LADDER: Language-Driven Slice Discovery and Error Rectification in Vision Classifiers

Anonymous ACL submission

Abstract

Slice discovery refers to identifying systematic biases in the mistakes of pre-trained vision models. Current slice discovery methods in computer vision rely on converting input images into sets of attributes and then testing hypotheses about configurations of these precomputed attributes associated with elevated error patterns. However, such methods face several limitations: 1) they are restricted by the predefined attribute bank; 2) they lack the *common sense* reasoning and domain-specific knowledge often required for specialized fields *e.g.,* radiology; 3) at best, they can only identify biases in image attributes while overlooking those introduced during preprocessing or data preparation. We hypothesize that bias-inducing variables leave traces in the form of language (*e.g.,* logs), which can be captured as unstructured text. Thus, we introduce LADDER, which leverages the reasoning capabilities and latent domain knowledge of Large Language Models (LLMs) to generate hypotheses about these mistakes. Specifically, we project the internal activations of a pre-trained model into text using a retrieval approach and prompt the LLM to propose potential bias hypotheses. To detect biases from preprocessing pipelines, we convert the preprocessing data into text and prompt the LLM. Finally, LADDER generates pseudo-labels for each identified bias, thereby mitigating all biases without requiring expensive attribute annotations. Rigorous evaluations on 3 natural and 3 medical imaging datasets, 200+ classifiers, and 4 LLMs with varied architectures and pretraining strategies – demonstrate that LADDER consistently outperforms current methods. Code is available: https://github.com/annonymous-vision/ACL.

1 Introduction

Error slices are data subsets on which vision classifiers systematically fail. Discovering such slices is critical for improving model robustness. Identifying such slices is challenging in vision classifers where biases are pervasive and can be traced through textual artifacts such as image captions, metadata, and medical imaging headers *e.g.,* DICOMs. However, their unstructured nature makes manual analysis impractical. Natural language, with its inherent flexibility, offers a powerful tool for capturing subtle biases beyond predefined attribute sets. LLMs, equipped with advanced reasoning capabilities and latent domain knowledge, excel at analyzing such free-form text to detect complex relationships and domain-specific biases. However, existing slice discovery methods often rely on predefined attribute banks or unsupervised clustering, both of which lack the reasoning ability to identify nuanced and domain-specific biases. This paper proposes LADDER, that leverages LLMs to systematically identify and mitigate error slices in vision classifiers by analyzing captions, metadata, and beyond – without relying on fixed attribute sets or clustering methods.

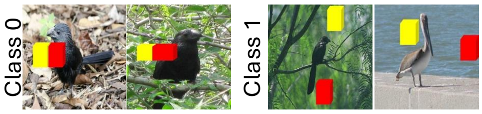


Figure 1: Synthetic dataset containing Class 0 images consistently with a yellow box to the left of a red box, while Class 1 images have boxes placed randomly. Captions encode the spatial bias, used by LADDER for slice discovery.

Prior slice discovery methods *e.g.,* DrML (Zhang et al., 2023) use text encoders to mitigate biases in CLIP by closing the modality gap through cross-modal transfer, which limits their applicability to non-multimodal models. Plus, DrML relies on user-defined prompts with fixed attribute sets, introducing human bias into the mitigation process. Similarly, Facts (Yenamandra et al., 2023) amplifies the spuriousness in the initial training stage by setting large weight decay, deviating from standard supervised learning practices. Methods like

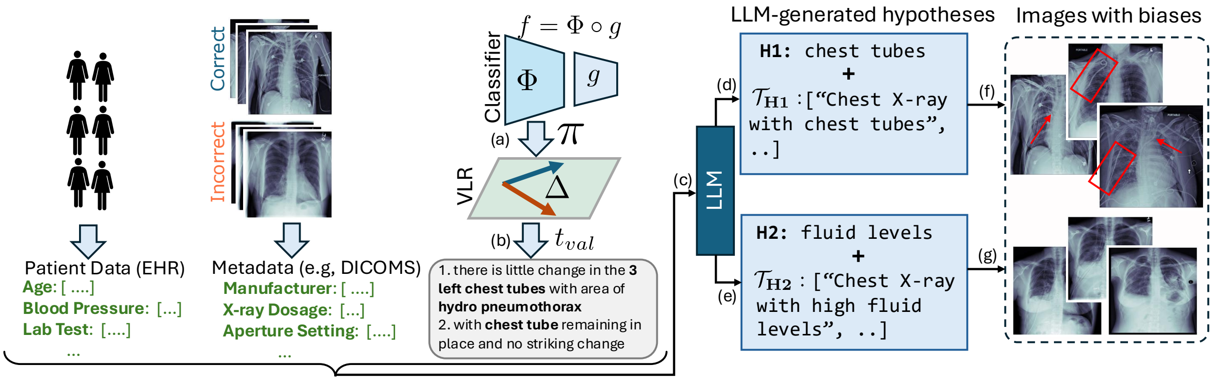


Figure 2: Schematic of LADDER. **(a)**: Projection (π) of model representation (Φ) to VLR space. **(b)**: Retrieval of `topK` sentences based on the image embeddings difference (Δ) of correct and incorrect groups in VLR space. **(c)**: LLM is invoked with `topK` sentences/other metadata. **(d-e)**: LLM generated hypotheses ($\{\mathcal{H}, \mathcal{T}\}$). **(f-g)**: Finding the clusters faithful to the hypotheses. In red, we highlight the chest tubes (ground truth bias for NIH) in this example.

Domino (Eyuboglu et al., 2022) and Facts discover slices by clustering samples with similar attributes within the vision-language representation (VLR) space. However, the slices often exhibit semantic inconsistencies – attributes within slices lack coherence, leading to unreliable interpretations of model errors. PRIME (Rezaei et al., 2023) relies on expensive tagging models, limited to detecting the presence/absence of a fixed set of attributes. All these methods lack the reasoning capabilities and domain knowledge required to capture complex error patterns, limiting their effectiveness in specialized tasks. Also, their dependence on pre-existing semantic labels (*e.g.,* visual tags) hinders the detection of biases in the metadata or domain-specific fields such as DICOM headers.

Prior mitigation methods (Sagawa et al., 2020; Liu et al., 2021; Kirichenko et al., 2022) rely on expensive and incomplete attributes. While they improve worst group accuracy (WGA), they amplify errors in other groups (Li et al., 2023b). Although Li et al. (2023b) addresses errors across multiple biases, it assumes prior knowledge of the number and types of biases to design specific data augmentations. This reveals a critical gap: the need for an automated method to discover and mitigate multiple biases without prior knowledge/annotations.

This paper proposes LADDER with the following contributions: **1. Using language for error slice discovery:** LADDER uses image captions/radiology reports to retrieve sentences indicative of model errors, utilizing the flexibility of natural language to capture deeper insights beyond the simple presence or absence of attributes, unlike tagging models. **2. Using LLMs' reasoning capabilities and latent domain knowledge:** To identify biases, LADDER leverages LLMs' advanced reasoning to generate testable hypotheses from these sentences, unlike traditional methods. For instance, in a synthetic dataset (Appendix A.11), where Class 0 images consistently feature a yellow box to the left of a red box (Fig 1), the classifier exhibits poor performance on test data without this bias. LADDER correctly identifies this reliance on spatial positioning by analyzing textual descriptions through LLM (Fig 11). Note, LLM in LADDER processes only text inputs without images (total cost of ~$28). In medical images, LADDER uses LLMs' domain knowledge to identify fine-grained biases, including disease subtypes and pathological patterns. **3. Slice discovery from any off-the-shelf model:** It detects slices from any supervised model, regardless of architecture/pretraining, overcoming specific training requirements of Facts and DrML. **4. Detecting biases beyond captions:** LADDER uses LLM to analyze metadata, such as Electronic Health Records (EHR) or DICOM headers, discovering biases beyond captions. **5. Mitigating multiple biases w/o any annotation:** LADDER mitigates biases by generating pseudo-labels for each hypothesis and fine-tuning the classifier's linear head through attribute rebalancing. By ensembling debiased model predictions, LADDER corrects multiple biases without requiring attribute annotations/prior knowledge of their number and type. **Additionally,** we explore the use of instruction-tuning models (*e.g.,* LLaVA) in applicable domains to reduce LADDER's reliance on captions. Rigorous evaluations on 6 datasets with 200+ classifiers and 4 LLMs across architectures and pretraining strategies show that LADDER outperforms slice discovery and mitigation baselines.

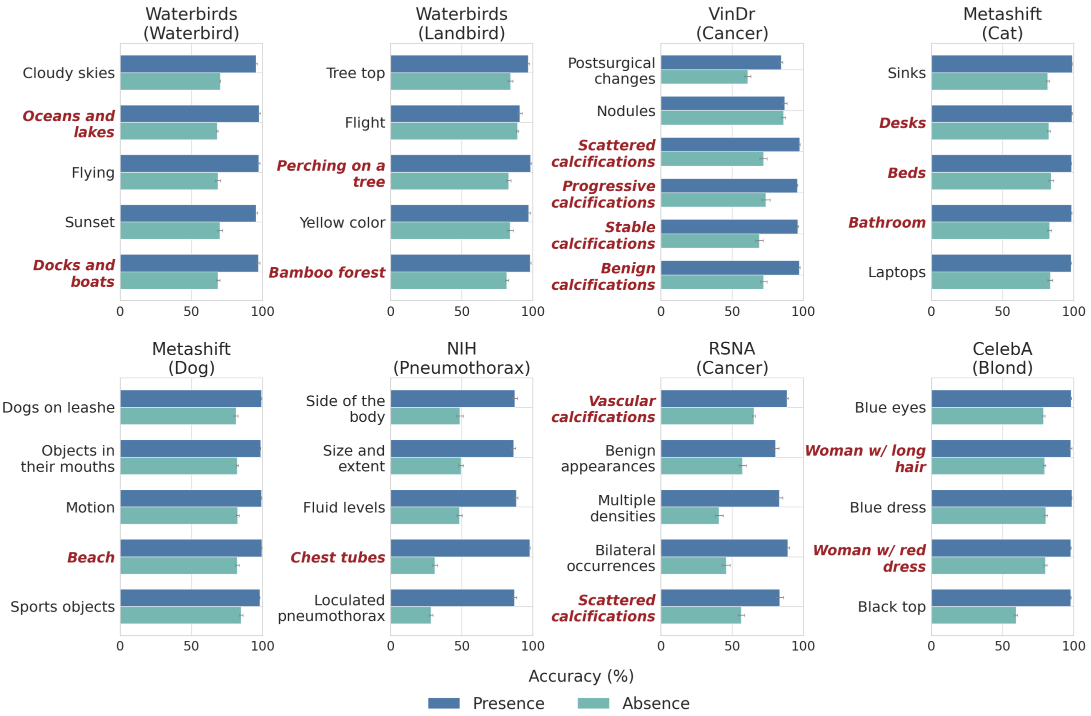


Figure 3: Bias identification by LADDER in RN Sup IN1k classifier. Each panel shows the classifier's performance for a specific dataset (RSNA) and class label (Cancer) when biased attributes in the identified hypotheses are present/absent. Hypotheses indicative of ground truth biases (*e.g.,* water for waterbirds) are shown in red.

2 Related Work

Slice discovery. Initial methods (d'Eon et al., 2022; Sohoni et al., 2020; Kim et al., 2019; Singla et al., 2021) on slice discovery utilize dimensionality reduction, lacking comprehensive evaluation. Recent methods *e.g.,* Domino (Eyuboglu et al., 2022) projects data into VLR space, identifies slices via a mixture model, and captions them. Facts (Yenamandra et al., 2023) amplifies spurious correlations in the initial training phase by increasing weight decay and discovering slices in VLR space. Both approaches compromise visual semantics, resulting in attribute inconsistencies within slices. DrML (Zhang et al., 2023) probes only CLIP-based classifiers using modality gap geometry and user-defined prompts, introducing potential human biases. Also, Facts and DrML are restricted to specific training setups, limiting generalizability to standard ERM classifiers. PRIME (Rezaei et al., 2023) uses expensive tagging models to discover attributes for slice discovery. HiBug (Chen et al., 2024a) prompts LLM to suggest biases for model errors without any textual context from the data. Thus, it results in superficial keyword-based attributes derived purely from general user prompts, lacking the deeper contextual grounding needed for bias detection. Recently, OpenBias (D'Incà et al., 2024) detects biases in T2I models via LLM-driven keyword queries but is not designed for posthoc classifier error analysis. B2T (Kim et al., 2024) extracts keywords from captions. All these methods are limited by incomplete tags or keyword-based attributes and lack reasoning or latent *domain knowledge*, essential in fields *e.g.,* radiology. **Bias mitigation.** Mitigation methods *e.g.,* GroupDRO (Sagawa et al., 2020) optimizes for worst-performing groups, while JTT (Liu et al., 2021) reweights minority groups. DFR (Kirichenko et al., 2022) retrains the final layer using a balanced validation set. All of them require group annotations and focus on mitigating errors in the worst-performing group, amplifying errors in other subgroups. Li et al. (2023b) mitigates multiple biases using an ensemble-based approach but relies on predefined bias types, which limits its adaptability to unknown biases. LADDER overcomes all these limitations. For discovery, LADDER incorporates the *domain knowledge* of LLMs, reason about model errors, and generates hypotheses identifying biases from any pretrained model without external attributes, unlike existing methods. For mitigation, LADDER leverages pseudo-labels for each bias to finetune the classifier's last layer – without any group annotations, predefined bias types, or human intervention.

3 Method

Assume the classifier $f = g \circ \Phi$ is trained using ERM to predict the labels $\mathcal{Y}$ from the images $\mathcal{X}$, where Φ and g are the representation and classification head, respectively. $\{\Psi^I, \Psi^T\}$ denote the image and text encoders of the joint VLR space. For a set of images $\mathcal{X}_Y$ of a class $Y \in \mathcal{Y}$, LADDER identifies error slices where f underperforms and mitigates it. Throughout the paper, $\langle \cdot, \cdot \rangle$ denotes the dot product to estimate the similarity between two representations. Fig. 2 shows the schematic of LADDER. We do not rely on sample-specific paired annotations, human-generated prompts, or prior knowledge of bias types or their numbers. We utilize a text corpus t_{val} from radiology reports or image captions from the validation dataset to discover and mitigate errors. **Error slice.** An error slice for a class Y includes subsets $\mathcal{X}_Y$ where the model performs significantly worse than its overall performance on the entire class Y, formally defined as: $\mathbb{S}_Y = \{\mathcal{S}_{Y,\neg\texttt{attr}} \subseteq \mathcal{X}_Y | e(\mathcal{S}_{Y,\neg\texttt{attr}}) \gg e(\mathcal{X}_Y), \exists \texttt{attr}\}$, where $e(\cdot)$ is the error rate on the specific data subset and $\mathcal{S}_{Y,\neg\texttt{attr}}$ denotes the subset of $\mathcal{X}_Y$ without the attribute `attr`. Alternatively, f is biased on the attribute `attr`, resulting in better performance on the subpopulation with `attr` *e.g.,* error rate in pneumothorax patients w/o chest tubes is higher than overall pneumothorax patients (Docquier and Rapoport, 2012).

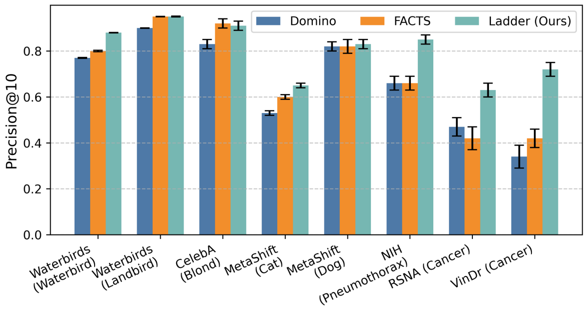


Figure 4: Precision@10 for CNN models (f) quantifying slice discovery. LADDER outperforms the baselines, especially for medical imaging datasets.

3.1 Retrieving Sentences Indicative of Biases

First, for a particular class, LADDER retrieves the sentences that describe the visual attributes contributing to correct classifications but missing in misclassified ones, leading to model errors. Following Moayeri et al. (2023), it learns a projection function $\pi : \Phi \rightarrow \Psi^I$ (Appendix A.4) to align the representation of the classifier, Φ, with the image representation Ψ^I of the VLR space. Then, for a class label Y, we estimate the difference in mean of the projected representations of the correct and misclassified samples as $\Delta^I = \mathbb{E}_{X,Y|f(X)=Y}[\pi(\Phi(X)] - \mathbb{E}_{X,Y|f(X)\neq Y}[\pi(\Phi(X)]$. Assuming the mean representations preserve semantics, this difference captures key attributes contributing to correct classifications but are poorly captured or misrepresented in misclassified ones. Denoting the text embedding of t_{val} as $\Psi^T(t_{val})$, we retrieve the `topK` sentences as: $\texttt{topK} = \mathscr{R}\big(\langle \Delta^I, \Psi^T(t_{val}) \rangle, t_{val}\big)$, where $\mathscr{R}$ is a retrieval function retrieving `topK` sentences from the text corpus having the highest similarity score with the mean difference of the projected image representations. Next, the LLM analyzes the sentences and constructs hypotheses to find error slices.

3.2 Discovering Error Slices via LLM

Generating hypothesis. To form the set of hypotheses, LADDER invokes an LLM with the `topK` sentences. Formally, $\{\mathcal{H}, \mathcal{T}\} = \texttt{LLM}\big(\texttt{topK}\big)$, where $\mathcal{H}$ is a set of hypotheses with attributes that f may be biased and $\mathcal{T}$ is a set of sentences to be used to test each hypothesis. f underperforms on the subpopulation without the attributes in $\mathcal{H}$. Each hypothesis $H \in \mathcal{H}$ is paired with $\mathcal{T}_H \in \mathcal{T}$, a set of sentences that provide diverse contextual descriptions of the hypothesis-specific attribute as it appears in various images. Representations of images with the attribute specified in H, are highly similar to the mean text embedding of $\mathcal{T}_H$. Refer to Appendix A.7 for the prompt used by LLM to generate the hypothesis. **Identifying error slices.** For each hypothesis $H \in \mathcal{H}$, we first compute the mean embedding of the set of sentences $\mathcal{T}_H$ as $\Psi^T(\mathcal{T}_H) = \frac{1}{|\mathcal{T}_H|} \sum_{t \in \mathcal{T}_H} \Psi^T(t)$. For an image $X \in \mathcal{X}_Y$, we obtain the projected representation $\pi(\Phi(X))$ in VLR space and compute the similarity score, $s_H(X) = \langle \pi(\Phi(X)), \Psi^T(\mathcal{T}_H) \rangle$ Finally, for a class label Y, we retrieve images with similarity scores below a threshold τ as $\mathcal{S}_{Y,\neg H} = \{X \in \mathcal{X}_Y | s_H(X) < \tau\}$. The hypothesis H fails in these images as they lack the attribute specified in the H. The subset $\mathcal{S}_{Y,\neg H}$ may be a potential error slice if the error $e(\mathcal{S}_{Y,\neg H})$ is greater than $\mathcal{X}_Y$. Formally, $\hat{\mathbb{S}}_Y$, the predicted slice for a class Y is: $\hat{\mathbb{S}}_Y = \{\mathcal{S}_{Y,\neg H} \subseteq \mathcal{X}_Y | e(\mathcal{S}_{Y,\neg H}) \gg e(\mathcal{X}_Y), \exists H \in \mathcal{H}\}$

3.3 Mitigate Multi-bias w/o Annotation

For the attributes linked to a hypothesis, LADDER treats s_H as a logit and converts it to a probability. If the probability exceeds a threshold (0.5 in all experiments), LADDER assigns a pseudo-label 1 to the attribute and 0 otherwise. Thus, it generates pseudo-labels for all relevant attributes, enabling error mitigation without annotations. To do so, LADDER adopts an ensemble-based strategy. Following DFR, we create a balanced dataset from a held-out validation set, for each pseudo-labeled attribute per hypothesis. We then fine-tune the classification head g using this balanced dataset, producing a debiased model per hypothesis. During inference, we again compute the similarity score s_H for all hypotheses and select the classifier head g_{H^*} associated with the hypothesis having maximum similarity: $H^* = \arg\max_{H \in \mathcal{H}} s_H(X)$.

4 Experiments

We perform experiments to answer the research questions: **RQ1.** How does LADDER perform in discovering error slices compared to baselines? **RQ2.** How does LADDER leverage reasoning and latent domain knowledge of LLMs for slice discovery? **RQ3.** How does LADDER discover biased attributes with different architectures and pretraining methods? **RQ4.** How does LADDER mitigate biases using the discovered attributes? **RQ5.** Can LADDER operate w/o captions? **RQ6.** Can LADDER detect biases beyond captions/reports?

Datasets. We evaluate LADDER on 6 datasets (AppendixA.1 for details): 1) **Waterbirds** (Wah et al., 2011): bird classification where background correlates with bird type. 2) **CelebA** (Liu et al., 2018): blond hair classification with gender as a spurious feature. 3) **MetaShift**: cat vs. dog classification with background correlation. 4) **NIH Chest-X-ray (CXR)** (Wang et al., 2017): pneumothorax detection with chest tubes as a shortcut (Docquier and Rapoport, 2012). 5) **RSNA-Mammo** and 6) **VinDr-Mammo** (Nguyen et al., 2023): breast cancer and abnormality detection from mammograms, with calcifications as a shortcut (Wen et al., 2024).

Experimental details. For natural images and CXRs, we use an ImageNet1k (IN1k)-initialized ResNet50 (RN Sup IN1k) as the model f that LADDER aims to probe, trained with a standard supervised loss. For mammograms, we use EfficientNet-B5 (EN-B5) as f. For the text corpus (t_{val}), we use BLIP-captioner (Li et al., 2022), radiology reports from MIMIC-CXR (Johnson et al., 2019) and the radiology texts from Mammo-FActOR (Ghosh et al., 2024) for natural images, CXRs and mammograms, respectively. For VLR space ($\{\Psi^I, \Psi^T\}$), we use CLIP (Radford et al., 2021), CXR-CLIP (You et al., 2023), and Mammo-CLIP (Ghosh et al., 2024) for natural images, CXR and mammograms, respectively. We use 200 and 100 sentences as `topK` for natural and medical images (CXR and mammo). We use GPT-4o (Wu et al., 2024) as the LLM. Error slices are defined as subsets where the error rate exceeds the overall class error by at least 10%. Refer to Appendix A.10 for further experimental details. All reported results are obtained from experiments conducted over 3 random seeds.

Baselines. For slice discovery, we compare LADDER with Domino and Facts (Appendix A.2). For mitigation, we compare with the baselines, including ERM (Vapnik, 1999), GroupDRO (Sagawa et al., 2020), JTT (Liu et al., 2021), DFR (Guo et al., 2019), CVaRDRO (Duchi and Namkoong, 2021) and LfF (Nam et al., 2020) (Appendix A.3).

Evaluation metrics. We use `Precision@10` (Appendix A.5) (Eyuboglu et al., 2022) to evaluate the slice discovery methods and the CLIP score (Kim et al., 2024) to quantify the effect of biased attributes. For mitigation, we report Worst Group Accuracy (WGA) for mitigation for natural images. We report mean AUROC and WGA for medical images, where WGA refers to model performance on pneumothorax patients w/o chest tubes (NIH) and cancer or abnormal patients w/o calcifications (RSNA & VinDr-Mammo).

5 Results

RQ1: Comparison of LADDER with slice discovery baselines. Following (Eyuboglu et al., 2022; Yenamandra et al., 2023), Fig. 4 compares the `Precision@10` of different slice discovery methods for CNN models (EN-B5 for mammograms & RN Sup IN1k for others). For medical images, LADDER achieves a substantial **50%** improvement over the baselines. Refer to Fig. 12 in Appendix A.12.1 for WGA evaluation using the slices discovered from Domino, Facts, and LADDER with our ensemble-based mitigation strategy. In all the experiments, LADDER outperforms the baselines. Facts and Domino cluster the images by projecting them directly into VLR space, often leading to incoherent slices. In contrast, LADDER first projects the model's representation into the VLR space, preserving the nuanced semantics of the classifier features. Instead of relying solely on unsupervised clustering, it leverages the reasoning capabilities of LLMs

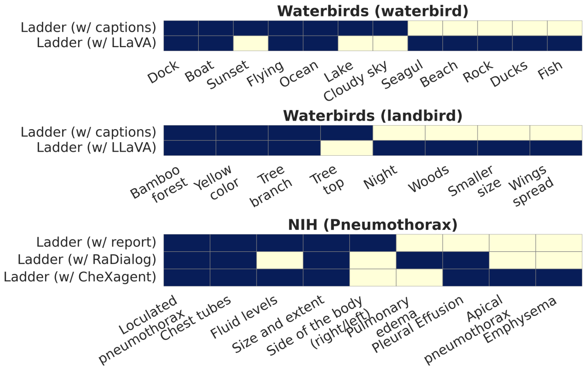


Figure 5: Biased attributes detected by LADDER w/ captions and w/ instruction-tuned models (w/o captions). Bright/light colors show presence/absence of attributes

and signals from the captions/radiology reports to identify the coherent-biased attributes within the discovered slices. Next, we assign pseudo-labels to the attributes using similarity scores ($s_H(X)$). The coherent slices produced by LADDER ensure that the pseudo-labeling process is more accurate than the baselines leading to superior bias mitigation performance (Appendix A.12.1).

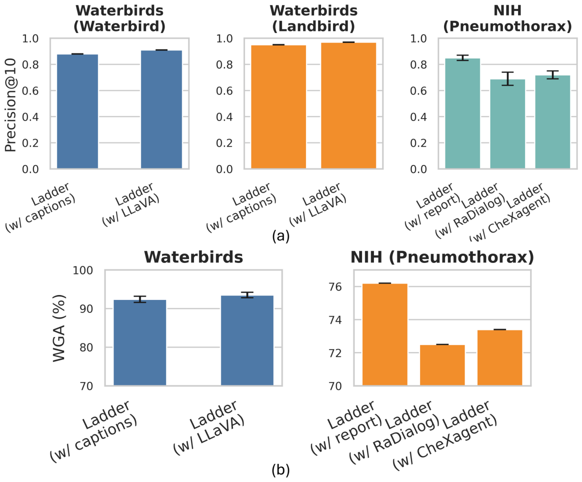


Figure 6: (a) Precision@10 for slice discovery and (b) WGA for bias mitigation using LADDER w/ captions vs. instruction-tuned models.

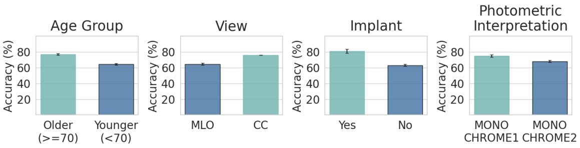
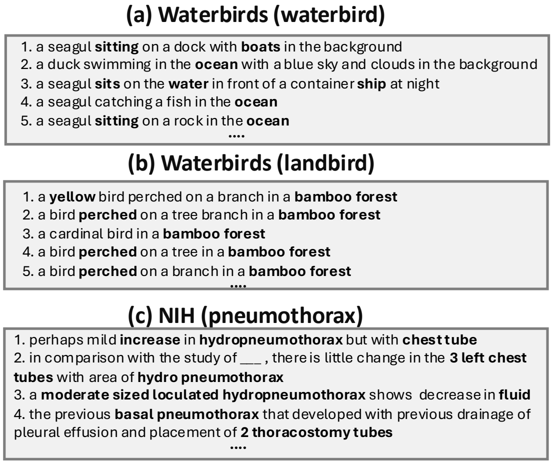


Figure 7: LADDER detects biases beyond reports, identifying biases from metadata (age, view and implant) and DICOM headers (Photometric interpretation).



(a) Waterbirds (waterbird)

1. a seagul **sitting** on a dock with **boats** in the background
2. a duck swimming in the **ocean** with a blue sky and clouds in the background
3. a seagul **sits** on the **water** in front of a container **ship** at night
4. a seagul catching a fish in the **ocean**
5. a seagul **sitting** on a rock in the **ocean**

....

(b) Waterbirds (landbird)

1. a **yellow** bird perched on a branch in a **bamboo forest**
2. a bird **perched** on a tree branch in a **bamboo forest**
3. a cardinal bird in a **bamboo forest**
4. a bird **perched** on a tree in a **bamboo forest**
5. a bird **perched** on a branch in a **bamboo forest**

....

(c) NIH (pneumothorax)

1. perhaps mild **increase** in **hydropneumothorax** but with **chest tube**
2. in comparison with the study of ___ , there is little change in the **3 left chest tubes** with area of **hydro pneumothorax**
3. a **moderate sized loculated hydropneumothorax** shows decrease in **fluid**
4. the previous **basal pneumothorax** that developed with previous drainage of pleural effusion and placement of **2 thoracostomy tubes**

....

Figure 8: Sentences retrieved by LADDER in Sec. 3.1 encoding model biases (in bold) for LLM to analyze. Each panel denotes a class label of a specific dataset.

RQ2. Leveraging LLM's reasoning and domain knowledge for bias discovery. Fig. 8 displays the sentences retrieved by LADDER indicating the different model biases. Fig. 3 shows the biased attributes discovered by LADDER. The presence of these attributes correlates with f's performance, while their absence results in error slices where f's performance drops. Recall, LADDER uses LLM to generate hypotheses from the sentences, indicative of biases. The similarity score ($s_H(X)$) tests these hypotheses to validate if the absence of specific attributes linked to each hypothesis results in a drop in f's performance. For *e.g.,* waterbirds `flying` vs. `not flying` achieve 97.3% vs. 68.6% accuracy. In NIH, pneumothorax patients with and without `chest tubes` achieve an accuracy of ∼98%, compared to 31%. For all tasks, LADDER effectively detects ground truth biases. In the Waterbirds dataset, LADDER identifies diverse water-related biases such as `boat` and `lake`. Also, Fig.3 reports that LADDER identifies domain-specific biases (*e.g.,* `chest tubes`, `loculated pneumothorax` for NIH; subtypes of `calcifications` for RSNA & VinDr Mammo), capturing a more granular characterization of biases. Unlike the keyword extraction or tagging models, which struggle with missing or insufficient attributes, LADDER leverages LLM-driven latent medical knowledge to generate comprehensive hypotheses. Such fine-grained detection of contextual biases, including subtypes, allows LADDER to for the detection of patterns that would be difficult to detect without domain expertise. Refer to Appendix A.12.3, A.12.2 and A.12.6 for detailed qualitative results, the hypotheses closest to the ground truth biases, and the influence of biased attributes via CLIP score, respectively.

RQ3: Biased attributes discovery across

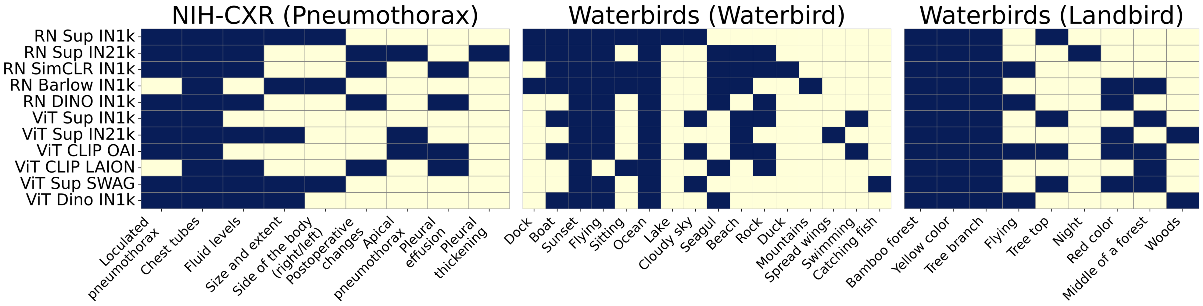


Figure 9: Biased attributes discovered by LADDER show consistent biases across architectures and pretraining. Several attributes (*e.g.,* ocean, lake, beach etc.) represent the same visual concepts (water bodies) denoting the groundtruth bias. Bright and light colors indicate attribute presence and absence, respectively.

Table 1: Impact of captioners on LADDER's performance for RN Sup IN1k classifier. Though GPT-4o is expensive, its quality is better than others.

	Waterbirds		CelebA	
Method	Mean Acc	WGA	Mean Acc	WGA
BLIP (Li et al., 2022)	93.1	91.4	89.8	88.9
BLIP2 (Li et al., 2023a)	93.3	91.6	89.8	89.2
ClipCap (Mokady et al., 2021)	93.7	91.8	88.3	87.4
GPT-4o (Wu et al., 2024)	**94.2**	**93.1**	**91.4**	**90.3**

architectures/pre-taining methods. In this setup, we extract biases using LADDER on a range of model architectures (both ResNet50 and ViT), initializing f (the model to be probed) with diverse pretraining methods, including SimCLR (Chen et al., 2020), Barlow Twins (Zbontar et al., 2021), DINO (Caron et al., 2021), and CLIP (Radford et al., 2021). These methods are pretrained on datasets *e.g.,* ImageNet-1K (IN1k) (Deng et al., 2009), ImageNet-21K (IN21k) (Ridnik et al., 2021), SWAG (Singh et al., 2022), LAION-2B (Schuhmann et al., 2022), and OpenAI-CLIP (OAI) (Radford et al., 2021). Yang et al. (2023) shows that every ERM-trained classifier (f) exhibits low WGA irrespective of architecture/pre-training due to consistently learning similar biases. Figure 9 shows that LADDER, leveraging LLM-driven reasoning and *domain knowledge*, consistently identifies similar biases across different architectures, pretraining methods, and datasets. In the NIH dataset, LADDER identifies mostly key attributes such as `chest tubes`, `fluid levels` etc. Also, in the Waterbirds dataset, LADDER detects attributes *e.g.,* `ocean` and `bamboo forest` consistently, showing the correlation of the spurious backgrounds with class labels and the ground truth biases. Appendix A.12.8 lists more results.

RQ4: Mitigating biases using LADDER. Tab. 2 shows that LADDER outperforms other bias mitigation baselines in estimating WGA, without requiring the expensive ground truth shortcut attributes, for both training and validation datasets across CNN models (EN-B5 for Mammograms and RN Sup IN1k for the rest). LADDER achieves a WGA of 91.4%, 76.4% and 82.5% – a 3.6%, 7.3% and 21.1% improvement (↑) over DFR in the Waterbirds, RSNA, and VinDr datasets, respectively. For NIH, LADDER outperforms JTT and DFR by 8.2% and 7.4%, respectively. Appendix A.12.4 illustrates further analysis with an additional 9 baselines. Fig. 15 shows LADDER's consistent performance gain across various architectures and pre-training methods. Tab. 11 in Appendix A.12.7 shows that LADDER outperforms Li et al. (2023b) on multi-shortcut benchmark UrbanCars. Leveraging LLMs' advanced reasoning, LADDER accurately derives pseudo labels for the biased attributes from hypotheses to identify true model biases. LADDER then applies targeted bias mitigation by fine-tuning the last layer, resulting in a systematic debiased model per hypothesis. This efficient strategy effectively enhances model performance across the biases, modalities, and architectures.

Table 2: Error mitigation results (WGA) for EN-B5 for mammograms and RN Sup IN1k for the rest. We boldface and underline the best and second-best results. We compare with 9 additional baselines in A.12.4.

Method	Waterbirds	CelebA	NIH	RSNA	VinDr
ERM	69.1$_{\pm 1.2}$	62.2$_{\pm 1.5}$	60.3$_{\pm 0.0}$	69.8$_{\pm 0.0}$	45.6$_{\pm 0.0}$
JTT	84.5$_{\pm 0.3}$	87.2$_{\pm 7.5}$	70.4$_{\pm 0.0}$	68.5$_{\pm 0.0}$	66.1$_{\pm 0.0}$
GroupDRO	87.1$_{\pm 1.3}$	<u>88.1</u>$_{\pm 0.7}$	71.1$_{\pm 0.0}$	<u>72.3</u>$_{\pm 0.0}$	67.1$_{\pm 0.0}$
CVaRDRO	85.4$_{\pm 2.3}$	83.1$_{\pm 1.5}$	<u>71.3</u>$_{\pm 0.0}$	71.7$_{\pm 0.0}$	67.1$_{\pm 0.0}$
LfF	75.2$_{\pm 0.7}$	63.0$_{\pm 4.4}$	61.6$_{\pm 0.0}$	66.4$_{\pm 0.0}$	64.5$_{\pm 0.0}$
DFR	<u>88.2</u>$_{\pm 0.3}$	87.1$_{\pm 1.1}$	70.5$_{\pm 0.0}$	71.2$_{\pm 0.0}$	<u>68.1</u>$_{\pm 0.0}$
LADDER	**91.4**$_{\pm 0.8}$	**88.9**$_{\pm 0.4}$	**76.2**$_{\pm 0.0}$	**76.4**$_{\pm 0.0}$	**82.5**$_{\pm 0.0}$

RQ5: Relaxing the dependency on captions. To reduce LADDER 's reliance on captions/reports, we leverage instruction-tuned models to generate tex-

tual descriptions for the correctly classified samples. Specifically, we use LLaVA-1.5 7B (Liu et al., 2024) for natural images and RaDialog (Pellegrini et al., 2023) and cheXagent (Chen et al., 2024b) for CXRs to probe RN Sup IN1k classifier. Refer to Appendix A.8 for the utilized prompts. LADDER ’s LLM pipeline utilizes these generated descriptions to identify biased attributes. Recall we aim to detect biases consistently present in correctly classified instances. Figure 5 compares the biases identified using LADDER ’s retrieval pipeline (captions/reports) vs. those detected via instruction-tuned models. Figure 6(a) compares Precision@10 for LADDER under both settings, while Figure 6(b) evaluates the WGA metric, evaluating the bias discovery and mitigation quantitatively, respectively. For natural images, LADDER with instruction-tuned models perform comparably to the standard pipeline using captions. For CXRs, the retrieval-based approach utilizing actual reports outperforms methods using cheXagent and RaDialog, highlighting the importance of domain-specific reports in medical imaging. Thus, using models *e.g.,* LLaVA can eliminate LADDER’s need for captions. However, this approach is challenging for 2D mammograms and dermatology imaging (Alzubaidi et al., 2021) etc.where robust instruction-tuned models are lacking. In such cases, LADDER’s retrieval pipeline remains highly adaptable and shows broad applicability. Thus, a trade-off emerges: models can either leverage explicit radiology reports for bias identification or develop robust VLRs to reduce dependence on reports.

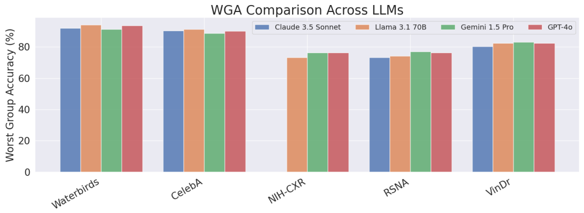


Figure 10: WGA comparison across different LLMs for bias mitigation by LADDER with RN Sup IN1k for natural images and CXRs, and EN-B5 for mammograms. GPT-4o and Gemini excel in medical imaging tasks.

RQ6: Detecting biases beyond captions/reports. While prior work (Boyd et al., 2023) highlights biases in EHR and medical imaging headers (*e.g.,* DICOMs), LADDER extends bias detection beyond captions. We use metadata from the RSNA-Mammo dataset, which includes metadata *e.g.,* BIRADS(0-2), age, implant, view (CC or MLO), laterality (left or right breast), machine_id, and site_id. Also, the DICOM headers provide attributes *e.g.,* photometric interpretation, VOI LUT, and pixel intensity relationships. We probe the same EN-B5 classifier to find attributes consistently present in correctly classified samples, whose absence results in a performance drop. By listing each sample’s metadata to a Python dictionary (refer to Appendix A.9) and using LADDER ’s LLM pipeline (Sec. 3.2), we generate hypotheses about the biased attributes; we then validate their impact on the classifier’s performance based on the presence/absence of these attributes, with their ground truth values from the metadata. Figure 7 shows that LADDER detects an age bias (a 19.5% accuracy gap for patients aged 70+ vs. the rest) and a 10% gap to different photometric interpretations (Monochrome 1 vs. Monochrome 2). This finding aligns with existing evidence of age bias in oncology (Tasci et al., 2022). Existing methods lack LLM-based reasoning, limiting them to fixed attributes or clustering, while LADDER uses LLMs to reason across metadata for comprehensive analysis.

6 Ablations and Additional Results

Table 1 compares LADDER’s performance across different captioning methods, while Fig.10 presents the WGA of LADDER for various LLMs. Due to space constraints, we provide detailed analyses in Appendices A.12.10 and A.12.12. Additionally, Appendices A.12.11, A.12.13 and A.12.14 include ablation studies on slices discovered using different LLMs, their computational costs, and the impact of different VLRs on LADDER. Appendix A.12.9 demonstrates LADDER’s ability to identify biases in the ImageNet dataset (multiclass classification), while Appendix A.12.5 shows how these identified attributes improve CLIP’s zero-shot accuracy.

7 Conclusion

We introduce LADDER, a novel LLM-driven method for error slice discovery and bias mitigation for vision classifiers. Unlike prior methods that rely on predefined attributes or unsupervised clustering, LADDER leverages LLM’s reasoning to detect coherent error slices without requiring explicit annotations from any off-the-shelf pretrained classifier. Next, it mitigates multiple biases through pseudo-label generation and attribute rebalancing. Extensive evaluations on 6 datasets show LADDER ’s effectiveness, outperforming existing baselines.

Limitations

While LADDER demonstrates superior performance in bias discovery and mitigation, we outline the limitations of our work and potential areas for improvement: **1. Dependence on captions for bias discovery:** LADDER primarily relies on captions to identify biases, which may not be suitable for domains with sparse or limited textual descriptions. While we introduce a workaround using instruction-tuned models *e.g.,* LLaVA for specific applications, future research will explore reducing language dependence across broader domains. **2. Potential bias in pretrained models:** LADDER utilizes pretrained models such as CLIP and LLMs, which inherently reflect biases present in their training data. This dependency may influence the bias discovery process and potentially undermine fairness objectives. Addressing and mitigating these inherent biases in foundational models is an important direction for future research. **3. Lack of human oversight in bias discovery:** To prevent the introduction of additional bias, LADDER automates the discovery phase without human intervention. Instead, domain experts (*e.g.,* clinicians) validate the identified biases prior to mitigation. While this strategy minimizes human-induced bias during discovery, it introduces subjectivity in the validation phase. Enhancing and standardizing this validation process remains a key focus for future work.

Ethical Considerations

We strongly adhere to ethical standards in the handling of medical data, the use of language models, and the implementation of machine learning methods. We provide the following details: **1. Medical datasets:** All medical datasets used in this study, including MIMIC-CXR, RSNA-Mammo, and VinDr-Mammo, are anonymized and publicly available. We strictly follow the respective data-use agreements and ethical guidelines associated with each dataset. **2. Language models for medical tasks:** The large language models (LLMs) employed for medical applications adhere to the guidelines established for MIMIC[1]. Specifically, we use GPT-4o (Wu et al., 2024) via Azure OpenAI service as LLM for NIH in the main experiments. For ablations, we use Google's Gemini via Vertex AI. For LLaMA, we set up the model on a local machine. No information from NIH datasets was processed using language models not covered by these guidelines, such as Claude. **3. Classifier models and codebase:** All classifiers used in this research are standard architectures and publicly available models, ensuring reproducibility and transparency. We list them in detail in Appendix A.10. **4. Vision-Language representations (VLRs):** All VLRs utilized in this study are publicly available, and we list the corresponding resources in Appendix A.10. We adhere strictly to the license terms specified by the creators of these resources.

Broader Impact

The development and deployment of LADDER have potential implications for AI applications in medical and general computer vision tasks. We outline the broader impacts as follows: **1. Medical applications and patient outcomes:** LADDER can improve the robustness and interpretability of vision models in medical imaging. By identifying and mitigating biases, it can lead to more reliable diagnostic tools, ultimately enhancing patient care and reducing diagnostic disparities. **2. Bias detection and fairness:** LADDER offers a generalizable approach to uncovering and addressing systematic biases across datasets. This can contribute to the development of fairer AI models, particularly in domains prone to dataset biases, such as healthcare and social applications. **3. Continuous auditing and bias mitigation:** LADDER can act as an auditor for any pretrained network in a continuous manner. By running it on a dataset, it can identify and mitigate biases using language. Whenever a bias can be traced in language, LADDER can detect it with its superior reasoning capabilities and domain knowledge.

References

Laith Alzubaidi, Muthana Al-Amidie, Ahmed Al-Asadi, Amjad J Humaidi, Omran Al-Shamma, Mohammed A Fadhel, Jinglan Zhang, Jesus Santamaría, and Ye Duan. 2021. Novel transfer learning approach for medical imaging with limited labeled data. *Cancers*, 13(7):1590.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2020. Invariant risk minimization. *Preprint*, arXiv:1907.02893.

Andrew D Boyd, Rosa Gonzalez-Guarda, Katharine Lawrence, Crystal L Patil, Miriam O Ezenwa, Emily C O'Brien, Hyung Paek, Jordan M Braciszewski, Oluwaseun Adeyemi, Allison M

[1] https://physionet.org/news/post/gpt-responsible-use

Cuthel, et al. 2023. Potential bias and lack of generalizability in electronic health record data: reflections on health equity from the national institutes of health pragmatic trials collaboratory. *Journal of the American Medical Informatics Association*, 30(9):1561–1566.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660.

Muxi Chen, Yu Li, and Qiang Xu. 2024a. Hibug: on human-interpretable model debug. *Advances in Neural Information Processing Systems*, 36.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. 2024b. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Greg d'Eon, Jason d'Eon, James R Wright, and Kevin Leyton-Brown. 2022. The spotlight: A general method for discovering systematic errors in deep learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1962–1981.

Moreno D'Incà, Elia Peruzzo, Massimiliano Mancini, Dejia Xu, Vidit Goel, Xingqian Xu, Zhangyang Wang, Humphrey Shi, and Nicu Sebe. 2024. Openbias: Open-set bias detection in text-to-image generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12225–12235.

Frédéric Docquier and Hillel Rapoport. 2012. Globalization, brain drain, and development. *Journal of economic literature*, 50(3):681–730.

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

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

John Duchi and Hongseok Namkoong. 2021. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49.

Sabri Eyuboglu, Maya Varma, Khaled Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Ré. 2022. Domino: Discovering systematic errors with cross-modal embeddings. *arXiv preprint arXiv:2203.14960*.

Shantanu Ghosh, Clare B. Poynton, Shyam Visweswaran, and Kayhan Batmanghelich. 2024. Mammo-clip: A vision language foundation model to enhance data efficiency and robustness in mammography. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 632–642, Cham. Springer Nature Switzerland.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. 2019. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. 2020. Certified data removal from machine learning models. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

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

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. 2021. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. 2020. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*.

Michael P Kim, Amirata Ghorbani, and James Zou. 2019. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254.

Younghyun Kim, Sangwoo Mo, Minkyu Kim, Kyungmin Lee, Jaeho Lee, and Jinwoo Shin. 2024. Discovering and mitigating visual biases through keyword explanation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11082–11092.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. 2022. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision*, 123(1):32–73.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. 2018. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pretraining for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Zhiheng Li, Ivan Evtimov, Albert Gordo, Caner Hazirbas, Tal Hassner, Cristian Canton Ferrer, Chenliang Xu, and Mark Ibrahim. 2023b. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20071–20082.

Weixin Liang and James Zou. 2022. Metashift: A dataset of datasets for evaluating contextual distribution shifts and training conflicts. *arXiv preprint arXiv:2202.06523*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. 2021. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. *Proceedings of the IEEE International Conference on Computer Vision (ICCV).*

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2018. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11.

Mazda Moayeri, Keivan Rezaei, Maziar Sanjabi, and Soheil Feizi. 2023. Text-to-concept (and back) via cross-model alignment. In *International Conference on Machine Learning*, pages 25037–25060. PMLR.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Nihal Murali, Aahlad Puli, Ke Yu, Rajesh Ranganath, and Kayhan Batmanghelich. 2023. Beyond distribution shift: Spurious features through the lens of training dynamics. *arXiv preprint arXiv:2302.09344*.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. 2020. Learning from failure: Debiasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684.

Hieu T Nguyen, Ha Q Nguyen, Hieu H Pham, Khanh Lam, Linh T Le, Minh Dao, and Van Vu. 2023. Vindr-mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *Scientific Data*, 10(1):277.

Priya K Palanisamy, Bhawna Dev, and MC Sheela. 2023. Reporting template: Mammogram, usg, mri. In *Holistic Approach to Breast Disease*, pages 71–75. Springer.

Chantal Pellegrini, Ege Özsoy, Benjamin Busam, Nassir Navab, and Matthias Keicher. 2023. Radialog: A large vision-language model for radiology report generation and conversational assistance. *arXiv preprint arXiv:2311.18681*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Keivan Rezaei, Mehrdad Saberi, Mazda Moayeri, and Soheil Feizi. 2023. Prime: Prioritizing interpretability in failure mode extraction. *arXiv preprint arXiv:2310.00164*.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. 2021. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*.

Shiori Sagawa, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks. In *International Conference on Learning Representations*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, 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.

Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens Van Der Maaten. 2022. Revisiting weakly supervised pre-training of visual perception models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 804–814.

Sahil Singla, Besmira Nushi, Shital Shah, Ece Kamar, and Eric Horvitz. 2021. Understanding failures of deep networks via robust feature extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12853–12862.

Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. 2020. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33:19339–19352.

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Erdal Tasci, Ying Zhuge, Kevin Camphausen, and Andra V Krauze. 2022. Bias and class imbalance in oncologic data—towards inclusive and transferrable ai in large scale oncology data sets. *Cancers*, 14(12):2897.

Gemini Team, M Reid, N Savinov, D Teplyashin, Lepikhin Dmitry, T Lillicrap, JB Alayrac, R Soricut, A Lazaridou, O Firat, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. in arxiv [cs. cl]. arxiv.

Vladimir Vapnik. 1999. *The Nature of Statistical Learning Theory*. Springer.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*.

Xuesong Wen, Jianjun Li, and Liyuan Yang. 2024. Breast cancer diagnosis method based on cross-mammogram four-view interactive learning. *Tomography*, 10(6):848–868.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21372–21383.

Yiqi Wu, Xiaodan Hu, Ziming Fu, Siling Zhou, and Jiangong Li. 2024. Gpt-4o: Visual perception performance of multimodal large language models in piglet activity understanding. *arXiv preprint arXiv:2406.09781*.

Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. 2023. Change is hard: A closer look at subpopulation shift. *arXiv preprint arXiv:2302.12254*.

Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. 2022. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning*, pages 25407–25437. PMLR.

Sriram Yenamandra, Pratik Ramesh, Viraj Prabhu, and Judy Hoffman. 2023. Facts: First amplify correlations and then slice to discover bias. In *IEEE/CVF International Conference in Computer Vision (ICCV)*.

Kihyun You, Jawook Gu, Jiyeon Ham, Beomhee Park, Jiho Kim, Eun K Hong, Woonhyuk Baek, and Byungseok Roh. 2023. Cxr-clip: Toward large scale chest x-ray language-image pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 101–111. Springer.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. 2021. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

Yuhui Zhang, Jeff Z HaoChen, Shih-Cheng Huang, Kuan-Chieh Wang, James Zou, and Serena Yeung. 2023. Diagnosing and rectifying vision models using language. In *International Conference on Learning Representations (ICLR)*.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

A Appendix

A.1 Extended details on datasets

Waterbirds

The **Waterbirds** dataset (Wah et al., 2011) is frequently employed in studies addressing spurious correlations. This binary classification dataset overlaps images from the Caltech-UCSD Birds-200-2011 (CUB) dataset with backgrounds sourced from the Places dataset (Zhou et al., 2017). The primary task involves determining whether a bird depicted in an image is a landbird or a waterbird, with the background (water or land) as the spurious attribute. For consistency and comparability, we adhere to the train/validation/test splits utilized in prior research (Guo et al., 2020).

CelebA

The **CelebA** dataset (Liu et al., 2015) comprises over 200,000 images of celebrity faces. In the context of spurious correlations research, this dataset is typically used for the binary classification task of predicting hair color (blond vs. non-blond), with gender serving as the spurious correlation. In alignment with previous studies (Guo et al., 2020), we use the standard dataset splits. The CelebA dataset is available under the Creative Commons Attribution 4.0 International license.

MetaShift

The **MetaShift** dataset (Liang and Zou, 2022) offers a flexible platform for generating image datasets based on the Visual Genome project (Krishna et al., 2017). Our experiments utilize the pre-processed *Cat vs. Dog* dataset, designed to differentiate between cats and dogs. The dataset features the image background as a spurious attribute, with cats typically appearing indoors and dogs outdoors. We use the "unmixed" version of this dataset, as provided by the authors' codebase.

NIH chestXrays

The **NIH** ChestX-ray dataset (Wang et al., 2017), also known as ChestX-ray14, is a large dataset of chest radiographs (X-rays) provided by the National Institutes of Health (NIH). The dataset comprises 112,120 frontal-view X-ray images of 30,805 unique patients. Each image is associated with one or more of the 14 labeled thoracic diseases, which include atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, pneumothorax, consolidation, edema, emphysema, fibrosis, pleural thickening, and hernia. Previous works (Docquier and Rapoport, 2012) show that most pneumothorax patients have a spurious correlation with the chest drains. Chest drains are used to treat positive Pneumothorax cases. We adopt the strategy discussed in Murali et al. (2023) to annotate chest drains for each sample. We use the official train/val/test split (Wang et al., 2017).

RSNA breast mammograms

The **RSNA-Mammo** dataset[2] is a publicly available dataset containing 2D mammograms from 11,913 patients, with 486 diagnosed cancer cases. The task is to classify malignant cases from screening mammograms. We use a 70/20/10 train/validation/test split for evaluation as Ghosh et al. (2024).

VinDr breast mammograms

The **VinDr-Mammo** dataset[3] (Nguyen et al., 2023) is a publicly available 2D mammogram dataset of 5,000 exams (20,000 images) from Vietnam, each with four views. It includes breast-level BI-RADS assessment categories (1-5), breast density categories (A-D), and annotations for mammographic attributes (*e.g.,* mass, calcifications). Following Wen et al. (2024), we classify patients with BI-RADS scores between 1 and 3 as normal and those with scores of 4 and 5 as abnormal. We adopt the train-test split from Nguyen et al. (2023).

A.2 Extended details on slice discovery algorithms

Domino. Domino (Eyuboglu et al., 2022) identifies systematic errors in machine learning models by leveraging cross-modal embeddings. It operates in three main steps: embedding, slicing, and describing.

[2] https://www.kaggle.com/competitions/rsna-breast-cancer-detection

[3] https://www.physionet.org/content/vindr-mammo/1.0.0/

1. **Embedding**: Domino uses cross-modal models (*e.g.,* CLIP) to embed inputs and text in the same latent space. This enables the incorporation of semantic meaning from text into input embeddings, which is crucial for identifying coherent slices.

2. **Slicing**: It employs an error-aware mixture model to detect underperforming regions within the embedding space. This model clusters the data based on embeddings, class labels, and model predictions to pinpoint areas where the model performance is subpar. The mixture model ensures that identified slices are coherent and relevant to model errors.

3. **Describing**: Domino generates natural language descriptions for the discovered slices. It creates prototype embeddings for each slice and matches them with text embeddings to describe the common characteristics of the slice. This step provides interpretable insights into why the model fails on these slices.

Domino's approach improves slice coherence and generates meaningful slice descriptions.

Facts. Facts (Yenamandra et al., 2023) (First Amplify Correlations and Then Slice) aims to identify bias-conflicting slices in datasets through a two-stage process:

1. **Amplify Correlations**: This stage involves training a model with a high regularization term to amplify its reliance on spurious correlations present in the dataset. This step helps segregate biased-aligned from bias-conflicting samples by making the model fit a simpler, biased-aligned hypothesis.

2. **Correlation-aware Slicing**: In this stage, FACTS uses clustering techniques on the bias-amplified feature space to discover bias-conflicting slices. The method identifies subgroups where the spurious correlations do not hold, highlighting areas where the model underperforms due to these biases.

Facts leverages a combination of bias amplification and clustering to reveal underperforming data slices, providing a foundation for understanding and mitigating systematic biases in machine learning models.

A.3 Extended details on error mitigation baselines

We categorize the various bias mitigation algorithms and provide detailed descriptions for each category below.

Vanilla

The empirical risk minimization (ERM) algorithm, introduced by Vapnik (Vapnik, 1999), seeks to minimize the cumulative error across all samples.

Subgroup Robust Methods

GroupDRO: GroupDRO (Sagawa et al., 2020) propose Group Distributionally Robust Optimization (GroupDRO), which enhances ERM by prioritizing groups with higher error rates. **CVaRDRO:** Duchi and Namkoong (Duchi and Namkoong, 2021) introduce a variant of GroupDRO that dynamically assigns weights to data samples with the highest losses. **LfF:** LfF (Nam et al., 2020) concurrently trains two models: the first model is biased, and the second is de-biased by re-weighting the loss gradient. **Just Train Twice (JTT):** JTT (Liu et al., 2021) propose an approach that initially trains an ERM model to identify minority groups in the training set, followed by a second ERM model where the identified samples are re-weighted. **LISA:** LISA (Yao et al., 2022) utilizes invariant predictors through data interpolation within and across attributes. **Deep Feature Re-weighting (DFR):** DFR (Kirichenko et al., 2022) suggests first training an ERM model and then retraining the final layer using a balanced validation set with group annotations.

Data Augmentation

Mixup: Mixup (Zhang et al., 2018) proposes an approach that performs ERM on linear interpolations of randomly sampled training examples and their corresponding labels.

Domain-Invariant Representation Learning

Invariant Risk Minimization (IRM): IRM (Arjovsky et al., 2020) learns a feature representation such that the optimal linear classifier on this representation is consistent across different domains. **Maximum Mean Discrepancy (MMD):** MMD (Li et al., 2018) aims to match feature distributions across domains. **Note: All methods in this category necessitate group annotations during training**.

Imbalanced Learning

Focal Loss (Focal): Focal (Lin et al., 2017) introduces Focal Loss, which reduces the loss for well-classified samples and emphasizes difficult samples. **Class-Balanced Loss (CBLoss):** CBLoss (Cui et al., 2019) suggests re-weighting by the inverse effective number of samples. **LDAM Loss (LDAM):** LDAM (Cao et al., 2019) employs a modified margin loss that preferentially weights minority samples. **Classifier Re-training (CRT):** CRT (Kang et al., 2020) decomposes representation learning and classifier training into two distinct stages, re-weighting the classifier using class-balanced sampling during the second stage. **ReWeightCRT:** ReWeightCRT (Kang et al., 2020) proposes a re-weighted variant of CRT.

A.4 Learning Projection from classifier to VLR space

π is a learnable projection function, $\pi : \Phi \to \Psi^I$, projecting the image representation of the classifier $\Phi(x)$ to the VLR space, $\Psi(x)$, where $x \in \mathcal{D}_{train}$. $\mathcal{D}_{train}$ denotes the training set. We follow (Moayeri et al., 2023) to learn π. Specifically, π is an affine transformation, *i.e.*, $\pi_{W,b}(z) = W^T z + b$, where W and b are the learnable weights and biases of the projector π. To retain the original semantics in the classifier representation space, we optimize the following objective:

$$W, b = \arg\min_{W,b} \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{x \in \mathcal{D}_{\text{train}}} \left\| W^T \Phi(x) + b - \Psi(x) \right\|_2^2 \tag{1}$$

A.5 Precision@k

`Precision@k` (Eyuboglu et al., 2022; Yenamandra et al., 2023) measures the degree to which the predicted slices overlap with the ground truth slices in a dataset.

Let $S = \{s_1, s_2, \ldots, s_l\}$ represent the ground truth bias-conflicting slices in a dataset $\mathcal{D}$. A slice discovery algorithm A predicts a set of slices $\hat{S} = \{\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_m\}$. For each predicted slice $\hat{s}_j$, let $O_j = \{o_{j1}, o_{j2}, \ldots, o_{jn}\}$ denote the sequence of sample indices ordered by the decreasing likelihood that each sample belongs to the predicted slice $\hat{s}_j$.

Given a ground truth slice s_i and a predicted slice $\hat{s}_j$, we compute their similarity as:

$$P_k(s_i, \hat{s}_j) = \frac{1}{k} \sum_{i=1}^{k} \mathbb{I}[x_{o_{ji}} \in s_i],$$

where $P_k(s_i, \hat{s}_j)$ is the proportion of the top k samples in the predicted slice $\hat{s}_j$ that overlap with the samples in the ground truth slice s_i, and $\mathbb{I}$ is an indicator function that returns 1 if the sample belongs to s_i and 0 otherwise.

For each ground truth slice s_i, we map it to the most similar predicted slice $\hat{s}_j$ by maximizing $P_k(s_i, \hat{s}_j)$. We then compute the average similarity score between the ground truth slices and their best-matching predicted slices. Specifically, the `Precision@k` for a slice discovery algorithm A is given by:

$$\texttt{Precision@k}(A) = \frac{1}{l} \sum_{i=1}^{l} \max_{j \in [m]} P_k(s_i, \hat{s}_j),$$

where l is the number of ground truth slices, m is the number of predicted slices, and $P_k(s_i, \hat{s}_j)$ is the similarity score for the ground truth slice s_i and predicted slice $\hat{s}_j$.

This metric evaluates how well the algorithm's predicted slices match the bias-conflicting slices in the dataset, with higher scores indicating better alignment between predicted and ground truth slices. By computing the `Precision@k`, we can assess the effectiveness of slice discovery algorithms in identifying and isolating the most significant bias-conflicting regions in the data.

A.6 Clip Score

Kim et al. (2024) introduces the CLIP score, a metric that leverages the similarity between language and vision embeddings to quantify the influence of specific attributes on misclassified samples. In their method, attributes frequently present in misclassified images receive a high CLIP score, while absent ones score lower. For instance, in the Waterbirds dataset, the CLIP score for "bamboo" is high, as many misclassified waterbirds appear with bamboo in the background.

We propose a modification to the CLIP score. As discussed in Sec. 3.1, our goal is to identify visual attributes that are prevalent in correctly classified samples but absent in misclassified ones. This approach provides deeper insights into the attributes contributing to correct classifications, which is particularly valuable for medical images. In scenarios such as pneumothorax detection in the NIH dataset, understanding biases incorrectly classified cases—such as the presence of chest tubes—can help isolate features that lead to reliable diagnoses while addressing spurious correlations. Formally we define the CLIP score corresponding to the attribute `attr` and a dataset $\mathcal{D}$ as,

$$s_{CLIP}(\mathsf{attr}, \mathcal{D}) = \mathsf{sim}(\mathsf{attr}, \mathcal{D}_{correct}) - \mathsf{sim}(\mathsf{attr}, \mathcal{D}_{wrong}),$$

where `attr` is the attribute obtained from the specific hypothesis by LLM, described in Sec. 3.2, $\mathcal{D}_{correct}$ and $\mathcal{D}_{wrong}$ are the correctly classified and misclassified samples. Also, $\mathsf{sim}(\mathsf{attr}, \mathcal{D})$ is the similarity between the attribute `attr` and the dataset $\mathcal{D}$, estimated as the average cosine similarity between normalized embedding of a word $\Psi^T(\mathsf{attr})$ and images $\Psi^I(\mathsf{x})$ for $x \in \mathcal{D}$, where

$$\mathsf{sim}(\mathsf{attr}, \mathcal{D}) = \frac{1}{\mathcal{D}} \sum_{x \in \mathcal{D}} \Psi^I(x) \Psi^T(\mathsf{attr})$$

Refer to Appendix A.12.6 for the results.

A.7 Prompts used by LLM for hypotheses generation

The following is a general template of the prompt utilized to generate the hypotheses from LLM, discussed in Sec. 3.2. In this template, we substitute the <task> placeholders with bird species, hair color, animal species, pneumothorax, cancer, and abnormality based on the corresponding dataset – Waterbirds, CelebA, MetaShift, NIH, RSNA-Mammo, and VinDr-Mammo. The modalities are natural images, chest-x-rays, and 2D mammograms. **Crucially, we only replace these two placeholders. We never include the actual dataset names or words like “water”, “land”, “gender”, “tube”, “background” or any other attributes leading to model’s mistakes in the prompt, as these may bias the LLM’s output**. For medical images, we also add: `Ignore ‘___’ as they are due to anonymization. We focus only on positive <disease> patients`, as many reports consist of ‘___’ for clarity. top <K> depends on the dataset discussed in the experiment section (Sec. 4).

Prompt for Hypothesis Generation

Context: <task> classification from <modality> using a deep neural network.

Analysis Post-Training: On a validation set:

a. Get the difference between the image embeddings of correct and incorrectly classified samples to estimate the features present in the correctly classified samples but missing in the misclassified samples.

b. Retrieve the top <K> sentences from the <captions/radiology report> that match closely to the embedding difference in step a.

c. The sentence list is given below:

> TopK Sentence List
>
> **Retrieved using Sec. 3.1**

These sentences represent the features present in the correctly classified samples but missing in the misclassified samples.
Task: Consider the consistent attributes present in the descriptions of correctly classified and misclassified samples regarding <task>. Formulate hypotheses based on these attributes. Attributes include all concepts (e.g., explicit or implicit anatomies, observations, symptoms of change related to the disease, concepts leading to potential bias in medical images, or visual cues in natural images) in the sentences. Assess how these characteristics might influence the classifier's performance.
Your response should only contain the list of top hypotheses, formatted as follows:

```
hypothesis_dict = {
    'H1': 'The classifier is making mistake as it is biased toward <attribute>',
    'H2': 'The classifier is making mistake as it is biased toward <attribute>',
    'H3': 'The classifier is making mistake as it is biased toward <attribute>',
    ...
}
```

To effectively test Hypothesis 1 (H1) using the CLIP language encoder, create prompts explicitly validating H1. These prompts will help generate text embeddings that capture the essence of the hypothesis, which can be compared with the image embeddings from the dataset. The goal is to verify alignment with or violation of H1. Prompts must focus only on the <task>. Each hypothesis must have five prompts, formatted as:

```
prompt_dict = {
    'H1_<attribute>': [List of prompts],
    'H2_<attribute>': [List of prompts],
    ...
}
```

Final response format strictly:

```
hypothesis_dict
prompt_dict
```

Table 3: Detailed description of the prompt for hypothesis generation and analysis for the <task> classification problem.

A.8 Prompts and details on the experiments in RQ5 with instruction-tuned models (*e.g.,* LLaVA)

In this setup, we don't use CLIP as VLR for the retrieval step discussed in Sec. 3.1. Instead, using the instruction-tuned vision language models (LLaVA-1.5 7B for natural images; cheXagent and RaDialog for CXRs), we first select the correctly classified images by the classifier f. Next, for each of the images, we pass them through the vision encoder in LLaVA and use the prompt for the natural images: "`Describe the image`" for the language model in LLaVA. For NIH, we use the prompt:

```
You are a radiologist. Based on the provided
    Chest X-Ray image and generate a structured
    report. The report should include sections
    for `Findings,' `Impression,' and
    `Recommendations,' emphasizing relevant
    findings like consolidation, effusion,
    cardiomegaly, pneumonia, or pneumothorax.
    Use a formal radiology reporting style.
```

We select the texts for all the correctly classified images and follow LADDER's pipeline discussed in Sec 3.2 to generate the hypothesis (results shown in Fig. 5). Finally, we utilize LADDER's mitigation strategy, discussed in Sec. 3.3 to mitigate the biases (results shown in Tab. 6). **Note: in this experiment, we did not use any language explicitly. However, there is always a trade-off between getting language or using an instruction-tuning model like LLaVA.**

A.9 Prompts and examples of metadata for detecting biases beyond radiology reports in RQ6

Refer to Tab. 4 for the prompt and the example of Python dictionary of metadata details of the correctly classified cancer patients to detect biases using LADDER.

A.10 Extended details on general experiments

A.10.1 Implementation details of the source model f using ERM

For natural images and chest X-rays (CXRs), we resize the images to 224×224 and train ResNet-50 (RN)(He et al., 2016) and Vision Transformer (ViT)(Dosovitskiy et al., 2020) models as f to predict labels. We explore various pretraining methods for initializing model weights, including supervised learning (Sup), SIMCLR(Chen et al., 2020), Barlow Twins (Zbontar et al., 2021), DINO (Caron et al., 2021), and CLIP-based pretraining (Radford et al., 2021). The pretraining datasets utilized include ImageNet-1K (IN1)(Deng et al., 2009), ImageNet-21K (IN-21K)(Ridnik et al., 2021), SWAG (Singh et al., 2022), LAION-2B (Schuhmann et al., 2022), and OpenAI-CLIP (OAI) (Radford et al., 2021). For instance,"RN Sup IN1k" refers to a ResNet model pretrained using supervised learning and ImageNet-1K.

We train both ResNet and ViT models as f for natural images and NIH-CXR following the setup in Yang et al. (2023)[4]. Preprocessing steps include resizing the images to 224×224, applying center-cropping, and normalizing the images using ImageNet channel statistics. Consistent with prior work (Guo et al., 2020, 2019), we apply stochastic gradient descent (SGD) with momentum for optimization across all image datasets. Each model is trained for a total of 30,000 steps across all datasets, with specific training on Waterbirds and MetaShift for 5,000 steps each. For NIH, we utilize the Adam optimizer with a learning rate of 0.0001 and train for 60 epochs to achieve optimal convergence.

For RSNA-Mammo, we leverage the setting from one of the leading Kaggle competition solutions[5]. In this setup, the images are resized to 1520×912, and we train an EfficientNet-B5 model (Tan and Le, 2019) for 9 epochs using the SGD optimizer, with a learning rate of 5e-5 and a weight decay of 1e-4.

Additionally, for CXR-CLIP, we use their pretrained models[6], which were trained on MIMIC-CXR and CheXpert (MC) datasets. For Mammo-CLIP, we utilize their EN-B5 variant[7].

A.10.2 Ablations

For the captioning ablations, we compare the performance of LADDER using BLIP (Li et al., 2022), BLIP-2 (Li et al., 2023a), ClipCap (Mokady et al., 2021), and GPT-4o (Wu et al., 2024). Additionally, for LLMs, we compare the performance of LADDER with GPT-4o (Wu et al., 2024), Claude 3.5 Sonnet, Llama 3.1 70B (Dubey et al., 2024), and Gemini 1.5 Pro (Team et al., 2024).

[4] https://github.com/YyzHarry/SubpopBench

[5] https://github.com/Masaaaato/RSNABreast7thPlace

[6] https://github.com/kakaobrain/cxr-clip

[7] https://huggingface.co/shawn24/Mammo-CLIP/blob/main/Pre-trained-checkpoints/b5-model-best-epoch-7.tar

Context: Breast cancer classification from mammograms using a deep neural network
Analysis post-training: On a validation set, you are provided with the metadata details for the correctly classified positive cancer patients in a Python dictionary, as follows

Metadata Dictionary (Sample Entries)

- **Patient 1:** `{site_id: 1, laterality: L, view: MLO, age: 71, biopsy: 1, invasive: 1, BIRADS: 0, implant: 0, density: B, machine_id: 49, photometric_interpretations: Monochrome 1, voi_lut_function: SIGMOID, pixel_intensity_relationship: LOG}`
- **Patient 2:** `{site_id: 2, laterality: L, view: CC, age: 83, biopsy: 0, invasive: 0, BIRADS: 0, implant: 1, density: D, machine_id: 49, photometric_interpretations: Monochrome 1, voi_lut_function: SIGMOID, pixel_intensity_relationship: LOG}`
- ··· `(Additional metadata entries omitted for brevity)`

Task: Consider the consistent attributes present in the dictionary regarding the positive cancer patients. Formulate hypotheses based on these attributes. Assess how these characteristics might be influencing the classifier's performance. Your response should contain only the list of top hypothesis, nothing else. For the response, you should be the following python dictionary template, no extra sentence:

```
hypothesis_dict = {
    'H1': 'The classifier is making mistake as it is biased toward <attribute>',
    'H2': 'The classifier is making mistake as it is biased toward <attribute>',
    'H3': 'The classifier is making mistake as it is biased toward <attribute>',
    ...
}
```

Table 4: Prompts and examples of metadata for detecting biases beyond radiology reports in the experiment RQ6.

A.10.3 Radiology text synthesis for 2D Mammograms

In Ghosh et al. (2024), the authors generate mammography reports using labeled mammographic attributes from the VinDr dataset in collaboration with a board-certified radiologist. This approach leverages the templated nature of breast mammogram reports, which are more standardized than those for other medical imaging modalities. This standardized structure follows protocols like BI-RADS (Breast Imaging-Reporting and Data System), which promotes uniformity in reporting (Palanisamy et al., 2023). Specifically, they focus on the following attributes: `mass`, `architectural distortion`, `calcification`, `asymmetry (focal, global)`, `density`, `suspicious lymph nodes`, `nipple retraction`, `skin retraction`, and `skin thickening`. Then they follow the report templates with radiologist-defined prompts in Ghosh et al. (2024), describing key parameters such as:

Attribute Value: Positive, negative, etc.

Subtype: Suspicious, obscured, spiculated, etc.

Laterality: Left or right breast.

Position: Upper, lower, inner, outer quadrant.

Depth: Anterior, mid, or posterior.

Finally, they generate concise report-like sentences by substituting these values into the templates. The authors leverage these sentences in Mammo-FActOR to perform weakly supervised localization of mammographic findings. In our work, we collect all these sentences to probe the EN-B5 classifier f, analyzing its errors during the retrieval step (Sec. 3.1) for the RSNA-Mammo and VinDr-Mammo datasets.

Below are some examples of mammography report sentences corresponding to the specific mammographic attributes.

Mass:

```
1. there is a mass in the right breast
2. there is a mass in the right breast at
   anterior depth
3. there is a mass in the upper right breast
   at mid-depth
                    ...
```

Architectural distortion:

```
1. there is architectural distortion in the
   right breast
2. there is architectural distortion in the
   right breast at anterior depth
3. there is architectural distortion in the
   right breast at mid-depth
                    ...
```

Calcification:

```
1. there is calcification in the right breast
2. there is calcification in the right
   breast at anterior depth
3. there is calcification in the right
   breast at mid depth
                    ...
```

Asymmetry:

```
1. there is a developing asymmetry in the
   outer right breast
2. there is an asymmetry in the inner right
   breast at anterior depth
3 .there is an asymmetry in the right breast
   at mid-depth
                    ...
```

Global Asymmetry:

```
1. there is a global asymmetry in the right
   breast
2. there is a new global asymmetry in the
   right breast
3. there is a global asymmetry in the inner
   right breast
                    ...
```

Focal Asymmetry:

```
1. mthere is a focal asymmetry in the right
   breast
2. there is a focal asymmetry in the right
   breast at anterior depth
3. there is a focal asymmetry in the right
   breast at mid depth
                    ...
```

Density:

```
1. the breasts being almost entirely fatty
2. scattered areas of fibroglandular density
3. the breast tissue is heterogeneously dense
4. the breasts are extremely dense
                    ...
```

Suspicious lymph node:

```
1. there is a suspicious lymph node in the
   right axilla
2. there is a hyperdense lymph node in the
   right axillary tail
3. there is an increased lymph node in the
   right axillary tail
                    ...
```

Suspicious lymph node:

1. there is a suspicious lymph node in the
   right axilla
2. there is a hyperdense lymph node in the
   right axillary tail
3. there is an increased lymph node in the
   right axillary tail
                    ...

Nipple retraction:

```
1. there is a new nipple retraction in the
   right breast
2. there is an increased nipple retraction
   in the right breast
3. there is a possible nipple retraction in
   the right breast
                    ...
```

Skin retraction:

```
1. there is skin retraction in the right
   breast
2. there is skin retraction in the inner
   right breast
3. there is skin retraction in the lower
   right breast
                    ...
```

Skin thickening:

```
1. there is increasing skin thickening of
   the periareolar right breast
2. there is asymmetric skin thickening of
   the lower right breast
3. there is asymmetric skin thickening of
   the inner right breast
                    ...
```

A.11 Toy dataset construction

We construct a synthetic dataset based on the **CUB-200-2011** (Wah et al., 2011) dataset, classifying bird species into two categories: **Class 0** ($y = 0$) and **Class 1** ($y = 1$). Class 1 consists of the following bird species: *Albatross*, *Auklet*, *Cormorant*, *Frigatebird*, *Fulmar*, *Gull*, *Jaeger*, *Kittiwake*, *Pelican*, *Puffin*, *Tern*, *Gadwall*, *Grebe*, *Mallard*, *Merganser*, *Guillemot*, and *Pacific Loon*. All remaining bird species are assigned to Class 0. To introduce spurious correlations, we overlay two 3D boxes on each image. In the training set for Class 0, the majority of samples (95%) were biased, with the yellow box consistently placed to the left of the red box. For Class 1, the boxes were randomly placed, introducing variability in their positioning. In the validation and test sets, we split the positioning evenly, with 50% biased and 50% random samples across both classes, ensuring a balanced evaluation of the model's reliance on spurious cues.

The primary goal of this dataset is to introduce a form of *reasoning* beyond the mere presence or absence of spurious correlations. Unlike prior datasets that rely on background cues (*e.g.,* Waterbirds or Metashift) or attributes like gender (*e.g.,* CelebA), our dataset integrates positional reasoning. Specifically, for Class 0, the yellow box is consistently placed to the left of the red box, creating a spurious correlation. For Class 1, the boxes are randomly positioned, removing this shortcut. The relative positioning of the boxes allows the captions to encode spatial relationships, which can be consumed by large language models (LLMs) to reason about these spatial cues. We train an ImageNet pretrained-ResNet model (RN Sup IN1k) on this dataset. Predictably, the classifier latches onto the spurious correlation of rectangle position, leading to underperformance on subsets where the shortcut is absent. The model achieves a mean accuracy of 85.6% and a worst-group accuracy (WGA) of 65.2%.

To analyze the model's errors, we generate a corpus of rich captions for the validation set using a GPT-4o-based captioner. These captions describe both the presence of the rectangle and its position relative to the bird. Using LADDER, we aim to detect the reason for the classifier's mistakes and mitigate it. LADDER leverages the reasoning capabilities of LLMs to capture both the presence of the rectangles and their relative spatial position. In contrast, methods *e.g.,* PRIME, rely on external tagging models, which only detect the presence or absence of shortcuts. Furthermore, since LADDER discovers biased attributes via LLM-generated reasoning, it can effectively mitigate these biases without requiring ground truth annotations or prior knowledge of the attributes.

The data is split into training, validation, and test sets, with all metadata (including labels, rectangle positions) saved for future analysis.

A.12 Extended main results

A.12.1 Results on WGA for using all slice discovery methods:

Fig.12 shows that LADDER improves WGA compared to other slice discovery methods for natural images and CXRs. In this experimental setup, we first discover the slices with Domino (Eyuboglu et al., 2022), Facts (Yenamandra et al., 2023) and LADDER's hypothesis-driven approaches. Next, we apply LADDER's mitigation approach for each discovered slice to mitigate the biases and compute the WGA for each slice discovery method. As LADDER detects the slices precisely, it achieves better WGA compared to Domino and Facts. Fig. 13 shows LADDER improves WGA compared to other slice discovery methods for RSNA-Mammo and VinDr-Mammo datasets.

A.12.2 Closest hypothesis to the ground truth attribute

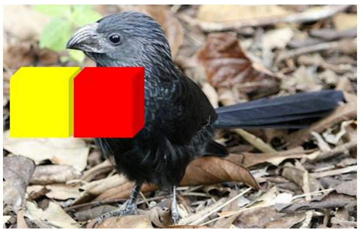
Tab. 6 and Tab. 5 show the `top3` hypotheses for RN Sup IN1K (convolution-based) and ViT Sup IN1K (transformer-based) architectures, respectively. These hypotheses are the most similar to the ground truth attribute on which the source model f is biased.

Table 5: Top 3 associated hypotheses for the ground truth biased attribute for ViT Sup IN1K model on various datasets

Dataset (Label)	**Attribute**	**Top 3 hypotheses**
Waterbirds (waterbird)	Water	1. activities like swimming or flying 2. conditions like cloudy or sunny 3. presence of objects like boats or rocks
Waterbirds (landbird)	Land	1. bird in the middle of a forest 2. yellow bird 3. bird sitting on top of a tree
CelebA (Blonde)	Women	1. woman wearing red dress 2. woman with red top 3. black jacket
MetaShift (Dog)	Outdoor	1. presence of a leash 2. presence of a ball 3. presence of a car
MetaShift (Cat)	Indoor	1. beds 2. windows 3. televisions

A.12.3 Extended qualitative results for our slice discovery method on various datasets

Figures 24 and 19 report LLM-generated the list of hypotheses and the prompts to test them discussed in the Sec. 5. Figures 20, 21, 22, 23, and 25 illustrate qualitative results of our method applied on various datasets using RN Sup IN1k models. Specifically, they showcase the classification of pneumothorax patients from NIH, "landbird" from the Waterbirds, "blond" from CelebA, "cat" and "dog" from MetaShift, and "cancer" from the RSNA-Mammo datasets, respectively. In all the cases, LADDER correctly identifies the hypothesis with true attribute causing biases in the given classifier f.



Extracted hypotheses by Ladder
The classifier is making mistake as it is biased toward:
H1: relative positioning of red and yellow box
H2: images with small birds
H3: images with overlapping boxes
H4: the position of boxes relative to the bird
H5: images with bird on branches

Figure 11: Sample images of our toy dataset to validate the reasoning of LLM utilized by LADDER. The dataset has two classes. Images with class 0 are biased, with the yellow box always placed to the left of the red box. For images with class 1, the boxes are randomly placed.

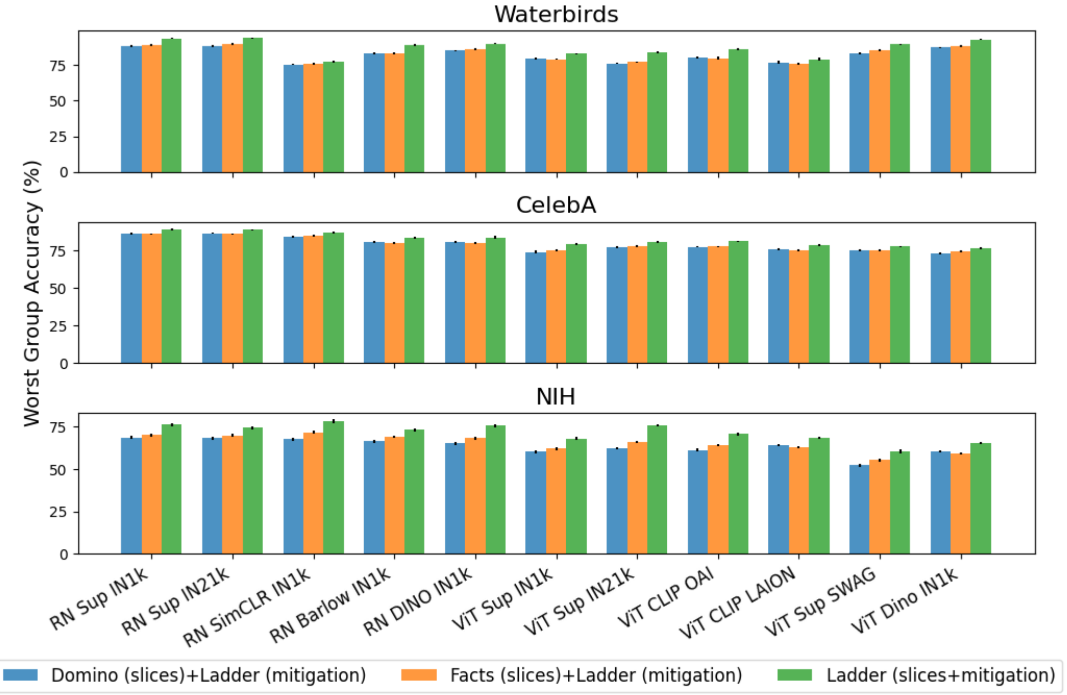


Figure 12: LADDER slices consistently outperform those from Domino and Facts when combined with LADDER's bias mitigation strategy across various settings.

A.12.4 Comparing the performance of LADDER for error mitigation across architectures

Tab. 8 compares LADDER with additional bias mitigation baselines for CNN-based models. Tab. 9 compares different error mitigation algorithms for ViT Sup IN1K-based models (f), for all the SOTA mitigation baselines discussed in Appendix A.3. For natural images (Waterbirds and CelebA), we report mean accuracy. For medical images (NIH, RSNA and VinDr), we report mean AUROC. Fig. 15 reports the WGA and shows that LADDER outperforms the other slice discovery baselines across the different architectures and pre-training strategies.

A.12.5 Application: Improvement on the zero-shot accuracy of Vision Language models using the attributes from the extracted hypothesis by LADDER

To evaluate the impact of LADDER's attribute-based slice discovery on zero-shot performance, we conducted experiments using a CLIP-based vision-language model across multiple datasets. LADDER extracts fine-grained attributes from error-prone data slices, which we incorporated as detailed prompts for zero-shot classification. These prompts were generated from hypotheses produced by the LADDER framework and reflect nuanced characteristics of the data that a model might otherwise overlook. We compare these attribute-driven prompts against standard, baseline prompts typically used for zero-shot tasks.

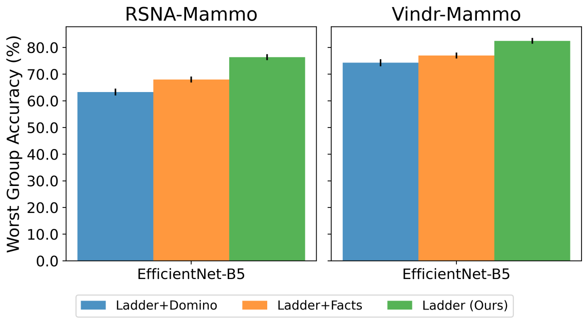


Figure 13: LADDER improves WGA compared to other bias mitigation methods for RSNA-Mammo and VinDr-Mammo datasets.

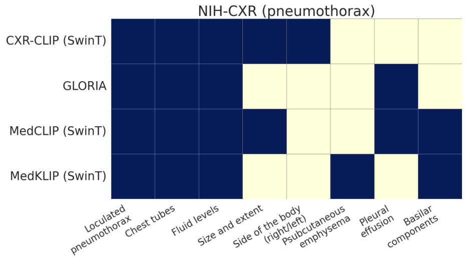


Figure 14: Effect of different VLRs for CXRs on biased attribute discovery by LADDER. Bright/light colors denote presence/absence of the attributes.

Experimental Process. For each dataset, we implemented two types of zero-shot prompts:

- **Baseline prompts**: CLIP-based prompts (Radford et al., 2021) *e.g.,* `[a photo of a landbird and a photo of a waterbird]` for the Waterbirds dataset for natural images, CXR-CLIP (You et al., 2023) prompts *e.g.,* `[no pneumothorax, pneumothorax]` for NIH, Mammo-CLIP (Ghosh et al., 2024) prompts *e.g.,* `[{no cancer, no malignancy},{cancer, malignancy}]` for RSNA-Mammo and VinDr-Mammo.

- **LADDER-derived prompts**: These prompts were generated based on the attributes extracted from LADDER's hypotheses, providing a more detailed description of the data. For example, in the Waterbirds dataset, we used prompts like `a photo of a waterbird on docks and boats` or `a photo of a landbird inside on bamboo forest`. In this experiment, we use the attributes from the hypotheses extracted from RN Sup IN1k (Resnet 50 pretrained with ImageNet 1K and supervised learning) classifier.

Table 6: Top 3 associated hypotheses for the ground truth biased attribute for RN Sup IN1K model on various datasets

Dataset (Label)	Attribute	Top 3 hypotheses
Waterbirds (waterbird)	Water	1. water bodies like oceans and lakes 2. actions like flying or sitting 3. conditions, e.g., cloudy skies
Waterbirds (landbird)	Land	1. bird being in flight 2. bird perching on top of a tree 3. bird perching on a tree branch
CelebA (Blonde)	Women	1. woman with long hair 2. woman wearing red dress 3. a black jacket
MetaShift (Dog)	Outdoor	1. dogs in motion 2. dogs on leashes 3. beach environments
MetaShift (Cat)	Indoor	1. televisions 2. windows 3. beds
NIH (pneumothorax)	Chest tube	1. the presence of chest tubes 2. loculated pneumothorax 3. size and extent of pneumothorax
RSNA-Mammo (cancer)	Calcification	1. scattered calcifications 2. vascular calcifications 3. bilateral occurrences

Table 7: **Token Usage and Cost for Each LLM.** Each row shows the breakdown for an LLM extracting hypotheses across all 6 datasets, using RN Sup IN1k (natural images / CXRs) and EN-B5 (mammograms).

Model Name	Input Tokens	Output Tokens	Total Cost
GPT-4o	33,217	4,284	$2.51
Claude 3.5 Sonnet	34,888	4,473	$0.17
Gemini 1.5 Pro	33,872	4,378	$0.32
Llama 3.1 70B	32,688	4,176	$0.05
Total	134,665	17,311	**$3.05**

We evaluated the zero-shot classification performance of the model using both prompt types. The results are shown in Tab. 10.

Results. The results demonstrate a significant improvement in zero-shot accuracy when using LADDER-extracted attributes as prompts. Across all datasets, the attribute-driven prompts outperformed the baseline, indicating the effectiveness of using detailed, hypothesis-driven attributes to enhance zero-shot performance. In the **Waterbirds** dataset, LADDER prompts improved accuracy by +8.56%, rising from 50.40% with baseline prompts to 58.96% with LADDER attributes. The improvement was even more pronounced for the **NIH** dataset, with a +19.05% gain (49.17% to 68.22%). The **RSNA** dataset also saw a notable improvement, with a +5.81% gain in accuracy (60.17% to 65.98%). The improvements for **CelebA** (+0.32%) and **VinDr** (+1.41%) were more modest but still indicate that using LADDER's attribute-based prompts provides consistent gains across various domains. These results highlight the ability of LADDER to extract meaningful attributes that guide the vision-language model to

Table 8: Benchmarking error mitigation methods over 3 seeds for CNN models (EN-B5 for mammograms and RN Sup IN1k for the rest). For natural images (Waterbirds and CelebA), we report mean accuracy. For medical images (NIH, RSNA and VinDr), we report mean AUROC. We bold-face and underline the best and second-best results, respectively.

Method	**Waterbirds**		**CelebA**		**NIH**		**RSNA**		**VinDr**	
	Mean(%)	**WGA(%)**	**Mean(%)**	**WGA(%)**	**Mean(%)**	**WGA(%)**	**Mean(%)**	**WGA(%)**	**Mean(%)**	**WGA(%)**
Vanilla (ERM)	88.2$_{\pm 0.7}$	69.1$_{\pm 1.2}$	94.1$_{\pm 0.2}$	62.2$_{\pm 1.5}$	**87.4**$_{\pm 0.0}$	60.3$_{\pm 0.0}$	**86.5**$_{\pm 0.0}$	69.8$_{\pm 0.0}$	**86.9**$_{\pm 0.0}$	45.6$_{\pm 0.0}$
Mixup	88.5$_{\pm 0.5}$	77.3$_{\pm 0.5}$	94.5$_{\pm 0.1}$	57.8$_{\pm 0.8}$	85.1$_{\pm 0.0}$	67.6$_{\pm 0.8}$	84.5$_{\pm 0.0}$	64.8$_{\pm 0.0}$	83.2$_{\pm 0.0}$	65.3$_{\pm 0.0}$
IRM	88.1$_{\pm 0.2}$	74.3$_{\pm 0.1}$	94.5$_{\pm 0.5}$	63.3$_{\pm 2.5}$	83.2$_{\pm 0.0}$	63.4$_{\pm 0.0}$	83.3$_{\pm 0.0}$	68.4$_{\pm 0.0}$	83.5$_{\pm 0.0}$	65.2$_{\pm 0.0}$
MMD	92.5$_{\pm 0.1}$	83.5$_{\pm 1.1}$	92.5$_{\pm 0.6}$	22.7$_{\pm 2.5}$	84.6$_{\pm 0.0}$	65.4$_{\pm 0.0}$	84.2$_{\pm 0.0}$	69.1$_{\pm 0.0}$	81.2$_{\pm 0.0}$	64.8$_{\pm 0.0}$
Focal	89.3$_{\pm 0.2}$	71.6$_{\pm 0.8}$	**94.9**$_{\pm 0.3}$	59.3$_{\pm 2.0}$	85.5$_{\pm 0.0}$	68.9$_{\pm 0.7}$	83.6$_{\pm 0.0}$	65.5$_{\pm 0.0}$	82.6$_{\pm 0.0}$	63.7$_{\pm 0.0}$
CBLoss	91.3$_{\pm 0.7}$	86.1$_{\pm 0.3}$	91.2$_{\pm 0.7}$	87.3$_{\pm 0.5}$	85.5$_{\pm 0.0}$	63.4$_{\pm 0.0}$	83.2$_{\pm 0.0}$	65.1$_{\pm 0.0}$	81.7$_{\pm 0.0}$	62.5$_{\pm 0.0}$
LDAM	91.3$_{\pm 0.7}$	86.1$_{\pm 0.3}$	94.5$_{\pm 0.2}$	58.3$_{\pm 2.5}$	84.3$_{\pm 0.0}$	69.4$_{\pm 0.2}$	81.6$_{\pm 0.0}$	63.5$_{\pm 0.0}$	81.2$_{\pm 0.0}$	62.2$_{\pm 0.0}$
CRT	90.5$_{\pm 0.0}$	79.7$_{\pm 0.3}$	92.5$_{\pm 0.1}$	87.3$_{\pm 0.3}$	82.7$_{\pm 0.0}$	68.5$_{\pm 0.0}$	82.7$_{\pm 0.0}$	68.8$_{\pm 0.0}$	82.9$_{\pm 0.0}$	63.3$_{\pm 0.0}$
ReWeightCRT	91.3$_{\pm 0.1}$	78.4$_{\pm 0.1}$	92.5$_{\pm 0.2}$	87.2$_{\pm 0.5}$	83.0$_{\pm 0.0}$	69.5$_{\pm 0.0}$	82.4$_{\pm 0.0}$	68.3$_{\pm 0.0}$	82.9$_{\pm 0.0}$	63.3$_{\pm 0.0}$
JTT	88.8$_{\pm 0.7}$	84.5$_{\pm 0.3}$	90.6$_{\pm 2.2}$	87.2$_{\pm 7.5}$	85.1$_{\pm 0.0}$	70.4$_{\pm 0.0}$	84.6$_{\pm 0.0}$	68.5$_{\pm 0.0}$	83.7$_{\pm 0.0}$	66.1$_{\pm 0.0}$
GroupDRO	88.8$_{\pm 1.7}$	87.1$_{\pm 1.3}$	91.4$_{\pm 0.6}$	88.1$_{\pm 0.7}$	85.2$_{\pm 0.0}$	71.1$_{\pm 0.0}$	85.1$_{\pm 0.0}$	72.3$_{\pm 0.0}$	82.7$_{\pm 0.0}$	67.1$_{\pm 0.0}$
CVaRDRO	89.8$_{\pm 0.4}$	85.4$_{\pm 2.3}$	94.5$_{\pm 0.1}$	83.1$_{\pm 1.5}$	85.7$_{\pm 0.1}$	71.3$_{\pm 0.0}$	85.4$_{\pm 0.0}$	71.7$_{\pm 0.0}$	82.7$_{\pm 0.0}$	67.1$_{\pm 0.0}$
LfF	87.0$_{\pm 0.3}$	75.2$_{\pm 0.7}$	81.1$_{\pm 5.6}$	63.0$_{\pm 4.4}$	75.9$_{\pm 0.0}$	61.6$_{\pm 0.0}$	79.8$_{\pm 0.0}$	66.4$_{\pm 0.0}$	82.4$_{\pm 0.0}$	64.5$_{\pm 0.0}$
LISA	92.8$_{\pm 0.3}$	88.7$_{\pm 0.6}$	92.6$_{\pm 0.1}$	86.2$_{\pm 1.1}$	85.2$_{\pm 0.0}$	66.6$_{\pm 0.0}$	85.1$_{\pm 0.0}$	64.4$_{\pm 0.0}$	82.8$_{\pm 0.0}$	63.1$_{\pm 0.0}$
DFR	92.3$_{\pm 0.2}$	88.2$_{\pm 0.3}$	89.3$_{\pm 0.2}$	87.1$_{\pm 1.1}$	86.1$_{\pm 0.0}$	70.5$_{\pm 0.0}$	85.1$_{\pm 0.0}$	71.2$_{\pm 0.0}$	83.8$_{\pm 0.0}$	68.1$_{\pm 0.0}$
LADDER **(ours)**	**93.1**$_{\pm 0.8}$	**91.4**$_{\pm 0.8}$	89.8$_{\pm 1.2}$	**88.9**$_{\pm 0.4}$	86.8$_{\pm 0.0}$	**76.2**$_{\pm 0.0}$	85.3$_{\pm 0.0}$	**76.4**$_{\pm 0.0}$	86.2$_{\pm 0.0}$	**82.5**$_{\pm 0.0}$

Table 9: Benchmarking error mitigation methods over 3 seeds for ViT models pretrained with IN1k using the supervised method (RN Sup IN1k). We bold-face and underline the best and second-best results, respectively.

Method	**Waterbirds**		**CelebA**	
	Mean(%)	**WGA(%)**	**Mean(%)**	**WGA(%)**
Vanilla (ERM)	82.7$_{\pm 1.4}$	51.2$_{\pm 1.3}$	95.2$_{\pm 0.4}$	46.8$_{\pm 1.1}$
Mixup	81.8$_{\pm 0.4}$	44.9$_{\pm 0.3}$	**95.8**$_{\pm 0.3}$	48.3$_{\pm 0.3}$
IRM	79.8$_{\pm 0.3}$	54.5$_{\pm 0.3}$	85.1$_{\pm 1.2}$	48.7$_{\pm 0.3}$
MMD	83.6$_{\pm 2.7}$	42.5$_{\pm 1.1}$	95.6$_{\pm 0.4}$	54.2$_{\pm 0.4}$
JTT	81.7$_{\pm 0.5}$	49.1$_{\pm 0.5}$	94.8$_{\pm 0.3}$	52.7$_{\pm 0.6}$
GroupDRO	82.2$_{\pm 0.8}$	53.1 $_{\pm 1.2}$	93.5$_{\pm 0.1}$	80.1$_{\pm 0.4}$
CVaRDRO	83.5$_{\pm 0.3}$	46.6$_{\pm 2.8}$	95.6$_{\pm 0.1}$	55.1$_{\pm 1.8}$
LISA	83.7$_{\pm 0.1}$	48.8$_{\pm 0.1}$	95.6$_{\pm 0.2}$	60.2$_{\pm 0.1}$
DFR	85.0$_{\pm 0.3}$	76.2$_{\pm 0.3}$	91.3$_{\pm 1.1}$	81.1$_{\pm 0.5}$
LADDER **(ours)**	**85.3**$_{\pm 0.5}$	**86.5**$_{\pm 0.4}$	90.7$_{\pm 0.1}$	83.4$_{\pm 0.1}$

more accurate predictions, even in zero-shot settings where explicit training on the target data is absent. By leveraging these hypotheses, LADDER enables more precise alignment between image representations and class descriptions, significantly enhancing zero-shot performance.

A.12.6 CLIP score comparison of various attributes extracted by LADDER

Refer to Fig. 16 for the CLIP scores (discussed in Appendix A.6) of various attributes extracted from the hypotheses by LADDER. For *e.g.*, the correctly classified samples for the waterbird class in the Waterbirds dataset have a bias on the water-related backgrounds. As a result, the CLIP score of `ocean, boat, lake` is high. We observe consistent results for other datasets as well.

Table 10: Application: Boost in Zero-shot accuracy results using attributes from the hypotheses extracted from RN Sup IN1k (Resnet 50 pretrained with ImageNet 1K and supervised learning) classifier

Dataset	CLIP Prompts	LADDER Hypotheses	Gain
Waterbirds	50.40	**58.96**	+8.56 ↑
CelebA	86.69	**87.01**	+0.32 ↑
NIH	49.17	**68.22**	+19.05 ↑
RSNA	60.17	**65.98**	+5.81 ↑
VinDr	90.92	**92.33**	+1.41 ↑

A.12.7 Improvement on different slices of UrbanCars benchmark

Tab. 11 shows that LADDER achieves higher accuracy compared to the Whac-A-Mole method(Li et al., 2023b) across multiple shortcut benchmarks on the Urbancars dataset, without prior knowledge of the number or types of possible shortcuts.

Table 11: LADDER achieves higher accuracy compared to the Whac-A-Mole method (Li et al., 2023b) across multiple shortcut benchmarks on the Urbancars dataset without prior knowledge of the number or types of possible shortcuts.

Method	Mean Acc	BG gap	CoObj Gap	BG+CoObj Gap
ERM	96.4	-15.3	-11.2	-69.2
Whac-A-Mole	95.2	-2.4	-2.9	-5.8
LADDER	92.2	-1.1	-1.6	-3.8

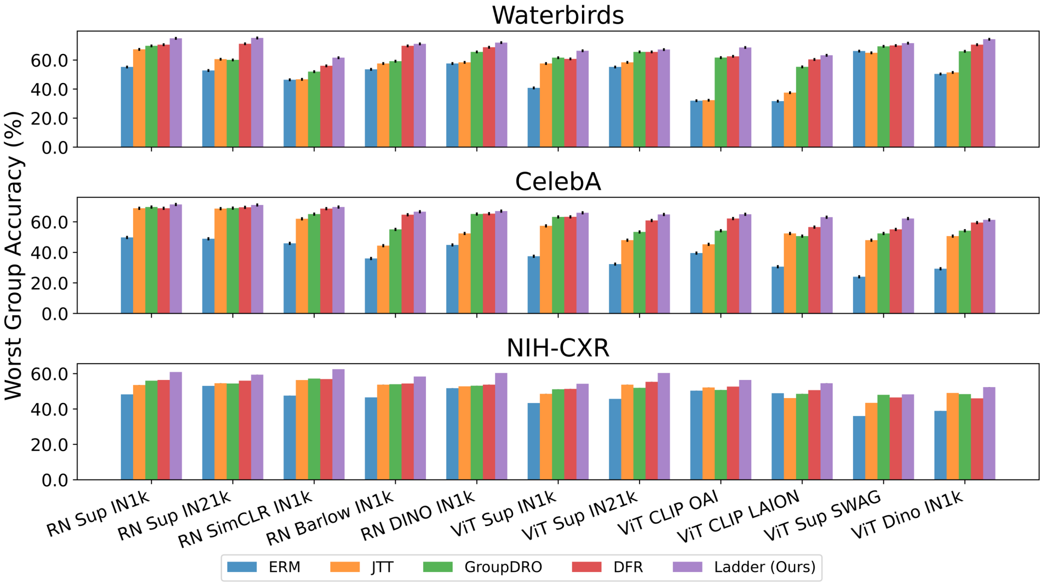


Figure 15: WGA across bias mitigation methods. LADDER consistently outperforms other bias mitigation baselines (ERM, JTT, GroupDRO, and DFR) across different model architectures and pre-training strategies.

A.12.8 Extended results on discovered hypothesis by LADDER for various architectures and pre-training methods

Fig. 17 illustrates additional results for the CelebA and Metashift datasets, demonstrating that LADDER accurately captures various sources of bias, regardless of the underlying architectures or pre-training methods.

A.12.9 Results on Imagenet

Tables 12, 13, 14 shows that LADDER identifies unique biases for the Imagenet dataset for a stethoscope, ant, and horizontal bar, respectively.

A.12.10 Ablation 1: WGA of LADDER using other captioning methods

Tab. 1 presents an ablation study evaluating the effect of various captioning models on LADDER's performance in mitigating biases. The quality of captions directly affects LADDER's ability to effectively generate hypotheses, as these captions are analyzed by LLMs to identify biased attributes contributing to model errors. LADDER then pseudo-labels these attributes to systematically mitigate the identified biases. We consider different captioning models, including BLIP (Li et al., 2022), BLIP2 (Li et al., 2023a), ClipCap (Mokady et al., 2021), and GPT-4o (Wu et al., 2024), with **ResNet Sup IN1k** as the classifier.

The results indicate that the more advanced captioning model, GPT-4o, significantly improves LADDER's performance, achieving the highest Worst Group Accuracy (WGA) and mean accuracy across both datasets. Specifically, GPT-4o achieves a WGA of 94.5% on Waterbirds and 91.9% on CelebA, which is substantially better than the other models. BLIP and BLIP2 demonstrate comparable results, with BLIP slightly outperforming BLIP2 in the Waterbirds dataset, while BLIP2 performs better on CelebA in WGA. In contrast, ClipCap consistently yields the lowest scores, implying that simpler captioning methods are less effective for enhancing LADDER's bias identification capabilities. Overall, the results underscore the importance of selecting a high-quality captioning model to maximize LADDER's effectiveness. While more sophisticated models like GPT-4o entail higher costs, their significant impact on bias mitigation performance, particularly on WGA, makes them an indispensable choice in scenarios where accuracy is critical.

A.12.11 Ablation 2: Slice discovery by LADDER using different LLMs

In this ablation study, we explore how different LLMs impact the effectiveness of LADDER in discovering data slices and generating hypotheses for bias identification. We aim to discover the biases from RN Sup IN1k classifier for natural images and CXRs, and EN-B5 classifier for mammograms. We utilize four LLMs: GPT-4o, Claude 3.5 Sonnet, LLaMA 3.1 70B, and Gemini 1.5 Pro. Fig. 18

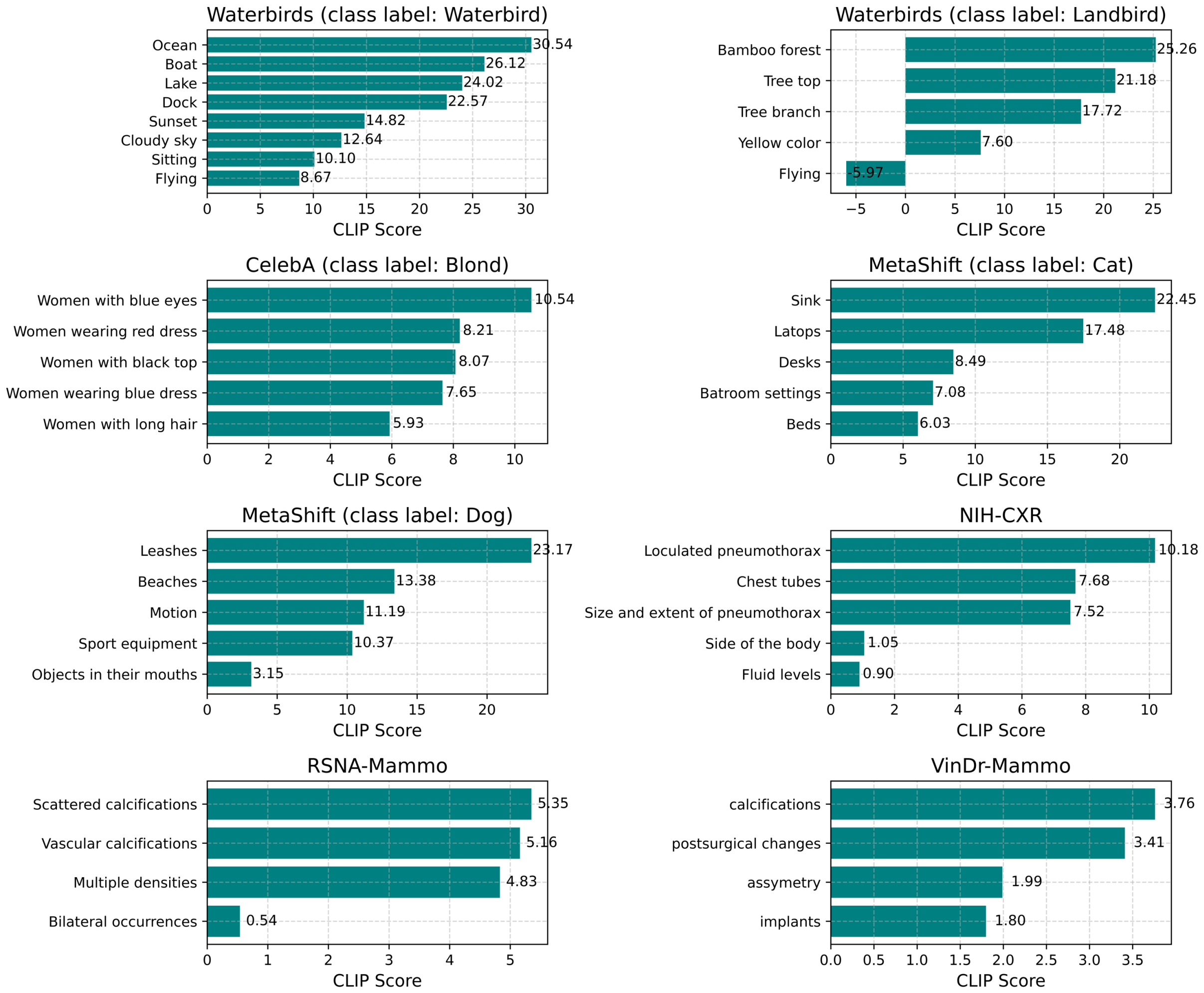


Figure 16: CLIP Score(Appendix A.6) for various attributes extracted from the hypotheses by LADDER. CLIP scores of the attributes are high signifying that they induce biases on the correctly classified samples.

illustrates the different attributes these models highlight across multiple datasets, including Waterbirds, CelebA, NIH, RSNA, VinDr, and MetaShift. Each LLM aims to extract a hypothesis related to an attribute, signifying the classifier's mistake. These attributes potentially lead to systematic model biases. As shown in Fig. 18, each LLM focuses on distinct subsets of attributes, reflecting their unique interpretation capabilities. Despite these differences, there is significant overlap in the overall hypotheses generated across the models, indicating consistency in identifying the attributes contributing to model errors.

For instance, in the Waterbirds dataset, all LLMs frequently highlight attributes like `ocean` and `boat` for the waterbird class and `bamboo forest` and `tree branch` for the landbird class. These attributes align closely with the ground truth bias in this dataset, which relates to water and land backgrounds being associated with the respective bird classes. This suggests that LLMs effectively identify these underlying environmental biases that lead to systematic errors. Similarly, in medical datasets, such as NIH-CXR for pneumothorax, all LLMs consistently highlight `chest tube` as a common attribute for misclassified samples. This reflects a true bias, as the presence of a chest tube often strongly correlates with pneumothorax cases. Identifying this attribute helps understand the systematic bias that models may develop when chest tubes are spuriously correlated in pneumothorax images.

This consistency across various LLMs demonstrates the robustness of LADDER for systematic bias detection, irrespective of the underlying LLM used. The results highlight that LADDER is effective at leveraging the strengths of different LLMs to produce meaningful insights into model behavior, regardless of which LLM is utilized. Moreover, it emphasizes the versatility of using LLMs for extracting domain-specific attributes—whether

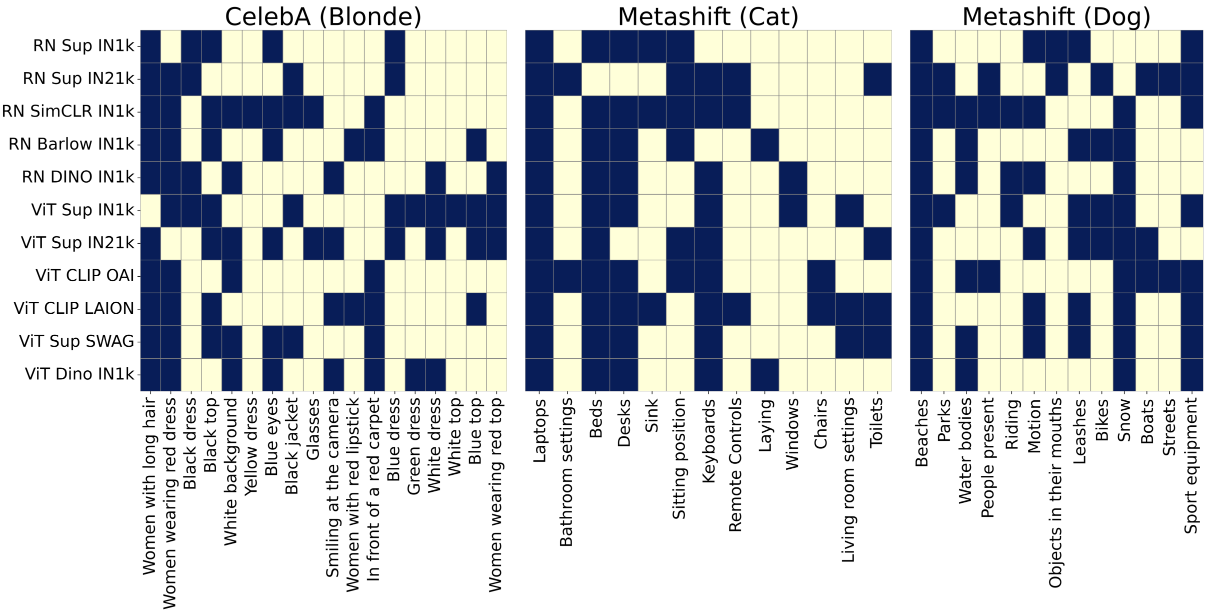


Figure 17: LADDER accurately captures various sources of bias, regardless of the underlying architectures or pre-training methods for the CelebA and Metashift datasets. Bright colors indicate attributes in LADDER's hypotheses, while light colors indicate their absence.

the focus is on natural images, chest X-rays, or mammography scans – while maintaining cost efficiency and avoiding manual annotation. Overall, this ablation shows that the specific choice of LLM slightly influences which attributes are emphasized, but all models effectively support the generation of comprehensive hypotheses that capture the biases inherent in different datasets.

A.12.12 Ablation 3: WGA by LADDER using the hypothesis by different LLMs

Fig. 10 illustrates the worst group accuracy (WGA) achieved across multiple datasets when utilizing LADDER to mitigate biases with different LLMs. The LLMs compared in this study include Claude 3.5 Sonnet, LLaMA 3.1 70B, Gemini 1.5 Pro, and GPT-4o. We consider the RN Sup IN1k classifier for natural images and CXRs, as well as the EN-B5 classifier for mammograms. The primary aim of this ablation is to assess how well LADDER can mitigate biases when generating hypotheses using different LLMs. As shown in Fig. 10, the WGA values remain consistently high across all LLMs, indicating that LADDER is effective in mitigating biases irrespective of the choice of LLM for hypothesis generation. Specifically, all LLMs achieve WGA scores of over 80% for most datasets, with only slight variations between models. This consistency demonstrates the robustness of LADDER in leveraging different LLMs to address model biases effectively. For datasets like Waterbirds and CelebA, the performance across all LLMs is nearly identical, suggesting that the generated hypotheses successfully capture the underlying biases and lead to similar improvements in fairness. In medical datasets, such as NIH and RSNA, the trend is also maintained, with LLMs like GPT-4o and Gemini 1.5 Pro achieving better results than other LLMs. These findings emphasize that the specific choice of LLM has only a minor impact on the overall ability of LADDER to mitigate bias. This makes LADDER a flexible and cost-effective solution, as it can work effectively with a range of LLMs, each with different computational costs and capabilities. Using different LLMs ensures flexibility based on resource availability while effectively identifying and mitigating dataset biases.

A.12.13 Ablation 4: Overall cost and choice of LLMs

Tab. 7 shows the cost of using various LLMs. Each row shows the total breakdown for an LLM extracting hypotheses across all 6 datasets, using RN Sup IN1k (natural images or CXRs) and EN-B5 (mammograms). LADDER invokes LLM once using sentences only (no images). The total cost incurred is ∼\$28 across all architectures and pretraining used in the experiments. Thus, LLMs are far more cost-effective than developing new tagging models for unexplored domains *e.g.,* radiology, or manually annotating shortcuts. Fig. 18 in Appendix A.12.11 shows the attributes identified by each LLM while generating hypotheses. Different LLMs capture distinct sets of attributes, yet substantial overlap exists, with many attributes consistently revealing actual biases across models. Ablation studies in Appendix A.12.12 indicate that using different LLMs to compute WGA shows that Gemini and GPT-4o achieve higher WGA for medical images than the others.

A.12.14 Ablation 5: Choice of VLR on LADDER

Fig.14 demonstrates that LADDER consistently detects well-known biases in CXRs, such as `chest tube`, across various VLRs (CXR-CLIP (SwinT), GLORIA (Huang et al., 2021), MedCLIP (SwinT) (Wang et al., 2022), and MedKLIP (SwinT) (Wu et al., 2023)) on the NIH dataset. This consistency suggests that the choice of VLR does not significantly impact LADDER's ability to identify biased attributes.

Table 12: LADDER identifies unique biases in **ImageNet** for the "Stethoscope" class. The table shows accuracy for subpopulations where the hypothesis failed (Error Slice) and where it passed (Bias-Aligned).

Biases	**Accuracy of the subpopulation where hypothesis failed (Error Slice) (%)**	**Accuracy of the subpopulation where hypothesis passed (Bias-Aligned) (%)**
Littmann branding	51.3	95.2
Dual-head stethoscopes	53.7	95.2
Medical settings	51.3	93.3
Colors *e.g.,* yellow or copper	55.6	87.8
Children interacting with stethoscopes	58.2	93.6

Table 13: LADDER identifies unique biases in **ImageNet** for the "Ant" class. The table shows accuracy for subpopulations where the hypothesis failed (Error Slice) and where it passed (Bias-Aligned).

Biases	**Accuracy of the subpopulation where hypothesis failed (Error Slice) (%)**	**Accuracy of the subpopulation where hypothesis passed (Bias-Aligned) (%)**
Close up settings	62.6	73.3
Textured surface	59.6	74.5
Green Leaves	67.5	76.5
Yellow flower	62.4	69.8
Black ant	63.8	73.1

Table 14: LADDER identifies unique biases in **ImageNet** for the "Horizontal bar" class. The table shows accuracy for subpopulations where the hypothesis failed (Error Slice) and where it passed (Bias-Aligned).

Biases	**Accuracy of the subpopulation where hypothesis failed (Error Slice) (%)**	**Accuracy of the subpopulation where hypothesis passed (Bias-Aligned) (%)**
Child	66.4	82.4
Playground	61.4	82.7
Green Leaves	67.7	76.5
Yellow flower	62.5	69.8
Black ant	63.5	73.8

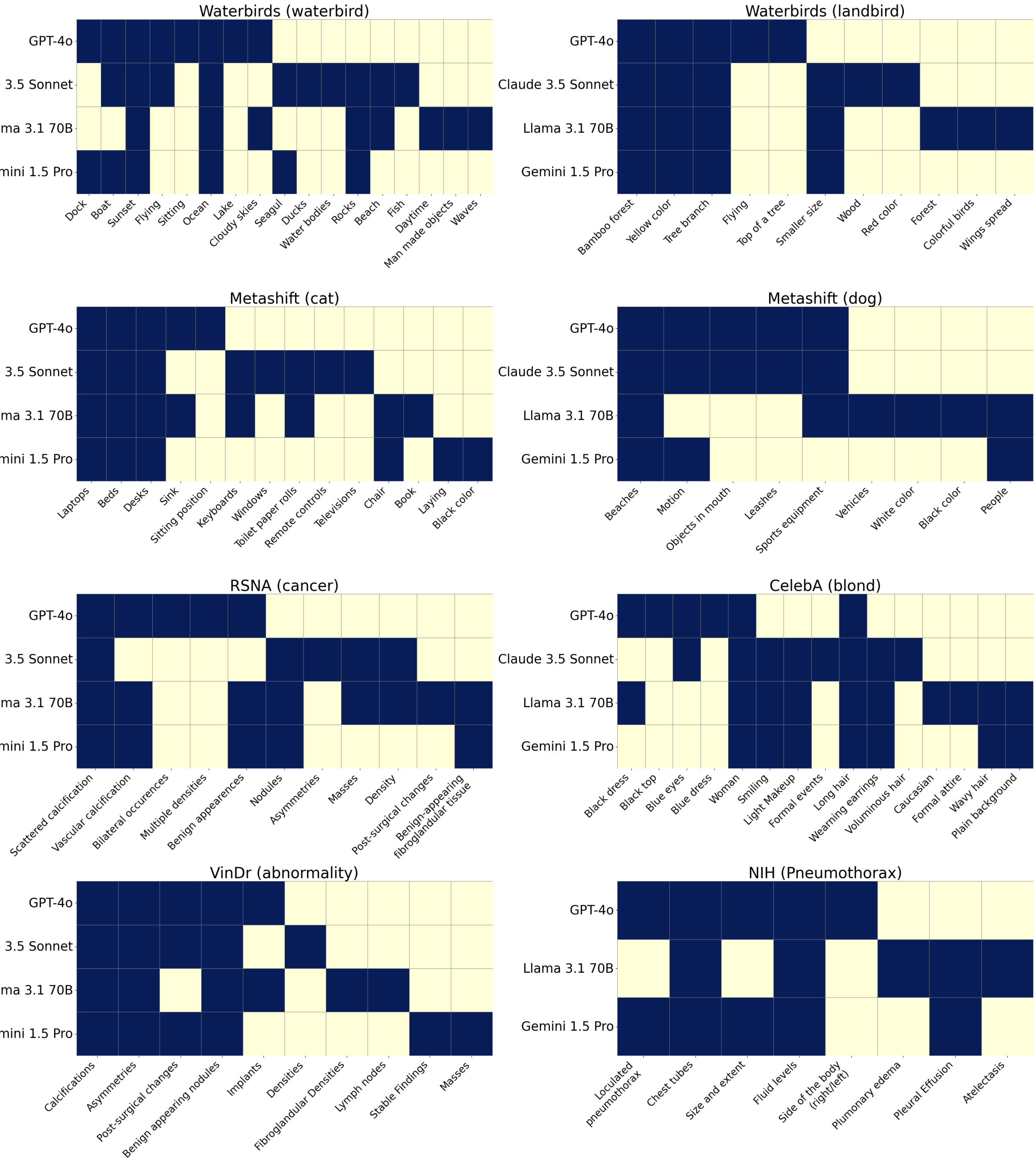
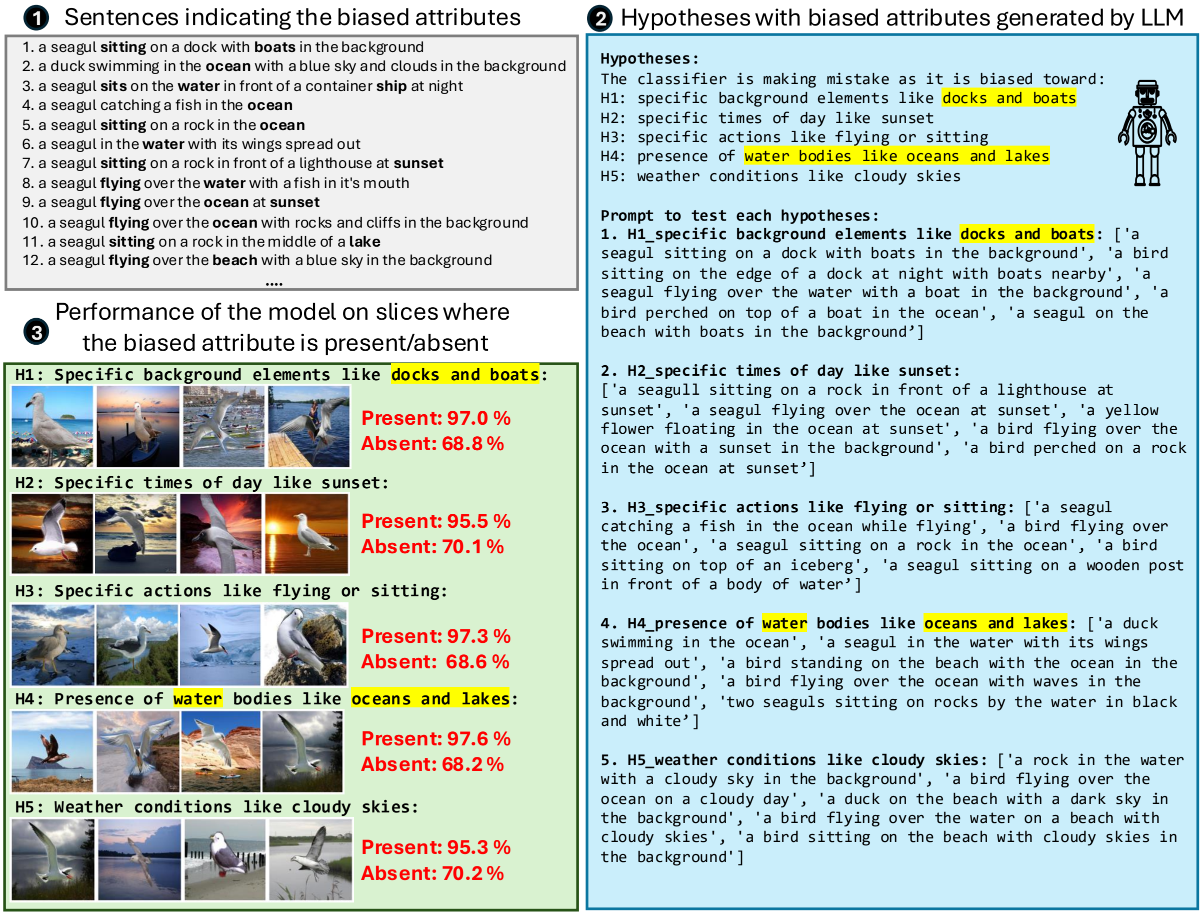


Figure 18: Ablation 2: Attributes identified by different LLMs while generating hypotheses across datasets for bias identification: RN Sup IN1k for natural images and CXRs, and EN-B5 for mammograms. Each LLM (GPT-4o, Claude 3.5 Sonnet, LlaMA 3.1 70B, Gemini 1.5 Pro) focuses on distinct attributes, yet the overall hypotheses are consistent across datasets, showing LADDER's robust bias detection. Bright colors indicate attributes in LADDER's hypotheses, while light colors indicate their absence. Following MIMIC's regulations, we use Gemini 1.5 Pro (via Vertex AI on Google Cloud Platform), GPT-4o via Azure OpenAI service, and Llama 3.1 70B (running locally) for NIH. Bright colors indicate attributes in LADDER's hypotheses.



❶ Sentences indicating the biased attributes

1. a seagul **sitting** on a dock with **boats** in the background
2. a duck swimming in the **ocean** with a blue sky and clouds in the background
3. a seagul **sits** on the **water** in front of a container **ship** at night
4. a seagul catching a fish in the **ocean**
5. a seagul **sitting** on a rock in the **ocean**
6. a seagul in the **water** with its wings spread out
7. a seagul **sitting** on a rock in front of a lighthouse at **sunset**
8. a seagul **flying** over the **water** with a fish in it's mouth
9. a seagul **flying** over the **ocean** at **sunset**
10. a seagul **flying** over the **ocean** with rocks and cliffs in the background
11. a seagul **sitting** on a rock in the middle of a **lake**
12. a seagul **flying** over the **beach** with a blue sky in the background

....

❷ Hypotheses with biased attributes generated by LLM

Hypotheses:
The classifier is making mistake as it is biased toward:
H1: specific background elements like docks and boats
H2: specific times of day like sunset
H3: specific actions like flying or sitting
H4: presence of water bodies like oceans and lakes
H5: weather conditions like cloudy skies

Prompt to test each hypotheses:

1. H1_specific background elements like docks and boats: ['a seagul sitting on a dock with boats in the background', 'a bird sitting on the edge of a dock at night with boats nearby', 'a seagul flying over the water with a boat in the background', 'a bird perched on top of a boat in the ocean', 'a seagul on the beach with boats in the background']

2. H2_specific times of day like sunset: ['a seagull sitting on a rock in front of a lighthouse at sunset', 'a seagul flying over the ocean at sunset', 'a yellow flower floating in the ocean at sunset', 'a bird flying over the ocean with a sunset in the background', 'a bird perched on a rock in the ocean at sunset']

3. H3_specific actions like flying or sitting: ['a seagul catching a fish in the ocean while flying', 'a bird flying over the ocean', 'a seagul sitting on a rock in the ocean', 'a bird sitting on top of an iceberg', 'a seagul sitting on a wooden post in front of a body of water']

4. H4_presence of water bodies like oceans and lakes: ['a duck swimming in the ocean', 'a seagul in the water with its wings spread out', 'a bird standing on the beach with the ocean in the background', 'a bird flying over the ocean with waves in the background', 'two seaguls sitting on rocks by the water in black and white']

5. H5_weather conditions like cloudy skies: ['a rock in the water with a cloudy sky in the background', 'a bird flying over the ocean on a cloudy day', 'a duck on the beach with a dark sky in the background', 'a bird flying over the water on a beach with cloudy skies', 'a bird sitting on the beach with cloudy skies in the background']

❸ Performance of the model on slices where the biased attribute is present/absent

H1: Specific background elements like docks and boats:
Present: 97.0 %
Absent: 68.8 %

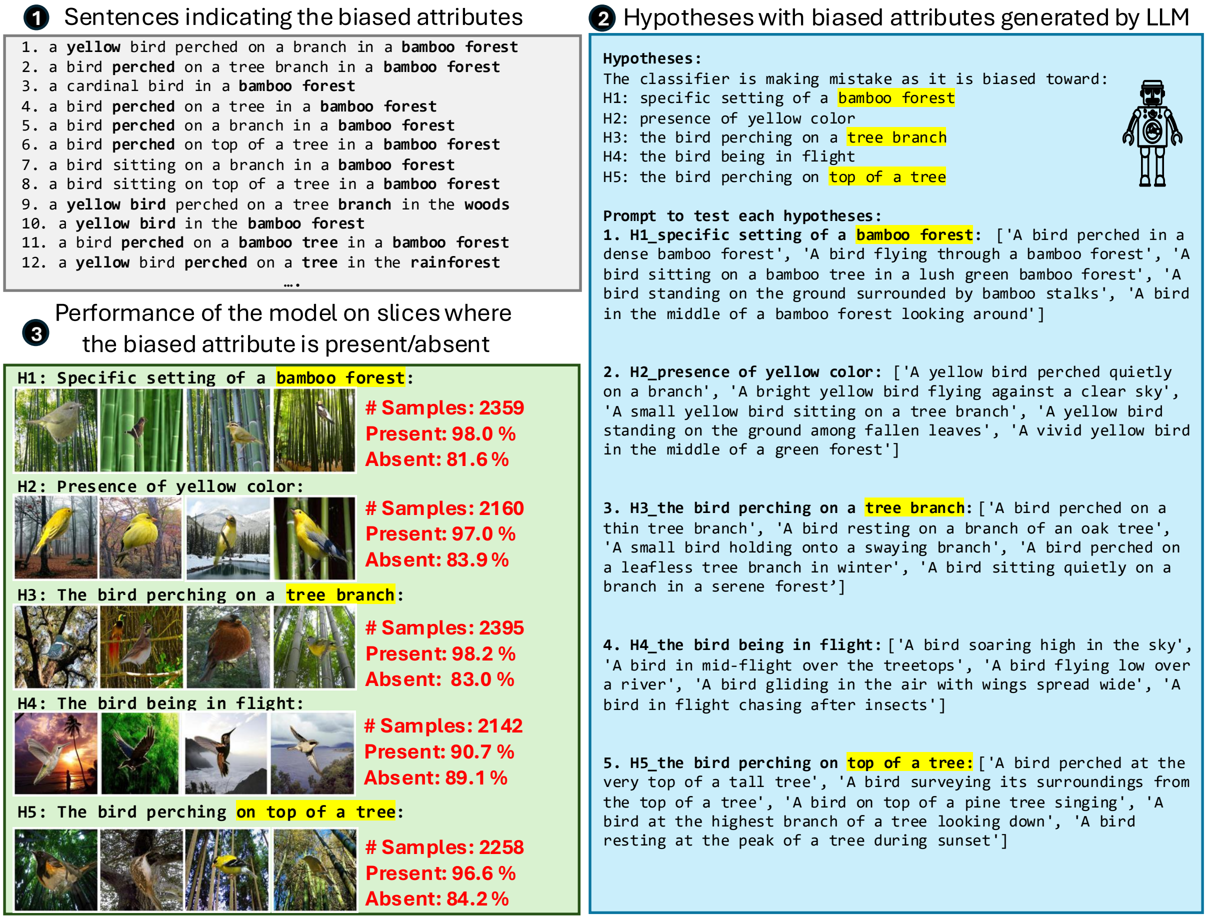
H2: Specific times of day like sunset:
Present: 95.5 %
Absent: 70.1 %

H3: Specific actions like flying or sitting:
Present: 97.3 %
Absent: 68.6 %

H4: Presence of water bodies like oceans and lakes:
Present: 97.6 %
Absent: 68.2 %

H5: Weather conditions like cloudy skies:
Present: 95.3 %
Absent: 70.2 %

Figure 19: LADDER discovers slices for biased attributes in RN Sup IN1k-based classifier for *waterbird* classification in **Waterbirds** dataset. This figure details the slice discovery process for biased attributes involving sentence analysis, hypothesis generation by an LLM, and the model's performance on slices where attributes are present or absent, demonstrating how biases affect classifier accuracy. We highlight the hypothesis generated by LADDER that corresponds to the ground truth biased attribute (*e.g.,* `water` for landbirds) in **yellow**.



1 Sentences indicating the biased attributes

1. a **yellow** bird perched on a branch in a **bamboo forest**
2. a bird **perched** on a tree branch in a **bamboo forest**
3. a cardinal bird in a **bamboo forest**
4. a bird **perched** on a tree in a **bamboo forest**
5. a bird **perched** on a branch in a **bamboo forest**
6. a bird **perched** on top of a tree in a **bamboo forest**
7. a bird sitting on a branch in a **bamboo forest**
8. a bird sitting on top of a tree in a **bamboo forest**
9. a **yellow bird** perched on a tree **branch** in the **woods**
10. a **yellow bird** in the **bamboo forest**
11. a bird **perched** on a **bamboo tree** in a **bamboo forest**
12. a **yellow** bird **perched** on a **tree** in the **rainforest**

….

2 Hypotheses with biased attributes generated by LLM

Hypotheses:
The classifier is making mistake as it is biased toward:
H1: specific setting of a bamboo forest
H2: presence of yellow color
H3: the bird perching on a tree branch
H4: the bird being in flight
H5: the bird perching on top of a tree

Prompt to test each hypotheses:
1. H1_specific setting of a bamboo forest: ['A bird perched in a dense bamboo forest', 'A bird flying through a bamboo forest', 'A bird sitting on a bamboo tree in a lush green bamboo forest', 'A bird standing on the ground surrounded by bamboo stalks', 'A bird in the middle of a bamboo forest looking around']

2. H2_presence of yellow color: ['A yellow bird perched quietly on a branch', 'A bright yellow bird flying against a clear sky', 'A small yellow bird sitting on a tree branch', 'A yellow bird standing on the ground among fallen leaves', 'A vivid yellow bird in the middle of a green forest']

3. H3_the bird perching on a tree branch: ['A bird perched on a thin tree branch', 'A bird resting on a branch of an oak tree', 'A small bird holding onto a swaying branch', 'A bird perched on a leafless tree branch in winter', 'A bird sitting quietly on a branch in a serene forest’]

4. H4_the bird being in flight: ['A bird soaring high in the sky', 'A bird in mid-flight over the treetops', 'A bird flying low over a river', 'A bird gliding in the air with wings spread wide', 'A bird in flight chasing after insects']

5. H5_the bird perching on top of a tree: ['A bird perched at the very top of a tall tree', 'A bird surveying its surroundings from the top of a tree', 'A bird on top of a pine tree singing', 'A bird at the highest branch of a tree looking down', 'A bird resting at the peak of a tree during sunset']

3 Performance of the model on slices where the biased attribute is present/absent

H1: Specific setting of a bamboo forest:
# Samples: 2359
Present: 98.0 %
Absent: 81.6 %

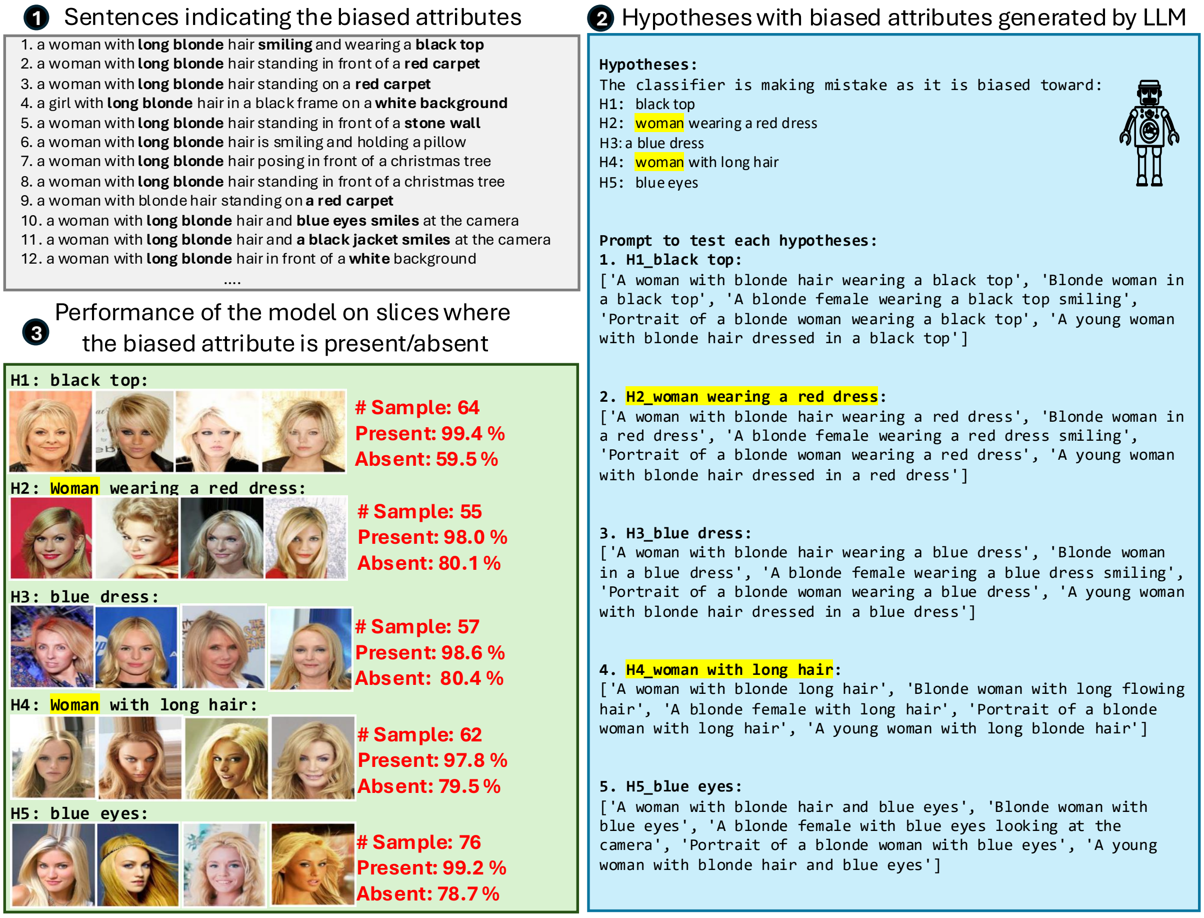
H2: Presence of yellow color:
# Samples: 2160
Present: 97.0 %
Absent: 83.9 %

H3: The bird perching on a tree branch:
# Samples: 2395
Present: 98.2 %
Absent: 83.0 %

H4: The bird being in flight:
# Samples: 2142
Present: 90.7 %
Absent: 89.1 %

H5: The bird perching on top of a tree:
# Samples: 2258
Present: 96.6 %
Absent: 84.2 %

Figure 20: LADDER discovers slices for biased attributes in RN Sup IN1k-based classifier for *landbird* classification in **Waterbirds** dataset. This figure details the slice discovery process for biased attributes involving sentence analysis, hypothesis generation by an LLM, and the model's performance on slices where attributes are present or absent, demonstrating how biases affect classifier accuracy. We highlight the hypothesis generated by LADDER that corresponds to the ground truth biased attribute (*e.g.,* `land` for landbirds) in **yellow**.



1 Sentences indicating the biased attributes

1. a woman with **long blonde** hair **smiling** and wearing a **black top**
2. a woman with **long blonde** hair standing in front of a **red carpet**
3. a woman with **long blonde** hair standing on a **red carpet**
4. a girl with **long blonde** hair in a black frame on a **white background**
5. a woman with **long blonde** hair standing in front of a **stone wall**
6. a woman with **long blonde** hair is smiling and holding a pillow
7. a woman with **long blonde** hair posing in front of a christmas tree
8. a woman with **long blonde** hair standing in front of a christmas tree
9. a woman with blonde hair standing on **a red carpet**
10. a woman with **long blonde** hair and **blue eyes smiles** at the camera
11. a woman with **long blonde** hair and **a black jacket smiles** at the camera
12. a woman with **long blonde** hair in front of a **white** background

....

2 Hypotheses with biased attributes generated by LLM

```
Hypotheses:
The classifier is making mistake as it is biased toward:
H1:  black top
H2:  woman wearing a red dress
H3: a blue dress
H4:  woman with long hair
H5:  blue eyes

Prompt to test each hypotheses:
1. H1_black top:
['A woman with blonde hair wearing a black top', 'Blonde woman in
a black top', 'A blonde female wearing a black top smiling',
'Portrait of a blonde woman wearing a black top', 'A young woman
with blonde hair dressed in a black top']

2. H2_woman wearing a red dress:
['A woman with blonde hair wearing a red dress', 'Blonde woman in
a red dress', 'A blonde female wearing a red dress smiling',
'Portrait of a blonde woman wearing a red dress', 'A young woman
with blonde hair dressed in a red dress']

3. H3_blue dress:
['A woman with blonde hair wearing a blue dress', 'Blonde woman
in a blue dress', 'A blonde female wearing a blue dress smiling',
'Portrait of a blonde woman wearing a blue dress', 'A young woman
with blonde hair dressed in a blue dress']

4. H4_woman with long hair:
['A woman with blonde long hair', 'Blonde woman with long flowing
hair', 'A blonde female with long hair', 'Portrait of a blonde
woman with long hair', 'A young woman with long blonde hair']

5. H5_blue eyes:
['A woman with blonde hair and blue eyes', 'Blonde woman with
blue eyes', 'A blonde female with blue eyes looking at the
camera', 'Portrait of a blonde woman with blue eyes', 'A young
woman with blonde hair and blue eyes']
```

3 Performance of the model on slices where the biased attribute is present/absent

H1: black top:
# Sample: 64
Present: 99.4 %
Absent: 59.5 %

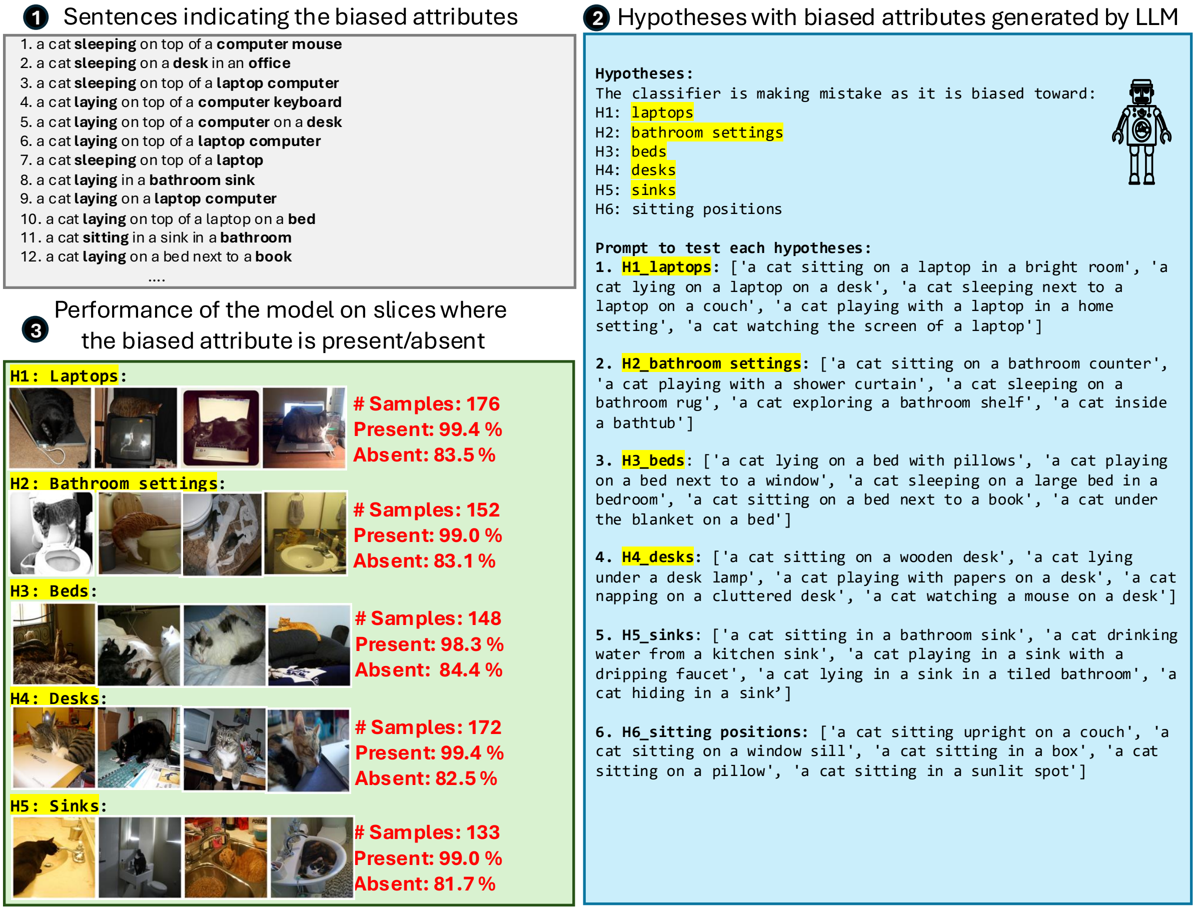
H2: Woman wearing a red dress:
# Sample: 55
Present: 98.0 %
Absent: 80.1 %

H3: blue dress:
# Sample: 57
Present: 98.6 %
Absent: 80.4 %

H4: Woman with long hair:
# Sample: 62
Present: 97.8 %
Absent: 79.5 %

H5: blue eyes:
# Sample: 76
Present: 99.2 %
Absent: 78.7 %

Figure 21: LADDER discovers slices for biased attributes in RN Sup IN1k-based classifier for *blond* classification in **CelebA** dataset. This figure details the slice discovery process for biased attributes involving sentence analysis, hypothesis generation by an LLM, and the model's performance on slices where attributes are present or absent, demonstrating how biases affect classifier accuracy. We highlight the hypothesis generated by LADDER that corresponds to the ground truth biased attribute (*e.g.,* woman for blond) in **yellow**.



❶ Sentences indicating the biased attributes

1. a cat **sleeping** on top of a **computer mouse**
2. a cat **sleeping** on a **desk** in an **office**
3. a cat **sleeping** on top of a **laptop computer**
4. a cat **laying** on top of a **computer keyboard**
5. a cat **laying** on top of a **computer** on a **desk**
6. a cat **laying** on top of a **laptop computer**
7. a cat **sleeping** on top of a **laptop**
8. a cat **laying** in a **bathroom sink**
9. a cat **laying** on a **laptop computer**
10. a cat **laying** on top of a laptop on a **bed**
11. a cat **sitting** in a sink in a **bathroom**
12. a cat **laying** on a bed next to a **book**

....

❷ Hypotheses with biased attributes generated by LLM

```
Hypotheses:
The classifier is making mistake as it is biased toward:
H1: laptops
H2: bathroom settings
H3: beds
H4: desks
H5: sinks
H6: sitting positions

Prompt to test each hypotheses:
1. H1_laptops: ['a cat sitting on a laptop in a bright room', 'a
cat lying on a laptop on a desk', 'a cat sleeping next to a
laptop on a couch', 'a cat playing with a laptop in a home
setting', 'a cat watching the screen of a laptop']

2. H2_bathroom settings: ['a cat sitting on a bathroom counter',
'a cat playing with a shower curtain', 'a cat sleeping on a
bathroom rug', 'a cat exploring a bathroom shelf', 'a cat inside
a bathtub']

3. H3_beds: ['a cat lying on a bed with pillows', 'a cat playing
on a bed next to a window', 'a cat sleeping on a large bed in a
bedroom', 'a cat sitting on a bed next to a book', 'a cat under
the blanket on a bed']

4. H4_desks: ['a cat sitting on a wooden desk', 'a cat lying
under a desk lamp', 'a cat playing with papers on a desk', 'a cat
napping on a cluttered desk', 'a cat watching a mouse on a desk']

5. H5_sinks: ['a cat sitting in a bathroom sink', 'a cat drinking
water from a kitchen sink', 'a cat playing in a sink with a
dripping faucet', 'a cat lying in a sink in a tiled bathroom', 'a
cat hiding in a sink’]

6. H6_sitting positions: ['a cat sitting upright on a couch', 'a
cat sitting on a window sill', 'a cat sitting in a box', 'a cat
sitting on a pillow', 'a cat sitting in a sunlit spot']
```

❸ Performance of the model on slices where the biased attribute is present/absent

H1: Laptops:
# Samples: 176
Present: 99.4 %
Absent: 83.5 %

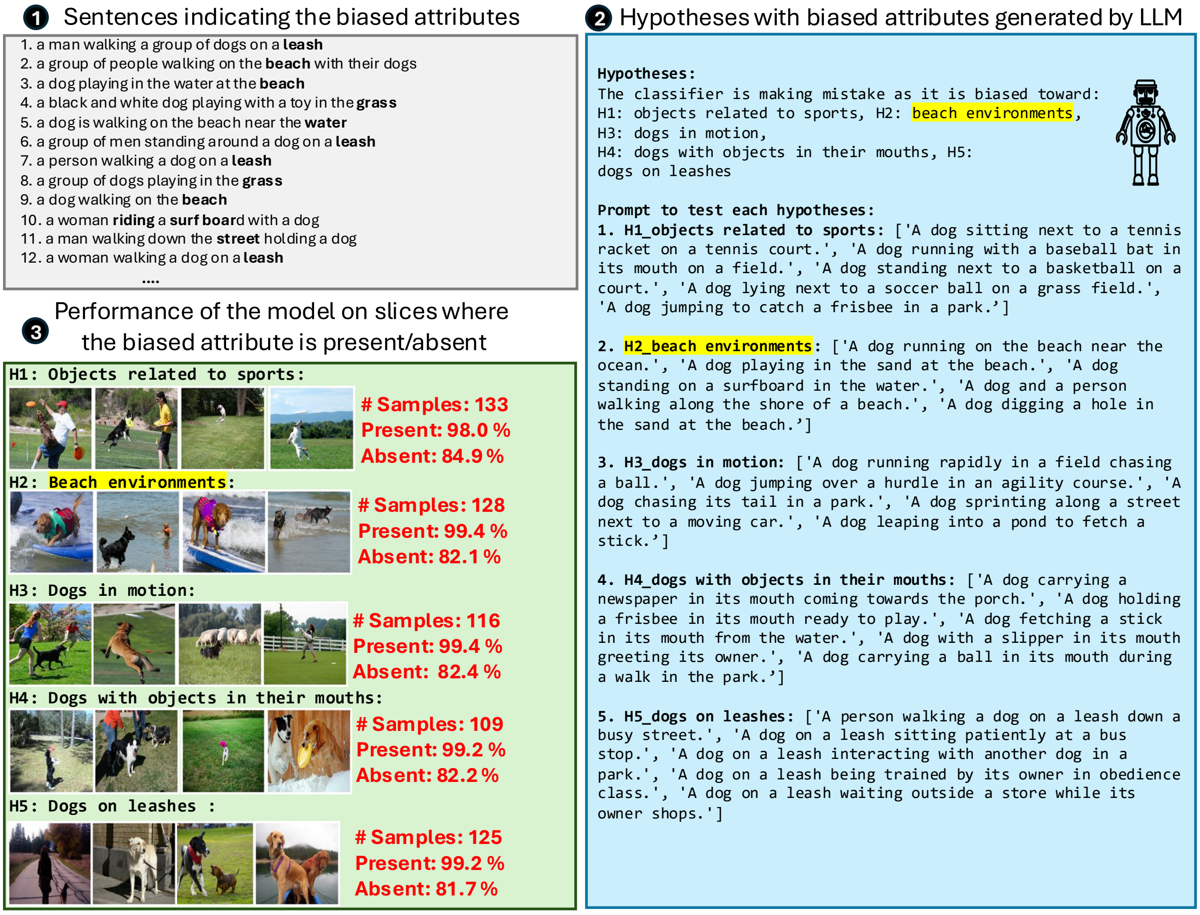
H2: Bathroom settings:
# Samples: 152
Present: 99.0 %
Absent: 83.1 %

H3: Beds:
# Samples: 148
Present: 98.3 %
Absent: 84.4 %

H4: Desks:
# Samples: 172
Present: 99.4 %
Absent: 82.5 %

H5: Sinks:
# Samples: 133
Present: 99.0 %
Absent: 81.7 %

Figure 22: LADDER discovers slices for biased attributes in RN Sup IN1k-based classifier for *cat* classification in **MetaShift** dataset. This figure details the slice discovery process for biased attributes involving sentence analysis, hypothesis generation by an LLM, and the model's performance on slices where attributes are present or absent, demonstrating how biases affect classifier accuracy. We highlight the hypothesis generated by LADDER that corresponds to the ground truth biased attribute (*e.g.,* `indoor` for cat) in **yellow**.



❶ Sentences indicating the biased attributes

1. a man walking a group of dogs on a **leash**
2. a group of people walking on the **beach** with their dogs
3. a dog playing in the water at the **beach**
4. a black and white dog playing with a toy in the **grass**
5. a dog is walking on the beach near the **water**
6. a group of men standing around a dog on a **leash**
7. a person walking a dog on a **leash**
8. a group of dogs playing in the **grass**
9. a dog walking on the **beach**
10. a woman **riding** a **surf board** with a dog
11. a man walking down the **street** holding a dog
12. a woman walking a dog on a **leash**

....

❷ Hypotheses with biased attributes generated by LLM

Hypotheses:
The classifier is making mistake as it is biased toward:
H1: objects related to sports, H2: beach environments,
H3: dogs in motion,
H4: dogs with objects in their mouths, H5:
dogs on leashes

Prompt to test each hypotheses:

1. H1_objects related to sports: ['A dog sitting next to a tennis racket on a tennis court.', 'A dog running with a baseball bat in its mouth on a field.', 'A dog standing next to a basketball on a court.', 'A dog lying next to a soccer ball on a grass field.', 'A dog jumping to catch a frisbee in a park.’]

2. H2_beach environments: ['A dog running on the beach near the ocean.', 'A dog playing in the sand at the beach.', 'A dog standing on a surfboard in the water.', 'A dog and a person walking along the shore of a beach.', 'A dog digging a hole in the sand at the beach.’]

3. H3_dogs in motion: ['A dog running rapidly in a field chasing a ball.', 'A dog jumping over a hurdle in an agility course.', 'A dog chasing its tail in a park.', 'A dog sprinting along a street next to a moving car.', 'A dog leaping into a pond to fetch a stick.’]

4. H4_dogs with objects in their mouths: ['A dog carrying a newspaper in its mouth coming towards the porch.', 'A dog holding a frisbee in its mouth ready to play.', 'A dog fetching a stick in its mouth from the water.', 'A dog with a slipper in its mouth greeting its owner.', 'A dog carrying a ball in its mouth during a walk in the park.’]

5. H5_dogs on leashes: ['A person walking a dog on a leash down a busy street.', 'A dog on a leash sitting patiently at a bus stop.', 'A dog on a leash interacting with another dog in a park.', 'A dog on a leash being trained by its owner in obedience class.', 'A dog on a leash waiting outside a store while its owner shops.']

❸ Performance of the model on slices where the biased attribute is present/absent

H1: Objects related to sports:
# Samples: 133
Present: 98.0 %
Absent: 84.9 %

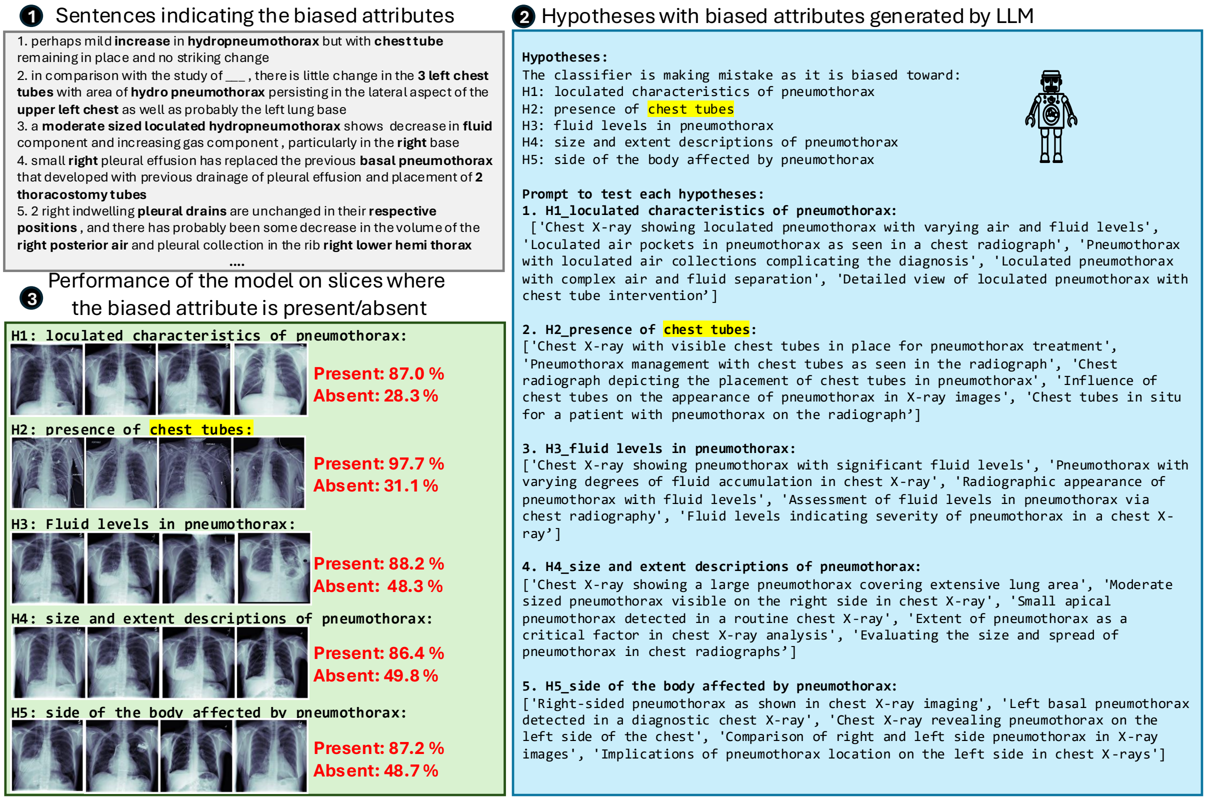
H2: Beach environments:
# Samples: 128
Present: 99.4 %
Absent: 82.1 %

H3: Dogs in motion:
# Samples: 116
Present: 99.4 %
Absent: 82.4 %

H4: Dogs with objects in their mouths:
# Samples: 109
Present: 99.2 %
Absent: 82.2 %

H5: Dogs on leashes :
# Samples: 125
Present: 99.2 %
Absent: 81.7 %

Figure 23: LADDER discovers slices for biased attributes in RN Sup IN1k-based classifier for *dog* classification in **MetaShift** dataset. This figure details the slice discovery process for biased attributes involving sentence analysis, hypothesis generation by an LLM, and the model's performance on slices where attributes are present or absent, demonstrating how biases affect classifier accuracy. We highlight the hypothesis generated by LADDER that corresponds to the ground truth biased attribute (*e.g.,* outdoor for cat) in **yellow**.



❶ Sentences indicating the biased attributes

1. perhaps mild **increase** in **hydropneumothorax** but with **chest tube** remaining in place and no striking change
2. in comparison with the study of ___ , there is little change in the **3 left chest tubes** with area of **hydro pneumothorax** persisting in the lateral aspect of the **upper left chest** as well as probably the left lung base
3. a **moderate sized loculated hydropneumothorax** shows decrease in **fluid** component and increasing gas component , particularly in the **right** base
4. small **right** pleural effusion has replaced the previous **basal pneumothorax** that developed with previous drainage of pleural effusion and placement of **2 thoracostomy tubes**
5. 2 right indwelling **pleural drains** are unchanged in their **respective positions** , and there has probably been some decrease in the volume of the **right posterior air** and pleural collection in the rib **right lower hemi thorax**

....

❷ Hypotheses with biased attributes generated by LLM

```
Hypotheses:
The classifier is making mistake as it is biased toward:
H1: loculated characteristics of pneumothorax
H2: presence of chest tubes
H3: fluid levels in pneumothorax
H4: size and extent descriptions of pneumothorax
H5: side of the body affected by pneumothorax

Prompt to test each hypotheses:
1. H1_loculated characteristics of pneumothorax:
 ['Chest X-ray showing loculated pneumothorax with varying air and fluid levels',
'Loculated air pockets in pneumothorax as seen in a chest radiograph', 'Pneumothorax
with loculated air collections complicating the diagnosis', 'Loculated pneumothorax
with complex air and fluid separation', 'Detailed view of loculated pneumothorax with
chest tube intervention’]

2. H2_presence of chest tubes:
['Chest X-ray with visible chest tubes in place for pneumothorax treatment',
'Pneumothorax management with chest tubes as seen in the radiograph', 'Chest
radiograph depicting the placement of chest tubes in pneumothorax', 'Influence of
chest tubes on the appearance of pneumothorax in X-ray images', 'Chest tubes in situ
for a patient with pneumothorax on the radiograph’]

3. H3_fluid levels in pneumothorax:
['Chest X-ray showing pneumothorax with significant fluid levels', 'Pneumothorax with
varying degrees of fluid accumulation in chest X-ray', 'Radiographic appearance of
pneumothorax with fluid levels', 'Assessment of fluid levels in pneumothorax via
chest radiography', 'Fluid levels indicating severity of pneumothorax in a chest X-
ray’]

4. H4_size and extent descriptions of pneumothorax:
['Chest X-ray showing a large pneumothorax covering extensive lung area', 'Moderate
sized pneumothorax visible on the right side in chest X-ray', 'Small apical
pneumothorax detected in a routine chest X-ray', 'Extent of pneumothorax as a
critical factor in chest X-ray analysis', 'Evaluating the size and spread of
pneumothorax in chest radiographs’]

5. H5_side of the body affected by pneumothorax:
['Right-sided pneumothorax as shown in chest X-ray imaging', 'Left basal pneumothorax
detected in a diagnostic chest X-ray', 'Chest X-ray revealing pneumothorax on the
left side of the chest', 'Comparison of right and left side pneumothorax in X-ray
images', 'Implications of pneumothorax location on the left side in chest X-rays']
```

❸ Performance of the model on slices where the biased attribute is present/absent

H1: loculated characteristics of pneumothorax:

Present: 87.0 %
Absent: 28.3 %

H2: presence of chest tubes:

Present: 97.7 %
Absent: 31.1 %

H3: Fluid levels in pneumothorax:

Present: 88.2 %
Absent: 48.3 %

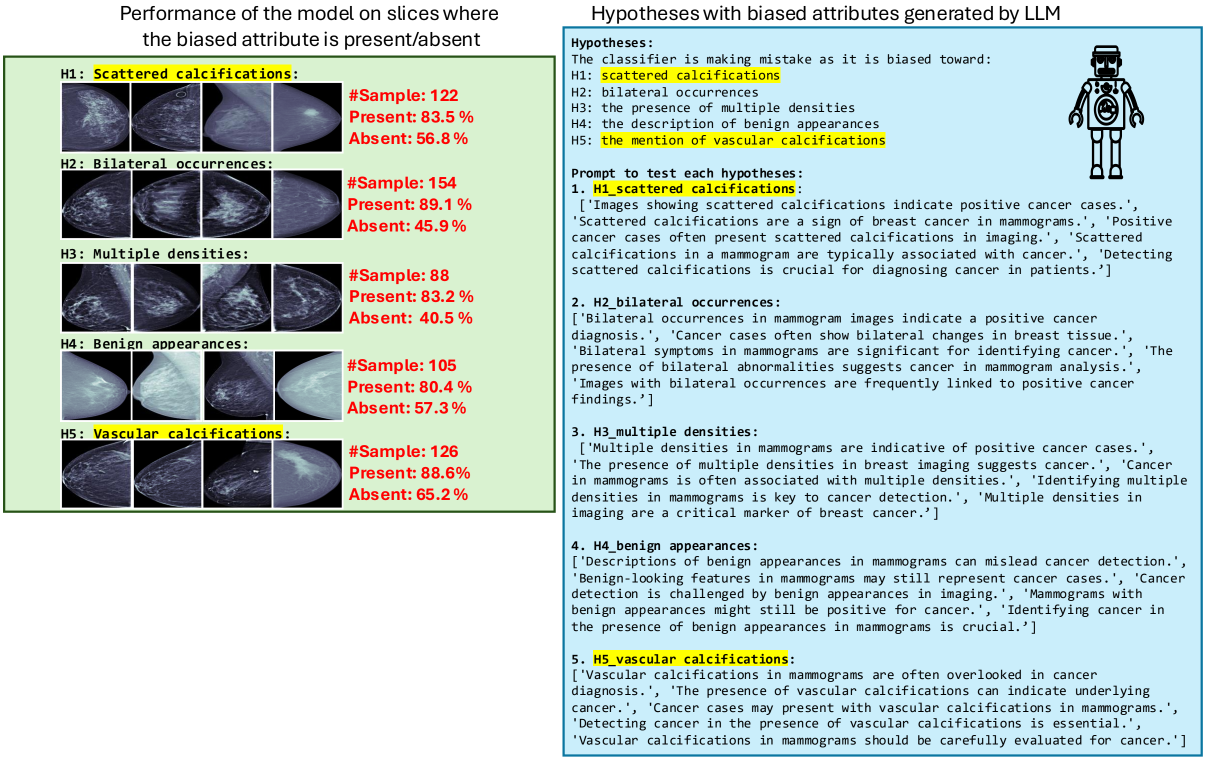
H4: size and extent descriptions of pneumothorax:

Present: 86.4 %
Absent: 49.8 %

H5: side of the body affected by pneumothorax:

Present: 87.2 %
Absent: 48.7 %

Figure 24: LADDER discovers slices for biased attributes in RN Sup IN1k-based classifier for *pneumothorax* classification in **NIH-CXR** dataset. This figure details the slice discovery process for biased attributes involving sentence analysis, hypothesis generation by an LLM, and the model's performance on slices where attributes are present or absent, demonstrating how biases affect classifier accuracy. We highlight the hypothesis generated by LADDER that corresponds to the ground truth biased attribute (*e.g.,* `chest-tube` for landbirds) in **yellow**.



```
Hypotheses:
The classifier is making mistake as it is biased toward:
H1: scattered calcifications
H2: bilateral occurrences
H3: the presence of multiple densities
H4: the description of benign appearances
H5: the mention of vascular calcifications

Prompt to test each hypotheses:
1. H1_scattered calcifications:
 ['Images showing scattered calcifications indicate positive cancer cases.',
'Scattered calcifications are a sign of breast cancer in mammograms.', 'Positive
cancer cases often present scattered calcifications in imaging.', 'Scattered
calcifications in a mammogram are typically associated with cancer.', 'Detecting
scattered calcifications is crucial for diagnosing cancer in patients.’]

2. H2_bilateral occurrences:
['Bilateral occurrences in mammogram images indicate a positive cancer
diagnosis.', 'Cancer cases often show bilateral changes in breast tissue.',
'Bilateral symptoms in mammograms are significant for identifying cancer.', 'The
presence of bilateral abnormalities suggests cancer in mammogram analysis.',
'Images with bilateral occurrences are frequently linked to positive cancer
findings.’]

3. H3_multiple densities:
 ['Multiple densities in mammograms are indicative of positive cancer cases.',
'The presence of multiple densities in breast imaging suggests cancer.', 'Cancer
in mammograms is often associated with multiple densities.', 'Identifying multiple
densities in mammograms is key to cancer detection.', 'Multiple densities in
imaging are a critical marker of breast cancer.’]

4. H4_benign appearances:
['Descriptions of benign appearances in mammograms can mislead cancer detection.',
'Benign-looking features in mammograms may still represent cancer cases.', 'Cancer
detection is challenged by benign appearances in imaging.', 'Mammograms with
benign appearances might still be positive for cancer.', 'Identifying cancer in
the presence of benign appearances in mammograms is crucial.’]

5. H5_vascular calcifications:
['Vascular calcifications in mammograms are often overlooked in cancer
diagnosis.', 'The presence of vascular calcifications can indicate underlying
cancer.', 'Cancer cases may present with vascular calcifications in mammograms.',
'Detecting cancer in the presence of vascular calcifications is essential.',
'Vascular calcifications in mammograms should be carefully evaluated for cancer.']
```

Figure 25: LADDER discovers slices for biased attributes for *cancer* classification in **RSNA-Mammo** dataset. This figure details the slice discovery process for biased attributes involving sentence analysis, hypothesis generation by an LLM, and the model's performance on slices where attributes are present or absent, demonstrating how biases affect classifier accuracy. We highlight the hypothesis generated by LADDER that corresponds to the ground truth biased attribute (*e.g.,* `calcification` for cancer) in **yellow**.