SafeFix: Targeted Model Repair via Controlled Image Generation

Anonymous authors
Paper under double-blind review

Abstract

Deep learning models for visual recognition often exhibit systematic errors due to underrepresented semantic subpopulations. While existing debugging frameworks can identify these failure slices, effectively repairing them remains difficult. Current solutions often rely on manually designed prompts to generate synthetic images—an approach that introduces distribution shift and semantic errors, often resulting in new bugs. To address these issues, we introduce SafeFix, a framework for distribution-consistent model repair via controlled generation that employs a diffusion model to generate semantically faithful images that modify only specific failure attributes while preserving the underlying data distribution. To ensure the reliability of the repair data, we implement a verification mechanism using a large vision–language model (LVLM) to enforce semantic consistency and label preservation. By retraining models on the synthetic data, we significantly reduce errors in rare cases and improve overall performance. Our experiments show that SafeFix achieves superior robustness by maintaining high precision in attribute editing without introducing additional bugs.

1 Introduction

Despite strong performance on standard benchmarks, computer vision models often fail on rare semantic subpopulations (Barbu et al., 2019; Gao et al., 2023; Leclerc et al., 2022). Such errors usually arise from dataset bias: some attribute combinations (for example, individuals with red hair color) are rarely represented in the training data (Buolamwini & Gebru, 2018). Finding and fixing these bug slices is important for deploying AI systems in safety-critical applications.

Recent research has focused on identifying these failures using interpretable debugging pipelines (Gao et al., 2023; Chen et al., 2023; Singla et al., 2024). For instance, HiBug (Chen et al., 2023) uses vision–language models to discover failure slices based on shared visual attributes. However, the subsequent repair process remains a major bottleneck. Existing repair strategies generally fall into two categories: retrieval-based and generation-based. Retrieval-based methods (Gao et al., 2023; Singla et al., 2024) collect samples from external datasets, which often introduces a domain gap. Generation-based methods, such as those suggested by HiBug, use text-to-image models to synthesize training data. However, standard generative augmentation suffers from two critical flaws: (1) **Semantic Drift**: Generative models often fail to preserve attributes unrelated to the intended edit, and (2) **Distribution Mismatch**: Synthetic images may not align with the original data distribution, leading the model to learn generative artifacts rather than the intended semantic concepts.

To overcome these challenges, we propose SafeFix, a targeted model repair method designed to correct failures caused by underrepresented semantic subpopulations through controlled image generation—without degrading overall performance or introducing new errors. Compared to AdaVision (Gao et al., 2023) and DCD (Singla et al., 2024), which retrieve samples from another dataset, our method generates new examples grounded in the training set to ensure distributional consistency. Unlike HiBug (Chen et al., 2023), which simply composes failure attributes into language prompts, SafeFix anchors generation on real instances in the training set to preserve unrelated attributes and avoid unintended changes. While the latent-space-based filtering method (Jain et al., 2023) explores repair via diffusion models, their approach relies solely on caption-based

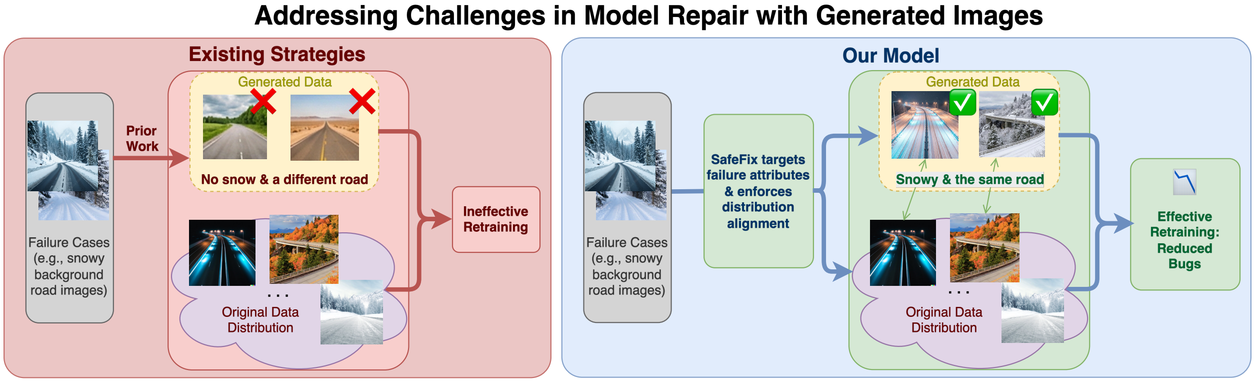

Addressing Challenges in Model Repair with Generated Images
Existing Strategies
Failure Cases (e.g., snowy background road images)
Prior Work
Generated Data
No snow & a different road
Original Data Distribution
. . .
Ineffective Retraining
Our Model
Failure Cases (e.g., snowy background road images)
SafeFix targets failure attributes & enforces distribution alignment
Generated Data
Snowy & the same road
Original Data Distribution
. . .
Effective Retraining: Reduced Bugs

Figure 1: SafeFix addresses model repair challenges by generating images that target specific failure attributes, such as snowy backgrounds. While prior work (Zhang et al., 2024; Chen et al., 2023) may produce data that are insufficiently represented in the source dataset, such methods often miss true failure attributes and cause distribution shifts. In contrast, SafeFix generates images under snowy conditions that maintain the same road geometry to enforce distribution alignment, keeping generated samples consistent with the original data distribution. This approach reduces bugs by improving training coverage of underrepresented semantic subpopulations where model failures stem from insufficient data. It ensures that the augmented dataset effectively repairs the model without introducing new errors.

prompting and does not guarantee the semantic accuracy of critical attributes. In contrast, SafeFix mitigates the unreliability of generative models in rendering sensitive attributes—such as skin tone or hair color—by further filtering out semantically inconsistent images using a large vision–language model (LVLM) (Bai et al., 2025; Liu et al., 2023), and the LVLM is verified by human audit. This system ensures that the final outputs faithfully reflect the intended attribute change (e.g., generating a sad expression on a darker-skinned woman with red hair, whereas standard diffusion models often mistakenly produce a darker-skinned man with black hair instead), enabling precise and robust repair of underrepresented semantic subpopulation failures.

As shown in Figure 1, SafeFix directly addresses two major challenges in existing model repair strategies: **1)** ensuring the generated images reflect the true failure attributes, and **2)** aligning them with the original data distribution. By retraining models on this augmentation set, we reduce errors associated with underrepresented semantic subpopulations, where model failures stem from insufficient training coverage of specific attributes. The contributions of this work are as follows:

- We formulate a targeted model repair pipeline, SafeFix, which leverages conditional text-to-image generation and LVLM-based filtering to synthesize high-quality, attribute-faithful data for correcting model failures arising from underrepresented subpopulations.
- We introduce a verification mechanism that leverages LVLMs to mitigate the inherent unreliability of generative models, ensuring that the synthesized repair data is both semantically accurate and remains aligned with the original dataset distribution.
- We demonstrate that our approach significantly improves model performance on rare-case failure slices across multiple architectures and datasets, leading to higher accuracy and robustness in underrepresented scenarios without introducing new errors.

2 Related Work

Failure Pattern Discovery. HiBug (Chen et al., 2023) proposes a pioneering pipeline that identifies interpretable failure cases in vision models by clustering semantically meaningful attributes, revealing both rare categories and spurious correlations. HiBug2 (Chen et al., 2025) extends this approach with more efficient error-slice discovery and a closed-loop debugging mechanism, improving coherence and coverage of discovered model bugs. Several follow-up works attempt to strengthen interpretability in model debugging (Adebayo et al., 2020; 2022), though often limited by static failure patterns or inadequate visual grounding. Beyond HiBug, MODE (Vendrow et al., 2023) introduces a state-differential analysis framework that locates internal model faults and proposes data-driven remedies. TCAV (Kim et al., 2018) further enriches interpretability by testing the model's sensitivity to high-level concepts and by correcting spurious activations at the concept level. 3DB (Leclerc et al., 2022) complements these efforts by constructing structured attribute spaces over failure modes, allowing discovery of underrepresented attributes through unsupervised analysis of visual model errors.

Targeted Synthetic Augmentation via Diffusion Models. Diffusion models such as Stable Diffusion (Rombach et al., 2022) and classifier-free guidance (Ho & Salimans, 2022) enable controllable, semantically faithful image synthesis. These models have proven useful in debugging (Casper et al., 2022; Fang et al., 2024; Huang et al., 2024), conditional text-to-image visualization (Augustin et al., 2022; Boreiko et al., 2022), and training data augmentation (Trabucco et al., 2023; Dunlap et al., 2023), especially in few-shot and fine-grained recognition tasks. Recent work like DiGA (Zhang et al., 2024) shows how editing spurious attributes while preserving class semantics can mitigate bias without requiring new annotations. These advances highlight how targeted generation can shape training distributions to address model weaknesses. Building on these insights, our method forms a debugging pipeline that not only identifies failure cases but also synthesizes and integrates targeted images to improve model performance. Compared to prior augmentation or error discovery pipelines (Fang et al., 2024; Huang et al., 2024; Chen et al., 2023), our approach is more generalizable and less reliant on predefined attribute sets. We draw inspiration from efforts like StylizedImageNet (Geirhos et al., 2019), debiasing pipelines (Jin & Rinard, 2021), and adaptive augmentation methods (Mikołajczyk-Bareła et al., 2023; Zhao et al., 2022; Wang et al., 2024b), but focus on semantically controllable generation tailored to discovered bugs.

Multimodal Filtering via LVLMs. LVLMs like Flamingo (Alayrac et al., 2022) have demonstrated strong capabilities in semantically grounding visual concepts. Recent studies show that using LVLMs as filters to select high-quality image-text pairs enhanced dataset quality for downstream tasks (Wang et al., 2024a; Li et al., 2024). These approaches outperform traditional methods like CLIP-based filtering by providing fine-grained, attribute-aware analysis of generated samples. In line with these trends, we adopt an LVLM as an automated filtering component, grounding our approach in established methods that leverage LVLMs for semantic validation of generated data.

Targeted Repair for Rare-Case Bugs. Recent work explores how underrepresented subpopulations induce systematic errors in vision models and how targeted interventions can mitigate such failure modes. DOMINO (Eyuboglu et al., 2022) discovers coherent failure slices by clustering model errors in a cross-modal embedding space, enabling automated identification of rare-case bugs without manual slice definitions. REAL (Parashar et al., 2024) focuses on rare visual concepts that are neglected in large-scale vision–language datasets, augmenting them by retrieving semantically similar examples and fine-tuning lightweight classifiers on the retrieved subsets to improve model robustness on these rare categories. This work reflects a broader shift toward targeted model repair using interpretable diagnostics and subpopulation-aware interventions, aligning with our controlled synthesis and refinement approach.

3 Background

Attribute-based Model Debugging. Let $x \in \mathcal{X}$ be an input image with its corresponding ground-truth label $y(x) \in \mathcal{Y}$. A computer vision model $f_\theta \colon \mathcal{X} \to \mathcal{Y}$ produces a prediction $f_\theta(x)$ for each input image x. Denote the training, validation, and test splits by $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{val}}$, and $\mathcal{D}_{\text{test}}$, respectively. The full dataset is then given by $\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{val}} \cup \mathcal{D}_{\text{test}}$. The overall validation accuracy is computed as:

$$\mathrm{Acc}(\mathcal{D}_{\text{val}}) = \frac{1}{|\mathcal{D}_{\text{val}}|} \sum_{(x,y) \in \mathcal{D}_{\text{val}}} \mathbf{1}\{f_\theta(x) = y(x)\}. \tag{1}$$

In attribute-based model debugging (Chen et al., 2023), each image is assigned several attributes that represent different subpopulations. In particular, let $\mathcal{A} = \{a_1, \dots, a_m\}$ be a set of attribute functions, where

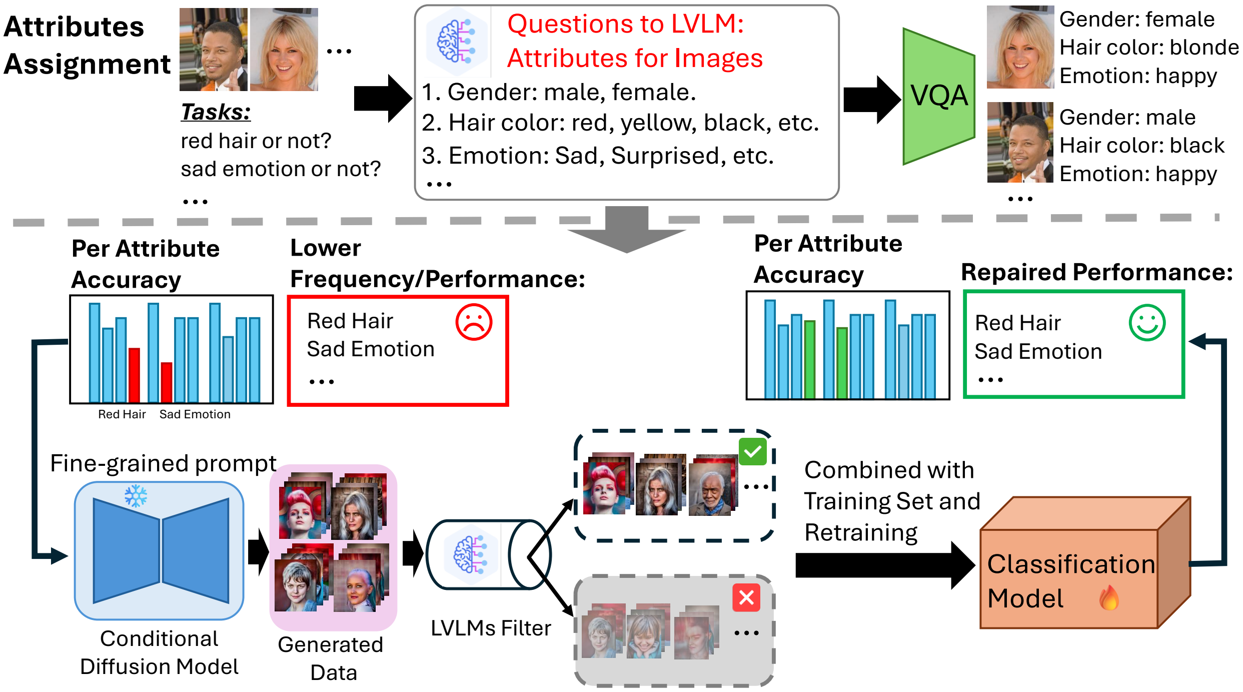


Figure 2: **Overview of SafeFix**. We propose a targeted model repair pipeline that identifies rare-case failures, generates attribute-specific synthetic images using a conditional diffusion model, filters them via a large vision–language model, and retrains the model to improve accuracy and fix rare-case bugs.

each $a_i\colon \mathcal{X} \to \mathcal{V}_i$ maps an image x to a discrete value $v \in \mathcal{V}_i$ and $\mathcal{V}_i$ represents the set of all possible values for attribute i. A slice S is defined as the set of images that satisfy a conjunction of attribute-value assignments:

$$S = \{x \mid a_{j_1}(x) = v_{j_1}, \dots, a_{j_k}(x) = v_{j_k}\}, \tag{2}$$

where $\mathcal{J} = \{j_1, \dots, j_k\} \subseteq [m]$ is the index set of attributes involved in the slice S. The slice accuracy is

$$\mathrm{Acc}(S) = \frac{1}{|S|} \sum_{x \in S} \mathbf{1}\{f_\theta(x) = y(x)\}. \tag{3}$$

Underrepresented Semantic Subpopulations. Rare-case bugs, often caused by underrepresented semantic subpopulations (Eyuboglu et al., 2022), are attribute-based failure modes for which the model's error rate on validation samples matching a target description significantly exceeds its overall error and those samples appear infrequently in the training dataset. We say a slice S_r is a *rare-case slice* if it constitutes less than a fraction ρ of the training data, i.e.,

$$|S_r| < \rho\,|\mathcal{D}_{\text{train}}|, \tag{4}$$

where ρ is a *rare threshold* (e.g., 0.05). We flag a candidate slice S_r as a bug slice S_e if

$$\mathrm{Acc}(S_e) < \mathrm{Acc}(\mathcal{D}_{\text{val}}) - \epsilon. \tag{5}$$

where ϵ is an *accuracy difference threshold*. Thus, the model shows significantly lower accuracy on the bug slice—which represents an underrepresented semantic subpopulation—relative to its average validation accuracy. A bug slice can be converted to a human-readable bug description (e.g., "people with red hair smiling tend to have low accuracy on the 'wearing lipstick' classification task").

4 SafeFix

To address rare-case bugs caused by underrepresented semantic subpopulations, we propose a targeted model repair strategy that leverages controlled image generation and semantic filtering enabled by a large vision–language model. Our goal is to generate synthetic images that accurately represent underrepresented semantic subpopulations while ensuring these images remain aligned with the training distribution. The overall workflow is summarized in Figure 2.

We begin by generating visually controlled images using a text-to-image diffusion model, *Stable Diffusion*, guided by a structure-conditioned controller, ControlNet (Zhang et al., 2023). Instead of relying solely on language prompts, we generate images conditioned on training data while modifying specific attributes to reflect failure slice semantics. Next, we employ an LVLM to automatically verify whether each generated image accurately reflects the intended attributes. This filtering step ensures that only semantically faithful samples are retained. Finally, we augment the original training dataset with the validated images and retrain the model. SafeFix enhances performance on error-prone regions without compromising overall accuracy or introducing new bugs.

4.1 Model Diagnosis for Identifying Rare-Case Bugs

We begin by training a standard computer vision model f_θ on the original training set $\mathcal{D}_{\text{train}}$. After obtaining predictions on the validation set $\mathcal{D}_{\text{val}}$, similar to HiBug, we use a large vision–language model (LVLM; e.g., GPT-4 with vision (OpenAI, 2023)) to propose candidate attributes and a VQA model (e.g., BLIP (Li et al., 2022)) to assign attribute values $a_i(x)$ to each image. We extract a set of rare-case bug slices $\{S_e\}$, each defined by a conjunction of attribute conditions that exhibit both high error rates and low coverage in the dataset. Specifically, to identify underrepresented and error-prone subpopulations, we analyze each attribute a_i and its associated values $v \in \mathcal{V}_i$. For each value v, we examine the slice $S = \{x \mid a_i(x) = v\}$ and compute two quantities: (1) its proportion in the training set, and (2) the model's accuracy on the corresponding validation samples.

We flag the slice S as a rare-case bug if:

- The training support $|S \cap \mathcal{D}_{\text{train}}|$ is below the threshold $\rho \cdot |\mathcal{D}_{\text{train}}|$.
- The validation accuracy $\text{Acc}(S)$ is significantly below the overall validation accuracy $\text{Acc}(\mathcal{D}_{\text{val}})$ (by a margin ϵ).

For example, consider the attribute *hair color*. Suppose only 3% of the dataset have *red hair*, and the model performs poorly on this group (e.g., 60% accuracy versus 85% overall). This makes *red hair* a rare-case bug. In contrast, if *yellow hair* is infrequent but achieves high accuracy, it is not considered a bug. This analysis helps isolate specific attribute values that contribute to systematic model failures.

4.2 Targeted Generation with Conditional Diffusion Models (CDMs)

Next, we aim to produce attribute-preserving edits on each original image x, focusing on targeted attributes identified in problematic slices from the previous diagnostic stage. Specifically, we generate visually controlled synthetic images x' using a text-to-image diffusion model, *Stable Diffusion*, guided by the structure-conditioned controller ControlNet (Zhang et al., 2023).

To determine which attribute–value pairs to edit, we first identify rare-case slices S_e by computing slice support and validation accuracy across attribute conjunctions, as defined in Eq. equation 4 and Eq. equation 5. Each S_e contains one or more attribute–value pairs $\{(a_j, v'_j)\}_{j\in\mathcal{J}}$ that are both infrequent and underperforming. These attributes yield significantly lower accuracy than average on the validation set. We rank all such attributes by validation error and select the top-k attributes for augmentation. For each original image x that does not satisfy all conditions in S_e, we construct an edited variant x' by modifying the selected attributes $\{a_j\}_{j\in\mathcal{J}}$ to match the error-prone configuration defined by S_e, while preserving all other visual characteristics and keeping the original label unchanged, i.e., $y(x') = y(x)$.

Given an original image x in the training set with attribute assignment $\{(a_i(x) = v_i)\}_{i=1}^{m}$, we construct a modified image x' by replacing a subset $\{(a_j, v_j)\}_{j\in\mathcal{J}}$ with $\{(a_j, v'_j)\}_{j\in\mathcal{J}}$, where $\mathcal{J} \subseteq [m]$ indexes attributes satisfying the slice condition S_e. In practice, this subset is small (i.e., $|\mathcal{J}| \ll m$), and the remaining attribute assignments $\{(a_i, v_i)\}_{i \in [m] \setminus \mathcal{J}}$ are left unchanged. For example, suppose x has attribute assignment (*black hair*, *happy emotion*) and label "not wearing lipstick." If both attributes are part of the rare-case slice S_e, then the synthetic variant x' is generated with attributes (*red hair*, *sad emotion*) while retaining the same label "not wearing lipstick." This attribute-preserving image generation aligns with fairness-driven augmentation in attribute classification, where rare attributes (e.g., hair color or emotion) are modified while the primary label is held fixed. By retraining the model on these synthetically augmented samples $\{x'\}$, we encourage the model to correct rare-case bugs without introducing new bugs.

4.3 Filtering via Large Vision–Language Models

We found that synthetic images generated by the Conditional Diffusion Model (CDM) can sometimes fail to accurately reflect the intended attribute modifications. To address this, we employ large vision–language models (LVLMs), Qwen2.5-VL-7B (Bai et al., 2025) and LLaVA-v1.5-7B (Liu et al., 2023), to automatically verify that each generated image correctly exhibits the desired attributes and retains the original label.

Specifically, for each synthetic image x' generated to satisfy a bug slice S_e, which contains the error attribute-value pairs responsible for bugs, we iterate over each edited pair (a_j, v'_j) for $j = 1, \ldots, k$, and query the LVLM with:

$$\texttt{"Does the object have attribute } a_j \texttt{ equal to } v'_j\texttt{?"}$$
$$\texttt{"Is the object in this picture labeled } y(x)\texttt{?"}$$

We retain only those images for which the LVLM answers "yes" to all queries and confirms that the label matches the original ground truth $y(x)$. Thus, after the LVLM filtering, the generated image x' satisfies the desired attributes and preserves the original label, i.e., $y(x') = y(x)$. These validated samples are then added to $\mathcal{D}_{\text{train}}$ for model retraining. We conduct human-audit experiments to verify that the LVLMs can reliably filter out low-quality generated samples.

4.4 Combining Generated Images with the Original Dataset and Retraining

To repair the rare-case bugs while maintaining the model's performance on the overall training distribution (Lee et al., 2024), we augment the training set by adding the validated synthetic images $\{x'\}$, yielding an updated training set $\mathcal{D}'_{\text{train}} = \mathcal{D}_{\text{train}} \cup \{x'\}$. We then retrain the vision model f_θ on $\mathcal{D}'_{\text{train}}$ and evaluate its performance on $\mathcal{D}_{\text{val}}$. We report both the overall accuracy improvement and the reduction in failure rates (fix rate) on critical rare-attribute slices S_e, before and after augmentation.

5 Results

5.1 Experimental Setup

Datasets. We evaluate our method on two classification tasks that exhibit attribute-based failure modes and we use an 8:1:1 split for training, validation, and testing, respectively.

Lipstick-wearing classification. We use the CelebA dataset (Liu et al., 2015) and follow the same data split protocol as (Chen et al., 2023) (80,000, 10,000, 10,000 for train/val/test). The task is to predict whether a person is wearing lipstick, a label known to be correlated with other attributes such as gender and hair color. In the following experiments, we refer to this dataset simply as *CelebA*.

ImageNet-10 classification. We construct a 10-class subset of ImageNet (Deng et al., 2009) containing the following categories: `backpack`, `barber chair`, `coffee mug`, `desk`, `electric guitar`, `park bench`, `pitcher`, `purse`, `rocking chair`, and `water bottle`. Each class contains 1,300 images. This subset is selected to study classification failures related to visual attributes such as texture and color. In the following experiments, we refer to this dataset simply as *ImageNet10*.

Baselines. We compare our method with six recent baselines that are either data augmentation or use generative augmentation strategies:

- **Data Augmentation**. We apply on-the-fly augmentations to each training image, including random resizing, flipping, color jittering, grayscale conversion, erasing, and normalization, while keeping the dataset size fixed. At test time, images are resized, center-cropped, and normalized deterministically.
- **DiGA (Zhang et al., 2024)**. This method utilizes a two-stage framework to automatically detect spurious attributes and modify them with varying degrees of intensity. It keeps the target attribute constant while diversifying other features to mitigate the effect of spurious correlations on model performance.
- **DA-CDM (Fang et al., 2024)** is a data augmentation method for object detection. It uses a controllable diffusion model guided by visual priors from original images, which enables direct reuse of existing bounding box annotations. It then applies a category-calibrated CLIP score to filter generated data and ensure high-quality, text-aligned samples.
- **Mask-ControlNet (Huang et al., 2024)** is a text-guided image generation pipeline that uses ControlNet with facial occlusion masks to synthesize diverse face images under specific occlusions as a data augmentation method to improve model robustness.
- **HiBug_Class (Chen et al., 2023).** A **Class**-level method that augments training data with synthetic images. For each class, it uses a diffusion model with a prompt:

  *"A photo of (*label)."*
- **HiBug_Task (Chen et al., 2023).** A **Task**-level variant of HiBug_Class that targets failure-prone attributes. It selects attribute slices with the highest validation error and generates prompts to guide a diffusion model:

  - *CelebA*: *"A photo of a {gender} {beard clause} {makeup clause} {lipstick clause} (*label), with {hair} hair and {skin} skin, looking {emotion}, appearing {age}."*

  - *ImageNet10*: *"A photo of a {color} {class name} (*label) with {texture} texture, located {object position}, appearing {object size}, in a {background}, under {lighting} lighting, during {time}, from a {perspective} perspective."*

  Note: Unless otherwise specified, "HiBug" refers to this optimized "HiBug_Task" variant in the paper. We do not compare with HiBug2 (Chen et al., 2025), which is a data selection method and differs from the synthetic generation strategy.

Metrics. We focus on **targeted improvements for rare slices**, which are critical for fairness and safety, while maximizing overall classification accuracy. Following prior work (Lai et al., 2023), we report the ***Fix Rate*** **(FR)**, defined as

$$FR = \frac{Acc_{\text{after}} - Acc_{\text{before}}}{1 - Acc_{\text{before}}}, \tag{6}$$

which measures the fraction of previously misclassified samples that are corrected. Here, Acc_{after} denotes the accuracy obtained after applying the proposed method, and Acc_{before} represents the baseline accuracy from standard training using the vision model.

Implementation Details. All experiments use an NVIDIA A100 GPU. We take the CelebA "wearing lipstick" classification task as an example. We evaluate three backbone architectures for this classification task: ResNet-18 (He et al., 2016), ViT-B/16 (Dosovitskiy et al., 2020), and CLIP (ViT-B/32) (Radford et al., 2021), all of which are initialized with random weights. This design allows us to isolate the effectiveness of SafeFix from pre-existing biases inherent in pre-trained weights, such as the latent knowledge of ImageNet. Moreover, retraining on the original large-scale pre-training datasets is often infeasible as they may be unknown or inaccessible. All vision models are trained using cross-entropy loss. For rare-case bug discovery, we set the rarity threshold $\rho = 0.05$ and the accuracy difference threshold $\epsilon = 0.03$. For synthetic augmentation, we use ControlNet (Zhang et al., 2023), a CDM based on Stable Diffusion 1.5 (Rombach et al., 2022) and conditioned on soft HED boundaries (Xie & Tu, 2015), with 30 DDIM inference steps. Soft HED boundaries preserve structural details, making this approach suitable for attribute-preserving edits like recoloring and stylizing. To filter generated images, we employ the large vision–language models (LVLMs) Qwen2.5-VL-7B (Bai et al.,

Table 1: Fix Rate (FR%) on CelebA for varying numbers of added images, models, and methods. Ours (L) and Ours (Q) denote LLaVA-7B and Qwen-7B as the large vision–language model filters, respectively. The highest FR in each column is marked in bold.

Method	**ResNet** (base acc: 90.57%)			**ViT** (base acc: 85.02%)			**CLIP** (base acc: 88.32%)		
	1k Images	5k Images	10k Images	1k Images	5k Images	10k Images	1k Images	5k Images	10k Images
Data Augmentation	2.45	1.82	3.14	4.27	2.56	3.89	1.12	4.73	2.31
DiGA	2.34	4.18	3.57	5.91	4.82	5.16	3.44	2.89	6.03
DA-CDM	4.03	3.29	3.08	4.07	10.21	6.81	15.32	16.52	15.58
Mask-ControlNet	5.41	7.32	6.15	11.28	10.48	12.88	16.10	15.92	16.78
HiBug_Class	5.09	5.83	3.40	5.14	7.74	7.21	14.64	14.98	13.96
HiBug_Task	6.79	4.88	3.92	18.69	11.75	7.74	15.75	15.15	14.64
Ours (L)	10.39	11.45	11.66	18.09	**18.42**	15.35	22.26	22.52	**21.75**
Ours (Q)	**14.32**	**12.09**	**14.95**	**19.29**	15.42	**15.95**	**22.77**	**23.12**	20.46

Table 2: Fix Rate (FR %) on ImageNet10 under varying image counts, models, and methods.

Method	**ResNet** (base acc: 71.73%)			**ViT** (base acc: 97.42%)			**CLIP** (base acc: 93.78%)		
	100 Images	500 Images	1k Images	100 Images	500 Images	1k Images	100 Images	500 Images	1k Images
Data Augmentation	3.52	2.19	4.67	1.05	3.44	2.78	4.12	1.56	3.93
DiGA	2.15	5.67	3.98	4.41	3.22	6.12	5.49	4.33	2.76
DA-CDM	5.24	7.32	8.53	6.20	13.18	13.57	2.89	4.34	5.95
Mask-ControlNet	1.73	2.87	2.33	-6.59	10.47	3.88	-5.47	1.61	-1.29
HiBug_Class	1.80	1.49	0.74	-11.63	-13.57	-5.43	2.25	6.91	0.80
HiBug_Task	6.30	4.07	5.77	9.30	13.95	6.20	-7.88	6.43	5.95
Ours (L)	8.38	**8.70**	7.75	30.62	**29.07**	26.74	11.41	16.40	10.61
Ours (Q)	**9.13**	6.01	**11.14**	**31.01**	25.97	**38.37**	**12.70**	**17.68**	**18.81**

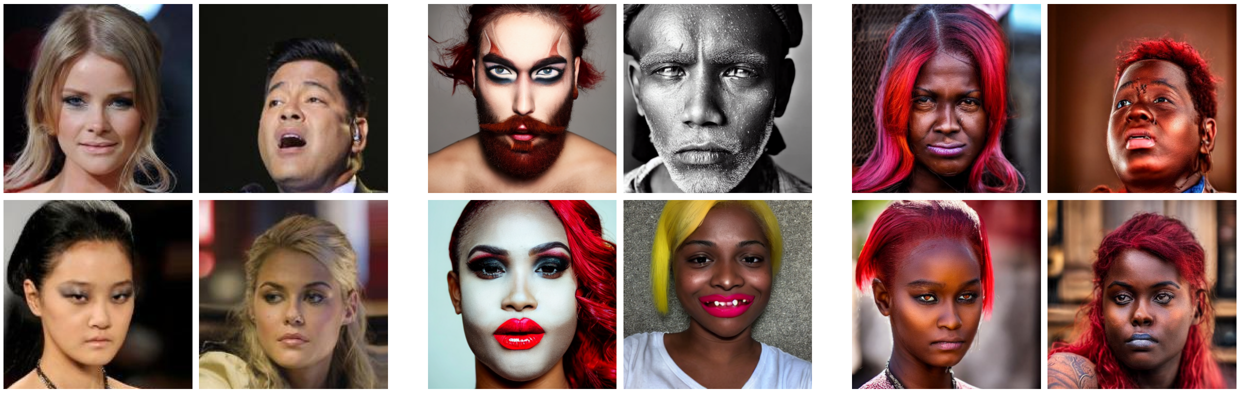
2025) and LLaVA-v1.5-7B (Liu et al., 2023). Generating 1,000 images with ControlNet takes about one hour, and filtering these 1,000 images with the LVLM takes ten minutes. Filtering accuracy for most attributes (e.g., hair color, skin tone) exceeds 90%, which is consistent with the human audit in Section 5.6. Combining three attributes yields at least a 70% pass rate for filtered images, showing that the diffusion model produces high-quality samples and that the LVLM filtering is reliable.

5.2 Main Results

We summarize the main results on the CelebA and ImageNet10 datasets in Tables 1 and 2, respectively.

On **CelebA**, we select rare-case bugs defined by attribute–value combinations `red hair`, `brown skin`, and `sad emotion` for ResNet and ViT, and `yellow hair`, `brown skin`, and `sad emotion` for CLIP, based on the most frequent patterns identified among failure slices. SafeFix consistently achieves the highest test accuracy, i.e., the highest FR, across all models (ResNet, ViT, and CLIP) and varying levels of synthetic augmentation. For example, with 1,000 added images, our method improves FR by +14.32% (ResNet), +10.29% (ViT), and +22.77% (CLIP) relative to their base accuracies. These results show that our attribute-targeted augmentation and filtering pipeline is effective in repairing rare-case failure slices, outperforming both CDM-based and HiBug baselines.

On **ImageNet10**, similar trends emerge, as shown in Table 2. For ResNet and ViT, we target rare-case bugs involving `pink color` and `fabric texture`, while for CLIP we use `orange color` and `fabric texture`. Across all models, our proposed method consistently surpasses the baseline methods. Specifically, our method improves ResNet accuracy by +11.14% (with 1,000 images) compared to the base model. ViT and CLIP also exhibit a steady improvement compared to other methods. Table 3 shows that the proposed SafeFix also achieves better overall accuracy on **ImageNet10** compared with the baselines.



(a) Original Images (b) HiBug-Generated (c) SafeFix-Generated

Figure 3: Comparison of generated images from different methods with edited attributes `red hair`, `brown skin`, and `sad emotion`. HiBug often produces invalid or imprecise samples due to the lack of conditional generation and semantic filtering. In contrast, SafeFix generates attribute-faithful images that specifically target rare-case bugs.

Analysis. Combined with Tables 1, 2, 3, and *all other test-accuracy results* in Appendix F, our experiments show that baseline performance is often unstable. For instance, baselines like DA-CDM and Mask-ControlNet show some improvement on object and facial attribute tasks, respectively, due to their use of conditional diffusion models. However, they are **not targeted model repair methods**, so their overall performance is inferior to both HiBug and SafeFix. This lack of a targeted strategy means their gains are task-specific and **not robustly transferable**. Similarly, the accuracy of HiBug does not improve substantially. This is likely because it does not precisely target failure-critical attributes; instead, it applies general augmentations that can introduce noise or modify unintended features. As shown in Figure 3, this issue is compounded by the lack of LVLM filtering, resulting in generated data that is often misaligned with the intended rare-case fixes.

Table 3: Test accuracy (%) on ImageNet10 using ResNet-18.

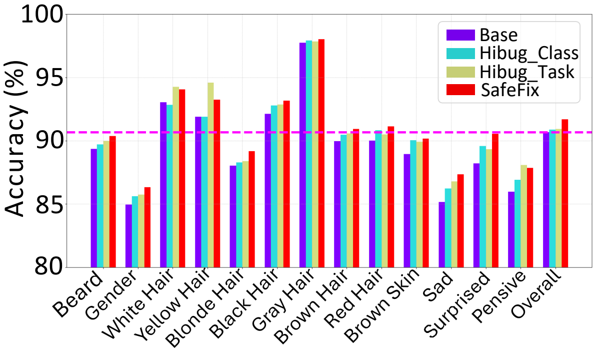
Method	ResNet		
	100 Images	500 Images	1k Images
Base	71.73		
Data Augmentation	72.73	72.35	73.05
DiGA	72.34	73.33	72.86
DA-CDM	73.21	73.80	74.14
Mask-ControlNet	72.22	72.54	72.39
HiBug_Class	72.24	72.15	71.94
HiBug_Task	73.51	72.88	73.36
Ours (L)	74.10	**74.19**	73.92
Ours (Q)	**74.31**	73.43	**74.88**

In contrast, **our method (SafeFix) consistently shows stable or improving performance**. This robustness indicates our attribute-targeted augmentation and LVLM filtering are highly effective at correcting failure-prone subpopulations with meaningful data.

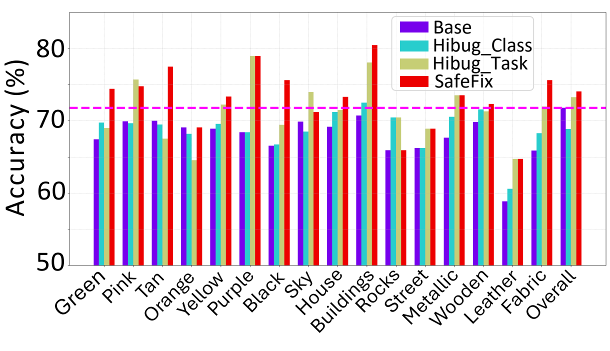
5.3 SafeFix Can Effectively Fix Rare-case Bugs

To verify that SafeFix's improvements specifically address targeted rare-case bugs rather than merely enhancing overall performance, we analyze attribute-level validation accuracy changes on CelebA and ImageNet10, as shown in Figure 4. For clarity, the ImageNet10 plot includes only three representative attributes—`color`, `background`, and `texture`—as the dataset contains many attribute dimensions. Specifically, the left plot highlights improvements for CelebA after adding 5,000 synthetic images to the original 80,000 training samples. The right plot demonstrates accuracy gains on ImageNet10, achieved by adding 100 synthetic images to the original training set of 10,400 samples.

Taking ImageNet10 as an example, our targeted synthetic augmentation on `pink color` and `fabric texture` significantly improved accuracy for these selected rare-case attributes. Accuracy for the `pink color` attribute increased from 69.90% to 74.76%, surpassing the overall accuracy of 74.31%. Similarly, accuracy for the



(a) CelebA



(b) ImageNet10

Figure 4: Accuracy comparison of ResNet-18 trained using different augmentation methods. The dashed line represents the average overall accuracy without additional synthetic training data.

`fabric texture` attribute improved from 65.85% to 75.61%, also exceeding the overall accuracy. In contrast, attributes not explicitly targeted by augmentation, such as the `rocks` background, exhibited minimal or no improvement—its accuracy remained unchanged—highlighting that the gains from SafeFix are concentrated on the intended rare-case bugs rather than uniformly distributed across all attributes.

These substantial attribute-specific improvements confirm that SafeFix effectively repairs identified rare-case failure slices rather than providing a generalized performance boost. SafeFix also shows that all attributes improve across both datasets and introduces **no new bugs**, which indicates that the method remains stable outside the targeted regions. Attributes not targeted by augmentation show negligible accuracy changes, further reinforcing that SafeFix precisely and safely addresses the **targeted** rare-case failures.

5.4 SafeFix Directly Addresses Diagnosed Failure Modes, Not Merely Augments Data

To verify that SafeFix's performance gains stem from accurately fixing rare-case bugs rather than from generic data augmentation, we compare red-hair and yellow-hair augmentations on CLIP. As discussed in Section 5.2, we select red hair for ResNet and yellow hair for CLIP in the CelebA dataset, based on which attribute is more likely to trigger rare-case bugs. For CLIP, both `red hair` and `yellow hair` are low-frequency attributes that meet the rarity criterion in Eq. equation 4. However, only the `yellow hair` slice additionally satisfies the low-accuracy criterion in Eq. equation 5, making it a true rare-case bug slice for CLIP.

Figure 5 confirms that augmenting 1000 images using yellow-hair samples—the diagnosed failure mode for CLIP—substantially improves accuracy on its target slice (**+4.51%**). Furthermore, this targeted augmentation provides positive gains across all other attributes, including on the "Red Hair" slice (+0.91%) and a significant boost to **"Overall" accuracy (+2.66%)**. Conversely, augmenting with red hair, which is not a diagnosed bug, reveals a harmful outcome. Most notably, it degrades performance on its own target slice by **-0.79%**. This counter-intuitive result contrasts with the positive gains seen for ResNet (Figure 4a), highlighting that different architectures can react to synthetic

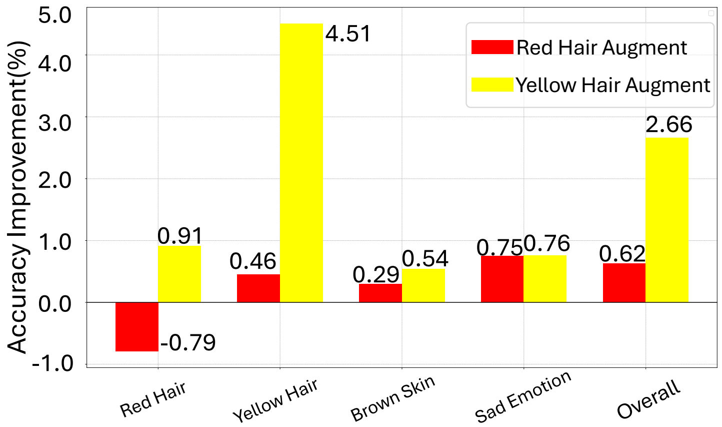


Figure 5: Test accuracy (%) improvements for red-hair vs. yellow-hair augmentation on CLIP.

Table 4: Fix Rate (FR %) for attribute-variant selections. Ours is the RH_BS_SE combination.

# images	RH	BS	SE	RH_BS	RH_SE	RH_BS_SE
1,000	7.85	8.59	11.77	7.74	10.39	**14.32**
5,000	10.82	12.62	11.77	**14.42**	6.57	12.69
10,000	11.03	10.29	12.62	9.97	9.33	**14.95**

data in unpredictable ways. While red-hair augmentation does provide minor gains on other attributes (e.g., +0.75% on `sad emotion`), its minimal overall accuracy impact (+0.62%) and significant negative side effect on the targeted class itself underscore the importance of our diagnosis-driven approach. **Augmenting the *correctly diagnosed* bug** (`yellow hair`) leads to **effective and safe repair**, while augmenting an *undiagnosed* attribute (`red hair`) can be ineffective and harmful.

5.5 Effect of Alternative Bug Slice Selections

Table 4 reports FR on CelebA using ResNet augmenting with different subsets of rare-case attributes. **RH**, **BS**, and **SE** refer to `red hair`, `brown skin`, and `sad emotion`, respectively. Combined settings like **RH_BS** and **RH_BS_SE** (Ours) augment images with multiple attributes. Augmentation improves over the base model (90.57%) across all settings. Notably, our three-way combination, **RH_BS_SE**, achieves the highest FR at $1,000$ and $10,000$ images and remains competitive at $5,000$ images, showing that targeting intersecting rare-case conditions is more effective than augmenting isolated attributes alone. Although **RH_BS** alone achieves the highest FR at $5,000$ images, augmenting only that slice does not address concurrent failure modes, indicating persistent errors. Firstly, our **RH_BS_SE** combination improves all three slices concurrently (Figure 4a). Secondly, for the **RH_BS_SE** attribute combination, the number of bugs decreases from 10 to 4, while augmenting only **SE** or **RH_BS** reduces the count to 7 and 6 respectively (using $\rho = 0.05$ and $\epsilon = 0.03$). A similar pattern holds on ImageNet10: selecting the `pink color` + `fabric texture` slice reduces rare-case bugs from 13 to 7, whereas augmenting only `pink color` reduces them to 9 and only `fabric texture` reduces them to 10.

After further verification, this reduction does not introduce any new bugs that were not present before, indicating that SafeFix repairs existing failures without adding new bugs on both datasets. The number of identified bugs is sensitive to these thresholds and using different threshold values can change the number of bugs fixed, as shown in Appendix H.5. We note that larger attribute combinations (4 to 7 attributes) also operate correctly but reach lower FR (around 9% on CelebA and 7% on ImageNet10), falling below our selected attribute groups.

5.6 Effect of the LVLM Filter and Human Audit

We use an LVLM as the attribute filter, and a human-based test yields nearly identical results. In a human audit conducted with 5 AI graduate students on 300 CelebA images, the pass rates are 97% for "red hair", 99% for "brown skin", 78% for "sad emotion", and 98% for the original label, with an average overlap of about 95% with Qwen2.5-VL(Bai et al., 2025) and 93% with LLaVA-v1.5(Liu et al., 2023), showing that the LVLMs closely match human verification. The main errors arise from background color changes (**BG**) (7%) and the edited image failing to express the target attribute (**ATTR**) (24%); the LVLM fixes most of them, as shown in Table 5.

Table 5: Average error rates (%) before and after LVLM filtering.

Error Type	Before	After
BG	7	2
ATTR	24	3

5.7 Ablation Study

We ablate the contributions of the two core components in our pipeline: the Conditional Diffusion Model (CDM) and the LVLM filter. Table 6 reports the fix rate on CelebA when using different combinations of these components across three augmentation scales. Note that configuration (a) corresponds to the baseline strategy from HiBug (Chen et al., 2023). We observe several trends: adding either the LVLM (b) or the CDM (c) individually improves accuracy across all scales; SafeFix combining both the CDM and the LVLM (d) yields the highest performance in all cases.

Table 6: Ablation on the impact of CDM and LVLM components at different scales, reported as Fix Rate (FR %).

	Components		**CelebA Fix Rate (FR %)**		
	CDM	LVLM	1,000	5,000	10,000
a)			6.79	4.88	3.92
b)		✓	7.21	4.77	6.99
c)	✓		8.06	8.27	10.07
d)	✓	✓	**14.32**	**12.09**	**14.95**

6 Conclusion

We presented an automated model repair pipeline for vision tasks that combines failure-attribute diagnostics with targeted synthetic augmentation. We use a Conditional Diffusion Model (CDM) to generate attribute-preserving variants and apply LVLM-based filtering to ensure semantic correctness, thereby focusing augmentation on true rare-case failure slices. Experiments on CelebA and ImageNet10 with ResNet, ViT, and CLIP backbones show consistent gains in accuracy and reduced bugs on underrepresented subpopulations, outperforming CDM-based and HiBug baselines. Ablations confirm that the CDM and the LVLM contribute complementary benefits, highlighting the importance of targeted, validated augmentation for robust model repair.

Limitations. Our method's effectiveness is constrained by its components. Crucially, the pipeline can inherit biases from the diffusion model or the LVLM, potentially perpetuating the fairness issues if these components have demographic bias. Other limitations include the computational cost of generation and filtering, potential image artifacts, and a fixed attribute vocabulary that cannot address unmodeled failure modes. Future work will focus on mitigating inherited biases, improving efficiency, and dynamic attribute discovery.

References

Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. Debugging tests for model explanations. In *NeurIPS*, 2020.

Julius Adebayo, Michael Muelly, Hal Abelson, and Been Kim. Post hoc explanations may be ineffective for detecting unknown spurious correlations. In *ICLR*, 2022.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

Maximilian Augustin, Valentyn Boreiko, Francesco Croce, and Matthias Hein. Diffusion visual counterfactual explanations. In *NeurIPS*, 2022.

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*, 2025.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Joshua Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Valentyn Boreiko, Maximilian Augustin, Francesco Croce, Philipp Berens, and Matthias Hein. Sparse visual counterfactual explanations in image space. In *GCPR*, 2022.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023. URL https://arxiv.org/abs/2211.09800.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pp. 77–91. PMLR, 2018.

Stephen Casper, Max Nadeau, Dylan Hadfield-Menell, and Gabriel Kreiman. Robust feature-level adversaries are interpretability tools. In *NeurIPS*, 2022.

Muxi Chen, Yu Li, and Qiang Xu. Hibug: On human-interpretable model debug. In *Advances in Neural Information Processing Systems*, 2023.

Muxi Chen, Chenchen Zhao, and Qiang Xu. Hibug2: Efficient and interpretable error slice discovery for comprehensive model debugging. *arXiv preprint arXiv:2501.16751*, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Lisa Dunlap, Alyssa Umino, Han Zhang, Jiezhi Yang, Joseph E Gonzalez, and Trevor Darrell. Diversify your vision datasets with automatic diffusion-based augmentation. In *Advances in Neural Information Processing Systems*, 2023.

Sabri Eyuboglu, Maya Varma, Khaled Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Ré. Domino: Discovering systematic errors with cross-modal embeddings, 2022. URL https://arxiv.org/abs/2203.14960.

Haoyang Fang, Boran Han, Shuai Zhang, Su Zhou, Cuixiong Hu, and Wen-Ming Ye. Data augmentation for object detection via controllable diffusion models. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1246–1255, 2024. doi: 10.1109/WACV57701.2024.00129.

Irena Gao, Gabriel Ilharco, Scott Lundberg, and Marco Tulio Ribeiro. Adaptive testing of computer vision models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4003–4014, 2023.

Robert Geirhos, Claudio Michaelis Rubisch, Felix A Wichmann, Matthias Bethge, 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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Zhiqi Huang, Hao Geng, Haoyu Wang, Huixin Xiong, and Zhiheng Li. Data augmentation for facial recognition with diffusion model. In *CVPR 2024 Workshop SyntaGen: Harnessing Generative Models for Synthetic Visual Datasets*, 2024. URL https://openreview.net/forum?id=GXmlanJ6rC.

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

Wangchunshu Jin and Martin Rinard. Uncovering and mitigating spurious correlations for robust image classification. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), 2018. URL https://arxiv.org/abs/1711.11279.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *International Conference on Learning Representations (ICLR)*, 2015.

Yi-An Lai, Elman Mansimov, Yuqing Xie, and Yi Zhang. Improving prediction backward-compatiblility in nlp model upgrade with gated fusion, 2023. URL https://arxiv.org/abs/2302.02080.

Guillaume Leclerc, Hadi Salman, Andrew Ilyas, Sai Vemprala, Logan Engstrom, Vibhav Vineet, Kai Xiao, Pengchuan Zhang, Shibani Santurkar, Greg Yang, Ashish Kapoor, and Aleksander Madry. 3db: A framework for debugging computer vision models. *Advances in Neural Information Processing Systems*, 35: 8498–8511, 2022.

Hansang Lee, Haeil Lee, and Helen Hong. Genmix: Combining generative and mixture data augmentation for medical image classification, 2024. URL https://arxiv.org/abs/2405.20650.

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*, pp. 12888–12900. PMLR, 2022.

X Li et al. Your vision-language model itself is a strong filter. *arXiv preprint arXiv:2402.12501*, 2024.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild, 2015. URL https://arxiv.org/abs/1411.7766.

Agnieszka Mikołajczyk-Bareła, Maria Ferlin, and Michał Grochowski. Targeted data augmentation for bias mitigation. *arXiv preprint arXiv:2308.11386*, 2023.

OpenAI. Gpt-4 technical report. https://openai.com/research/gpt-4, 2023. Accessed: 2025-05-16.

Shubham Parashar, Zhiqiu Lin, Tian Liu, Xiangjue Dong, Yanan Li, Deva Ramanan, James Caverlee, and Shu Kong. The neglected tails of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL https://arxiv.org/abs/2401.12425.

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*, pp. 8748–8763. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Sahil Singla, Atoosa Malemir Chegini, Mazda Moayeri, and Soheil Feizi. Data-centric debugging: Mitigating model failures via targeted image retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4991–5000, 2024.

Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*, 2023.

Joshua Vendrow, Saachi Jain, Logan Engstrom, and Aleksander Madry. Dataset interfaces: Diagnosing model failures using controllable counterfactual generation. In *NeurIPS*, 2023.

Weizhi Wang, Khalil Mrini, Linjie Yang, Sateesh Kumar, Yu Tian, Xifeng Yan, and Heng Wang. Finetuned multimodal language models are high-quality image-text data filters. *arXiv preprint arXiv:2403.02677*, 2024a.

Yinong Oliver Wang, Younjoon Chung, Chen Henry Wu, and Fernando De la Torre. Domain gap embeddings for generative dataset augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

Saining Xie and Zhuowen Tu. Holistically-nested edge detection, 2015. URL `https://arxiv.org/abs/1504.06375`.

Fengda Zhang, Qianpei He, Kun Kuang, Jiashuo Liu, Long Chen, Chao Wu, Jun Xiao, and Hanwang Zhang. Distributionally generative augmentation for fair facial attribute classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Lvmin Zhang, Maneesh Wu, Chengyue Zhang, Yijun He, Siwei Zhang, Xiaolong Xu, Jingwan Yang, Kevin Shih, Xintao Wang, Yingying Zhang, et al. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Chen Zhao, Feng Mi, Xintao Wu, Kai Jiang, Latifur Khan, and Feng Chen. Adaptive fairness-aware online meta-learning for changing environments. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2565–2575, 2022.