Attribution Analysis-based Concept Alignment: A Human-in-the-loop Data Debugging Framework

Lei Chai[1,2], **Lu Qi**[1], **Hailong Sun**[1,2*], **Jing Zhang**[1], **Jingxuan Xu**[1]

[1]State Key Laboratory of Complex & Critical Software Environment (CCSE), Beihang University, China
[2]Hangzhou Innovation Institute of Beihang University, China
chailei@buaa.edu.cn, qiluxt@buaa.edu.cn, sunhl@buaa.edu.cn, zhang_jing@buaa.edu.cn, e1553956@u.nus.edu

Abstract

Ensuring consistently high-quality training data is essential for developing reliable machine learning systems. Recent research demonstrates that incorporating human supervision into training set debugging effectively improves model performance, especially for text classification tasks. However, such methods often prove inapplicable to image understanding tasks, where inherently unstructured pixel data presents challenges in understanding and correcting biases. Inspired by 'human-AI alignment', we introduce AACA (**A**ttribution **A**nalysis-based **C**oncept **A**lignment), a human-in-the-loop framework that mitigates bias in the training set by aligning the concepts focused by humans and AI during decision-making. Specifically, AACA comprises two primary stages: interpretable data bug discovery and targeted data augmentation. During the data bug discovery stage, AACA identifies confounded and valid concepts to explain 'why prediction failure occurs' and 'what concept the model should focus', using interpretability methods and human annotation. In the stage of targeted data augmentation, AACA adopts these concept-level attributions as clues to synthesize debugging instances via text-to-image generative model. The prediction model is then retrained on the augmented set to correct prediction failures. Comparative experiments conducted on crowdsourced annotations and real-world datasets demonstrate that AACA can accurately identifies data bugs and effectively repairs prediction failures, thereby significantly improving prediction performance.

Code — https://github.com/Anonymous4AACA/AACA

Introduction

Deep learning models (DNN) have made remarkable advancements in recent years (Goodfellow 2016; Sun et al. 2016). Nonetheless, they are susceptible to capturing spurious correlation and shortcuts from the biased training data, thus exhibiting systematic failures on specific subsets of data with similar visual attributes (Ye et al. 2024). Such susceptibility can lead to potentially serious consequences for various applications in high-stake domains, such as autonomous driving (Wang et al. 2024), safety monitoring (Dong et al. 2025) and medical image processing (Chen et al. 2019).

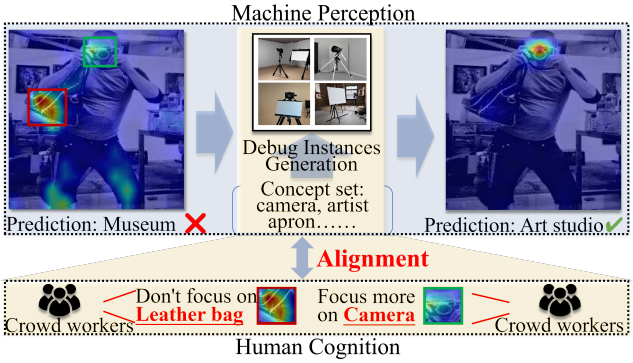


Figure 1: The image 'art studio' was incorrectly classified as 'museum', with the model incorrectly focus on the confounding concept 'leather bag'. Human cognition helps the model shift focus to the valid concept 'camera', which makes the model classify the image correctly.

Meanwhile, most existing image classification models are constructed based on neural networks, which are generally considered as unexplained 'black boxes' (Pedreschi et al. 2019). The inherent opacity of these models presents challenges for revealing and repairing the failure modes.

The current data debugging methods (Wu et al. 2020; Li et al. 2022) can be broadly classified into automatic debugging methods and human-in-the-loop debugging methods based on the difference in human involvement.

Automated data debugging methods (Li et al. 2022; Singla et al. 2022) are widely adopted due to their low cost and high scalability (Chen, Li, and Xu 2024). However, they typically focus on local attributions (Mahendran and Vedaldi 2015), such as identifying key pixels within an image or key data instances, which often lack unbiased, global explanations and thus limit the effectiveness.

Human-in-the-loop (Kondyli, Suchan, and Bhatt 2022; Teso and Kersting 2019; Chai et al. 2024) or human-centered debugging (Lertvittayakumjorn and Toni 2021; Biswas et al. 2022) approaches incorporate human knowledge (Sharifi Noorian et al. 2022; Chai, Sun, and Zhang 2024) and expertise (Balayn et al. 2021) into the debugging process, and are able to explain and optimize prediction failures in a human-understandable way. However, most

*Corresponding Author

existing human-in-the-loop debugging methods face challenges in effectively debugging image classification models due to several limitations. Firstly, instance-level annotations require extensive human involvement to optimize complex neural network models, resulting in significant costs (Han and Ghosh 2021). Secondly, these debugging methods for failures repair primarily focus on constraining the model's learning process using a fixed training set, which limits their effectiveness when the training data contain biases (Jain et al. 2022). Unfortunately, biases are common in real-world datasets (Whang et al. 2023).

To address these issues, we introduce an Attribution Analysis-based Concept Alignment (AACA) framework for human-centered data debugging. Specifically, we explore to identify the systematic biases in model predictions with a series of concepts, and further leverage concept-level human feedback as guides for updating the training set.

An example is introduced in Figure 1. A trained image classification model incorrectly classifies a test image as 'museum'. Analysis of the saliency map reveals that the model relies on the confounded concept 'leather bag' as its prediction rationale, failing to recognize the correct correlation between the valid concept 'camera' and the image category 'art studio'. Consequently, AACA generates prompts based on the human-annotated valid concept 'camera' and class label 'art studio'. These prompts are fed into a text-to-image generative model to synthesize debug instances. After retraining on the augmented training set, the model correctly classifies the test image, attributing its decision to the valid concept 'camera'.

In summary, our work makes the following contributions:

- We investigate the influence of concept alignment on the effectiveness of data bug discovery and training set augmentation. We then formulate this problem as interpretable data debugging;
- We introduce AACA, a data debugging framework that identifies and repairs systematic prediction failures through attribution analysis-based concept alignment;
- We present a concept-level data attribution method that identifies data bugs in a human-understandable way, and synthesize debug images using T2I models to mitigate biases inherent in the initial training set;
- We conduct comparative experiments using crowdsourced annotations and real-world datasets to demonstrate the effectiveness of AACA. Additionally, we conduct a series of intermediate analysis to explore why AACA works and how to optimize it.

Related Works

In this section, we will review current data debugging methods and discuss their limitations.

Automatic Debugging Methods

Most of automatic methods utilize machine learning interpretability methods (like saliency maps (Simonyan, Vedaldi, and Zisserman 2013) and activation maps (Jiang et al. 2021)) to generate instance-level explanations of model failures. These failures are then rectified using a range of data debugging strategies, which include removing image patches (Jain et al. 2022), finding influential functions (Koh and Liang 2017), generating counterfactual inputs (Mahendran and Vedaldi 2015; Vendrow et al. 2023), collecting similar data from the search engine (Singla et al. 2022) and generating target data by large language models (Chen, Li, and Xu 2024).

These automated debugging methods typically use pixel-level prediction attribution as the foundation for identifying and repairing data bugs. However, these localized explanations are generally not sufficiently reliable due to the lack of consistency across instances. Additionally, their incomprehensibility to humans further reduces their subjective reliability.

Human-in-the-loop Debugging Methods

In light of the limitations of automated data debugging methods on reliability and interpretability, recent studies have started to incorporate human annotations into the debugging process. For instance, Jie Yang et al. integrate machine learning interpretability methods with human annotation tasks to interpret model behavior (Sharifi Noorian et al. 2022; Balayn et al. 2021; Chai, Sun, and Wang 2022). Furthermore, recent research on Explanation-based Human Debugging (EBHD) utilizes human feedback to enhance training data and retrain models to repair bugs. These approaches include rectifying mislabeled training examples (Koh and Liang 2017; Han and Ghosh 2021), assigning noisy labels to unlabeled instances (Yao et al. 2021), filtering irrelevant words from input texts (Ribeiro, Singh, and Guestrin 2016), and generating augmented sentences to mitigate artifact effects (Teso and Kersting 2019). Although current human-in-the-loop approaches can somewhat uncover and address data bugs, their performance and scalability are often limited by the size of training sets and the opacity of pixel-level explanations.

To address these challenges, we introduce AACA, a human-centered data debugging framework inspired by human-AI alignment. AACA integrates automated attribution algorithms, human annotation tasks, and text-to-image generative models to identify and rectify data bugs at conceptual level. This approach effectively alleviates the performance limitations of current data debugging methods due to the challenges such as 'local clues', 'human-machine divergence' and 'fixed training sets'.

Problem Statement

To model the debugging process depicted in figure 1, we identify a two-stage problem called interpretable data debugging. The goal of this problem is to correct systematic prediction biases through interpretable attribution analysis and targeted data augmentation.

Definition 1: (Interpretable Data Debugging.) Given an initial training set D_{init} and an initial model M_{init} that exhibits systematic bias, leading to suboptimal predictions on the validation set D_{val}. The objective of interpretable data debugging is to construct an enhanced training set D_{aug} =

$\mathcal{M}$ (D_{init}, D_{debug}) according to the analysis of prediction attribution I and optimization feedback C_{feed} on the seed set D_{seed}, in a human-interpretable way. After retrained on D_{aug}, an optimized model M' is obtained with improved performance on the validation set D_{val}.

Specifically, $\mathcal{M}$ refers to the strategy for integrating datasets, while D_{seed} is a seed set random selected from the validation D_{val}, used for subsequent data bugs discovery and data synthetic.

Effective data debugging hinges on accurately characterizing training set biases and leveraging feedback to resolve them. Thus, we propose converting model prediction biases and optimization feedback into human-annotated concepts. This creates a unified representation linking upstream error discovery and downstream augmentation, bridging model attribution with human rationale.

Problem 1 (Data Debugging through Concept Alignment.) Concept alignment seeks to 'pursuing alignment between humans and AI at the conceptual level, so that AI systems understand the world using the same concepts humans use to comprehend it' (Rane et al. 2024). It emphasizes enhancing the predictability and interpretability of AI systems through conceptual communication between humans and machines.

Inspired by this, we formalize bias identification using concepts and concept-level human feedback for dataset updates:

$$\arg\min_{D_{debug}} \sum_i \sum_j \|P(D_{aug}\ (s_i, c_j)) - P(Human(s_i, c_j))\|,$$

where D_{debug}= $\{d_1, d_2,...,d_n\}$ minimizes distributional divergence between human and AI cognition for instance category s_i and concept c_j.

Under the definition of data debugging, the problem of collecting effective debugging instances can be formulated as follows:

$$d_i = \mathcal{C}\ (\ \mathcal{S}e,\ C_{feed}),$$

$$s.t.\ C_{feed} = \mathcal{F}\ (\mathcal{I}\ (M_{init},\ D_{seed})\ ),$$

$$and\ \ \mathcal{Q}(\mathcal{C}) > \epsilon,\ (quality\ constraint)$$

Here, $\mathcal{C}$ (e.g., synthesis/retrieval methods) generates debug instances $d \in D$ using concept-level feedback and selection strategy Se, improving concept alignment between D_{init} and human rationale. Additionally, $\mathcal{F}$ designs feedback mechanisms (e.g., human annotation tasks) that combine D_{seed} and M_{init} to produce concept-level clues C_{feed} about 'what the model actually knows' and 'what it should knows', aided by a failure attribution algorithm $\mathcal{I}$. Notably, the quality of the data collection method $\mathcal{C}$ also play a crucial role in effective data debugging.

Method

The primary objective of AACA is to attribute systematic prediction failures to concept-level data bugs and repair them through targeted training set augmentation and model retraining.

As illustrated in Figure 2, AACA consists of seven components, which are categorized into two stages based on their functions: interpretable data bug discovery and targeted data augmentation. The stage of interpretable data bug discovery includes steps **a** and **b**, while the targeted data augmentation stage includes steps **c**, **d**, **e** and **f**. A brief overview of the execution process is provided in Algorithm 1 for clarity.

Characterized by high flexibility and modularity, AACA allows users to replace any component with a more effective alternative if needed.

Interpretable Data Bug Discovery

Interpretable bug discovery aim to identify human-understandable concepts that explain the causes of model failures.

(a) Failure attribution. Initially, a machine learning interpretability method (Chefer, Gur, and Wolf 2021) $\mathcal{I}$ is used to generate saliency maps $A(i)$:

$$A(i) = \int_{\alpha=0}^{1} \frac{\partial M_{init}(x + \alpha \cdot (x_{baseline} - x))}{\partial x} d\alpha$$

To alleviate the uncertainty of attribution results, we compute Monte Carlo dropout variance (Milanés-Hermosilla et al. 2021):

$$\sigma_A(i) = \text{Var}_{\theta \sim p(\theta)}\left[A(i;\theta)\right]$$

Regions with $\sigma_A(i) > \tau$ ($\tau = 0.1$) are flagged as ambiguous and prioritized for human verification.

The saliency maps provide pixel-level prediction attributions to help humans understand the areas the model focuses on when classifying images.

(b1) Concept-level confound set. The confounded set $C(s_n)$ contains concepts indicating the visual features that the current model struggles to recognize in class s_n.

(b2) Concept-level valid set. Conversely, the valid set $V(s_n)$ includes valid concepts $v(s_n)$ that represent the visual features contributing to the correct classification of images in category s_n. Human annotations are incorporated to convert pixel-level model attributions into comprehensible concepts. Figure 3 presents the interface for the human annotation task, where participants inspect these two images and write down the confounded and valid concepts in designated answer boxes. Details about the annotation task, the quality control strategy and cost analysis are introduced in Appendix.A.

Targeted Data Augmentation

Part 1: Debug images Generation. Subsequently, leveraging the valid set $V(s_n)$, the confounded set $C(s_n)$ and the controllable image generation capability of the text-to-image generative model (Vendrow et al. 2023), debug images are synthesized to construct the debug set for failure repair.

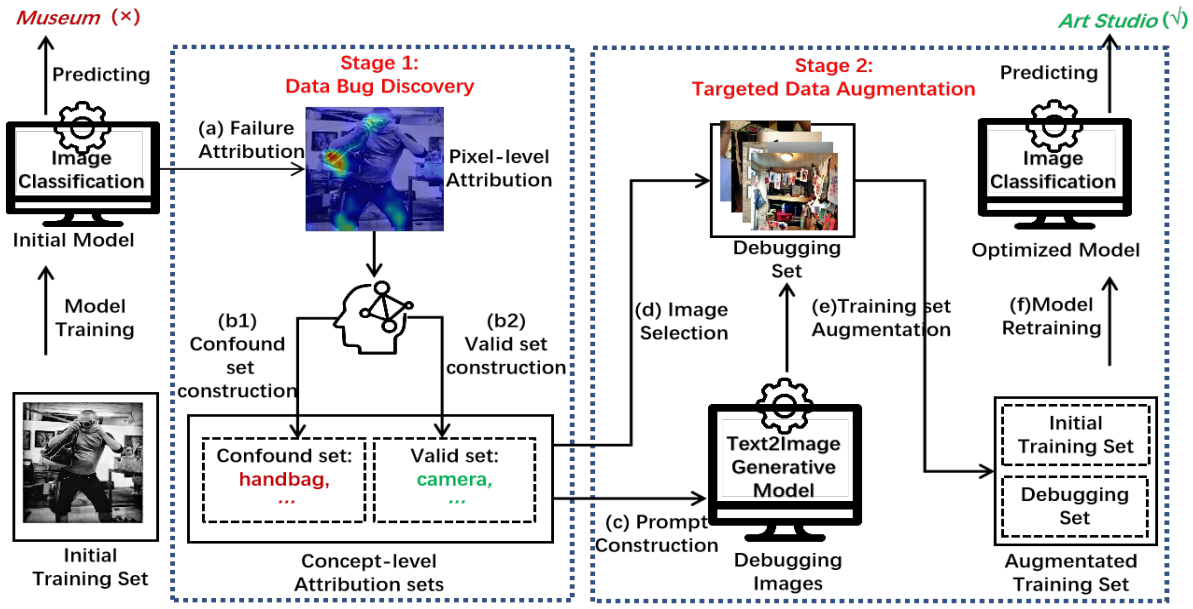


Figure 2: Framework of AACA: It takes an initial training set D_{init} and an initial prediction model M_{init} as input, and produces sets of valid and confounded concepts, a set of debugging images D_{debug} and an optimized prediction model M' as output.

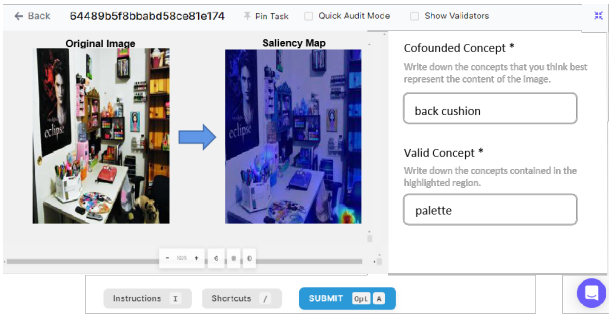


Figure 3: Interface for the Concept Annotation Task.

(c) Prompt construction. Prompt p describe the synthetic debug images d. Assume $V(s_n)$ is the set of concepts for which objects are present in the synthetic image d. We construct the prompt $p\langle s_n, v(s_n)\rangle$ that includes class label s_n and concepts $v(s_n) \in V(s_n)$. e.g., p = "Please generate a photo of art studio containing a camera", when s_n = 'art studio' and $v(s_n)$ = 'camera'. Prompts p are then fed into text to the image generative model (Vendrow et al. 2023) to produce debug images D'_{debug}.

(d) Images selection via visual similarity. Simply adding debug images to the training set may not effectively enhance the prediction performance due to the diverse quality of synthetic images and a large extent of noise. Our data selection strategy creates a debug-train set D_{debug} by selecting images from D'_{debug} with higher predictive entropy (Gal, Islam, and Ghahramani 2017) and lower semantic correlation with the confounded concepts in C. Specifically, each synthetic image is prioritized based on the following formula:

$$Score_{pri}(d_i) = \lambda H[s_n|d_i, D_{debug}] - (1-\lambda) [\sum_{j=0}^{k} sim(emb(d_i), emb(c_j(s_n)))], \quad (1)$$

where H refers to cross entropy, sim is cosine similarity, $emb(c_j(s_n))$ denotes the feature vector of the image salience region where the concept is located, $\lambda \in (0,1)$. The synthesized images are arranged in descending order of priority scores, adding each image category s_n in turn to construct the debug set $D_{debug}(s_n)$.

Part 2: Failure Repair. Systematic biases in the initial model M_{init} are repaired through **(e) training set augmentation** and **(f) model retraining**. The augmented training set D_{aug} is constructed by adding the synthesized images from the debug set D_{debug} to the original training set D_{init}. The model is then retrained on D_{aug} using the same training method. The updated model is recorded as M'.

Experiments

Datasets

Crowd annotations. We recruited crowd workers on Scale AI [1] to undertake concept annotation tasks for 240 images. A pre-testing strategy was employed to ensure high-quality annotations by excluding participants whose accepted rate was less than 80%. Each participant received $1.35 per task, equivalent to an average hourly wage of $10.65.

Places365 [2] (Zhou et al. 2014) contains 1.8 million training images spanning 365 diverse scene categories. Within

[1] https://scale.com/
[2] http://places2.csail.mit.edu/

Algorithm 1: Pseudo code of AACA

Require: Initial training set D_{init}, Seed image set D_{seed}, Failure model M_{init}, Attribution Model $\mathcal{I}$, Text to image generative model M_g;

Ensure: Confounded set C, Valid set V, Augmented set D_{aug}, Repaired model M';

for Subset of seed images D_m on class m **do**
  for image i in D_m **do**
    Generate saliency map $A(i)$
    Collect confounded concept c_i and valid concept v_i displayed on $A(i)$ through human annotation tasks;
    C'_m = C_m.append (c_i)
    V'_m = V_m.append (v_i)
  Select top 10 concepts through majority voting:
  c_m = MV (C'_m)
  v_m = MV (V'_m)
for Subset of valid set V_m on class m **do**
  for Valid concept v_m in Valid set V_m **do**
    Construct prompt p_m with v_m and m
    Synthesis images by M_g and p_m:
    $D_{debug}(v_m)$ = M_g (p_m)
    $D_{debug}(m)'$.append ($D_{debug}(v_m)$)
    Select debug images by priority score:
    $D_{debug}(m)$ = Top1000 $Score_{pri}$ ($D_{debug}(m)'$)
for m in Class **do**
  D_{debug}.append ($D_{debug}(m)$)
Construct Augmented set D_{aug}:
D_{aug}= concatenate (D_{init}, D_{debug})
Model retraining: M' = train (D_{aug})
Return C, V, D_{aug}, M'

the training set, each category is capped at a maximum of 5000 images, ensuring a diverse and representative distribution within each class. To mitigate potential biases stemming from diverse interpretations of scenes by crowd workers, we selected a subset with similar visual attributes for comparative experiments and model evaluation. This subset comprises 12 scene categories that encompass indoor concepts.

Experimental Settings

In this study, MAE (He et al. 2022) was used as the baseline, which is built on the Vision Transformer architecture and achieves the state-of-the-art performance on Places365. The effectiveness of AACA is assessed by comparing the classification accuracy to baseline models. The advancement of AACA is verified by comparing the average incremental accuracy of AACA with various data augmentation methods, including search-based methods (including random selection (Ghiasi et al. 2021), similarity-based selection (Schuhmann et al. 2022) and active learning-based selection (Gal, Islam, and Ghahramani 2017)) and generation-based methods (including directly generating (Rombach et al. 2022), Tailor generation (Vendrow et al. 2023) and generation with concept (Gandikota et al. 2024)). A detailed introduction of the baselines and experimental settings are provided in Appendix.B.

Debug-Train method	**Avg. Acc.(%) vs. Debug Image No.**				
	0	500	1000	1500	2000
Baseline	86.75	-	-	-	-
S-Random	-	86.92	84.39	85.72	84.66
S-Similarity	-	86.56	86.89	87.11	87.45
S-Active Learning	-	**87.68**	87.32	88.08	87.94
G-LLM	-	87.52	85.44	86.36	87.17
G-TG	-	86.22	85.31	86.83	86.77
G-LLM-prompt	-	86.91	86.56	87.42	87.21
AACA_app.	-	86.91	87.57	86.44	85.25
AACA_sel.	-	87.36	**88.52**	**89.92**	**89.04**

Table 1: Comparison of classification accuracy with research-based and generation-based data augmentation methods.

The Effectiveness of AACA

Part 1: The accuracy improvement of AACA. Both $AACA_{app.}$ (direct addition of generated debug images) and $AACA_{sel.}$ (priority-score-based selection) significantly outperform the baseline (Table 1). $AACA_{sel.}$ achieves peak improvement of **3.17%** at 1,500 debug images, demonstrating the value of selective augmentation. We further validated the generalizability of AACA through image captioning tasks (Appendix C).

Part 2: The Advancement of AACA. As shown in Table 1, $AACA_{sel.}$ consistently surpasses search-based methods (S-Random, S-Similarity, S-Active Learning) across experimental settings, with maximal average gain of **2.95%** at 1500 images.

Additionally, AACA significantly outperforms generation-based methods (rows G-LLM, G-TG and G-LLM-prompt). Compared to directly generating debug instances, the debug images generated by AACA capture the distribution of seed images more accurately and better aligns the primary concepts within the images, which may be an important factor contributing to the observed performance enhancement. An visual comparison among the initial images, directed generated images and debug images generated by AACA are introduced in Appendix.D to illustrate this phenomenon.

Evaluation of Concept Alignment

Part 1: Objective evaluation of concept alignment. To quantitatively assess AACA's concept alignment efficacy, we employed the current state-of-the-art object detection model, YOLO11 [3], to extract concepts from $D_{val'}$, D_{init}, and D_{aug}. We analyzed occurrence probabilities for the top 10 most frequent concepts in $D_{val'}$ (covering 90% of concept occurrences). As visualized in Figure 4:

- **Distributional alignment** between D_{aug} and $D_{val'}$ significantly improved, with underrepresented concepts

[3]https://github.com/ultralytics/ultralytics

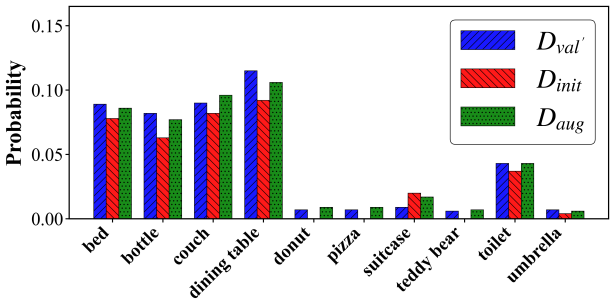


Figure 4: The concept ratios across the validation set $D_{val'}$, the initial training set D_{init}, and the augmented training set D_{aug}.

(e.g., 'donut', 'pizza', 'teddy bear') increasing in probability to match validation set proportions

- **Overrepresented concepts** (e.g., 'suitcase') decreased from 0.02 to about 0.015.
- **Jensen-Shannon Divergence (JSD) (Menéndez et al. 1997)** confirmed these improvements:

$$JS(Distr.(D_{init})||Distr.(D_{val'})) = 0.152836,$$

$$JS(Distr.(D_{aug})||Distr.(D_{val'})) = 0.084219,$$

This 45% reduction in distributional divergence demonstrates AACA's effectiveness in aligning training concepts with target distributions.

Part 2: Subjective evaluation of concept alignment. Since data debugging aims to minimize the discrepancy between the concept distribution in D_{debug} and human cognition, we recruited 20 crowd workers on Scale AI to evaluate concepts in D_{init} and D_{aug} using a 5-point Likert scale (Joshi et al. 2015). They assessed three dimensions:

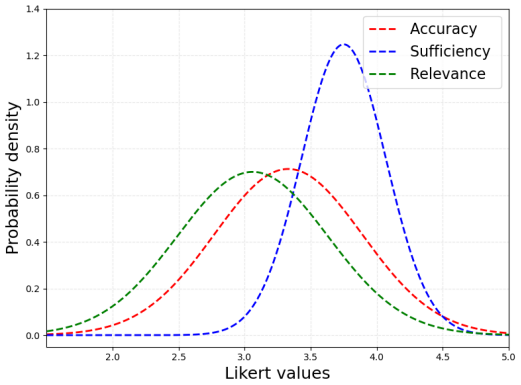
- **Accuracy**: how well the concepts correctly identify the image categories.
- **Sufficiency**: Whether all important concepts are included to enable human understanding of scenes.
- **Relevance**: Whether the significance of identified concepts matches their proportional ranking.

Evaluation results (fitted to normal distributions in Figure 5) show AACA improved both the mean and variance of scores for D_{aug}, making it more accurate, sufficient, and relevant. Notably, "Relevance" saw a significant increase (average from **3.060** to **3.520**).

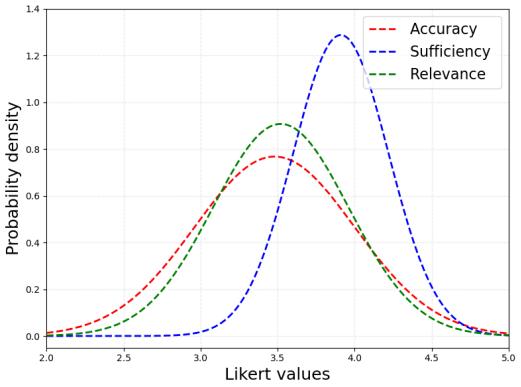
Further analysis of Figure 5 highlights differences across dimensions: "Sufficiency" scored highest (mean 3.910, variance 0.310), indicating strong agreement that concepts adequately described scenes. "Relevance" scored lowest (mean 3.060, variance 0.570), reflecting inconsistency between concept importance and their ranking.

Ablation Study

Part 1: Generalization Across Text-to-Image Models. To test AACA's adaptability, we used three open-source models: Stable Diffusion (Rombach et al. 2022), FLUX.1[4], and DeepSeek Janus-Pro-7B[5]. For each, 1500 debugging images were generated using AACA_app. (direct addition) and AACA_sel. (priority-based selection). Results (Table 2) show AACA consistently outperforms the baseline (86.75%): AACA_sel. achieves 88.52% (Stable Diffusion), 89.92% (FLUX.1), and 89.08% (DeepSeek Janus), confirming strong generalization.



(a) Distributions of Human Evaluation on D_{init}



(b) Distributions of Human Evaluation on D_{aug}

Figure 5: Subjective evaluations of concept alignment on Accuracy, Sufficiency and Relevance.

Baseline	T2I Model	AACA_app.	AACA_sel.
86.75	Stable Diffusion	87.57	88.52
	Flux.1	87.58	89.92
	Deepseek-Janus	86.92	89.08

Table 2: Performance comparison of AACA variants across different text-to-image generative models.

Part 2: Necessity of Human Annotation vs. LLMs. In this section, we explore whether a large language model can replace human effort in data debugging. Specifically,

[4]https://flux1.ai/

[5]https://huggingface.co/deepseek-ai/Janus-Pro-7B

We asked GPT-4 to generate a list of 10 concepts for each scene that best determine the scene's category, denoted as V_{gpt}. We then applied the same approach introduced in Section to construct prompts, generate debug instances, retrain model and evaluate its performance. The experiment results for G-LLM-prompt and $AACA_{sel.}$ in Table 1 indicate that human involvement remains crucial for effective data debugging. We posit that large language models, such as GPT-4, typically provide only coarse-grained common concepts (e.g. 'computer' for 'office'), whereas human annotators can offer finer-grained concepts (e.g. 'laptop'). These finer distinctions may play a crucial role in determining scene categories.

Part 3: Evaluating the contribution of similarity-based selection strategy. We observe from Table 1 and Table 2 that $AACA\_sel.$ significantly outperforms $AACA\_app.$. This result indicates that: 1) directly incorporating generated images into the training set can introduce substantial noise, adversely affecting prediction performance; 2) selecting images based on semantic similarity and priority scores effectively reduces the impact of noisy data and enhances prediction performance.

Additional experiment results about "annotation cost and efficiency", "visual analysis of generated instances"and "hyperparameter analysis" are introduced in Appendix A.2, Appendix.D, and Appendix.E respectively.

Discussion and Limitations

Despite validating the potential of attribution analysis, concept alignment, and human-in-the-loop methods in data debugging, our work faces limitations in scalability and quality control. Below, we discuss these aspects alongside future directions for human-centered data debugging.

Motivations and potential applications

AACA draws inspiration from the development and maintenance processes of ML systems in industrial applications, as well as the concept of data-centric AI (Whang et al. 2023). **Rather than proposing one powerful algorithm or model to fit all tasks, our primary focus is on proposing a model-agnostic framework to ensuring consistently high-quality training data.**

Factors Affecting AACA's Effectiveness

Given that AACA comprises multiple modules, each with the potential to produce biased results that limit the validity of AACA. To ensure the generalizability of AACA, we carefully examined the outputs of each module to identify the factors that might cause quality issues and proposed solutions.

The bias/subjectivity in crowdsourcing annotation Human annotators may misidentify concepts due to saliency map misinterpretation or cognitive bias. Mitigations include: (1) pre-testing to ensure familiarity with scenarios; (2) selecting top 10 frequent concepts per scene to reduce subjectivity; and (3) unifying synonymous concepts via semantic matching.

Reliable issue on T2I Models Generated images may exhibit distribution drift. We address this by: (1) validating across multiple models (Stable Diffusion, FLUX.1, DeepSeek Janus); (2) filtering noisy images via similarity and priority scores; and (3) using text inversion (Rombach et al. 2022) to align generated data with target distributions.

Uncertainty in Interpretability Models Saliency maps may highlight ambiguous regions. Mitigations include: (1) leveraging human annotators to judge ambiguous attributions (e.g., ignoring irrelevant highlighted pixels); and (2) enabling modular replacement of interpretability methods $\mathcal{I}$ with more robust alternatives.

The size of debugging set Table 1 shows performance improvement plateaus with increasing debugging images (added in descending priority order per Formula 1), indicating marginal gains beyond a threshold.

The limitations in scalability

Human-in-the-loop methods face inherent challenges in cost and scalability, especially for large ML models. While concept-level attribution reduces annotation costs, derived insights are task/category-specific, limiting cross-domain applicability—why our experiments focus on representative tasks/datasets.

To address these exciting and challenging problems, we plan to work on two topics in the future:

Enhancing Synthetic-Real Distribution Alignment Generated data often contains noise; advancing semantic similarity-guided synthesis (e.g., refining text-to-image alignment with validation set distributions) remains critical.

Amplifying Human Effort via Annotation Task Design Human involvement poses challenges regarding scalability, cost, and efficiency. Designing efficient annotation tasks to amplify human cognition (e.g., enabling human annotators to provide task-level causal knowledge rather than instance-level local attributions) emerges as a promising solutions to this issue.

Conclusion

This paper introduced the **Attribution Analysis-based Concept Alignment (AACA)** approach, a novel human-in-the-loop framework for interpretable data debugging. By bridging model attribution analysis with human-understandable concept feedback, AACA addresses a critical gap in current data-centric AI systems: the **lack of semantic alignment between machine-learned patterns and human cognitive frameworks**. Our experiments demonstrate that AACA significantly enhances debugging efficacy—accurately identifies and repairs data bias, further **boosting classification accuracy by 3.17%**. Additionally, our fine-grained experimental analyses—including concept distribution evaluation, synthetic instance validation, and ablation studies—indicate that semantic similarity-guided data synthesis and causal-driven human annotation design may play a key role in advancing data debugging in future research.

Acknowledgements

This work was supported partly by National Natural Science Foundation of China under Grant No. 62472017, and partly by Guangxi Collaborative Innovation Center of Multi-source Information Integration and Intelligent Processing.

References

Balayn, A.; Soilis, P.; Lofi, C.; Yang, J.; and Bozzon, A. 2021. What do you mean? Interpreting image classification with crowdsourced concept extraction and analysis. In *Proceedings of the Web Conference 2021*, 1937–1948.

Biswas, S.; Corti, L.; Buijsman, S.; and Yang, J. 2022. CHIME: Causal Human-in-the-Loop Model Explanations. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, 27–39.

Chai, L.; Qi, L.; Sun, H.; and Li, J. 2024. RA 3: A Human-in-the-loop Framework for Interpreting and Improving Image Captioning with Relation-Aware Attribution Analysis. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, 330–341. IEEE.

Chai, L.; Sun, H.; and Wang, Z. 2022. An error consistency based approach to answer aggregation in open-ended crowdsourcing. *Information Sciences*, 608: 1029–1044.

Chai, L.; Sun, H.; and Zhang, J. 2024. Quality Control in Open-Ended Crowdsourcing: A Survey. *arXiv preprint arXiv:2412.03991*.

Chefer, H.; Gur, S.; and Wolf, L. 2021. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 782–791.

Chen, C.; Dou, Q.; Chen, H.; Qin, J.; and Heng, P.-A. 2019. Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 865–872.

Chen, M.; Li, Y.; and Xu, Q. 2024. HiBug: On Human-Interpretable Model Debug. *Advances in Neural Information Processing Systems*, 36.

Dong, S.; Ma, M.; Lamp, J.; Elbaum, S.; Dwyer, M. B.; and Feng, L. 2025. Quantitative Predictive Monitoring and Control for Safe Human-Machine Interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 26203–26210.

Gal, Y.; Islam, R.; and Ghahramani, Z. 2017. Deep bayesian active learning with image data. In *International conference on machine learning*, 1183–1192. PMLR.

Gandikota, R.; Orgad, H.; Belinkov, Y.; Materzyńska, J.; and Bau, D. 2024. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5111–5120.

Ghiasi, G.; Cui, Y.; Srinivas, A.; Qian, R.; Lin, T.-Y.; Cubuk, E. D.; Le, Q. V.; and Zoph, B. 2021. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2918–2928.

Goodfellow, I. 2016. Deep learning.

Han, X.; and Ghosh, J. 2021. Model-agnostic explanations using minimal forcing subsets. In *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

Jain, S.; Salman, H.; Wong, E.; Zhang, P.; Vineet, V.; Vemprala, S.; and Madry, A. 2022. Missingness bias in model debugging. *arXiv preprint arXiv:2204.08945*.

Jiang, P.-T.; Zhang, C.-B.; Hou, Q.; Cheng, M.-M.; and Wei, Y. 2021. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30: 5875–5888.

Joshi, A.; Kale, S.; Chandel, S.; and Pal, D. K. 2015. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4): 396.

Koh, P. W.; and Liang, P. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*, 1885–1894. PMLR.

Kondyli, V.; Suchan, J.; and Bhatt, M. 2022. Grounding Embodied Multimodal Interaction: Towards Behaviourally Established Semantic Foundations for Human-Centered AI. In *Proceedings of the 1st International Workshop on Knowledge Representation for Hybrid Intelligence (KR4HI 2022)*.

Lertvittayakumjorn, P.; and Toni, F. 2021. Explanation-based human debugging of nlp models: A survey. *Transactions of the Association for Computational Linguistics*, 9: 1508–1528.

Li, Y.; Meng, L.; Chen, L.; Yu, L.; Wu, D.; Zhou, Y.; and Xu, B. 2022. Training data debugging for the fairness of machine learning software. In *Proceedings of the 44th International Conference on Software Engineering*, 2215–2227.

Mahendran, A.; and Vedaldi, A. 2015. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5188–5196.

Menéndez, M. L.; Pardo, J. A.; Pardo, L.; and Pardo, M. d. C. 1997. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2): 307–318.

Milanés-Hermosilla, D.; Trujillo Codorniú, R.; López-Baracaldo, R.; Sagaró-Zamora, R.; Delisle-Rodriguez, D.; Villarejo-Mayor, J. J.; and Nunez-Alvarez, J. R. 2021. Monte carlo dropout for uncertainty estimation and motor imagery classification. *Sensors*, 21(21): 7241.

Pedreschi, D.; Giannotti, F.; Guidotti, R.; Monreale, A.; Ruggieri, S.; and Turini, F. 2019. Meaningful explanations of black box AI decision systems. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 9780–9784.

Rane, S.; Bruna, P. J.; Sucholutsky, I.; Kello, C.; and Griffiths, T. L. 2024. Concept alignment. *arXiv preprint arXiv:2401.08672*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD interna-*

tional conference on knowledge discovery and data mining, 1135–1144.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

Sharifi Noorian, S.; Qiu, S.; Gadiraju, U.; Yang, J.; and Bozzon, A. 2022. What Should You Know? A Human-In-the-Loop Approach to Unknown Unknowns Characterization in Image Recognition. In *Proceedings of the ACM Web Conference 2022*, 882–892.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Singla, S.; Chegini, A. M.; Moayeri, M.; and Feiz, S. 2022. Data-Centric Debugging: mitigating model failures via targeted data collection. *arXiv preprint arXiv:2211.09859*.

Sun, S.; Chen, W.; Wang, L.; Liu, X.; and Liu, T.-Y. 2016. On the depth of deep neural networks: A theoretical view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Teso, S.; and Kersting, K. 2019. Explanatory interactive machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 239–245.

Vendrow, J.; Jain, S.; Engstrom, L.; and Madry, A. 2023. Dataset interfaces: Diagnosing model failures using controllable counterfactual generation. *arXiv preprint arXiv:2302.07865*.

Wang, T.; Kim, S.; Wenxuan, J.; Xie, E.; Ge, C.; Chen, J.; Li, Z.; and Luo, P. 2024. Deepaccident: A motion and accident prediction benchmark for v2x autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5599–5606.

Whang, S. E.; Roh, Y.; Song, H.; and Lee, J.-G. 2023. Data collection and quality challenges in deep learning: A data-centric ai perspective. *The VLDB Journal*, 32(4): 791–813.

Wu, W.; Flokas, L.; Wu, E.; and Wang, J. 2020. Complaint-driven training data debugging for query 2.0. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 1317–1334.

Yao, H.; Chen, Y.; Ye, Q.; Jin, X.; and Ren, X. 2021. Refining language models with compositional explanations. *Advances in Neural Information Processing Systems*, 34: 8954–8967.

Ye, W.; Zheng, G.; Cao, X.; Ma, Y.; and Zhang, A. 2024. Spurious correlations in machine learning: A survey. *arXiv preprint arXiv:2402.12715*.

Zhou, B.; Lapedriza, A.; Xiao, J.; Torralba, A.; and Oliva, A. 2014. Learning deep features for scene recognition using places database. *Advances in neural information processing systems*, 27.