Believing without Seeing: Quality Scores for Contextualizing Vision-Language Model Explanations

Anonymous ACL submission

Abstract

When people query Vision-Language Models (VLMs) but cannot see the accompanying visual context (e.g. for blind and low-vision users), augmenting VLM predictions with natural language explanations can signal which model predictions are reliable. However, prior work has found that explanations can easily convince users that inaccurate VLM predictions are correct. To remedy undesirable overreliance on VLM predictions, we propose evaluating two complementary qualities of VLM-generated explanations via two quality scoring functions. We propose *Visual Fidelity*, which captures how faithful an explanation is to the visual context, and *Contrastiveness*, which captures how well the explanation identifies visual details that distinguish the model's prediction from plausible alternatives. On the A-OKVQA, VizWiz, and MMMU-Pro tasks, these quality scoring functions are better calibrated with model correctness than existing explanation qualities. We conduct a user study in which participants have to decide whether a VLM prediction is accurate without viewing its visual context. We observe that showing our quality scores alongside VLM explanations improves participants' accuracy at predicting VLM correctness by 11.1%, including a 15.4% reduction in the rate of falsely believing incorrect predictions. These findings highlight the utility of explanation quality scores in fostering appropriate reliance on VLM predictions.

1 Introduction

Vision-Language Models (VLMs) are being deployed in applications where users who do not have access to the VLM's visual context; for example, in assisting blind and low-vision individuals (Huh et al., 2024; An et al., 2025; Kim et al., 2025), autonomous multimodal digital agents (Koh et al., 2024), and in human-robot collaboration (Lukin et al., 2018). However, despite recent advances in VLM capabilities, they often exhibit unreliable behavior, including hallucinating visual details (Li et al., 2023; Gunjal et al., 2024) and making overconfident predictions (Valdenegro-Toro, 2024). In scenarios where users cannot directly observe the visual context of the VLM, it becomes imperative to enable users to accurately trust model outputs. How can we provide adequate context for users to establish appropriate reliance on model predictions and explanations?

Prior work has explored the utility of model explanations to support user decision making (Wang and Yin, 2021; Bansal et al., 2021). However, natural language explanations can be misleading for inaccurate model predictions (Joshi et al., 2023; Chaleshtori et al., 2024; Si et al., 2024a; Sieker et al., 2024). In Figure 1, we consider the question "*What period of the day does this photo reflect?*" where a VLM (incorrectly) answers "*Noon*" and generates a plausible explanation referencing cues like "*a clock on the building*" or "*lighting and shadows*." This reasoning may appear highly convincing to a user without access to the image, even though the prediction is incorrect. Users are particularly susceptible to *overreliance* on explanations in scenarios with such information asymmetry.

Existing explanation metrics (Hase et al., 2020; Chen et al., 2023; Liu et al., 2023b) are designed to evaluate explanations generated by text-only LLMs, and we find these are *poorly calibrated* to VLM prediction accuracy. However, considering the explanation with respect to the accompanying visual context can reveal the reliability of the VLM prediction; for instance, identifying hallucinated details in the explanation (*"shadows suggesting that the sun is at its peak"*) or that the explanation does not mention the critical detail for identifying the correct answer (the time on the clock).

We propose evaluating two new qualities of VLM-generated explanations: *Visual Fidelity*—how faithfully a VLM explanation reflects the accompanying visual context—and *Contrastiveness*—

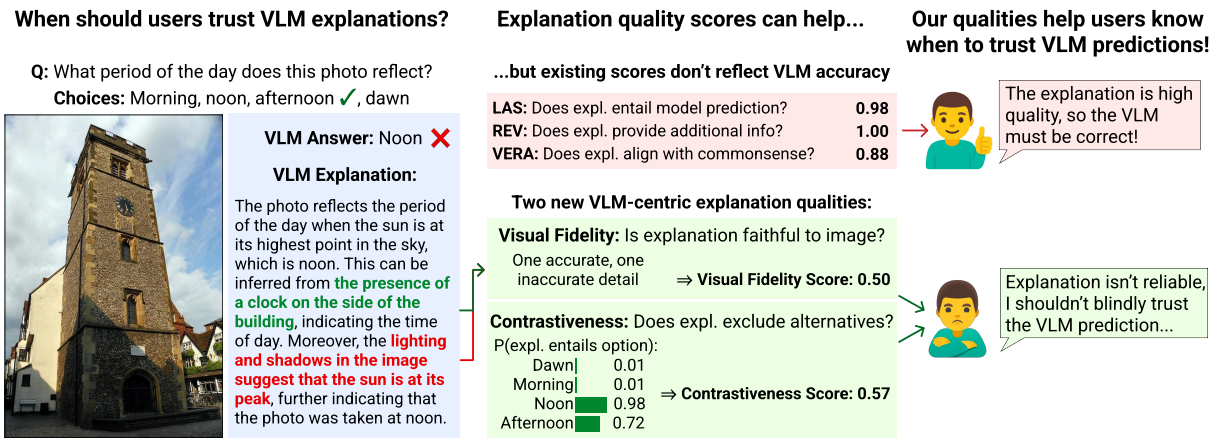


Figure 1: VLM explanations that sound plausible can mislead users. Contextualizing VLM explanations with quality scores may help users know when to rely, but existing explanation qualities are not calibrated with VLM prediction accuracy. We propose Visual Fidelity and Contrastiveness to contextualize these explanations, which are better calibrated with correctness, and also help users make better decisions about when to believe VLM predictions.

how well the explanation rules out alternative answers. We introduce training-free methods to estimate both these qualities without relying on ground-truth quality annotations.

In experiments with three popular VLMs and three visual reasoning tasks, A-OKVQA (Schwenk et al., 2022), VizWiz (Gurari et al., 2018), and MMMU-Pro (Yue et al., 2025), we find that our Visual Fidelity and Contrastiveness qualities better distinguish correct VLM predictions than do existing evaluation qualities (Hase et al., 2020; Chen et al., 2023; Liu et al., 2023b), and are also better calibrated with VLM correctness (§3). We conduct several user studies in which participants evaluated VLM predictions without access to visual inputs (§4). We find that augmenting VLM explanations with our proposed qualities improves participants' ability to distinguish between correct and incorrect VLM predictions. In particular, showing the product of Visual Fidelity and Contrastiveness as a single quality score leads to an 11.1% absolute improvement in user accuracy, along with a 15.4% absolute reduction in the rate of users falsely believing inaccurate predictions. Finally, we show that presenting explanation qualities via natural language descriptions instead of scores further improves user decision making accuracy (§4.2). Overall, our findings highlight the utility of evaluating the quality of VLM-generated explanations and communicating these qualities to users relying on VLM assistants.

2 Quality Scores for VLM Explanations

Visual reasoning is the task of answering a textual question about an image by drawing inferences from what is visually observed. A vision-language model (VLM) is given an input $x = (I, Q) \in \mathcal{X}$ consisting of an image I and a question Q, and produces an answer $a \in \mathcal{A}$, a (closed or open) set of options. Additionally, the VLM generates a natural language explanation E for its prediction a. This explanation can be either a chain-of-thought rationale (Wei et al., 2023) that the VLM prediction is conditioned on $(IQ \rightarrow Ea)$ or a post hoc justification generated after the VLM prediction $(IQ \rightarrow aE)$ (Wiegreffe et al., 2022b).

We define an explanation quality scoring function $S_\varphi(E; I, Q, a, \mathcal{A})$[1] that maps a VLM-generated explanation E to a score in $[0, 1]$ reflecting a specific quality φ of the reliability of an explanation. Existing qualities for natural language explanations, primarily developed and evaluated in text-only settings, typically look at how well an explanation supports the model's prediction (Hase et al., 2020), its informativeness (Chen et al., 2023), and its plausibility according to common sense priors (Liu et al., 2023b). However, evaluating *only* these qualities is insufficient for VLM-generated explanations since they do not consider the relationship between the explanation and the visual context. Further, we find that these explanation qualities are poorly calibrated with prediction accuracy (§3) and therefore not useful for a user deciding whether to trust the VLM's prediction or not.

To address these gaps, we introduce two *novel trust cues*: **Visual Fidelity** and **Contrastiveness**. (Figure 1).

[1]For brevity, we note arguments for S only when relevant.

Algorithm 1 Evaluate Visual Fidelity

Input: Image I, question Q, model prediction a, explanation E

Output: Visual Fidelity score $S_{\text{VF}}(E)$

$\mathcal{Q}^V = \{q_1^V, \ldots, q_K^V\} \leftarrow m_{\texttt{QGen}}(E, Q, a)$
for all $q_j^V \in \mathcal{Q}^V$ **do**
  $a_j^V \leftarrow m_{\texttt{Verif}}(q_j^V, I)$
end for
$S_{\text{VF}}(E) \leftarrow \frac{\sum_{i=1}^{K} \mathbb{1}\{a_i^V = \text{“yes”}\}}{K}$

Algorithm 2 Evaluate Contrastiveness

Input: Question Q, explanation E, set of possible answers $\mathcal{A}$, model prediction $a_0 \in \mathcal{A}$

Output: Contrastiveness score $S_{\text{Contr.}}(E)$

$P \leftarrow \text{Mask}(E, \mathcal{A})$
for all $a_j \in \mathcal{A}$ **do**
  $h_j \leftarrow m_{\texttt{QA}\rightarrow\texttt{S}}(Q, a_j)$
end for
$S_{\text{Contr.}}(E) \leftarrow \frac{\text{PR}_{\texttt{NLI}}(P \text{ entails } h_0)}{\sum_{a_j \in \mathcal{A}} \text{PR}_{\texttt{NLI}}(P \text{ entails } h_j)}$

2.1 Visual Fidelity

Visual Fidelity quality captures how faithful the VLM's explanation is to the contents of the image. The explanation, typically spanning multiple sentences, may mention several details about the image in order to explain its prediction. However, it is possible that some of these details are hallucinated by the model (Gunjal et al., 2024). For example, in Figure 1, the VLM justifies its prediction by incorrectly claiming that the shadows indicate that the sun is at its peak. Such misleading details can sway users who cannot observe the VLM's visual context. Evaluating the Visual Fidelity of an explanation could potentially alert users when important explanation details may be hallucinated.

We measure the visual fidelity of an explanation E with respect to the image I by decomposing the explanation into a set of facts about the image and individually verifying each fact (Algorithm 1). Conceptually, this is an *oracle* scoring function that would check each atomic claim against ground-truth visual Q&A; we approximate it with an LLM-as-judge, using a verifier $m_{\texttt{Verif}}$ (GPT-4o by default), and we audit this approximation with expert annotations and verifier swaps at Section J. Concretely, we prompt a question generation LLM $m_{\texttt{QGen}}$ to generate a set of verification questions $\mathcal{Q}^V$ that confirm visual details that have been mentioned in the VLM-generated explanation. The model has to generate questions such that answering "yes" confirms the presence of the detail in the image (see Table 6 for the full LLM prompt). The number of questions generated by $m_{\texttt{QGen}}$ depends on the explanation and may vary across instances.

For each verification question $q_i^V \in \mathcal{Q}^V$, a verifier VLM $m_{\texttt{Verif}}$ produces an answer $a_i^V \in \{\text{yes}, \text{no}\}$. The Visual Fidelity score $S_{\text{VF}}(E)$ is calculated as the fraction of questions for which the verifier $m_{\texttt{Verif}}$ answers "yes":

$$S_{\text{VF}}(E) = \frac{\sum_{i=1}^{|\mathcal{Q}^V|} \mathbb{1}\{a_i^V = \text{“yes”}\}}{|\mathcal{Q}^V|}$$

A high Visual Fidelity score indicates that the rationale faithfully represents the contents of the image.

2.2 Measuring Contrastiveness

Contrastiveness captures whether the VLM's explanation mentions all relevant visual details to identify the correct answer. The evidences and reasoning in the explanation may not identify the visual details that correctly distinguish the correct answer from alternatives. For instance, the explanation in Figure 1 fails to mention the time on the clock—the key visual cue for correctly answering the question. When such relevant cues are missed, the explanation may inadvertently support other alternative answers apart from the VLM's prediction. Evaluating the Contrastiveness of an explanation could potentially alert users to when the model explanation fails to identify relevant visual cues.

However, evaluating whether all relevant visual details have been mentioned in the explanation is difficult if we do not know what the relevant visual cues are. For example, if the explanation never mentioned a clock, then we would not know that the time on the clock is the critical detail for predicting the correct answer. Therefore, we estimate sufficiency using a proxy measure, based on the insight that if the model did capture all relevant details then it would have eliminated other plausible alternatives. Specifically, for a task where a VLM has to predict an answer a_0 from a closed set of possible answers $\mathcal{A}$, we measure the Contrastiveness of an explanation E by evaluating how strongly it entails the model prediction a_0 relative to the set of alternative answers $\mathcal{A} \backslash \{a_0\}$ (Algorithm 2).

We begin by masking mentions of all answers $a \in \mathcal{A}$ in the explanation E to prevent label leak-

age from affecting the entailment model. Then, for each possible answer $a_j \in \mathcal{A}$, a paraphraser LLM $m_{\texttt{QA}\rightarrow\texttt{S}}$ paraphrases the question-answer pair (Q, a_j) into a declarative sentence h_j. Finally, we calculate the probability that the masked explanation P entails the hypothesis h_j using an entailment model $m_{\texttt{NLI}}$. The Contrastiveness $S_{\text{Contr.}}(E)$ score is computed as the relative entailment probability for the predicted hypothesis h_0, compared to the entailment probability over all possible hypotheses:

$$S_{\text{Contr.}}(E) = \frac{\text{PR}_{\texttt{NLI}}(P \text{ entails } h_0)}{\sum_{a_j \in \mathcal{A}} \text{PR}_{\texttt{NLI}}(P \text{ entails } h_j)}.$$

A high Contrastiveness score indicates the explanation identifies visual cues that not only support the prediction but also eliminate plausible alternatives.

2.3 Combining Visual Fidelity and Contrastiveness

We study how complementary our scoring functions are by computing a single quality score that combines the two explanation quality scores $S_{\text{VF}}(E)$ and $S_{\text{Contr.}}(E)$. We evaluate three combinations: the average, product, and minimum of the Visual Fidelity and Contrastiveness quality scores.

$$S_{\text{avg}}(E) = \frac{S_{\text{VF}} + S_{\text{Contr.}}}{2};$$
$$S_{\text{Prod}}(E) = S_{\text{VF}} \times S_{\text{Contr.}};$$
$$S_{\text{min}} = \min(S_{\text{VF}}, S_{\text{Contr.}}).$$

3 Are Visual Fidelity and Contrastiveness Indicative of VLM Correctness?

We evaluate quality scoring functions on their ability to indicate the accuracy of a VLM prediction.

Visual Reasoning Tasks. We evaluate our explanation qualities on three visual reasoning tasks: A-OKVQA (Schwenk et al., 2022)—a multiple-choice VQA benchmark that requires reasoning over images using external knowledge and commonsense; VizWiz (Gurari et al., 2018)—an open-ended VQA task consisting of questions asked by blind and low-vision users about images they captured on their mobile phones; and MMMU-Pro (Yue et al., 2025)—a massive multi-discipline multimodal understanding and reasoning benchmark. We sample 500 questions from the validation set of each task. Section K.2 contains details about the tasks' pre-processing and evaluation. For VizWiz, which lacks a specified set of possible answers, we generate hard negatives using an LLM to enable Contrastiveness evaluation. These negatives are used for post-hoc evaluation and are not provided to the model during answer generation.[2]

Models. We experiment with three popular vision-language models: LLaVA-v1.5-7B (Liu et al., 2023a), Qwen2.5-VL-7B (Bai et al., 2025), and GPT-4o (Hurst et al., 2024)[3]. Answers and rationales are generated in a two-step process: the model is first prompted to predict an answer to the question based on the image, and then prompted to generate a natural language explanation for its prediction. Section C contains the exact prompts we use for answer and explanation generation.

Additionally, our quality scoring functions use several model-based tools. For computing Visual Fidelity, GPT-4o is used both for generating verification questions and answering those questions. For computing Contrastiveness, we use GPT-4o to mask answers in the explanation and paraphrase question-answer pairs into declarative sentences, and the entailment model from Sanyal et al. (2024).

Baseline Explanation Qualities. We compare our proposed qualities against three established text-only explanation baselines: Simulatability (Hase et al., 2020), Informativeness (Chen et al., 2023) and Commonsense Plausibility (Liu et al., 2023b). See Section D for details.

Evaluating Quality Score Calibration. We evaluate each quality scoring function on its ability to indicate whether a VLM prediction is accurate based on its explanation. Specifically, we consider two evaluation metrics: **Discriminability (Disc)** and **Expected Calibration Error (ECE)**.

On an evaluation set of N visual reasoning instances, $\text{Disc}(S_\varphi)$ evaluates a quality scoring function S_φ by calculating the difference between the mean quality score assigned to instances with accurate predictions ($\text{Acc}(a_i) = 1$) and inaccurate predictions ($\text{Acc}(a_i) = 0$):

$$\text{Disc}(S_q) = \mathbb{E}_{\substack{1 \le i \le N \\ \text{Acc}(a_i)=1}} \left[S_q(E_i)\right] - \mathbb{E}_{\substack{1 \le i \le N \\ \text{Acc}(a_i)=0}} \left[S_q(E_i)\right].$$

We further calculate whether the difference in means between these two distributions ($S_\varphi(E)$, for accurate predictions, versus $S_\varphi(E)$, for inaccurate predictions) is significant using an unpaired t-test.

[2]We discuss the specific challenges and results of this open-set evaluation in Section F and validate the strategy via ablation in Section G.

[3]We use the `gpt-4o-2024-05-13` checkpoint.

	Discriminability (higher is better)												
	A-OKVQA				VizWiz				MMMU-Pro				**Overall Avg**
	LLaVA	Qwen2.5	GPT-4o	Avg	LLaVA	Qwen2.5	GPT-4o	Avg	LLaVA	Qwen2.5	GPT-4o	Avg	
Simulatability	0.106^{**}	0.247^{***}	0.098	0.150	0.088^{*}	0.158^{**}	-0.130^{*}	0.039	0.059	0.195^{***}	0.079^{**}	0.111	0.100
Informativeness	-0.018	0.000	0.069	0.017	0.093^{*}	-0.007	-0.113	-0.009	-0.008	0.006	0.011	0.003	0.004
Plausibility	0.043^{**}	0.022	0.028	0.031	0.004	-0.020	-0.002	-0.006	0.006	0.012	0.004	0.007	0.011
Visual Fidelity	0.181^{***}	0.080^{***}	0.085^{***}	0.115	0.240^{***}	0.042^{*}	0.024	**0.102**	0.169^{**}	0.038	0.090^{***}	0.099	0.106
Contrastiveness	0.243^{***}	0.283^{***}	0.248^{***}	0.258	0.041	0.054	-0.097	-0.001	0.047	0.154^{***}	0.101^{***}	0.101	0.119
Avg(VF, Contr.)	0.212^{***}	0.181^{***}	0.166^{***}	0.187	0.141^{***}	0.048	-0.037	0.051	0.108^{**}	0.096^{***}	0.096^{***}	0.100	0.112
Prod(VF, Contr.)	0.320^{***}	0.315^{***}	0.266^{***}	**0.301**	0.168^{***}	0.082	-0.082	0.056	0.090^{**}	0.146^{***}	0.143^{***}	**0.126**	**0.161**
Min(VF, Contr.)	0.295^{***}	0.298^{***}	0.255^{***}	0.283	0.165^{***}	0.077	-0.084	0.053	0.082^{**}	0.127^{***}	0.124^{***}	0.111	0.149

Table 1: Discriminability (Disc., higher is better) for different explanation quality scoring functions across three models and three visual reasoning tasks. Significance is evaluated using an unpaired t-test: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Averages are not evaluated for significance. Notably, VizWiz–GPT-4o exhibits significantly lower and predominantly negative discriminability scores across all metrics; see Section I for a discussion of the underlying causes.

	Expected Calibration Error (lower is better)												
	A-OKVQA				VizWiz				MMMU-Pro				**Overall Avg**
	LLaVA	Qwen2.5	GPT-4o	Avg	LLaVA	Qwen2.5	GPT-4o	Avg	LLaVA	Qwen2.5	GPT-4o	Avg	
Simulatability	0.288	0.219	0.302	0.270	0.305	0.292	0.494	0.364	0.446	0.418	0.357	0.407	0.347
Informativeness	0.332	0.152	0.228	0.237	0.429	0.168	0.446	0.348	0.216	0.706	0.538	0.487	0.357
Plausibility	0.162	0.270	0.317	0.250	0.110	0.310	0.325	0.248	0.318	0.391	0.272	0.327	0.275
Visual Fidelity	0.197	0.137	0.099	0.144	0.271	0.139	0.160	0.190	0.558	0.535	0.412	0.501	0.279
Contrastiveness	0.176	0.136	0.210	0.174	0.294	0.292	0.424	0.337	0.221	0.276	0.256	0.251	0.254
Avg(VF, Contr.)	0.109	0.054	0.090	**0.084**	0.139	0.138	0.216	**0.164**	0.360	0.389	0.279	0.343	**0.197**
Prod(VF, Contr.)	0.164	0.154	0.233	0.184	0.290	0.296	0.426	0.337	0.144	0.203	0.200	**0.182**	0.234
Min(VF, Contr.)	0.147	0.144	0.227	0.173	0.265	0.284	0.425	0.325	0.154	0.234	0.215	0.201	0.233

Table 2: Expected Calibration Error (ECE, lower is better) for different explanation quality scoring functions across three models and three visual reasoning tasks.

Expected Calibration Error (ECE) (Guo et al., 2017) is typically used to evaluate how accurately a model's confidence estimate reflects the true accuracy of its predictions. We evaluate our quality scoring functions using ECE by interpreting the quality scoring functions as confidence estimates.

Results. Table 1 and Table 2 compare three sets of explanation qualities: existing ones developed for text-only explanations (Simulatability, Informativeness, Plausibility), our proposed qualities (Visual Fidelity, Contrastiveness), and combinations of proposed qualities. We observe that across all datasets and models, our proposed quality scoring functions almost always achieve higher Discriminability scores and lower Expected Calibration Errors than existing scoring functions across all models. Most differences in Discriminability on these datasets are statistically significant ($p < 0.001$), confirming that our metrics effectively distinguish accurate from inaccurate predictions. We also find that our explanation-based metrics significantly outperform standard logit-based uncertainty baselines on complex reasoning tasks (see Section H). On

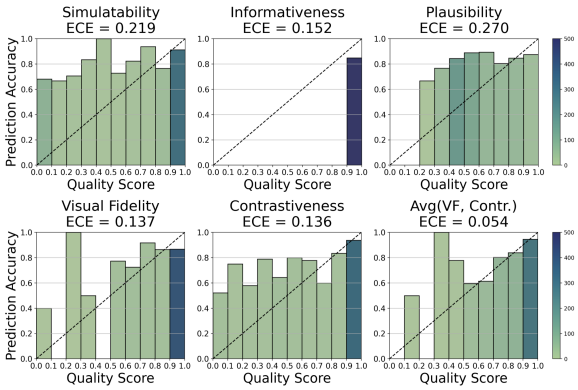


Figure 2: Calibration curves for various quality scoring functions when evaluating explanations generated by Qwen2.5-VL-7B on the A-OKVQA dataset.

VizWiz, while LLaVA-v1.5-7B shows strong discriminability, GPT-4o and Qwen2.5-VL-7B exhibit lower or negative scores on all scoring functions; as discussed in Section I, this is primarily due to valid answers being penalized by the dataset's strict ground truth. Crucially, **combining the Visual Fidelity and Contrastiveness quality scores results in the best overall Discriminability and Calibration**, highlighting their complementary utility.

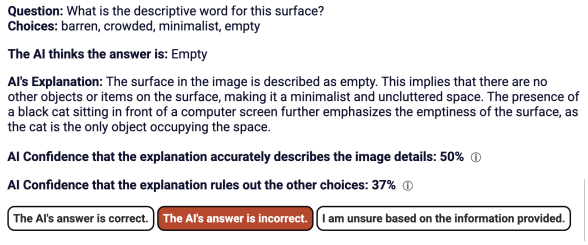


Figure 3: Our study interface where users are shown the visual reasoning question, VLM prediction and explanation, and optionally one or more quality scores (using simplified language descriptions). Here, based on the explanation's quality scores, the user correctly believes that the VLM prediction is incorrect.

Figure 2 shows calibration curves for the baseline qualities, our proposed qualities, and an average of Visual Fidelity and Contrastiveness, for explanations generated by Qwen2.5-VL-7B on the A-OKVQA task. We observe that Informativeness's low calibration error is achieved by assigning high scores to all explanations, while our proposed quality scores achieve a lower calibration error while assigning a wider spread of scores.

Verifier robustness. To test whether VF's behavior depends on the specific verifier model, we replace the default GPT-4o with two smaller open verifiers (Gemma-3n-E4B and Qwen2.5-VL-7B). Section J.3 shows that all three yield comparable discriminability and calibration, indicating VF remains stable across verifier architectures.

4 Are VLM Explanation Quality Scores Helpful to Users?

Calibration alone does not necessarily translate to downstream utility for users in real-world scenarios (Vodrahalli et al., 2022). We now investigate whether providing VF and Contr. scores alongside VLM explanations helps users more accurately decide when to believe VLM predictions.

4.1 User Study Setup

We design a user study simulating a real-world setting of a user relying on a VLM assistant where the user cannot directly view the VLM's visual input. Study participants are tasked with assessing the accuracy of VLM predictions.

For a visual reasoning question (I, Q) from the A-OKVQA and VizWiz tasks, participants are shown the question Q, the VLM's prediction a and explanation E, and optionally an explanation quality score $S_\varphi(E)$. Importantly, participants *cannot see the image I accompanying the question*. Participants have to decide whether they believe the model's answer is correct or incorrect based on the provided context. They can also indicate "unsure" if they feel the information is insufficient for making a reliable judgment. This design helps us measure how presenting different quality scores to a user affects user decision making. Figure 3 contains an example of how information is presented in our user study interface.

Study Details. We conducted a between-subjects study on a balanced subset of 100 questions each from A-OKVQA and VizWiz (total 300 annotations per condition). Participants were recruited via Prolific and paid $2 base + performance bonuses (see Section K for full protocol and demographics).

Evaluation Metrics. Since participants are asked to judge whether the VLM's prediction is correct or incorrect, we first exclude responses marked as "unsure." Following prior work on evaluating human-AI collaboration (Joshi et al., 2023; Ma et al., 2024; Srinivasan and Thomason, 2025; Wiegreffe et al., 2022a), we then calculate **User Accuracy** over the remaining responses, and the degree of *appropriate reliance* using two metrics: Over-Reliance and Under-Reliance. **Over-Reliance** represents the fraction of interactions where the user believed the VLM prediction was correct, when in fact it was incorrect. Similarly, **Under-Reliance** is the fraction of interactions where the user mistakenly believed the VLM was incorrect. For both reliance metrics, lower values are preferred.

4.2 RQ1: Does providing explanation quality scores improve user reliance?

We first evaluate whether showing users our proposed quality scores alongside VLM explanations helps users more accurately assess VLM correctness. We compare communicating: 1) only the Visual Fidelity score, 2) only the Contrastiveness score, 3) both scores, and 4) a product of the two scores. Figure 9 contains examples of how these quality scores are communicated to users. We additionally include a *control* setting where only an explanation is shown without a quality score, and a *random* setting where users are shown a quality score randomly sampled between 0 and 1.

Results. Presenting explanation quality scores, either individually or in combination, consistently improves user performance across all three metrics (Figure 4). Relative to the control explanation-

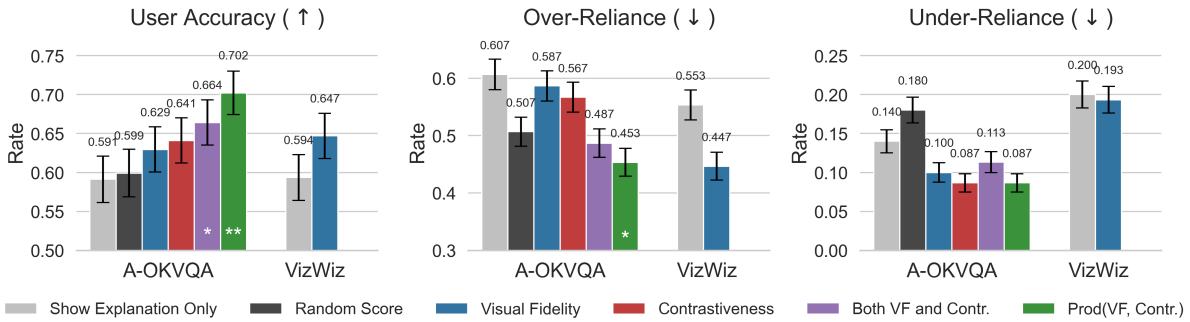


Figure 4: Effect of showing users different quality scores on User Accuracy, Over-Reliance and Under-Reliance. Error bars represent the standard deviation of the data. Asterisks denote improvements over the explanation-only baseline using a bootstrap significance test (*: $p < 0.05$, **: $p < 0.01$).

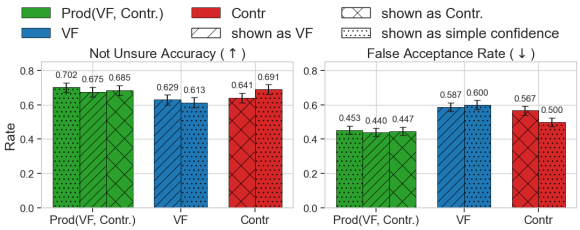


Figure 5: Ablation on *presentation types* on A-OKVQA dataset: holding the numeric signal fixed (Prod(VF, Contr.), VF, or Contr.), we vary the on-screen label (*shown as VF*, *shown as Contr*, or *simple confidence*).

only condition and the Random Score baseline, all treatment conditions improve user accuracy. On A-OKVQA, the most pronounced improvements are observed when showing the product of Visual Fidelity and Contrastiveness, which leads to a 11.1% increase in user accuracy, a 15.4% reduction in over-reliance, and a 5.3% reduction in under-reliance. These gains suggest that quality signals help users better identify correct model predictions and avoid being misled by incorrect ones.

Other treatment variants also show meaningful improvements. Visual Fidelity and Contrastiveness alone both reduce over-reliance by approximately 2–4%, and displaying both scores side-by-side further improves user outcomes. On VizWiz, the trends are similar; addition of quality scores leads to improvements in both accuracy and over-reliance. Together, **these findings highlight the utility of explanation quality scores in helping users make more informed and calibrated decisions about when to trust VLM predictions**.

4.3 RQ2: Which aspects of the presentation affect user decision accuracy?

Building on our findings in RQ1, we explore which *presentation choices* are significant. We explore two dimensions:

(i). Score labeling (RQ2a): We examine whether presenting the same numeric value, but varying its labeling on the study screen (e.g., presenting the score as VF, Contr., or a generic "AI confidence") affects user reliance.

(ii). Score descriptiveness (RQ2b): Here, we examine the effect of presenting a scalar quality score versus a descriptive summary that presents evidence (e.g., verified visual claims for VF or plausible alternatives for Contr.).

4.3.1 RQ2a: Does framing of the quality score affect user reliance?

One concern is whether our observed gains stem from *what* the score communicates versus *how* it is labeled to users. For instance, presenting a quality score as the visual fidelity of the explanation may yield different user reliance behavior than if the same score is presented as a generic AI confidence in the explanation.

In this ablation study, using the same underlying score distributions for A-OKVQA predictions (Prod(VF, Contr.), VF, and Contr.), we permuted their on-screen presentation: (i) shown under the correct name, (ii) mislabeled as the other qualities or shown as a generic "AI confidence" that does not specifically mention either explanation quality.

Results. As shown in Figure 5, renaming/genericizing the numeric signal yields nearly identical outcomes in both User Accuracy and Over-Reliance: differences are small and largely within error bars across all three bases. Thus, we observed that user behavior is driven by the numeric signal's distribution rather than its label.

4.3.2 RQ2b: Does providing language descriptions of quality scores improve user decision accuracy?

Our proposed quality scores are interpretable. For Visual Fidelity, we can observe which visual details

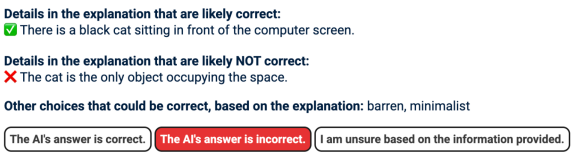


Figure 6: The descriptive quality presentation. The question and explanation display are identical to Figure 3, but numeric scores are replaced with text descriptions of the Visual Fidelity and Contrastiveness qualities.

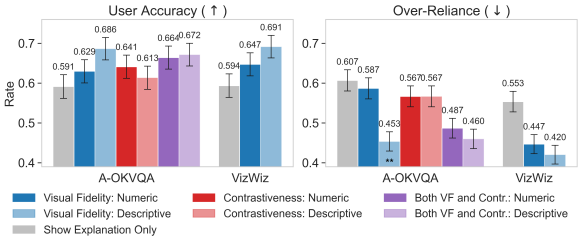


Figure 7: Effect of showing numeric and descriptive qualities on User Accuracy and Over-Reliance. Error bars show standard deviation of the data. Asterisks denote improvements between descriptive setting against the corresponding numeric setting with a bootstrap significance test (**: $p < 0.01$).

in the explanation were and were not verified by the verifier VLM. For Contrastiveness, we can examine which alternative answers were also assigned a high entailment probability. We compare the effect of showing explanation qualities as numeric scores versus interpretable text descriptions on user decision making. Figure 6 shows an example of descriptive presentations of our proposed qualities.

Results. As shown in Figure 7, we observe that descriptive formats perform comparably to, and in some cases slightly better than, their numeric counterparts. While showing both quality scores and for VF individually, the descriptive versions lead to improved user accuracy and reduced over-reliance. However, descriptive treatments hinder performance for Contr., leading to reduced accuracy. These findings suggest that while numeric scores offer a compact signal, **descriptive formats may help reduce over-reliance by making the explanation evaluation more transparent**.

4.4 RQ3: How does calibration of quality scores impact user decision making?

Our results in Figure 4 compared quality scores with different levels of calibration, which were also presented with different messages to users. To isolate the relationship between calibration error and user decision accuracy from the presentation of explanation qualities, we run user studies by showing scores from different distributions (Visual Fidelity

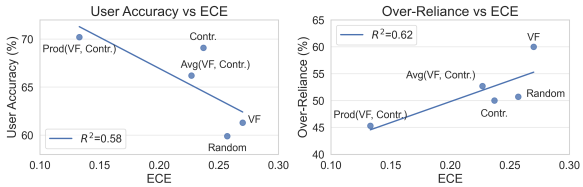


Figure 8: Relationship between ECE of different quality scores and their downstream utility to users.

scores, Contrastiveness scores, and average and product combinations of the two) while presenting the scores to users using the same message (*"AI Confidence that the explanation is accurate"*).

Results. In Figure 8, we compare the ECE of different explanation qualities[4] and the resulting User Accuracy and Over-Reliance. We observe that there is a negative correlation between calibration error and the downstream user decision making accuracy, with ECE explaining $\approx 60\%$ of the variance in User Accuracy and Over-Reliance respectively. These findings indicate that **calibration error of the quality scores is an important determiner of downstream utility to users**.

5 Conclusions and Future Directions

While explanations are thought to help users discern when to rely on VLM predictions, providing explanations alone can mislead users into believing the model even when it is incorrect. We propose evaluating two unexplored, complementary explanation qualities: Visual Fidelity and Contrastiveness, and also introduce scoring functions for measuring both qualities. We find that our proposed quality scoring functions are well calibrated with model correctness, compared to existing notions of explanation quality. Through several user studies, we demonstrate real-world utility of these scores, as presenting users with quality scores alongside explanations improves their task accuracy and reduces over-reliance on the VLM.

An important direction is to develop adaptive human-AI reliance strategies that learn when to present explanations and quality scores and when to suppress them, potentially utilizing bandit algorithms to automate system-side abstention (Feng and Boyd-Graber, 2022). Additionally, explanation quality scores could be used as training objectives to improve the generation of explanations. Finally, future works should study how explanation quality influences user trust over time.

[4] ECE is computed over the 100 samples used in the user study, not the full evaluation set of 500 samples.

Limitations

One limitation of our approach lies in the open-ended Contrastiveness evaluation. Its score quality is inherently partially bounded by the semantic quality of the generated hard negatives. While we validate that our generator produces challenging distractors in Section I, the metric remains sensitive to their plausibility; if distractors are too easily distinguished, the metric may overestimate confidence, whereas if they are semantically ambiguous, it may penalize valid reasoning. Thus, our metric's resolution is coupled with the generator's capability. Second, while we demonstrate effectiveness on A-OKVQA, its multiple-choice format may not strictly be representative of real-world, open-ended visual queries. We attempt to bridge this by evaluating on open-ended VizWiz examples and the diverse MMMU-Pro benchmark, but future work should explore even broader task distributions. Finally, we only evaluate on English-language datasets and conduct user studies with fluent English speakers. Future work should explore multilingual generalization.

Ethics Statement

This work was conducted with careful consideration of ethical implications. Human studies were conducted with informed consent through Prolific, with participants compensated at rates exceeding US minimum wage ($2 base + performance bonuses, see Section K.3.2). Participants were native English speakers from the US with higher education credentials (Table 17). We acknowledge this may limit generalizability across demographic groups.

While our work aims to improve accessibility for blind and low-vision users by calibrating trust in VLM outputs, we recognize potential risks: (1) users might over-rely on quality scores themselves, (2) errors in quality assessment could disproportionately affect vulnerable populations, and (3) current English-only implementation excludes non-English speakers. Future work should address these limitations through multilingual support and extensive testing with blind/low-vision communities.

The VizWiz dataset used contains questions from blind and low-vision users but is fully anonymized. In our human studies, the annotators were anonymized. We filtered NSFW content (Section K.2) and did not collect any additional personal information. All model outputs were manually reviewed for offensive content before presentation to study participants.

A-OKVQA and MMMU-Pro are released under Apache-2.0 license. VizWiz is released under a CC BY 4.0 license. GPT-4o is accessed via OPENAI API under their terms of service. LLaVA-v1.5-7B and Qwen2.5-VL-7B are both under Apache-2.0 license. soumyasanyal/nli-entailment-verifier-xxl model we used for the NLI task (see Section D.1) is under a CC By 4.0 license.

Our use of A-OKVQA and VizWiz aligns with their intended research purposes for evaluating VQA systems. We use language models consistent with their intended purposes.

We used AI assistants (e.g., ChatGPT) for grammar checking and code debugging, but not for producing substantive scientific content or analysis. All model outputs were reviewed and verified by the authors.

References

Na Min An, Eunki Kim, Wan Ju Kang, Sangryul Kim, Hyunjung Shim, and James Thorne. 2025. Can lvlms and automatic metrics capture underlying preferences of blind and low-vision individuals for navigational aid? *Preprint*, arXiv:2502.14883.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–16.

Marc Brysbaert. 2019. How many words do we read per minute? a review and meta-analysis of reading rate. *Journal of memory and language*, 109:104047.

Fateme Hashemi Chaleshtori, Atreya Ghosal, Alexander Gill, Purbid Bambroo, and Ana Marasović. 2024. On evaluating explanation utility for human-ai decision making in nlp. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Hanjie Chen, Faeze Brahman, Xiang Ren, Yangfeng Ji, Yejin Choi, and Swabha Swayamdipta. 2023. Rev: Information-theoretic evaluation of free-text rationales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2007–2030.

Shi Feng and Jordan Boyd-Graber. 2022. Learning to explain selectively: A case study on question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. Roscoe: A suite of metrics for scoring step-by-step reasoning. *Preprint*, arXiv:2212.07919.

Ana Valeria González, Gagan Bansal, Angela Fan, Yashar Mehdad, Robin Jia, and Srinivasan Iyer. 2021. Do explanations help users detect errors in open-domain qa? an evaluation of spoken vs. visual explanations. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1103–1116.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *AAAI Conference on Artificial Intelligence*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. 2020. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4351–4367.

Mina Huh, Fangyuan Xu, Yi-Hao Peng, Chongyan Chen, Hansika Murugu, Danna Gurari, Eunsol Choi, and Amy Pavel. 2024. Long-form answers to visual questions from blind and low vision people. In *First Conference on Language Modeling*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Weina Jin, Xiaoxiao Li, and Ghassan Hamarneh. 2023. Why is plausibility surprisingly problematic as an xai criterion? *arXiv preprint arXiv:2303.17707*.

Brihi Joshi, Ziyi Liu, Sahana Ramnath, Aaron Chan, Zhewei Tong, Shaoliang Nie, Qifan Wang, Yejin Choi, and Xiang Ren. 2023. Are machine rationales (not) useful to humans? measuring and improving human utility of free-text rationales. In *61st Annual Meeting of the Association for Computational Linguistics*.

Junhyeok Kim, Jaewoo Park, Junhee Park, Sangeyl Lee, Jiwan Chung, Jisung Kim, Ji Hoon Joung, and Youngjae Yu. 2025. Guidedog: A real-world egocentric multimodal dataset for blind and low-vision accessibility-aware guidance. *Preprint*, arXiv:2503.12844.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Russ Salakhutdinov, and Daniel Fried. 2024. VisualWebArena: Evaluating multimodal agents on realistic visual web tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 881–905, Bangkok, Thailand. Association for Computational Linguistics.

Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. 2023. Aligning text-to-image models using human feedback. *Preprint*, arXiv:2302.12192.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Jiacheng Liu, Wenya Wang, Dianzhuo Wang, Noah A Smith, Yejin Choi, and Hannaneh Hajishirzi. 2023b. Vera: A general-purpose plausibility estimation model for commonsense statements. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1264–1287.

Jiaying Lu, Xin Ye, Yi Ren, and Yezhou Yang. 2022. Good, better, best: Textual distractors generation for multiple-choice visual question answering via reinforcement learning. *Preprint*, arXiv:1910.09134.

Stephanie M. Lukin, Felix Gervits, Cory J. Hayes, Pooja Moolchandani, Anton Leuski, John G. Rogers III, Carlos Sanchez Amaro, Matthew Marge, Clare R. Voss, and David Traum. 2018. ScoutBot: A dialogue system for collaborative navigation. In *Proceedings of ACL 2018, System Demonstrations*, pages 93–98, Melbourne, Australia. Association for Computational Linguistics.

Shuai Ma, Xinru Wang, Ying Lei, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2024. "are you really sure?" understanding the effects of human self-confidence calibration in ai-assisted decision making. *Preprint*, arXiv:2403.09552.

Pranava Madhyastha, Josiah Wang, and Lucia Specia. 2019. Vifidel: Evaluating the visual fidelity of image descriptions. *Preprint*, arXiv:1907.09340.

Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. 2023. Receval: Evaluating reasoning chains via correctness and informativeness. *Preprint*, arXiv:2304.10703.

Soumya Sanyal, Tianyi Xiao, Jiacheng Liu, Wenya Wang, and Xiang Ren. 2024. Are machines better at complex reasoning? unveiling human-machine inference gaps in entailment verification. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10361–10386.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162.

Chenglei Si, Navita Goyal, Sherry Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé III, and Jordan Boyd-Graber. 2024a. Large language models help humans verify truthfulness – except when they are convincingly wrong. *Preprint*, arXiv:2310.12558.

Chenglei Si, Navita Goyal, Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé Iii, and Jordan Boyd-Graber. 2024b. Large language models help humans verify truthfulness–except when they are convincingly wrong. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1459–1474.

Judith Sieker, Simeon Junker, Ronja Utescher, Nazia Attari, Heiko Wersing, Hendrik Buschmeier, and Sina Zarrieß. 2024. The illusion of competence: Evaluating the effect of explanations on users' mental models of visual question answering systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19459–19475.

Tejas Srinivasan and Jesse Thomason. 2025. Adjust for trust: Mitigating trust-induced inappropriate reliance on ai assistance. *arXiv preprint arXiv:2502.13321*.

Tobias Groot Matias Valdenegro-Toro. 2024. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. In *The Fourth Workshop on Trustworthy Natural Language Processing (TrustNLP)*.

Kailas Vodrahalli, Tobias Gerstenberg, and James Y Zou. 2022. Uncalibrated models can improve human-ai collaboration. *Advances in Neural Information Processing Systems*, 35:4004–4016.

Xinru Wang and Ming Yin. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, pages 318–328.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022a. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States. Association for Computational Linguistics.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2022b. Measuring association between labels and free-text rationales. *Preprint*, arXiv:2010.12762.

Weihao Xuan, Qingcheng Zeng, Heli Qi, Junjue Wang, and Naoto Yokoya. 2025. Seeing is believing, but how much? a comprehensive analysis of verbalized calibration in vision-language models. *Preprint*, arXiv:2505.20236.

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. 2025. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *Preprint*, arXiv:2409.02813.

A Related Work

Our work builds on insights from evaluation of explanation quality and their utility for decision support, and adapt them for a vision-language setting.

Evaluating Explanation Quality. Prior work on evaluating explanation quality has largely focused on text-only domains, proposing metrics that assess informativeness (Chen et al., 2023), prediction support (Wiegreffe et al., 2022b; Hase et al., 2020), acceptability or helpfulness to users (Wiegreffe et al., 2022a), and consistency across reasoning chains (Golovneva et al., 2023; Prasad et al., 2023). Our work goes beyond textual plausibility by introducing two novel explanation qualities tailored for vision-language settings. Visual Fidelity has been explored for evaluating faithfulness of model-generated image captions (Madhyastha et al., 2019; Lee et al., 2023). Our work builds on these ideas to do zero-shot evaluation of explanations. Separately, recent work has investigated the calibration of verbalized confidence from models directly (Xuan et al., 2025), whereas we derive calibration signals from the explanation content itself.

Explanations for Decision Support. While some studies in open-domain QA suggest that explanations can help users detect model errors (González et al., 2021), many researches on utility of natural language explanations for decision support have found that plausible-sounding explanations (Jin et al., 2023) often mislead users into overrelying on inaccurate model predictions (Joshi et al., 2023; Si et al., 2024b; Chaleshtori et al., 2024; Sieker et al., 2024). Our work adds to these findings by showing that evaluating and communicating explanation qualities to users can reduce overreliance.

B Code and Data Availability

Our codebase is available at `https://anonymous.4open.science/r/vlm-quality-scores-31D9`.

C Prompts used for Answers and Rationales Generation

We adopt a two-step, post-hoc justification strategy ($IQ \rightarrow aE$) inspired by (Wiegreffe et al., 2022b). In the first step, we prompt the model to produce only its answer prediction, without any accompanying rationale. In the second step, we present that predicted answer back to the model and ask it to "Please explain the reasoning behind your answer." Note that for MMMU-Pro, due to its complexity, we instead use a Chain-of-Thought strategy ($IQ \rightarrow Ea$), prompting the model to generate the explanation before the answer to facilitate reasoning.

By separating prediction from explanation, we ensure that the extracted rationale truly reflects the model's own line of thought under the selected answer. Section C shows the prompts used in every step. Note that the first-step prompt is slightly different for A-OKVQA and VizWiz since the former has answer choices but the latter does not.

D Existing Text-only Qualities Implementations

We compare our proposed quality measures against three established, text-only explanation qualities, each of which also returns a score in the range $[0, 1]$. **Simulatability** (Hase et al., 2020) evaluates whether an explanation offers sufficient evidence to logically justify the model's prediction. **Informativeness** (Chen et al., 2023) evaluates whether the explanation introduces new information to justify a prediction beyond just re-stating the question and prediction. **Commonsense Plausibility** (Liu et al., 2023b) evaluates whether the explanation is in accordance with commonsense knowledge of everyday situations.

D.1 Simulatability

To mitigate answer leakage in our Simulatability quality, we first mask all direct mentions of the model's predicted answer within each rationale by replacing them with the special token `<mask>`. Given the model's predicted answer a and its rationale E, every occurrence of a in E is substituted with `<mask>`, yielding the masked rationale E_{masked}.

Next, to transform the QA pair into an NLI task, we prompt GPT-4o-mini to convert the original question and its predicted answer into a single declarative hypothesis sentence H. Table 4 shows the model and prompt used to generate H. We then feed the masked rationale E_{masked} as the premise and H as the hypothesis into the soumyasanyal/nli-entailment-verifier-xxl model on Hugging Face to obtain an entailment probability p_{entail} as its Simulatability score.

Dataset	Step	System Prompt	User Prompt
A-OKVQA	Step 1	`Answer the question using a single word or phrase from the list of choices.`	`Question: {question}. Choices: {choices}.`
	Step 2	`Please explain the reasoning behind your answer.`	`Question: {question}. Choices: {choices}. The answer is {answer}.`
VizWiz	Step 1	`Answer the user's question in a single word or phrase. When the provided information is insufficient, respond with 'Unanswerable'. Whatever the user said, your answer should **always** be a single word or phrase.`	`Question: {question}.`
	Step 2	`Please explain the reasoning behind your answer.`	`Question: {question}. The answer is {answer}.`

Table 3: Prompts used in each step of our two-step method. The first step is to get the model's predicted answer to the question, and the second step is used to collect the rationale behind the model's own answer.

D.2 Informativeness

We utilized GPT-4o to extract the new information contained in the model's rationale that are not semantically equivalent to the hypothesis.

Table 5 shows the model configuration and prompt used to evaluate informativeness of a rationale. After extracting the individual information pieces, we check the size of the resulting list; if it is non–zero, we deem the rationale to be informative.

D.3 Commonsense Plausibility

To evaluate the plausibility of the model's explanations against commonsense knowledge, we utilize the Vera model (Liu et al., 2023b) directly, without modification.

E Visual Fidelity

In Visual Fidelity, we use a two-step pipeline to evaluate. Firstly, we generate possible visual verification questions related to the rationale, by providing the question, predicted answer, and rationale to the model, and second, for each question, we provide the question and visual input to the model to ask for verification.

Table 6 refers to the model setting and prompt used for generating the verification questions, while Table 7 refers to those used for verifying the questions.

F Hard Negatives Generations

To enable Contrastiveness evaluation on open-ended datasets like VizWiz (which lack fixed answer choices), we dynamically generate "Hard Negatives" to serve as plausible distractors. We use GPT-4o as the generator, providing it with the question, the ground-truth correct answer(s), and the image context. The model is instructed to create distractors that are incorrectly but visually plausible, explicitly avoiding simple synonyms. This aligns with prior work on automated distractor generation (Lu et al., 2022); however, unlike previous learning-based methods, we employ a training-free approach using an off-the-shelf LLM. Table 8 shows the exact prompt used.

G Validation of Hard Negatives Strategy

To validate our strategy of dynamically generating hard negatives for open-set evaluation (as used in VizWiz), we conducted an ablation study on the A-OKVQA dataset. A-OKVQA typically provides 4 ground-truth answer choices (Closed-Set). We stripped these choices and instead generated hard negatives using the same pipeline applied

Config	Assignment
Model	`gpt-4o-mini-2024-07-18`
Max Tokens	1024
Temperature	0.1

User Prompt	`Integrate the question and the answer into one sentence. For example, given the question "What is the man waiting for?" and the answer "taxi", you should output "The man is waiting for taxi."`
	`Question: {question}`
	`Answer: {answer}`

Table 4: Model configuration and prompt used to generate descriptive sentence

to VizWiz, effectively creating a "Reconstructed Closed-Set" version of A-OKVQA.

Table 9 compares the performance of our Contrastiveness quality on the original Closed-Set vs. the Reconstructed Closed-Set. We observe a consistent drop in Discriminability (~ 0.13) and a rise in ECE (~ 0.08) across all models. This indicates that the generated hard negatives successfully make the task more challenging, likely because the generated distractors are semantically closer to the correct answer than the original multiple-choice options. However, crucially, the discriminability scores remain positive and statistically significant (e.g., 0.158^{***} for Qwen2.5-VL-7B), confirming that our metric remains valid and robust even in this harder, constructed setting.

To further understand why the generated negatives increase difficulty, we analyzed the semantic proximity of the distractors to the correct answer using `all-mpnet-base-v2`. Across the 500-sample set, the mean cosine similarity for Original Negatives was 0.424, whereas for Generated Negatives it rose to 0.468 ($\Delta + 0.044$).

While A-OKVQA provides a rigorous baseline with plausible distractors, our Reconstructed Closed-Set method produces even more semantically proximal candidates. As shown in Table 10, the generated options (e.g., "buttercream" vs "frosting") often require finer-grained visual discrimination than original distractors (e.g., "paint" vs "frosting"). This suggests that LLM-generated hard negatives serve as a harder mode for evaluation, explaining the lower discriminability scores while verifying the metric's robustness efficiently.

H Uncertainty Quantification Baselines

To situate our explanation quality metrics within broader uncertainty quantification literature, we compare them against a standard logit-based baseline. Specifically, for the LLaVA-v1.5-7B model (where we have access to output probabilities), we compute the softmax probability of the first token of predicted answer as a measure of model confidence.

Table 11 presents the comparison on two contrasting datasets: A-OKVQA (simpler visual reasoning) and MMMU-Pro (complex, multi-discipline reasoning).

On A-OKVQA, the model's intrinsic "intuition" (logits) is highly effective, achieving strong Discriminability (0.252) and excellent ECE (0.046). This suggests that for simpler tasks, raw confidence is a sufficient reliability indicator.

On MMMU-Pro, however, this signal breaks down. Logits yield poor discriminability (0.055, not significant) and high calibration error (0.356). In contrast, our explanation quality metrics—particularly the product of Visual Fidelity and Contrastiveness (S_{Prod})—remain robust, achieving statistically significant discriminability (0.090^{**}) and much better calibration (0.144).

These results indicate that logits (which measure *answer* confidence) and our metrics (which measure *explanation* reliability) are distinct signals. While logits suffice for simpler tasks, explanation quality offers a necessary and robust alternative for complex reasoning, particularly when model confidence becomes unreliable or inaccessible.

From Table 12, we show our evaluation framework on four representative VQA instances. For each example, we provide the original question (and answer choices where applicable), image,

Config	Assignment
Model	`gpt-4o-2024-05-13`
Max Tokens	1024
Temperature	0.1

User Prompt	`Please break the following rationale into distinct pieces, and keep only the ones that are not semantically equivalent to the hypothesis. Output the final answer in a Python list format.`
	
	`Example:`
	`Hypothesis: The man by the bags is waiting for a delivery.`
	`Rationale: The man by the bags is waiting for a delivery, as indicated by the presence of the suitcases and the fact that he is standing on the side of the road. The other options, such as a skateboarder, train, or cab, do not seem to be relevant to the situation depicted in the image.`
	`Output: ["Suitcases are present in the image.", "The man is standing on the side of the road.", "The other options, such as a skateboarder, train, or cab, do not seem to be relevant to the situation depicted in the image."]`
	
	`Task:`
	`Hypothesis: {hypothesis}`
	`Rationale: {rationale}`

Table 5: Model configuration and prompt used to evaluate informativeness of a rationale.

ground truth answer, the model's selected answer and rationale, and the corresponding scores for Simulatability, Informativeness, Commonsense Plausibility, Visual Fidelity, and Contrastiveness. Note that for VizWiz, the ground-truth answer is defined as the majority vote among the 10 user annotations from the original dataset, and since VizWiz provides no answer choices, we do not compute a Contrastiveness score.

I VizWiz Dataset Limitations and Negative Discriminability

We observed that for VizWiz, particularly with GPT-4o, the discriminability scores for our proposed metrics (and baselines) are often low or negative (see Table 1). A negative discriminability score implies that the model assigns higher quality scores to "incorrect" predictions than to "correct" ones.

Upon qualitative inspection, we found this phenomenon is largely driven by the "Ground Truth Problem" inherent in the VizWiz dataset. The ground truth answers are often narrow or specific, causing valid model predictions to be labeled as incorrect. Table 13 shows three representative examples:

- **Setup Windows Example**: The ground truth includes "windows 7 home premium", but the GPT-4o model predicts "set up windows", which accurately describes the visible title. This is marked incorrect, yet receives high quality scores.
- **Toronto Example**: The ground truth is generic ("tall buildings"), but the GPT-4o model correctly identifies "downtown Toronto". This specific identification is penalized, again resulting in high quality scores for a "wrong" answer.
- **Keyboard Example**: The ground truth lists various terms for a screen (e.g., "monitor", "laptop"), but Qwen2.5-VL-7B predicts "key-

board", identifying the other prominent object in the blurry image. The explanation correctly describes the keyboard's features, leading to high fidelity despite the answer mismatch.

To quantify this, an expert annotator manually inspected 20 randomly sampled instances where the model prediction was marked "incorrect" but achieved a Visual Fidelity score > 0.9. We found that for GPT-4o, 60% (12/20) of these cases were actually valid answers rejected by the ground truth (e.g., specific vs generic). This rate was lower for Qwen2.5-VL-7B (25%, 5/20) and LLaVA-v1.5-7B (15%, 3/20), confirming that GPT-4o's "hallucinations" on VizWiz are frequently artifacts of valid but mismatched reasoning.

Because GPT-4o often provides these "over-answering" or semantically valid but lexically distinct responses, it accumulates a large number of high-quality "incorrect" samples, which inverts the discriminability metric. This effect is less pronounced in Qwen2.5-VL-7B and least in LLaVA-v1.5-7B, which tend to generate shorter, simpler answers closer to the crowd-sourced ground truths. Thus, the negative discriminability reflects a limitation of the evaluation ground truth rather than a failure of the explanation quality metrics.

J Reliability and Validity of the Trust Cues

J.1 Validation on Agreement with Human Judgements

To validate the reliability of our automated Visual Fidelity and Contrastiveness qualities, we conducted an internal sanity check on a held-out sample. We randomly selected 100 examples from our evaluation sets and had a single expert annotator manually assess each instance. The instruction provided to the expert annotator is shown at Table 16. We binarized our scores at a threshold of 0.5 (scores ≥ 0.5 map to 1, else 0) and compared these labels to the human judgments using Cohen's κ:

$$\kappa_{\text{VF}} = 0.51 \quad \text{and} \quad \kappa_{\text{Contr.}} = 0.44,$$

indicating moderate agreement.

J.2 Accuracy of LLM-generated verification questions and answers

To quantify hallucination risk in the VF pipeline, we sampled 50 explanations each from A-OKVQA and VizWiz (total 210 verification questions produced by m_{QGen} and corresponding answers from m_{Verif}). Three expert annotators judged (i) whether each generated verification question was correctly grounded in the rationale and image, and (ii) whether the verifier's answer was correct for questions deemed evaluable from the image. We observed that 95.2% of generated questions were correctly grounded and 95.0% of verifier answers were correct. These results support using LLM-as-judge as a cost-efficient approximation to the ideal ground-truth VF scorer.

J.3 Verifier robustness for Visual Fidelity

We use GPT-4o by default for both question generation (m_{QGen}) and verification (m_{Verif}) to isolate pipeline effects. To assess robustness to verifier choice, we swap in smaller open models (Gemma-3n-E4B, Qwen2.5-VL-7B) and report changes in Discriminability (Disc; higher is better) and Expected Calibration Error (ECE; lower is better) on held-out slices.

As shown in Table 14 and Table 15, across A-OKVQA and VizWiz, all three verifiers yield positive discriminability on average (GPT-4o: 0.109; Gemma-3n-E4B: 0.061; Qwen2.5-VL-7B: 0.154), and ECE remains comparably low (0.168/0.197/0.196 respectively). Thus, VF continues to separate correct from incorrect predictions and stays well-calibrated even with using smaller open models as verifiers.

K User Studies

K.1 Annotator filtering

To ensure high-quality annotations, we recruited annotators on Prolific using the criteria listed in Table 17. All participants identities were fully anonymized. All annotators provided informed consent through Prolific's platform, which clearly described the task purpose, data use, and compensation prior to participation.

This study protocol was reviewed and approved by our institution's Institutional Review Board (the name of the institution will be included in the camera-ready version).

All participants were required to be native English speakers residing in the United States, with approval rates between 98–100% and at least 1000 prior submissions. We further restricted enrollment to individuals holding at least an undergraduate degree. To prevent contamination across experi-

mental conditions, each participant was confined to a single "setting" (i.e., one quality-type configuration).

K.2 Dataset Splits and User Study Design

Metric Evaluation Sets (500 instances) For the quality evaluations detailed in Section 3, we selected three 500-instance sets. From A-OKVQA, we used the first 500 examples from the official validation split. From VizWiz, we selected the first 500 validation examples where the majority human annotation was not "unanswerable" and which contained no NSFW content, as filtered by GPT-4o. From MMMU-Pro, we used the first 500 examples with single-image contexts from the *standard (10 options)* set.

Human Study Subsets (100 instances) We conducted a between-subjects user study where participants predicted VLM correctness based on varying levels of explanation quality scores. To facilitate this, we derived a subset of 100 questions each from the A-OKVQA and VizWiz evaluation sets described above.

These subsets were balanced to contain 50 correct and 50 incorrect predictions by LLaVA-v1.5-7B. To ensure the subsets remained representative of the larger distribution, we generated 50 random candidate subsets and selected the one with the lowest Expected Calibration Error (ECE) relative to the full 500-instance set. Specifically, for A-OKVQA, we minimized the average ECE across Visual Fidelity and Contrastiveness; for VizWiz, we minimized the ECE for Visual Fidelity (as Contrastiveness scores were unavailable during the subset creation phase). As shown in Table 18 and Table 19, this selection maintains a distribution of quality scores consistent with the full set from Section 3.

Study Protocol In the study, each question was annotated by three unique participants, yielding 300 annotations per dataset per condition. Each participant completed a batch of 10 questions, balanced to include five instances where the model was correct and five where it was incorrect; participants were not informed of this distribution. Users were randomly assigned to a single study condition (control or a treatment variant) and did not participate in multiple conditions. In total, we collect 300 human annotations per dataset and condition (3 users × 100 questions), capturing user reliance behavior under varying explanation and quality signal conditions.

K.3 Attention Incentives

To ensure that participants engaged carefully with each instance's annotation, we combined a per-item timer with small monetary bonuses and penalties.

K.3.1 Timer Implementation

To ensure that participants spent sufficient time on each stage (and did not simply skim and click through), we imposed a per-item timer in all of our human studies. The user may only start their selection after the timer ends. For each question, we computed the explanation's "reading time" as

$$\text{reading_time} = \frac{\#\text{words in explanation}}{238 \text{ words/minute}}$$

where 238 wpm is the average adult reading speed (Brysbaert, 2019).

We then capped the total display time at

$$\text{reading_time} + 10 \text{ seconds}$$

to cover the question, answers, and any qualites shown.

K.3.2 Bonus Payments

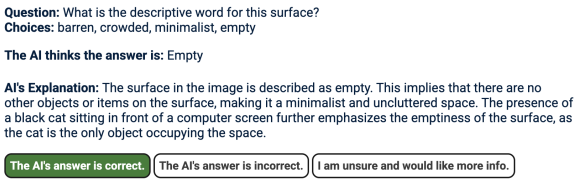
To further motivate careful reading and discourage guessing, we tied each response to a small bonus bank: correct answers earned $+\$0.10$; incorrect answers incurred $-\$0.10$ (with the bonus bank floored at $\$0$ so it would not harm their base payments); selecting "I'm unsure based on the information provided" resulted in no change. Participants were paid their accumulated bonus in addition to the base participation fee of $\$2$ for annotating 10 instances.

K.4 User Studies Settings & Examples

In the main user study, we present 14 settings in total. Table 20 summarizes all 14 different settings and Figure 9 shows examples of all 8 different categories of messages presented to users. Table 21 contains the main user-study results of user behavior patterns.

K.5 Supplementary Human Studies

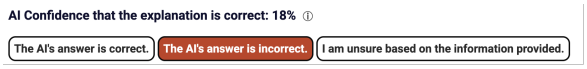
In addition to the study described in §4.2 and Section K.4, we ran a supplementary user study to evaluate how explanations and qualities affect users' trust in model predictions. Each of the 10 questions was presented in three successive stages; after each stage, participants indicated whether they believed the model's answer was correct or were unsure:



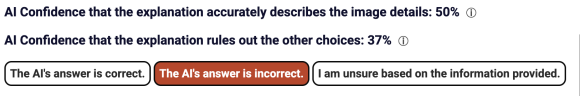
(a) Control (no quality)

(b) VF – Numeric quality

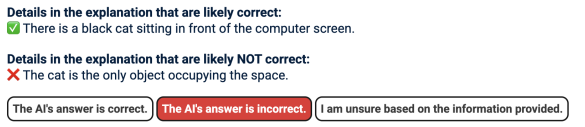
(c) Contr. – Numeric quality



(d) VF×Contr. (shown as a simple confidence score) – Numeric quality

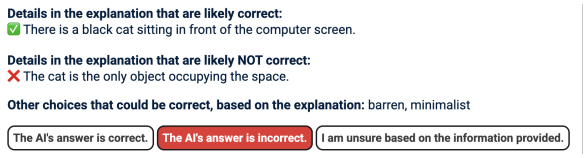


(e) Both VF and Contr. – Numeric quality



(f) VF – Descriptive quality

(g) Contr. – Descriptive quality



(h) Both VF and Contr. – Descriptive quality

Figure 9: Explanation quality messages for each instruction condition. Subfigure a is the baseline with no quality displayed; the next three (b–e) show numeric qualities; the remaining four (f–h) show descriptive ones.

1. Answer Only: Participants viewed only the question, answer choices (if available), and model prediction.

2. With Explanation: Participants were provided with AI-generated rationales alongside predictions.

3. With Explanation + Quality: Qualities (varied from our experiment settings, e.g. Visual Fidelity and Contrastiveness) were displayed alongside explanations.

This three-stage design of the user study enables us to track how users' confidence in the model's correctness evolves as they receive additional information.

Timed Stages in Supplementary Human Studies To standardize attention across conditions, we also enforced a stage-specific timer, where the users can only make their selections at a stage after the timer at that stage ends:

1. Answer Only: fixed 5 seconds

2. With Explanation: explanation reading time (words / 238 wpm) (roughly 10–40 seconds)

3. Explanation + Quality: fixed 5 seconds

Bonus Payments in Supplementary Human Studies Participants were paid a $2 base fee and could earn up to $1 in performance-based bonuses, which were awarded only during Stage 3 (Explanation + Quality; see Section K.3.2).

Table 22 shows that as users progress from seeing only the model's answer to viewing explanations and then explanations with quality scores, their unsure rate steadily decreases while user accuracy correspondingly increases, with the largest gains observed when descriptive qualities are provided alongside explanations. These findings support our findings that richer, more interpretable quality signals can meaningfully improve users' trust calibration.

L Computational Resources Spent & Total Cost

The complete breakdown of all monetary expenditures across the study is given in Table 23. In brief, the computational expenses include six hours of A100 GPU time for running the vision-language models, plus OpenAI API calls for generating predicted answers, explanations, and computing qual-

ities. Human evaluation costs cover two Prolific studies—a one-stage study as shown in Section K.4 and a three-stage follow-up from Section K.5 (including up to $1.00 in per-participation bonuses; see Section K.3.2).

Altogether, these sum to an overall expense of approximately $2,827.

Config	Assignment
Model	gpt-4o-2024-08-06
Max Tokens	1024
Temperature	0.1

User prompt: You will be shown a question about an image, along with an answer, and a rationale that explains the answer based on details from the image. Your task is to generate a list of yes/no questions that verify the details about the image that are **explicitly** mentioned in the rationale. Your questions should be phrased such that the answer to that question being yes means that the detail in the rationale is correct. Focus on creating questions that can be visually verified or refuted based on the details provided in the rationale. Ensure the questions are specific and directly pertain to aspects that are visually relevant and mentioned in the rationale. Avoid generating questions about elements that are not mentioned in the rationale, or the rationale explicitly states are not relevant or present. Also avoid generating multiple questions that check for the same visual detail.

Here is one example:
Input:
Question: Why is the person wearing a helmet?
Answer: For safety
Rationale: The person is wearing a helmet because they are riding a bicycle on a busy city street. Helmets are commonly used to protect against head injuries in case of accidents, especially in areas with heavy traffic.

Good Questions:
1. Is the person wearing a helmet while riding a bicycle?
Reason: This question is directly answerable by observing whether the person on the bicycle is wearing a helmet in the image.
2. Is the street in the image busy with traffic?
Reason: This question can be visually verified by looking at the amount of traffic on the street in the image.

Bad Questions:
1. Is the person wearing the helmet because they are concerned about head injuries?
Reason: This question is not good because it assumes the person's intentions or concerns, which cannot be visually verified from the image.
2. Does wearing a helmet suggest that the person is highly safety-conscious?
Reason: This question relies on inference and external knowledge about the person's mindset, rather than on observable details from the image.
3. Is there any indication that the person is wearing a helmet for safety reasons?
Reason: This question verifies the answer to the original question, rather than verifying a detail about the image that's mentioned in the rationale.
4. Is the person wearing a safety vest?
Reason: This question is not good because it tries to verify details about the image that are not explicitly mentioned in the rationale.
5. Is the person not wearing sunglasses?
Reason: This question is not good because it asks for verification by absence and can only be answered with a "no," which is not the preferred type of question.

Respond with a list of (good) questions (without the reasons), starting from '1. '

Table 6: Model configuration and prompt used to generate verification visual questions of a rationale.

Config	Assignment
Model	gpt-4o-2024-08-06
Max Tokens	1024
Temperature	0.1

User prompt: Question: {question}. Based on the information provided in the image, answer with 'yes' or 'no'. Provide one-word answer only.

Table 7: Model configuration and prompt used to verify the visual questions generated from Table 6.

Hard Negatives Generation Prompt
Given the following question, its correct answer(s), and the accompanying image, generate exactly THREE (3) different "Hard Negative" answers. Context: Question: {question} Correct Answer(s): {valid_answers_str} Your Goal: Generate 3 "Hard Negative" distractors. A "Hard Negative" must be: 1. **Incorrect**: It must NOT be a synonym or a close grammatical variation of the correct answer(s). 2. **Visually Plausible**: Ideally, it refers to something else *visible in the image* or common in this context. 3. **Challenging**: It should be a reasonable guess for a Model that didn't pay perfect attention effectively. Constraints: - **NO Synonyms**: If the answer is "bike", do NOT generate "bicycle". - **Format**: Provide exactly THREE (3) comma-separated phrases. No numbering, no extra text. - **Length**: Short, concise (1-3 words).

Table 8: System prompt used to generate hard negatives for Contrastiveness evaluation on open-ended datasets (VizWiz).

Model	**Metric**	**Closed-Set** (Original)	**Reconstructed Closed-Set** (Generated Hard Negatives)	**Delta**
GPT-4o	Disc (↑)	0.248^{***}	0.111^{*}	-0.137
	ECE (↓)	0.210	0.312	$+0.102$
Qwen2.5-VL-7B	Disc (↑)	0.283^{***}	0.158^{***}	-0.125
	ECE (↓)	0.136	0.235	$+0.099$
LLaVA-v1.5-7B	Disc (↑)	0.243^{***}	0.112^{***}	-0.131
	ECE (↓)	0.176	0.233	$+0.057$

Table 9: Ablation study on A-OKVQA comparing standard Closed-Set evaluation (using ground-truth options) vs. Reconstructed Closed-Set evaluation (using generated hard negatives) for the Contrastiveness quality. The generated negatives significantly increase task difficulty.

Image	**Question**	**Correct**	**Original Negatives** (Sim)	**Generated Negatives** (Sim)
	What kind of coating has been used?	frosting	polish, paint, varnish ($S_{\text{sim}} = 0.27$)	buttercream, whipped cream, glaze ($S_{\text{sim}} = 0.58$)
	In what type of location are they playing with the body board?	room	beach, park, store ($S_{\text{sim}} = 0.28$)	living room, hallway, garage ($S_{\text{sim}} = 0.58$)

Table 10: Comparison of Original (Closed-Set) vs. Generated (Reconstructed Closed-Set) distractors for matching A-OKVQA questions. The generated negatives are often semantically closer to the correct answer (higher similarity S_{sim}), making the discrimination task harder.

	A-OKVQA (Simpler)		**MMMU-Pro** (Harder)	
Metric	**Disc.** ↑	**ECE** ↓	**Disc.** ↑	**ECE** ↓
Model Confidence (Logits)	0.252***	**0.046**	0.055	0.356
VF	0.181***	0.197	**0.169****	0.558
Contr.	0.243***	0.176	0.047	0.221
Avg(VF, Contr.)	0.212***	0.109	0.108**	0.360
Prod(VF, Contr.)	**0.320*****	0.164	0.090**	**0.144**

Table 11: Comparison of our explanation quality metrics against a **Logit-based Baseline** (softmax probability of the predicted answer's first token) on LLaVA-v1.5-7B. Significance is evaluated using an unpaired t-test: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

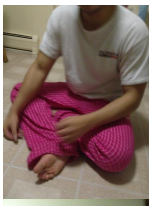
Dataset	Question	Answer Choices	Image	Ground Truth	Predicted Answer	LLaVA-v1.5-7B Explanation	Sim.	Info.	Plau.	VF	Contr.
A-OKVQA	What is the white substance on top of the cupcakes?	butter, mayo, ice cream, icing		icing	icing	The white substance on top of the cupcakes is icing. This can be inferred from the fact that icing is commonly used as a topping for cupcakes to enhance their flavor and appearance. In the image, there are three cupcakes with icing on top, which indicates that they have been decorated with icing as part of their presentation. Additionally, icing is a common ingredient in cupcake recipes, so it is reasonable to assume that the white substance on top of the cupcakes is indeed icing.	0.391	1.000	0.473	1.000	0.996
A-OKVQA	What is the descriptive word for this surface?	barren, crowded, minimalist, empty		crowded	empty	The surface in the image is described as empty. This implies that there are no other objects or items on the surface, making it a minimalist and uncluttered space. The presence of a black cat sitting in front of a computer screen further emphasizes the emptiness of the surface, as the cat is the only object occupying the space.	0.964	1.000	0.566	0.500	0.393
VizWiz	What color are those pants?	blue, red, green		pink, pink white checkered	pink	The man is wearing pink pajamas, which is an unusual choice for a casual setting.	0.025	1.000	0.397	1.000	0.073
VizWiz	What is in this bottle?	pen, laser pointer, small torch		bacon bits, hormel bacon bits, bacon	spices	The person holding the bottle of spices is likely using it to season their food, adding flavor and aroma to various dishes. Spices are commonly used in cooking as they can enhance the taste of ingredients and create unique flavor combinations. The hand holding the bottle suggests that the person might be preparing a meal or experimenting with different spice blends for their culinary creations.	0.999	1.000	0.566	0.500	0.555

Table 12: Four VQA examples from both A-OKVQA and VizWiz datasets with model predictions, explanations, and qualities. The predicted answer and explanation are from LLaVA-v1.5-7B model. Answer choices in VizWiz examples (shown in gray) are generated negatives from the ground truth, which are not shown to the VLM model when generating the predictions and explanations and are only provided to compute the contrastiveness (shown in gray).

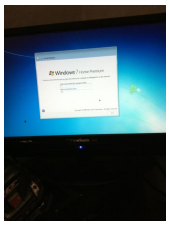
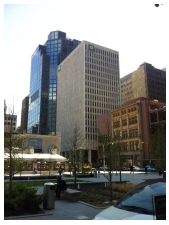
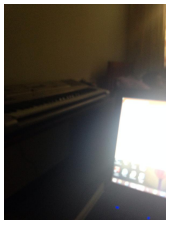
Question	Image	Ground Truth List	Model	Prediction	Explanation	VF	Contr.
What does the screen say?		windows 7 home premium, windows, windows home premium	GPT-4o	set up windows	The screen displays the setup process for Windows 7 Home Premium. It prompts the user to choose a user name for their account and name their computer to distinguish it on the network. The text on the screen includes: - "Set Up Windows" - "Windows 7 Home Premium" - "Choose a user name for your account and name your computer to distinguish it on the network." - "Type a user name (for example, John):" - "Type a computer name:" - "PC" - "Copyright © 2009 Microsoft Corporation. All rights reserved."	1.000	0.585
What is this?		tall buildings, buildings, city landscape, downtown city, city buildings, skyscrapers, city	GPT-4o	downtown Toronto	The image shows a cityscape with several tall buildings, one of which has the TD (Toronto-Dominion) logo prominently displayed. This suggests that the location is likely in downtown Toronto, where TD Bank has a significant presence. The mix of modern and older architecture is also characteristic of downtown Toronto. Therefore, the answer is indeed downtown Toronto.	1.000	0.659
What's this?		room, half computer screen, monitor, laptop, computer screen, computer monitor, tv	Qwen2.5-VL-7B	keyboard	The image shows a piece of electronic equipment that appears to be an electronic keyboard or synthesizer. It has keys similar to a piano and is positioned next to what looks like a computer monitor, suggesting it might be used for music production or performance. The lighting is dim, and the image quality is low, making it difficult to discern finer details, but the shape and layout of the keys are consistent with a keyboard.	1.000	0.999

Table 13: Examples of "Incorrect" VizWiz predictions that received high Explanation Quality scores. In these cases, the model provides a semantically valid or more specific answer than the ground truth, leading to a "False Negative" error judgment despite high Visual Fidelity and Contrastiveness.

Dataset	Model	Accuracy	Disc VF(GPT)	Disc VF(Gemma)	Disc VF(Qwen)
A-OKVQA	LLaVA-v1.5-7B	0.696	0.181^{***}	0.151^{***}	0.247^{***}
A-OKVQA	Qwen2.5-VL-7B	0.848	0.080^{***}	0.096^{**}	0.131^{***}
A-OKVQA	GPT-4o	0.910	0.085^{***}	0.044	0.156^{***}
VizWiz	LLaVA-v1.5-7B	0.557	0.240^{***}	0.098^{***}	0.193^{***}
VizWiz	Qwen2.5-VL-7B	0.838	0.042^{*}	0.024	0.092^{***}
VizWiz	GPT-4o	0.840	0.024	-0.048	0.104^{**}
Average		0.781	0.109	0.061	0.154

Table 14: Discriminability of VF scores across verifier models (higher is better). Significance: * $p<.05$, ** $p<.01$, *** $p<.001$.

Dataset	Model	Accuracy	ECE VF(GPT)	ECE VF(Gemma)	ECE VF(Qwen)
A-OKVQA	LLaVA-v1.5-7B	0.696	0.207	0.156	0.181
A-OKVQA	Qwen2.5-VL-7B	0.848	0.137	0.174	0.177
A-OKVQA	GPT-4o	0.910	0.099	0.209	0.173
VizWiz	LLaVA-v1.5-7B	0.557	0.271	0.303	0.310
VizWiz	Qwen2.5-VL-7B	0.838	0.136	0.134	0.161
VizWiz	GPT-4o	0.840	0.160	0.205	0.174
Average		0.781	0.168	0.197	0.196

Table 15: Expected Calibration Error (ECE) of VF scores across verifier models (lower is better).

Criterion	**Instruction**
Visual Fidelity	In this task, you'll view an image, and a question about the image. You will then see an answer to the question given by an AI model, along with an explanation. Your job is to evaluate whether the details in the explanation are consistent with the image that is shown. You'll select one of these choices: 0: The explanation does not mention any details / elements that are directly visible in the image (apart from the prediction itself). Or the explanation mentions details about the image, but one or more of those details are incorrect (they contradict what is visible in the image). 1: The explanation mentions details about the image, and all the details are consistent with the image. IMPORTANT NOTE: Your job is NOT to check the correctness of the AI model's answer. It could be that the answer or the logic in the response is incorrect, but the explanation talks about something that is directly in the image. It can also be that the answer is correct, but the explanation does not refer to the image or presents inconsistent details! IMPORTANT: See examples from this form (https://forms.gle/nvpcQ7fpLoFqsLiJ9) before proceeding!
Contrastiveness	In this task, you'll view a question about an image (without seeing the image). You will also see an answer from an AI model, along with an explanation (the model has access to the image). Your job is to evaluate whether the explanation meets these two key qualities: • You check if the explanation is consistent with the predicted answer. • You can ask yourself this question: "Does the explanation provide evidence that matches with the answer it gives?" You check if the explanation covers enough details to reject all other possible answers. You can ask yourself this question: "Does the explanation eliminate all other answers with proper justifications?" By eliminate, we mean that the explanation should provide a strong argument for the selection answer, or strong counter arguments for the other options. You'll decide whether the explanation has these qualities. IMPORTANT: See examples from this form (https://forms.gle/HYp6N44TpsjAoxHa6) before proceeding!

Table 16: Instructions Provided to the Expert Annotator

Config	Criteria
Location	United States
Current Country of Residence	United States
Primary Language	English
Approval Rate	98-100
Number of previous submissions	1000-10000
Highest education level completed	Undergraduate degree (BA/BSc/other), Graduate degree (MA/MSc/MPhil/other), Doctorate degree (PhD/other)
Exclude participants from other studies	Prohibit any user who took part in a different setting

Table 17: Prolific annotator filtering

Set	Accuracy	VF	Contr.	Prod(VF, Contr.)	Min(VF, Contr.)	Avg(VF, Contr.)	Support	Informative	Plausibility
Full (500)	0.696	0.207	0.176	0.164	0.147	0.109	0.288	0.332	0.162
Subset (100)	0.500	0.270	0.237	0.133	0.130	0.227	0.372	0.490	0.075

Table 18: Accuracy and Expected Calibration Error (ECE) for different qualities on the **A-OKVQA** dataset. "Subset" rows correspond to the 100-question user study subsets (50 correct / 50 incorrect), while "Full" rows represent the 500-instance evaluation set.

Set	Accuracy	VF	Support	Informative	Plausibility
Full (500)	0.557	0.271	0.305	0.429	0.110
Subset (100)	0.500	0.236	0.379	0.390	0.071

Table 19: Accuracy and Expected Calibration Error (ECE) for different qualities on the **VizWiz** dataset.

Setting	**Description**
Control	Show explanation only; no quality scores are shown.
Random Score	Random scoring baseline (uniform random distribution from $[0, 1]$), shown as a simple confidence score "AI Confidence that the explanation is correct".
Prod(VF, Contr)	Product of VF and Contrast scores, shown as a simple confidence score "AI Confidence that the explanation is correct".
Avg(VF, Contr)	Average of VF and Contrast scores, shown as a simple confidence score "AI Confidence that the explanation is correct".
VF num.	Show Visual Fidelity numeric score "AI Confidence that the explanation accurately describes the image details"
VF desc.	Show at most two descriptive sentences converted from visual questions which are verified by verifier VLM $m_{\texttt{Verif}}$, and at most two from questions which are not verified by $m_{\texttt{Verif}}$
Contr. num.	Show Contrastiveness score "AI Confidence that the explanation rules out the other choices"
Contr. desc.	Show the other answer options $a_j \neq a_0$ s.t. $\text{PR}_{\texttt{NLI}}(P \text{ entails } h_j) \geq 0.5$
Both num.	Display both VF num. and Contr. num. messages
Both desc.	Display both VF desc. and Contr. desc. messages
VF shown as Conf.	VF score displayed as a simple confidence score "AI Confidence that the explanation is correct"
Contr shown as Conf.	Contr. score displayed as a simple confidence score "AI Confidence that the explanation is correct"
Prod shown as VF	VF×Contr. score presented as a VF score "AI Confidence that the explanation accurately describes the image details"
Prod shown as Contr.	Combined VF×Contr. score presented as a Contr. score "AI Confidence that the explanation rules out the other choices"

Table 20: Summary of the 14 one-step human-study settings.

Dataset	Setting	Unsure Rate	Not Unsure Accept Rate	Not Unsure Accuracy	False Accept Rate	False Reject Rate
A-OKVQA	Control	08.7% ± 1.6%	76.3% ± 2.6%	59.1% ± 3.0%	60.7% ± 2.7%	14.0% ± 1.5%
	Random Score	14.3% ± 2.0%	70.0% ± 2.9%	59.9% ± 3.1%	50.7% ± 2.5%	18.0% ± 1.7%
	Prod(VF, Contr.)	09.3% ± 1.7%	69.9% ± 2.8%	70.2% ± 2.8%	45.3% ± 2.4%	08.7% ± 1.2%
	Avg(VF, Contr.)	07.3% ± 1.5%	74.8% ± 2.6%	66.2% ± 2.8%	52.7% ± 2.5%	10.0% ± 1.3%
	VF num.	07.3% ± 1.5%	76.3% ± 2.6%	62.9% ± 2.9%	58.7% ± 2.6%	10.0% ± 1.3%
	VF desc.	09.7% ± 1.7%	69.4% ± 2.8%	68.6% ± 2.8%	45.3% ± 2.4%	11.3% ± 1.3%
	Contr. num.	09.0% ± 1.7%	77.7% ± 2.5%	64.1% ± 2.9%	56.7% ± 2.6%	08.7% ± 1.2%
	Contr. desc.	06.0% ± 1.4%	71.6% ± 2.7%	61.3% ± 2.9%	56.7% ± 2.6%	16.0% ± 1.6%
	Both num.	10.7% ± 1.8%	72.8% ± 2.7%	66.4% ± 2.9%	48.7% ± 2.5%	11.3% ± 1.3%
	Both desc.	08.7% ± 1.6%	68.2% ± 2.8%	67.2% ± 2.8%	46.0% ± 2.4%	14.0% ± 1.5%
	VF shown as Conf.	07.0% ± 1.5%	77.8% ± 2.5%	61.3% ± 2.9%	60.0% ± 2.7%	12.0% ± 1.4%
	Contr. shown as Conf.	08.3% ± 1.6%	74.5% ± 2.6%	69.1% ± 2.8%	50.0% ± 2.5%	06.7% ± 1.0%
	Prod shown as VF	07.7% ± 1.5%	65.7% ± 2.9%	67.5% ± 2.8%	44.4% ± 2.4%	16.0% ± 1.6%
	Prod shown as Contr.	08.0% ± 1.6%	70.3% ± 2.8%	68.5% ± 2.8%	44.7% ± 2.4%	13.3% ± 1.4%
VizWiz	Control	07.3% ± 1.5%	70.5% ± 2.7%	59.4% ± 3.0%	55.3% ± 2.6%	20.0% ± 1.7%
	VF num.	09.3% ± 1.7%	65.1% ± 2.9%	64.7% ± 2.9%	44.7% ± 2.4%	19.3% ± 1.7%
	VF desc.	10.3% ± 1.8%	67.3% ± 2.9%	69.1% ± 2.8%	42.0% ± 2.4%	13.3% ± 1.4%

Table 21: User-study results on A-OKVQA and VizWiz in the one-stage setting. Each row received 300 annotations (100 questions $\times$ 3 annotations per question).

Dataset	VLM	User study setting	#Ann.	After Stage 1		After Stage 2		After Stage 3	
				Unsure Rate	User Accuracy	Unsure Rate	User Accuracy	Unsure Rate	User Accuracy
A-OKVQA	LLaVA	VF num.	300	$65.3\% \pm 2.8\%$	$67.3\% \pm 4.6\%$	$19.0\% \pm 2.3\%$	$63.4\% \pm 3.1\%$	$5.0\% \pm 1.3\%$	$63.5\% \pm 2.9\%$
		Contr num.	300	$65.7\% \pm 2.7\%$	$58.3\% \pm 4.9\%$	$17.7\% \pm 2.2\%$	$60.7\% \pm 3.1\%$	$6.3\% \pm 1.4\%$	$61.2\% \pm 2.9\%$
		Both num.	300	$77.3\% \pm 2.4\%$	$58.8\% \pm 6.0\%$	$38.0\% \pm 2.8\%$	$54.8\% \pm 3.7\%$	$9.7\% \pm 1.7\%$	$64.2\% \pm 2.9\%$
		Avg(VF, Contr.)	300	$86.0\% \pm 2.0\%$	$54.8\% \pm 7.8\%$	$35.3\% \pm 2.8\%$	$63.9\% \pm 3.5\%$	$6.0\% \pm 1.4\%$	$63.5\% \pm 2.9\%$
		VF desc.	300	$67.3\% \pm 2.7\%$	$59.2\% \pm 5.0\%$	$19.7\% \pm 2.3\%$	$59.3\% \pm 3.2\%$	$3.7\% \pm 1.1\%$	$62.3\% \pm 2.9\%$
		Contr desc.	300	$63.0\% \pm 2.8\%$	$64.9\% \pm 4.6\%$	$17.0\% \pm 2.2\%$	$63.1\% \pm 3.1\%$	$8.7\% \pm 1.6\%$	$59.5\% \pm 3.0\%$
		Both desc.	290	$70.7\% \pm 2.7\%$	$47.1\% \pm 5.4\%$	$23.8\% \pm 2.5\%$	$57.9\% \pm 3.3\%$	$8.3\% \pm 1.6\%$	$64.7\% \pm 2.9\%$
VizWiz	Qwen2.5	VF num.	300	$57.7\% \pm 2.9\%$	$59.8\% \pm 4.4\%$	$5.7\% \pm 1.3\%$	$65.7\% \pm 2.8\%$	$1.7\% \pm 0.7\%$	$65.1\% \pm 2.8\%$
		VF desc.	300	$58.7\% \pm 2.8\%$	$66.1\% \pm 4.3\%$	$17.7\% \pm 2.2\%$	$65.2\% \pm 3.0\%$	$4.7\% \pm 1.2\%$	$67.5\% \pm 2.8\%$

Table 22: User-study results on A-OKVQA and VizWiz in the three-stages setting.

Study & Step	**Activity**	**Total Cost**
Predicted answer & Explanation Generation (LLaVA-v1.5-7B & Qwen2.5-VL-7B)	GPU inference (A100, 6h)	—
Predicted answer & Explanation Generation (GPT-4o)	OpenAI API call	≈$30
Qualities Computation	OpenAI API call	≈$120
Human Study (One-stage)	Prolific participant pay	$1,561
Supplementary Human Study (Three-stages)	Prolific participant pay	$1,116
		≈$2,827

Table 23: Monetary Cost Breakdown