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

Keyu He Tejas Srinivasan Brihi Joshi
Xiang Ren Jesse Thomason Swabha Swayamdipta
University of Southern California
{frankhe, tejass55, brihijos, xiangren, jessetho, swabhas}@usc.edu

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 and VizWiz 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 & 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 such *overreliance* on explanations in scenarios with such information asymmetry.

Preprint. Under review.

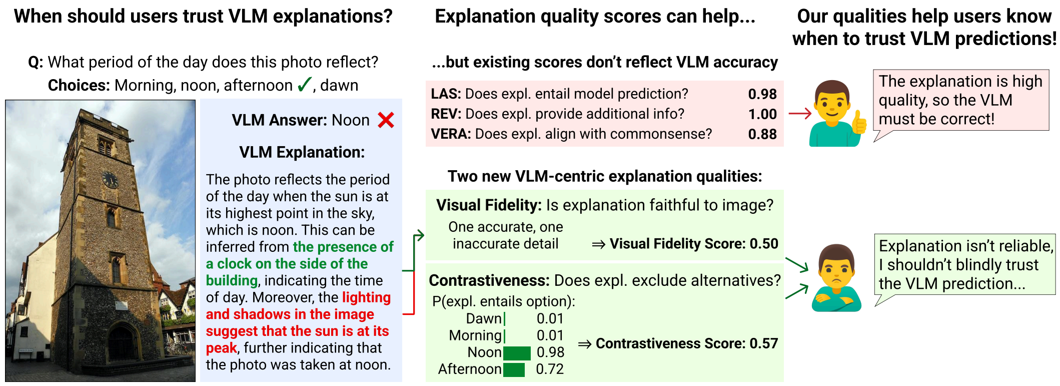


Figure 1: VLM explanations that sound plausible can mislead users. Contextualizing VLM explanations with quality scores may help users know when to rely on VLM outputs, but existing explanation qualities are not calibrated with VLM prediction accuracy. We propose evaluating two new qualities of VLM explanations: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

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*—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 two visual reasoning tasks, A-OKVQA (Schwenk et al., 2022) and VizWiz (Gurari et al., 2018), 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).

Algorithm 1 Evaluate Visual Fidelity

Input: Image I, question Q, model prediction a, explanation E
Output: Visual Fidelity score $S_{\mathrm{VF}}(E)$

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

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 $S_{\mathrm{Contr.}}(E)$

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

We further 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 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 in determining whether to trust the VLM's prediction or not.

To address the shortcomings of existing explanation qualities, we propose evaluating two new qualities of VLM-generated explanations: **Visual Fidelity** and **Contrastiveness** (Figure 1).

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). 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 $m_{\texttt{QGen}}$ is dependent 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 $a_i^V \in \{\text{yes, no}\}$. The Visual Fidelity score $S_{\mathrm{VF}}(E)$ is calculated as the fraction of questions for which the verifier $m_{\texttt{Verif}}$ answers "yes":

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

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

2.2 Measuring Contrastiveness

This quality 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

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

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 wouldnt 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 leakage from affecting our entailment model. Then, for each possible answer $a_j \in \mathcal{A}$, a paraphraser LLM $m_{\texttt{QA}\to\texttt{S}}$ 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 that the explanation identifies visual cues that not only support the VLM 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.

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

3 Are Visual Fidelity and Contrastiveness Indicative of VLM Correctness?

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

Visual Reasoning Tasks. We evaluate our explanation qualities on two visual reasoning tasks: A-OKVQA (Schwenk et al., 2022)—a multiple-choice VQA benchmark that requires reasoning over images using external knowledge and commonsense—and 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. We sample 500 questions from the validation set of each task. Appendix G.2 contains details about the tasks' pre-processing and evaluation. We do not evaluate Contrastiveness or combined quality scoring functions on VizWiz, which lacks a specified set of possible answers for each question.[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. Table 3 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).

[2]We tried turning this into a multiple-choice task by mining negatives from GPT-4o, however we found that the quality of negatives was not very good.

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

Table 1: Discriminability scores (higher is better) for different Quality Scoring Functions across three models and two visual reasoning tasks. We evaluate significance of Discriminability scores using an unpaired t-test; *** indicates significance at $p < 0.001$, ** indicates $p < 0.01$, * indicates $p < 0.05$.

	A-OKVQA			VizWiz			
	LLaVA	Qwen2.5	GPT-4o	LLaVA	Qwen2.5	GPT-4o	Average
Simulatability	0.106^{**}	0.247^{***}	0.098	0.088^{*}	0.158^{**}	-0.130^{*}	0.094
Informativeness	-0.018	0.000	0.069	0.093^{*}	-0.007	-0.113	0.017
Plausibility	0.043^{**}	0.022	0.028	0.004	-0.020	-0.002	0.031
Visual Fidelity	0.181^{***}	0.080^{***}	0.085^{***}	0.240^{***}	0.042^{*}	0.024	**0.115**
Contrastiveness	0.243^{***}	0.283^{***}	0.248^{***}	–	–	–	**0.258**
Avg (VF, Contr.)	0.212^{***}	0.181^{***}	0.166^{***}	–	–	–	**0.186**
Prod(VF, Contr.)	0.320^{***}	0.315^{***}	0.266^{***}	–	–	–	**0.300**
Min (VF, Contr.)	0.295^{***}	0.298^{***}	0.255^{***}	–	–	–	**0.283**

Table 2: Expected Calibration Error (ECE, lower is better) for explanation qualities.

	A-OKVQA			VizWiz			
	LLaVA	Qwen2.5	GPT-4o	LLaVA	Qwen2.5	GPT-4o	Average
Simulatability	0.288	0.219	0.302	0.305	0.292	0.494	0.316
Informativeness	0.332	0.152	0.228	0.429	0.168	0.446	0.293
Plausibility	0.162	0.270	0.317	0.110	0.310	0.325	0.248
Visual Fidelity	0.207	0.137	0.099	0.271	0.136	0.160	**0.168**
Contrastiveness	0.176	0.136	0.209	–	–	–	**0.174**
Avg(VF, Contr.)	0.109	0.053	0.090	–	–	–	**0.084**
VF * Contr.	0.164	0.154	0.233	–	–	–	**0.183**
Min(VF, Contr.)	0.147	0.144	0.227	–	–	–	**0.173**

Baseline Explanation Qualities. 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-starting the question and prediction. **Commonsense Plausibility** (Liu et al., 2023b) evaluates whether the explanation is in accordance with commonsense knowledge of everyday situations.

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, Disc(S_φ) evaluates a quality scoring function S_φ by calculating the difference between the mean quality score assigned to instances with accurate predictions (Acc$(a_i) = 1$) and inaccurate predictions (Acc$(a_i) = 0$):

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

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.

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. Tables 1 and 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 Visual Fidelity and Contrastiveness.

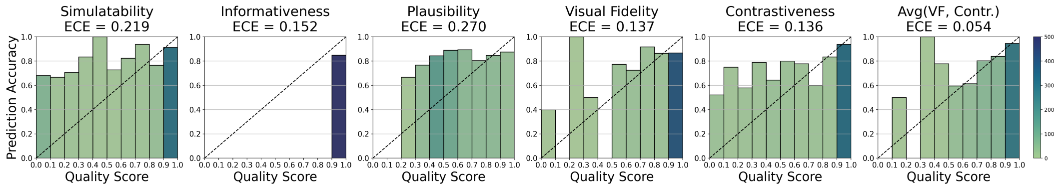


Figure 2: Calibration curves for various quality scoring functions when evaluating explanations generated by Qwen2.5-VL-7B on the A-OKVQA dataset.

We observe that on A-OKVQA, across all three VLMs, our proposed quality scoring functions achieve higher Discriminability scores and lower Expected Calibration Errors than existing scoring functions. We also see that all Discriminability scores are significant at $p < 0.001$, indicating that both our quality scoring functions are capable of distinguishing accurate and inaccurate VLM predictions. Further, **combining the Visual Fidelity and Contrastiveness quality scores results in even higher Discriminability scores and lower calibration errors**, highlighting complementary utility. On VizWiz, we similarly find that Visual Fidelity is better calibrated than baseline quality scoring functions, and is better at distinguishing correct and incorrect predictions made by LLaVA.

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. Our proposed quality scores achieve a lower calibration error while assigning a wider spread of scores across all evaluation samples.

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). To what extent can providing VF and $\mathtt{Contr.}$ scores alongside VLM explanations help users more accurately decide when to believe VLM predictions?

4.1 User Study Setup

We simulate a real-world setting of a user relying on a VLM assistant when the user cannot directly view the VLM's visual input. Human participants assess 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.

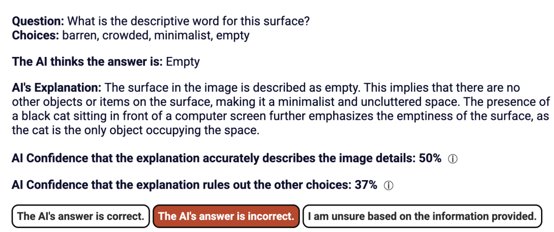


Figure 3: Our study interface where users are shown the visual reasoning question, VLM prediction and explanation, and optionally one or more quality scores (using simplified language descriptions of both qualities). Here, the user believes that the VLM prediction is incorrect.

Study Details. We conduct a between-conditions user study, where users makes predictions about VLM correctness based on either none or a subset of the explanation quality scores. We conduct user studies on a subset of 100 questions each from the A-OKVQA and VizWiz tasks, with each subset containing 50 correct and 50 incorrect predictions by LLaVA-v1.5-7B. We select these subsets to approximately match the ECE of VF and $\mathtt{Contr.}$ scores (if available) with the full evaluation set to maintain a similar distribution of quality scores in our subset.

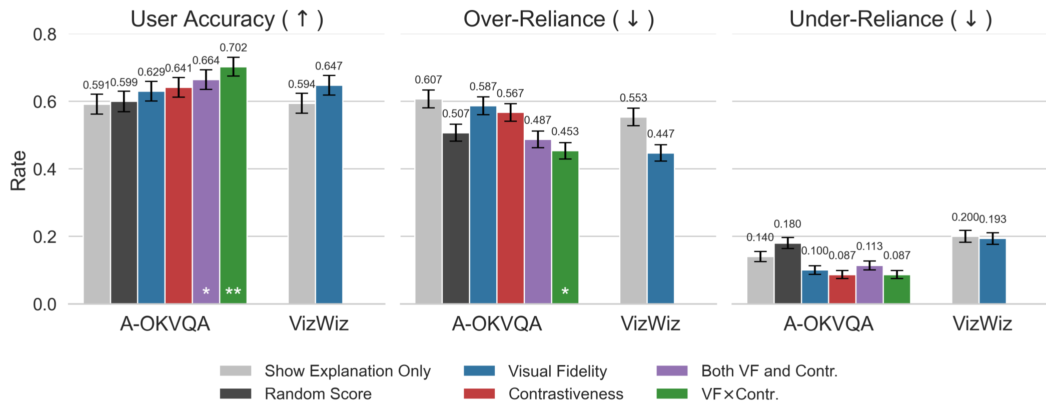


Figure 4: Effect of showing users different quality scores on User Accuracy, Over-Reliance and Under-Reliance. Error bars represent the standard deviation of the data. Asterisks denote improvements over the explanation-only baseline using a bootstrap significance test (*: $p < 0.05$, **: $p < 0.01$).

In a given setting (either none or some of the quality scores are presented to users), each question is annotated by three participants, yielding 300 annotations. Each participant completes 10 questions, also balanced to contain five instances where the model is correct and five where it is incorrect; participants are not informed of this balancing. Participants are randomly assigned to one of the study conditions (control or one of the treatment variants) and do not participate in more than one condition. 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.

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 & 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 8 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-only condition and the Random Score baseline, all treatment conditions lead to higher user accuracy. On A-OKVQA, the most pronounced improvements are observed when showing the product of Visual Fidelity and Contrastiveness, which results in an 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 not only better identify correct model predictions, but also 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**.

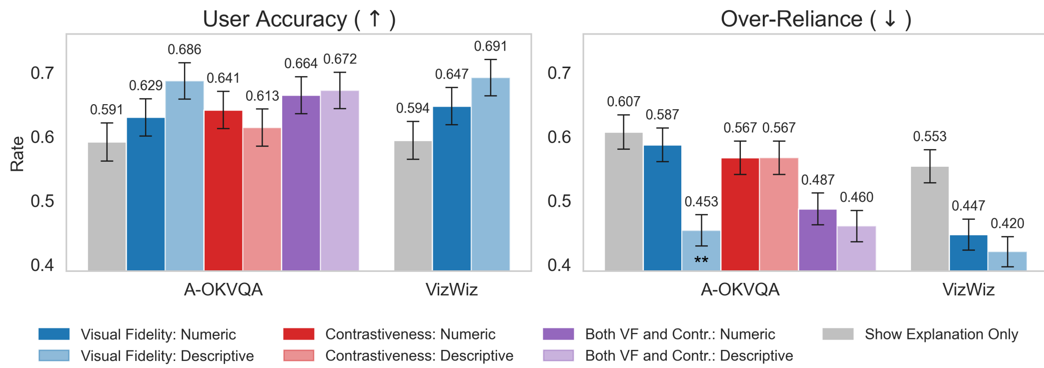


Figure 5: Effect of showing numeric and descriptive qualities on User Accuracy and Over-Reliance. Error bars show standard deviation of the data. Asterisks denote improvements between descriptive setting against the corresponding numeric setting with a bootstrap significance test (**: $p < 0.01$).

4.3 RQ2: Do descriptive expressions of quality scores affect user decision accuracy?

Our proposed quality scores are interpretable. For Visual Fidelity, we can observe which visual details 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 our proposed qualities as text descriptions.

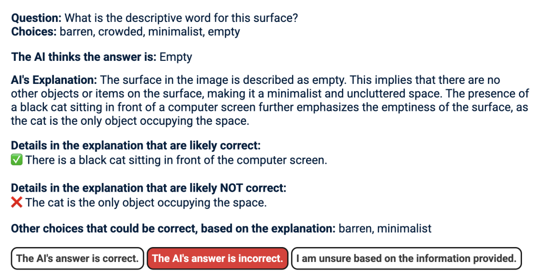


Figure 6: Here, the study interface shows the participant descriptive text versions of the Visual Fidelity and Contrastiveness qualities.

Results. As shown in Figure 5, 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 $\mathrm{Contr.}$, 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 process more transparent and interpretable**.

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 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 7, 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**.

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

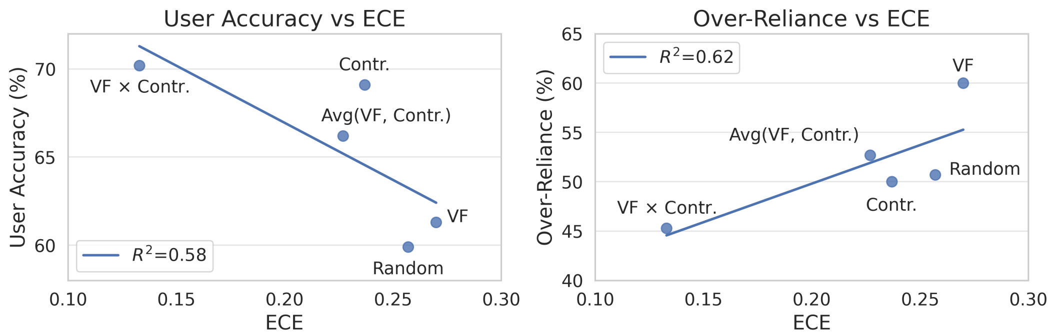


Figure 7: Relationship between ECE of different quality scores and their downstream utility to users.

5 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.

Explanations for Decision Support. Most research on utility of natural language explanations for decision support has 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.

6 Conclusions and Future Work

As VLMs are increasingly deployed in settings where users lack access to the model's visual inputs, users must be able to discern when to rely on model predictions using secondary cues like explanations. However, 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 also demonstrate real-world utility of these scores, as users presented with these quality scores alongside VLM predictions and explanations are able to improve their task accuracy and reduce over-reliance on the VLM.

Limitations. A-OKVQA multiple-choice questions may not be representative of real-world visual queries. Further, the design of our Contrastiveness quality limits its application to tasks with a closed-set of possible answers, which is often not available in realistic settings. Finally, we only evaluate on English-language datasets, and conduct user studies with only fluent English speakers.

Future Work. An important direction is to develop adaptive human-AI reliance strategies that learn *when* to present explanations, *when* to suppress them, and *when* to show only quality scores, depending on the context, task difficulty, or user uncertainty. 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 in order to better understand how users adapt to model signals and how trust dynamics evolve over repeated human-AI interactions.

Acknowledgments

This research is supported in part by the National Science Foundation (NSF) under grant IIS2403437, the Simons Foundation, and the Allen Institute for AI. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of NSF. B. Joshi was supported by the Apple Scholars in AI/ML PhD Fellowship. This work was partially done when S. Swayamdipta was a visitor at the Simons Institute for the Theory of Computing.

References

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

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

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 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*, pp. 1–16, 2021.

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

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

Hanjie Chen, Faeze Brahman, Xiang Ren, Yangfeng Ji, Yejin Choi, and Swabha Swayamdipta. 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)*, pp. 2007–2030, 2023.

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. Roscoe: A suite of metrics for scoring step-by-step reasoning, 2023. URL https://arxiv.org/abs/2212.07919.

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

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

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

Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. 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*, pp. 4351–4367, 2020.

Mina Huh, Fangyuan Xu, Yi-Hao Peng, Chongyan Chen, Hansika Murugu, Danna Gurari, Eunsol Choi, and Amy Pavel. Long-form answers to visual questions from blind and low vision people. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=z7FvXbyyrM.

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

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

Brihi Joshi, Ziyi Liu, Sahana Ramnath, Aaron Chan, Zhewei Tong, Shaoliang Nie, Qifan Wang, Yejin Choi, and Xiang Ren. 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*, 2023.

Junhyeok Kim, Jaewoo Park, Junhee Park, Sangeyl Lee, Jiwan Chung, Jisung Kim, Ji Hoon Joung, and Youngjae Yu. Guidedog: A real-world egocentric multimodal dataset for blind and low-vision accessibility-aware guidance, 2025. URL https://arxiv.org/abs/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. 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)*, pp. 881–905, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.50. URL https://aclanthology.org/2024.acl-long.50/.

Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback, 2023. URL https://arxiv.org/abs/2302.12192.

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

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

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

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. ScoutBot: A dialogue system for collaborative navigation. In Fei Liu and Thamar Solorio (eds.), *Proceedings of ACL 2018, System Demonstrations*, pp. 93–98, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-4016. URL https://aclanthology.org/P18-4016/.

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

Pranava Madhyastha, Josiah Wang, and Lucia Specia. Vifidel: Evaluating the visual fidelity of image descriptions, 2019. URL https://arxiv.org/abs/1907.09340.

Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. Receval: Evaluating reasoning chains via correctness and informativeness, 2023. URL https://arxiv.org/abs/2304.10703.

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

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

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

Chenglei Si, Navita Goyal, Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé Iii, and Jordan Boyd-Graber. 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)*, pp. 1459–1474, 2024b.

Judith Sieker, Simeon Junker, Ronja Utescher, Nazia Attari, Heiko Wersing, Hendrik Buschmeier, and Sina Zarrieß. 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*, pp. 19459–19475, 2024.

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

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

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

Xinru Wang and Ming Yin. 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*, pp. 318–328, 2021.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. Reframing human-AI collaboration for generating free-text explanations. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 632–658, Seattle, United States, July 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.47. URL https://aclanthology.org/2022.naacl-main.47/.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. Measuring association between labels and free-text rationales, 2022b. URL https://arxiv.org/abs/2010.12762.

A Code and Data Availability

Our full codebase is available at `https://anonymous.4open.science/r/quality_scores_for_VLM_explanations-364C/`.

B 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."

By separating prediction from explanation, we ensure that the extracted rationale truly reflects the model's own line of thought under the selected answer. Table 3 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.

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.

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}.`

C Existing Text-only Qualities Implementations

C.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.

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

Config	Assignment
Model	`gpt-4o-mini-2024-07-18`
Max Tokens Temperature	1024 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 5: Model configuration and prompt used to evaluate informativeness of a rationale.

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

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

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

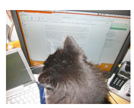
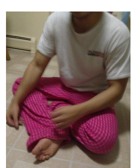
Config	Assignment
Model	`gpt-4o-2024-08-06`
Max Tokens	1024
Temperature	0.1

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

E Example of Model Explanations with Different Qualities

From Table 8, we show our evaluation framework on four representative VQA instances. For each example, we provide the original question (and answer choices where applicable), image, 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.

Table 8: 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.

Dataset	Question	Answer Choices	Image	Ground Truth	Predicted Answer	LLaVA-v1.5-7B Model 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?	—		pink	pink	The man is wearing pink pajamas, which is an unusual choice for a casual setting.	0.025	1.000	0.397	1.000	—
VizWiz	What is in this bottle?	—		bacon bits	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	—

F Internal Sanity Check on Visual Fidelity and Contrastiveness

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 9. After collecting the human judgments, we compared them to our model’s binary VF and Contrastiveness labels, computing Cohen’s κ for each quality. We observed:

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

Table 9: Instructions Provided to the Expert Annotator

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 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 before proceeding!

Table 10: Prolific annotator filtering

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

indicating moderate agreement with human annotation.

Note that both VF and Contrastiveness scores were cast to binary labels using a threshold of 0.5 (scores ≥ 0.5 mapped to 1, otherwise 0).

G User Studies

G.1 Annotator filtering

To ensure high-quality annotations, we recruited annotators on Prolific using the criteria listed in Table 10. 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 experimental conditions, each participant was confined to a single "setting" (i.e., one quality-type configuration).

G.2 Creating the subset splits

For our qualities evaluations from Section 3, we selected two 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 whose majority human annotation was not "unanswerable" and which contained no NSFW content, as filtered by GPT-4o.

To create the subset used for human studies, we create a subset of 100 questions each from the A-OKVQA and VizWiz 500-instance datasets above. These subsets are chosen by sampling 50 times from the validation set; each sampled subset contains randomly sampled 50 correct and 50 incorrect instances from the validation set. On the A-OKVQA dataset, we picked the subset for which the average Expected Calibration Error across Visual Fidelity and Contrastiveness is the lowest in 50 samples; on the VizWiz dataset, we picked the subset for which the Expected Calibration Error for Visual Fidelity is the lowest in our 50 samples, as Contrastiveness is not available on the VizWiz dataset. From Table 11 this maintains a similar ECE distribution of these quality measures to the full set from Section 3.

Table 11: Accuracy and Expected Calibration Error (ECE) for different qualities. . "Subset selected" rows correspond to the 100-question subsets (50 correct / 50 incorrect) selected and the other two rows represent LLaVA-v1.5-7B on the A-OKVQA and VizWiz 500-instance datasets.

Dataset	Accuracy	VF	Contr.	VF×Contr.	Min(VF, Contr.)	Avg(VF, Contr.)	Support	Informative	Plausibility
AOKVQA	0.696	0.207	0.176	0.164	0.147	0.109	0.288	0.332	0.162
AOKVQA (subset selected)	0.500	0.270	0.237	0.133	0.130	0.227	0.372	0.490	0.075
VizWiz	0.557	0.271	—	—	—	—	0.305	0.429	0.110
VizWiz (subset selected)	0.500	0.236	—	—	—	—	0.379	0.390	0.071

G.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.

G.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

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

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.

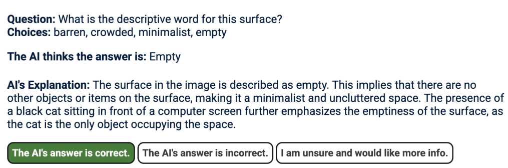
G.3.2 Bonus Payments

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.

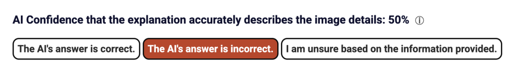
G.4 User Studies Settings & Examples

Table 12: Summary of the 14 one-step human-study settings.

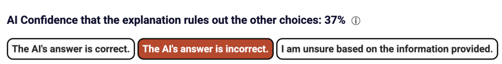
Setting	Description
Show Explanation Only	No quality scores shown
Random Score	Random scoring baseline (uniform random distribution from $[0, 1]$), shown as a simple confidence score "AI Confidence that the explanation is correct".
VF×Contr	Product of VF and Contrast scores, shown as a simple confidence 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 Numeric	Display both VF num and Contr. num messages
Both Descriptive	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"



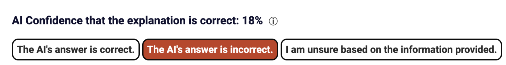
(a) Control (no quality)



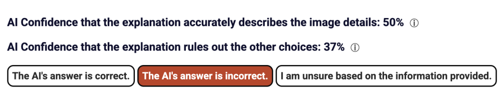
(b) VF – Numeric quality



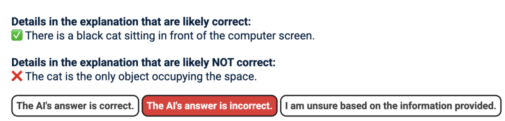
(c) Contr. – Numeric quality



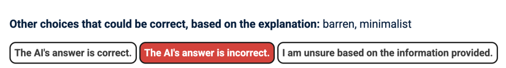
(d) VF×Contr. (shown as a simple confidence score) – Numeric quality



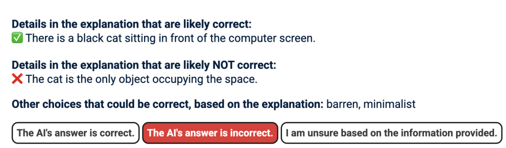
(e) Both VF and Contr. – Numeric quality



(f) VF – Descriptive quality



(g) Contr. – Descriptive quality



(h) Both VF and Contr. – Descriptive quality

Figure 8: 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.

In the main user study, we present 14 settings in total. Table 12 summarizes all 14 different settings and Figure 8 shows examples of all 8 different categories of messages presented to users. Table 13 contains the main user-study results of user behavior patterns.

Table 13: 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	Setting	Unsure Rate	Not Unsure Accept Rate	Not Unsure Accuracy	False Accept Rate	False Reject Rate
A-OKVQA	Control	8.7% ± 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%
	VF$\times$Contr.	9.3% ± 1.7%	69.9% ± 2.8%	70.2% ± 2.8%	45.3% ± 2.4%	8.7% ± 1.2%
	AVG(VF, Contr.)	7.3% ± 1.5%	74.8% ± 2.6%	66.2% ± 2.8%	52.7% ± 2.5%	10.0% ± 1.3%
	VF num	7.3% ± 1.5%	76.3% ± 2.6%	62.9% ± 2.9%	58.7% ± 2.6%	10.0% ± 1.3%
	VF desc	9.7% ± 1.7%	69.4% ± 2.8%	68.6% ± 2.8%	45.3% ± 2.4%	11.3% ± 1.3%
	Contr. num	9.0% ± 1.7%	77.7% ± 2.5%	64.1% ± 2.9%	56.7% ± 2.6%	8.7% ± 1.2%
	Contr. desc	6.0% ± 1.4%	71.6% ± 2.7%	61.3% ± 2.9%	56.7% ± 2.6%	16.0% ± 1.6%
	Both Numeric	10.7% ± 1.8%	72.8% ± 2.7%	66.4% ± 2.9%	48.7% ± 2.5%	11.3% ± 1.3%
	Both Descriptive	8.7% ± 1.6%	68.2% ± 2.8%	67.2% ± 2.8%	46.0% ± 2.4%	14.0% ± 1.5%
	VF shown as Prod	7.0% ± 1.5%	77.8% ± 2.5%	61.3% ± 2.9%	60.0% ± 2.7%	12.0% ± 1.4%
	Contr. shown as Prod	8.3% ± 1.6%	74.5% ± 2.6%	69.1% ± 2.8%	50.0% ± 2.5%	6.7% ± 1.0%
	Prod shown as VF	7.7% ± 1.5%	65.7% ± 2.9%	67.5% ± 2.8%	44.4% ± 2.4%	16.0% ± 1.6%
	Prod shown as Contr.	8.0% ± 1.6%	70.3% ± 2.8%	68.5% ± 2.8%	44.7% ± 2.4%	13.3% ± 1.4%
VizWiz	Control	7.3% ± 1.5%	70.5% ± 2.7%	59.4% ± 3.0%	55.3% ± 2.6%	20.0% ± 1.7%
	VF num	9.3% ± 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%

G.5 Supplementary Human Studies

In addition to the study described in §4.2 and Appendix G.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:

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 Appendix G.3.2).

Table 14 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.

Table 14: User-study results on A-OKVQA and VizWiz in the three stages setting.

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-v1.5-7B	VF (numeric)	300	65.3% ± 2.8%	67.3% ± 4.6%	19.0% ± 2.3%	63.4% ± 3.1%	5.0% ± 1.3%	63.5% ± 2.9%
		Contr (numeric)	300	65.7% ± 2.7%	58.3% ± 4.9%	17.7% ± 2.2%	60.7% ± 3.1%	6.3% ± 1.4%	61.2% ± 2.9%
		Both VF and Contr (numeric)	300	77.3% ± 2.4%	58.8% ± 6.0%	38.0% ± 2.8%	54.8% ± 3.7%	9.7% ± 1.7%	64.2% ± 2.9%
		Avg(VF, Contr)	300	86.0% ± 2.0%	54.8% ± 7.8%	35.3% ± 2.8%	63.9% ± 3.5%	6.0% ± 1.4%	63.5% ± 2.9%
		VF (descriptive)	300	67.3% ± 2.7%	59.2% ± 5.0%	19.7% ± 2.3%	59.3% ± 3.2%	3.7% ± 1.1%	62.3% ± 2.9%
		Contr (descriptive)	300	63.0% ± 2.8%	64.9% ± 4.6%	17.0% ± 2.2%	63.1% ± 3.1%	8.7% ± 1.6%	59.5% ± 3.0%
		Both VF and Contr (descriptive)	290	70.7% ± 2.7%	47.1% ± 5.4%	23.8% ± 2.5%	57.9% ± 3.3%	8.3% ± 1.6%	64.7% ± 2.9%
VizWiz	Qwen2.5-VL-7B	VF (numeric)	300	57.7% ± 2.9%	59.8% ± 4.4%	5.7% ± 1.3%	65.7% ± 2.8%	1.7% ± 0.7%	65.1% ± 2.8%
		VF (descriptive)	300	58.7% ± 2.8%	66.1% ± 4.3%	17.7% ± 2.2%	65.2% ± 3.0%	4.7% ± 1.2%	67.5% ± 2.8%

H Computational Resources Spent & Total Cost

The complete breakdown of all monetary expenditures across the study is given in Table 15. 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 qualities. Human evaluation costs cover two Prolific studies—a one-stage study as shown in Appendix G.4 and a three-stage follow-up from Appendix G.5 (including up to $1.00 in per-participation bonuses; see Appendix G.3.2).

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

Table 15: Monetary Cost Breakdown

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