Hypothesize and Verify: Natural-Language Explanations of Vision Model Errors

Anonymous authors
Paper under double-blind review

Abstract

LLM- and agent-based assistants now bring non-experts into direct ML work, where they probe model failures by asking the assistant in plain language. When such a classifier misclassifies an image, the non-expert needs a faithful account of *why*. Two obstacles stand in the way. No benchmark scores free-form natural-language explanations of such errors, and existing retrieval-based methods can only return sentences from a fixed error corpus. We close both. NEMO is a task and benchmark of 1,200 misclassified images across ImageNet-R, ObjectNet, and ImageNet-D, each varying along a controlled factor (artistic style, viewpoint, low-level attributes), scored by an LLM-as-a-Judge (LLM Match) protocol that asks whether the explanation describes that factor. SciTX is a generation-based method that emulates the scientific method: retrieve observations, hypothesize candidate causes, verify each with a counterfactual intervention, and retain the hypothesis whose intervention shifts the model's prediction farthest toward the ground-truth class. The shift is captured by our Counterfactual Explanation Impact (CEI), which serves as both SciTX's selection signal and a complementary evaluation metric. On NEMO, SciTX outperforms retrieval-based and MLLM-augmented baselines on both LLM Match and CEI, and 30 AI practitioners rank it first across all five helpfulness dimensions, including factuality, specificity, and actionability.

1 Introduction

With the advent of the LLM and Agent era, even non-experts without sufficient machine learning (ML) expertise—such as product managers, domain stakeholders, and newcomers to the ML field—can now easily use ML workflows. For example, users who are new to machine learning can build ML models and debug them using only natural language interfaces within agent-based coding environments. In this changing landscape, providing faithful natural-language explanations of why a model failed is increasingly important to users. Unreliable natural-language explanations can lead to incorrect decisions during subsequent model improvement processes or in stakeholder decision-making.

Although various methods have been developed for debugging models, most do not provide users with a direct natural-language interface. For example, saliency-based methods such as CAM (Selvaraju et al., 2017) and LIME (Ribeiro et al., 2016) visually highlight the image regions that influence a model's prediction, but interpreting and reasoning about the actual cause of failure in natural language remains the user's responsibility. Additionally, Concept Bottleneck Models (Koh et al., 2020) restrict explanations to a predefined, fixed concept vocabulary, which makes it difficult to flexibly describe new failure modes. Meanwhile, there are also approaches that retrieve error explanations from a natural language-based error corpus (Jain et al., 2023; Eyuboglu et al., 2022), but these are limited by the scope of the pre-constructed, fixed error corpus. As a result, if a model failure falls outside the categories covered by the existing corpus, it cannot provide an appropriate explanation. Finally, it is possible to generate explanations by directly querying a multi-modal LLM (MLLM), but this approach largely relies on the MLLM's parametric knowledge, which poses the risk of generating plausible-sounding, yet unfaithful, natural language explanations that are unrelated to the model's actual failure mechanism.

From the perspective of explanation evaluation, prior studies have mainly focused on assessing explanations retrieved from predefined sentence pools or error corpora. For example, LBEE (Csurka et al., 2024) proposes metrics for error-related sentence retrieval, but the explanations are chosen from a predefined sentence set and do not evaluate free-form natural-language explanations. Moreover, these approaches emphasize how well the retrieved explanations identify predefined error categories, rather than assessing whether the explanations are truly faithful to the model's actual failure mechanism. Therefore, existing approaches have limitations in describing novel failure modes outside the corpus or in evaluating the faithfulness of open-ended, natural-language explanations.

In this work, we propose **Sci**entific Method-Inspired **T**extual E**X**planation (SciTX), a generation-based method that produces verified natural-language explanations of vision-classifier errors. For the evaluation of free-form natural language explanation, we construct the evaluation protocol, **N**atural-language **E**xplanations for **MO**del Errors (NEMO), for this task, paired with LLM-as-a-Judge protocol (LLM Match) scores and the Counterfactual Explanation Impact (CEI) metric.

The key challenge for generation-based explanation methods is ensuring that the generated explanations are not just plausible, but actually testable. We interpret an explanation as a hypothesis about the cause of the model's failure. A faithful explanation should cause the model's prediction to shift toward the ground-truth class when an intervention targeting the hypothesized cause is performed, while an unfaithful explanation should lead to no significant change in prediction. SciTX implements this principle from the perspective of scientific methodology, proceeding in four stages. **(1) Observation.** Given a misclassified image, SciTX extracts contrastive observations by comparing correctly classified samples from the ground-truth and predicted classes. **(2) Hypothesis formation.** Based on these observations, SciTX generates candidate causes for the model's failure as natural-language hypotheses. **(3) Experimentation.** For each candidate hypothesis, SciTX performs a counterfactual intervention. **(4) Conclusion.** SciTX selects the explanation whose intervention shifts the model's prediction most toward the ground-truth class, ensuring that the explanation is grounded in the model's actual failure mechanism.

Given target classifiers, NEMO pairs 1,200 misclassified images with a controlled failure factor per dataset. We draw 400 images from each of ImageNet-R, ObjectNet, and ImageNet-D, which vary in artistic style, viewpoint, and low-level attributes, respectively. Because free-form natural-language explanations admit no single reference answer, reference-based text metrics such as BLEU or ROUGE are inapplicable. We therefore adopt the LLM-as-a-Judge paradigm (Dunlap et al., 2024; Yu et al., 2024; Liu et al., 2023b), which has become the standard for evaluating free-form generation through its strong correlation with human judgment. Our protocol, LLM Match, scores each explanation by whether it names the controlled factor, yielding a reproducible automated measure. We complement LLM Match with CEI, which measures the change in ground-truth probability after applying the intervention derived from an explanation, providing a causal-effect signal that protocol-based scoring alone cannot capture.

On NEMO, with the Qwen-VL family as the explanation-generating MLLM and three target classifiers (ViT, CLIP, SigLIP), SciTX outperforms every retrieval-based and MLLM-augmented baseline on both LLM Match and CEI. We compute LLM Match with three judges drawn from non-Qwen families, namely GPT, Claude, and Gemini, so that the explanation generator and the evaluator never share a model family. The lead persists when baselines adopt the same multi-candidate generation and CEI-argmax selection as SciTX, indicating that the gains stem from explanation content rather than from matched selection. The method ranking is also stable across two distinct image editors used for counterfactual interventions, and the absolute scores scale with backbone capacity from Qwen2.5-VL-7B to Qwen3-VL-30B. A user study with AI practitioners further ranks SciTX first on every one of the five helpfulness dimensions. In summary, we make the following contributions:

- We introduce NEMO, an evaluation protocol for free-form natural-language explanations of vision-model errors and an automated LLM Match scoring procedure.
- We propose SciTX, a method that selects each explanation by counterfactual validation following the scientific method.

- We introduce CEI, a counterfactual-impact metric that serves both as an evaluation signal and as SciTX's hypothesis-selection criterion.
- On NEMO, SciTX outperforms all retrieval- and MLLM-based baselines across three target classifiers and multiple MLLM backbones, and 30 AI practitioners rank it first on every helpfulness dimension.

2 Related Work

Automatic Model Debugging. Model debugging (Ribeiro et al., 2016; Koh et al., 2020; Nguyen et al., 2024) is a key step in designing and improving machine learning models. Diagnosing model errors guides model improvement and informs deployment decisions (Kirichenko et al., 2023; Darcet et al., 2023; Hyeon-Woo et al., 2023; Geirhos et al., 2019). Class Activation Map (CAM) (Selvaraju et al., 2017; Zhou et al., 2016; Fernandez, 2020) highlights image regions that drive a prediction, while Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016) and SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017; Lundberg et al., 2020) attribute a decision to specific input features. Concept Bottleneck Models (CBMs) (Koh et al., 2020; Yang et al., 2023) instead design the model to predict human-understandable concepts at intermediate layers. These methods do not describe the cause of an error in free-form natural language, which is what SciTX targets.

Language-based Model Debugging. Natural language is among the most accessible forms of explanation for non-experts (Doshi-Velez & Kim, 2017), and LLM-based interfaces have made it the primary medium through which non-experts interact with machine learning models. This motivates a growing line of work that explains classifier failures in natural language, progressing from retrieving sentences out of a fixed corpus toward generating them with language models, and from group-level summaries toward per-sample explanations.

Prior work (Eyuboglu et al., 2022; Jain et al., 2023; Rezaei et al., 2024; Dunlap et al., 2024; Csurka et al., 2024; Shaham et al., 2024) produces natural-language explanations of classifier failures. Retrieval-based methods (Eyuboglu et al., 2022; Jain et al., 2023; Rezaei et al., 2024; Csurka et al., 2024) score multimodal similarities between misclassified samples and a fixed corpus of error sentences and return the closest match. They require manual curation of that corpus and cannot describe errors absent from it.

Recently, various methods have attempted to use language models to verbally explain the failure modes of classifiers at the slice or dataset level. For example, LADDER (Ghosh et al., 2025) uses a text-only LLM to generate hypothesis sentences about classifier biases, and B2T (Kim et al., 2024) extracts bias-related keywords from caption corpora. However, these approaches primarily focus on group-level failure analysis or provide results in forms such as attributes or keywords, rather than complete natural-language explanations. In contrast, SciTX focuses on the still under-explored sample-level generation setting. Specifically, we use an MLLM conditioned on the misclassified image to generate free-form natural language explanations, rather than retrieved templates or keywords. Each candidate's explanation is then validated through a counterfactual intervention, thereby evaluating its faithfulness.

Large Language Models. LLMs (Dubey et al., 2024; Achiam et al., 2023; Bai et al., 2023; Abdin et al., 2024; Guo et al., 2025) and multi-modal LLMs (MLLMs) (Liu et al., 2023a; Li et al., 2022; 2023; Touvron et al., 2023; Bai et al., 2025a) have advanced rapidly, with strong reasoning capabilities across diverse tasks (Kojima et al., 2022; Wei et al., 2022). Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Guu et al., 2020; Gao et al., 2023) grounds their outputs in external knowledge. MLLMs combine a vision encoder (Radford et al., 2021; Zhai et al., 2023) with an LLM to understand vision and text jointly.

Counterfactual Reasoning. Counterfactual reasoning (Pearl, 2009) asks how a model would behave when one factor of the input is changed, and has become a standard lens for diagnosing what drives a model's decision. Prior counterfactual XAI work mostly produces counterfactual examples in pixel or feature space and measures how the model's prediction shifts (Goyal et al., 2019; Yan & Wang, 2023; Jeanneret et al., 2022; Farid et al., 2023; Chiquier et al., 2025), delivering the explanation as a visual diff rather than as a language-level statement of the responsible factor. We adapt this principle to evaluate natural-language explanations. CEI derives an intervention from an explanation through an MLLM and reports the change

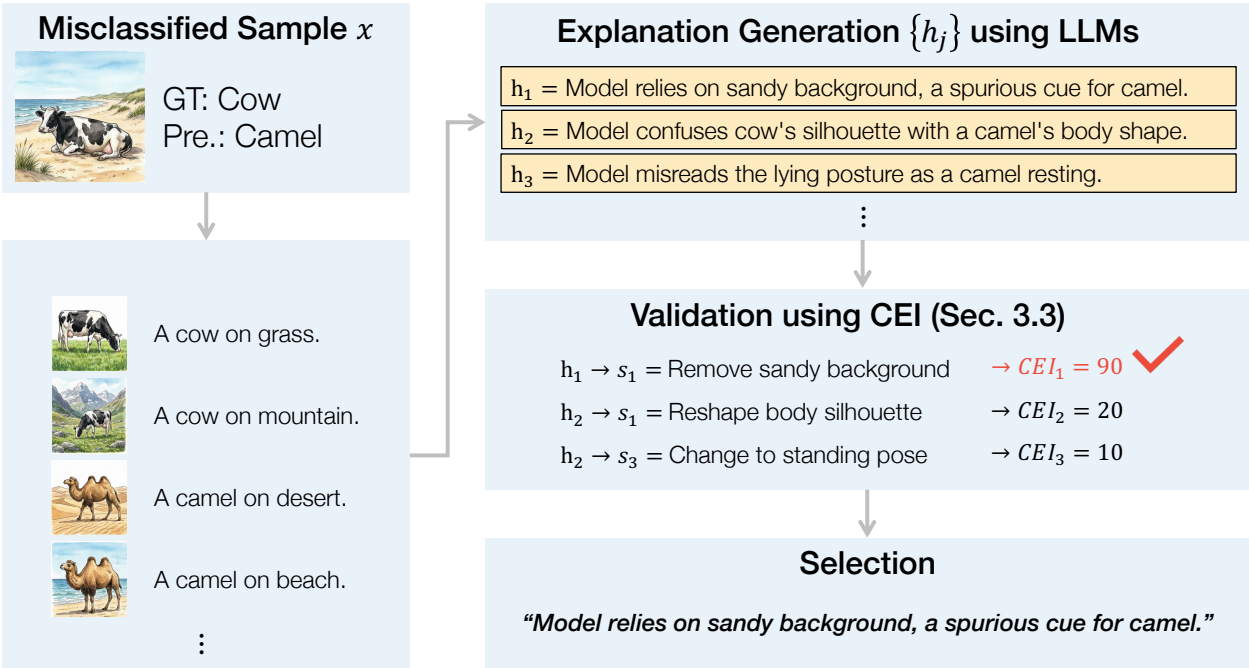


Figure 1: **SciTX.** **(1)** Observation Retrieval. Reference samples from the ground-truth and predicted classes form contrastive context. **(2)** Explanation Generation. An LLM produces candidate hypotheses $\{h_j\}$ from the query and the observations. **(3)** Explanation Validation. Each hypothesis is scored by the CEI of its counterfactual intervention. **(4)** Explanation Selection. The highest-CEI hypothesis is kept as the final explanation. The example is intentionally simple for visual clarity.

in the ground-truth probability as a single scalar, providing a quantitative, explanation-conditioned signal. The intervention is applied in image space for standard vision classifiers and in prompt space for VLM-based zero-shot classifiers such as CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023), where the zero-shot template is rewritten to incorporate the cue named by the explanation. Recent instruction-following image editors such as FLUX.1-Kontext (Labs et al., 2025), Qwen Image Edit (Wu et al., 2025), and Nano Banana apply targeted edits with high fidelity, making image-space counterfactuals feasible at scale.

3 Method

3.1 Task Formulation

Let $f : \mathcal{X} \rightarrow \mathcal{Y}$ be a discriminative classifier mapping an image $x \in \mathcal{X}$ to a class label, and let $y \in \mathcal{Y}$ denote its ground-truth label. We define the set of misclassified samples as

$$\mathcal{D}_{\text{err}} = \{(x, y) \mid f(x) \neq y\}. \tag{1}$$

Given a sample $(x, y) \in \mathcal{D}_{\text{err}}$ with predicted label $\hat{y} = f(x)$, the **N**atural-language **E**xplanations for **MO**del Errors (NEMO) task is to produce a natural-language explanation

$$e = g(x, y, \hat{y}, f) \in \mathcal{E}, \tag{2}$$

where g is an explanation generator and $\mathcal{E}$ denotes the space of natural-language strings.

We want e to identify the *causal factor* of the misclassification, not just describe the visual content of x. We measure this with CEI (Sec. 3.3) and LLM Match (Sec. 3.4).

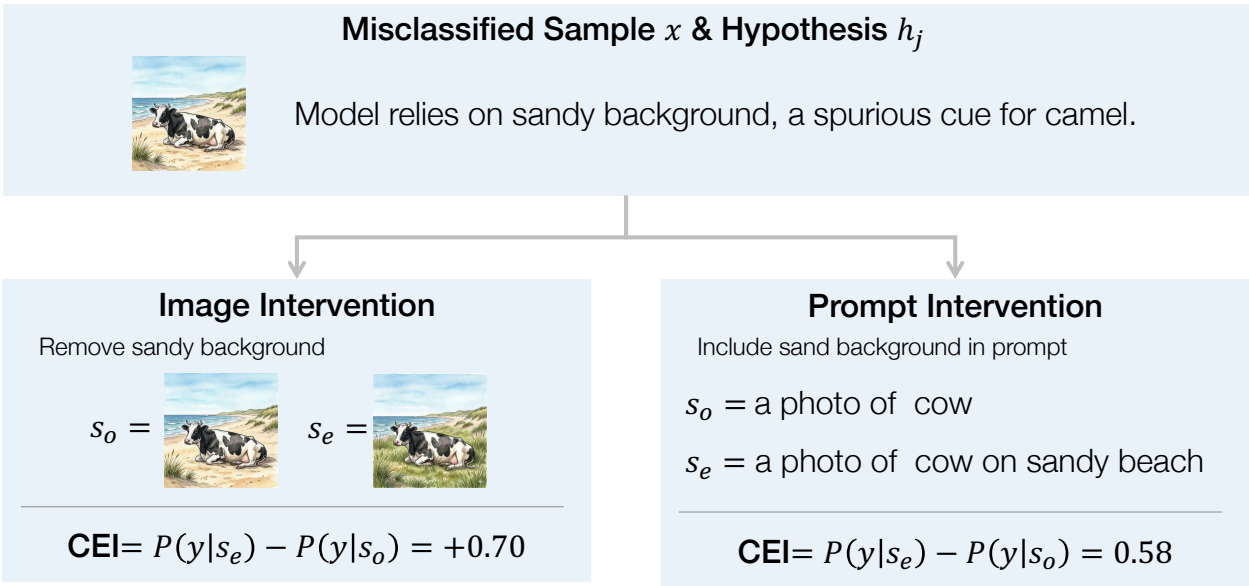


Figure 2: **Counterfactual Explanation Impact (CEI).** CEI measures the change in ground-truth probability after neutralizing the cause described by an explanation, $P(y \mid s_e) - P(y \mid s_o)$. **(Left)** Image Intervention. An image editor transforms the original image s_o into a counterfactual s_e following an MLLM-derived edit. **(Right)** Prompt Intervention. An MLLM rewrites the zero-shot prompt s_o into s_e to incorporate the identified cue.

3.2 SciTX

To diagnose the root causes of model errors, we propose **Sci**entific Method-Inspired **T**extual E**X**planation (SciTX), which implements the explanation generator g from Sec. 3.1. SciTX follows the four steps of the scientific method, namely observation, hypothesis, experiment, and conclusion (Fig. 1).

Observation Retrieval (Survey). To avoid explanations that rely solely on the MLLM's parametric knowledge, we build a support set $\{o_i\}_{i=1}^{K}$ of correctly classified samples from a reference database $\mathcal{D}$, so that the MLLM can contrast them with the misclassified image x. We retrieve observations under three criteria:

- *Correctness*: f classifies o_i correctly, ensuring o_i serves as a reliable reference.
- *Class Relevance*: the label of o_i matches either the ground-truth y or the prediction $\hat{y}$.
- *Semantic Similarity*: o_i is among the K nearest neighbors of x in the visual feature space of f.

We denote this process as

$$\{o_i\}_{i=1}^{K} = \text{Retrieve}(x, \mathcal{D}). \tag{3}$$

Like Retrieval-Augmented Generation (RAG; Lewis et al., 2020), this step grounds the MLLM in external evidence so that it can compare the error against valid references. Feeding multiple images to an MLLM can cause *cross-image information leakage* (Park et al., 2025), where the model unintentionally mixes information across them. To avoid this, we represent each observation o_i as a textual caption that the same MLLM produces from the raw image, rather than as the image itself.

Explanation Generation (Hypothesis). Using the retrieved observations as context, SciTX generates candidate explanations. Following the broader role-prompting literature (Park et al., 2023; Wang et al., 2024a;b), we assign the MLLM an explicit failure-analyst role so that the hypotheses $\{h_j\}_{j=1}^{M}$ describe possible causes of the misclassification rather than the image content:

$$\{h_j\}_{j=1}^{M} = \text{Hypothesize}\left(x, \{o_i\}_{i=1}^{K}\right). \tag{4}$$

Validation and Selection (Experiment and Conclusion). In the final stage, SciTX tests each hypothesis h_j and keeps the most effective one. The MLLM translates h_j into a counterfactual intervention s_j, either an image-editing instruction or a modified prompt, and we score h_j with the Counterfactual Explanation Impact (CEI) metric defined in Sec. 3.3. The hypothesis with the largest CEI becomes the final explanation e:

$$e = \underset{h_j \in \{h_j\}_{j=1}^{M}}{\operatorname{argmax}} \ \mathrm{CEI}(h_j), \quad \text{s.t.} \ \ s_j = \mathrm{Intervention}(h_j). \tag{5}$$

Algorithm 1 summarizes SciTX.

Algorithm 1 SciTX (overview).

Require: Misclassified image x, reference database $\mathcal{D}$, labels $(y, \hat{y})$
Ensure: Explanation e

```
{o_i}_{i=1}^K ← Retrieve(x, D)                          ▷ Observation
{c_i}_{i=1}^K ← Caption({o_i}_{i=1}^K)                  ▷ via the same MLLM
{h_j}_{j=1}^M ← Hypothesize(x, {c_i}_{i=1}^K)           ▷ Hypothesis
for each h_j do
    s_j ← Intervention(h_j)                             ▷ Experiment
    score_j ← CEI(h_j)
end for
e ← argmax_j score_j                                    ▷ Conclusion
return e
```

3.3 Counterfactual Explanation Impact (CEI)

We propose the **Counterfactual Explanation Impact (CEI)**, a counterfactual-inference-based metric (Pearl, 2009; Goyal et al., 2019; Yan & Wang, 2023). If e identifies a causal factor of the error, neutralizing that factor should move the model's prediction toward the ground-truth class. CEI measures this shift directly (Fig. 2).

Definition. CEI evaluates an explanation through a three-step counterfactual procedure. (i) *Derivation* extracts an intervention instruction from e. (ii) *Intervention* applies the instruction to the inference source and constructs a counterfactual scenario. (iii) *Estimation* measures the resulting change in the ground-truth probability. To make the metric reflect e, we never reveal the true causal factor of the error to the MLLM; the intervention is derived from e alone.

Formally, let s_o denote the original inference source and s_e its counterfactual counterpart derived from e. CEI is defined as

$$\mathrm{CEI}(e) = P(y \mid s_e) - P(y \mid s_o), \tag{6}$$

where $P(y \mid \cdot)$ is the model's probability assigned to the ground-truth label y, so a higher CEI provides stronger evidence that e captures a causal factor of the misclassification. We define CEI in two settings, one for standard vision classifiers and one for vision-language models.

Counterfactual Image Intervention. For standard discriminative vision models, we intervene directly in the image space (Chiquier et al., 2025; Farid et al., 2023; Jeanneret et al., 2022). As illustrated in Fig. 2 (Left), an MLLM-guided image editor removes the factor described by e from the input image. The original source is the input image $s_o = \mathrm{Img}$, and its counterfactual is the edited image

$$s_e = \mathcal{G}(\mathrm{Img}, \tau), \tag{7}$$

where $\mathcal{G}$ is a subject-driven image-editing model (Labs et al., 2025) and the editing instruction τ (e.g., "*Remove the sandy background*") is derived from e by an MLLM.

Counterfactual Prompt Intervention. For VLM-based zero-shot classifiers such as CLIP, we instead intervene in the textual prompt space, which modulates the classifier's decision boundary without altering

image pixels. As illustrated in Fig. 2 (Right), an MLLM rewrites the zero-shot prompt to incorporate the cue described by e. Let s_o be a standard zero-shot prompt template (e.g., *"a photo of a {class}"*). The counterfactual prompt is

$$s_e = \mathcal{L}(s_o, e), \tag{8}$$

where the MLLM $\mathcal{L}$ rewrites s_o to explicitly incorporate the cue identified by e (e.g., *"a photo of a {class} on a sandy beach"*).

3.4 LLM Match

We report two metrics. The primary one is the **LLM Match**, which we use for cross-method comparison. The second is CEI (Sec. 3.3), which directly measures the causal effect of an explanation. Since SciTX uses CEI as its hypothesis-selection signal, we treat CEI as a *supplementary* metric.

NEMO comprises misclassified samples from three robustness benchmarks, namely ImageNet-R (Hendrycks et al., 2021), ObjectNet (Barbu et al., 2019), and ImageNet-D (Zhang et al., 2024), each constructed to vary along a specific factor. ImageNet-R varies artistic style and rendition; ObjectNet varies viewpoint, rotation, and background; ImageNet-D varies background, texture, and material via diffusion generation. We sample 400 misclassified images from each, yielding 1,200 in total. By construction, the controlled factor is the dominant driver of misclassification on each benchmark, so a faithful explanation should describe it.

We accordingly define a dataset-specific factor criterion for each benchmark and assess whether the explanation describes it. Given an explanation and the dataset-specific criterion, an LLM-as-a-Judge assigns one of three scores, namely 1.0 (full description), 0.5 (partial description), or 0.0 (no description). To reduce single-judge bias, we average scores from three LLM judges (Sec. 4.1.1).

4 Experiments

4.1 Setting

4.1.1 Models

Target Models. We employ two VLMs, CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023), sourced from Hugging Face, alongside a standard ViT fine-tuned on ImageNet (Steiner et al., 2022; Dosovitskiy et al., 2021) sourced from Timm library (Wightman, 2019).[1]

MLLM Backbone. For explanation generation, we adopt the Qwen-VL series (Bai et al., 2025b;a) as a representative high-performing open-weight MLLM family, which keeps the pipeline reproducible without proprietary API access.

Image Editor. For counterfactual image intervention (Sec. 3.3), we use FLUX.1-Kontext-dev (Labs et al., 2025), an instruction-following image editor.

Evaluation Judges. We distinguish reported metrics from SciTX's internal selection signal. For reporting, both LLM Match and CEI are averaged over three non-Qwen judges, namely `gpt-5-mini-2025-08-07`, `claude-haiku-4-5`, and `gemini-3-flash-preview`, so the explanation generator and the evaluator never share a model family. For hypothesis selection inside SciTX (Sec. 3.2), CEI is computed with the same Qwen-VL backbone used for explanation generation.

4.1.2 Baselines

ErrorRetrieval (Jain et al., 2023; Csurka et al., 2024) retrieves a textual description of the failure from a predefined error corpus, given the misclassified image. Since no prior method tackles per-sample natural-language explanation of classifier errors (Ghosh et al., 2025; Kim et al., 2024), we construct two MLLM-based baselines (CamMLLM, ChangeMLLM). **CamMLLM** feeds Grad-CAM (Selvaraju et al., 2017) saliency maps

[1] CLIP: openai/clip-vit-base-patch32, SigLIP: google/siglip-base-patch16-224, ViT: vit_base_patch16_224.augreg2_in21k_ft_in1k.

Table 1: **Quantitative results across three datasets.** SciTX outperforms all baselines on both metrics. **Bold** marks the best per column.

(a) LLM Match averaged over GPT, Claude, and Gemini

MLLM	ImageNet-R			ObjectNet			ImageNet-D		
	ViT	CLIP	SigLIP	ViT	CLIP	SigLIP	ViT	CLIP	SigLIP
Retrieval	0.277	0.270	0.268	0.499	0.492	0.493	0.142	0.147	0.145
CamMLLM	0.111	0.066	0.044	0.640	0.632	0.603	0.498	0.522	0.467
ChangeMLLM	0.279	0.309	0.291	0.761	0.744	0.758	0.390	0.373	0.358
SciTX	**0.558**	**0.632**	**0.594**	**0.772**	**0.781**	**0.797**	**0.562**	**0.619**	**0.644**

(b) CEI

MLLM	ImageNet-R			ObjectNet			ImageNet-D		
	ViT	CLIP	SigLIP	ViT	CLIP	SigLIP	ViT	CLIP	SigLIP
Retrieval	11.29	11.22	1.81	16.14	7.86	0.82	6.35	3.31	0.70
CamMLLM	13.26	3.17	-0.02	15.25	7.67	0.26	7.05	5.55	0.78
ChangeMLLM	13.39	9.59	0.76	19.02	11.94	1.84	10.88	11.92	7.43
SciTX	**13.63**	**17.52**	**2.37**	**19.06**	**21.29**	**5.57**	**11.97**	**19.27**	**16.94**

to an MLLM, which then verbalizes the regions that drove the misclassification. **ChangeMLLM** follows change captioning (Park et al., 2019), asking an MLLM to describe the differences between the misclassified image and a reference image from the ground-truth or predicted class, and then to infer the error cause from those differences.

4.2 Results

4.2.1 Effectiveness of Our Method

Table 1 reports performance across ImageNet-R, ObjectNet, and ImageNet-D under both metrics. *SciTX outperforms baselines on both metrics across all three datasets.* The baselines instead show dataset-specific weaknesses. Error Retrieval performs comparatively better on ImageNet-R, where its predefined corpus covers rendition-style errors, but collapses on ObjectNet and ImageNet-D, where the corpus does not. ChangeMLLM and CamMLLM also trade ranks across the three datasets. SciTX remains consistently strong, indicating robustness to varying error factors.

4.2.2 Fair Comparison and Generalization of Validation Module

SciTX uses CEI both as its hypothesis-selection signal and as one of our reported metrics, so its CEI lead in Table 1 could in principle be explained by selection alone. To rule this out, we augment each baseline with the same multi-candidate generation and CEI-argmax selection used by SciTX on ImageNet-D / CLIP. Table 2 shows that CEI improves for all three baselines and LLM Match for two, yet SciTX remains higher. The lead is therefore not produced by matched selection. The same module also helps the baselines, suggesting that causal-impact selection is broadly useful for explanation generation, not just for SciTX.

4.2.3 Robustness to Image Intervention Model

To verify that the CEI ranking reflects explanation quality rather than depending on a particular image editor, we re-evaluate on ImageNet-D / ViT with FLUX.1-Kontext-dev (Labs et al., 2025) and Qwen Image Edit (Wu et al., 2025). Table 3 preserves the method ranking across both editors, indicating that SciTX's explanations specify more effective counterfactual interventions than the baselines.

4.2.4 Effect of MLLM Backbone

Table 4 reports performance on ImageNet-R / CLIP across three MLLM backbones (Qwen2.5-VL-7B, Qwen3-VL-8B, Qwen3-VL-30B). Both SciTX and the baselines improve consistently on both LLM Match and CEI

Table 2: **Fair comparison under matched candidate selection (ImageNet-D / CLIP).** Each baseline is augmented with the same multi-candidate generation and CEI-argmax selection used by SciTX. Deltas vs. (w/o) are shown next to (w/) values. Green marks gains and red marks regressions.

Method	w/o Validation		w/ Validation	
	LLM Match	CEI	LLM Match	CEI
Retrieval	0.147	3.31	0.150 (+0.003)	4.72 (+1.41)
CamMLLM	0.522	5.55	0.513 (-0.009)	6.26 (+0.71)
ChangeMLLM	0.373	11.92	0.532 (+0.159)	12.00 (+0.08)
SciTX	–	–	**0.619**	**19.27**

Table 3: **Robustness to image intervention model on ImageNet-D / ViT.** The method ranking is preserved across two distinct image editors (FLUX.1-Kontext-dev and Qwen Image Edit).

Method	CEI	
	Kontext	Qwen Image Edit
Retrieval	$6.351^{(4)}$	$11.542^{(4)}$
CamMLLM	$7.050^{(3)}$	$14.290^{(3)}$
ChangeMLLM	$10.879^{(2)}$	$15.115^{(2)}$
SciTX	$11.967^{(1)}$	$20.397^{(1)}$

Table 4: **Effect of MLLM backbone on ImageNet-R / CLIP.** MLLMs with higher reasoning capability yield higher LLM Match and CEI. **Bold** marks the best per column.

Method	Qwen2.5-VL-7B		Qwen3-VL-8B		Qwen3-VL-30B	
	LLM Match	CEI	LLM Match	CEI	LLM Match	CEI
CamMLLM	0.066	3.17	0.216	4.94	0.213	7.31
ChangeMLLM	0.309	9.59	0.403	12.24	0.483	18.60
SciTX	**0.632**	**17.52**	**0.725**	**20.76**	**0.816**	**25.09**

Table 5: Ablation on observation retrieval strategies (ImageNet-D, CEI). Adding class relevance to similarity-based retrieval consistently improves CEI across all target classifiers. **Bold** marks the best per column.

Observation	ImageNet-D		
	ViT	CLIP	SigLIP
Correctness + Semantic Similarity	10.31	19.05	16.84
+ Class Relevance	**11.97**	**19.27**	**16.94**

as the backbone grows stronger, indicating that stronger MLLM reasoning translates directly into better explanations of model errors.

4.2.5 Retrieved Observations

Fig. 3 shows the observations $\{o_i\}$ retrieved by SciTX for the ground-truth and predicted classes, which serve as comparative visual cues (woven texture vs. metal bars) for the subsequent explanation. We further ablate the retrieval strategy. Replacing class-aware retrieval with pure visual similarity drops CEI across all target classifiers (Table 5). The gain is largest on ViT, suggesting that grounding observations in both the ground-truth and predicted classes is helpful, especially for non-VL classifiers.

4.2.6 Qualitative Analysis

Generated Explanations. In the cases shown in Fig. 6, SciTX's explanations refer to multiple visual factors from the retrieved observations, while the baselines describe only one. For the ostrich graffiti, CamMLLM only highlights the head/beak region, whereas SciTX attributes the error to graffiti colors blending with duck-like features. For the umbrella-on-doormat, ChangeMLLM cites only texture similarity, whereas SciTX further invokes the spatial context of a folded umbrella on a doormat-like surface.

Counterfactual Intervention Examples. Fig. 5 shows how SciTX's explanations translate into counterfactual interventions. A vase misclassified as confectionery is explained by its colorful grocery-store background, and the intervention replaces the background with a neutral backdrop. A jeep misclassified as a tank is explained by its roof-mounted gun, and the intervention removes it.

Pose- and Affect-Aware Reasoning. We identify representative qualitative cases where SciTX invokes high-level cues beyond pixel-level similarity (Fig. 4). For example, an espresso illustration is misclassified as a mobile phone because the figure's holding pose mirrors that of phone-using subjects, and a stylized wheelbarrow drawing is misclassified as a guillotine because both evoke a violent emotional tone.

Figure 3: Observations retrieved by SciTX. Reference images for the ground-truth (Hamper) and predicted (Prison) classes that provide comparative visual cues for the explanation.

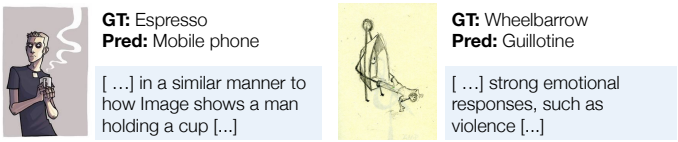


Figure 4: Beyond pixel-level similarity: (left) an espresso illustration confused with a mobile phone; (right) a stylized wheelbarrow drawing confused with a guillotine.

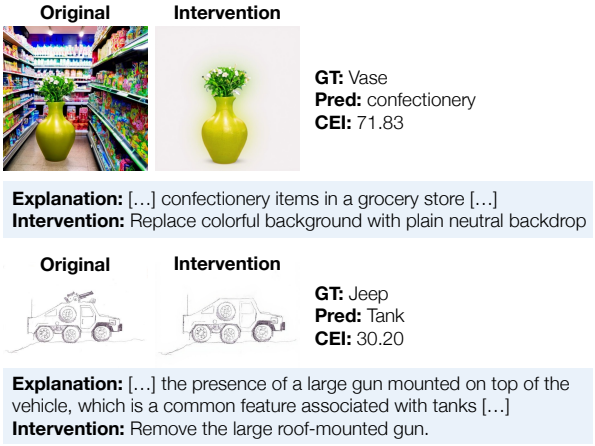


Figure 5: SciTX's explanations translated into counterfactual interventions. CEI measures the change in the ground-truth probability after applying each intervention.

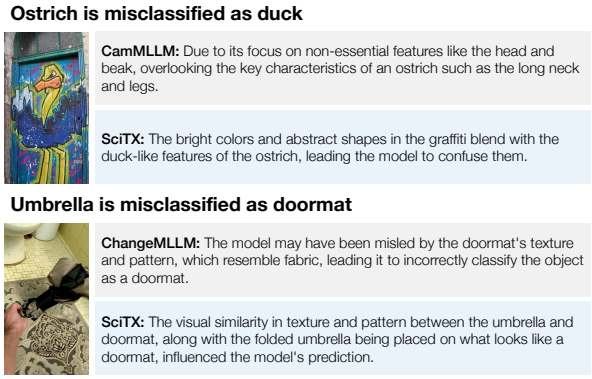


Figure 6: Generated textual explanations. CamMLLM describes the localization information; ChangeMLLM and ours describe the differences or commonalities between the misclassified sample and observations.

4.2.7 Application: Global Explanation

SciTX can also be used to characterize a model's failure modes at the dataset level. We use an MLLM to aggregate all sample-wise explanations produced by SciTX into a single concise global explanation. For CLIP on ImageNet-R, the resulting summary identifies the dominant failure mode as follows:

> *"[. . .] struggles with domain shifts, frequently misclassifying stylized representations—such as origami, tattoos, and cartoons—based on* ***their artistic style rather than their actual structural characteristics****."*

This description is consistent with ImageNet-R being a rendition benchmark, suggesting that the per-sample explanations carry a distribution-level signal that emerges naturally under aggregation.

As a single-condition probe on CLIP / ImageNet-R with Qwen2.5-VL-7B, we plug it back into evaluation as the explanation supplied to the CEI scorer. As reported in Table 6, CEI improves from 17.52 with per-sample explanations to 19.42 with the single global explanation, indicating that the aggregated description preserves the per-sample causal signal under this single condition. We view this as preliminary evidence that SciTX can also serve as a dataset-level failure-mode summarizer.

Table 6: CEI of SciTX on CLIP / ImageNet-R.

Local	17.52
+ Global	19.42 (+1.9)

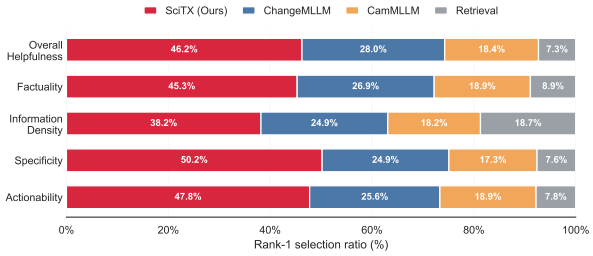


Figure 7: Rank-1 selection ratio of different helpfulness dimensions of human evaluation. We observe that AI practitioners rank our method first more often than the others.

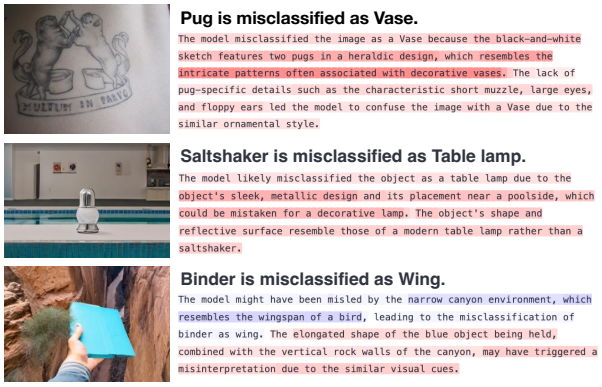


Figure 8: Human highlight result. Red means that humans perceive the explanation is good; blue means that humans perceive the explanation is bad. We average the highlight results of participants and visualize using a color spectrum.

4.3 Human Evaluation

Study Design. We evaluate on 30 misclassified samples (10 from each of ImageNet-R, ObjectNet, and ImageNet-D). To ensure that all methods produce non-trivial explanations, we restrict to samples whose harmonic mean of CEI across methods is relatively high. Each sample is paired with four candidate explanations, and 30 AI practitioners recruited via Prolific[2] rank them along five dimensions (Overall Helpfulness, Factuality, Information Density, Specificity, and Actionability) using a forced ranking without ties. See Appendix D for the evaluation interface and dimension definitions.

Results. As shown in Fig. 7, SciTX consistently achieves the highest rank-1 selection ratio across all five evaluation dimensions (see detailed results in Appendix Fig. 14). Qualitatively, users highlight good (red, $+1.0$) or bad (blue, -1.0) parts of SciTX's explanations, averaged across participants and visualized on a color spectrum (Fig. 8). Users mark concrete visual cues such as "heraldic design" and "sleek, metallic design" positively, but mark logical leaps such as "narrow canyon . . . resembles the wingspan of a bird" negatively.

4.4 Failure Cases

We observe two failure modes in SciTX's outputs (Fig. 9). First, SciTX may build the explanation on top of an object that the MLLM has misidentified, as in the padlock case where the explanation discusses a basket–hamper handle resemblance without ever referring to the padlock. Second, it may rely on shared scene context rather than discriminative object features, as in the chair-vs-bench case where the confusion is attributed to a common brick-wall background.

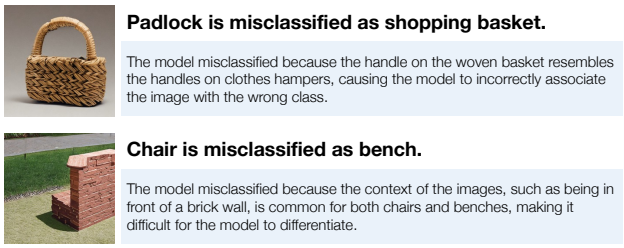


Figure 9: Failure cases.

5 Conclusion

We introduce NEMO, a task and benchmark for natural-language explanations of vision-model errors. We further propose SciTX, a generation-based method that follows a four-stage scientific-method-inspired pipeline and validates each candidate explanation against two complementary metrics, Counterfactual Explanation

[2]https://www.prolific.com/. Participants were compensated at £10/h; the average completion time was 57 minutes.

Impact (CEI) and LLM Match. On NEMO, SciTX outperforms every retrieval-based and MLLM-augmented baseline on both metrics. A user study with 30 AI practitioners further ranks SciTX first on all five helpfulness dimensions. Future work includes extending the pipeline to broader debugging settings such as generative-model failures and multi-modal classifiers.

Broader Impact Statement

As LLM-based interfaces become standard for interacting with deployed models, explanations that describe plausible failure causes can help practitioners and non-expert users decide whether to trust, escalate, or retrain a model. As with any MLLM-generated text, the explanations are best treated as a starting point for inspection rather than a verified diagnosis, and our pipeline is intended to operate within the review-and-iterate workflow that already governs model deployment. Beyond this, we do not foresee additional societal consequences that warrant separate discussion.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Mia Chiquier, Orr Avrech, Yossi Gandelsman, Berthy Feng, Katherine Bouman, and Carl Vondrick. Teaching humans subtle differences with diffusion. *arXiv preprint arXiv:2504.08046*, 2025.

Gabriela Csurka, Tyler L Hayes, Diane Larlus, and Riccardo Volpi. What could go wrong? discovering and describing failure modes in computer vision. *European Conference on Computer Vision Workshops (ECCVW)*, 2024.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Lisa Dunlap, Yuhui Zhang, Xiaohan Wang, Ruiqi Zhong, Trevor Darrell, Jacob Steinhardt, Joseph E Gonzalez, and Serena Yeung-Levy. Describing differences in image sets with natural language. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Sabri Eyuboglu, Maya Varma, Khaled Kamal Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Re. Domino: Discovering systematic errors with cross-modal embeddings. In *International Conference on Learning Representations (ICLR)*, 2022.

Karim Farid, Simon Schrodi, Max Argus, and Thomas Brox. Latent diffusion counterfactual explanations. *arXiv preprint arXiv:2310.06668*, 2023.

François-Guillaume Fernandez. Torchcam: class activation explorer. `https://github.com/frgfm/torch-cam`, March 2020.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2019.

Shantanu Ghosh, Rayan Syed, Chenyu Wang, Vaibhav Choudhary, Binxu Li, Clare B Poynton, Shyam Visweswaran, and Kayhan Batmanghelich. Ladder: Language-driven slice discovery and error rectification in vision classifiers. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165*, 2019.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International Conference on Machine Learning (ICML)*, 2020.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

Nam Hyeon-Woo, Kim Yu-Ji, Byeongho Heo, Dongyoon Han, Seong Joon Oh, and Tae-Hyun Oh. Scratching visual transformer's back with uniform attention. In *IEEE International Conference on Computer Vision (ICCV)*, 2023.

Saachi Jain, Hannah Lawrence, Ankur Moitra, and Aleksander Madry. Distilling model failures as directions in latent space. In *International Conference on Learning Representations (ICLR)*, 2023.

Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Diffusion models for counterfactual explanations. In *Asia Conference on Computer Vision (ACCV)*, 2022.

Mark T. Keane, Eoin M. Kenny, Eoin Delaney, and Barry Smyth. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conferences on Artificial Intelligence Organization, 2021.

Younghyun Kim, Sangwoo Mo, Minkyu Kim, Kyungmin Lee, Jaeho Lee, and Jinwoo Shin. Discovering and mitigating visual biases through keyword explanation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations (ICLR)*, 2023.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning (ICML)*, 2020.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023b.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2020.

Elisa Nguyen, Johannes Bertram, Evgenii Kortukov, Jean Y Song, and Seong Joon Oh. Towards user-focused research in training data attribution for human-centered explainable ai. *arXiv preprint arXiv:2409.16978*, 2024.

Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, 2023.

Yeji Park, Minyoung Lee, Sanghyuk Chun, and Junsuk Choe. Mitigating cross-image information leakage in lvlms for multi-image tasks. *arXiv preprint arXiv:2508.13744*, 2025.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Keivan Rezaei, Mehrdad Saberi, Mazda Moayeri, and Soheil Feizi. PRIME: Prioritizing interpretability in failure mode extraction. In *International Conference on Learning Representations (ICLR)*, 2024.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

Tamar Rott Shaham, Sarah Schwettmann, Franklin Wang, Achyuta Rajaram, Evan Hernandez, Jacob Andreas, and Antonio Torralba. A multimodal automated interpretability agent. In *International Conference on Machine Learning (ICML)*, 2024.

Andreas Peter Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research (TMLR)*, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024a.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2024b.

Greta Warren, Ruth M. J. Byrne, and Mark T. Keane. Categorical and continuous features in counterfactual explanations of ai systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 2023.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research (TMLR)*, 2022.

Ross Wightman. Pytorch image models. https://github.com/huggingface/pytorch-image-models, 2019.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

Jingquan Yan and Hao Wang. Self-interpretable time series prediction with counterfactual explanations. In *International Conference on Machine Learning (ICML)*, 2023.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: evaluating large multimodal models for integrated capabilities. In *International Conference on Machine Learning (ICML)*, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE International Conference on Computer Vision (ICCV)*, 2023.

Chenshuang Zhang, Fei Pan, Junmo Kim, In So Kweon, and Chengzhi Mao. Imagenet-d: Benchmarking neural network robustness on diffusion synthetic object. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Contents

A. Prompts
B. Experimental Details
C. More Experimental Results
D. Human Study

A Prompts

A.1 SciTX

The prompt of SciTX is given below.

```
You are an expert in visual reasoning and model error analysis.

The first image has the true label **{true_label_1}**, but the model incorrectly predicted it as
**{prediction}**.

***Task***
Your goal is to identify visual differences between these sets of images and analyze the likely
reason for the models misclassification.

***Steps***
1. Examine all provided sets (A, B, and/or C).
   - If a set is missing, use the remaining sets for comparison or inference.
2. Identify visual or semantic differences relevant to the misclassification.
3. Consider whether factors like background, object texture, pose, or context influenced the error.
4. Formulate a concise, evidence-based explanation.

***Evidence Sets***
- **Set A** (True Label: {true_label_1})
  {set_a_captions}

- **Set B** (Predicted Label: {prediction})
  {set_b_captions}

- **Set C** (Random Examples)
  {set_c_captions}

***Output Format***
Write your reasoning **strictly** in JSON format as shown below.

‘‘‘json
{
   "cause_of_error": "A concrete, evidence-based conclusion directly explaining why the model
   predicted ’{prediction}’ instead of ’{true_label_1}’. Mention the most influential visual cue"
}
‘‘‘

Do not include any text outside of the JSON object in your output.
```

A.2 Baseline Prompts

For completeness, we list the generation prompts used by the MLLM-augmented baselines. ErrorRetrieval is non-generative and does not use a prompt.

CamMLLM.

```
You are an expert visual reasoning assistant.

Given the Grad-CAM visualization, the model prediction, and the correct label, identify the main
cause of the misclassification.

Your task is to analyze why the model made the wrong prediction.

Follow these reasoning steps:
1. Describe what regions the Grad-CAM highlights (what the model focused on).
2. Assess whether these regions are relevant or misleading for the true label.
3. Identify any important visual cues for the true label that were ignored.

Finally, summarize your reasoning in **strict JSON format** as follows:
‘‘‘json
{
  "cause_of_error": "concise and precise sentences explaining the most probable cause of
  misclassification"
}
‘‘‘

Inputs:
Prediction: {prediction_cls_name}
True label: {true_cls_name}

Do not include any text outside of the JSON object in your output.
```

ChangeMLLM.

```
You are an expert visual reasoning assistant.

The first image is {true_cls_name}, but the model predicted it as {prediction_cls_name} (incorrect).

The second image is {src_cls_name}, and the model predicted it correctly.

Your task is to identify the differences between these images and analyze why the model made the
wrong prediction.

Follow these steps:
1. Describe key visual differences between the two images.
2. Suggest what kind of visual or contextual cue the model failed to capture.

Finally, summarize your reasoning in **strict JSON format** as follows:
‘‘‘json
{
  "visual_differences": "Describe the main differences between the two images",
  "cause_of_error": "1-2 concise and precise sentences explaining the most probable cause of
  misclassification"
}
‘‘‘

Do not include any text outside of the JSON object in your output.
```

A.3 Evaluation

For counterfactual image intervention:

```
An image with the true label **{true_label_1}** was incorrectly predicted as **{prediction}**.

**Context for Error:**
- **Hypothesis for Misclassification:** {hypothesis}

Your task is to write a concise editing prompt that would modify the image to correct this error.
The prompt should describe a minimal visual change that helps a generative model fix the issue (e.g
., "Add a hat to the cat").

**Guidelines:**
1. **Address the Hypothesis:** The edit must directly resolve the hypothesized cause of the
misclassification.
2. **Be Specific & Semantic:** The edit must be a *content-based* visual change (e.g., "Add a stem
to the apple," "Change the shirt color to blue"). Do NOT suggest generic, non-specific edits like "
make {true_label_1} larger," "sharpen the image," or "increase contrast."
3. **Avoid the Class:** Do NOT reference or imply the incorrect label **{prediction}** in your
solution.
4. **Be Concise:** Output only one short, imperative command (ideally under 10 words).

**Output Format:**
Write your reasoning strictly in JSON format as shown below.

‘‘‘json
{
   "reasoning": "Briefly explain why this new prompt is better. The reasoning should be grounded in
    Hypothesis for Misclassification.",
   "solution": "Short, concise, imperative editing prompt" # 1 sentence and lower than 10 words
}
‘‘‘
Do not include any text outside of the JSON object in your output.
```

For counterfactual zero-shot prompt intervention:

```
An image with the true label **{true_label_1}** was incorrectly predicted as **{prediction}**.

**Diagnostic Context:**
1. **Specific Hypothesis:** {hypothesis}
   *(The specific visual reason this image was misclassified.)*
2. **Original Failed Prompt:** "{original_prompt}"

**Task:**
Generate a **concise, CLIP-friendly zero-shot prompt** to correct this error.
Unlike human descriptions, CLIP prefers simple, direct captions. Avoid complex sentences or lists
of features.

**Strict Guidelines for CLIP:**
1. **Keep it Short:** Use fewer than 15 words.
2. **Focus on the Scene:** Instead of listing body parts (e.g., "showing legs, neck, beak"),
describe the overall scene or action (e.g., "standing in a grassy field").
3. **No "Checklists":** Do NOT use phrases like "clearly showing," "distinctive features," or "
characterized by."
4. **No Negatives:** Do NOT use words like "not," "no," or "unlike." (e.g., instead of "not a close
-up," say "a full body shot").
5. **Address the Hypothesis:** Subtlely include the missing visual cue identified in the Hypothesis
, but blend it naturally into the caption.

**Output Format:**
Write your reasoning strictly in JSON format:
```

```
```json
{
   "reasoning": "Briefly explain how this simplified prompt targets the hypothesis without over-
   complicating.",
   "solution": "a photo of [adjective] {true_label_1} [context/action]"
}
```
```

B Experimental Details

B.1 Datasets

Dataset-level scope (controlled factor, 400-per-dataset sampling, 1,200 total) is described in Sec. 3.4. Below we record source and reproducibility detail.

ImageNet-R ImageNet-R (Hendrycks et al., 2021) covers 200 classes across 30,000 rendition images of ImageNet classes.

ImageNet-D ImageNet-D (Zhang et al., 2024) is released under the MIT License. Its images are generated by diffusion models.

ObjectNet ObjectNet (Barbu et al., 2019) contains images captured in challenging conditions, *e.g.*, varied backgrounds, rotations, and imaging viewpoints.

Observation Source For retrieval, the observation source $\mathcal{D}$ pools the un-sampled examples from these three benchmarks together with a 100-class subset of ImageNet (Deng et al., 2009).

B.2 Models

Multimodal Large Language Models For explanation generation, we employ the Qwen-VL series (Bai et al., 2025b;a). To evaluate the generated outputs, we average scores from three LLM judges, namely `gpt-5-mini-2025-08-07`, `claude-haiku-4-5`, and `gemini-3-flash-preview`. For all models, the top-p and temperature parameters are set to 0 to ensure deterministic results.

Image Editing Model We utilize two instruction-following image editing models for our counterfactual interventions, namely FLUX.1-Kontext-dev (Labs et al., 2025) and Qwen Image Edit (Wu et al., 2025).

Computational Resources All experiments are conducted on a single NVIDIA RTX A6000 GPU. This setup is sufficient for our study as we use efficient inference APIs to minimize computational overhead.

C Experimental Results

Qualitative Result More samples of generated explanations from baselines and ours are given. See Fig. 10 for ImageNet-D, Fig. 11 for ImageNet-R, and Fig. 12 for ObjectNet.

D Human Study

Recognizing that mathematically optimized metrics do not always translate into good explanations for humans (Keane et al., 2021; Warren et al., 2023), we conducted a human study to evaluate the practical helpfulness of generated error explanations. We use 30 randomly selected misclassified samples, 10 each from ImageNet-R, ObjectNet, and ImageNet-D. To ensure a fair comparison, we selected samples with a high harmonic mean of CEI across all methods. Each participant evaluated a total of 15 images, with 5 samples drawn from each dataset.

As illustrated in Fig. 13, the evaluation interface provides participants with instructions, the input image, the ground truth, the model's prediction, and four candidate explanations. To mitigate bias, both the order of

samples and the explanations were randomized. Participants were asked to rank the explanations across five dimensions without ties:

- Overall Helpfulness: Assessing which explanation is most helpful for understanding the error.
- Factuality: Evaluating the factual accuracy and reliability of the explanation.
- Information Density: Identifying which explanation is more concise yet information-dense.
- Specificity: Determining which explanation provides the most specific and relevant details about the model's error.
- Actionability: Judging which explanation provides the most actionable and useful insights for model improvement.

We recruited 30 AI practitioners via Prolific, a platform specialized in high-quality human data collection. Participants were compensated at a rate of £10/h, and the average completion time for the study was 57 minutes. Figure 14 shows the result; the ratio of rank-1 of ours is higher than others.

Ground Truth: Pill bottle
Prediction: Water bottle

Retrieval
The ground truth label for the image is actually incorrect.

CamMLLM
The model focused on the transparent part of the bottle, which is more characteristic of a water bottle, while ignoring the pill bottle's shape and contents, which are crucial for identifying it.

ChangeMLLM
The model might have been misled by the bottle's shape and the presence of a cap, which are common features of both pill bottles and water bottles. However, the content inside the bottle in the first image is distinct from typical water bottle contents, which could have confused the model.

Ours
The model misclassified the image because of the visual similarity between a pill bottle and a water bottle, particularly the blue color and the white cap. The background, which is a grassy field, might have led the model to associate the image more with a water bottle due to its common outdoor usage.

Ground Truth: Spatula
Prediction: Shovel

Retrieval
Only the skeletal remains of the object are shown.

CamMLLM
The model focused on the handle of the shovel, which is a common feature for both shovels and spatulas. However, the GradCAM also highlighted the blade area, which is more characteristic of a spatula. The model's prediction was misled by the handle, as it is a shared feature between shovels and spatulas.

ChangeMLLM
The model likely confused the spatula's ridged edge with the texture of a shovel blade, leading to the incorrect classification.

Ours
The model's misclassification can be attributed to the context and background. The spatula is placed in a dirt pit, which might have led the model to associate it more with a tool used for digging, thus predicting 'shovel'. The visual similarity in texture and shape between the spatula and the shovel could also contribute to this confusion.

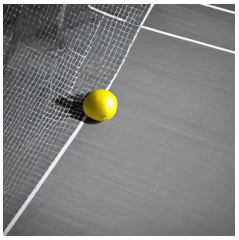
Ground Truth: Binder
Prediction: Envelope

Retrieval
Image is a print on a curved surface like a mug or shirt.

CamMLLM
The model focused on the yellow gradient pattern, which is not relevant to the true label 'binder'. The true label 'binder' would require attention to the functional shape and structure of the object, which was not highlighted by the GradCAM.

ChangeMLLM
The model may have been misled by the folded nature of the items, which can sometimes resemble envelopes due to their flat presentation. Additionally, the context of the first image being labeled as a 'binder' might have influenced the model's prediction.

Ours
The model likely misclassified the image as an 'envelope' due to the presence of the letter 'S' logo, which resembles the design often found on envelopes. The neat folding and the visible stitching along the edges also contribute to the envelope-like appearance, leading the model to confuse it with an envelope rather than a binder.



Ground Truth: Lemon
Prediction: Ping pong ball

Retrieval
Object is viewed from directly above, obscuring the side profile.

CamMLLM
The model focused on the bright yellow region near the center, which is consistent with the color of a ping-pong ball. However, this region is also characteristic of lemons, leading to the misclassification.

ChangeMLLM
The model may have been misled by the bright yellow color, which is common for both lemons and ping-pong balls, leading it to incorrectly classify the lemon as a ping-pong ball.

Ours
The model likely misclassified the lemon as a ping-pong ball due to the bright yellow color and round shape, which are common characteristics of both objects. The context of the image, showing a tennis court with a net, may have led the model to associate the yellow object with a ping-pong ball rather than a lemon.

Ground Truth: Frying pan
Prediction: Stove

Retrieval
Typically co-occurring objects are missing from the scene.

CamMLLM
The model focused on the colorful checkered floor, which is irrelevant to the true label 'frying_pan'. The frying pan itself was not highlighted by the GradCAM, leading to the misclassification.

ChangeMLLM
The model might have been misled by the presence of the frying pan in the first image, which is a common kitchen tool often associated with stoves. However, the pan's position and the overall context of the image suggest that it is being used for cooking rather than the stove itself.

Ours
The model may have misclassified the frying pan as a stove due to the visual similarity in the checkered countertop and the industrial setting. The checkered pattern on the countertop could be mistaken for the tiled backsplash typically found behind a stove, leading the model to predict the presence of a stove.

Figure 10: Generated explanation from ImageNet-D samples.

Ground Truth: Broom
Prediction: Military aircraft

Retrieval
Object is viewed from below, causing extreme foreshortening.

CamMLLM
The model focused on the geometric shapes and lines within the image, which are characteristic of a military aircraft, rather than the overall context or the presence of a broom. This led to the misclassification as the model did not recognize the broom's shape and structure.

ChangeMLLM
The model may have been misled by the white silhouette in the first image, which resembles the shape of a military aircraft, leading it to incorrectly predict 'military_aircraft' instead of 'broom'. The simplicity and distinctiveness of the broom in the second image likely helped the model recognize it correctly.

Ours
The model misclassified the image due to the strong visual similarity between the glowing silhouette of a military aircraft and the broom. The aircraft's streamlined, angular shape and the way it is positioned against the wall closely resemble the broom's form, leading the model to confuse them.

Ground Truth: Bell pepper
Prediction: Banana

Retrieval
An unnatural color filter alters the object's appearance.

CamMLLM
The model focused on the bright, glowing areas around the banana, which may have been interpreted as a sign of ripeness or sweetness, leading it to predict 'banana' instead of recognizing the distinct shape and color of the bell pepper.

ChangeMLLM
The model may have been trained on a dataset where the bell pepper was often depicted with a different color or shape, leading to confusion when presented with the yellow bell pepper costume.

Ours
The model may have confused the large, curved shape of the bell pepper with the similar contour of a banana. The vibrant yellow color and elongated form strongly resemble a banana, even though the texture and details are distinctly different.

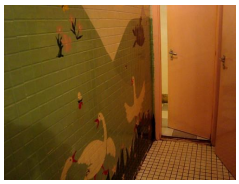
Ground Truth: Dragonfly
Prediction: Canoe

Retrieval
Object is a plush toy lacking rigid structure.

CamMLLM
The model focused on the red-highlighted regions, which appear to be the body parts of the dragonfly, but it misclassified them as a canoe due to the similarity in shape and color gradient. The model did not consider the distinct features of the dragonfly's wings and body, which are crucial for identifying the true label.

ChangeMLLM
The model might have been trained on a dataset that did not include images of dragonflies, leading to incorrect classification.

Ours
The model's misclassification from 'dragonfly' to 'canoe' is likely due to the visual similarity between the wooden toy airplane and the canoe. Both objects have a streamlined, elongated shape with a single central body and two side structures resembling wings or oars. The simplicity and symmetry of the shapes, along with the lack of distinct features that differentiate a dragonfly from a canoe, contributed to the model's confusion.



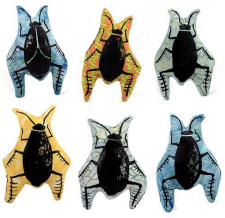
Ground Truth: Goose
Prediction: Bathtub

Retrieval
Object pattern blends perfectly into the background (camouflage).

CamMLLM
The model focused on the colorful, abstract patterns on the wall, which are not relevant to the true label 'goose'. The model did not consider the presence of the goose, which was likely obscured by the vibrant background.

ChangeMLLM
The model may have been misled by the context of the image, assuming that the presence of a door and a tiled floor indicated a bathtub rather than recognizing the mural as the primary subject.

Ours
The model misclassified the goose as a bathtub due to the presence of bathroom-related elements such as the tiled floor, shower, and sink in the background. The wall art depicting geese did not provide enough distinguishing features to override the contextual clues suggesting a bathroom setting.



Ground Truth: Goose
Prediction: Bathtub

Retrieval
A different species has evolved to mimic the appearance of the object.

CamMLLM
The model focused on the body and wings of the insect, which are common features in both electric guitars and cockroaches. However, the presence of antennae and the overall shape of the insect are distinctive features of a cockroach that the model did not emphasize, leading to the misclassification.

ChangeMLLM
The model may have been misled by the variety of colors and patterns in the first image, which could have distracted it from recognizing the commonality of the cockroach shapes. It might have focused more on the unique visual elements rather than the overall form.

Ours
The model's misclassification can be attributed to the presence of artistic and decorative elements in the cockroach designs that resemble the stylized and colorful features of the electric guitars. The vibrant colors, intricate patterns, and overall aesthetic of both the cockroaches and guitars share similar artistic qualities, leading the model to confuse them.

Figure 11: Generated explanation from ImageNet-R samples.

Ground Truth: Bench
Prediction: Binder

Retrieval
Camera is too close, showing only texture without shape.

CamMLLM
The model may have been misled by the texture and color patterns in the central and upper parts of the image, which are not indicative of a bench, leading to the incorrect prediction of 'Binder (closed).'

ChangeMLLM
The model likely failed to recognize the overall shape and context of the object in the first image due to its close-up nature and lack of surrounding elements that typically define a bench.

Ours
The model misclassified the image due to the similarity in the flat, rectangular shape and the presence of a metal fastener, which are common features of binders. The reddish-brown color and the texture of the surface also contributed to the confusion, as they resemble the appearance of a closed binder.

Ground Truth: Toaster
Prediction: Portable heater

Retrieval
Camera is too close, showing only texture without shape.

CamMLLM
The model focused on the control panel, which is characteristic of a portable heater, while ignoring the key feature of the toaster, i.e., the slots for bread.

ChangeMLLM
The model might have been misled by the shape and size of the device, which resembles a toaster, but the presence of a label and the context of the room (a bed) could have confused it into predicting it as a portable heater.

Ours
The model misclassified the object due to the similar rectangular shape of the appliance and the lack of distinguishing features that differentiate a toaster from a portable heater. The bed setting might have contributed to the model's uncertainty, as toasters are typically kitchen appliances, not bedroom items.

Ground Truth: Measuring cup
Prediction: Bucket

Retrieval
Image is overexposed, washing out the object's contours.

CamMLLM
The model focused too broadly on the bucket's body and handle, missing the distinctive spout characteristic of a measuring cup, leading to the misclassification.

ChangeMLLM
The model might have been misled by the handle on the blue container in the first image, which resembles a bucket's handle, leading it to incorrectly classify the object as a bucket instead of recognizing the measurement markings typically associated with a measuring cup.

Ours
The model's misclassification can be attributed to the lack of distinct features that differentiate a measuring cup from a bucket. The blue object in both images appears to have a handle and a spout-like structure, which are common to both utensils. However, the context of the first image, featuring a cluttered workspace with various electronic equipment, suggests a more utilitarian environment where buckets could be commonly used for storage or water collection. In contrast, the second image shows a single blue object without additional context, leading the model to categorize it as a bucket due to its shape and lack of specific measuring markings.

Ground Truth: Bath towel
Prediction: Soap dispenser

Retrieval
Object is viewed from below, causing extreme foreshortening.

CamMLLM
The model misclassified the image due to its strong focus on the soap dispenser, overlooking the presence of the bath towel.

ChangeMLLM
The model may have been misled by the towel's appearance alone without recognizing the context of its placement, which is typically found in a bathroom setting.

Ours
The model likely misclassified the object due to the presence of a blue fabric piece at the top of the towel, which resembles the spout of a soap dispenser. The texture and shape of this piece, combined with the overall appearance of the object, may have led the model to confuse it with a soap dispenser.

Ground Truth: Binder
Prediction: Weight

Retrieval
Incorrect white balance makes the object look unnatural.

CamMLLM
The model focused on the blue fabric-like areas due to their color similarity with the exercise equipment, leading it to predict 'Weight (exercise)' instead of the correct 'Binder (closed).'

ChangeMLLM
The model likely misclassified the first image because it included a person's legs, which can be mistakenly associated with a weight exercise context, despite the actual object being a closed binder.

Ours
The model misclassified the closed purple binder as a 'Weight (exercise)' due to the presence of a blue folder-like object on the wooden floor, which resembles a weight plate often used in gym settings. The blue folder's shape and color might have been mistaken for a weight plate, leading to the incorrect prediction.

Figure 12: Generated explanation from objectnet samples.

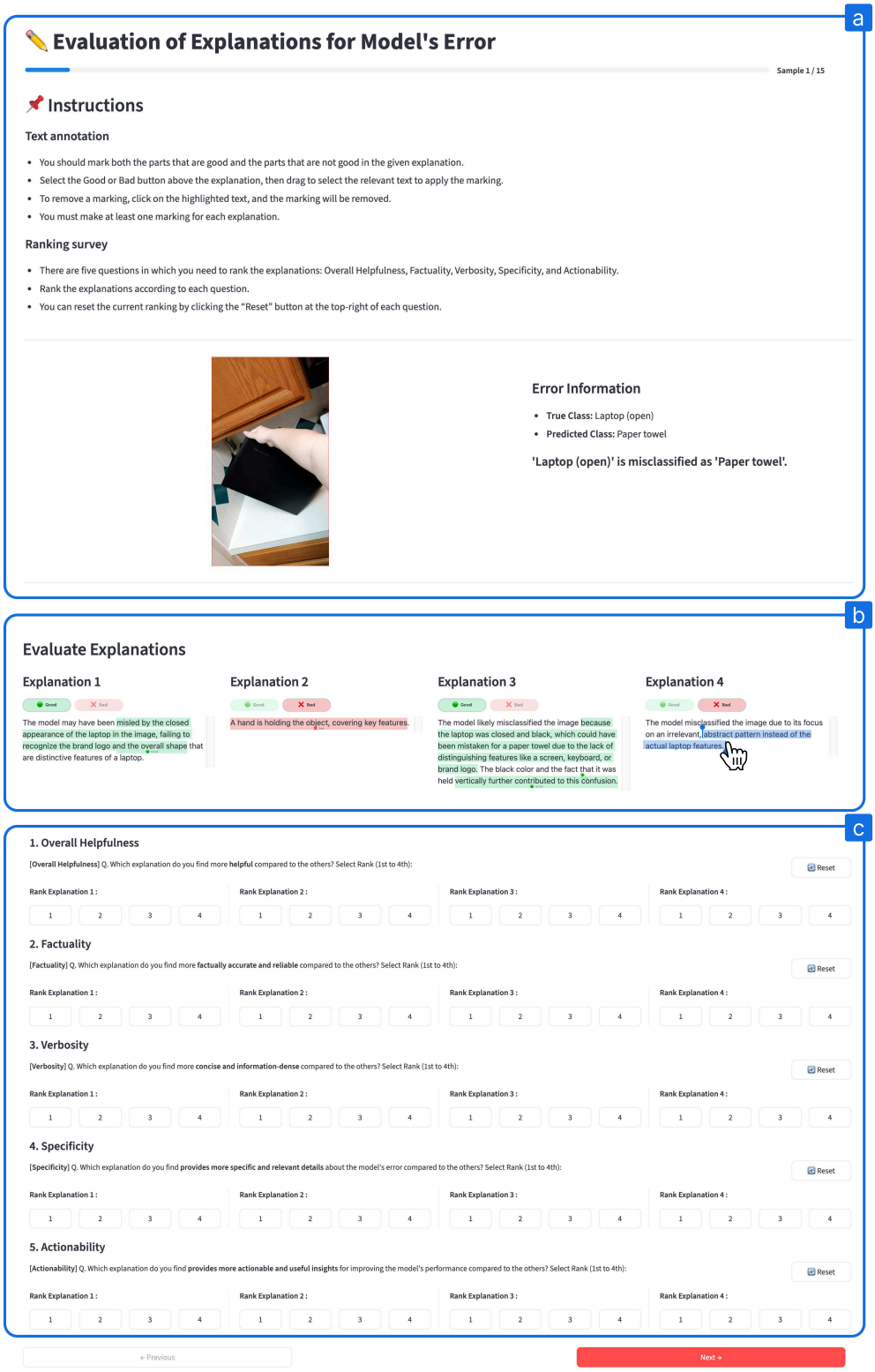


Figure 13: Interface for human evaluation. (a) The task instruction and an example showing the ground-truth and predicted classes. (b) An interface that allows participants to directly annotate helpful and unhelpful parts of the explanation. (c) An interface for ranking explanations across evaluation criteria.

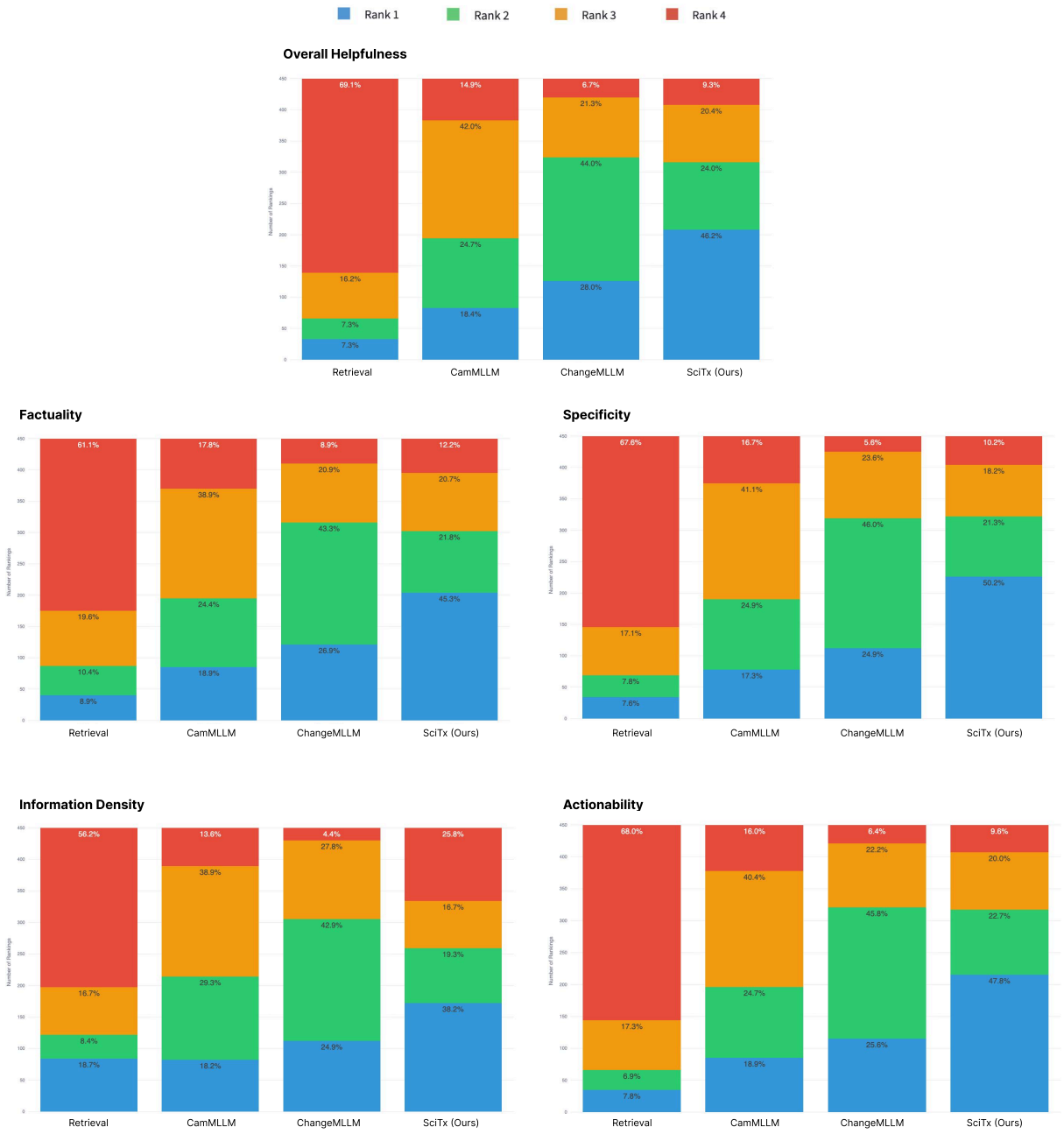


Figure 14: Detailed human evaluation results for five categories