as t increases. We use the final perturbed input $\mathbf{v}_{i,T}$ as I'_i and define the distorted input as $x'_i = (I'_i, q_i)$.

Confidence conflict selction. For the first token $z_{m,i,1}$ generated under the original input, we compute its distorted-input NLL as $\ell^{F'}_{m,i} = -\log p_m(z_{m,i,1} \mid x'_i)$. The NLL shift $\Delta\ell^F_{m,i} = \ell^{F'}_{m,i} - \ell^F_{m,i}$ measures how visual perturbation changes the model's support for the original prediction. Similarly, we define the entropy shift as $\Delta h^F_{m,i} = h^{F'}_{m,i} - h^F_{m,i}$, where $h^F_{m,i}$ and $h^{F'}_{m,i}$ denote the first-token entropy under x_i and x'_i, respectively.

The NLL shift measures the change in support for the original token, while the entropy shift measures the change in uncertainty of the entire first-token distribution. Based on their sign patterns, we divide the confidence responses into four regimes, as shown in Table 1.

Conflict-V ($E^- N^+$): According to Theorem 3, an NLL increase under visual degradation indicates that weakening visual information reduces support for the original prediction. The simultaneous decrease in entropy provides additional evidence that the model is not simply becoming more uncertain: instead, the output distribution becomes sharper as probability shifts toward competing predictions. This pattern suggests that visual evidence is important for maintaining the original prediction's confidence, and we therefore regard Conflict-V as a diagnostic pattern of potentially excessive visual dependence.

Conflict-L ($E^+ N^-$): Conflict-L captures the complementary behavior. Here, the original prediction becomes more probable after visual degradation, even though the overall output distribution becomes less concentrated. By Theorem 3, the NLL decrease indicates that reducing visual information increases the relative support for the original prediction. Since this increase occurs while distribution-level uncertainty grows, the resulting confidence is more likely to be sustained by non-visual information rather than visual evidence. We therefore use Conflict-L as a diagnostic pattern of potentially excessive non-visual, including linguistic, dependence. We further measure how often the original-image prediction matches the prediction under text-only or visually degraded conditions, as reported in Appendix D.6. **Conflict-V** exhibits lower prediction agreement under text-only and visually degraded conditions, whereas **Conflict-L** exhibits higher agreement. These observations are consistent with our interpretation of **Conflict-V** and **Conflict-L** as diagnostic patterns of potentially excessive visual and non-visual dependence, respectively.

Accordingly, we define the Confidence Conflict Selection region as $\alpha_{m,i} = \mathbb{I}\left[\Delta h^F_{m,i} \Delta\ell^F_{m,i} < 0\right]$, where $\alpha_{m,i}$ indicates whether instance i is selected as a confidence-conflict sample for model f_m, and $\mathbb{I}[\cdot]$ denotes the indicator function.

Table 2: Ranking performance on four benchmarks, measured by Spearman's correlation (ρ), Kendall's correlation (τ), and R^2 between proxy scores and model performance. Higher values indicate better ranking performance. Best results are in **bold**, and second-best results are underlined.

Method	MMStar			MMBench			AI2D			ScienceQA		
	ρ	τ	R^2	ρ	τ	R^2	ρ	τ	R^2	ρ	τ	R^2
G-NLL	0.5995	0.4825	0.1819	0.5066	0.3879	0.0331	0.7678	0.6413	0.0504	0.7697	0.6211	0.0041
MSP	0.7071	0.5333	0.4724	0.5846	0.4420	0.6934	0.8561	0.7111	0.8434	0.8132	0.6561	0.5882
Perplexity / PPL	0.5554	0.4114	0.3695	0.6034	0.4261	0.3003	0.6950	0.4984	0.4724	0.8028	0.6211	0.5014
Mean token entropy	0.5065	0.3828	0.1120	0.5958	0.4102	0.1282	0.6821	0.4794	0.2280	0.6156	0.5163	0.0777
Energy	0.5294	0.3869	0.3491	0.5062	0.3498	0.1666	0.5552	0.3683	0.2731	0.3945	0.2907	0.1438
SC	0.7133	0.6236	0.3889	0.5832	0.4464	0.6976	0.7254	0.5905	0.6921	**0.8923**	**0.7673**	**0.7962**
$\mathrm{NLL}_{\max}$	0.6773	0.5544	0.2614	0.5803	0.4515	0.6368	0.8831	0.7175	0.8563	0.8295	0.6751	0.6318
$\mathrm{Ent}_{\max}$	0.6396	0.4909	0.2829	0.6172	0.4388	0.4146	0.8363	0.6540	0.5103	0.6241	0.5290	0.2487
Ent_F	0.7205	0.5830	0.3085	0.7067	0.5215	0.3632	0.8463	0.6603	0.5633	0.6505	0.5576	0.2487
VCRU (Ent_F)	0.8257	0.6974	0.5137	0.8076	0.6455	0.4175	0.8659	0.6984	0.5844	0.8478	0.7069	0.6831
Δ (%)	+14.60%	+19.62%	+66.52%	+14.28%	+23.78%	+14.95%	+2.32%	+5.77%	+3.75%	+30.33%	+26.78%	+174.67%
NLL_F	0.8273	0.6783	0.6680	0.7375	0.5501	0.7605	0.8559	0.6889	0.8456	0.8602	0.7006	0.7334
VCRU (NLL_F)	**0.9437**	**0.8054**	**0.8554**	**0.8506**	**0.7059**	**0.8351**	**0.9344**	**0.7937**	**0.8680**	0.8779	0.7164	0.7424
Δ (%)	+14.07%	+18.74%	+28.05%	+15.32%	+28.32%	+9.81%	+9.17%	+15.21%	+2.65%	+2.06%	+2.26%	+1.23%

Drift Gate. Confidence Conflict Selection identifies potential visually grounded and language-prior confidence failures, but the sign pattern alone may also capture negligible fluctuations. We therefore introduce a drift gate that emphasizes originally high-confidence responses with substantial confidence changes after visual degradation. For each model f_m, we compute the mean μ_m^F and standard deviation σ_m^F of its first-token NLL under the original visual input. Similarly, $\mu_m^{F'}$ and $\sigma_m^{F'}$ denote the corresponding statistics under the distorted visual input. We define the model-relative confidence values as $c_{m,i}^F = (\mu_m^F - \ell_{m,i}^F)/(\sigma_m^F + \epsilon)$ and $c_{m,i}^{F'} = (\mu_m^{F'} - \ell_{m,i}^{F'})/(\sigma_m^{F'} + \epsilon)$. For high confidence instances satisfying $c_{m,i}^F > 0$, the drift gate is defined as $g_{m,i}^{\mathrm{drift}} = c_{m,i}^F \frac{\left|c_{m,i}^F - c_{m,i}^{F'}\right|}{1+\left|c_{m,i}^F - c_{m,i}^{F'}\right|}$.

$c_{m,i}^F > 0$ indicates that the original-input NLL is below the model average, while $|c_{m,i}^F - c_{m,i}^{F'}|$ measures the confidence change after visual degradation. Normalizing by each model's own NLL distribution reduces sensitivity to cross-model scale differences. We then aggregate the drift gate over the samples selected by Confidence Conflict Selection: $P_m = \frac{\sum_{i=1}^N \alpha_{m,i} g_{m,i}^{\mathrm{drift}}}{\sum_{i=1}^N \alpha_{m,i} + \epsilon}$.

Here, P_m denotes the penalty for model f_m. A larger P_m indicates that selected conflict samples have higher original confidence and exhibit larger confidence changes after visual degradation.

VCRU score. The final VCRU score $s_m^{\mathrm{VCRU}}(\lambda)$ is defined as:

$$s_m^{\mathrm{VCRU}}(\lambda) = \mathrm{base}_m + \lambda P_m, \tag{4}$$

where the correction coefficient $\lambda \leq 0$ controls the correction strength. When $\lambda = 0$, VCRU reduces to the baseline score base_m (e.g., NLL-based ranking), while more negative values of λ impose stronger penalties on models with larger P_m.

3 Experiments

3.1 Experimental Setup

This subsection describes the benchmarks, models, and baselines used in our experiments. Details of the evaluation metrics and implementation are provided in Appendix C.1.

Benchmarks. We evaluate our method on four MCQA benchmarks: MMStar (Chen et al., 2024b), MMBench (Liu et al., 2024b), AI2D (Kembhavi et al., 2016), and ScienceQA (Lu et al., 2022).

Models. Our goal is to evaluate whether label-free ranking scores can reliably identify stronger VLMs from a diverse candidate pool. We include models from LLaVA-V1.5 (Liu et al., 2024a), ShareGPT4V (Chen et al., 2024a), LLaVA-NeXT (Liu et al., 2024a), InstructBLIP (Dai et al., 2023), LLaVA-NeXT-Interleave (Li et al., 2024), LLaVA-OneVision (Li et al., 2025), Eagle (Shi et al., 2025), mPLUG-Owl (Ye et al., 2024), InternVL (Chen et al., 2024c; Zhu et al., 2025), PaliGemma (Beyer et al., 2024), Mantis (Jiang et al., 2024), DeepSeek-VL2 (Wu et al., 2024),

Qwen2-VL (Wang et al., 2024b), and Qwen2.5-VL (Bai et al., 2025). In total, we evaluate 36 VLMs, all publicly available through Hugging Face (Wolf et al., 2020). The complete model list is provided in Table 8 of the Appendix.

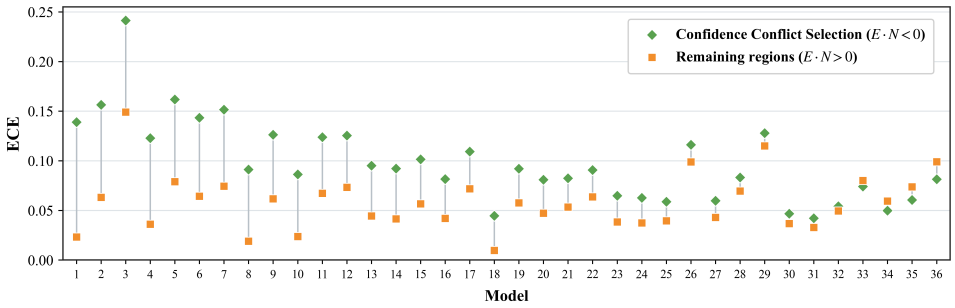

Figure 3: Expected calibration error (ECE) across 36 VLMs on MMBench. Green diamonds represent the ECE of samples identified by Confidence Conflict Selection ($E \cdot N < 0$), while orange squares represent the ECE of the remaining samples for each model. Observation reveals samples identified by Confidence Conflict Selection generally have higher ECE across different models.

Table 3: Ranking performance under randomly sampled 10% subsets of MMStar and MMBench over 50 seeds. The results indicate that VCRU still exhibits better performance with limited samples.

Method	MMStar			MMBench		
	ρ	τ	R^2	ρ	τ	R^2
NLL_F	0.813 ± 0.025	0.653 ± 0.029	0.647 ± 0.038	0.7426 ± 0.0268	0.5634 ± 0.0264	0.7581 ± 0.0146
Ent_F	0.706 ± 0.023	0.563 ± 0.025	0.303 ± 0.019	0.7071 ± 0.0123	0.5204 ± 0.0138	0.3630 ± 0.0105
VCRU	$\mathbf{0.870 \pm 0.029}$	$\mathbf{0.701 \pm 0.037}$	$\mathbf{0.720 \pm 0.056}$	$\mathbf{0.7702 \pm 0.0489}$	$\mathbf{0.6089 \pm 0.0504}$	$\mathbf{0.7914 \pm 0.0322}$

Baselines. We compare VCRU with label-free ranking baselines from three categories: logit-based, consistency-based, and prompt-based methods. For logit-based uncertainty, we consider maximum sequence probability (MSP) (Wang et al., 2026), G-NLL (Aichberger et al., 2026), energy score (Liu et al., 2020), entropy (Kadavath et al., 2022; Kuhn et al., 2023), negative log-likelihood (NLL) (Manakul et al., 2023; Tu et al., 2025; Aichberger et al., 2026; Buffa & Corro, 2026), perplexity (PPL), and Mean Token Entropy (Fomicheva et al., 2020). Following Tu et al. (2025), we further evaluate several NLL- and entropy-based variants: NLL_F and Ent_F are computed at the first generated token, while $\mathrm{NLL}_{\max}$ and $\mathrm{Ent}_{\max}$ use the maximum token-level uncertainty over the response. For consistency-based ranking, we use self-consistency (SC) (Xiong et al., 2024; Wang & Stengel-Eskin, 2026), which estimates confidence from agreement among multiple sampled responses. Further details of SC are provided in Appendix C.2.

3.2 Experiment Results

Superior performance and generalization. Experimental results on different benchmarks are shown in Table 2. On MMStar, VCRU achieves the best performance across all metrics, obtaining 0.9437 ρ, 0.8054 τ, and 0.8554 R^2. Compared with the strongest logit-based baseline, VCRU improves ρ by +0.1164, τ by +0.1271, and R^2 by +0.1874. Among the logit-based baselines, NLL_F performs best, confirming that first-token NLL is a strong label-free ranking signal.

It is worth noting that SC slightly outperforms VCRU on ScienceQA.

Table 4: Model-ranking performance on OCRVQA and VCR_EN_HARD across 36 VLMs. MMStar Acc. and MMBench Acc. rank models directly by their accuracies on the corresponding source benchmarks, using these accuracies as proxies for performance on each target dataset. The best results are **bolded**.

Method	OCRVQA			VCR_EN_HARD		
	ρ	τ	R^2	ρ	τ	R^2
MMStar Acc.	0.0246	0.0365	0.0203	0.3132	0.2220	0.1466
MMBench Acc.	0.2319	0.1876	0.0034	0.3989	0.2811	0.1267
Ent_F	0.5300	0.4032	0.0862	0.5039	0.3748	0.1337
VCRU (Ent_F)	0.8242	0.6063	0.3408	0.5479	0.3882	0.1780
NLL_F	0.6587	0.4952	0.1670	0.5426	0.4050	0.1504
VCRU (NLL_F)	**0.8494**	**0.6474**	**0.6111**	**0.5840**	**0.4487**	**0.1890**

Table 5: Evaluation on the newly added Qwen3.5-9B. AE denotes absolute accuracy prediction error. Rank (Err) denotes the rank obtained directly from proxy scores among 37 models, with absolute rank errors in parentheses. Best results are **bolded**.

Method	MMStar		MMBench	
	AE ↓	Rank (Err) ↓	AE ↓	Rank (Err) ↓
Ent_F	18.18	4 (3)	12.67	4 (3)
NLL_F	12.62	**1 (0)**	7.49	2 (1)
VCRU (Ent_F)	16.39	**1 (0)**	11.65	**1 (0)**
VCRU (NLL_F)	**9.11**	**1 (0)**	**5.91**	**1 (0)**

Table 6: Ablation study of VCRU. We evaluate the contributions of Confidence Conflict Selection and the Drift Gate on MMStar and MMBench. The full VCRU method achieves the best results across all three metrics. Best results are **bolded**.

Variant	MMStar			MMBench		
	ρ	τ	R^2	ρ	τ	R^2
Base (NLL_F)	0.8273	0.6783	0.6680	0.7375	0.5501	0.7605
w/o Selection	0.8927	0.7196	0.7585	0.8020	0.6423	0.8036
w/o Drift Gate	0.8532	0.6974	0.6886	0.7633	0.5882	0.7956
VCRU (NLL_F)	**0.9437**	**0.8054**	**0.8554**	**0.8506**	**0.7059**	**0.8351**

Specifically, SC obtains a ρ of 0.8923, while VCRU achieves 0.8779. This can be attributed to the relatively weaker visual dependency of ScienceQA compared with other benchmarks (Lu et al., 2022; Chen et al., 2024b). In such cases, SC can benefit from repeated sampling consistency, since stable answer generation may already provide a strong ranking signal without visual contrast.

Confidence Conflict Selection captures miscalibrated predictions. We report the Expected Calibration Error (ECE) separately for the samples selected by Confidence Conflict Selection and for the remaining samples on MMBench. The selected samples generally have higher ECE across models (Figure 3). The higher ECE of the selected samples reflects that the Conflict-V and Conflict-L regimes are associated with the two confidence failure modes.

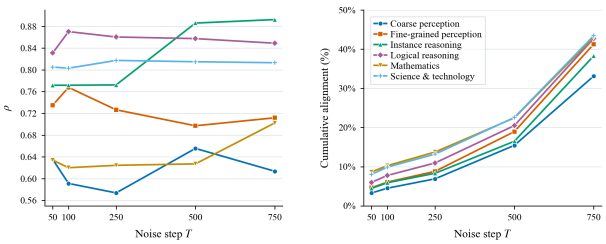

Figure 4: (a) Category-specific ranking performance of VCRU across noise steps, measured by the Spearman correlation ρ between VCRU scores and model performance across MMStar categories. (b) Cumulative alignment with the text-only prediction across noise steps. We report the cumulative proportion of samples whose first alignment with the text-only prediction occurs at a tested noise step T. These results suggest that milder perturbations may better preserve fine-grained visual cues and confidence differences, thereby helping distinguish model capabilities on fine-grained perception tasks.

Limited samples preserve ranking performance. To verify whether reliable model rankings can be obtained using a small set of samples, we evaluate ranking performance under the 10% sampling setting over 50 seeds on MMStar. As shown in Table 3, VCRU consistently outperforms the baselines, demonstrating that VCRU can preserve reliable model rankings with only a small subset of samples, reducing evaluation cost while retaining its advantage over raw uncertainty baselines.

Reliable model rankings from MCQ to other tasks. We further analyze whether VCRU also exhibit reliable ranking capability on other tasks. As shown in Table 4, we compare VCRU with its baseline results on OCRVQA (Mishra et al., 2019) and VCR_EN_HARD (Zhang et al., 2025). We additionally consider two cross-benchmark proxies that rank these models directly by their accuracies on MMStar and MMBench. These proxies assess whether ranking obtained on general multimodal benchmarks remain informative for a new target task. Compared with VCRU and its baseline, both accuracy proxies exhibit weak correlations with ranking performance, while VCRU continues to outperform its baselines.

Different categories favor different perturbation strengths. We analyze the ranking performance of VCRU across different categories on MMStar. As shown in Figure 4 (a), Fine-grained Perception(FP) prefers mild perturbations, whereas Coarse Perception(CP) favors stronger perturbations. We further report cumulative alignment with the text-only prediction across noise steps. As shown in Figure 4 (b), FP exhibits higher cumulative alignment than CP at different tested steps. Together, these observations suggest that milder perturbations may better preserve informative differences in

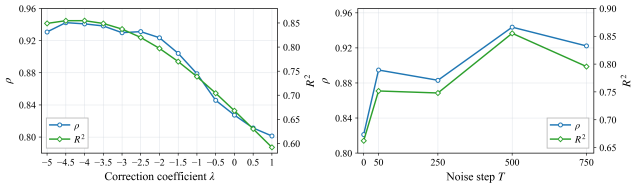

Figure 5: Hyperparameter Sensitivity analysis of VCRU on MMStar. (a) Sensitivity to the correction coefficient λ. A more negative λ applies a stronger penalty to the baseline score. (b) Sensitivity to the noise steps T. A larger T corresponds to stronger perturbations.

Table 7: Efficiency comparison on MMStar using Qwen2.5-VL-3B-Instruct. We report runtime and GPU memory usage; lower values are better. VCRU incurs modest overhead over NLL_F while requiring less runtime and GPU memory than SC.

Method	T(s/sample)↓	GPU Memory↓
NLL_F	0.54 s (× 1.00)	11.12 GiB (× 1.00)
SC	1.91 s (× 3.54)	22.96 GiB (× 2.07)
VCRU	0.66 s (× 1.22)	14.42 GiB (× 1.30)

confidence responses for FP, whereas stronger noise may remove subtle but relevant visual details and reduce the usefulness of these responses for distinguishing model capabilities.

Evaluation on a New VLM. To evaluate VCRU in a practical scenario where a new VLM becomes available, we use Qwen3.5-9B (Team, 2026) as a newly added model and examine whether VCRU remains effective. As shown in Table 5, VCRU correctly ranks Qwen3.5-9B on both benchmarks and exhibits lower absolute accuracy prediction compared to the baselines.

3.3 Ablation Study

Both Selection and Drift Gate contribute. Table 6 presents an ablation study of VCRU. We evaluate the contributions of Confidence Conflict Selection and the Drift Gate by removing each component separately. On MMStar, the full VCRU method achieves $\rho = 0.9437$, $\tau = 0.8054$, and $R^2 = 0.8554$. Removing Selection reduces ρ to 0.8927, while removing the Drift Gate reduces it to 0.8532. These results support the contributions of both components to model ranking. Additional analysis is provided in Appendix D.7.

Stable performance across a wide hyperparameter range. Figure. 5 analysis the sensitivity of VCRU on MMStar to the correction coefficient λ and noise steps T. We further evaluate the generalization of the correction coefficient through leave-one-model-out validation in Appendix D.1.

Efficiency-Performance trade-off. To assess the computational cost of VCRU, we evaluate its runtime and peak GPU memory usage on MMStar benchmark with Qwen2.5-VL-3B-Instruct. As shown in Table 7, VCRU maintains acceptable inference overhead. In particular, it runs substantially faster than SC and provides a more favorable trade-off between uncertainty ranking performance and computational efficiency compared to NLL_F.

4 Conclusion

In this work, we investigate how to reliably leverage model confidence for unsupervised ranking of VLM performance. Through empirical and theoretical analyses, we show that confidence driven primarily by a single information source is more susceptible to overconfidence than confidence jointly supported by visual and linguistic evidence. Motivated by this finding, we propose Visual Contrast for Reliable Uncertainty (VCRU), a label-free method that identifies confidence conflicts by comparing the negative log-likelihood (NLL) of the original first-token prediction and the change in output entropy between the original and visually degraded inputs. VCRU further employs a drift gate to emphasize predictions that exhibit both high confidence on the original input and substantial changes in relative confidence after visual degradation. Extensive experiments demonstrate the superiority of VCRU compared with existing baselines.

Limitations and Future Work. For API-only VLMs, token-level generation probabilities are typically inaccessible, which limits the direct application of our current method to black-box systems. In future work, we will investigate extensions that operate using only observable outputs or query-based signals. We also plan to extend the proposed framework beyond the current evaluation setting to open-ended generation and calibration tasks.

AI USE STATEMENT

In this work, we used generative AI tools to polish and improve the readability of the paper. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Lukas Aichberger, Kajetan Schweighofer, and Sepp Hochreiter. Rethinking uncertainty estimation in LLMs: A principled single-sequence measure. In *The Fourteenth International Conference on Learning Representations*, 2026.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report, 2025.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

Pedro Memoli Buffa and Luciano Del Corro. Entropy sentinel: Probing entropy traces for llm monitoring, 2026. URL https://arxiv.org/abs/2601.09001.

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

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024b.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024c.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Weijian Deng, Yumin Suh, Stephen Gould, and Liang Zheng. Confidence and dispersity speak: Characterizing prediction matrix for unsupervised accuracy estimation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7658–7674. PMLR, July 2023.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11198–11201, 2024.

Hao Fang, Changle Zhou, Jiawei Kong, Kuofeng Gao, Bin Chen, Tao Liang, Guojun Ma, and Shu-Tao Xia. Grounding language with vision: A conditional mutual information calibrated decoding strategy for reducing hallucinations in LVLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=W5HnIf2jla.

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

539–555, 2020. doi: 10.1162/tacl_a_00330. URL https://aclanthology.org/2020.tacl-1.35/.

Malcolm R Forster. Key concepts in model selection: Performance and generalizability. *Journal of mathematical psychology*, 44(1):205–231, 2000.

Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, et al. Mini-internvl: a flexible-transfer pocket multi-modal model with 5% parameters and 90% performance. *Visual Intelligence*, 2(1):32, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Qiang Hu, Jin Wen, Yao Zhang, Maxime Cordy, and Yongqiang Lyu. On the evaluation of capability estimation methods for large language models. In *Proceedings of the 40th AAAI Conference on Artificial Intelligence (AAAI)*, pp. 31077–31085. AAAI Press, 2026.

Ta Duc Huy, Trang Nguyen, Townim Chowdhury, Ankit Yadav, Minh-Son To, Zhibin Liao, Johan W. Verjans, and Vu Minh Hieu Phan. Visual semantic entropy: Do vision language models recognize visual ambiguity?, 2026. URL https://arxiv.org/abs/2606.31407.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024.

Chaeyoung Jung, Youngjoon Jang, and Joon Son Chung. AVCD: Mitigating hallucinations in audio-visual large language models through contrastive decoding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=3zx087XPtz.

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.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision – ECCV 2016*, pp. 235–251, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46493-0.

Maurice George Kendall. *Rank Correlation Methods.* Griffin, 1948.

Reza Khanmohammadi, Erfan Miahi, Simerjot Kaur, Charese H. Smiley, Ivan Brugere, Kundan Thind, and Mohammad M. Ghassemi. Grounded or guessing? lvlm confidence estimation via blind-image contrastive ranking, 2026. URL https://arxiv.org/abs/2605.10893.

Jieun Kim, Jinmyeong Kim, Yoonji Kim, and Sung-Bae Cho. Fuzzy contrastive decoding to alleviate object hallucination in large vision-language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20572–20581. IEEE, 2025. doi: 10.1109/ICCV51701.2025.01913.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Tarald O. Kvålseth. Cautionary note about R2. *The American Statistician*, 39(4):279–285, 1985. ISSN 00031305, 15372731.

Jihoon Lee and Min Song. Retrieval visual contrastive decoding to mitigate object hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8200–8219, 2025.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882, Seattle, WA, USA, June 2024. IEEE. ISBN 979-8-3503-5300-6. doi: 10.1109/CVPR52733.2024.01316.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=zKv8qULV6n.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.

Jiaming Li, Jiacheng Zhang, Zequn Jie, Lin Ma, Ming Li, Xiaonan Luo, and Guanbin Li. Cross-modal attention calibration for lvlm hallucination mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40186–40196, 2026.

Bingqian Liu, Fu Zhang, Guoqing Chen, and Jingwei Cheng. Multi-frequency contrastive decoding: Alleviating hallucinations for large vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28556–28572. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1452.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024a.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pp. 216–233. Springer, 2024b.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Jingying Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, et al. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in vlms. In *International Conference on Learning Representations*, volume 2026, pp. 72849–72876, 2026.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jN5y-zb5Q7m.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557.

John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pp. 7721–7735. PMLR, 2021.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 947–952, 2019. doi: 10.1109/ICDAR.2019.00156.

Benjamin Plaut, Khanh Xuan Nguyen, and Tu Trinh. Probabilities of chat LLMs are miscalibrated but still predict correctness on multiple-choice q&a. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=E6LOh5vz5x.

Mario Sanz-Guerrero, Minh Duc Bui, and Katharina von der Wense. Mind the gap: A closer look at tokenization for multiple-choice question answering with llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19584–19594, 2025.

Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, Yilin Zhao, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, et al. Eagle: Exploring the design space for multimodal llms with mixture of encoders. In *International Conference on Learning Representations*, volume 2025, pp. 92628–92646, 2025.

Grace S. Shieh. A weighted Kendall's tau statistic. *Statistics & Probability Letters*, 39(1):17–24, 1998. ISSN 0167-7152. doi: 10.1016/S0167-7152(98)00006-6.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Weijie Tu, Weijian Deng, Dylan Campbell, Yu Yao, Jiyang Zheng, Tom Gedeon, and Tongliang Liu. Ranked from within: Ranking large multimodal models without labels. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=F1BCivEK57.

Chenyu Wang, Tianle Chen, HM Ahmad, Kayhan Batmanghelich, and Wenchao Li. VLM-uqbench: A benchmark for modality-specific and cross-modality uncertainties in vision language models. *arXiv preprint arXiv:2602.09214*, 2026.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *Advances in Neural Information Processing Systems*, 37:75392–75421, 2024a.

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

Victor Wang and Elias Stengel-Eskin. Calibrating verbalized confidence with self-generated distractors. In *International Conference on Learning Representations*, 2026.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. "my answer is c": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 7407–7416, 2024c.

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 *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, 2020.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

Wenyi Xiao, XU Xinchi, and Leilei Gan. Vl-calibration: Decoupled confidence calibration for large vision-language models reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 44791–44815, 2026.

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. In *International Conference on Learning Representations*, volume 2024, pp. 23650–23678, 2024.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. Mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the Ieee/Cvf Conference on Computer Vision and Pattern Recognition*, pp. 13040–13051, 2024.

Tianyu Zhang, Suyuchen Wang, Lu Li, Ge Zhang, Perouz Taslakian, Sai Rajeswar, Jie Fu, Bang Liu, and Yoshua Bengio. VCR: Pixel-level complex reasoning by restoring occluded text. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=s0Z4csHOoE.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Zhihong Zhu, Yunyan Zhang, Fan Zhang, Bowen Xing, and Xian Wu. CMID: Towards medical visual question answering via contrastive mutual information decoding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(41):35275–35283, 2026. doi: 10.1609/aaai.v40i41.40835.

A Related Work

Label-free model ranking. Label-free model ranking aims to estimate the relative performance of candidate models without ground-truth annotations. Early studies considered unsupervised model selection for conventional machine learning models using prediction statistics on unlabeled data (Forster, 2000; Miller et al., 2021; Deng et al., 2023). Recent work extends this setting to foundation models and VLMs, where model responses and uncertainty provide practical signals for estimating relative performance (Plaut et al., 2025; Tu et al., 2025; Hu et al., 2026). Uncertainty-based ranking methods generally follow two paradigms. **Consistency-based methods** measure agreement among multiple responses generated from the same input (Wang et al., 2023; Xiong et al., 2024), based on the intuition that stronger models produce more stable outputs. **Logit-based methods** use statistics from the output distribution, such as token probability (Malinin & Gales, 2021), entropy (Kadavath et al., 2022; Kuhn et al., 2023), maximum softmax probability (Manakul et al., 2023), and negative log-likelihood (NLL) (Manakul et al., 2023; Tu et al., 2025; Aichberger et al., 2026; Buffa & Corro, 2026). These methods generally assume that lower uncertainty indicates stronger model capability. However, this assumption can become unreliable when a VLM depends excessively on a single information source. Such reliance may produce highly consistent or high-confidence predictions without corresponding multimodal capability (Huy et al., 2026; Wang et al., 2024a; Liu et al., 2026; Li et al., 2026).

Visual contrast for multimodal confidence analysis. Visual contrast has been explored in VLMs as a mechanism to analyze model behaviors under different visual conditions. A representative line of work uses visual contrast to improve multimodal generation. Contrastive decoding methods (Leng et al., 2024; Lee & Song, 2025; Fang et al., 2025; Jung et al., 2025; Kim et al., 2025; Liu et al., 2025; Zhu et al., 2026) compare output distributions under original and visually perturbed inputs to mitigate object hallucinations. Beyond that, recent studies explore visual contrast for analyzing confidence reliability in VLMs. BICR (Khanmohammadi et al., 2026) introduces a contrast between the original image and a blacked-out image to address visual ungroundedness. VL-Calibration (Xiao et al., 2026) measures confidence changes under image perturbations to decouple confidence into visual and reasoning components through reinforcement learning. These studies demonstrate that contrasting model behaviors across visual conditions provides useful signals for understanding multimodal reliability. Different from previous contrastive approaches, VCRU does not aim to calibrate confidence estimates or improve generation quality. Instead, it uses visual contrast as a diagnostic signal for label-free VLM ranking. By measuring NLL and entropy changes between original and visually perturbed inputs, VCRU analyzes confidence patterns associated with excessive reliance on a single information source and improves the reliability of uncertainty-based model ranking.

B Additional Theoretical Details

B.1 Proof of Theorem 1

Proof. For any competing token $k \neq z$, the logit difference between the predicted token z and token k is

$$\begin{aligned} s_z(\rho) - s_k(\rho) &= \big(a_z + \rho v_z\big) - \big(a_k + \rho v_k\big) \\ &= (a_z - a_k) + \rho(v_z - v_k). \end{aligned} \tag{5}$$

Under either form of predominance, one information source supports z over every competing token by at least γ, while any opposing contribution from the other source is bounded by B. Therefore, regardless of whether the predominant information source is visual or linguistic,

$$s_z(\rho) - s_k(\rho) \geq \gamma - B, \qquad \forall k \neq z. \tag{6}$$

Since $\gamma > B$, this logit difference is strictly positive. Equivalently,

$$s_k(\rho) - s_z(\rho) \leq -(\gamma - B), \qquad \forall k \neq z.$$

We next use this bound to derive a lower bound on the confidence assigned to z. By the softmax definition,

$$p_z(\rho) = \frac{\exp(s_z(\rho))}{\sum_{j \in \mathcal{V}} \exp(s_j(\rho))}.$$

Dividing both the numerator and denominator by $\exp(s_z(\rho))$ gives

$$p_z(\rho) = \frac{1}{1 + \sum_{k \neq z} \exp\big(s_k(\rho) - s_z(\rho)\big)}.$$

From Eq. 6,

$$\exp\big(s_k(\rho) - s_z(\rho)\big) \leq \exp\big(-(\gamma - B)\big), \qquad \forall k \neq z.$$

Since there are $|\mathcal{V}| - 1$ competing tokens,

$$\sum_{k \neq z} \exp\big(s_k(\rho) - s_z(\rho)\big) \leq (|\mathcal{V}| - 1) \exp\big(-(\gamma - B)\big).$$

It therefore follows that

$$p_z(\rho) \geq \frac{1}{1 + (|\mathcal{V}| - 1) \exp\big(-(\gamma - B)\big)},$$

which proves Eq. 1.

Now consider the group $\mathcal{G}$ of predictions satisfying these conditions. Since the lower bound above holds for every prediction in $\mathcal{G}$, taking the conditional expectation gives

$$\mathbb{E}[p_z(\rho) \mid \mathcal{G}] \geq \frac{1}{1 + (|\mathcal{V}| - 1) \exp\big(-(\gamma - B)\big)}.$$

Together with the assumption

$$\Pr(\hat{y} = y^\star \mid \mathcal{G}) \leq r,$$

we obtain

$$\begin{aligned} &\mathbb{E}[p_z(\rho) \mid \mathcal{G}] - \Pr(\hat{y} = y^\star \mid \mathcal{G}) \\ &\quad \geq \frac{1}{1 + (|\mathcal{V}| - 1) \exp\big(-(\gamma - B)\big)} - r, \end{aligned}$$

which proves Eq. 2.

Finally, define

$$C(\gamma - B) = \frac{1}{1 + (|\mathcal{V}| - 1) \exp\big(-(\gamma - B)\big)}.$$

Since

$$\frac{\partial C(\gamma - B)}{\partial(\gamma - B)} = \frac{(|\mathcal{V}| - 1) \exp\big(-(\gamma - B)\big)}{\big[1 + (|\mathcal{V}| - 1) \exp\big(-(\gamma - B)\big)\big]^2} > 0,$$

the confidence lower bound increases monotonically as $\gamma - B$ increases and approaches 1. □

B.2 Proof of Theorem 3

Proof. For the fixed original token z, its probability under the visual condition u is

$$p_z(u) = \frac{\exp(a_z + u v_z)}{\sum_{k \in \mathcal{V}} \exp(a_k + u v_k)}.$$

Therefore, its NLL can be written as

$$\begin{aligned} \ell_z(u) &= -\log p_z(u) \\ &= -a_z - u v_z + \log \sum_{k \in \mathcal{V}} \exp(a_k + u v_k). \end{aligned}$$

We first compute how $\ell_z(u)$ changes with u. Differentiating the first two terms gives $-v_z$. For the last term,

$$\begin{aligned} &\frac{\partial}{\partial u} \log \sum_{k \in \mathcal{V}} \exp(a_k + u v_k) \\ &\quad = \frac{\sum_{k \in \mathcal{V}} v_k \exp(a_k + u v_k)}{\sum_{j \in \mathcal{V}} \exp(a_j + u v_j)} \\ &\quad = \sum_{k \in \mathcal{V}} p_k(u) v_k = \mathbb{E}_{p(u)}[v]. \end{aligned}$$

Hence,

$$\frac{\partial \ell_z(u)}{\partial u} = \mathbb{E}_{p(u)}[v] - v_z.$$

Since the original and distorted inputs correspond to ρ and ρ', respectively, with $\rho' < \rho$, the NLL change of the original token is

$$\begin{aligned}\Delta\ell_z &= \ell_z(\rho') - \ell_z(\rho) \\ &= \int_{\rho}^{\rho'} \frac{\partial \ell_z(u)}{\partial u}\,\mathrm{d}u \\ &= \int_{\rho}^{\rho'} \big(\mathbb{E}_{p(u)}[v] - v_z\big)\,\mathrm{d}u.\end{aligned}$$

Reversing the integration limits gives

$$\Delta\ell_z = \int_{\rho'}^{\rho} \big(v_z - \mathbb{E}_{p(u)}[v]\big)\,\mathrm{d}u,$$

which proves Eq. 3.

We next consider the sign of this NLL change. If

$$v_z > \mathbb{E}_{p(u)}[v], \qquad \forall u \in [\rho', \rho],$$

then

$$v_z - \mathbb{E}_{p(u)}[v] > 0$$

throughout the interval. Since $\rho > \rho'$, it follows that

$$\Delta\ell_z > 0.$$

By the definition of $\Delta\ell_z$,

$$-\log p_z(\rho') > -\log p_z(\rho),$$

and therefore

$$p_z(\rho') < p_z(\rho).$$

Thus, weakening visual information reduces the probability assigned to the original token z.

Conversely, if

$$v_z < \mathbb{E}_{p(u)}[v], \qquad \forall u \in [\rho', \rho],$$

then

$$v_z - \mathbb{E}_{p(u)}[v] < 0$$

throughout the interval, which gives

$$\Delta\ell_z < 0.$$

Equivalently,

$$p_z(\rho') > p_z(\rho).$$

Thus, the original token becomes more probable as visual information is weakened. □

C Implementation Details

C.1 More information on Experiments Setup

Evaluation metrics. We evaluate ranking quality using Spearman's rank correlation coefficient ρ (Kendall, 1948) to measure the relationship between proxy scores and model performance. We also report Kendall's rank correlation τ (Shieh, 1998) which measures pairwise ordering agreement. Both metrics range from -1 to 1, with larger positive values indicating better rank agreement. We further report the coefficient of determination R^2 (Kvålseth, 1985) to measure the proportion of variance in true model performance explained by the proxy scores.

Implementation Details. We implement our evaluation pipeline based on VLMEvalKit (Duan et al., 2024). We use greedy decoding for the response generation. For the distorted visual input, we apply

Gaussian noise to the original image. The noise step T controls the distortion strength. We use $T = 500$ in the experiments. For the results reported on MMStar and MMBench, we set correction coefficient $\lambda = -4.2$ and $\lambda = -3$, respectively. A more detailed hyperparameter sensitivity analysis are shown in Figure. 5. Experiments are conducted on NVIDIA L40 GPUs (48 GB per GPU) for 13B-scale models (Eagle-X5-13B, LLaVA-NeXT-13B, etc.), and NVIDIA GeForce RTX 4090 GPUs (24 GB per GPU) for 7B and smaller models.

Details on models. All 36 evaluated models can be downloaded from the Hugging Face Hub (Wolf et al., 2020). We use different `transformers` library versions for each model series:

- `transformers==4.31.0`: mPLUG-Owl2 (Ye et al., 2024)
- `transformers==4.37.2`: LLaVA-v1.5 (Liu et al., 2024a), ShareGPT4V (Chen et al., 2024a), InternVL2 (Chen et al., 2024c), Mini-InternVL (Gao et al., 2024)
- `transformers==4.38.2`: DeepSeek-VL2-Tiny (Wu et al., 2024)
- `transformers==4.45.2`: Eagle (Shi et al., 2025), InstructBLIP (Dai et al., 2023), Mantis (Jiang et al., 2024)
- `transformers==4.49.0`: LLaVA-NeXT (Liu et al., 2024a), LLaVA-OneVision (Li et al., 2025), LLaVA-NeXT-Interleave (Li et al., 2024), PaliGemma (Beyer et al., 2024), Qwen2-VL (Wang et al., 2024b), Qwen2.5-VL (Bai et al., 2025), InternVL3 (Zhu et al., 2025)

Models Series. In Table 8, we present the detailed VLM series and models evaluated in our experiments.

Table 8: VLM series and models evaluated in our experiments.

Series	Model	Series	Model
Eagle (Shi et al., 2025)	Eagle-X5-7B	InstructBLIP (Dai et al., 2023)	InstructBLIP-Vicuna-7B
	Eagle-X5-13B		InstructBLIP-Vicuna-13B
	Eagle-X5-13B-Chat	LLaVA-OneVision (Li et al., 2025)	LLaVA-OneVision-Qwen2-0.5B-SI
InternVL2 (Chen et al., 2024c)	InternVL2-1B		LLaVA-OneVision-Qwen2-0.5B-OV
	InternVL2-2B		LLaVA-OneVision-Qwen2-7B-SI
	InternVL2-4B		LLaVA-OneVision-Qwen2-7B-OV
	InternVL2-8B	LLaVA-NeXT-Interleave (Li et al., 2024)	LLaVA-NeXT-Interleave-7B
InternVL3 (Zhu et al., 2025)	InternVL3-1B		LLaVA-NeXT-Interleave-7B-DPO
	InternVL3-2B	LLaVA-NeXT (Liu et al., 2024a)	LLaVA-NeXT-Mistral-7B
Mini-InternVL (Gao et al., 2024)	Mini-InternVL-Chat-2B-V1.5		LLaVA-NeXT-Vicuna-7B
	Mini-InternVL-Chat-4B-V1.5		LLaVA-NeXT-Vicuna-13B
Mantis (Jiang et al., 2024)	Mantis-8B-Idefics2	LLaVA-v1.5 (Liu et al., 2024a)	LLaVA-v1.5-7B
	Mantis-8B-CLIP-Llama3		LLaVA-v1.5-13B
	Mantis-8B-SigLIP-Llama3	mPLUG-Owl2 (Ye et al., 2024)	mPLUG-Owl2
Qwen2-VL(Wang et al., 2024b)	Qwen2-VL-2B-Instruct	PaliGemma (Beyer et al., 2024)	PaliGemma-3B-Mix-224
Qwen2.5-VL(Bai et al., 2025)	Qwen2.5-VL-3B-Instruct		PaliGemma-3B-Mix-448
	Qwen2.5-VL-7B-Instruct	ShareGPT4V (Chen et al., 2024a)	ShareGPT4V-7B
DeepSeek-VL2 (Wu et al., 2024)	DeepSeek-VL2-Tiny		ShareGPT4V-13B

C.2 Self-Consistency

Self-consistency (SC) estimates prediction confidence from agreement across repeated stochastic generations (Xiong et al., 2024; Wang & Stengel-Eskin, 2026). For each model f_m and sample x_i, let $\hat{y}_{m,i}$ denote the original prediction obtained through greedy decoding. Following Wang & Stengel-Eskin (2026) five additional predictions are obtained by stochastic sampling from the same input, with the number of sampled predictions denoted by $B = 5$. Treating the original prediction as a deterministic self-match, its SC confidence is computed as

$$c_{m,i}^{\mathrm{SC}} = \frac{1 + \sum_{b=1}^{B} \mathbb{I}\left[\hat{y}_{m,i}^{(b)} = \hat{y}_{m,i}\right]}{B + 1} \tag{7}$$

The SC confidence scores are then averaged across all samples to obtain the ranking signal for model f_m. A higher average SC confidence indicates more stable predictions and therefore results in a higher model ranking.

D Additional Experiments

D.1 Leave-One-Model-Out Validation

To examine whether VCRU's improvements are driven by a small subset of models, we perform leave-one-model-out validation, fitting the correction coefficient on the remaining 35 models and evaluating it on the held-out model.As shown in Table 9 VCRU consistently improves ρ, τ, and R^2 over the NLL_F baseline, demonstrating that the gains generalize across models.

Table 9: Leave-one-model-out validation of VCRU on MMStar and MMBench. For each of the 36 VLMs, we fit the correction coefficient λ using the remaining 35 models and compute the VCRU score for the excluded model.

Method	MMStar			MMBench		
	ρ	τ	R^2	ρ	τ	R^2
NLL_F	0.8273	0.6783	0.6680	0.7375	0.5501	0.7605
VCRU (NLL_F)	**0.9401**	**0.7895**	**0.8378**	**0.8302**	**0.6677**	**0.8243**

D.2 Analysis of correction coefficient λ choice

In practical applications, applying VCRU to a new benchmark requires choosing a correction coefficient λ. We therefore examine whether VCRU remains effective with a fixed correction coefficient. Specifically, we use the same $\lambda = -3$ as in our MMBench experiments for other datasets. As shown in Table 10, VCRU still outperforms the baseline across five datasets.

Table 10: Ranking performance of VCRU with a fixed correction coefficient $\lambda = -3$, the same coefficient used on MMBench. We compare VCRU with the NLL_F across 36 VLMs on different datasets. Best results are **bolded**.

Method	MMStar			AI2D			ScienceQA			OCRVQA			VCR_EN_HARD		
	ρ	τ	R^2	ρ	τ	R^2	ρ	τ	R^2	ρ	τ	R^2	ρ	τ	R^2
NLL_F	0.8273	0.6783	0.6680	0.8559	0.6889	0.8456	0.8602	0.7006	0.7334	0.6587	0.4952	0.1670	0.5426	0.4050	0.1504
VCRU ($\lambda = -3$)	**0.9406**	**0.7895**	**0.8496**	**0.8888**	**0.7333**	**0.8597**	**0.8739**	**0.7101**	**0.7361**	**0.7920**	**0.5841**	**0.3278**	**0.5782**	**0.4555**	**0.1811**

D.3 Limitations of Raw Logits for Model ranking

Besides the baselines we already discussed in Table 2, we also examine the effectiveness of raw logits for model ranking. As shown in Table 11, Top-1 logit Top-2 Logit Margin perform worse than other baselines, suggesting that raw logits may be less reliable for model ranking, as their scales can more differ across models,limiting their comparability.

Table 11: Additional analysis of logit-based methods for model ranking on MMStar and MMBench.

Method	MMStar			MMBench		
	ρ	τ	R^2	ρ	τ	R^2
Top-1 Logit	0.6343	0.4750	0.4007	0.5566	0.3847	0.2146
Top-2 Logit Margin	0.7688	0.5989	0.4282	0.5867	0.4293	0.3433
Top-2 Probability Gap	0.7922	0.6402	0.6175	0.7022	0.5246	0.7361
Ent_F	0.7205	0.5830	0.3085	0.7067	0.5215	0.3632
NLL_F	**0.8273**	**0.6783**	**0.6680**	**0.7375**	**0.5501**	**0.7605**

D.4 Detail results of ranking

Figure 6 shows the detail results of ranking on MMStar.

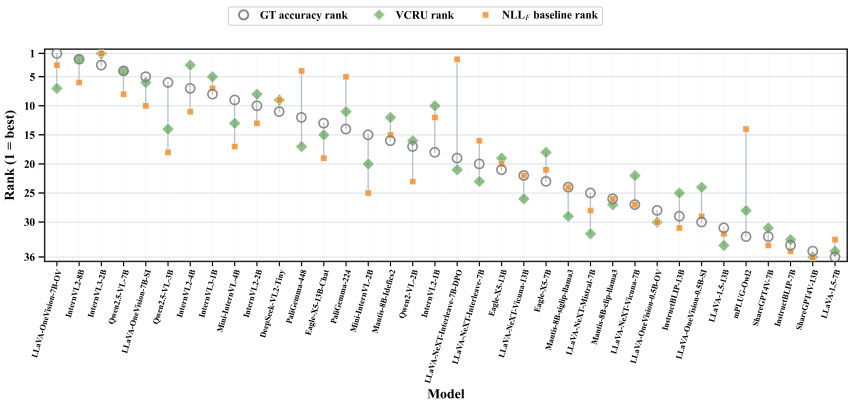

Figure 6: Model ranking comparison on MMStar.We compare model rankings based on VCRU and NLL_F with the ground-truth accuracy across 36 VLMs. Gray circles, green diamonds, and orange squares denote ground-truth, VCRU, and NLL_F rankings, respectively. Models are ordered by ground-truth accuracy, with rank 1 indicating the best performance. VCRU rankings align more closely overall with the ground-truth ranking than NLL_F baseline.

D.5 Alignment Between First-Token Scoring and MCQ Response Format

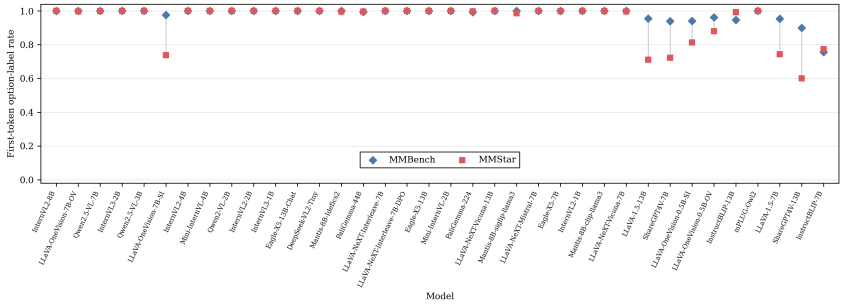

Figure 7: First-Token Option-Label Rate Across 36 VLMs on MMStar and MMBench. We report the proportion of samples whose first token represents an MCQ option lablel. Blue diamonds and red squares denote MMBench and MMStar, respectively. Models are ordered from best to worst by their average ground-truth accuracy rank across the two benchmarks.

VCRU rank the model performance by the first token negative log-likelihood (NLL). However, the first token may not faithfully represent the target option label in MCQ (Wang et al., 2024c; Sanz-Guerrero et al., 2025), since the model can produce irrelevant leading tokens prior to the answer. To evaluate the sensitivity of VCRU to this potential mismatch,we report the proportion of model responses whose first token is an MCQ option label. As shown in Figure 7, most models exhibit high option-label rates, although some show mismatches between the first scored token and the MCQ answer format. Models are ordered from best to worst by their average ground-truth accuracy rank across MMStar and MMBench. Lower-ranked models tend to exhibit lower option-label rates.

D.6 Prediction Agreement and NLL Shift Across Different Regions

We next examine whether the two conflict regions exhibit behaviors consistent with the confidence failure modes motivating VCRU. As shown in Figure 8.(a) and Figure 8.(b), **Conflict-V** (E^-N^+) exhibits the lowest prediction agreement between the original-image condition and both the text-only and visually degraded conditions. This suggest that its original prediction depend more strongly on visual information. Figure 8(c) further shows that its NLL more substantially than in the other regions as visual information is progressively degraded, indicating that the original prediction increasingly loses support under visual distortion. Conversely, **Conflict-L** E^+N^- exhibits the opposite

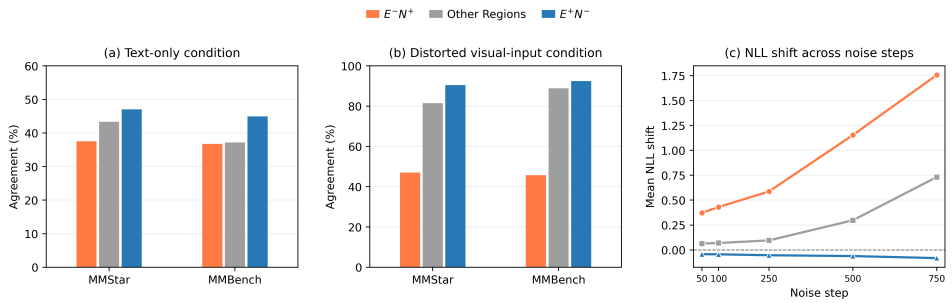

Figure 8: Prediction agreement and NLL shifts across sign regions. (a) We report prediction agreement between the original-image and text-only conditions. (b) We report prediction agreement between the original-image and visually degraded conditions. (c) We report the mean first-token NLL shift across noise steps. Regions are defined by the signs of NLL changes ($N = \Delta\ell^F_{m,i}$) and entropy changes ($E = \Delta h^F_{m,i}$).

Table 12: Correlation of different uncertainty-shift regimes with model performance. For each model, we compute the proportion of samples satisfying each selection rule and use this proportion as a ranking score. We report Spearman's ρ and Kendall's τ between these scores and ground-truth model accuracy.

Group	**Selection rule**	**MMStar**		**MMBench**		**AI2D**		**ScienceQA**	
		ρ	τ	ρ	τ	ρ	τ	ρ	τ
Uncertainty ↓	$E < 0$	-0.4106	-0.3126	-0.1678	-0.1145	-0.1579	-0.1382	-0.3444	-0.2448
	$N < 0$	-0.3016	-0.2024	-0.1418	-0.0986	0.0371	-0.0064	-0.2546	-0.1781
Uncertainty ↑	$E > 0$	0.4245	0.3235	0.1678	0.1145	0.1579	0.1382	0.3444	0.2448
	$N > 0$	0.2966	0.2021	0.1429	0.1049	-0.0347	0.0095	0.2369	0.1575
Same-sign signal	$E < 0,\ N < 0$	0.0209	0.0096	-0.0375	-0.0095	0.2793	0.1890	-0.1323	-0.1113
	$E > 0,\ N > 0$	0.6135	0.4713	0.2362	0.1622	0.4247	0.3079	0.4953	0.3357
Opposite-sign signal	$E < 0,\ N > 0$	-0.4967	-0.3490	-0.4902	-0.3662	-0.8142	-0.6180	-0.7796	-0.5997
	$E > 0,\ N < 0$	-0.5822	-0.4032	-0.5522	-0.4179	-0.6202	-0.4885	-0.7544	-0.5998
Ours	$\boldsymbol{E \cdot N < 0}$	-0.7424	-0.5426	-0.6491	-0.5191	-0.8776	-0.7059	-0.8428	-0.7091

pattern across these analyses: its predictions remain more stable and its NLL does not increase as visual information is weakened.

D.7 Analysis of Selection Rules

We examine whether the opposite-sign selection regimes also provide a model-level ranking signal. **Conflict-V** (E^-N^+) and **Conflict-L** (E^+N^-) reflect unreliable confidence conflict associated with excessive reliance on a single information source. Models exhibiting these patterns more frequently may perform worse. We therefore compute the proportion of samples satisfying each selection rule for every model and correlate these proportions with ground-truth accuracy. As shown in Table 12, the combined opposite-sign rule ($E \cdot N < 0$) yields the strongest negative correlations among the evaluated selection rules across all four benchmarks, supporting its use as a model-level ranking signal.

D.8 Accuracy Analysis across different Sign Regions

Figure 9 shows the distribution of samples across the different regions. The height of each bar indicates the fraction of samples assigned to the corresponding region, while the solid and hatched portions represent correct and incorrect predictions, respectively. Across all four benchmarks, accuracy is consistently lower in the opposite-sign regions.

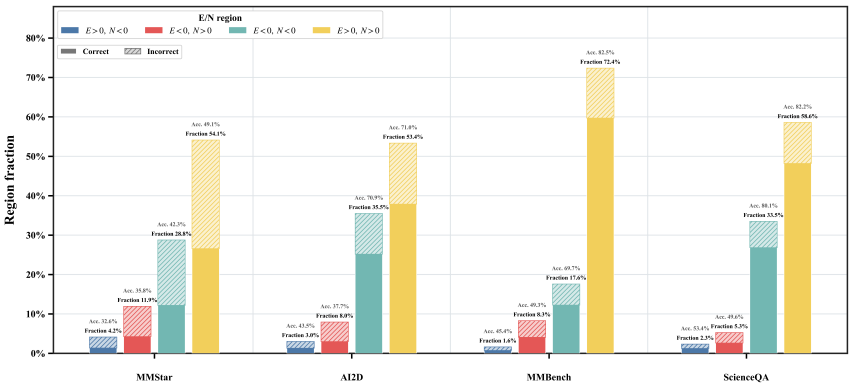

Figure 9: Sample distribution and prediction accuracy across sign regions on four benchmarks.Samples are grouped by the signs of NLL changes ($N = \Delta\ell^F_{m,i}$) and entropy changes ($E = \Delta h^F_{m,i}$) under visual degradation.Each bar shows the proportion of samples in the corresponding region, with solid and hatched portions representing correct and incorrect predictions, respectively. The opposite-sign regions ($E \cdot N < 0$), corresponding to **Conflict-V** and **Conflict-L**, exhibit lower accuracy than the other regions.

D.9 Expected calibration error (ECE) across sign regions.

We examine the Expected Calibration Error (ECE) by binning the baseline ranking scores over samples and comparing bin-wise mean ranking scores with accuracy as illustrated in Figure 10. Except on MMStar, the regions identified by Confidence Conflict Selection already exhibit higher raw ECE than the other regions. MMStar highlights a limitation of relying solely on the directions of entropy and NLL changes: models may already assign lower confidence to predictions in the selected regions, which can result in lower raw ECE instead of failure modes such as overconfidence. The drift gate further distinguishes these cases by emphasizing predictions that have low original-input NLL and undergo substantial confidence drift after visual degradation. After drift gate reweighting, the selected regions generally exhibit higher ECE than the other regions across all four benchmarks.

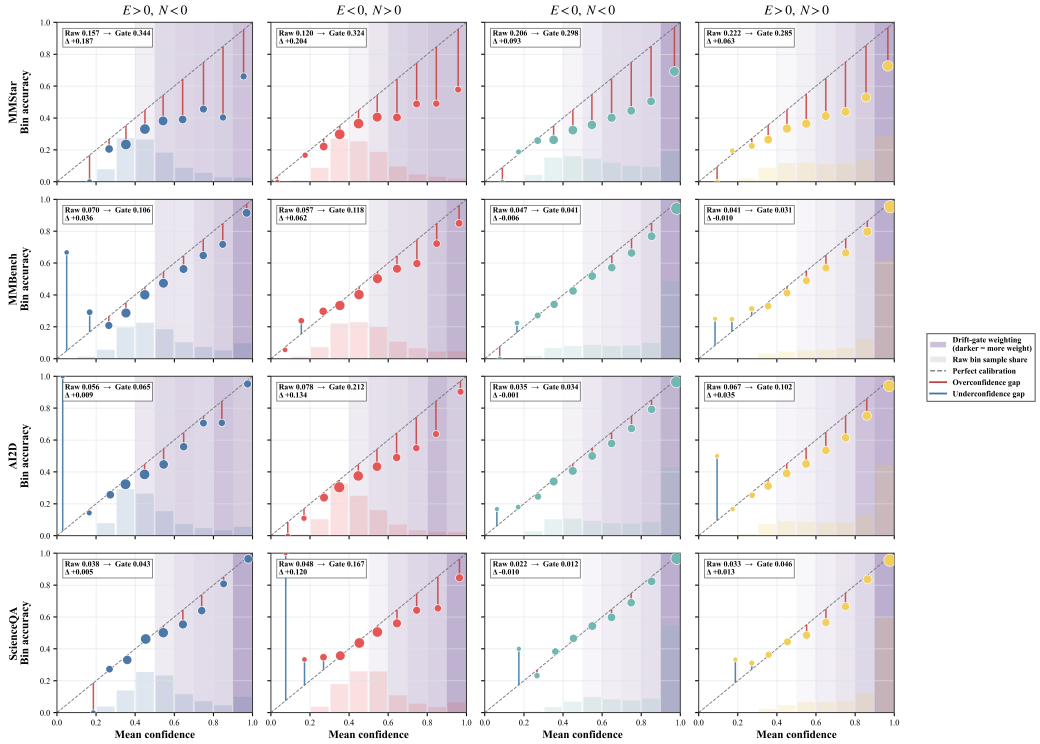

Figure 10: Expected calibration error (ECE) across sign regions on four benchmarks. Confidence is computed from the original-image first-token. Points show bin-wise mean confidence and accuracy and darker purple shading indicates greater drift-gate weight.