NEMO: Benchmarking Natural-Language Explanations of Vision Model Errors

Nam Hyeon-Woo [1] **Yoonsu Kim** [2] **Kihoon Son** [2] **Juho Kim** [2] **Tae-Hyun Oh** [2]

Abstract

With the rise of agentic LLM systems, non-experts increasingly interact with vision classifiers through natural language. When a classifier misclassifies an image, users need a faithful account of *why*. Such explanations help users diagnose failure modes and debug the model. Progress on this need is blocked by two gaps. First, no benchmark evaluates free-form natural-language explanations of vision model errors. Second, existing retrieval-based methods are limited to a fixed corpus of error sentences and cannot describe failure modes outside it. We address both gaps. We introduce NEMO, a task and benchmark, paired with an LLM-as-a-Judge protocol that scores explanations for whether they describe the failure factor. We then propose SciTX, a generation-based method powered by Multimodal Large Language Models (MLLMs). SciTX is composed of a four-stage pipeline: observation, hypothesis, experiment, and conclusion. The pipeline retrieves contrastive observations, generates candidate hypotheses, validates each via a counterfactual intervention, and selects the hypothesis whose intervention shifts the model's prediction most toward the ground-truth class. SciTX outperforms retrieval-based and MLLM-augmented baselines. A human study with AI practitioners also ranks SciTX first.

1. Introduction

Recent advances in large language models (LLMs) and agentic AI systems have made natural-language interfaces a common way for non-experts to engage with machine learning models. Product managers, domain specialists, and other stakeholders often use natural-language outputs to understand model behavior and make decisions, frequently without direct access to the model's internals. This shift gives natural-language explanations of *failure cases* a central role. When a vision classifier misclassifies an image, a non-expert reader needs a faithful account of *why* to diagnose the failure mode and debug the model. Existing retrieval-based approaches (Jain et al., 2023; Eyuboglu et al., 2022) are limited to the explanations in their predefined corpus and cannot describe error contexts that fall outside it.

Furthermore, no benchmark exists for evaluating free-form natural-language explanations of model errors. To address both gaps, we first introduce NEMO, a task for producing such explanations paired with a benchmark of 1,200 misclassified images drawn from ImageNet-R, ObjectNet, and ImageNet-D, each varying along a distinct factor, namely artistic style, viewpoint, and low-level attributes. To assess explanation quality at scale, we adopt an LLM-as-a-Judge (LLM Match) protocol that scores whether each explanation describes the dataset-specific failure factor, providing an automated and reproducible measure of NEMO performance.

We propose SciTX (**Sci**entific Method-Inspired **T**extual E**X**planation), a generation-based method using MLLMs. We design SciTX after the scientific method: gather evidence, form hypotheses, test each, and select the one with the largest measured effect. SciTX has four stages namely observation, hypothesis, experiment, and conclusion. Given a misclassified image, it retrieves reference observations from correctly classified samples in the ground-truth and predicted classes, generates candidate explanations grounded in this evidence, validates each via a counterfactual intervention, and selects the candidate whose intervention shifts the model's prediction most toward the ground-truth class. We measure this shift with the Counterfactual Explanation Impact (CEI). This explicit counterfactual test favors explanations whose intervention shifts the model's prediction toward the ground-truth class, rather than merely plausible-sounding ones.

Across all three datasets in NEMO, three target classifiers (ViT, CLIP, SigLIP), and the Qwen-VL family of MLLM backbones, SciTX outperforms every retrieval-based and MLLM-augmented baseline on both LLM Match and CEI. A human study with AI practitioners further ranks SciTX first on every one of the five helpfulness dimensions, including factuality, specificity, and actionability. In summary,

[1]Electrical Engineering, POSTECH, Pohang, South Korea [2]School of Computing, KAIST, Daejeon, Republic of Korea. Correspondence to: Tae-Hyun Oh <taehyun.oh@kaist.ac.kr>.

Accepted at the Trustworthy AI for Good (AI4GOOD) Workshop at the 43rd International Conference on Machine Learning (ICML 2026), Seoul, South Korea.

we (i) introduce NEMO, a benchmark for free-form natural-language explanations of vision model errors paired with an LLM-as-a-Judge protocol; (ii) propose SciTX, a scientific-method-inspired pipeline that selects each explanation by validating it through a counterfactual intervention; and (iii) show that this counterfactual-validated design produces explanations that human practitioners consistently prefer over both retrieval-based and direct-MLLM baselines.

2. Related Work

Automatic Model Debugging. Model debugging (Ribeiro et al., 2016a; Koh et al., 2020; Nguyen et al., 2024) is a key step in designing and improving machine learning models. Diagnosing the causes of model errors enables informed decisions and guides future development (Kirichenko et al., 2022; Darcet et al., 2023; Hyeon-Woo et al., 2023; Geirhos et al., 2018).

Various debugging methods have been proposed. Class Activation Map (CAM) (Selvaraju et al., 2017; Zhou et al., 2016; Fernandez, 2020), Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016b), and SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017; Lundberg et al., 2020) share the goal of visually illustrating which features the model attends to. Concept Bottleneck Models (CBMs) (Koh et al., 2020; Yang et al., 2023) instead design the model to predict human-understandable concepts at intermediate layers. In contrast, our work infers error causes and provides natural-language explanations.

Language-based Model Debugging. Natural language is the most human-understandable form of explanation (Doshi-Velez & Kim, 2017), and recent LLM-based interfaces have established it as a primary interface between non-experts and machine learning models. This motivates a growing line of work that explains classifier failures in natural language.

In AI research, previous works (Eyuboglu et al., 2022; Jain et al., 2023; Rezaei et al., 2024; Dunlap et al., 2024; Csurka et al., 2024; Shaham et al., 2024) have adopted natural language. Among them, retrieval-based approaches (Eyuboglu et al., 2022; Jain et al., 2023; Rezaei et al., 2024; Csurka et al., 2024) retrieve appropriate explanations from predefined error sentences. In particular, multimodal embedding models compute similarities between error samples and predefined explanations to perform retrieval. However, these approaches require manual curation of the corpus and cannot describe errors absent from it.

More recently, several methods leverage language models to verbalize classifier failures at the slice or dataset level: LADDER (Ghosh et al., 2025) generates hypothesis sentences about classifier biases via a text-only LLM; and B2T (Kim et al., 2024) extracts bias keywords from caption corpora. These approaches operate at the group level and either rely on a text-only LLM or output non-sentence forms (attributes, keywords). SciTX instead generates a free-form natural-language explanation per sample with an MLLM that takes the image as input.

Large Language Models. Large Language Models (LLMs) (Dubey et al., 2024; Achiam et al., 2023; Bai et al., 2023; Abdin et al., 2024; DeepSeek-AI, 2025) and Multimodal LLMs (MLLMs) (Liu et al., 2023; 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, while MLLMs combine a vision encoder (Radford et al., 2021; Zhai et al., 2023) with an LLM to handle vision and text jointly.

Counterfactual Reasoning. Counterfactual reasoning (Pearl, 2009; Goyal et al., 2019; Yan & Wang, 2023) identifies causal factors of model behavior. We adopt this principle in CEI, intervening on the factor each explanation identifies and measuring the resulting change in the model's prediction. 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.

3. Method

3.1. Task Formulation

Let $f : \mathcal{X} \to \mathcal{Y}$ 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 $\hat{y} = f(x)$, **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. 4.1.2).

3.2. SciTX

To diagnose the root causes of model errors, we propose SciTX (**Sci**entific Method-Inspired **T**extual E**X**planation),

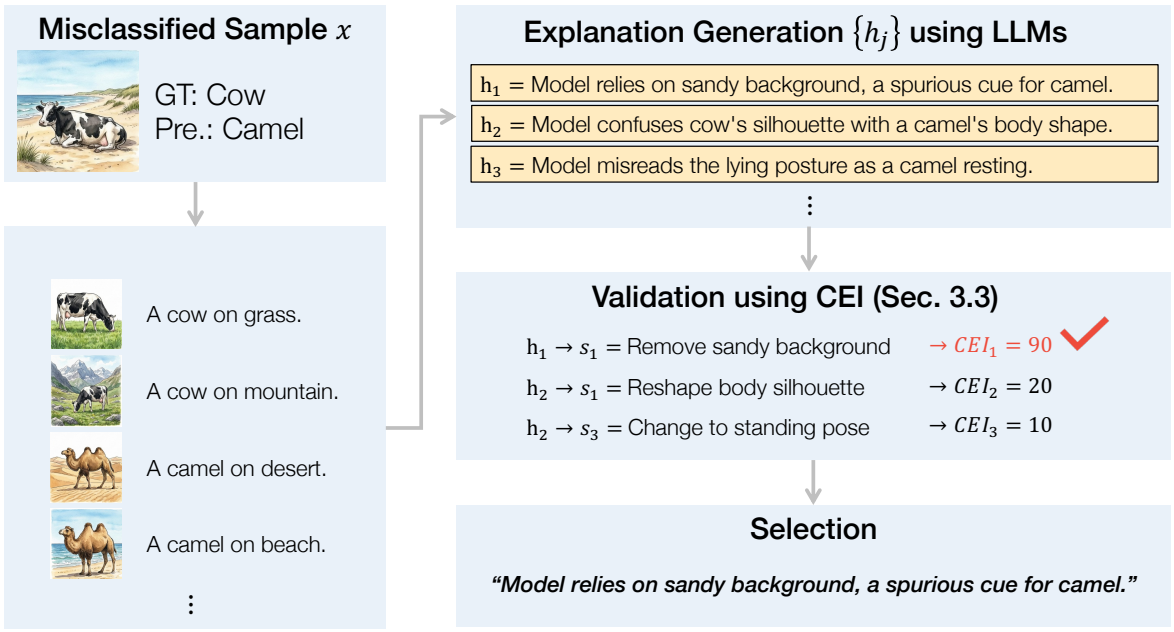


Figure 1. **SciTX for Language Explanation.** SciTX follows a systematic four-stage process: (1) **Observation Retrieval (*Survey*):** Relevant reference samples from both the ground-truth and predicted classes are retrieved to provide contrastive context. (2) **Explanation Generation (*Hypothesis*):** An LLM generates multiple candidate hypotheses $(h_1, h_2, h_3, \cdots)$ by comparing the input query with retrieved observations. (3) **Explanation Validation (*Experiment*):** Each hypothesis is translated into a counterfactual intervention to empirically measure its impact via the CEI metric. (4) **Explanation Selection (*Conclusion*):** The hypothesis that yields the highest CEI score is formally selected as the final, verified explanation.

an instantiation of the explanation generator g defined in Sec. 3.1. SciTX follows the four steps of the scientific method: observation, hypothesis, experiment, and conclusion (Fig. 1).

Observation Retrieval (Survey). To avoid generic or hallucinated explanations, 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. In practice, we represent each observation o_i as a textual caption rather than a raw image. We use captions because of *cross-image information leakage* (Park et al., 2025): when given multiple input images, MLLMs unintentionally mix information across them.

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., 2023; 2024), we assign the MLLM an explicit failure-analyst role so that the hypotheses $\{h_j\}_{j=1}^M$ describe possible failure modes rather than the image content:

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

Each candidate h_j articulates a specific potential failure mode.

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 (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 gain becomes the final explanation e:

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

3.3. Counterfactual Explanation Impact (CEI)

We propose the **Counterfactual Explanation Impact (CEI)**, a counterfactual-inference-based metric (Pearl, 2009;

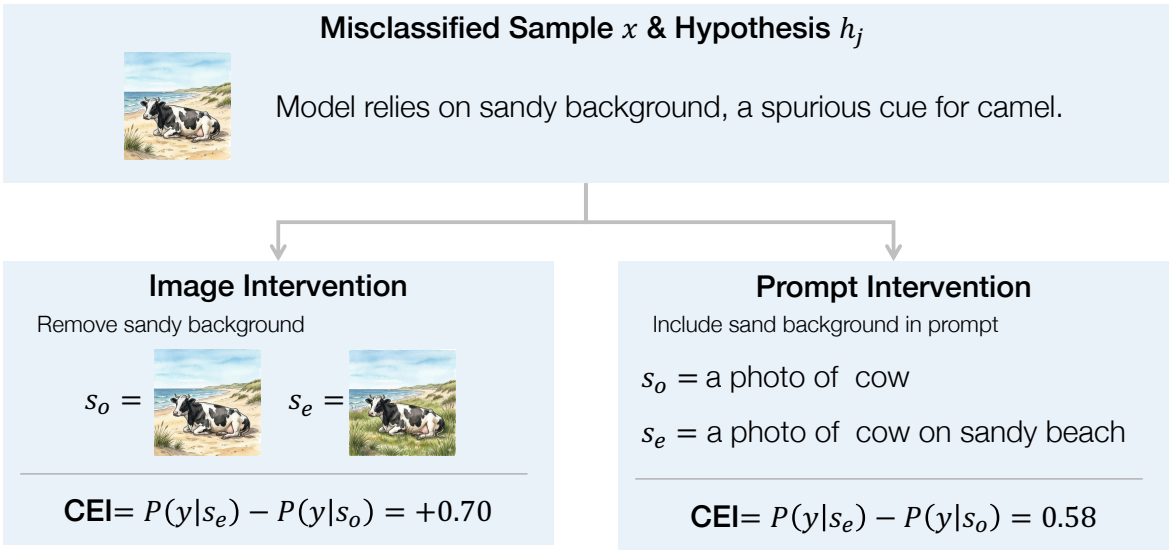


Figure 2. **Counterfactual Explanation Impact (CEI).** CEI quantifies the quality of a natural-language explanation e by measuring how much the model's confidence in the ground-truth label y improves once the cause described in e is neutralized. Given a hypothesis h_j derived from e, we apply two complementary interventions. **Image Intervention** (left): an MLLM derives an editing instruction from h_j, and an image-editing model transforms the original image s_o into a counterfactual image s_e. **Prompt Intervention** (right): an MLLM rewrites the zero-shot prompt s_o into s_e to reflect the identified cue, modulating the classifier's decision boundary without altering pixels. CEI is then computed as the change in the ground-truth probability, $P(y \mid s_e) - P(y \mid s_o)$.

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.

Definition. CEI evaluates an explanation through a three-step counterfactual procedure: (i) *Derivation*—extract an intervention instruction from e; (ii) *Intervention*—apply the instruction to the inference source to construct a counterfactual scenario; and (iii) *Estimation*—measure the resulting change in the ground-truth probability. To make the metric reflect e, we derive the intervention only from e, without giving the MLLM any other information about the ground-truth class.

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

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

where $P(y \mid \cdot)$ is the model's probability assigned to the ground-truth label y. We report CEI in percentage points throughout (i.e., values are multiplied by 100). A positive CEI means that the intervention derived from e shifts the model's prediction toward the ground-truth class; larger values indicate a stronger such shift. We define CEI in two settings, one for standard vision classifiers and one for vision-language models.

We note that CEI is not a strict counterfactual in the do-calculus sense; it measures the probability shift induced by an MLLM-derived intervention, which may alter aspects of the input beyond the factor identified by e. We therefore treat CEI as a directional, intervention-conditional signal and pair it with LLM Match (Sec. 4.1.2) and human evaluation (Sec. 4.3) for cross-validation.

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). The original source is the input image $s_o = \text{Img}$, and its counterfactual is the edited image

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

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 vision-language classifiers such as CLIP, we instead intervene in the textual prompt space, which modulates the classifier's decision boundary without altering image pixels. 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"*).

4. Experiments

4.1. Setting

4.1.1. Models and Dataset

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., 2021; Dosovitskiy et al., 2021) sourced from Timm library (Wightman, 2019).[1]

Generative Models. We use generative models for explanation generation and counterfactual image intervention (Sec. 3.3). We adopt the Qwen-VL series (Bai et al., 2025b;a) for explanation and FLUX.1-Kontext-dev (Labs et al., 2025) for intervention. We pick the Qwen-VL series as a representative high-performing open-weight MLLM family, which keeps the pipeline reproducible without proprietary API access.

Datasets. Our analysis focuses on misclassified samples from three benchmarks: ImageNet-R (Hendrycks et al., 2021), ObjectNet (Barbu et al., 2019), and ImageNet-D (Zhang et al., 2024). ImageNet-R contains stylized renditions such as art and cartoons. ObjectNet contains images with diverse rotations, viewpoints, and complex backgrounds. ImageNet-D contains diffusion-generated samples with controlled background/texture/material shifts. For each benchmark, we sample 400 misclassified images, yielding 1,200 in total.

4.1.2. Evaluation

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* diagnostic. Table 2 verifies that the CEI lead of SciTX persists when baselines are equipped with the same selection.

The benchmarks used in our experiments are robustness datasets, each constructed to vary along a specific axis. For ImageNet-R, the factor is artistic style/rendition; for ObjectNet, the factor is non-canonical viewpoint, rotation, or background; for ImageNet-D, the factor is background, texture, and material. 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: 1.0 (full description), 0.5 (partial description), 0.0 (no description). To reduce single-judge bias, we average scores from three judges: `gpt-5-mini-2025-08-07`, `claude-haiku-4-5`, and `gemini-3-flash-preview`.

4.1.3. Baselines

Error Retrieval. 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).

Grad-CAM. CamMLLM integrates a pixel-attribution method (Grad-CAM) with an MLLM. By leveraging Grad-CAM (Selvaraju et al., 2017) to highlight regions that influence model decisions, the MLLM can generate the reasoning behind misclassifications.

Change of Caption. ChangeMLLM adopts a change captioning approach (Park et al., 2019) to generate descriptions of the differences between a misclassified image and a reference image from either the ground-truth or predicted class. From these visual discrepancies, the MLLM infers the underlying cause of the error.

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 generally outperforms baselines.* 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

[1] CLIP (`openai/clip-vit-base-patch32`), SigLIP (`google/siglip-base-patch16-224`), and 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**

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	$11.967^{(1)}$	$20.397^{(1)}$

SciTX remains substantially higher (e.g., LLM Match 0.619 vs. 0.532, CEI 19.27 vs. 12.00 against ChangeMLLM). The lead is therefore not produced by matched selection. The same module also helps the baselines, suggesting that selecting hypotheses by causal impact is generally useful.

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 baselines do under both editors.

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**

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.

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 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

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**

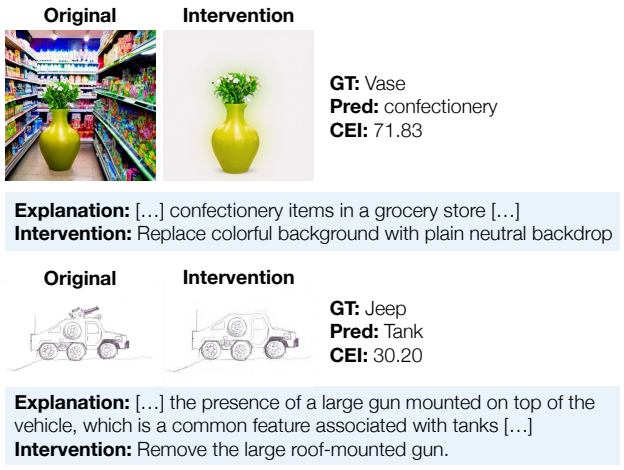


Figure 4. SciTX's explanations translated into counterfactual interventions. CEI measures the change in the ground-truth probability after applying each intervention.

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

Counterfactual Intervention Examples. Fig. 4 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.

Generated Explanations. In the cases shown in Fig. 5, SciTX's explanations refer to multiple visual factors from the retrieved observations, while the baselines describe only one (CamMLLM points to image regions, ChangeMLLM cites a single visual difference). 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.

Pose- and Affect-Aware Reasoning. We identify representative qualitative cases where SciTX invokes high-level cues beyond pixel-level similarity (Fig. 6). 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 on the basis of its violent emotional connotation.

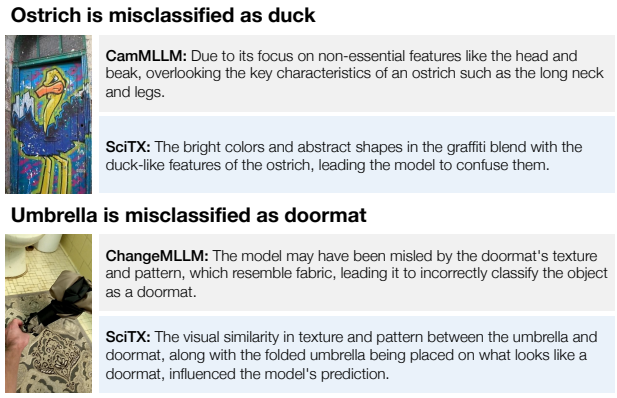


Figure 5. Generated textual explanations. While CamMLLM describes the localization information, ChangeMLLM and ours describe the differences or commonalities between the given misclassified sample and retrieved observations.

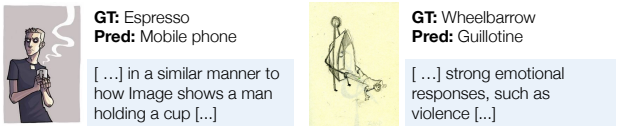


Figure 6. SciTX leveraging high-level cues beyond pixel-level similarity: (left) an espresso illustration confused with a mobile phone via holding pose; (right) a stylized wheelbarrow drawing confused with a guillotine via violent emotional connotation.

4.2.7. Application: Global Explanation

Beyond per-sample diagnosis, 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 distribution-level signal that surfaces 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

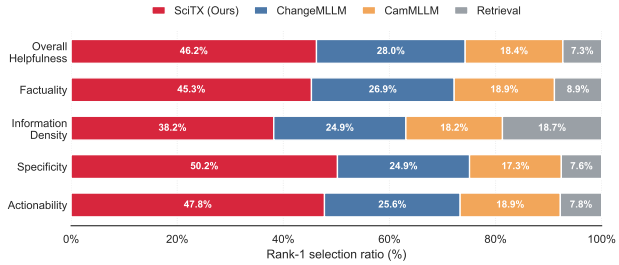
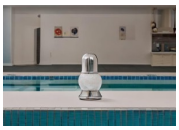


Figure 7. Rank-1 selection ratio of different helpfulness dimensions of human evaluation. We observe that AI practitioners rank our method first more often than the others.

Pug is misclassified as Vase.

The model misclassified the image as a Vase because the black-and-white sketch features two pugs in a heraldic design, which resembles the intricate patterns often associated with decorative vases. The lack of pug-specific details such as the characteristic short muzzle, large eyes, and floppy ears led the model to confuse the image with a Vase due to the similar ornamental style.



Saltshaker is misclassified as Table lamp.

The model likely misclassified the object as a table lamp due to the object's sleek, metallic design and its placement near a poolside, which could be mistaken for a decorative lamp. The object's shape and reflective surface resemble those of a modern table lamp rather than a saltshaker.

Binder is misclassified as Wing.

The model might have been misled by the narrow canyon environment, which resembles the wingspan of a bird, leading to the misclassification of binder as wing. The elongated shape of the blue object being held, combined with the vertical rock walls of the canyon, may have triggered a misinterpretation due to the similar visual cues.

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.

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, and leave broader validation across backbones and datasets to future work.

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

Local	17.52
+ Global	19.42 (+1.9)

4.3. Human Evaluation

4.3.1. 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.

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

Padlock is misclassified as shopping basket.

The model misclassified because the handle on the woven basket resembles the handles on clothes hampers, causing the model to incorrectly associate the image with the wrong class.

Chair is misclassified as bench.

The model misclassified because the context of the images, such as being in front of a brick wall, is common for both chairs and benches, making it difficult for the model to differentiate.

Figure 9. Failure cases of SciTX: (top) an explanation built around a misperceived object; (bottom) reliance on shared scene context over discriminative object shape.

4.3.2. 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—e.g., 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.

5. Conclusion

We introduced NEMO, a task and benchmark for natural-language explanations of vision-model errors. We also proposed SciTX, a generation-based method that follows a four-stage scientific-method-inspired pipeline and validates each candidate explanation with our Counterfactual Explanation Impact (CEI) metric. Across three robustness benchmarks, SciTX outperforms both retrieval-based and MLLM-augmented baselines on LLM Match and CEI, and a human study with AI practitioners ranks SciTX first on all five helpfulness dimensions. As discussed in our failure analysis, SciTX can amplify upstream perception errors and overweight scene-level context. Addressing these limitations and extending the pipeline to broader model-debugging settings are promising directions for future work.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00124, No. RS-2022-II220124, Development of Artificial Intelligence Technology for Self-Improving Competency-Aware Learning Capabilities; No. RS-2026-25518317, Development of AI memory mechanism that reflects human cognitive principles; No. RS-2024-00457882, National AI Research Lab Project); and by the InnoCORE program of the Ministry of Science and ICT (N10250156, KAIST InnoCore LLM).

Impact Statement

This work introduces a task, benchmark, method, and metric for natural-language explanation of vision-classifier errors. As LLM-based interfaces are increasingly used by non-experts to interact with deployed models, explanations that describe plausible failure causes can help users decide whether to trust, escalate, or retrain a model. We see two risks worth flagging. First, MLLM-generated explanations can be confidently wrong; users who treat them as ground truth may misdiagnose model behavior. Our failure analysis (Sec. 4.4) shows that explanations may anchor on spurious cues, and we therefore advocate treating such explanations as a starting point for human inspection rather than a substitute for it. Second, the counterfactual editing pipeline relies on instruction-following image editors that can be repurposed for content manipulation beyond the diagnostic setting; we use them only on misclassified examples within our benchmark and rely on safeguards from their releases. Beyond these specific points, we do not foresee additional societal consequences that warrant separate discussion.

References

Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., Harrison, M., Hewett, R. J., Javaheripi, M., Kauffmann, P., et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., Hui, B., Ji, L., Li, M., Lin, J., Lin, R., Liu, D., Liu, G., Lu, C., Lu, K., Ma, J., Men, R., Ren, X., Ren, X., Tan, C., Tan, S., Tu, J., Wang, P., Wang, S., Wang, W., Wu, S., Xu, B., Xu, J., Yang, A., Yang, H., Yang, J., Yang, S., Yao, Y., Yu, B., Yuan, H., Yuan, Z., Zhang, J., Zhang, X., Zhang, Y., Zhang, Z., Zhou, C., Zhou, J., Zhou, X., and Zhu, T. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., and Zhu, K. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., and Lin, J. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., and Katz, B. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019.

Chiquier, M., Avrech, O., Gandelsman, Y., Feng, B., Bouman, K., and Vondrick, C. Teaching humans subtle differences with diffusion. *arXiv preprint arXiv:2504.08046*, 2025.

Csurka, G., Hayes, T. L., Larlus, D., and Volpi, R. What could go wrong? discovering and describing failure modes in computer vision. *arXiv preprint arXiv:2408.04471*, 2024.

Darcet, T., Oquab, M., Mairal, J., and Bojanowski, P. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

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

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N.

An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Dunlap, L., Zhang, Y., Wang, X., Zhong, R., Darrell, T., Steinhardt, J., Gonzalez, J. E., and Yeung-Levy, S. Describing differences in image sets with natural language. In *CVPR*, 2024.

Eyuboglu, S., Varma, M., Saab, K. K., Delbrouck, J.-B., Lee-Messer, C., Dunnmon, J., Zou, J., and Re, C. Domino: Discovering systematic errors with cross-modal embeddings. In *ICLR*, 2022.

Farid, K., Schrodi, S., Argus, M., and Brox, T. Latent diffusion counterfactual explanations. *arXiv preprint arXiv:2310.06668*, 2023.

Fernandez, F.-G. Torchcam: class activation explorer. https://github.com/frgfm/torch-cam, March 2020.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H., and Wang, H. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1), 2023.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*, 2018.

Ghosh, S., Syed, R., Wang, C., Choudhary, V., Li, B., Poynton, C. B., Visweswaran, S., and Batmanghelich, K. Ladder: Language-driven slice discovery and error rectification in vision classifiers. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

Goyal, Y., Feder, A., Shalit, U., and Kim, B. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165*, 2019.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M. Retrieval augmented language model pre-training. In *International conference on machine learning*, pp. 3929–3938. PMLR, 2020.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., and Gilmer, J. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021.

Hyeon-Woo, N., Yu-Ji, K., Heo, B., Han, D., Oh, S. J., and Oh, T.-H. Scratching visual transformer's back with uniform attention. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5807–5818, 2023.

Jain, S., Lawrence, H., Moitra, A., and Madry, A. Distilling model failures as directions in latent space. In *ICLR*, 2023.

Jeanneret, G., Simon, L., and Jurie, F. Diffusion models for counterfactual explanations. In *Proceedings of the Asian conference on computer vision*, pp. 858–876, 2022.

Keane, M. T., Kenny, E. M., Delaney, E., and Smyth, B. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques. *arXiv preprint arXiv:2103.01035*, 2021.

Kim, Y., Mo, S., Kim, M., Lee, K., Lee, J., and Shin, J. Discovering and mitigating visual biases through keyword explanation. In *CVPR*, 2024.

Kirichenko, P., Izmailov, P., and Wilson, A. G. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. Concept bottleneck models. In *ICML*, 2020.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. *NeurIPS*, 2022.

Labs, B. F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *NeurIPS*, 2020.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *NeurIPS*, 2023.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 4765–4774. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):2522–5839, 2020.

Nguyen, E., Bertram, J., Kortukov, E., Song, J. Y., and Oh, S. J. Towards user-focused research in training data attribution for human-centered explainable ai. *arXiv preprint arXiv:2409.16978*, 2024.

Park, D. H., Darrell, T., and Rohrbach, A. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4624–4633, 2019.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, 2023.

Park, Y., Lee, M., Chun, S., and Choe, J. Mitigating cross-image information leakage in lvlms for multi-image tasks. *arXiv preprint arXiv:2508.13744*, 2025.

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

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Rezaei, K., Saberi, M., Moayeri, M., and Feizi, S. PRIME: Prioritizing interpretability in failure mode extraction. In *ICLR*, 2024.

Ribeiro, M. T., Singh, S., and Guestrin, C. "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*, 2016a.

Ribeiro, M. T., Singh, S., and Guestrin, C. " 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*, pp. 1135–1144, 2016b.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

Shaham, T. R., Schwettmann, S., Wang, F., Rajaram, A., Hernandez, E., Andreas, J., and Torralba, A. A multimodal automated interpretability agent. In *ICML*, 2024.

Steiner, A., Kolesnikov, A., , Zhai, X., Wightman, R., Uszkoreit, J., and Beyer, L. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Wang, Z., Mao, S., Wu, W., Ge, T., Wei, F., and Ji, H. 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)*, 2024.

Wang, Z. M., Peng, Z., Que, H., Liu, J., Zhou, W., Wu, Y., Guo, H., Gan, R., Ni, Z., Yang, J., Zhang, M., Zhang, Z., Ouyang, W., Xu, K., Huang, S. W., Fu, J., and Peng, J. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv: 2310.00746*, 2023.

Warren, G., Byrne, R. M., and Keane, M. T. Categorical and continuous features in counterfactual explanations of ai systems. *ACM Transactions on Interactive Intelligent Systems*, 14(4):1–37, 2024.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. Emergent abilities of large language models. *TMLR*, 2022.

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

Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., and Liu, Z. Qwen-image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

Yan, J. and Wang, H. Self-interpretable time series prediction with counterfactual explanations. In *International Conference on Machine Learning*, pp. 39110–39125. PMLR, 2023.

Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., and Yatskar, M. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *CVPR*, 2023.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Zhang, C., Pan, F., Kim, J., Kweon, I. S., and Mao, C. Imagenet-d: Benchmarking neural network robustness on diffusion synthetic object. In *CVPR*, 2024.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. Learning deep features for discriminative localization. In *CVPR*, 2016.

Contents

A. Pseudo Algorithms and Prompts

A.1. SciTX

Our algorithm is composed of four stages, namely observation retrieval, explanation generation, validation, and selection. Pseudo algorithm is shown in Algorithm 1.

Algorithm 1 SciTX Pseudo Code

Require: Misclassified image x, Reference database $\mathcal{D}$, Discriminator f, Labels $\{y, \hat{y}\}$
Ensure: Best explanation e

Stage 1: Observation Retrieval
$\mathcal{D}_{filtered} \leftarrow \{d \in \mathcal{D} \mid f(d) = \text{label}(d) \wedge (\text{label}(d) \in \{y, \hat{y}\})\}$
$\{o_i\}_{i=1}^{K} \leftarrow \text{Retrieve}(x, \mathcal{D}_{filtered}, K)$

Stage 2: Explanation Generation
$\{c_i\}_{i=1}^{K} \leftarrow \text{Caption}(\{o_i\}_{i=1}^{K})$
$\text{Prompt} \leftarrow \text{Prompt}_{\text{hypo}} + x + \{c_i\}_{i=1}^{K}$
$\{h_j\}_{j=1}^{M} \leftarrow \text{MLLM}(\text{Prompt})$

Stage 3 & 4: Validation and Selection
$best\_score \leftarrow -\infty$
for each $h_j \in \{h_j\}_{j=1}^{M}$ **do**
  $s_j \leftarrow \text{Intervention}(h_j)$
  $current\_score \leftarrow \text{CEI}(h_j)$
  if $current\_score > best\_score$ **then**
    $best\_score \leftarrow current\_score$
    $e \leftarrow h_j$
  end if
end for
return e =0

The prompt of ours 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

ImageNet-R ImageNet-R (Hendrycks et al., 2021) is the robustness benchmark. The images in this dataset are renditions of ImageNet classes. This dataset is composed of 200 classes and 30,000 images.

ImageNet-D ImageNet-D (Zhang et al., 2024) is the robustness benchmark under MIT License. The images in this dataset are generated using the diffusion models, with variations in background, texture, and material properties.

ObjectNet ObjectNet (Barbu et al., 2019) is the robustness benchmark. The images in this dataset are taken in the difficult setting, *e.g.*, random object backgrounds, rotations, and imaging viewpoints.

We sample 400 misclassified samples from ImageNet-R, ImageNet-D, and ObjectNet (1,200 in total). The remaining samples and a 100-class subset of ImageNet (Deng et al., 2009) serve as the observation source for retrieval.

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 leverage 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., 2024), 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.

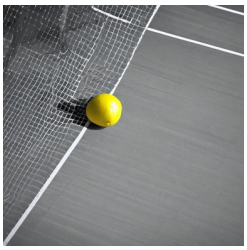
Ground Truth: Binder
Prediction: Envelope

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.

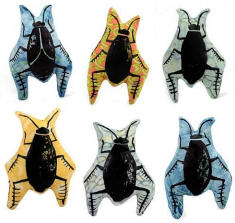
Ground Truth: Goose
Prediction: Bathtub

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.

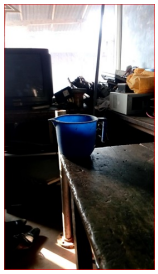
Ground Truth: Toaster
Prediction: Portable heater

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.

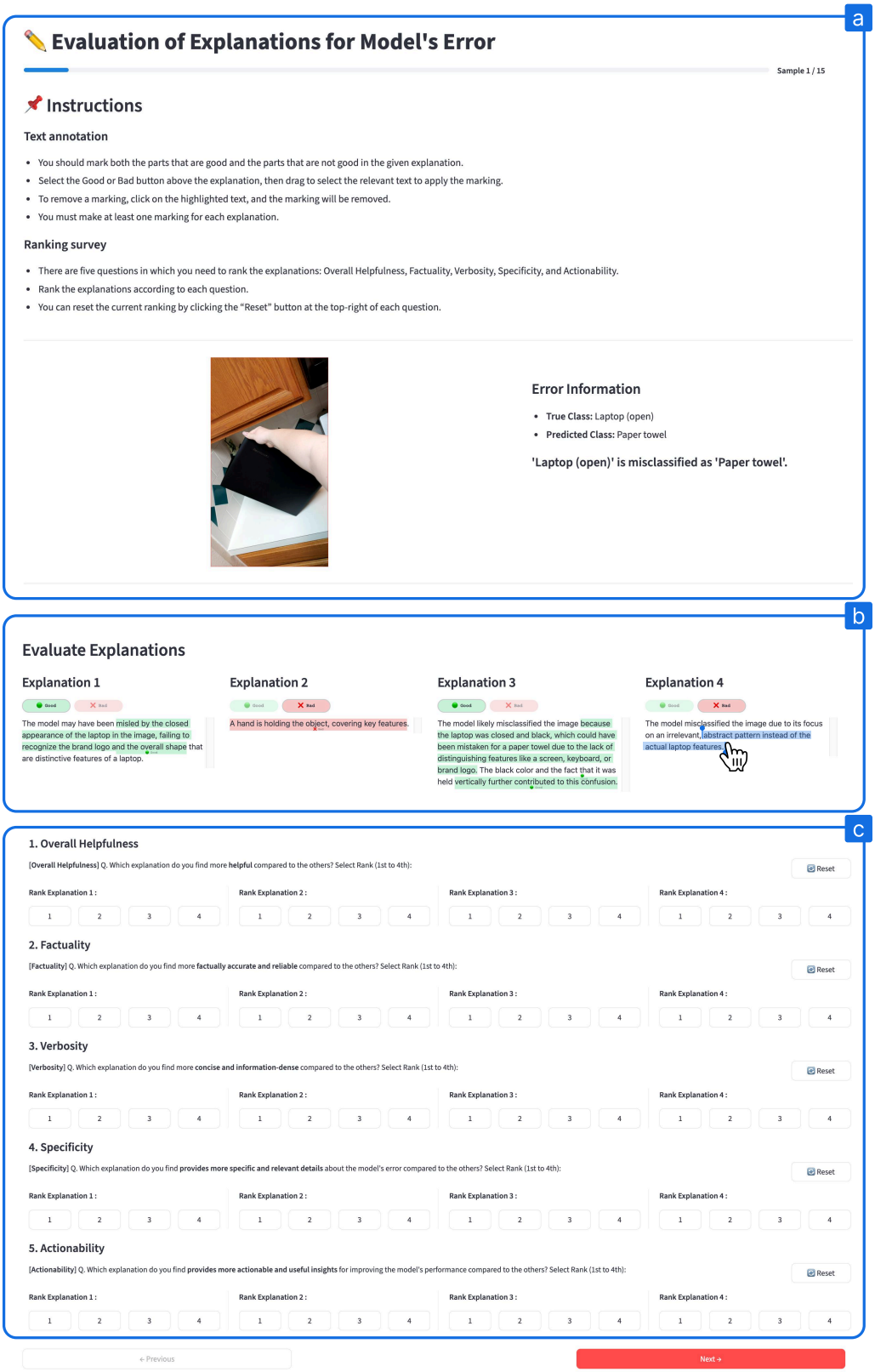


Figure 13. Interface for human evaluation. (a) The task instruction and an example showing the ground-truth and predicted classes. (b) An interface that allows participants to directly annotate helpful and unhelpful parts of the explanation. (c) An interface for ranking explanations across evaluation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

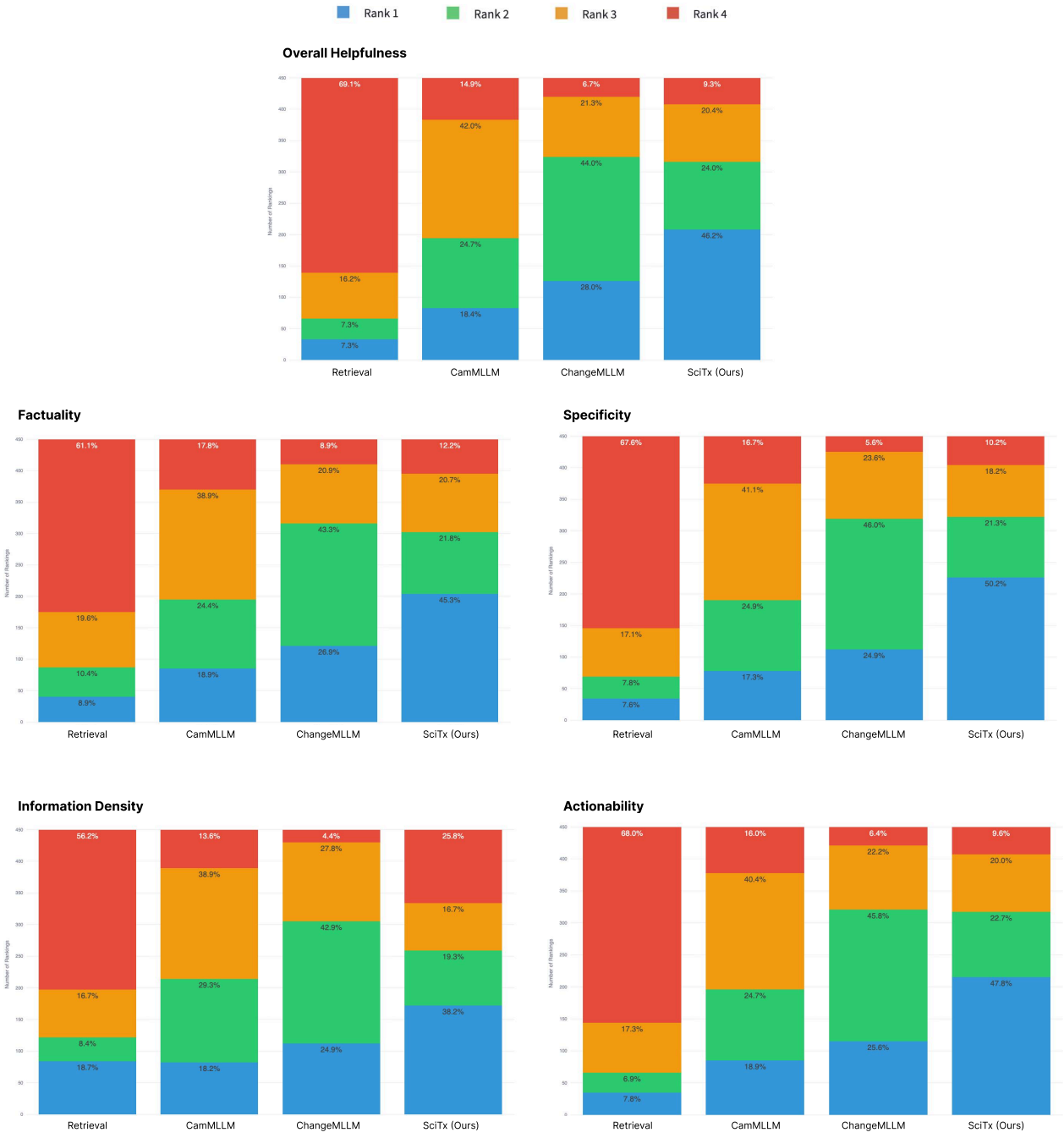


Figure 14. Detailed human evaluation results for five categories