SPELUNKA: A FRAMEWORK FOR FINDING UNANTICIPATED ASSOCIATIONS IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language Models (VLMs) are the backbone of many state-of-the-art multimodal systems, but often encode spurious and harmful associations from their training data. While prior work focuses on removing known biases from VLMs, methods for discovering unanticipated associations for retrieval tasks remain underexplored. We introduce SPELUNKA, an automated, inference-time framework for finding unanticipated distributional biases for retrieval tasks in VLMs without pre-specified labeled subgroups, inference-time human supervision, or model fine-tuning. Given a concept of interest, SPELUNKA generates captions for related images, extracts relevant entities from captions, and selects semantically meaningful associations. A frequency-based detection rule then identifies spurious associations by measuring over- or under-representation of these attributes in a VLM's top-ranked outputs relative to their full dataset distribution. We evaluate SPELUNKA across four datasets, and find that it 1) successfully recovers both synthetic and naturally occurring spurious correlations, 2) achieves an average AUROC of 0.93 across four datasets in identifying images strongly associated with a query, and 3) outperforms baselines in finding features that are strongly associated with a given query.

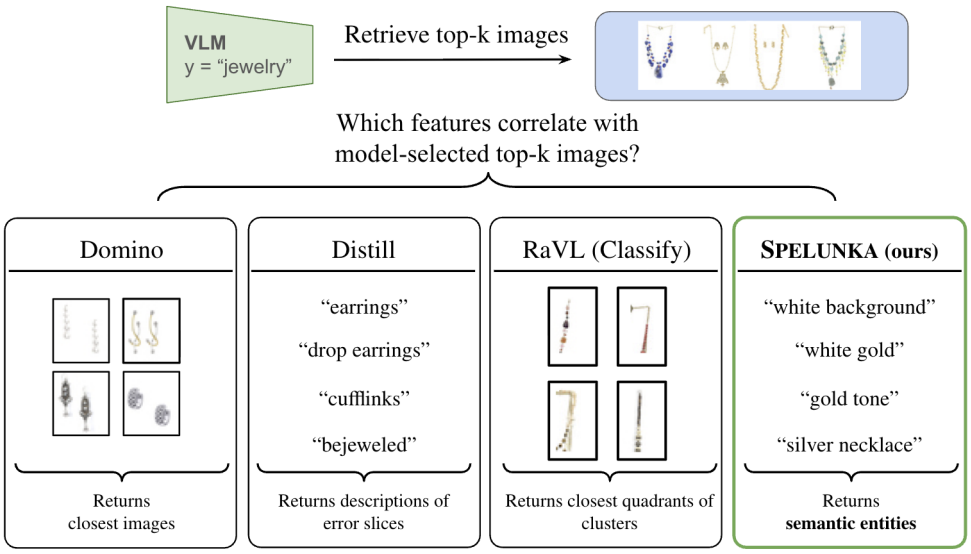
1 INTRODUCTION

Vision-Language Models (VLMs) have quickly become the state-of-the-art for multimodal tasks (Radford et al., 2021; Li et al., 2022; Liu et al., 2024b), such as image retrieval (Lahajal et al., 2024), zero and few-shot classification (Radford et al., 2021; An et al., 2023), and open-vocabulary object detection (Du et al., 2022). However, VLMs can encode both spurious and harmful associations present in training data, leading to undesired behavior such as co-occurring object bias (Varma et al., 2024), failure to recognize objects in atypical settings (Arias-Duart et al., 2023), and perpetuating negative social stereotypes (Alabdulmohsin et al., 2024; Kim et al., 2024; Friedrich et al., 2023; Hamidieh et al., 2024). Given that VLMs are the backbones for such a wide range of multi-modal applications, it is crucial to remove these undesired associations, otherwise known as spurious correlations (Ye et al., 2024).

Prior work has largely focused on *eliminating* known biases (Alabdulmohsin et al., 2024; Kim et al., 2024; Berg et al., 2022; Chuang et al., 2023; Kong et al., 2024; Wang et al., 2021; Luo et al., 2024; Zhu et al., 2023; Shen et al., 2023; Jung et al., 2024; Gerych et al., 2024; Dehdashtian et al., 2023), with relatively less targeting *finding* unanticipated correlations in VLMs (Varma et al., 2024; Kruspe, 2024). Many techniques assume prior knowledge of which attributes or concepts are spurious, in which case bias can be removed via finetuning (Zhu et al., 2023; Alabdulmohsin et al., 2024; Shen et al., 2023) or post-hoc intervention (Jung et al., 2024; Gerych et al., 2026; Chuang et al., 2023; Dehdashtian et al., 2023). However, prior work is limited to clearly defined tasks, e.g., classification, where differences in accuracy can be measured, but fails to explore tasks with less-defined metrics, such as retrieval (Eyuboglu et al., 2022; Jain et al., 2022; Kim et al., 2024; Varma et al., 2024).

Our work focuses on finding *unanticipated associations* in learned models, with three criteria. First, we do not assume access to existing labels or a classification task to generate spurious features or subgroups. Second, we avoid finetuning or auxiliary model training as this is computationally infeasible in many settings. Third, we seek a solution that provides semantically-meaningful identified features that can be easily assessed and removed by downstream debiasing techniques.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Identified biases on "jewelry" concept with `clip-vit-base-patch32`. Compared to other methods, SPELUNKA is able to pull out meaningful entities that map to the top retrieved images.

Figure 1: **SPELUNKA discovers and enables mitigation of unanticipated biases in VLM retrieval.** SPELUNKA differentiates itself by being able to work on tasks that (a) do not have labels and (b) are semantically meaningful for downstream users compared to other methods.

To address these criteria, we propose SPELUNKA (**S**emantic **P**attern **E**xtraction for **L**ocating **Unk**nown **A**ssociations). To identify spurious correlations for a *concept*, such as "person", SPELUNKA compares the top-k distribution of images related to the concept (e.g. top-k images of people) to an overall distribution for that concept (e.g. a large dataset of people). Specifically, SPELUNKA compares the relative frequency of different *elements* of the images, such as co-occurring objects, expressions, and backgrounds. If an element is over- or under-represented in the top-k retrieved images compared to its frequency within the full set of images for the target concept, then we can identify that the VLM is biased towards or against images where this element is present. To avoid the prohibitively expensive task of hand-labeling the elements of all images, SPELUNKA uses a captioning model, and SPELUNKA then extracts element labels and determines their significance. Importantly, we prove that SPELUNKA does *not* require its caption-generating component to be unbiased; the difference in concept frequency between the distributions can still identify spurious associations even with a biased image-captioning model.

Contributions. The primary contributions of this work are:

- We propose SPELUNKA, an automated spurious-correlation detection pipeline that can provide semantically-meaningful biases in retrieval tasks with no dataset labels.
- We demonstrate SPELUNKA's ability to accurately identify biases in both synthetic spurious correlations (e.g., waterbirds and injected grayscale bias) (Rojas et al., 2022; Sagawa et al., 2020), and naturally occurring dataset biases (Yang et al., 2023; Rojas et al., 2022; Aggarwal, 2019; Kärkkäinen & Joo, 2019).
- We validate that entities discovered by SPELUNKA can identify whether an image is strongly associated with a query on a range of embedding models (Radford et al., 2021; Fini et al., 2024; Zhai et al., 2023) and datasets (Yang et al., 2023; Rojas et al., 2022; Aggarwal, 2019; Kärkkäinen & Joo, 2019). On `clip-base-patch-16`, We find that entities identified across four datasets and queries that describe the image type, i.e., "jewelry" in an image set of jewelry are predictive of top-k images that will be associated with the term (average AUROC of 0.93 on 16 entities). On this task, SPELUNKA outperforms four baselines: (`Cluster`, adapted from Domino (Eyuboglu et al., 2022), `Distill` (Jain et al., 2022)), and retrieval and classification versions of `RaVL` (Varma et al., 2024).

2 Related Work

Spurious Correlations in VLMs Spurious correlations, features that are coincidentally or confounded to be correlated with data labels, have been known to cause misleading or incorrect outcomes in machine learning models (Ye et al., 2024). VLMs are a powerful model type that has grown in popularity due to state-of-the-art performance in both retrieval (Lahajal et al., 2024) and zero- and few-shot classification tasks (Radford et al., 2021; An et al., 2023). However, VLMs are still prone to spurious correlations which can lead to unsafe downstream performance, such as generating harmful content when faced with unsafe prompts (Chen et al., 2025). Other undesired behavior from associations present in training data include co-occuring object bias (e.g. butterflies being associated with flowers (Varma et al., 2024)), failure to recognize objects in atypical settings (e.g. removing backgrounds) (Arias-Duart et al., 2023), and perpetuating negative social stereotypes, (e.g. occupational gender bias) (Alabdulmohsin et al., 2024; Kim et al., 2024; Friedrich et al., 2023). While advances have been made in debiasing VLM embeddings for target concepts (Gerych et al., 2026; Chuang et al., 2023), determining the spurious correlations that VLM encodes is an ongoing challenge, as we do not have *a priori* knowledge of all of the biases that the model encodes.

Finding Spurious Correlations Prior work in spurious correlation detection focuses on studying subsets of data where a model performs poorly on a defined classification task (Eyuboglu et al., 2022; Jain et al., 2022; Kim et al., 2024; Varma et al., 2024). For instance, Domino (Eyuboglu et al., 2022) finds the images that perform poorly and then uses that class' proximity to cross-modal embeddings to describe that poor performing slice of data. While these approaches improve performance in the defined classification tasks, there remains a gap in finding spurious correlations in datasets where a clear performance assessment does not exist. Dammu & Shah (2023) developed a method that generates saliency maps to identify spurious correlations in an unlabeled set of images, but does not map these features to meaningful labels that can be understood and used for downstream tasks. This lack of semantically-meaningful features remains a challenge across all spurious correlation detection approaches where post-hoc explanations are unable to account for discovered artifacts that are unknown at test-time or are non-visible (Adebayo et al., 2022). SPELUNKA aims to both identify spurious correlations with meaningful labels without specifying a classification task.

3 Methodology

3.1 Preliminaries

Let M_θ denote a pretrained Vision-Language Model (VLM), which takes an image $x \in \mathcal{X}$ and a text query $y \in \mathcal{Y}$ and returns a relevance score $M_\theta(x, y) \in \mathbb{R}$. Let $\mathcal{D} = \{x_i\}_{i=1}^n$ be a set of images associated with a query y (e.g., `TV`), from which we aim to identify unanticipated, spuriously correlated visual or textual attributes. We define some additional terms specific to our problem set-up:

Definition 1 (Caption Vocabulary)**.** The *Caption Vocabulary* C is a set of all words that are used by the captioner to describe a set of images. Within this vocabulary, there exist a set of entities $\{e_1, e_2, \dots e_m\}$.

Definition 2 (Entity Indicator)**.** We define $\ell_e(x) \in \{0, 1\}$ to be the (unobserved) indicator of whether image $x \in \mathcal{D}$ contains entity e or not, i.e. $\ell(x) := \mathbb{1}\{e \in x\}$.

We assume no knowledge of any labels or *C*. Specifically, we assume no access to $\ell_e(x)$. Our goal then is to detect entities $e \in C$ that are over- or under-represented in the top-k images ranked by M_θ for query y, relative to the full dataset. We determine over and under-representation using *frequency bias*.

Definition 3 (Frequency Bias)**.** *Frequency bias* is the shift of entity frequencies between $\mathcal{D}$ and $\mathcal{D}_{\text{top-}k}$, the top-$k$ images according to $M_\theta(\cdot, y)$.

3.2 SPELUNKA: Uncovering Unanticipated Associations in VLMs

SPELUNKA consists of two stages: first we extract and group entities for a set of images. Then we identify any spurious correlations the VLM has between a target query y and the identified entities.

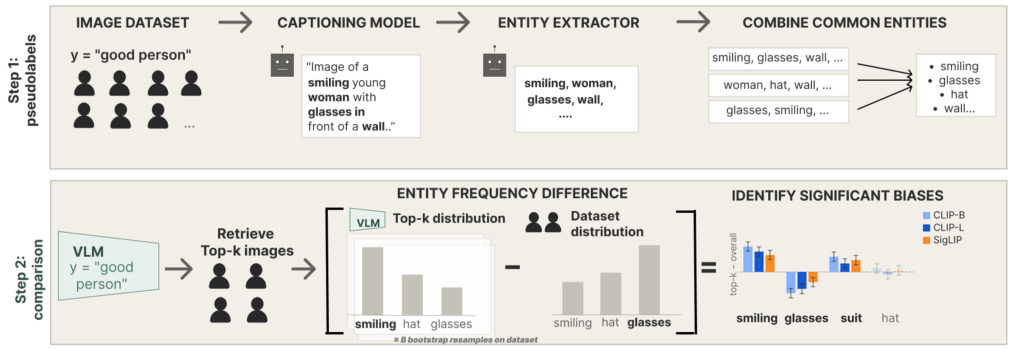


Figure 2: **Overview of the SPELUNKA pipeline.** Step 1 generates pseudolabels by captioning each image, extracting entities, and combines common entities into a final set. Step 2 retrieves the top-k images for a query using the VLM and identifies spurious correlations as the difference between the entity distribution in the top-k set and the full dataset distribution.

3.2.1 STAGE 1: GENERATING PSEUDOLABELS

Let f_{cap} be a multimodal LLM image captioning model. We first apply f_{cap} to each image $x \in \mathcal{D}$, generating a natural language caption $c_x = f_{\text{cap}}(x)$. The captions from f_{cap} describe objects, attributes, and contextual elements in an open-vocabulary fashion.

Next, we extract entity mentions from each caption c_x using an instruction-tuned language model f_{ent}. We prompt the model to identify discrete entities such as co-occurring objects (e.g., "smartphone"), descriptors (e.g., "flat-screen"), or settings (e.g., "brick wall"). Our entity-extraction prompt is given in the Supplement. This produces a candidate set of entity mentions $E_x = \{e_{1_x}, \ldots, e_{m_x}\}$ per image.

To organize these entities, we use agglomerative clustering to collapse entities of identical meanings, i.e. "television", "tv", and "TVs" are equivalent. For each image x and entity $e \in C$, we define a binary pseudolabel $\hat{\ell}_e(x) \in \{0, 1\}$ indicating the presence of e in the caption c_x. The pseudolabel function is defined as $\hat{\ell}_e(x) := \mathbb{1}\{e \in E_x\}$.

This step results in a weakly supervised binary matrix $L \in \{0, 1\}^{|\mathcal{D}| \times |E|}$, where $L_{ij} = \hat{\ell}_{e_j}(x_i)$ captures the presence of entity e_j in image x_i. This scalable and automatic extraction of labels from VLM-generated captions does not rely on predefined subgroups or human annotation (Figure 2).

3.2.2 STAGE 2: IDENTIFYING OVER-REPRESENTATION

Given a distribution of pseudolabels for the presence of entities across the image set $\mathcal{D}$ for query y, we now identify which entities are spuriously over-represented in the VLM's top-ranked outputs.

Let M_θ be the vision-language model under audit. Given the query y, we compute the VLM relevance score $M_\theta(x, y)$ for each image $x \in \mathcal{D}$. We then rank all images in descending order of their relevance score and select the top-k images: $\mathcal{D}_{\text{top-}k} := \text{Top-}k\left(\{M_\theta(x, y) : x \in \mathcal{D}\}\right)$.

For each entity $e \in E_C$, we compare its empirical prevalence in $\mathcal{D}_{\text{top-}k}$ against its prevalence in the full dataset $\mathcal{D}$. These quantities are estimated from the pseudolabel matrix $L \in \{0, 1\}^{n \times |E|}$, where $L_{ij} = \hat{\ell}_{e_j}(x_i)$.

Define:

$$p_{\text{obs}}^{\mathcal{D}}(e) := \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \hat{\ell}_e(x), \quad p_{\text{obs}}^{\mathcal{D}_{\text{top-}k}}(e) := \frac{1}{|\mathcal{D}_{\text{top-}k}|} \sum_{x \in \mathcal{D}_{\text{top-}k}} \hat{\ell}_e(x), \quad \Delta_{\text{obs}}(e) := p_{\text{obs}}^{\mathcal{D}_{\text{top-}k}}(e) - p_{\text{obs}}^{\mathcal{D}}(e)$$

The over-representation score $\Delta_{\text{obs}}(e)$ quantifies the degree to which e appears more frequently in the top-k ranked images than would be expected based on its global frequency, i.e., a positive $\Delta_{\text{obs}}(e)$ is positive spurious association between entity e and query y in the model M_θ.

This comparison is performed independently for each entity e, allowing us to identify unanticipated correlations across categories without relying on any prior knowledge of what the "spurious" attributes

might be. Further, to report entities that are relevant we perform N=1000 bootstraps with replacement on the dataset on the identified entities. In practice, we report the associations we identify as significant, i.e. 95% confidence intervals generated from the bootstrap do not cross zero. All graphs are sorted by the entities with the largest magnitude in one of the tested models with some minimum threshold, 5% shift unless stated otherwise.

3.3 Robustness to Captioning Model Bias

SPELUNKA's pseudolabels come from a captioner that may itself be biased. It may over-mention some entities and under-mention others. Does this captioner bias corrupt the spurious correlations SPELUNKA reports? We show that it does not. As long as the captioner detects each entity better than chance, the sign of $\Delta_{\text{obs}}(e)$ matches the sign of $\Delta_{\text{true}}(e)$, the *direction* of bias is preserved.

We demonstrate that an imperfect pseudolabling process can identify frequency bias by assuming a stochastic pseudolabeler with entity-dependent true positive rate (TPR) and false positive rate (FPR):

$$TPR_e := \mathbb{E}[\hat{\ell}_e(x) \mid \ell_e(x) = 1] := \alpha_e, \quad FPR_e := \mathbb{E}[\hat{\ell}_e(x) \mid \ell_e(x) = 0] := \beta_e$$

We express the expected value of the pseudolabel as:

$$\mathbb{E}[\hat{\ell}_e(x)] = \alpha_e \cdot \ell_e(x) + \beta_e \cdot (1 - \ell_e(x)) = \beta_e + (\alpha_e - \beta_e) \cdot \ell_e(x).$$

Note that this allows for the pseudolabeling process to be biased; the TPR and FPR may be different for different entities. Let the expected observed prevalence be:

$$\mathbb{E}[p^{\mathcal{D}}_{\text{obs}}(e)] := \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \mathbb{E}[\hat{\ell}_e(x)] = \beta_e + (\alpha_e - \beta_e) \cdot p^{\mathcal{D}}_{\text{true}}(e).$$

We compare the prevalence of e between the full dataset $\mathcal{D}$ and a subset $\mathcal{D}_{\text{top-}k}$ of the k most relevant images for some query y. Let:

$$\Delta_{\text{true}}(e) := p^{\mathcal{D}_{\text{top-}k}}_{\text{true}}(e) - p^{\mathcal{D}}_{\text{true}}(e), \quad \Delta^{\mathbb{E}}_{\text{obs}}(e) := \mathbb{E}[p^{\mathcal{D}_{\text{top-}k}}_{\text{obs}}(e) - p^{\mathcal{D}}_{\text{obs}}(e)]$$

By linearity of expectation and the above noise model:

$$\Delta^{\mathbb{E}}_{\text{obs}}(e) = \left[\beta_e + (\alpha_e - \beta_e) \cdot p^{\mathcal{D}_{\text{top-}k}}_{\text{true}}(e)\right] - \left[\beta_e + (\alpha_e - \beta_e) \cdot p^{\mathcal{D}}_{\text{true}}(e)\right] = (\alpha_e - \beta_e) \cdot \Delta_{\text{true}}(e).$$

Thus, as long as the pseudolabeler performs better than random — that is, $\alpha_e > \beta_e$ — the expected observed over-representation $\Delta^{\mathbb{E}}_{\text{obs}}(e)$ has the same sign as the true over-representation $\Delta_{\text{true}}(e)$. This implies that the *direction* of bias is preserved under label noise. Even with imperfect heuristics, consistent over-representation in the observed labels (e.g., $\Delta^{\mathbb{E}}_{\text{obs}}(e) > 0$) implies true over-representation ($\Delta_{\text{true}}(e) > 0$). Further analyses on the mathematical guarantees are in the Appendix.

4 Experimental Design

SPELUNKA Pipeline Setup. In our implementation of SPELUNKA, we selected BLIP (Li et al., 2022) as the captioning model f_{cap} and Llama 3.1 (Grattafiori et al., 2024) as the entity extraction and category-labeling model f_{ent} as small, local models. For entity embeddings, all-MiniLM-L6 (Reimers & Gurevych, 2019) and SpaCy (Honnibal et al., 2020) were used as inputs into Scikit-learn's (Pedregosa et al., 2011) Agglomerative Clustering was sued with a 20% distance threshold. Entities must have appeared at least 10 times in each dataset. We note that while Theorem 3.1 demonstrates that meaningful biases can be found even under noisy processes, we demonstrate that a variety of captioning models (`llava-1.5-7b-hf` (Liu et al., 2024a) and `Qwen3-VL-8B-Instruct` (Yang et al., 2025)) yield similar results Figure 5. Additional details are available in the Appendix.

Models. We test SPELUNKA on a range of CLIP (Radford et al., 2021) (`clip-vit-base-patch16`, `clip-vit-base-patch32`,

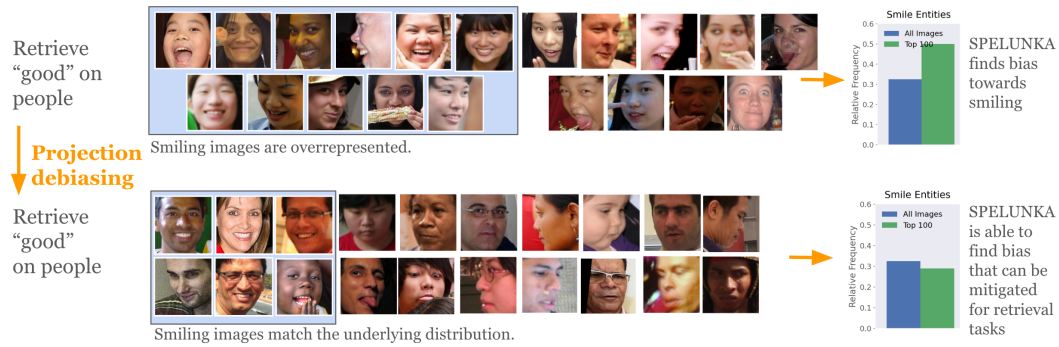


Figure 3: **SPELUNKA surfaces meaningful bias automatically by comparing entity frequencies between top-k retrieved images and the full dataset, with no labels, task definition, or finetuning required.** Querying AimV2 for "good" on `FairFace` overrepresents smiling faces relative to the dataset distribution (top). Projection debiasing, a common technique to remove a bias by orthogonalizing its embeddings to a concept (Bolukbasi et al., 2016), then aligns the retrieved distribution with the base rate (bottom).

`CLIP-ViT-L-14-laion2B-s32B-b82K`) an AIMV2 model (`aimv2-large-patch14-224-lit`) (Fini et al., 2024), and a SigLip model (`siglip-base-patch16-224`) (Zhai et al., 2023).

Data. We evaluate models on the `Fashion Product` (Aggarwal, 2019), `FairFace`, (Kärkkäinen & Joo, 2019), `Dollar Street` (Rojas et al., 2022), and `EmoSet` (Yang et al., 2023) each with a maximum of 128 tokens. We extract seven concepts across the four datasets to evaluate SPELUNKA. Concepts included were jewelry, topwear, and bottomwear from `Fashion Product`, TV, street view, and home from `Dollar Street`, and person from `FairFace`, and scene and person from `EmoSet`. Concepts were selected based on being an identifiable object with sufficient sample size with additional dataset descriptions in the Supplement. Note that `Fashion Product` and `EmoSet` differ from the other datasets by including ground truth subclass labels (e.g. rings, bracelets) for each concept (e.g. jewelry) in `Fashion Product` and different emotions for `EmoSet`.

5 Experimental Results

5.1 Finding Contextual Spurious Correlations

Unlike prior work, SPELUNKA is able to detect spurious correlations without a defined classification task or labels. This enables bias detection in real-world subjective cases. For instance, we find that SPELUNKA falsely attributes the entity "smiling" with "good person" in the `FairFace` dataset (Figure 3). Further, this bias, once identified, can be removed via projection debiasing (Bolukbasi et al., 2016). This demonstrates SPELUNKA's ability to find actionable, biases within models.

Beyond a meaningful example, we demonstrate that the extracted labels expose over/under representations of elements in VLMs. On the `FairFace` dataset, we set the query y to "person." There were 20 entities that had a greater than 5% over/under-representation and showed significance in at least once model Figure 4. Asian women, Asian men, and smiling were all entities that were overrepresented in the `AIM-V2` model, meaning that if the image was an Asian man or woman, it was more likely to be retrieved by the VLM. Results for all datasets can be found in the Appendix.

To show that the biases identified by SPELUNKA are biases found within the model, we take the significant entities and apply them to a different dataset of people images, `UTKFace` (Zhang & Qi, 2017). We kept the same query and only looked at entities that appeared in both datasets. We found that 14/15 overlapping entities that were significant in `FairFace` had sign agreement with `UTKFace` (Table 1). This demonstrates that the associations found by SPELUNKA are biases of the *model*, and that they can hold under varying applications.

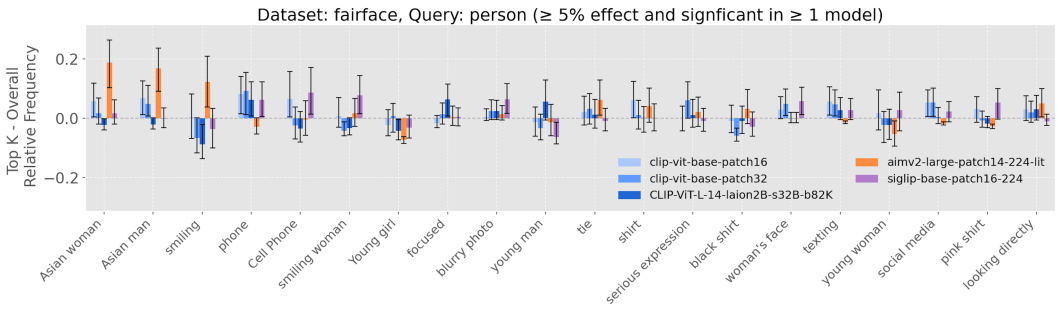


Figure 4: **Biases identified by SPELUNKA on Fairface when querying "person" across models.** Images with entities *Asian woman*, *Asian man*, and *smiling* appear more frequently than in the full dataset than for AIMV2.

Entity	Model	FairFace	UTKFace	Agreement
Asian woman	aimv2-large-patch14-224-lit	0.187	0.168	TRUE
Asian man	aimv2-large-patch14-224-lit	0.169	0.119	TRUE
smiling	aimv2-large-patch14-224-lit	0.123	0.060	TRUE
smiling	CLIP-ViT-L-14-laion2B-s32B-b82K	-0.087	0.020	FALSE
smiling woman	siglip-base-patch16-224	0.077	0.150	TRUE
Young girl	aimv2-large-patch14-224-lit	-0.073	-0.051	TRUE
Asian man	clip-vit-base-patch16	0.069	0.019	TRUE
young man	siglip-base-patch16-224	-0.064	-0.037	TRUE
tie	aimv2-large-patch14-224-lit	0.062	0.063	TRUE
shirt	clip-vit-base-patch16	0.061	0.004	TRUE
serious expression	clip-vit-base-patch32	0.061	0.063	TRUE
black shirt	clip-vit-base-patch32	-0.059	-0.019	TRUE
Asian woman	clip-vit-base-patch16	0.057	0.048	TRUE
young woman	aimv2-large-patch14-224-lit	-0.053	-0.024	TRUE
looking directly	aimv2-large-patch14-224-lit	0.050	0.051	TRUE

Table 1: **SPELUNKA finds biases that hold across models.** From the overlapping entities in `FairFace` and `UTKFace`, that met at least a 5% distribution shift and were significant in `FairFace`, 15/16 had sign agreement with `UTKFace`.

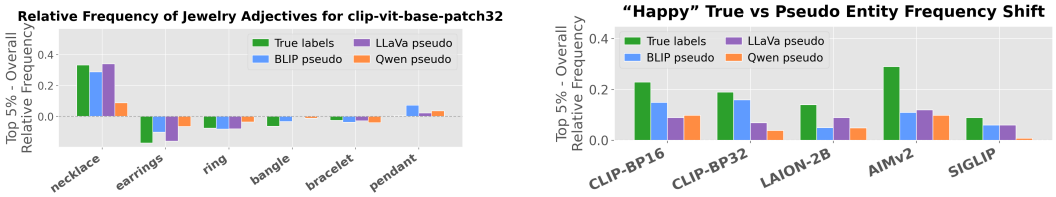
5.2 VALIDATION OF LABEL RECOVERY AND SPURIOUS CORRELATION DETECTION

Validation of Label Recovery. We validate that SPELUNKA correctly recovers ground-truth element labels. We use the jewelry images from the Fashion Product Images dataset, as this dataset contains labels for jewelry subtype (ring, necklace, etc). We also use the images from `EmoSet` labeled with a 'happy' facial expression for validation on large images. As shown in Figure 5, SPELUNKA's pseudolabels closely track ground-truth label frequencies, and this recovery is robust across captioning models (BLIP, Llava, Qwen) and VLM backbones (CLIP, AIMv2, SigLip) with models showing broad sign agreement. Results with additional datasets are in the Appendix.

Validation of Spurious Correlation Detection. We evaluated SPELUNKA's ability to identify a known bias, by examining `Waterbirds` (Sagawa et al., 2020). This dataset contains training data where where bird type matches background in the training data and testing data where bird and background type are mixed. We find that SPELUNKA extracts frequently occurring entities from the test dataset, such was 'water', 'ocean', 'trees' and 'Forest'. This shows SPELUNKA's ability to identify and known background spurious correlations with bird type (Figure 6) and demonstrates that SPELUNKA is able to uncover known spurious correlations from unlabeled datasets. We show an additional inserted bias experiment in the Appendix.

5.3 EVALUATING SPELUNKA FOR DOWNSTREAM CONCEPT FEATURE IDENTIFICATION

We compare SPELUNKA against other spurious correlation identification methods. For each method, we train a classifier on the identified spurious features to predict whether a given input will appear in the top 5% set of images. Higher classification accuracy indicates that the features identified by the method are more strongly correlated with the query.



(a) Pseudolabels and true label frequency for sub categories in the Jewelry concept. SPELUNKA accurately identifies true labels.

(b) Pseudolabels across models for `EmoSet` on 'happy' emotion. SPELUNKA identifies this subtle feature across models.

Figure 5: **SPELUNKA recovers ground-truth labels across datasets and captioning models.** Pseudolabel frequencies track true label frequencies on (a) `Fashion Product` jewelry subtypes and on (b) `EmoSet` 'happy' images across CLIP and AIMv2 backbones.

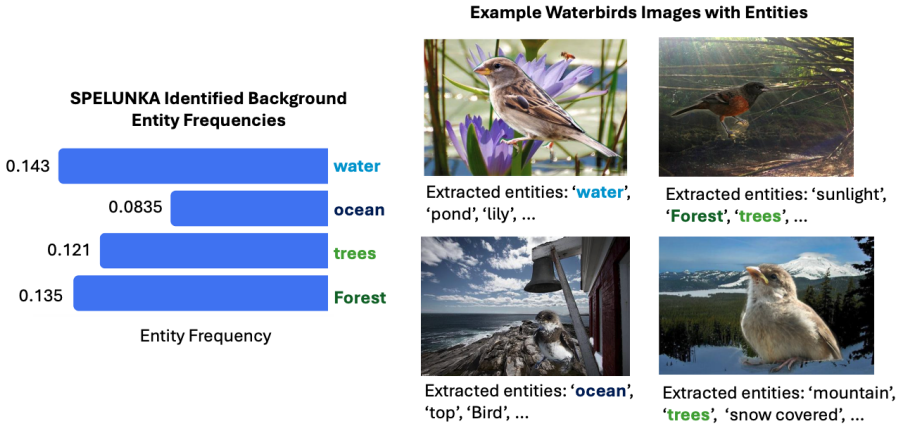


Figure 6: **SPELUNKA recovers the known background spurious correlation in `Waterbirds`.** Without access to bird-type or background labels, SPELUNKA extracts background entities (*water*, *ocean*, *trees*, *Forest*) at high frequency from the test set.

As no other automated methods exist for unlabeled bias detection in VLMs for retrieval tasks, we construct four baselines from prior works. `Domino` is inspired by (Eyuboglu et al., 2022). Here, we cluster image embeddings and then use the similarity to the *mean* embedding of each cluster as a feature. We vary the number of clusters from 2 to 16. `Distill` (Jain et al., 2022) is adapted by identifying spurious features through first identifying instances that are misclassified by a logistic regression classifier (e.g. misclassifications in jewelry types of rings, bracelets, necklaces, etc.). We then find natural language phrases from an LLM-generated corpus whose embeddings have the highest similarity to the normal vector. For each image embedding, we use similarity to these identified phrases as features, varying the number of phrases from 2 to 16. We use two versions of `RaVL` (Varma et al., 2024) where images are segmented into quadrants prior to clustering. For `RaVL (retrieve)`, we select the number of clusters based on the number of tested features. `RaVL (classify)` selects 256 features during clustering and then greedily selects the top features based on classifications of sublabels. Note, `Distill` and `RaVL (classify)` only work on the `Fashion` and `EmoSet` dataset which has subtype labels to construct error slices from.

For SPELUNKA, we define features as the similarity to the embeddings of the entities with the highest magnitude of difference between top-k and overall prevalence, according to SPELUNKA's pseudo-labeling process. We vary the number of entities from 2 to 16. We report AUROC (Melo, 2013) on identifying top-k images using a random forest (Breiman, 2001) classifier and report confidence intervals using 500 bootstrapped samples over the test set (50/50 split due to sample size). Table 2 quantitatively demonstrates SPELUNKA's ability to identify spurious features that are strongly correlated with the query, with an average AUROC of 0.93 across the full set of queries using 16 entities. We also find that the features are on average, more strongly correlated with the query

than those identified by other methods. Further, SPELUNKA's features are semantically meaningful. For instance, on `EmoSet`'s person concept, SPELUNKA tells us top potentially spurious features are 'smiling', 'young girl', and 'white shirt'. notably, these are spurious semantic concepts. In contrast, other methods return numeric vectors that are traced back to regions of the original image for interpretation or error descriptions from a pre-defined classification task (Figure 1a). These results indicate that SPELUNKA finds features that are both *quantitatively and qualitatively* more *informative* than the features identified by other methods. Additional results on the full set of queries are in the Appendix.

Table 2: **SPELUNKA outperforms baselines with an average AUROC of 0.93 in identifying images strongly associated with a query using a random forest classifier across 4 datasets and 6 queries with 16 entities** (full results in Appendix). Intervals denote 95% confidence intervals with 500 bootstrapped samples. Comparisons done with `clip-base-patch-16`. `Distill` and `RaVL (Classify)` require subtype labels and thus can not be applied to all datasets.

	# Entities	`Domino`	`RaVL` (Retrieve)	`Distill`	`RaVL` (Classify)	SPELUNKA
Fashion, Jewelry	2	0.61 [0.54, 0.68]	0.55 [0.50, 0.58]	**0.63 [0.56, 0.71]**	0.59 [0.54, 0.63]	0.56 [0.48, 0.65]
	4	0.68 [0.61, 0.75]	0.67 [0.63, 0.71]	0.67 [0.58, 0.75]	0.66 [0.57, 0.66]	**0.70 [0.63, 0.77]**
	8	0.73 [0.65, 0.79]	0.67 [0.63, 0.70]	**0.78 [0.72, 0.84]**	0.72 [0.68, 0.75]	0.75 [0.68, 0.81]
	16	0.73 [0.66, 0.80]	0.69 [0.65, 0.72]	0.80 [0.74, 0.87]	0.74 [0.70, 0.77]	**0.93 [0.90, 0.97]**
EmoSet, Person	2	0.77 [0.72, 0.81]	0.50 [0.47, 0.53]	0.76 [0.71, 0.81]	0.60 [0.57, 0.63]	**0.86 [0.81, 0.90]**
	4	0.85 [0.82, 0.89]	0.73 [0.71, 0.75]	0.83 [0.79, 0.87]	0.73 [0.71, 0.76]	**0.94 [0.91, 0.96]**
	8	0.864 [0.82, 0.90]	0.73 [0.71, 0.76]	0.87 [0.83, 0.91]	0.74 [0.71, 0.76]	**0.95 [0.93, 0.91]**
	16	0.86 [0.82, 0.91]	0.75 [0.72, 0.77]	0.88 [0.84, 0.92]	0.79 [0.77, 0.81]	**0.97 [0.95, 0.98]**
Dollar Street, TV	2	0.67 [0.55, 0.78]	0.54 [0.48, 0.61]	-	-	**0.89 [0.80, 0.96]**
	4	0.53 [0.40, 0.64]	0.50 [0.45, 0.57]	-	-	**0.87 [0.76, 0.95]**
	8	0.74 [0.62, 0.84]	0.59 [0.53, 0.65]	-	-	**0.92 [0.86, 0.96]**
	16	0.82 [0.73, 0.90]	0.61 [0.55, 0.67]	-	-	**0.92 [0.84, 0.96]**
FairFace, Face	2	0.63 [0.58, 0.69]	0.51 [0.48, 0.54]	-	-	**0.83 [0.78, 0.87]**
	4	0.71 [0.65, 0.76]	0.57 [0.55, 0.60]	-	-	**0.88 [0.84, 0.91]**
	8	0.73 [0.67, 0.78]	0.66 [0.63, 0.69]	-	-	**0.90 [0.87, 0.94]**
	16	0.72 [0.67, 0.77]	0.67 [0.64, 0.70]	-	-	**0.97 [0.96, 0.98]**

6 LIMITATIONS AND FUTURE WORK

Although SPELUNKA provides a practical way to surface unanticipated associations in VLMs, it has limitations. First, while we demonstrate that directional bias of captioning models is preserved under label noise and empirically demonstrate that varying captioning models yield similar pseudo-label results, the entities extracted will have variation dependent on caption quality. Future work could experiment with taking the super set of several captioning models to stabilize this. Second, SPELUNKA finds correlations based on frequency but does not assess their harmfulness or contextual relevance. It is a valuable auditing tool, but is not a substitute for human judgment in high-stakes applications. Despite these limitations, SPELUNKA has the potential for meaningful impact due to its automated ability to identify spurious correlations regardless of the task. As multimodal models see wider real-world deployment, accessible auditing tools become increasingly important. SPELUNKA lowers the barrier for auditing pretrained multimodal models, and we hope this work sparks further research on systematic, open set methods for identifying and addressing unwanted associations.

7 CONCLUSIONS

We introduced SPELUNKA, a fully automated framework for uncovering unanticipated associations in vision-language models (VLMs). SPELUNKA identifies spurious correlations without labeled subgroups, finetuning, or manual inspection. Our method leverages open-vocabulary captioning and large language models to construct semantically meaningful spurious correlations. Through theoretical analysis and empirical validation, we demonstrated that SPELUNKA can reliably identify meaningful biases across multiple domains, often outperforming baselines in identifying features predictive of being in the retrieval set. By enabling scalable and unsupervised auditing of VLMs, SPELUNKA represents a step towards ensuring the robustness, generalizability, and fairness of multimodal AI systems in real-world applications.

AI USE STATEMENT

In this work, we have not used generative AI tools for any of the tasks with required disclosure. Generative AI was used for software edits and LATEX formatting only. We have reviewed all AI-assisted work. We take responsibility for the final content of this work.

ETHICS STATEMENT

SPELUNKA can be used to detect biases that may have ethical implications for AI. However, the contents of this paper follow the Code of Ethics, and we believe there are no concerns raised.

REPRODUCIBILITY STATEMENT

The code for SPELUNKA is included in the submission and will be released publicly on GitHub upon acceptance. Results are produced on publicly-available datasets to further aid in reproducibility.

REFERENCES

Julius Adebayo, Michael Muelly, Hal Abelson, and Been Kim. Post hoc explanations may be ineffective for detecting unknown spurious correlation, 2022. URL https://arxiv.org/abs/2212.04629.

Param Aggarwal. Fashion product images dataset, 2019. URL https://www.kaggle.com/ds/139630.

Ibrahim Alabdulmohsin, Xiao Wang, Andreas Steiner, Priya Goyal, Alexander D'Amour, and Xiaohua Zhai. Clip the bias: How useful is balancing data in multimodal learning? *arXiv preprint arXiv:2403.04547*, 2024.

Bang An, Sicheng Zhu, Michael-Andrei Panaitescu-Liess, Chaithanya Kumar Mummadi, and Furong Huang. More context, less distraction: Visual classification by inferring and conditioning on contextual attributes. *arXiv preprint arXiv:2308.01313*, 2023.

Anna Arias-Duart, Victor Gimenez-Abalos, Ulises Cortés, and Dario Garcia-Gasulla. Assessing biases through visual contexts. *Electronics*, 12(14):3066, 2023.

Hugo Berg, Siobhan Hall, Yash Bhalgat, Hannah Kirk, Aleksandar Shtedritski, and Max Bain. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. In Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang (eds.), *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 806–822, Online only, November 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.aacl-main.61.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

Leo Breiman. Random forests, 2001.

Yiwei Chen, Yuguang Yao, Yihua Zhang, Bingquan Shen, Gaowen Liu, and Sijia Liu. Safety mirage: How spurious correlations undermine vlm safety fine-tuning, 2025. URL https://arxiv.org/abs/2503.11832.

Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023.

Preetam Prabhu Srikar Dammu and Chirag Shah. Detecting spurious correlations via robust visual concepts in real and ai-generated image classification, 2023. URL https://arxiv.org/abs/2311.01655.

Sepehr Dehdashtian, Lan Wang, and Vishnu Boddeti. Fairvlm: Mitigating bias in pre-trained vision-language models. In *The Twelfth International Conference on Learning Representations*, 2023.

Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14084–14093, 2022.

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

Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor Guilherme Turrisi da Costa, Louis Béthune, Zhe Gan, Alexander T Toshev, Marcin Eichner, Moin Nabi, Yinfei Yang, Joshua M. Susskind, and Alaaeldin El-Nouby. Multimodal autoregressive pre-training of large vision encoders, 2024. URL https://arxiv.org/abs/2411.14402.

Felix Friedrich, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Patrick Schramowski, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023.

Walter Gerych, Haoran Zhang, Kimia Hamidieh, Eileen Pan, Maanas Sharma, Thomas Hartvigsen, and Marzyeh Ghassemi. Bendvlm: Test-time debiasing of vision-language embeddings. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Walter Gerych, Cassandra Parent, Rafiya Javed Quinn Perian, Justin Solomon, and Marzyeh Ghassemi. Wring out the bias: Arotation-based alter-native to projection debiasing. In *The Fourteenth International Conference on Learning Representations*, 2026.

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

Kimia Hamidieh, Haoran Zhang, Walter Gerych, Thomas Hartvigsen, and Marzyeh Ghassemi. Identifying implicit social biases in vision-language models. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):547–561, Oct. 2024. doi: 10.1609/aies.v7i1.31657. URL https://ojs.aaai.org/index.php/AIES/article/view/31657.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spacy: Industrial-strength natural language processing in python. 2020. doi: 10.5281/zenodo.1212303.

Saachi Jain, Hannah Lawrence, Ankur Moitra, and Aleksander Madry. Distilling model failures as directions in latent space. *arXiv preprint arXiv:2206.14754*, 2022.

Hoin Jung, Taeuk Jang, and Xiaoqian Wang. A unified debiasing approach for vision-language models across modalities and tasks. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*, 2019.

Younghyun Kim, Sangwoo Mo, Minkyu Kim, Kyungmin Lee, Jaeho Lee, and Jinwoo Shin. Discovering and mitigating visual biases through keyword explanation, 2024.

Fanjie Kong, Shuai Yuan, Weituo Hao, and Ricardo Henao. Mitigating test-time bias for fair image retrieval. *Advances in Neural Information Processing Systems*, 36, 2024.

Anna Kruspe. Towards detecting unanticipated bias in large language models. *arXiv preprint arXiv:2404.02650*, 2024.

Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age, 2019. URL https://arxiv.org/abs/1908.04913.

Naresh Kumar Lahajal et al. Enhancing image retrieval: A comprehensive study on photo search using the clip mode. *arXiv preprint arXiv:2401.13613*, 2024.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024a. URL https://arxiv.org/abs/2310.03744.

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

Yan Luo, Min Shi, Muhammad Osama Khan, Muhammad Muneeb Afzal, Hao Huang, Shuaihang Yuan, Yu Tian, Luo Song, Ava Kouhana, Tobias Elze, et al. Fairclip: Harnessing fairness in vision-language learning. *arXiv preprint arXiv:2403.19949*, 2024.

Francisco Melo. *Area under the ROC Curve*, pp. 38–39. Springer New York, New York, NY, 2013. ISBN 978-1-4419-9863-7. doi: 10.1007/978-1-4419-9863-7_209. URL https://doi.org/10.1007/978-1-4419-9863-7_209.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

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

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

William A Gaviria Rojas, Sudnya Diamos, Keertan Ranjan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=qnfYsave0U4.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Xudong Shen, Chao Du, Tianyu Pang, Min Lin, Yongkang Wong, and Mohan Kankanhalli. Finetuning text-to-image diffusion models for fairness. *arXiv preprint arXiv:2311.07604*, 2023.

Maya Varma, Jean-Benoit Delbrouck, Zhihong Chen, Akshay Chaudhari, and Curtis Langlotz. Ravl: Discovering and mitigating spurious correlations in fine-tuned vision-language models. *Advances in Neural Information Processing Systems*, 37:82235–82264, 2024.

Jialu Wang, Yang Liu, and Xin Eric Wang. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. *arXiv preprint arXiv:2109.05433*, 2021.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Danny Cohen-Or, and Hui Huang. Emoset: A large-scale visual emotion dataset with rich attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20383–20394, 2023.

Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, and Aidong Zhang. Spurious correlations in machine learning: A survey, 2024. URL https://arxiv.org/abs/2402.12715.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. URL https://arxiv.org/abs/2303.15343.

Song Yang Zhang, Zhifei and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

Beier Zhu, Yulei Niu, Saeil Lee, Minhoe Hur, and Hanwang Zhang. Debiased fine-tuning for vision-language models by prompt regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

A Experiment Set-Up Additional Details

While we select specific models in our pipeline setup in Section 4, we note that SPELUNKA can work using a variety of captioning and extraction methods. We picked models that were relatively small and local for ease of reproducibility and low computational cost.

The following prompt was used for entity extraction 3.2.1: "I am going to give you a caption of a picture of {concept}. Please extract all of all objects and descriptors from the caption and return them as a comma separated list. Each list item should be a MAXIMUM of two words. Respond ONLY with this list. Here is the description for you to work with: "

B Dataset Additional Details

Three datasets were selected for this work to capture a diversity of image types. They were `Fashion Product`, (Aggarwal, 2019),`FairFace` (Kärkkäinen & Joo, 2019), and `Dollar Street` (Rojas et al., 2022). Concepts on `Fashion Product` and `Dollar Street` were selected based on nouns with large sample sizes. For `FairFace`, no sub selection of the dataset was used since they are all faces. Thus, entity extraction happened one time for all three concepts as the full dataset was covered. For all concepts, we took the first 2000 images if more than 2000 images were present. Example images and captions can be found in Figures 7 – 11.



Figure 7: Example image from 'TV' concept in `Dollar Street`.

Figure 7 Generated Caption: "The image depicts a cozy living room with a television and two chairs placed in front of it. The television is mounted on a wooden cabinet, which serves as a decorative piece in the room. The two chairs are positioned close to the television, creating a comfortable viewing area for the viewer. In addition to the television and chairs, there is a window in the room, allowing natural light to enter"

Extracted Entities: 'TV', 'window', 'cabinet', 'chair'

Figure 8 Generated Caption: "The image depicts a small hut with a thatched roof, situated in a dirt area. The hut is surrounded by various objects, such as pots, buckets, and other household items. There are several pots placed around the hut, some closer to the entrance and others further away. Additionally, there are two buckets positioned near the hut, one on the left side and another on the right side"

Extracted Entities: 'pot', 'household', 'hut', 'thatched', 'entrance', 'side', 'area', 'bucket', 'items', 'dirt'



Figure 8: Example image from ‘home’ concept in `Dollar Street`.



Figure 9: Example image from ‘topwear’ concept in `Fashion Product`.

Figure 9 Generated Caption: “In the image, a young man is standing against a white background, wearing a plaid shirt and jeans.”

Extracted Entities: ‘young man’, ‘white background’, ‘plaid shirt’, ‘jeans’

Figure 10 Generated Caption: “This is a gold and white diamond bracelet, featuring an intricate design that resembles a knot.”



Figure 10: Example image from ‘jewelry’ concept in `Fashion Product`.

Extracted Entities: ‘bracelet’, ‘gold’, ‘design’, ‘diamond’, ‘white’, ‘intricate’, ‘knot’



Figure 11: Example image from `FairFace`.

Figure 11 Generated Caption: “In the image, there is a man wearing glasses and a suit, standing in front of a blurry background.”

Extracted Entities: ‘man wearing’, ‘glasses’, ‘suit’, ‘standing’, ‘blurry background’

C Jewelry Pseudolabel Results for All Models

We include the full results from Figure 5a for all models used in this work. While individual models differ, Figure 12 demonstrates the similarity of true labels and pseudolabels across all models.

D Results Across Models and Concepts

The following Figures 13 – 19 show frequency shift results for all datasets and queries.

E Analysis

We formalize the analysis from the the methods section to present the mathematical guarantees of the biases identified by SPELUNKA when using a biased caption model. Theorem E.1 shows that with high probability, the direction of bias is preserved even with noisy pseudolabels. We then state Corollary E.2 that presents a bound for how much noisy pseudolabels may overestimate the identified bias direction. This motivates the usage of SPELUNKA as a robust spurious correlation detection framework even under imperfect setup conditions.

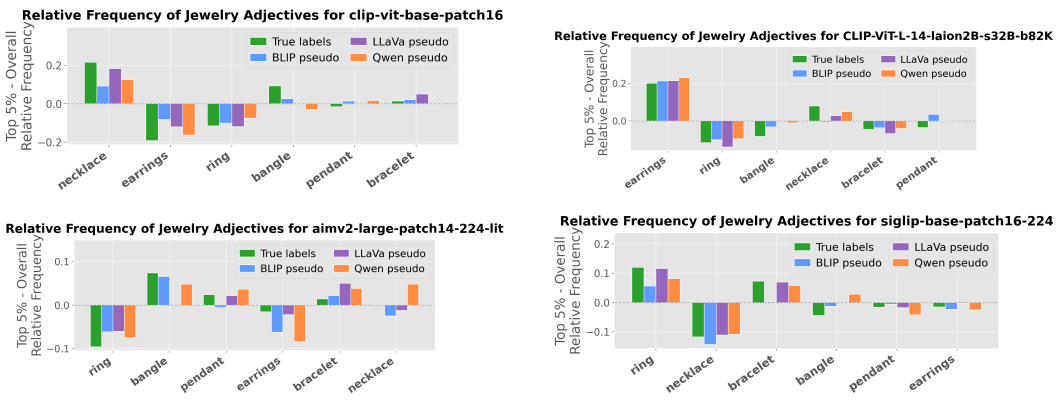


Figure 12: Pseudolabels and true label frequency for sub categories in the Jewelry concept. We find that SPELUNKA is able to accurately identify true labels.

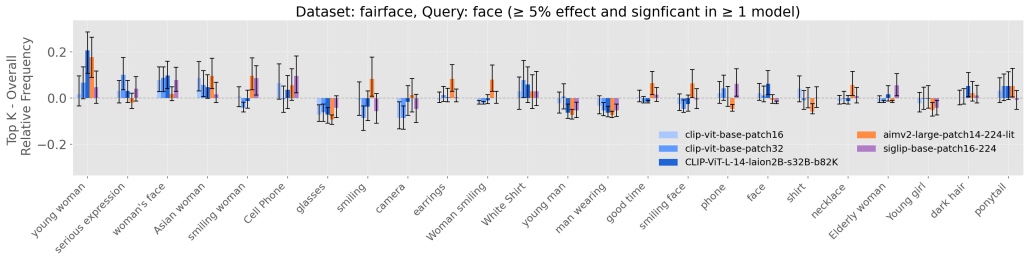


Figure 13: Frequency shift results for the 'face' concept in `FairFace`.

Theorem E.1 (High-Probability Sign Consistency under Noisy Pseudolabels)**.** *Let e be an entity with true overrepresentation*

$$\Delta_{true}(e) := p_{true}^{\mathcal{D}_{top\text{-}k}}(e) - p_{true}^{\mathcal{D}}(e),$$

and pseudolabel noise parameters $\alpha_e = \mathbb{E}[\hat{\ell}_e(x) \mid \ell_e(x) = 1]$, $\beta_e = \mathbb{E}[\hat{\ell}_e(x) \mid \ell_e(x) = 0]$.

Let $\Delta_{obs}(e)$ be the empirical observed overrepresentation computed from noisy pseudolabels over datasets $\mathcal{D}_{top\text{-}k}$ of size k, and $\mathcal{D}$ of size n.

Then with probability at least $1 - \delta$:

$$|\Delta_{obs}(e) - (\alpha_e - \beta_e) \cdot \Delta_{true}(e)| \leq \epsilon,$$

$$\textit{where } \epsilon := \sqrt{\frac{1}{2}\left(\frac{1}{k} + \frac{1}{n}\right)\log\left(\frac{4}{\delta}\right)}.$$

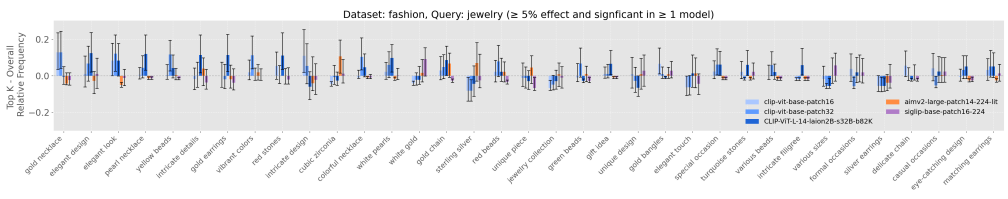


Figure 14: Frequency shift results for the 'jewelry' concept in `Fashion Data`.

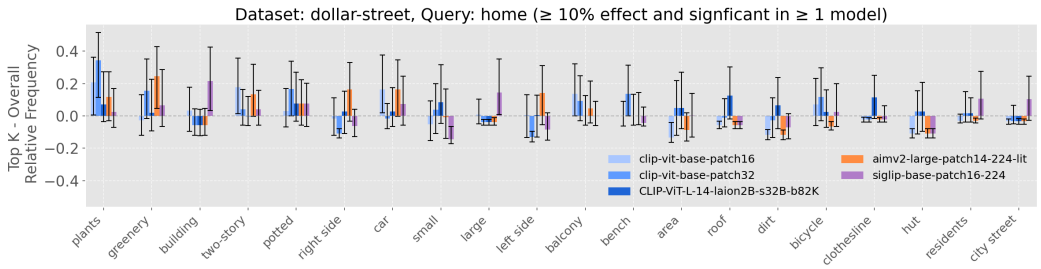


Figure 15: Frequency shift results for 'home' concept in `Dollar Street`.

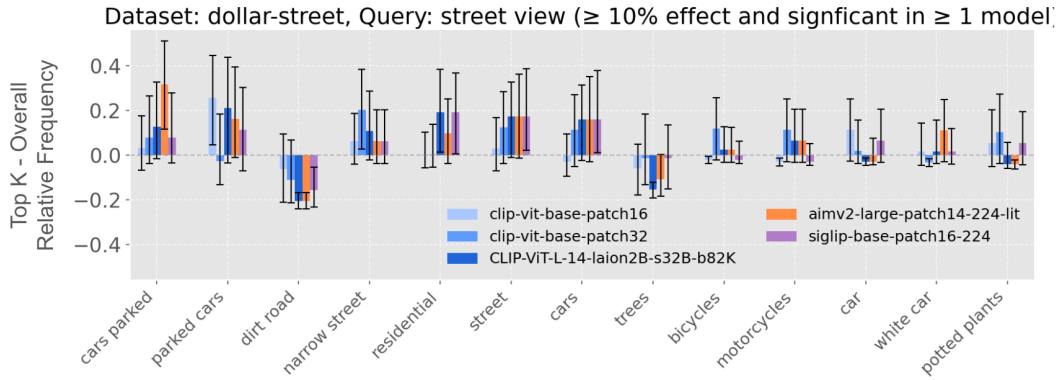


Figure 16: Frequency shift results for 'street view' concept in `Dollar Street`.

In particular, if

$$|(\alpha_e - \beta_e) \cdot \Delta_{true}(e)| > \epsilon,$$

then

$$\text{sign}(\Delta_{obs}(e)) = \text{sign}(\Delta_{true}(e)),$$

and the bias direction is preserved with probability $> 1 - \delta$.

Corollary E.2 (Bounding True Overrepresentation from Noisy Observations)**.** *Under the assumptions of Theorem E.1, define the estimator:*

$$\widehat{\Delta}_{true}(e) := \frac{\Delta_{obs}(e)}{\alpha_e - \beta_e}.$$

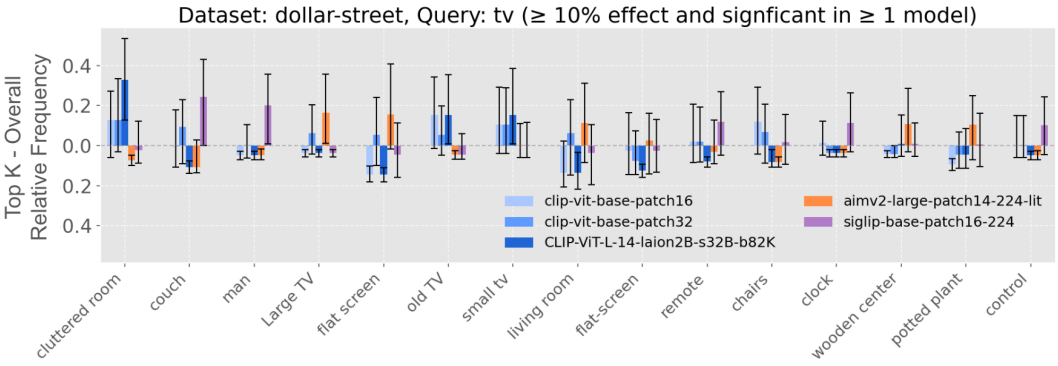


Figure 17: Frequency shift results for 'TV' concept in `Dollar Street`.

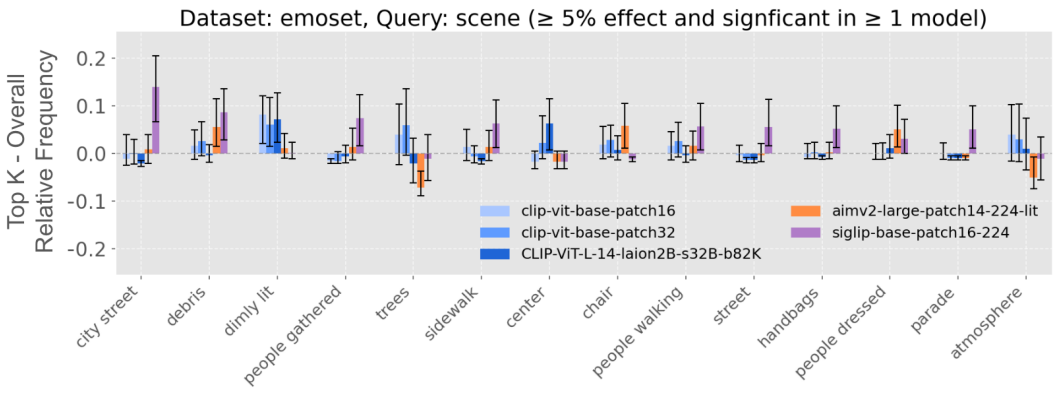


Figure 18: Frequency shift results for the 'scene' concept in `EmoSet`.

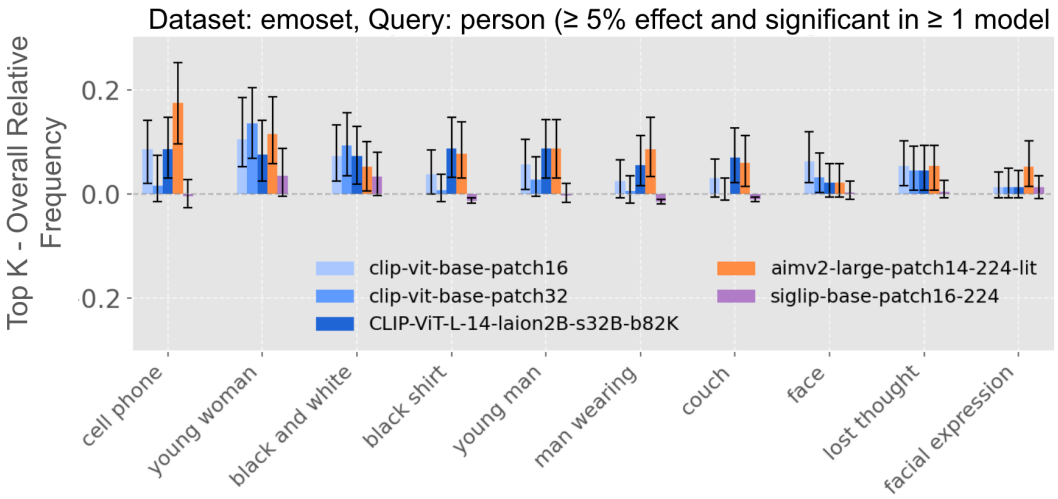


Figure 19: Frequency shift results for the 'person' concept in `EmoSet`.

Then with probability at least $1-\delta$*:*

$$\left|\widehat{\Delta}_{true}(e) - \Delta_{true}(e)\right| \leq \frac{1}{|\alpha_e - \beta_e|} \cdot \sqrt{\frac{1}{2}\left(\frac{1}{k} + \frac{1}{n}\right)\log\left(\frac{4}{\delta}\right)}.$$

In particular, if $\alpha_e > \beta_e$ *and* $\Delta_{true}(e) > 0$*, then* $\widehat{\Delta}_{true}(e)$ *provides an upwardly noisy but consistent estimate of the true frequency bias.*

F Full Table for Competitive Analysis

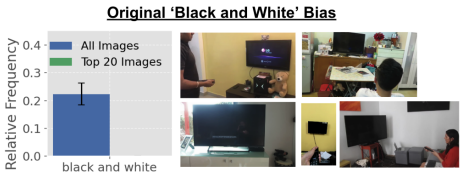
Full version of the competitive analysis Table 2 is shown here in Table 3.

G Inserted Bias Recovery

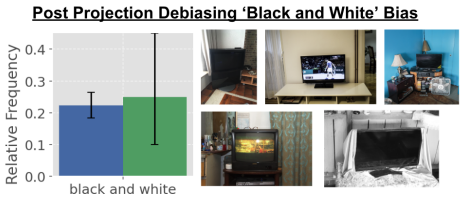
We inserted a bias on Dollar Street's TV images by randomly perturbing 50% of images to grayscale, ran the full SPELUNKA pipeline, and queried for "expensive" on `clip-vit-base-patch-16`. SPELUNKA correctly identified "black-and-white" as a relevant element, and determined that this was the most biased feature: in the top-20 retrieved images , only two were black-and-white despite half of the images in the dataset being black-and-white (see Figure 20).

Table 3: Comparison of methods with `clip-base-patch-16`. `Distill` and `RaVL (Classify)` require sub labels and thus can not be applied to all datasets.

	# Entities	Domino	RaVL (Retrieve)	Distill	RaVL (Classify)	SPELUNKA
Fashion, Jewelry	2	0.61 [0.54, 0.68]	0.55 [0.50, 0.58]	**0.63 [0.56, 0.71]**	0.59 [0.54, 0.63]	0.56 [0.48, 0.65]
	4	0.68 [0.61, 0.75]	0.67 [0.63, 0.71]	0.67 [0.58, 0.75]	0.66 [0.57, 0.66]	**0.70 [0.63, 0.77]**
	8	0.73 [0.65, 0.79]	0.67 [0.63, 0.70]	**0.78 [0.72, 0.84]**	0.72 [0.68, 0.75]	0.75 [0.68, 0.81]
	16	0.73 [0.66, 0.80]	0.69 [0.65, 0.72]	0.80 [0.74, 0.87]	0.74 [0.70, 0.77]	**0.93 [0.90, 0.97]**
EmoSet, Person	2	0.77 [0.72, 0.81]	0.50 [0.47, 0.53]	0.76 [0.71, 0.81]	0.60 [0.57, 0.63]	**0.86 [0.81, 0.90]**
	4	0.85 [0.82, 0.89]	0.73 [0.71, 0.75]	0.83 [0.79, 0.87]	0.73 [0.71, 0.76]	**0.94 [0.91, 0.96]**
	8	0.864 [0.82, 0.90]	0.73 [0.71, 0.76]	0.87 [0.83, 0.91]	0.74 [0.71, 0.76]	**0.95 [0.93, 0.91]**
	16	0.86 [0.82, 0.91]	0.75 [0.72, 0.77]	0.88 [0.84, 0.92]	0.79 [0.77, 0.81]	**0.97 [0.95, 0.98]**
Dollar Street, TV	2	0.67 [0.55, 0.78]	0.54 [0.48, 0.61]	-	-	**0.89 [0.80, 0.96]**
	4	0.53 [0.40, 0.64]	0.50 [0.45, 0.57]	-	-	**0.87 [0.76, 0.95]**
	8	0.74 [0.62, 0.84]	0.59 [0.53, 0.65]	-	-	**0.92 [0.86, 0.96]**
	16	0.82 [0.73, 0.90]	0.61 [0.55, 0.67]	-	-	**0.92 [0.84, 0.96]**
FairFace, Face	2	0.63 [0.58, 0.69]	0.51 [0.48, 0.54]	-	-	**0.83 [0.78, 0.87]**
	4	0.71 [0.65, 0.76]	0.57 [0.55, 0.60]	-	-	**0.88 [0.84, 0.91]**
	8	0.73 [0.67, 0.78]	0.66 [0.63, 0.69]	-	-	**0.90 [0.87, 0.94]**
	16	0.72 [0.67, 0.77]	0.67 [0.64, 0.70]	-	-	**0.97 [0.96, 0.98]**
Dollar Street, Home	2	0.59 [0.46, 0.72]	0.53 [0.47, 0.60]	-	-	**0.80 [0.72, 0.88]**
	4	0.72 [0.58, 0.84]	0.52 [0.46, 0.59]	-	-	**0.81 [0.70, 0.92]**
	8	0.75 [0.64, 0.86]	0.56 [0.51, 0.63]	-	-	**0.84 [0.73, 0.94]**
	16	0.79 [0.70, 0.87]	0.56 [0.50, 0.62]	-	-	**0.87 [0.79, 0.94]**
Dollar Street, Street View	2	0.67 [0.54, 0.79]	0.53 [0.47, 0.59]	-	-	**0.75 [0.65, 0.83]**
	4	0.78 [0.68, 0.85]	0.65 [0.59, 0.71]	-	-	**0.88 [0.81, 0.93]**
	8	0.81 [0.72, 0.88]	0.69 [0.62, 0.75]	-	-	**0.91 [0.83, 0.97]**
	16	0.84 [0.76, 0.91]	0.77 [0.71, 0.82]	-	-	**0.92 [0.85, 0.97]**



(a) Original bias distribution of black and white entity.



(b) Post-projection debiasing black and white entity distribution.

Figure 20: **SPELUNKA identifies grayscale bias and enables its correction.** On `Dollar Street` TVs with 50% of images randomly converted to grayscale, querying `clip-vit-base-patch-16` for "expensive" underrepresents grayscale images in the top-20 retrieved (a); projection debiasing on the black-and-white concept restores the base-rate distribution (b).

We demonstrate that the model was indeed relying on the presence or absence of grayscale when determining rankings, by using projection-based debiasing (Chuang et al., 2023; Bolukbasi et al., 2016) to make the query embedding orthogonal to the black-and-white concept. After projection debiasing, the model had approximately equal relative frequency for the target concept Figure 20. This analysis shows that **SPELUNKA captures visual attributes and map those attributes to corresponding biases in VLMs**.

H COMPUTE

All experiments were run on a GPU on an internal cluster. While multiple GPUs may have been used to speed up the slower pieces of experimentation, mainly captioning and entity extraction, multiple GPUs not necessary for the implementation of SPELUNKA.