# RANKING BLACK-BOX LMMS WITHOUT LABELS VIA VISUALLY GROUNDED CROSS-MODEL AGREEMENT

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Ranking large multimodal models (LMMs) on a new task typically requires labeled evaluation data, yet in practice target data may be unlabeled and candidate models may be accessible only as black boxes. Existing unsupervised ranking methods commonly rely on model internals, labeled proxy data, or repeated generations, which are costly and can favor models that are consistently wrong. We ask whether candidate LMMs can instead provide supervision for ranking one another. We introduce Cross-Model Agreement with Visual Evidence for Ranking (MAVER), a label-free and training-free framework for LMM ranking. Given one response from each candidate model, MAVER constructs an NLI-weighted response graph and characterizes agreement using semantic-cluster prevalence and within-cluster connectivity. However, agreement alone can be misleading when multiple models share the same language prior and converge on the same incorrect answer. MAVER therefore intervenes on the image while keeping the question fixed, and measures the resulting semantic response change to quantify how strongly each response depends on the visual input. This visual-dependence signal is used to refine the support assigned to each semantic answer, while cross-model semantic entropy gives greater weight to image–question pairs that better distinguish the candidate models. Across 32 LMMs and 15 datasets, MAVER outperforms 14 comparison methods in average ranking performance, showing that cross-model agreement becomes a reliable signal for label-free LMM ranking when complemented by evidence of visual dependence.

## 1 INTRODUCTION

The rapid growth of large multimodal models (LMMs) has created a practical model-selection problem: given a new task and a pool of candidate LMMs, which model should be deployed? No single LMM performs best across all tasks, and the relative strengths of different models can vary substantially with the target data. Reliable model selection therefore requires estimating each candidate's performance on the task of interest. In practice, however, this is difficult because target data are often unlabeled, while many strong commercial LMMs are accessible only as black boxes. Collecting task-specific annotations for every new deployment is costly, and black-box models do not expose logits, token probabilities, or internal representations. These constraints motivate unsupervised LMM performance estimation, which aims to rank candidate models directly on unlabeled target data using only information available from their outputs.

Existing approaches only partially address this setting. Probability-based methods estimate uncertainty from model output distributions (Tu et al., 2025), but cannot be applied when candidate models expose only decoded responses. Sampling-based methods avoid access to model internals by repeatedly querying each model and measuring self-consistency (Lin et al., 2024; Lyu et al., 2025). This comes at substantial inference cost and, more importantly, consistency is not equivalent to correctness. A model affected by systematic bias or reasoning-path collapse may repeatedly produce the same incorrect answer and therefore appear highly reliable (Dang et al., 2025; Gai et al., 2026; Tan et al., 2025). The central challenge is therefore to obtain a reliable ranking signal from black-box responses without relying on repeated generation or assuming that consistency implies correctness. We ask: **Can candidate LMMs provide useful supervision for ranking one another on unlabeled target data?**

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.

Our key idea is to exploit *cross-model agreement*. Instead of repeatedly sampling responses from each model, we compare one response from every candidate model on the same input. When model errors are not perfectly aligned, a reliable response is more likely to receive semantic support from the candidate-model population. Based on this intuition, we introduce Cross-Model Agreement with Visual Evidence for Ranking (MAVER), a label-free and training-free framework for unsupervised LMM ranking. MAVER constructs a natural language inference (NLI)-weighted response graph from one response per candidate model (Lin et al., 2024) and organizes the responses into semantic clusters (Kuhn et al., 2023). It characterizes cross-model agreement at two complementary levels. The cluster proportion measures how many candidate models support the same semantic answer, while the within-cluster degree measures how strongly an individual response agrees with the other responses in that cluster.

Cross-model agreement alone, however, is not always reliable. Candidate LMMs often share similar language priors and may converge on the same incorrect answer (Goyal et al., 2017; Deng et al., 2025). In such cases, strong agreement can reflect a shared prior rather than genuine use of the image. Existing approaches address related language-bias problems using counterfactual prediction (Niu et al., 2021) or vision-aware uncertainty estimation (Park et al., 2026), but these methods require model internals or output probabilities that may be unavailable for black-box LMMs. MAVER instead intervenes directly on the visual input while keeping the question fixed, and compares the original and intervened responses. The resulting semantic change measures how strongly the response depends on the visual input. We use this response-level evidence of visual dependence to refine the support assigned to each semantic cluster, so that a semantic answer receives greater support when it is backed by both more candidate models and stronger dependence on the image.

Finally, different examples provide different amounts of information for ranking. If all candidate models produce the same semantic answer, the example offers little evidence for distinguishing their relative performance. In contrast, examples with greater cross-model disagreement reveal larger differences among the candidates. MAVER therefore uses cross-model semantic entropy to assign greater weight to such examples when aggregating response-level scores into model-level ranking scores. We evaluate MAVER on 32 LMMs spanning diverse model families and scales across 15 datasets. MAVER achieves the highest average ranking performance among the evaluated methods and exhibits a strong correlation between its estimated scores and true model performance. Further analyses show that both visual-dependence weighting and semantic-entropy weighting improve ranking accuracy, with visual dependence providing larger gains when cross-model agreement alone is less reliable. MAVER also remains stable across a broad range of values for the visual-evidence coefficient. Our contributions are twofold: (1) We introduce MAVER, a black-box framework for unsupervised LMM ranking that combines cross-model agreement with visual evidence and emphasizes image–question pairs with greater cross-model disagreement. (2) Experiments and analyses on 32 LMMs and 15 datasets show that MAVER achieves higher average rank correlations than all 14 comparison methods.

## 2 RELATED WORK

**Unsupervised model ranking.** Unsupervised model evaluation estimates model performance or ranks candidate models on unlabeled target data (Deng et al., 2023; Xie et al., 2023; 2024; Tu et al., 2024; Xie et al., 2025). For language models, Dhurandhar et al. (2024) use triplet-based peer evaluation, and Zhang et al. (2025a) use reasoning consistency and consensus-based refinement. For multimodal models, UPME generates questions from unlabeled images and uses peer review to estimate model performance (Zhang et al., 2025b). On fixed unlabeled visual question answering (VQA) data, Tu et al. (2025) evaluate AoL (Miller et al., 2021) and ATC (Garg et al., 2022), which use known model performance on other datasets. They also evaluate NLL and Ent based on output probabilities (Malinin & Gales, 2021). Their strongest estimators require these probabilities, which closed-source LMMs may not expose. Sampling-based self-consistency is also evaluated by comparing repeated responses from each model. Related methods estimate uncertainty or select candidates through repeated sampling or input perturbations, using model likelihoods (Kuhn et al., 2023) or only generated responses (Lin et al., 2024; Lyu et al., 2025; Huy et al., 2026; Choi & Li, 2026).

**Cross-model disagreement.** Self-consistency uses repeated responses from the same model and can remain high even when the answers are incorrect (Tan et al., 2025). Hamidieh et al. (2026)

relate within-model inconsistency to aleatoric uncertainty and cross-model semantic disagreement to total uncertainty, which also includes epistemic uncertainty. Kruse et al. (2025) quantify disagreement among LLMs with Jensen–Shannon divergence, and Liao et al. (2026) use cross-model semantic divergence to audit VLM behavior. Gorbett & Jana (2026) use the token-level perplexity or entropy assigned by a separate verifier model as a label-free correctness signal. Sun et al. (2024) rank hallucination robustness by checking target outputs against evidence generated or judgments produced by peer models. We use cross-model semantic agreement to estimate the relative task performance of candidate LMMs on a fixed unlabeled VQA dataset. For each image–question pair, we compute this agreement from one original response per candidate LMM.

**Language bias and visual grounding in VQA.** VQA requires answering questions based on image content (Antol et al., 2015), but models can exploit language priors without sufficient visual grounding (Goyal et al., 2017; Deng et al., 2025). Similar modality neglect also occurs in LMM-based evaluators (Lee et al., 2026). Niu et al. (2021) reduce language bias by subtracting the question's direct causal effect on the answer during inference. Leng et al. (2024) use Visual Contrastive Decoding to contrast output distributions from original and distorted images, reducing reliance on unimodal priors during decoding. For self-evaluation, Park et al. (2026) quantify how much an image reduces predictive uncertainty and use core-region masking to emphasize salient visual evidence in the resulting score.

## 3 Problem Formulation

**Ranking black-box LMMs without labels.** Let $\mathcal{D} = \{(v_i, q_i)\}_{i=1}^{N}$ be an unlabeled VQA dataset, where $v_i$ denotes an image and $q_i$ its associated question. We consider a set of $M \geq 2$ candidate LMMs, denoted by $\mathcal{F} = \{f_m\}_{m=1}^{M}$. Given an image–question pair $(v_i, q_i)$, model $f_m$ generates a natural-language response $r_i^m = f_m(v_i, q_i)$. We study a black-box setting in which the ranking method can query each model with image–question pairs and observe only its generated responses, without access to model parameters, internal representations, logits, or token probabilities. Ground-truth labels and reference answers are also unavailable during ranking. Based solely on the observed responses, the method assigns each candidate model a score $S^m$ and produces a ranking $\hat{\pi}$ such that $S^{\hat{\pi}_1} \geq S^{\hat{\pi}_2} \geq \cdots \geq S^{\hat{\pi}_M}$. The goal is to recover the relative performance ordering of the candidate LMMs without using labeled data or model-internal information. To define the ground-truth model ranking, let $\mathrm{Eval}_{\mathcal{D}}$ denote the task-specific evaluator associated with dataset $\mathcal{D}$. Applying this evaluator to the responses of model $f_m$ gives its true task performance: $A^m = \mathrm{Eval}_{\mathcal{D}}\left(\{r_i^m\}_{i=1}^{N}\right)$. Sorting $\{A^m\}_{m=1}^{M}$ in descending order yields the reference ranking $\pi^*$, where $A^{\pi_1^*} \geq A^{\pi_2^*} \geq \cdots \geq A^{\pi_M^*}$. The scores $\{A^m\}_{m=1}^{M}$ and the reference ranking $\pi^*$ are hidden from the ranking method and are used only to assess the quality of its predicted ranking. We measure the agreement between $\hat{\pi}$ and $\pi^*$ using Spearman's rank correlation $\rho$ and weighted Kendall's rank correlation $\tau_w$, as defined in Appendix A. Spearman's $\rho$ measures agreement over the full ranking, whereas weighted Kendall's $\tau_w$ places greater emphasis on pairwise ordering errors involving higher-ranked models.

**Motivation for cross-model agreement.** Self-consistency measures agreement among multiple responses generated by the same model for the same input. While useful for characterizing response stability, it does not in general identify model quality. A model that repeatedly produces the same incorrect answer can be just as self-consistent as a model that repeatedly produces the correct one. Thus, self-consistency alone need not preserve the performance ordering of candidate models.

Cross-model agreement provides a different source of information by comparing a model's response with those produced by other candidate models. This perspective is related to the uncertainty decomposition of Hamidieh et al. (2026), who show that expected semantic disagreement among responses from the same model reflects aleatoric uncertainty, while expected disagreement across models reflects total uncertainty. Motivated by this result, we study cross-model agreement through Degree, defined as the average semantic similarity between a model's response and those of the other candidates. A lower Degree indicates greater semantic disagreement with the candidate-model population. Thus, when one response is obtained from each model, the complement of Degree captures the cross-model disagreement that is associated with total uncertainty.

More importantly for model ranking, cross-model agreement can contain information about relative accuracy that self-consistency does not. Intuitively, when candidate models are more likely than not to produce correct responses and their errors are not perfectly aligned, a more accurate model

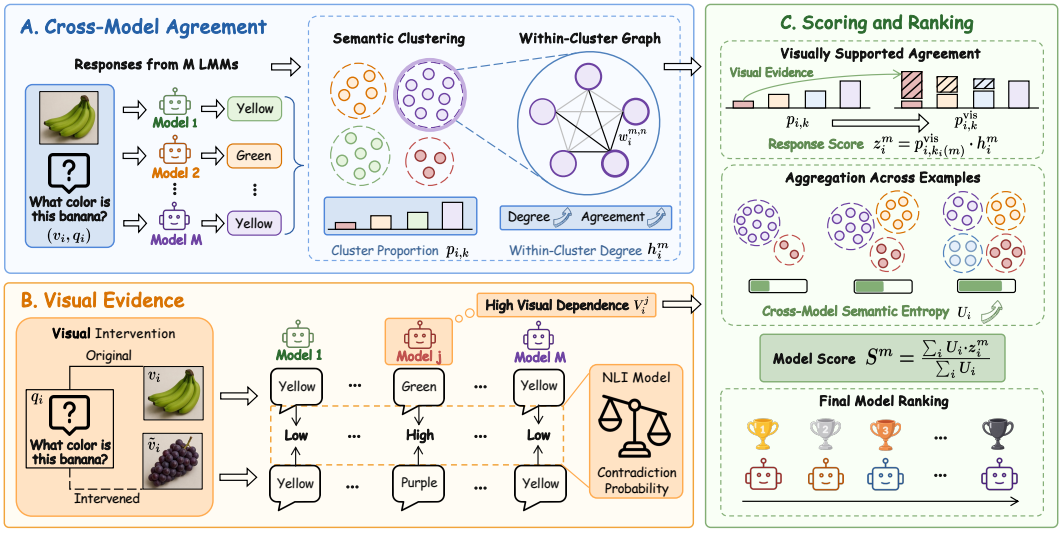



Figure 1: MAVER framework. (A) Cross-model agreement: pairwise NLI comparisons organize candidate responses into semantic clusters, where the cluster proportion $p_{i,k}$ measures support for each semantic answer and the normalized within-cluster degree $h_i^m$ measures response-level agreement. (B) Visual evidence: the image is intervened on while the question remains fixed, and semantic contradiction between the original and intervened responses yields $V_i^m$, measuring how strongly each response depends on the visual input. (C) Scoring and ranking: MAVER uses $V_i^m$ to refine cluster support into $p_{i,k}^{\text{vis}}$, combines it with $h_i^m$ to score each response, and uses cross-model semantic entropy $U_i$ to emphasize more discriminative examples when aggregating response scores into the final model score $S^m$.

is more likely to agree with the response population because correct responses receive support from other accurate models. This intuition can be formalized as follows.

**Proposition 1** (Expected ranking property of Degree)**.** *Consider at least three candidate models. Under the conditions in Appendix B, if one model has higher accuracy than another, then it also has higher expected Degree.*

The proposition provides a sufficient condition under which cross-model agreement preserves the accuracy ordering of candidate models in expectation. Appendix B gives the corresponding assumptions and proof, together with a counterexample showing that self-consistency does not generally enjoy this ranking property. Combined with the empirical results in Section 5.2 and Appendix B.3, these results provide both theoretical and empirical motivation for using cross-model agreement.

## 4 CROSS-MODEL AGREEMENT WITH VISUAL EVIDENCE FOR RANKING

**Method overview.** Our goal is to rank candidate LMMs on an unlabeled evaluation set using only their decoded responses. Since response correctness cannot be observed directly, MAVER infers reliability from two complementary signals: cross-model semantic support and dependence on the visual input. Figure 1 illustrates the three components of MAVER: **cross-model agreement**, **visual evidence**, and **scoring and ranking**. The cluster proportion $p_{i,k}$ measures how many models support the same semantic answer, while the normalized within-cluster degree $h_i^m$ measures how strongly an individual response agrees with others in that cluster. Cross-model agreement alone can be misleading because candidate LMMs may share similar language priors and agree on the same incorrect answer. We therefore intervene on the image while keeping the question fixed, and use the resulting semantic response change to obtain $V_i^m$, which measures how strongly the response depends on the visual input. MAVER uses $V_i^m$ to refine the support assigned to each semantic cluster, producing $p_{i,k}^{\text{vis}}$, which is combined with $h_i^m$ to score each response. Finally, normalized cross-model semantic entropy $U_i$ gives greater weight to examples with stronger model disagreement when aggregating response scores into the final model score $S^m$.

### 4.1 CROSS-MODEL AGREEMENT

For each image and question pair, we obtain one response from every candidate LMM. Our intuition is that, when model errors are not perfectly aligned, a reliable response is more likely to receive semantic support from the candidate-model population. We characterize this support at both the semantic-answer level and the individual-response level.

**Response graph.** Let $[M] = \{1, \dots, M\}$ index the candidate models. For each image–question pair, we construct a complete response graph whose vertex $m \in [M]$ represents response $r_i^m$.

Because semantically equivalent responses can differ substantially in wording, we compare response pairs using a Natural Language Inference (NLI) model (Lin et al., 2024). Let $p_{\mathsf{ent}}(r_i^m, r_i^n) = [\mathrm{softmax}(\boldsymbol{\ell}(r_i^m, r_i^n)/\tau)]_{\mathsf{ent}}$ denote the entailment probability from premise $r_i^m$ to hypothesis $r_i^n$. Here, $\boldsymbol{\ell}(x, y)$ denotes the NLI logits, and $\tau > 0$ is the temperature. Since entailment is directional, we define the edge weight $w_i^{m,n} = \begin{cases} \max[p_{\mathsf{ent}}(r_i^m, r_i^n), p_{\mathsf{ent}}(r_i^n, r_i^m)], & m \neq n, \\ 0, & m = n. \end{cases}$ Larger edge weights indicate stronger semantic agreement between two responses.

**Semantic clustering.** Pairwise similarities capture local relations between responses, but ranking also requires knowing how the candidate-model population is distributed among competing semantic answers. We therefore organize the responses into non-overlapping semantic clusters.

Exact duplicate responses are first assigned to the same cluster. A pair of distinct responses is considered compatible only when neither directional NLI comparison indicates a contradiction. This yields $K_i$ semantic clusters, represented by index sets $\mathcal{C}_{i,k} \subseteq [M]$, for $k = 1, \dots, K_i$. We write $k_i(m)$ for the cluster containing response $r_i^m$. The proportion of models in cluster $k$ is

$$p_{i,k} = \frac{|\mathcal{C}_{i,k}|}{M}. \tag{1}$$

Thus, it measures how much support the candidate-model population assigns to semantic answer $k$.

**Within-cluster agreement.** The cluster proportion assigns the same support to all responses in a cluster, although these responses may differ in how closely they agree with one another. We therefore use weighted within-cluster degree to obtain a response-specific measure of semantic support.

For $m \in \mathcal{C}_{i,k}$, $d_{i,\mathrm{in}}^m = \sum_{n \in \mathcal{C}_{i,k},\, n \neq m} w_i^{m,n}$. We normalize degrees by the cluster maximum to obtain $h_i^m$ ($h_i^m = 1$ for singleton clusters). A larger $h_i^m$ indicates stronger semantic agreement with the other responses in the same cluster.

**Agreement score.** We combine cluster-level support and response-level agreement as

$$z_{i,*}^m = p_{i,k_i(m)} \cdot h_i^m. \tag{2}$$

The first factor measures support for the semantic answer, while the second measures support for the individual response within that answer group. This score serves as the response-level signal for the agreement-only formulation, denoted MAVER*.

### 4.2 VISUAL EVIDENCE

Cross-model agreement indicates whether a response is supported by other candidate models, but it does not reveal whether that support is grounded in the image. Candidate LMMs may share similar pretraining distributions or language priors, causing several models to agree on the same incorrect answer. For example, when asked the color of the green bananas in Figure 1, most candidate LMMs answer "yellow," while fewer answer "green." The answer "yellow" therefore receives strong cross-model support even though it conflicts with the image. To distinguish visually dependent responses from those driven primarily by shared priors, we ask whether the response changes when the image changes while the question remains fixed. Niu et al. (2021) reduce language bias through factual and counterfactual comparisons, but their method requires internal model branches and branch-level output scores. We instead obtain a related signal directly from decoded responses.

For each image and question pair, let $\tilde{v}_i$ denote the intervened image while keeping $q_i$ unchanged. Candidate model $f_m$ produces $\tilde{r}_i^m = f_m(\tilde{v}_i, q_i)$. We measure the semantic conflict between the

original response $r_i^m$ and the intervened response $\tilde{r}_i^m$ using the same NLI model. Let $p_{\mathsf{con}}(x, y) = [\mathrm{softmax}(\boldsymbol{\ell}(x, y))]_{\mathsf{con}}$ denote the contradiction probability. We define

$$V_i^m = \max[p_{\mathsf{con}}(r_i^m, \tilde{r}_i^m), p_{\mathsf{con}}(\tilde{r}_i^m, r_i^m)]\,. \tag{3}$$

A larger $V_i^m$ indicates a stronger semantic change after modifying the visual input, and therefore stronger evidence that the response depends on the image. Importantly, $V_i^m$ does not measure correctness directly. It instead complements cross-model agreement by indicating whether a response is sensitive to the visual input.

### 4.3 Scoring and ranking

The cluster proportion $p_{i,k}$ measures how many models support semantic answer $k$, while $V_i^m$ measures how strongly each supporting response depends on the visual input. We combine these signals at the cluster level so that a semantic answer receives greater support when it is backed by both more candidate models and stronger visual dependence.

**Cluster weights.** For each response, we define $g_i^m = 1 + \lambda \cdot V_i^m$, where $\lambda \geq 0$ controls the contribution of visual dependence. The constant term retains the contribution of every response, while the second term increases its contribution when stronger visual dependence is observed. We sum these weights within each semantic cluster and normalize across clusters:

$$p_{i,k}^{\mathrm{vis}} = \frac{|\mathcal{C}_{i,k}| + \lambda \cdot \sum_{m \in \mathcal{C}_{i,k}} V_i^m}{M + \lambda \cdot \sum_{n=1}^{M} V_i^n}. \tag{4}$$

The quantity $p_{i,k}^{\mathrm{vis}}$ represents cross-model agreement after incorporating evidence of visual dependence. When $\lambda = 0$, it reduces to the original cluster proportion $p_{i,k}$.

**Visually supported agreement.** For model $f_m$, we define

$$z_i^m = p_{i,k_i(m)}^{\mathrm{vis}} \cdot h_i^m. \tag{5}$$

The first factor measures support for the semantic answer after incorporating visual dependence, while the second retains response-specific agreement within the cluster.

**Disagreement-aware aggregation.** The score $z_i^m$ evaluates a model on an individual example, but examples differ in how informative they are for ranking. If all candidate models produce the same semantic answer, the example provides little evidence for distinguishing their relative performance. In contrast, examples on which the models produce competing semantic answers reveal larger differences among the candidates. We quantify this disagreement using normalized semantic entropy over the cluster proportions (Kuhn et al., 2023):

$$U_i = -\frac{1}{\log M} \sum_{k=1}^{K_i} p_{i,k} \log p_{i,k}. \tag{6}$$

The entropy is zero when all responses belong to one semantic cluster and increases as the responses spread across multiple clusters of comparable size. We aggregate the example-level scores as

$$S^m = \frac{\sum_{i=1}^{N} U_i \cdot z_i^m}{\sum_{i=1}^{N} U_i}. \tag{7}$$

Thus, examples with greater cross-model disagreement contribute more strongly to the final ranking. If all semantic-entropy weights are zero, MAVER uses uniform aggregation. The agreement-only variant MAVER* applies the same aggregation rule using $z_{i,*}^m$. Candidate LMMs are ranked in descending order of $S^m$.

## 5 Experiments

### 5.1 Experimental setup

**Datasets and candidate models.** Our experiments use VLMEvalKit (Duan et al., 2024) and evaluate the same nine datasets and 32 candidate LMMs as Tu et al. (2025). The broader evaluation

Table 1: We compare unsupervised LMM ranking methods on nine datasets. We reproduce AoL and ATC and cite the NLL, Ent, and sampling-based results from Tu et al. (2025). MCVQ denotes multiple-choice VQA. Best and second-best results are bolded and underlined, respectively.

| Method | MCVQ | | | | | | | | VQA | | | | | | | | | | Overall | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SQA-I | | AI2D | | RWQA | | MMMU | | GQA | | ChartQA | | OCRVQA | | TextVQA | | DocVQA | | Average | |
| | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ |
| AoL | 0.55 | 0.50 | 0.74 | 0.67 | 0.73 | 0.63 | 0.55 | 0.52 | 0.55 | 0.29 | 0.73 | 0.66 | 0.27 | 0.17 | 0.71 | 0.61 | 0.72 | 0.65 | 0.62 | 0.52 |
| $\mathsf{NLL_F}$ | 0.83 | 0.78 | 0.84 | 0.76 | 0.56 | 0.58 | 0.60 | **0.59** | 0.71 | 0.57 | 0.63 | 0.41 | 0.78 | 0.64 | 0.74 | 0.63 | 0.84 | 0.74 | 0.73 | 0.63 |
| $\mathsf{NLL_P}$ | 0.64 | 0.60 | 0.61 | 0.58 | 0.25 | 0.37 | 0.16 | 0.18 | 0.58 | 0.52 | 0.79 | 0.67 | 0.81 | 0.68 | 0.71 | 0.67 | 0.88 | 0.78 | 0.60 | 0.56 |
| $\mathsf{NLL_{max}}$ | 0.80 | 0.76 | 0.91 | 0.81 | 0.63 | 0.63 | 0.49 | 0.44 | 0.72 | 0.59 | 0.94 | 0.80 | 0.64 | 0.63 | 0.83 | 0.69 | 0.92 | 0.82 | 0.76 | 0.69 |
| $\mathsf{NLL_{avg}}$ | 0.72 | 0.64 | 0.88 | 0.73 | 0.64 | 0.56 | 0.50 | 0.40 | 0.67 | 0.55 | 0.92 | 0.75 | 0.81 | 0.65 | 0.81 | 0.72 | 0.93 | 0.82 | 0.76 | 0.65 |
| $\mathsf{Ent_F}$ | 0.59 | 0.51 | 0.82 | 0.59 | 0.65 | 0.56 | 0.64 | 0.52 | 0.54 | 0.20 | 0.64 | 0.45 | 0.69 | 0.57 | 0.71 | 0.56 | 0.80 | 0.70 | 0.68 | 0.52 |
| $\mathsf{Ent_P}$ | 0.49 | 0.39 | 0.69 | 0.48 | 0.43 | 0.43 | 0.34 | 0.24 | 0.46 | 0.21 | 0.82 | 0.64 | 0.80 | 0.64 | 0.70 | 0.54 | 0.86 | 0.68 | 0.62 | 0.47 |
| $\mathsf{Ent_{max}}$ | 0.58 | 0.39 | 0.82 | 0.59 | 0.67 | 0.60 | 0.66 | 0.54 | 0.54 | 0.21 | 0.88 | 0.66 | 0.53 | 0.52 | 0.76 | 0.60 | 0.87 | 0.73 | 0.70 | 0.54 |
| $\mathsf{Ent_{avg}}$ | 0.58 | 0.33 | 0.80 | 0.56 | 0.62 | 0.50 | 0.59 | 0.41 | 0.57 | 0.26 | 0.91 | 0.68 | 0.77 | 0.62 | 0.74 | 0.58 | 0.88 | 0.74 | 0.72 | 0.52 |
| $\mathsf{Sample_{BLEU}}$ | 0.65 | 0.68 | 0.76 | 0.59 | 0.48 | 0.36 | 0.37 | 0.12 | 0.44 | 0.41 | 0.89 | 0.61 | 0.47 | 0.58 | 0.81 | 0.60 | 0.90 | 0.63 | 0.64 | 0.51 |
| $\mathsf{Sample_{BERT}}$ | 0.46 | 0.52 | 0.70 | 0.55 | 0.52 | 0.28 | 0.45 | 0.44 | 0.51 | 0.47 | 0.90 | 0.61 | 0.75 | 0.72 | 0.77 | 0.59 | 0.89 | 0.64 | 0.66 | 0.54 |
| $\mathsf{Sample^*_{BERT}}$ | 0.67 | 0.56 | 0.78 | 0.62 | 0.67 | 0.48 | 0.59 | 0.39 | 0.51 | 0.47 | 0.90 | 0.61 | 0.75 | 0.72 | 0.77 | 0.59 | 0.89 | 0.64 | 0.73 | 0.56 |
| ATC | 0.67 | 0.69 | 0.71 | 0.56 | 0.52 | 0.32 | 0.33 | 0.27 | 0.62 | 0.47 | 0.94 | 0.80 | 0.43 | 0.38 | 0.80 | 0.64 | 0.90 | 0.73 | 0.66 | 0.54 |
| Degree | **0.96** | 0.93 | 0.98 | 0.91 | 0.93 | 0.71 | 0.59 | 0.29 | 0.89 | 0.82 | **0.99** | 0.89 | 0.90 | 0.72 | **0.98** | **0.92** | **0.99** | **0.91** | 0.91 | 0.79 |
| MAVER* | **0.96** | **0.94** | **0.99** | 0.93 | 0.93 | 0.69 | 0.82 | 0.46 | **0.91** | **0.85** | **0.99** | **0.90** | 0.91 | 0.74 | **0.98** | 0.90 | **0.99** | 0.90 | 0.94 | 0.81 |
| MAVER | **0.96** | 0.91 | **0.99** | **0.94** | **0.97** | **0.86** | **0.86** | **0.59** | 0.90 | 0.75 | **0.99** | **0.90** | **0.92** | **0.83** | **0.98** | 0.91 | 0.98 | 0.90 | **0.95** | **0.84** |

covers MMMU-Pro (Yue et al., 2025), MathVista (Lu et al., 2024a), MMVet-Hard (Yang et al., 2025), WildVision (Lu et al., 2024b), CharXiv (Wang et al., 2024b), and OCRBench v2 (Fu et al., 2025). Appendices C.1 and C.2 list all evaluation files and model identifiers.

**Evaluation protocol.** For each dataset, we rank the candidate LMMs by their task-specific scores to obtain the reference ranking. We use exact matching for SQA-I, AI2D, RWQA, and MMMU and the official rule-based evaluators for GQA, ChartQA, OCRVQA, TextVQA, DocVQA, and OCRBench v2. We use Qwen3.5-9B to evaluate MMMU-Pro, MathVista, MMVet-Hard, WildVision, and CharXiv. We report Spearman's $\rho$ and weighted Kendall's $\tau_w$ for each dataset and their averages.

**Comparison methods.** We adopt the comparison methods evaluated by Tu et al. (2025). AoL (Miller et al., 2021) ranks models on the target dataset using their known performance on each other evaluated dataset. ATC (Garg et al., 2022) calibrates uncertainty thresholds using model performance on other datasets to estimate performance on the target dataset. The eight output-probability variants compute negative log-likelihood or predictive entropy at different token positions or across the generated sequence (Malinin & Gales, 2021). The three sampling-based variants measure consistency between the original response and repeated generations using BLEU (Papineni et al., 2002) or BERTScore (Zhang et al., 2020). Appendix C.3 details these comparison methods. We additionally include Degree, which ranks models by weighted node degree in the full cross-model response graph. MAVER* uses the agreement-only response score in Equation (2). MAVER incorporates visual evidence into the response score, as defined in Equation (5).

**Implementation details.** We use FineCat for all NLI comparisons and set the response-graph temperature to $\tau = 3$ following Lin et al. (2024). For each visual intervention, we keep the question fixed and draw an image from the full dataset image pool with replacement. We report MAVER with a fixed visual-evidence coefficient of $\lambda = 5$ on all datasets. Appendix C.4 provides the remaining clustering and intervention settings.

### 5.2 MAIN RESULTS

**Cross-model agreement outperforms the existing comparison methods.** On the nine datasets in Table 1, $\mathsf{NLL_{max}}$ and $\mathsf{NLL_{avg}}$ both obtain an average $\rho$ of 0.76, the highest among the existing

Table 2: We compare unsupervised LMM ranking methods in the broader evaluation and reproduce the comparison methods evaluated by Tu et al. (2025). Metrics and formatting follow Table 1.

| Method | MCVQ | | VQA | | | | | | | | | | Overall | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MMMU-Pro | | MathVista | | MMVet-Hard | | WildVision | | CharXiv | | OCRBench v2 | | Average | |
| | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ |
| AoL | 0.41 | 0.46 | 0.75 | **0.67** | 0.64 | 0.60 | 0.60 | 0.54 | 0.70 | 0.64 | 0.71 | 0.62 | 0.63 | 0.59 |
| $\mathsf{NLL_F}$ | 0.24 | 0.25 | 0.20 | 0.24 | 0.39 | 0.30 | 0.69 | 0.42 | 0.58 | 0.35 | 0.58 | 0.31 | 0.45 | 0.31 |
| $\mathsf{NLL_P}$ | 0.00 | 0.12 | 0.50 | 0.48 | 0.51 | 0.47 | 0.48 | 0.35 | 0.70 | 0.41 | 0.63 | 0.45 | 0.47 | 0.38 |
| $\mathsf{NLL_{max}}$ | 0.43 | 0.40 | -0.06 | 0.02 | -0.08 | 0.03 | -0.31 | -0.36 | 0.57 | 0.50 | 0.74 | 0.71 | 0.21 | 0.22 |
| $\mathsf{NLL_{avg}}$ | 0.22 | 0.29 | 0.57 | 0.57 | 0.53 | 0.55 | 0.43 | 0.24 | 0.66 | 0.37 | 0.69 | 0.48 | 0.52 | 0.42 |
| $\mathsf{Ent_F}$ | 0.55 | **0.52** | 0.24 | 0.25 | 0.38 | 0.33 | 0.73 | 0.44 | 0.62 | 0.44 | 0.55 | 0.30 | 0.51 | 0.38 |
| $\mathsf{Ent_P}$ | 0.25 | 0.26 | 0.67 | 0.51 | 0.56 | 0.53 | 0.58 | 0.42 | 0.76 | 0.46 | 0.68 | 0.46 | 0.58 | 0.44 |
| $\mathsf{Ent_{max}}$ | 0.56 | **0.52** | 0.15 | 0.28 | 0.22 | 0.42 | 0.29 | 0.20 | 0.69 | 0.53 | 0.81 | 0.68 | 0.45 | 0.44 |
| $\mathsf{Ent_{avg}}$ | 0.56 | 0.50 | 0.56 | 0.55 | 0.61 | 0.62 | 0.60 | 0.40 | 0.73 | 0.46 | 0.67 | 0.44 | 0.62 | 0.49 |
| $\mathsf{Sample_{BLEU}}$ | 0.47 | 0.42 | 0.16 | 0.36 | -0.06 | 0.04 | -0.02 | -0.13 | 0.50 | 0.44 | 0.61 | 0.58 | 0.28 | 0.29 |
| $\mathsf{Sample_{BERT}}$ | 0.40 | 0.43 | -0.05 | 0.23 | -0.23 | -0.02 | -0.39 | -0.45 | 0.48 | 0.46 | 0.56 | 0.60 | 0.13 | 0.21 |
| $\mathsf{Sample^*_{BERT}}$ | 0.37 | 0.40 | -0.07 | 0.21 | -0.23 | -0.02 | -0.39 | -0.45 | 0.48 | 0.46 | 0.56 | 0.60 | 0.12 | 0.20 |
| ATC | 0.41 | 0.29 | 0.14 | 0.13 | -0.28 | -0.12 | -0.33 | -0.35 | 0.47 | 0.41 | 0.61 | 0.58 | 0.17 | 0.16 |
| Degree | 0.41 | 0.24 | 0.48 | 0.40 | 0.80 | 0.64 | 0.63 | 0.60 | 0.88 | 0.63 | 0.93 | 0.79 | 0.69 | 0.55 |
| MAVER* | 0.54 | 0.26 | 0.66 | 0.53 | 0.82 | 0.62 | **0.81** | **0.72** | 0.92 | 0.74 | 0.95 | 0.86 | 0.78 | 0.62 |
| MAVER | **0.58** | 0.34 | **0.77** | 0.60 | **0.84** | **0.65** | **0.81** | 0.70 | **0.93** | **0.75** | **0.96** | **0.90** | **0.82** | **0.66** |

methods. $\mathsf{NLL_{max}}$ gives the highest average $\tau_w$ at $0.69$. Degree raises the average $\rho$ to $0.91$ and the average $\tau_w$ to $0.79$. MAVER* further raises them to $0.94$ and $0.81$. Degree exceeds $\mathsf{NLL_{max}}$ in $\rho$ on all nine datasets and in $\tau_w$ on eight. The largest improvement of MAVER* over Degree occurs on MMMU, where $\rho$ increases from $0.59$ to $0.82$ and $\tau_w$ from $0.29$ to $0.46$.

In the broader evaluation of six datasets reported in Table 2, AoL gives the strongest reproduced comparison results, with average correlations of $0.63$ for $\rho$ and $0.59$ for $\tau_w$. Degree obtains an average $\rho$ of $0.69$ and a $\tau_w$ of $0.55$. MAVER* raises these averages to $0.78$ and $0.62$, exceeding Degree in $\rho$ on all six datasets and in $\tau_w$ on five. These averages exceed those of every reproduced comparison method. $\mathsf{NLL_{max}}$ produces negative $\rho$ values on MathVista, MMVet-Hard, and WildVision. All three sampling-based methods also produce negative $\rho$ values on MMVet-Hard and WildVision.

**Visual evidence further improves ranking correlations.** On the nine datasets, MAVER achieves the highest average correlations, with $\rho = 0.95$ and $\tau_w = 0.84$. The clearest gains occur on RWQA and MMMU. Compared with MAVER*, $\tau_w$ increases by $0.17$ and $0.13$, respectively, and $\rho$ increases by $0.04$ on each dataset. GQA is the main exception. MAVER* and MAVER have similar $\rho$, and MAVER* gives a higher $\tau_w$.

In the broader evaluation, MAVER achieves the highest average correlations, with $\rho = 0.82$ and $\tau_w = 0.66$. It improves both metrics over MAVER* on five of the six datasets. WildVision is the only exception. Its $\rho$ remains at $0.81$, and its $\tau_w$ decreases from $0.72$ to $0.70$. The largest gains are the $0.11$ increase in $\rho$ on MathVista and the $0.08$ increase in $\tau_w$ on MMMU-Pro.

**MAVER scores correlate with model performance.** The left panel of Figure 2 shows strong correlations between MAVER scores and model performance in both score values ($r = 0.95$) and model rankings ($\rho = 0.99$) across the 32 candidate LMMs. We normalize both scores within each dataset and average them across the 15 datasets for each model. The predicted and reference ranks differ by at most one position for 24 of the 32 models. Details are provided in Appendix D.

### 5.3 Analysis and ablations

**When visual evidence helps.** To determine when visual evidence is most useful, we analyze ten datasets with binary response-level evaluation. Largest-cluster reliability is the percentage of pairs for which every largest semantic cluster contains a correct response. An image–question pair is

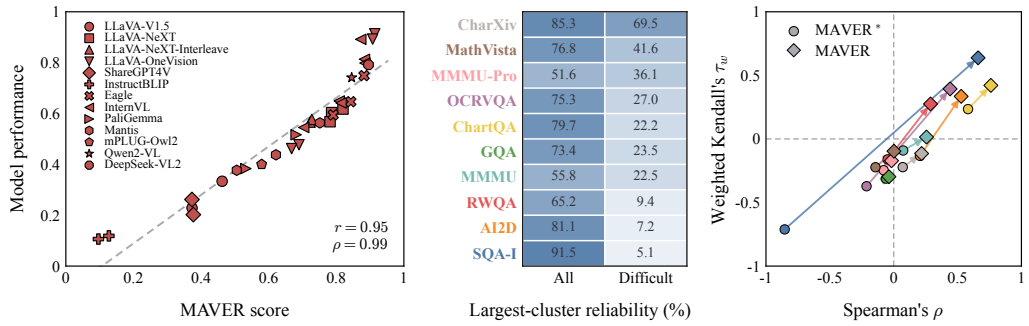



Figure 2: Left: correlation between MAVER scores and model performance across the 32 candidate LMMs. Middle: largest-cluster reliability on all and difficult image–question pairs. Right: changes in rank correlation from MAVER* to MAVER on agreement-failure pairs.

difficult when no more than $25\%$ of the candidate models answer correctly. An agreement failure occurs when at least one model is correct but none of the largest semantic clusters contains a correct response. The middle panel of Figure 2 shows that largest-cluster reliability on difficult pairs falls below $42\%$ on nine of the ten datasets. In the right panel, MAVER achieves higher $\rho$ and $\tau_w$ than MAVER* on the agreement-failure subset of every dataset. The results on low-accuracy MMMU disciplines, free-form MathVista questions, and higher-disagreement groups further show that visual evidence is especially useful when agreement alone is less reliable (Appendices E.1–E.3).

**Module analysis.** We remove visual evidence or replace $U_i$-weighted aggregation with uniform aggregation. In the broader evaluation, entropy weighting raises the average $\rho$ from $0.75$ to $0.82$ and $\tau_w$ from $0.59$ to $0.66$. Both ablations lower the average $\rho$ and $\tau_w$ across all 15 datasets, supporting the contribution of both components to model ranking (Appendix E.4).

**Ranking errors.** We also measure the absolute difference between predicted and reference model ranks. Across the 15 datasets, MAVER achieves an absolute rank error of at most five on $85.0\%$ of model–dataset pairs, compared with $81.9\%$ for MAVER* and $79.6\%$ for Degree. Thus, the higher correlations of MAVER are accompanied by smaller errors in individual model ranks (Appendix E.5).

**Parameter sensitivity.** We vary the visual evidence coefficient $\lambda$. Both average correlations improve over the agreement-only setting for every positive value tested and remain stable for $\lambda \in [2, 20]$. Thus, MAVER does not require a narrowly selected $\lambda$ (Appendix E.6). The overall ranking performance of MAVER also remains stable across visual intervention seeds (Appendix E.7).

**Query cost.** MAVER uses one additional intervened response per model–example pair, compared with five for sampling-based self-consistency. The runtime comparison also shows that MAVER requires less total time than these methods (Appendix F).

## 6 Conclusion and Future Work

We introduce MAVER, a label-free and training-free framework for ranking black-box LMMs using only decoded responses. The key idea is to use candidate models as comparative supervision: cross-model semantic agreement identifies responses supported by the model population, while visual intervention distinguishes image-dependent agreement from agreement driven by shared language priors. Cross-model semantic entropy further emphasizes examples that better separate candidate models. Across 32 LMMs and 15 datasets, MAVER achieves the strongest average ranking performance among the evaluated methods, supporting visually informed cross-model agreement as an effective signal for unsupervised LMM ranking.

Two potential limitations remain. First, cross-model agreement can be less reliable when candidate models share strongly correlated errors, motivating future work on diversity-aware agreement. Second, the visual-dependence signal depends on the quality of the intervention, suggesting more adaptive and task-aware intervention strategies. Future work can also extend MAVER beyond VQA to broader multimodal tasks and multimodal agents.

### AI USE STATEMENT

The authors develop the scientific ideas and claims, design the experiments, finalize the implementations, and write the manuscript. Generative AI tools were used for language polishing and organizing the code for public release. These tools were also used to check the sufficiency of the experiments and the consistency of mathematical derivations. All AI-assisted work was reviewed and verified by the authors before use. The authors remain fully responsible for the content of this paper.

### REPRODUCIBILITY STATEMENT

Section 5.1 and Appendices A and C provide the implementation and evaluation details needed to reproduce our results. Appendix B provides the assumptions and proof of Proposition 1. We will release the code upon acceptance.

## REFERENCES

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *ICCV*, 2015.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. PaliGemma: A versatile 3B VLM for transfer. *arXiv*, 2024.

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 *ECCV*, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024b.

Hyeong Kyu Choi and Sharon Li. ModeX: Evaluator-free best-of-n selection for open-ended generation. In *ACL*, 2026.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023.

Xingyu Dang, Christina Baek, Zico Kolter, and Aditi Raghunathan. Assessing diversity collapse in reasoning. In *ICLR Workshop on Scaling Self-Improving Foundation Models Without Human Supervision*, 2025.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. Words or vision: Do vision-language models have blind faith in text? In *CVPR*, 2025.

Weijian Deng, Yumin Suh, Stephen Gould, and Liang Zheng. Confidence and dispersity speak: Characterizing prediction matrix for unsupervised accuracy estimation. In *ICML*, 2023.

Amit Dhurandhar, Rahul Nair, Moninder Singh, Elizabeth Daly, and Karthikeyan Natesan Ramamurthy. Ranking large language models without ground truth. In *Findings of ACL*, 2024.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. VLMEvalKit: An open-source toolkit for evaluating large multi-modality models. In *ACM MM*, 2024.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, Jingqun Tang, Wei Chen, Lianwen Jin, Yuliang Liu, and

Xiang Bai. OCRBench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. In *NeurIPS*, 2025.

Jingchu Gai, Guanning Zeng, Huaqing Zhang, and Aditi Raghunathan. Differential smoothing mitigates sharpening and improves LLM reasoning. In *ICML*, 2026.

Saurabh Garg, Sivaraman Balakrishnan, Zachary Chase Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *ICLR*, 2022.

Matt Gorbett and Suman Jana. Cross-model disagreement as a label-free correctness signal. In *COLM*, 2026.

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 *CVPR*, 2017.

Kimia Hamidieh, Veronika Thost, Walter Gerych, Mikhail Yurochkin, and Marzyeh Ghassemi. Complementing self-consistency with cross-model disagreement for uncertainty quantification. In *ICLR*, 2026.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.

Ta Duc Huy, Trang Nguyen, Townim Chowdhury, Ankit Yadav, Minh-Son To, Zhibin Liao, Johan W. Verjans, and Vu Minh Hieu Phan. Do vision language models recognize visual ambiguity? In *ECCV*, 2026.

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

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016.

Maya Kruse, Majid Afshar, Saksham Khatwani, Anoop Mayampurath, Guanhua Chen, and Yanjun Gao. Simple yet effective: An information-theoretic approach to multi-LLM uncertainty quantification. In *EMNLP*, 2025.

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

Sua Lee, Sanghee Park, and Jinbae Im. MM-JudgeBias: A benchmark for evaluating compositional biases in MLLM-as-a-judge. In *ACL*, 2026.

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 *CVPR*, 2024.

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. *TMLR*, 2025a.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. LLaVA-Interleave: Tackling multi-image, video, and 3D in large multimodal models. In *ICLR*, 2025b.

Junchi Liao, Jiawen Deng, and Fuji Ren. VISTA: Auditing semantic divergence in vision-language models. *arXiv*, 2026.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *TMLR*, 2024.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. LLaVA project blog, 2024b. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Pan Lu, Swaroop Mishra, Tanglin 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 *NeurIPS*, 2022.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024a.

Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. WildVision: Evaluating vision-language models in the wild with human preferences. In *NeurIPS*, 2024b.

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. Calibrating large language models with sample consistency. In *AAAI*, 2025.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *ICLR*, 2021.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of ACL*, 2022.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for VQA on document images. In *WACV*, 2021.

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 *ICML*, 2021.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. OCR-VQA: Visual question answering by reading text in images. In *ICDAR*, 2019.

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 *CVPR*, 2021.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *ACL*, 2002.

Seongheon Park, Changdae Oh, Hyeong Kyu Choi, Sean Du, and Sharon Li. VAUQ: Vision-aware uncertainty quantification for LVLM self-evaluation. In *Findings of ACL*, 2026.

Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, Yilin Zhao, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, Bryan Catanzaro, Andrew Tao, Jan Kautz, Zhiding Yu, and Guilin Liu. Eagle: Exploring the design space for multimodal LLMs with mixture of encoders. In *ICLR*, 2025.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *CVPR*, 2019.

Guangzhi Sun, Potsawee Manakul, Adian Liusie, Kunat Pipatanakul, Chao Zhang, Phil Woodland, and Mark Gales. CrossCheckGPT: Universal hallucination ranking for multimodal foundation models. In *NeurIPS Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models*, 2024.

Hexiang Tan, Fei Sun, Sha Liu, Du Su, Qi Cao, Xin Chen, Jingang Wang, Xunliang Cai, Yuanzhuo Wang, Huawei Shen, and Xueqi Cheng. Too consistent to detect: A study of self-consistent errors in LLMs. In *EMNLP*, 2025.

Weijie Tu, Weijian Deng, Liang Zheng, and Tom Gedeon. What does softmax probability tell us about classifiers ranking across diverse test conditions? *TMLR*, 2024.

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 *ICML*, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv*, 2024a.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. CharXiv: Charting gaps in realistic chart understanding in multimodal LLMs. In *NeurIPS*, 2024b.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. DeepSeek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv*, 2024.

Renchunzi Xie, Hongxin Wei, Lei Feng, Yuzhou Cao, and Bo An. On the importance of feature separability in predicting out-of-distribution error. In *NeurIPS*, 2023.

Renchunzi Xie, Ambroise Odonnat, Vasilii Feofanov, Weijian Deng, Jianfeng Zhang, and Bo An. MaNo: Exploiting matrix norm for unsupervised accuracy estimation under distribution shifts. In *NeurIPS*, 2024.

Renchunzi Xie, Ambroise Odonnat, Vasilii Feofanov, Ievgen Redko, Jianfeng Zhang, and Bo An. Leveraging gradients for unsupervised accuracy estimation under distribution shift. *TMLR*, 2025.

Yue Yang, Shuibo Zhang, Kaipeng Zhang, Yi Bin, Yu Wang, Ping Luo, and Wenqi Shao. Dynamic multimodal evaluation with flexible complexity by vision-language bootstrapping. In *ICLR*, 2025.

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 *CVPR*, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *CVPR*, 2024.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-pro: A more robust multi-discipline multimodal understanding benchmark. In *ACL*, 2025.

Jiamu Zhang, Jiayi Yuan, Andrew Wen, Hoang Anh Duy Le, Yu-Neng Chuang, Soo-Hyun Choi, Rui Chen, and Xia Hu. ReasonerRank: Redefining language model evaluation with ground-truth-free ranking frameworks. In *Findings of ACL*, 2025a.

Qihui Zhang, Munan Ning, Zheyuan Liu, Yue Huang, Shuo Yang, Yanbo Wang, Jiayi Ye, Xiao Chen, Yibing Song, and Li Yuan. UPME: An unsupervised peer review framework for multimodal large language model evaluation. In *CVPR*, 2025b.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *ICLR*, 2020.

# APPENDICES

## A RANK-CORRELATION METRICS

For each dataset, let $R_m = \mathrm{rank}(A^m)$ and $\widehat{R}_m = \mathrm{rank}(S^m)$. Models tied on a score receive the average of the rank positions they occupy. Spearman's rank correlation is the Pearson correlation between the two rank vectors:

$$\rho = \frac{\sum_{m=1}^{M}(R_m - \bar{R})(\widehat{R}_m - \bar{\widehat{R}})}{\sqrt{\sum_{m=1}^{M}(R_m - \bar{R})^2 \sum_{m=1}^{M}(\widehat{R}_m - \bar{\widehat{R}})^2}}.$$

Weighted Kendall's rank correlation gives greater weight to pairs involving higher-ranked models. A model pair is concordant if the reference and predicted scores order the models in the same way, and discordant if they order them oppositely. For a given importance ranking, a model with zero-based rank $r_m$ receives weight $1/(r_m + 1)$, and each model pair receives the sum of its two model weights. Let $C_w$ and $D_w$ denote the total weights of concordant and discordant pairs. Let $T_{A,w}$ and $T_{S,w}$ denote the total weights of pairs tied only in the reference scores and only in the predicted scores, respectively. Pairs tied in both are excluded. For a given importance ranking, weighted Kendall's rank correlation is

$$\tau_w = \frac{C_w - D_w}{\sqrt{(C_w + D_w + T_{A,w})(C_w + D_w + T_{S,w})}}.$$

We use the default settings of `scipy.stats.weightedtau`. The first importance ranking orders models by decreasing reference score, using predicted scores to break ties. The second reverses the roles of the two scores. We report the average of the two coefficients as $\tau_w$. We compute both $\rho$ and $\tau_w$ separately for each dataset and report their arithmetic means within each evaluation setting.

## B RANKING PROPERTIES OF SELF-CONSISTENCY AND CROSS-MODEL AGREEMENT

Fix a pool of $M \geq 3$ candidate models. Let $X$ be an image–question pair drawn from the target input distribution, and write $r^m(X) = f_m(X)$ for the original response of model $m$. Let $C_m(X) \in \{0, 1\}$ indicate whether this response is correct. The population accuracy of model $m$ is

$$a_m = \Pr_X(C_m(X) = 1)\,.$$

For the Degree analysis, we fix the original response generation rule, so randomness comes only from the input $X$.

### B.1 SELF-CONSISTENCY AND ACCURACY

For a given input $x$, let $Y_m, Y'_m$ be independent responses from model $m$. Given a semantic similarity function satisfying $s(r, r) = 1$, within-model self-consistency is

$$C^m_{\mathrm{self}}(x) = \mathbb{E}[s(Y_m, Y'_m) \mid X = x]\,.$$

The corresponding aleatoric uncertainty and its expectation over the input distribution are

$$U^m_A(x) = 1 - C^m_{\mathrm{self}}(x), \qquad \bar{U}^m_A = \mathbb{E}_X[U^m_A(X)]\,.$$

**Example 1** (Identical self-consistency, different accuracies)**.** Consider a task in which each input has both a correct response and an incorrect response. Model $m$ deterministically produces a correct response to every input, and model $n$ deterministically produces an incorrect response to every input. Thus,

$$a_m = 1, \qquad a_n = 0.$$

Each model's response distribution is concentrated on a single response, so for every input $x$,

$$C^m_{\mathrm{self}}(x) = C^n_{\mathrm{self}}(x) = 1.$$

Consequently,

$$\bar{U}^m_A = \bar{U}^n_A = 0.$$

The models have different accuracies but identical self-consistency and corresponding aleatoric uncertainty scores. These scores therefore cannot generally identify the accuracy ordering of models.

### B.2 Expected ranking property of Degree

Let $w^{m,j}(X)$ denote the edge weight defined in Section 4.1 evaluated at input $X$. The normalized Degree of model $m$ is

$$d^m(X) = \frac{1}{M-1} \sum_{j \neq m} w^{m,j}(X). \tag{8}$$

When the same similarity function is used and the remaining candidate models are weighted uniformly, $1 - d^m(X)$ corresponds to empirical cross-model total uncertainty computed from one response per model.

We analyze the ranking properties of Degree under the following assumptions.

**Assumption 1** (Pairwise independence of correctness)**.** For any distinct models $m, j$, the correctness variables $C_m(X)$ and $C_j(X)$ are independent under the random input $X$.

**Assumption 2** (Common conditional mean similarities)**.** For any distinct models $m, j$ and any correctness combination $b, c \in \{0, 1\}$ with positive probability,

$$\mathbb{E}_X\left[w^{m,j}(X) \mid C_m(X) = b,\ C_j(X) = c\right] = \mu_{bc},$$

where $\mu_{01} = \mu_{10}$ and these conditional means do not vary across model pairs.

Thus, $\mu_{11}$, $\mu_{10}$, and $\mu_{00}$ denote the conditional mean similarities of correct–correct, correct–incorrect, and incorrect–incorrect response pairs, respectively.

For a model pair $m, n$, define the aggregate support difference from the remaining models as

$$B_{mn} = \sum_{j \notin \{m,n\}} \left[a_j(\mu_{11} - \mu_{10}) + (1 - a_j)(\mu_{10} - \mu_{00})\right]. \tag{9}$$

For a given model pair $m, n$, Proposition 1 requires Assumptions 1–2 and the positive support condition $B_{mn} > 0$.

**Proposition 1** (Full version)**.** *Under Assumptions 1–2, for any distinct models $m, n$,*

$$\mathbb{E}_X[d^m(X) - d^n(X)] = \frac{a_m - a_n}{M-1} B_{mn}. \tag{10}$$

*Hence, for any pair satisfying $B_{mn} > 0$,*

$$a_m > a_n \quad \Longrightarrow \quad \mathbb{E}_X[d^m(X)] > \mathbb{E}_X[d^n(X)].$$

*Proof.* For any third model $j \notin \{m, n\}$, the law of total expectation and Assumptions 1–2 give

$$\begin{aligned}\mathbb{E}_X[w^{m,j}(X)] = a_m a_j \mu_{11} + \left[a_m(1 - a_j) + (1 - a_m)a_j\right] \mu_{10} \\ + (1 - a_m)(1 - a_j)\mu_{00}.\end{aligned}$$

Writing the same expression for model $n$ and subtracting yields

$$\mathbb{E}_X\left[w^{m,j}(X) - w^{n,j}(X)\right] = (a_m - a_n)\left[a_j(\mu_{11} - \mu_{10}) + (1 - a_j)(\mu_{10} - \mu_{00})\right].$$

Since the edge weights are symmetric, the shared edge between models $m$ and $n$ cancels in the Degree difference:

$$d^m(X) - d^n(X) = \frac{1}{M-1} \sum_{j \notin \{m,n\}} \left[w^{m,j}(X) - w^{n,j}(X)\right].$$

Taking expectations and substituting the preceding expression gives

$$\mathbb{E}_X[d^m(X) - d^n(X)] = \frac{a_m - a_n}{M-1} B_{mn}.$$

When $B_{mn} > 0$, the expected Degree difference has the same sign as the accuracy difference. □

The condition $B_{mn} > 0$ requires the remaining candidate models to collectively provide more expected semantic support for correct than incorrect responses. When a third model answers correctly, $\mu_{11} - \mu_{10}$ measures its support advantage for correct responses. When it answers incorrectly, $\mu_{10} - \mu_{00}$ measures the corresponding support difference. Higher similarity between incorrect responses can reduce $B_{mn}$.

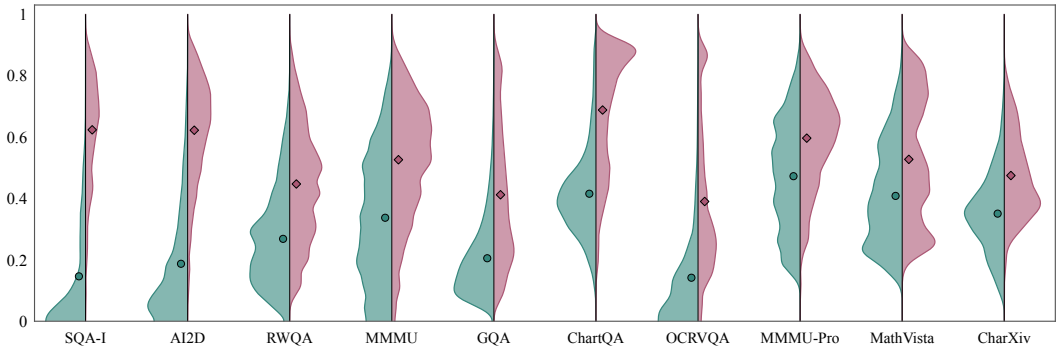



Figure B-1: TU distributions for correct (green) and incorrect (red) responses with zero empirical AU. Circles and diamonds mark the respective group means. The two densities are normalized separately and share a width scale within each dataset.

### B.3 Empirical analysis of self-consistent responses

We examine whether total uncertainty (TU) distinguishes correct from incorrect responses when empirical aleatoric uncertainty (AU) is zero. We use the ten datasets with binary response-level correctness labels and estimate AU and TU using the empirical formulas of Hamidieh et al. (2026). For each image–question pair, we use five sampled responses per model and treat the other 31 candidate models as the auxiliary model set. Both estimates use FineCat to compute semantic similarity, as in the response graph.

We retain model–input pairs with five identical, nonempty mapped sampled answers (empirical AU $= 0$) that match the original answer. We use the original response's evaluation label to determine correctness.

Figure B-1 shows the TU distributions in this subset. Incorrect responses have higher mean TU than correct responses on all ten datasets. These results show that cross-model disagreement provides information about response correctness beyond within-model self-consistency.

## C Experimental setup

We evaluate the nine datasets in the protocol of Tu et al. (2025) and six datasets in the broader evaluation. Both evaluation settings use the same 32 candidate LMMs. Table C-1 lists the evaluation files, and Table C-2 gives the corresponding model identifiers.

### C.1 Datasets

We use the TSV releases distributed through VLMEvalKit (Duan et al., 2024). Table C-1 records the exact evaluation files, with each filename linked to the corresponding download.

### C.2 Candidate LMMs

We evaluate 32 LMMs from 13 model families: LLaVA-V1.5 (Liu et al., 2024a), LLaVA-NeXT (Liu et al., 2024b), LLaVA-NeXT-Interleave (Li et al., 2025b), LLaVA-OneVision (Li et al., 2025a), ShareGPT4V (Chen et al., 2024a), InstructBLIP (Dai et al., 2023), Eagle (Shi et al., 2025), InternVL (Chen et al., 2024b), PaliGemma (Beyer et al., 2024), Mantis (Jiang et al., 2024), mPLUG-Owl2 (Ye et al., 2024), Qwen2-VL (Wang et al., 2024a), and DeepSeek-VL2 (Wu et al., 2024). Table C-2 provides the exact VLMEvalKit identifiers.

### C.3 Comparison methods

We follow the definitions of the comparison methods in Tu et al. (2025). AoL (Miller et al., 2021) ranks models on the target dataset using their known performance on each of the other 14 datasets. NLL measures token-level negative log-likelihood, and Ent measures predictive entropy normalized

Table C-1: Dataset files and corresponding TSV download links. MCVQ denotes multiple-choice VQA, and VQA includes open-ended and mixed-format datasets.

| Dataset | Format | Evaluation file |
|---|---|---|
| SQA-I (Lu et al., 2022) | MCVQ | `ScienceQA_TEST.tsv` |
| AI2D (Kembhavi et al., 2016) | MCVQ | `AI2D_TEST.tsv` |
| RWQA | MCVQ | `RealWorldQA.tsv` |
| MMMU (Yue et al., 2024) | MCVQ | `MMMU_DEV_VAL.tsv` |
| GQA (Hudson & Manning, 2019) | VQA | `GQA_TestDev_Balanced.tsv` |
| ChartQA (Masry et al., 2022) | VQA | `ChartQA_TEST.tsv` |
| OCRVQA (Mishra et al., 2019) | VQA | `OCRVQA_TESTCORE.tsv` |
| TextVQA (Singh et al., 2019) | VQA | `TextVQA_VAL.tsv` |
| DocVQA (Mathew et al., 2021) | VQA | `DocVQA_VAL.tsv` |
| MMMU-Pro (Yue et al., 2025) | MCVQ | `MMMU_Pro_10c.tsv` |
| MathVista (Lu et al., 2024a) | VQA | `MathVista_MINI.tsv` |
| MMVet-Hard (Yang et al., 2025) | VQA | `MMVet_Hard.tsv` |
| WildVision (Lu et al., 2024b) | VQA | `WildVision.tsv` |
| CharXiv (Wang et al., 2024b) | VQA | `CharXiv_descriptive_val.tsv` |
| OCRBench v2 (Fu et al., 2025) | VQA | `OCRBench_v2_MINI.tsv` |

Table C-2: Candidate LMM families and VLMEvalKit identifiers.

| Model family | Count | VLMEvalKit model identifiers |
|---|---|---|
| LLaVA-V1.5 | 2 | `llava_v1.5_7b`, `llava_v1.5_13b` |
| LLaVA-NeXT | 3 | `llava_next_mistral_7b`, `llava_next_vicuna_7b`, `llava_next_vicuna_13b` |
| LLaVA-NeXT-Interleave | 2 | `llava_next_interleave_7b`, `llava_next_interleave_7b_dpo` |
| LLaVA-OneVision | 4 | `llava_onevision_qwen2_0.5b_ov`, `llava_onevision_qwen2_7b_ov`, `llava_onevision_qwen2_0.5b_si`, `llava_onevision_qwen2_7b_si` |
| ShareGPT4V | 2 | `sharegpt4v_7b`, `sharegpt4v_13b` |
| InstructBLIP | 2 | `InstructBLIP_7b`, `InstructBLIP_13b` |
| Eagle | 3 | `Eagle-X5-7B`, `Eagle-X5-13B`, `Eagle-X5-13B-Chat` |
| InternVL | 6 | `Mini-InternVL-Chat-2B-V1-5`, `Mini-InternVL-Chat-4B-V1-5`, `InternVL2-1B`, `InternVL2-2B`, `InternVL2-4B`, `InternVL2-8B` |
| PaliGemma | 2 | `paligemma-3b-mix-224`, `paligemma-3b-mix-448` |
| Mantis | 3 | `Mantis-8B-Idefics2`, `Mantis-8B-clip-llama3`, `Mantis-8B-siglip-llama3` |
| mPLUG-Owl2 | 1 | `mPLUG-Owl2` |
| Qwen2-VL | 1 | `Qwen2-VL-2B-Instruct` |
| DeepSeek-VL2 | 1 | `deepseek_vl2_tiny` |

by the logarithm of the vocabulary size (Malinin & Gales, 2021; Tu et al., 2025). The subscripts $\mathsf{F}$ and $\mathsf{P}$ denote the first and penultimate tokens, respectively. The subscripts $\mathsf{max}$ and $\mathsf{avg}$ denote the maximum and mean token-level uncertainty across the generated sequence. These choices give four NLL variants and four Ent variants. The sampling-based methods average the similarities between the original response and five sampled responses. $\mathsf{Sample}_{\mathsf{BLEU}}$ and $\mathsf{Sample}_{\mathsf{BERT}}$ use unigram BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2020), respectively. For multiple-choice questions, $\mathsf{Sample}^{*}_{\mathsf{BERT}}$ includes the selected option's text when computing BERTScore, such as "A. Yellow" instead of "A" (Tu et al., 2025). When no option text is available, $\mathsf{Sample}^{*}_{\mathsf{BERT}}$ uses the same computation as $\mathsf{Sample}_{\mathsf{BERT}}$.

ATC (Garg et al., 2022) uses $-\mathsf{NLL}_{\mathsf{max}}$ as confidence. For each model, it selects a threshold so that the fraction of reference responses above the threshold approximately matches the model's accuracy on that dataset. The fraction of target responses above this threshold gives the predicted model score. We exclude WildVision from ATC calibration because its preference rewards do not specify the fraction of correct responses. ATC therefore uses 14 reference datasets when the target is WildVision and 13 otherwise.

### C.4 Clustering and intervention details

We first deduplicate the responses by exact string matching and sort the unique response strings lexicographically. Each unique response starts with a distinct label. We then examine all response pairs in the resulting order. If neither NLI direction predicts contradiction, the later response takes the current label of the earlier one. After all pairs have been examined, we map the labels back to the $M$ model responses and renumber them by first occurrence. The resulting labels define the non-overlapping semantic clusters used in Section 4.1.

For each visual intervention, we keep the question unchanged and draw an image from the full dataset image pool with replacement. The random seed for example $i$ is $s_i = \mathrm{int}(\mathsf{MD5}(\mathrm{str}(i)), 16) + 42$. This seed keeps the sampled image fixed across candidate models.

## D Score–performance correlation

We compare MAVER scores with the evaluation scores used to construct the reference ranking for each dataset. To account for differences in scale across datasets, we normalize the MAVER and evaluation scores separately over the $M = 32$ candidate LMMs. For dataset $d = 1, \ldots, D$, the normalized scores are

$$\widetilde{S}_d^m = \frac{S_d^m - \min_n S_d^n}{\max_n S_d^n - \min_n S_d^n}, \qquad \widetilde{A}_d^m = \frac{A_d^m - \min_n A_d^n}{\max_n A_d^n - \min_n A_d^n}.$$

For each model, we average both normalized scores over the $D = 15$ datasets:

$$\overline{\widetilde{S}}^m = \frac{1}{D}\sum_{d=1}^{D} \widetilde{S}_d^m, \qquad \overline{\widetilde{A}}^m = \frac{1}{D}\sum_{d=1}^{D} \widetilde{A}_d^m.$$

Across the 32 model-level score pairs, Pearson's correlation is $r = 0.95$ and Spearman's rank correlation is $\rho = 0.99$. Table D-1 reports the averaged scores, the rank obtained from each score, and the absolute rank error for every model. The predicted and reference ranks are identical for 14 models, differ by at most one position for 24 models, and differ by at most two positions for 29 models. The largest difference is four positions.

## E Analysis of visual evidence

### E.1 Agreement reliability and visual evidence

We first examine how often the largest semantic cluster contains a correct response. We then test whether visual evidence improves model ranking when this agreement fails. The analysis covers ten datasets with binary response-level evaluation. We define largest-cluster reliability as the percentage of image–question pairs whose largest semantic cluster contains a correct response. When several clusters tie for the largest size, a pair is counted as reliable only if every tied cluster contains a correct response. Difficult pairs are those answered correctly by no more than $25\%$ of the candidate models. We apply the same reliability criterion to this subset.

We define an agreement failure as an image–question pair for which at least one candidate model gives a correct response but none of the largest clusters contains one. $N_{\text{fail}}$ counts these pairs. Best $\uparrow$ is the percentage of agreement failures for which visual evidence improves the rank of the highest-ranked correct cluster. Top-1 is the percentage for which this cluster reaches first place. On the same subsets, $\Delta\rho_{\text{fail}}$ and $\Delta\tau_{w,\text{fail}}$ measure the changes in ranking correlations from MAVER* to MAVER.

Table E-1 and Figure 2 show that largest-cluster reliability ranges from $51.6\%$ on MMMU-Pro to $91.5\%$ on SQA-I across all image–question pairs. Reliability is lower on the difficult subset of every dataset. On difficult pairs, it is below $42\%$ on nine of the ten datasets and falls to $5.1\%$, $7.2\%$, and $9.4\%$ on SQA-I, AI2D, and RWQA. Thus, the largest semantic cluster is informative overall, but cross-model agreement becomes less reliable when few candidate models answer correctly.

On agreement-failure pairs, Table E-1 shows that visual evidence improves the rank of the highest-ranked correct cluster in $12.0\%$–$49.3\%$ of cases and moves it to first place in $7.4\%$–$42.9\%$. Both

Table D-1: Dataset-averaged normalized scores and ranks of the 32 candidate LMMs. Models are ordered by reference ranks. Ranks are computed from unrounded scores.

| Candidate LMM | $\widetilde{\widetilde{S}}^m$ | $\widetilde{\widetilde{A}}^m$ | Predicted rank | Reference rank | Absolute error |
|---|---|---|---|---|---|
| llava_onevision_qwen2_7b_ov | 0.91 | 0.91 | 1 | 1 | 0 |
| llava_onevision_qwen2_7b_si | 0.91 | 0.89 | 2 | 2 | 0 |
| InternVL2-8B | 0.87 | 0.89 | 7 | 3 | 4 |
| InternVL2-4B | 0.88 | 0.81 | 5 | 4 | 1 |
| Mini-InternVL-Chat-4B-V1-5 | 0.88 | 0.80 | 4 | 5 | 1 |
| deepseek_vl2_tiny | 0.90 | 0.79 | 3 | 6 | 3 |
| Eagle-X5-13B-Chat | 0.88 | 0.75 | 6 | 7 | 1 |
| Qwen2-VL-2B-Instruct | 0.85 | 0.74 | 8 | 8 | 0 |
| InternVL2-2B | 0.81 | 0.65 | 12 | 9 | 3 |
| Eagle-X5-13B | 0.84 | 0.65 | 9 | 10 | 1 |
| Mini-InternVL-Chat-2B-V1-5 | 0.81 | 0.64 | 11 | 11 | 0 |
| llava_next_vicuna_13b | 0.82 | 0.62 | 10 | 12 | 2 |
| llava_next_mistral_7b | 0.79 | 0.60 | 14 | 13 | 1 |
| Eagle-X5-7B | 0.79 | 0.59 | 13 | 14 | 1 |
| llava_next_interleave_7b_dpo | 0.73 | 0.58 | 17 | 15 | 2 |
| llava_next_vicuna_7b | 0.78 | 0.57 | 15 | 16 | 1 |
| llava_next_interleave_7b | 0.72 | 0.57 | 18 | 17 | 1 |
| Mantis-8B-Idefics2 | 0.75 | 0.56 | 16 | 18 | 2 |
| InternVL2-1B | 0.70 | 0.55 | 19 | 19 | 0 |
| paligemma-3b-mix-448 | 0.68 | 0.52 | 21 | 20 | 1 |
| llava_onevision_qwen2_0.5b_ov | 0.69 | 0.48 | 20 | 21 | 1 |
| llava_onevision_qwen2_0.5b_si | 0.67 | 0.46 | 22 | 22 | 0 |
| Mantis-8B-siglip-llama3 | 0.62 | 0.44 | 23 | 23 | 0 |
| mPLUG-Owl2 | 0.58 | 0.40 | 24 | 24 | 0 |
| paligemma-3b-mix-224 | 0.53 | 0.38 | 25 | 25 | 0 |
| Mantis-8B-clip-llama3 | 0.51 | 0.38 | 26 | 26 | 0 |
| llava_v1.5_13b | 0.46 | 0.33 | 27 | 27 | 0 |
| sharegpt4v_13b | 0.37 | 0.26 | 30 | 28 | 2 |
| llava_v1.5_7b | 0.37 | 0.23 | 29 | 29 | 0 |
| sharegpt4v_7b | 0.38 | 0.20 | 28 | 30 | 2 |
| instructblip_13b | 0.13 | 0.12 | 31 | 31 | 0 |
| instructblip_7b | 0.10 | 0.11 | 32 | 32 | 0 |

Table E-1: Agreement reliability and effects of visual evidence on ten datasets with binary response-level evaluation, where each response is labeled correct or incorrect.

| Dataset | $N_{\text{fail}}$ | Largest-cluster reliability (%) | | Effect of visual evidence (%) | | Ranking gain | |
|---|---|---|---|---|---|---|---|
| | | All | Difficult | Best ↑ | Top-1 | $\Delta\rho_{\text{fail}}$ | $\Delta\tau_{w,\text{fail}}$ |
| SQA-I | 140 | 91.5 | 5.1 | 49.3 | 42.9 | +1.51 | +1.35 |
| AI2D | 524 | 81.1 | 7.2 | 24.0 | 15.8 | +0.32 | +0.47 |
| RWQA | 222 | 65.2 | 9.4 | 32.9 | 29.3 | +0.34 | +0.44 |
| MMMU | 379 | 55.8 | 22.5 | 21.1 | 12.7 | +0.18 | +0.11 |
| GQA | 1589 | 73.4 | 23.5 | 21.5 | 17.6 | +0.02 | +0.01 |
| ChartQA | 358 | 79.7 | 22.2 | 19.3 | 12.8 | +0.18 | +0.19 |
| OCRVQA | 264 | 75.3 | 27.0 | 28.8 | 28.0 | +0.65 | +0.77 |
| MMMU-Pro | 493 | 51.6 | 36.1 | 20.3 | 10.3 | +0.06 | +0.07 |
| MathVista | 139 | 76.8 | 41.6 | 18.0 | 9.4 | +0.14 | +0.13 |
| CharXiv | 216 | 85.3 | 69.5 | 12.0 | 7.4 | +0.15 | +0.11 |

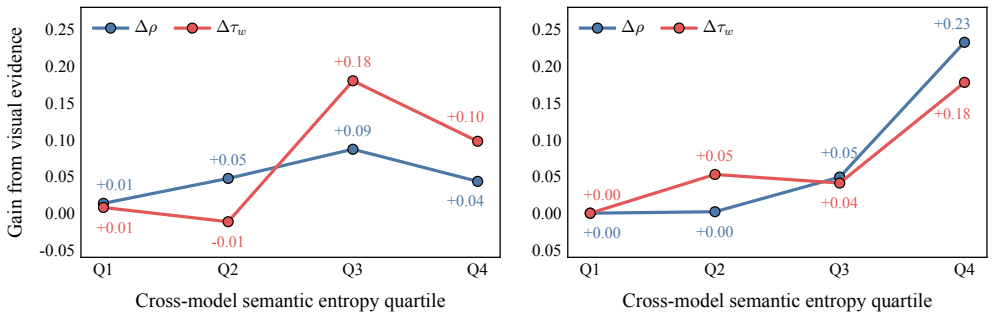
model-ranking correlations increase on all ten dataset-specific failure subsets. SQA-I and OCRVQA show the largest gains, with both correlations changing from negative to positive. These results show that visual evidence improves model ranking when the largest semantic cluster does not contain a correct response.

Table E-2: Effects of visual evidence across MMMU disciplines. Mean Acc. denotes accuracy averaged over models and questions. Acc.$\leq$ 0.5 is the percentage of questions for which no more than half of the candidate models answer correctly. $\Delta$ denotes the change from MAVER* to MAVER.

| Discipline | $N$ | Mean Acc. | Acc.$\leq$ 0.5 | Spearman $\rho$ | | | Weighted Kendall $\tau_w$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | MAVER* | MAVER | $\Delta$ | MAVER* | MAVER | $\Delta$ |
| Art & Design | 140 | 0.49 | 52.9% | 0.79 | 0.77 | -0.02 | 0.57 | 0.67 | +0.10 |
| Business | 175 | 0.30 | 84.0% | 0.65 | 0.72 | +0.07 | 0.58 | 0.65 | +0.07 |
| Health & Medicine | 175 | 0.35 | 76.0% | 0.47 | 0.55 | +0.08 | 0.20 | 0.19 | -0.01 |
| Humanities & Social Science | 140 | 0.52 | 47.9% | 0.84 | 0.85 | +0.01 | 0.66 | 0.61 | -0.05 |
| Science | 175 | **0.27** | **88.6%** | 0.43 | 0.53 | **+0.10** | 0.14 | 0.23 | **+0.09** |
| Tech & Engineering | 245 | 0.30 | 84.9% | 0.75 | 0.80 | +0.05 | 0.52 | 0.58 | +0.06 |

Table E-3: Effects of visual evidence across MathVista question types. Mean Acc. denotes accuracy averaged over models and questions. Acc.$\leq$ 0.5 is the percentage of questions for which no more than half of the candidate models answer correctly. $\Delta$ denotes the change from MAVER* to MAVER.

| Question type | $N$ | Mean Acc. | Acc.$\leq$ 0.5 | Spearman $\rho$ | | | Weighted Kendall $\tau_w$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | MAVER* | MAVER | $\Delta$ | MAVER* | MAVER | $\Delta$ |
| Free-form | 460 | **0.28** | **82.0%** | 0.69 | 0.84 | **+0.15** | 0.62 | 0.76 | **+0.14** |
| Multiple-choice | 540 | 0.46 | 62.4% | 0.38 | 0.38 | +0.00 | 0.33 | 0.34 | +0.01 |




(a) MMMU (b) MathVista

Figure E-1: Changes in rank correlations from MAVER* to MAVER across quartiles of cross-model semantic entropy. Q1 and Q4 represent the lowest and highest entropy groups, respectively.

### E.2 MMMU DISCIPLINES AND MATHVISTA QUESTION TYPES

**MMMU disciplines.**

Table E-2 highlights Business, Science, and Tech & Engineering. Mean model accuracy in these disciplines ranges from 0.27 to 0.30, lower than in the remaining disciplines. For at least 84.0% of the questions in each discipline, no more than half of the candidate models answer correctly. This proportion reaches 88.6% for Science. MAVER improves both ranking correlations in all three disciplines, with gains of 0.05–0.10 in $\rho$ and 0.06–0.09 in $\tau_w$. Science shows the largest improvements, 0.10 in $\rho$ and 0.09 in $\tau_w$.

**MathVista question types.**

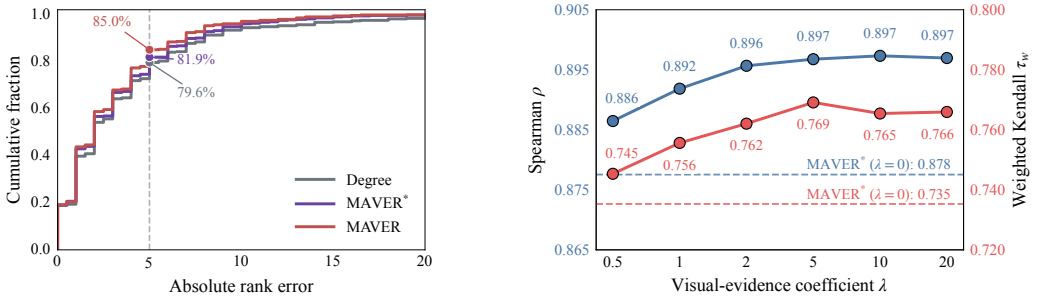
Table E-3 separates the 1,000 MathVista examples into free-form and multiple-choice questions. Candidate models achieve a mean accuracy of 0.28 on free-form questions and 0.46 on multiple-choice questions. For 82.0% of the free-form questions, no more than half of the candidate models answer correctly. The corresponding proportion is 62.4% for multiple-choice questions. On the free-form subset, MAVER raises $\rho$ from 0.69 to 0.84 and $\tau_w$ from 0.62 to 0.76. The corresponding changes on multiple-choice questions are 0.00 and 0.01. The gain from visual evidence is therefore concentrated on the more difficult, free-form portion of MathVista.

Table E-4: Effects of visual evidence and normalized semantic-entropy weighting on the nine datasets. Removing visual evidence gives MAVER*, and removing entropy weighting gives uniform aggregation. Higher is better.

| Variant | MCVQ | | | | | | | | VQA | | | | | | | | | | Overall | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SQA-I | | AI2D | | RWQA | | MMMU | | GQA | | ChartQA | | OCRVQA | | TextVQA | | DocVQA | | Average | |
| | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ |
| w/o. visual evidence | 0.96 | 0.94 | 0.99 | 0.93 | 0.93 | 0.69 | 0.82 | 0.46 | 0.91 | 0.85 | 0.99 | 0.90 | 0.91 | 0.74 | 0.98 | 0.90 | 0.99 | 0.90 | 0.94 | 0.81 |
| w/o. entropy weighting | 0.96 | 0.94 | 0.99 | 0.94 | 0.97 | 0.84 | 0.86 | 0.65 | 0.90 | 0.78 | 0.99 | 0.92 | 0.92 | 0.84 | 0.98 | 0.90 | 0.98 | 0.91 | 0.95 | 0.86 |
| Full (MAVER) | 0.96 | 0.91 | 0.99 | 0.94 | 0.97 | 0.86 | 0.86 | 0.59 | 0.90 | 0.75 | 0.99 | 0.90 | 0.92 | 0.83 | 0.98 | 0.91 | 0.98 | 0.90 | 0.95 | 0.84 |

Table E-5: Effects of visual evidence and normalized semantic-entropy weighting in the broader evaluation. Removing visual evidence gives MAVER*, and removing entropy weighting gives uniform aggregation. Higher is better.

| Variant | MCVQ | | VQA | | | | | | | | | | Overall | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MMMU-Pro | | MathVista | | MMVet-Hard | | WildVision | | CharXiv | | OCRBench v2 | | Average | |
| | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ |
| w/o. visual evidence | 0.54 | 0.26 | 0.66 | 0.53 | 0.82 | 0.62 | 0.81 | 0.72 | 0.92 | 0.74 | 0.95 | 0.86 | 0.78 | 0.62 |
| w/o. entropy weighting | 0.62 | 0.39 | 0.50 | 0.40 | 0.82 | 0.64 | 0.73 | 0.64 | 0.91 | 0.66 | 0.94 | 0.82 | 0.75 | 0.59 |
| Full (MAVER) | 0.58 | 0.34 | 0.77 | 0.60 | 0.84 | 0.65 | 0.81 | 0.70 | 0.93 | 0.75 | 0.96 | 0.90 | 0.82 | 0.66 |




(a) Ranking error distribution

(b) Sensitivity to $\lambda$

Figure E-2: Ranking errors and parameter sensitivity across 15 datasets. (a) ECDFs of absolute rank error for Degree, MAVER*, and MAVER. The vertical dashed line marks an absolute rank error of five. (b) Rank correlations averaged over datasets for different values of $\lambda$, which remain stable over $\lambda \in [2, 20]$. The horizontal dashed lines show the correlations of MAVER* ($\lambda = 0$).

### E.3 CROSS-MODEL SEMANTIC ENTROPY GROUPS

To test whether visual evidence is more useful when candidate models disagree, we sort the image–question pairs by $U_i$ and divide them into four approximately equal-sized groups. Q1 contains the pairs with the lowest cross-model semantic entropy, Q4 contains those with the highest entropy, and Q2–Q3 cover the intermediate ranges. We then compute model rankings within each group.

Figure E-1 shows that, from Q1 through Q4, the changes in $\rho$ on MMMU are $0.01$, $0.05$, $0.09$, and $0.04$. The corresponding changes in $\tau_w$ are $0.01$, $-0.01$, $0.18$, and $0.10$. Both metrics improve most in Q3 and remain above their agreement-only values in Q4. On MathVista, $\rho$ is unchanged in Q1 and Q2, increases by $0.05$ in Q3, and increases by $0.23$ in Q4. The $\tau_w$ gains are $0.00$, $0.05$, $0.04$, and $0.18$, respectively. The concentration of the largest gains in the higher-entropy groups supports the use of $U_i$ when aggregating image–question pairs.

Table E-6: MAVER ranking correlations across three visual intervention seeds. Seed 0 corresponds to the main experiments. We report the mean and sample standard deviation across seeds. The Average row summarizes the correlations averaged over all 15 datasets within each seed.

| Dataset | Seed 0 | | Seed 1 | | Seed 2 | | Mean ± Std. | |
|---|---|---|---|---|---|---|---|---|
| | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ | $\rho$ | $\tau_w$ |
| SQA-I | 0.96 | 0.91 | 0.96 | 0.91 | 0.96 | 0.92 | $0.96 \pm 0.00$ | $0.91 \pm 0.00$ |
| AI2D | 0.99 | 0.94 | 0.99 | 0.96 | 0.99 | 0.96 | $0.99 \pm 0.00$ | $0.95 \pm 0.01$ |
| RWQA | 0.97 | 0.86 | 0.96 | 0.83 | 0.96 | 0.86 | $0.96 \pm 0.01$ | $0.85 \pm 0.02$ |
| MMMU | 0.86 | 0.59 | 0.88 | 0.66 | 0.86 | 0.56 | $0.87 \pm 0.01$ | $0.60 \pm 0.05$ |
| GQA | 0.90 | 0.75 | 0.90 | 0.78 | 0.90 | 0.76 | $0.90 \pm 0.00$ | $0.76 \pm 0.01$ |
| ChartQA | 0.99 | 0.90 | 0.99 | 0.90 | 0.99 | 0.90 | $0.99 \pm 0.00$ | $0.90 \pm 0.00$ |
| OCRVQA | 0.92 | 0.83 | 0.92 | 0.83 | 0.93 | 0.86 | $0.93 \pm 0.01$ | $0.84 \pm 0.01$ |
| TextVQA | 0.98 | 0.91 | 0.98 | 0.91 | 0.98 | 0.91 | $0.98 \pm 0.00$ | $0.91 \pm 0.00$ |
| DocVQA | 0.98 | 0.90 | 0.99 | 0.91 | 0.99 | 0.91 | $0.99 \pm 0.00$ | $0.91 \pm 0.00$ |
| MMMU-Pro | 0.58 | 0.34 | 0.60 | 0.34 | 0.59 | 0.38 | $0.59 \pm 0.01$ | $0.35 \pm 0.02$ |
| MathVista | 0.77 | 0.60 | 0.76 | 0.60 | 0.75 | 0.61 | $0.76 \pm 0.01$ | $0.60 \pm 0.00$ |
| MMVet-Hard | 0.84 | 0.65 | 0.84 | 0.65 | 0.84 | 0.64 | $0.84 \pm 0.00$ | $0.65 \pm 0.01$ |
| WildVision | 0.81 | 0.70 | 0.80 | 0.70 | 0.81 | 0.71 | $0.81 \pm 0.00$ | $0.70 \pm 0.00$ |
| CharXiv | 0.93 | 0.75 | 0.93 | 0.75 | 0.93 | 0.76 | $0.93 \pm 0.00$ | $0.75 \pm 0.01$ |
| OCRBench v2 | 0.96 | 0.90 | 0.96 | 0.90 | 0.96 | 0.90 | $0.96 \pm 0.00$ | $0.90 \pm 0.00$ |
| Average | 0.90 | 0.77 | 0.90 | 0.77 | 0.90 | 0.78 | $0.90 \pm 0.00$ | $0.77 \pm 0.00$ |

### E.4 VISUAL EVIDENCE AND ENTROPY WEIGHTING BY DATASET

Tables E-4 and E-5 compare MAVER with variants that remove either visual evidence or normalized semantic-entropy weighting. On the nine datasets, adding visual evidence to MAVER* raises the average $\rho$ from $0.94$ to $0.95$ and $\tau_w$ from $0.81$ to $0.84$.

In the broader evaluation, using entropy weighting in MAVER raises the average $\rho$ from $0.75$ to $0.82$ and $\tau_w$ from $0.59$ to $0.66$. Both metrics improve on five of the six datasets. Across all 15 datasets, MAVER achieves the highest average $\rho$ and $\tau_w$. These results support using cross-model semantic entropy to weight response scores during aggregation.

### E.5 RANKING ERRORS

Rank correlations summarize the overall model ordering but not the displacement of individual models. The empirical cumulative distributions (ECDFs) in Figure E-2a show that MAVER achieves an absolute rank error of at most five on $85.0\%$ of model–dataset pairs, compared with $81.9\%$ for MAVER* and $79.6\%$ for Degree. Thus, MAVER improves the overall ordering and reduces individual rank errors.

### E.6 PARAMETER SENSITIVITY

Figure E-2b compares the agreement-only setting at $\lambda = 0$ with positive coefficients from $0.5$ to $20$. Every positive value improves both average correlations. Performance increases as $\lambda$ grows to 2 and remains stable over $\lambda \in [2, 20]$, showing that the gain from visual evidence is not confined to a narrow coefficient range.

### E.7 SENSITIVITY TO INTERVENTION SEEDS

We evaluate MAVER using three random seeds for visual intervention across all 15 datasets and 32 candidate LMMs. The original responses and all other settings remain fixed. We add offsets of 0, 1, and 2 to the per-example random seed defined in Appendix C.4. The run with zero offset corresponds to the main experiments. Table E-6 reports average correlations across all 15 datasets of $0.90$ for $\rho$ and $0.77$ for $\tau_w$. The standard deviations across seeds are below $0.01$ for both metrics. These results show that the overall ranking performance of MAVER is stable across intervention

Table F-1: Access requirements and additional LMM generations per model–example pair.

| Method | Proxy labels | Token probabilities | Additional LMM generations |
|---|---|---|---|
| AoL | Yes | No | 0 |
| ATC | Yes | Yes | 0 |
| NLL and Ent | No | Yes | 0 |
| Self-consistency | No | No | 5 |
| Degree | No | No | 0 |
| MAVER* | No | No | 0 |
| MAVER | No | No | 1 |

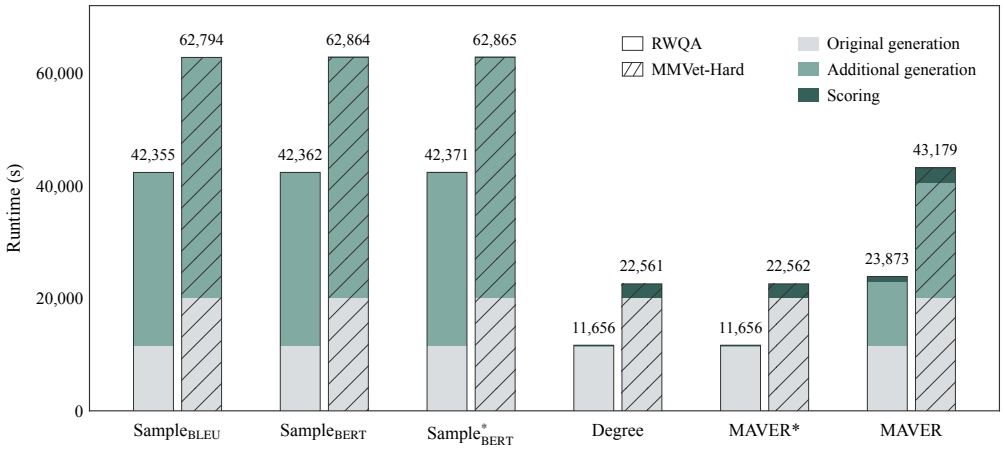



Figure F-1: Runtime comparison on RWQA (multiple-choice and short-answer VQA) and MMVet-Hard (open-ended VQA) with 32 candidate LMMs. Each bar sums original generation, additional generation, and scoring times. Values above the bars indicate total runtime.

seeds. Most datasets show small changes across seeds. MMMU shows greater variation in $\tau_w$, from 0.56 to 0.66.

## F Access requirements and computational cost

Table F-1 compares proxy-label access, token-probability access, and additional generations for each ranking method. Additional generations are counted relative to one original deterministic response per model–example pair. The sampling-based methods evaluated by Tu et al. (2025) use five additional stochastic generations. Degree and the agreement-only MAVER* reuse the original responses from the candidate pool, while the full MAVER adds one intervened response.

The generation counts exclude NLI inference. Degree and the agreement-only MAVER* construct the cross-model graph from $\binom{M}{2}$ unordered response pairs. The full MAVER uses the same pairs and additionally compares each of the $M$ original responses with its intervened counterpart. For $M = 32$, the two stages contain 496 and 32 pairs, respectively. We batch all NLI pairs and evaluate them in both directions without further candidate-LMM calls.

In Figure F-1, we compare runtime on RWQA and MMVet-Hard using all 32 candidate LMMs. For all three self-consistency methods, the five additional responses per model–example pair are generated in a batch. MAVER requires more scoring time but approximately $44\%$ less total time than the three self-consistency methods on RWQA and $31\%$ less on MMVet-Hard.