S^{2}COPE: SELF-SUPERVISED CONCEPT DISCOVERY VIA PREFERENCE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised learning can learn useful representations from unlabeled images. Yet these representations remain largely opaque, making it difficult to understand what visual concepts they encode. Can self-supervision instead produce useful visual concepts that are directly understandable to people? We introduce S^{2}COPE, a framework that represents images as sets of concepts expressed in natural language. A vision-language model proposes candidate concepts, and a self-supervised objective evaluates how well each concept describes the source image while distinguishing it from other images. This feedback is used to construct preferences and train the model to propose better concepts. Across eight diverse visual datasets, our self-supervised concept representations transfer without adaptation and improve downstream recognition accuracy over prior interpretable approaches. In medical imaging, our discovered concepts reveal spurious visual cues and enable a more robust classifier. These results show that self-supervision can discover useful, human-readable visual representations from unlabeled images.

1 INTRODUCTION

Self-supervised learning can discover useful visual representations without annotations (Chen et al., 2020; He et al., 2020; Caron et al., 2021; He et al., 2022), but what it discovers remains difficult to interpret. In scientific domains, from cellular pathology to astrophysics, the goal is often not only to classify observations, but also to understand the visual patterns that distinguish them. In many such settings, the relevant concepts may not be known in advance; more fundamentally, the category taxonomy itself may not yet be defined when the data are first collected. Identifying which visual concepts matter, without first defining categories or annotating images, is therefore itself part of the discovery process (Wang et al., 2023; Iten et al., 2020). This motivates our question: *how can models discover useful, human-readable visual concepts directly from unlabeled images?*

Prior work on interpretable visual representations often relies on predefined, annotated concepts (Koh et al., 2020) or image category labels (Oikarinen et al., 2023; Chiquier et al., 2024). Yet such supervision is not always available: concept annotations can be costly to obtain, and in many settings even the relevant category labels may not be known in advance. More recent approaches leverage embedding models (Radford et al., 2021; Zhai et al., 2023; Tschannen et al., 2025) or vision large language models (VLLMs) (Yang et al., 2025; Bai et al., 2025) to construct interpretable representations. These methods, however, typically treat the concept proposer as fixed: they use its descriptions for downstream prediction, but do not learn from the target images which concepts are most useful to propose (Menon & Vondrick, 2023; Oikarinen et al., 2023; Yang et al., 2023b). What is missing is a source of feedback that can teach the concept generator, directly from unlabeled images, which concepts are useful—without relying on predefined classes, annotated concepts, or human judgments.

In this paper, we introduce S^{2}COPE, a framework that learns to generate interpretable concepts from unlabeled data (Figure 1). Our key insight is that the same self-supervised signal used to learn visual representations can also teach a model which language concepts are useful to generate. Starting from a pretrained vision-language model, we sample candidate concept sets and evaluate them with a self-supervised contrastive objective that measures how well a concept set describes its source image while distinguishing it from other images. Higher- and lower-scoring candidates are then paired as preferences and used to train the concept generator with Direct Preference Optimization

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.

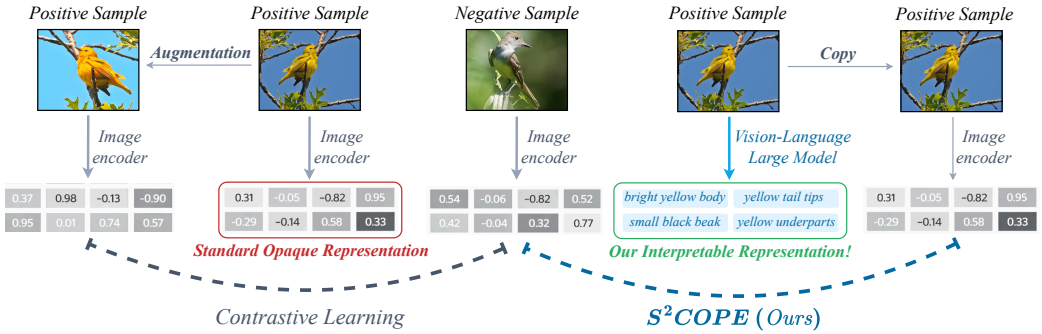


Figure 1: **Self-Supervised Visual Concept Discovery.** (Left) Standard self-supervised learning represents images as feature vectors, which are not directly understandable. (Right) S^{2}COPE learns to propose human-readable visual concepts from unlabeled images. A pretrained vision-language model samples candidate concepts, and a contrastive objective evaluates how well they describe the source image and distinguish it from other images. These scores define preferences for training the concept generator, without class labels or concept annotations. The learned concepts support downstream prediction and provide an explicit vocabulary for inspecting model behavior.

(DPO) (Rafailov et al., 2023). Rather than treating concept proposal as a fixed preprocessing step, S^{2}COPE adapts the model's proposals to the visual patterns present in the unlabeled training images. We call this process self-supervised because it requires neither target class names, image labels, concept annotations, nor human judgments.

We evaluate the learned concepts through two complementary applications: cross-domain recognition and shortcut mitigation. Across eight datasets spanning natural images (Van Horn et al., 2021; Wah et al., 2011), faces (Karkkainen & Joo, 2021), medical imaging (Tschandl et al., 2018; Yang et al., 2023a), and physics (Leung & Bovy, 2019; Zevin et al., 2017), our learned concepts outperform four concept-based baselines on every dataset, improving downstream top-1 accuracy by 7.37 percentage points on average over the best of these baselines per dataset. LLM-based and human evaluations further show that our method achieves the highest win rate in the quality of the discovered concepts. In ISIC dermoscopy (Nauta et al., 2021; Rieger et al., 2020), our self-supervised concepts expose colored calibration patches associated with benign lesions. Filtering training images using these concepts reduces the patch-induced error rate on malignant images by 15 points. Together, these results suggest that concept discovery need not begin with category labels: a language model can propose candidate hypotheses, while unlabeled visual data provides the evidence needed to determine which concepts are useful.

2 Related Work

Self-Supervised Learning. Self-supervised learning shows that useful visual representations can be learned without image annotations (Chen et al., 2020; He et al., 2020; Caron et al., 2021; He et al., 2022; Grill et al., 2020; Caron et al., 2020; Bardes et al., 2022; Zhou et al., 2022; Oquab et al., 2024; Caron et al., 2018). The training signal comes from the images themselves, but what the model learns is expressed as feature vectors rather than concepts with explicit meanings. Recent work brings related objectives into large vision-language models through reconstruction (Wang et al., 2025), jigsaw tasks (Wu et al., 2025), feature alignment (Yoon et al., 2025), and auxiliary visual tasks (Caffagni et al., 2025). These approaches improve visual learning without directly teaching the model which human-readable concepts to propose. Our work applies self-supervision to concept discovery itself. We evaluate candidate concept sets by how well they describe their source image and distinguish it from other images.

Concept-Based Models. Concept bottleneck models make predictions through named visual properties, allowing users to inspect the concepts behind a decision (Koh et al., 2020). The difficulty is choosing these concepts without first specifying what the model should look for. Early models rely on human-defined vocabularies and concept annotations. Later work reduces this requirement by using language models to propose descriptions (Yang et al., 2023b; Oikarinen et al., 2023), reusing pretrained concept representations (Yuksekgonul et al., 2023), or linking concepts to their visual

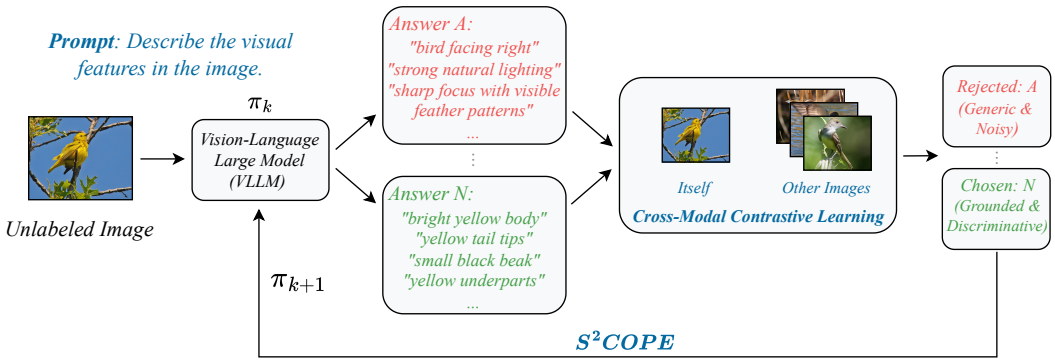


Figure 2: **Overview of the S^{2}COPE Discovery Loop.** At iteration k, the generator π_k samples several descriptions for each unlabeled image. A frozen CLIP model scores each description by how well it matches the source image relative to other images. We pair higher- and lower-scoring descriptions with the largest score gaps and train the generator on these preferences using DPO. The updated generator π_{k+1} then proposes candidates for the next iteration.

evidence (Srivastava et al., 2024). However, removing concept annotations does not necessarily remove the need for known categories. LaBo and LF-CBM use class names to guide concept generation (Yang et al., 2023b; Oikarinen et al., 2023), while PCBM can use class names to retrieve related concepts from a knowledge base (Yuksekgonul et al., 2023). These approaches therefore provide ways to build interpretable predictors for known categories, but do not directly address discovery when those categories have not been specified.

Another line of work extracts and describes concepts from the features of an existing model (Pach et al., 2025; Almudévar et al., 2025; Schrodi et al., 2025; Oikarinen & Weng, 2023). U-F^2-CBM (Sammani et al., 2025) discovers concepts from frozen features without class labels. M-CBM (De Santis et al., 2026) uses a sparse autoencoder to extract features from a pretrained classifier, then a multimodal language model to name and annotate the resulting concepts. Related to our shortcut analysis, B2T (Kim et al., 2024) extracts keywords from captions of misclassified images to identify spurious correlations; it uses class labels to identify these errors and keeps the captioner fixed. We instead train the concept generator itself using feedback from unlabeled images, rather than only extract, name, or select concepts from a fixed model. Neither concept generation nor selection in S^{2}COPE requires class names or image labels.

Preference Learning. Vision-language pretraining and instruction tuning connect visual content with language (Jia et al., 2021; Alayrac et al., 2022; Li et al., 2023a; Dai et al., 2023; Liu et al., 2023). However, a fluent description need not be accurate or capture the visual details that distinguish one image from another (Li et al., 2023b; Rohrbach et al., 2018; Guan et al., 2024; Snæbjarnarson et al., 2025). Preference learning offers a way to improve generated responses by training on comparisons between alternatives (Ding et al., 2025; Fu et al., 2025; Chen et al., 2025; Sun et al., 2025; Jia et al., 2025). A central question is how to obtain these comparisons without asking people or a stronger model to judge each output. SeVa (Zhu et al., 2024) and SMPRO (Swetha et al., 2026) construct visual preference data without external annotators to improve general-purpose multimodal responses. Our goal is instead to learn which visual concepts to propose. We score candidate concept sets by how well they describe their source image and distinguish it from other images. These scores define preference pairs for training the concept generator through direct preference optimization.

3 Methodology

Standard representation learning maps images to feature vectors that are hard to interpret. We instead seek a vocabulary of natural-language concepts as our representations. S^{2}COPE learns this vocabulary through three steps: sampling descriptions from the generator, scoring them against the unlabeled images, and training the generator to increase the resulting preference scores (Figure 2). We start from pretrained VLLMs, but use no class names, image labels, or concept annotations from the training images during our concept discovery.

3.1 Concept Proposal via Vision-Language Models

For each image x, we prompt the current vision-language generator π_k, a VLLM (Qwen3-VL-8B-Instruct in our experiments), to describe its visual attributes and sample N responses, $\mathcal{C}(x) = \{c_1, \ldots, c_N\}$. We use high-temperature sampling to obtain varied candidates. Each response may contain multiple concepts and is scored as a whole; we extract individual concepts later for downstream evaluation.

3.2 Cross-Modal Contrastive Reward

For each image x, the generator produces candidate descriptions $\mathcal{C}(x)$. We evaluate each candidate by how well it matches its source image relative to other images. A separate, frozen CLIP model (Radford et al., 2021; Cherti et al., 2023) provides this comparison: its image encoder f_{img} encodes x, and its text encoder f_{text} encodes each description $c \in \mathcal{C}(x)$. We normalize their outputs to obtain

$$v_x = \frac{f_{\text{img}}(x)}{\|f_{\text{img}}(x)\|_2}, \qquad t_c = \frac{f_{\text{text}}(c)}{\|f_{\text{text}}(c)\|_2}. \tag{1}$$

Thus, all candidates for an image share the same image embedding v_x, while each candidate description c has its own text embedding t_c.

For a batch $\mathcal{B}$ containing x, we encode every image in the same way and score each candidate as

$$R(c, x) = \log \frac{\exp(t_c^\top v_x / \tau)}{\sum_{x' \in \mathcal{B}} \exp(t_c^\top v_{x'} / \tau)}, \tag{2}$$

where τ is a temperature. The source image is the positive match, and the other batch images serve as negatives. A description c receives a high score when it matches its source image more strongly than the alternatives, rather than describing many images equally well. As described in the next subsection, this score is then used to train the generator to assign higher probability to such concepts. For the positive image corresponding to the query image being interpreted, we use the original image without augmentation. This differs from standard contrastive learning, where augmentations encode prior assumptions about which transformations the representation should ignore. In our setting, we do not assume in advance which visual variations the discovered concepts should be invariant to, and empirically find that using the original image works best. We study the effect of introducing additional augmentations in Table 4.

3.3 Self-Supervised Preference Optimization

Rather than only filtering proposals from a fixed generator (Oikarinen et al., 2023), we use the contrastive scores to improve what the VLLM generator proposes. For each image, we select candidate pairs with the largest score gaps. The higher-scoring description becomes c_w and the lower-scoring description becomes c_l, forming a preference dataset $\mathcal{D}_k = \{(x, c_w, c_l)\}$.

Starting from π_k, we train only the generator π_θ on these pairs using Direct Preference Optimization (DPO) (Rafailov et al., 2023):

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x, c_w, c_l) \sim \mathcal{D}_k} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(c_w \mid x)}{\pi_{\text{ref}}(c_w \mid x)} - \beta \log \frac{\pi_\theta(c_l \mid x)}{\pi_{\text{ref}}(c_l \mid x)} \right) \right]. \tag{3}$$

Here, π_{ref} is a frozen reference model which is the generator π_k from the previous iteration, σ is the sigmoid function, and β controls regularization toward the reference. The objective teaches the generator to favor higher-scoring descriptions over lower-scoring ones. We update the generator through the DPO loss, without differentiating through the sampled text or CLIP score. Figure 3 compares concepts before and after training.

Iterative Optimization. We repeat the above procedure for K rounds. After each DPO update, the optimized vision-language model generates a new set of candidate concepts. These concepts are then evaluated using the same self-supervised objective, converted into preference pairs, and used for the next round of DPO training π_{k+1}. Repeating this cycle allows the concept generator to progressively adapt its proposals to the visual structure of the unlabeled training images.

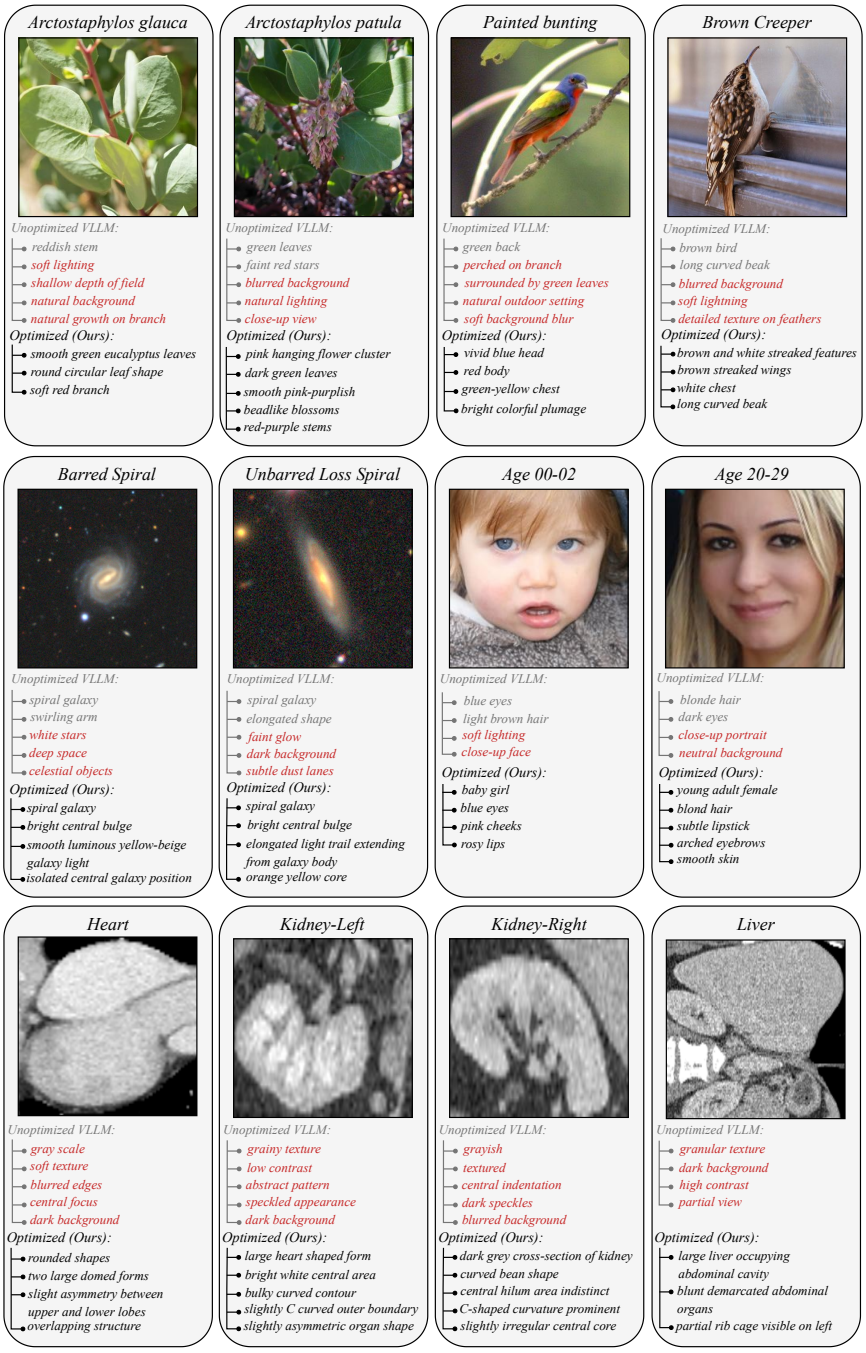


Figure 3: **Concepts Before and After Learning.** For each image, we compare concepts from the pretrained generator (top list) and S^{2}COPE (bottom list). Red marks concepts that are incorrect or uninformative for the category.

Evaluation. Following the linear evaluation protocol in self-supervised representation learning (Chen et al., 2020; He et al., 2020; Caron et al., 2021), we test whether the learned concepts support classification without further adapting the representation. After training, we freeze the model and generate descriptions for the training images and form a fixed concept bank. Each image is represented by its CLIP similarities to the concepts in this bank. We train a logistic regression classifier on these vectors and report top-1 accuracy on test images represented using the same bank. Class labels are used to train a linear probe classifier to evaluate the quality of the interpretable representation, not to discover concepts or construct the bank. Moreover, to assess the quality of the concepts, we report their win rates over the compared baselines, as judged by both a separate vision-language model and

Table 1: **Top-1 Concept-Based Classification Accuracy (%) Across Eight Datasets.** S^{2}COPE fine-tunes a VLLM on unlabeled iNaturalist-mini images, which is then transferred to other datasets without adaptation. "Labels" indicates class-label or class-name use in concept construction. Blue values denote percentage-point gains over the best of the four concept-based baselines per dataset; bold marks the highest accuracy. S^{2}COPE outperforms these baselines on all eight datasets, with an average gain of 7.37 percentage points.

Method	Labels	Nature		Physics	
		iNaturalist	CUB	Galaxy10	Gravity Spy
TextSpan	✓	20.00	81.08	24.07	4.25
DCLIP	✓	23.46	82.14	11.75	2.17
LF-CBM	✓	73.85	64.55	53.44	83.08
LaBo	✓	42.69	77.63	53.55	39.50
VLLM (Unoptimized)	✗	71.15	74.77	54.09	60.83
VLLM + S^{2}COPE (Ours)	✗	**85.00** (+11.15)	**85.11** (+2.97)	**66.52** (+12.97)	**83.55** (+0.47)

Method	Labels	Medical			Faces
		HAM10000	OrganCMNIST	OrganMNIST3D	FairFace
TextSpan	✓	17.10	7.40	12.42	50.95
DCLIP	✓	33.16	10.41	14.29	46.16
LF-CBM	✓	70.98	78.85	65.22	53.47
LaBo	✓	63.73	54.10	22.36	39.99
VLLM (Unoptimized)	✗	65.28	68.06	58.39	50.63
VLLM + S^{2}COPE (Ours)	✗	**79.27** (+8.29)	**91.43** (+12.58)	**70.19** (+4.97)	**59.01** (+5.54)

Table 2: **Pairwise Preference Win Rate for S^{2}COPE Concepts (%).** For each opponent, we report the percentage of the 39 images for which the evaluator prefers S^{2}COPE's anonymized concept list over the opponent's. GPT-5.6-Sol compares S^{2}COPE against five baselines, while human evaluators compare S^{2}COPE against the unoptimized VLLM, with results averaged across evaluators. Our evaluation confirms that S^{2}COPE is able to generate concepts that represent the image better.

Evaluator	GPT-5.6-Sol					**Human**
Opponent	Unoptimized VLLM	TextSpan	DCLIP	LF-CBM	LaBo	Unoptimized VLLM
S^{2}COPE preference	85.9	98.7	96.2	89.7	98.7	**96.41**

human evaluators. In addition, we show that removing training images flagged by the discovered bias concepts yields more robust classifiers on downstream tasks.

4 Experiments

4.1 Experimental Setup

Datasets. We optimize the concept generator on unlabeled **iNaturalist-mini** images (Van Horn et al., 2021), following the split of Chiquier et al. (2024), and evaluate frozen transfer across eight datasets in four domains. Except for the iNaturalist evaluation set, all datasets are unseen during training. **Nature**: **iNaturalist** (Van Horn et al., 2021; Chiquier et al., 2024) and **CUB** (Wah et al., 2011) (fine-grained biological categories). **Faces**: **FairFace** (Karkkainen & Joo, 2021) (nine age groups). **Medical**: **HAM10000** (Tschandl et al., 2018) (skin lesions), **OrganCMNIST**, and **OrganMNIST3D** (Yang et al., 2023a) (CT scans). **Physics**: **Galaxy10** (Leung & Bovy, 2019) (galaxy morphologies) and **Gravity Spy** (Zevin et al., 2017) (gravitational-wave spectrograms).

Baselines. We compare against four supervised concept-based methods that also produce interpretable concepts. **TextSpan** (Gandelsman et al., 2024) decomposes CLIP image representations into text-interpretable components via a spanning set of natural-language descriptions. **DCLIP** (Menon & Vondrick, 2023) performs zero-shot classification using LLM-generated per-class descriptions matched via CLIP. **LF-CBM** (Oikarinen et al., 2023) prompts an LLM with class names to build a

Table 3: **Concept-Guided Shortcut Mitigation on ISIC.** Colored calibration patches are spuriously associated with benign lesions. S^{2}COPE identifies affected training images; removing them and retraining the linear probe reduces the patch-induced error rate from 74.9% to 59.9%. S^{2}COPE thus enables targeted data curation to reduce shortcut reliance.

Setting	Patch detection (%)		Error rate (%) ↓
	Precision ↑	Recall ↑	
VLLM (Unoptimized)	—	—	74.9
VLLM + S^{2}COPE	95.5	83.4	**59.9**

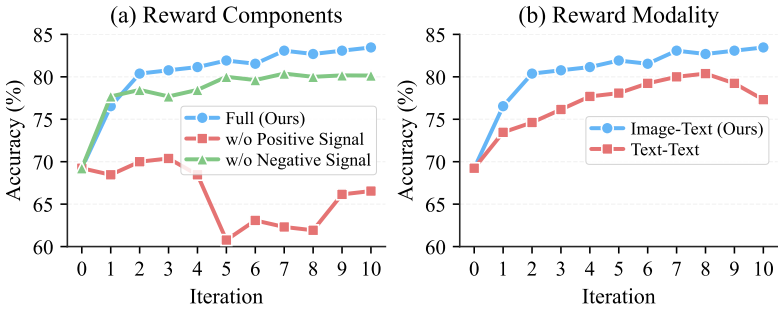


Figure 4: **Ablation Studies on Reward Formulation. (a) Reward Components:** Impact of isolating the positive and negative signals of the contrastive reward. Eliminating the positive signal causes a performance collapse, while removing the negative signal yields a suboptimal accuracy plateau. **(b) Reward Modality:** Comparison of cross-modal image-text grounding versus unimodal text-text consensus. Cross-modal grounding against physical image features achieves better performance than relying on unimodal textual consensus.

concept bank, then trains a linear classifier on CLIP concept activations. **LaBo** (Yang et al., 2023b) further applies submodular selection to retain a maximally discriminative concept subset.

Models. For fair comparison, all methods, both baselines and ours, employ the same Qwen3-VL-8B-Instruct (Yang et al., 2025; Bai et al., 2025) for concept generation (except TextSpan, which uses its released general description set) and OpenCLIP ViT-H/14 (Cherti et al., 2023) for concept-image similarity scoring, isolating each method's concept discovery strategy and the contribution of S^{2}COPE from differences in model capacity.

Implementation Details. We use a batch size of 1,300 for the contrastive reward calculation. We optimize with batch size of 512. We iterate the learning by $K = 10$ rounds. We use learning rate of 5e-6 for the Qwen3-VL vision tower and 1e-6 for the rest of Qwen3-VL parameters. Our training runs on one node with 8 RTX 6000 Blackwell Pro GPUs.

4.2 Main Results

We evaluate two uses of the learned concepts: representing images for recognition and exposing shortcuts for targeted data curation. In both settings, the concept generator is trained only on unlabeled iNaturalist-mini images and transferred without adaptation.

4.2.1 Cross-Domain Recognition and Preference Evaluation

Probing Results. Table 1 reports concept-based classification accuracy across eight datasets. S^{2}COPE outperforms all four concept-based baselines on every dataset, with a mean gain of 7.37 percentage points over the best of these baselines on each dataset. Gains range from +0.47 points on Gravity Spy to +12.97 points on Galaxy10, with a similarly large gain on OrganCMNIST (+12.58 points), demonstrating transfer to domains absent from concept-generator training. S^{2}COPE also consistently improves over the unoptimized VLLM, supporting the benefit of self-supervised preference learning for concept discovery.

Augmentation	Accuracy (%)
No augmentation	**83.46**
+ Horizontal flip, color jitter, Gaussian blur	82.69
+ Random crop	82.31
+ Grayscale, wider crop, stronger jitter and blur	81.92
Random crop only	80.77
Random crop, horizontal flip only	81.92
Color jitter, grayscale, Gaussian blur only	81.92

Table 4: **Effect of Data Augmentation.** We vary the image augmentations used to compute the contrastive reward, adding them cumulatively (top) or evaluating them separately (bottom). Using the original images achieves the highest accuracy among the evaluated settings.

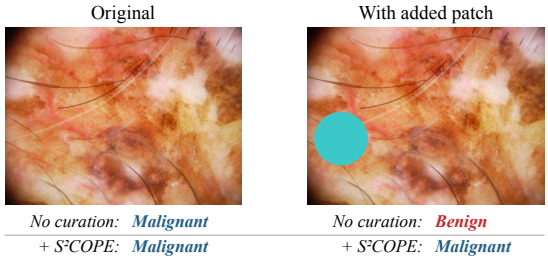


Figure 5: **Concept-Guided Shortcut Mitigation.** Adding a patch causes this malignant image to be misclassified as benign. "+ S^{2}COPE" denotes concept-guided data curation and probe retraining, which preserve the correct prediction. S^{2}COPE enables targeted data curation that removes bias in the data.

Pairwise Preference Evaluation. We evaluate concept quality through VLLM-based and human comparisons on a shared set of 39 images drawn from our evaluation domains (Table 2). GPT-5.6-Sol prefers S^{2}COPE concepts over the unoptimized VLLM in 85.9% of judgments and over each concept-based baseline in at least 89.7%, with each pair evaluated in both A/B orders (Appendix A.7). Ten volunteers select the more informative description of the depicted object, preferring S^{2}COPE over the unoptimized VLLM at a mean rate of 96.41%. These results support the improved descriptive quality of concepts learned through our self-supervised preference optimization.

4.2.2 Discovering and Mitigating Spurious Correlations

In addition to predictive accuracy, which measures the capability of our discovered concepts, we further test whether the discovered concepts can expose and help mitigate data bias. We study ISIC dermoscopy, where colored calibration patches are associated with benign lesions (Nauta et al., 2021; Rieger et al., 2020). On a 29,967-image ISIC dataset (Codella et al., 2018), our learned generator produces concepts such as "orange circular marker," "green circular patch," and "round sticker-like texture" without being supplied artifact names or patch annotations, which enables the inspection of the confounding factor.

In Table 3, "Unoptimized" denotes the pipeline before data curation, while "+ S^{2}COPE" denotes concept-guided filtering and probe retraining, where patch-group labels are used only to select which concept clusters to filter on, not to generate concepts; both use the same frozen VLLM and concept bank. Filtering identifies artifact-containing images with 95.5% precision and 83.4% recall. We measure the patch-induced error rate on malignant test images correctly classified before patch insertion. Retraining the probe on the filtered data reduces this rate from 74.9% to 59.9%, a 15.0-point reduction. Figure 5 illustrates the resulting robustness to an inserted patch. These results demonstrate that the discovered concepts support targeted data curation to mitigate shortcut reliance (Appendix A.6).

4.3 Ablation Studies and Analysis

Unless otherwise specified, we conduct ablations on the Qwen3-VL-8B-Instruct model with S^{2}COPE for $K = 10$ iterations on unlabeled iNaturalist-mini images and report top-1 accuracy with the linear probe trained on iNaturalist-mini rather than the full iNaturalist training set used in Table 1.

Contrastive Reward Components. To validate the necessity of each reward component, we ablate the positive and negative signals of the contrastive reward (Figure 4(a)). Removing the negative signal reduces the objective to measuring grounding in the source image alone, yielding generic concepts that lack distinctiveness from other instances. Eliminating the positive signal causes a sharp drop, as the optimization rewards distinctiveness indiscriminately and internalizes background noise as valid

Base Model	Top-1 Accuracy (%)	
	Unoptimized	+ S^{2}COPE
Qwen3-VL-2B-Instruct	67.31	**80.77**
Qwen3-VL-4B-Instruct	69.62	**82.31**
Qwen3-VL-8B-Instruct	69.23	**83.46**

Table 5: **Impact of Base Model Capacity.** Ablation over VLLM scale (2B, 4B, 8B). S^{2}COPE yields consistent gains (12.7–14.2 points) across all scales, with accuracy scaling monotonically with model size.

Table 6: **Selection of the Negatives.** We compare three levels of hardness on the negative samples: Hard Neg (smallest gap), Random, and Easy Neg (largest gap). Each strategy selects 16 preference pairs per image. Easy negatives yield the best accuracy.

Strategy	Unoptimized VLLM	VLLM + S^{2}COPE		
		Hard Neg	Random	Easy Neg
Top-1 Accuracy (%)	69.23	76.92	81.92	**83.46**

attributes. Both signals are necessary—the positive grounds concepts in the source image, while the negative promotes distinctiveness from other instances.

Reward Modality. To determine the optimal grounding space for the contrastive reward, we compare two modalities: *Text-Text* (distance between textual concepts of different images) and *Image-Text* (alignment between proposed concepts and raw visual embeddings). As shown in Figure 4(b), *Image-Text* achieves substantially higher accuracy, as anchoring candidates directly against physical image features produces a cleaner preference signal than relying on intermediate linguistic representations.

Data Augmentation. Standard contrastive learning for visual representation learning applies augmentations (Tian et al., 2020; Chen et al., 2020). In Table 4, we study which image augmentations work best for our algorithm. Surprisingly, we find that the highest accuracy is achieved with no augmentation (83.46%), exceeding the unoptimized VLLM (69.23%) by 14.23 percentage points. Cumulatively adding augmentations reduces accuracy to 81.92%, while random crop alone yields the lowest accuracy (80.77%). We therefore use original images to compute the reward. Since augmentations in contrastive learning encourage the model to be invariant to that augmentation, our result suggests that making no prior assumption about what features to be invariant is preferable for self-supervised concept discovery.

Base Model Capacity. To assess generalization across model scales, we evaluate the 2B, 4B, and 8B variants of Qwen3-VL-Instruct (Table 5). Despite similar zero-shot baselines (67%–69%), S^{2}COPE yields consistent gains of 12.7–14.2 points across all scales. Accuracy scales monotonically with model size (83.46% for 8B), suggesting that S^{2}COPE benefits from stronger base models.

Importance of Easy Negatives. To study the effect of preference pair construction, we compare three strategies defined by the reward gap between chosen and rejected descriptions: *Hard Neg.* (smallest gap), *Random*, and *Easy Neg.* (largest gap), each selecting 16 pairs per image (Table 6). Easy negatives achieve the best accuracy (83.46%), while hard negatives degrade to 76.92%. The self-supervised reward inherently contains noise from the continuous embedding space; forcing the model to distinguish near-identical scores amplifies this noise. Easy negatives sidestep this by isolating the widest reward discrepancies, providing an unambiguous corrective signal.

Qualitative Analysis. Figure 3 compares concepts before and after S^{2}COPE optimization (see Appendix A.10 for extended visualizations). Red marks errors and visually plausible descriptions that are uninformative for the category, including generic captions and photographic filler. After S^{2}COPE optimization, the model suppresses these concepts and produces more specific descriptions of visible attributes in these examples.

5 Conclusion

We introduce S^{2}COPE, a framework that discovers human-readable visual concepts from unlabeled images through self-supervised preference learning. The learned concepts improve recognition across diverse domains and expose spurious cues that guide data curation. As accuracy continues to grow with more unlabeled images (Appendix A.5), S^{2}COPE offers a scalable path toward self-supervised interpretable visual representations.

AI USE STATEMENT

We used generative AI tools only to polish the writing of this paper (grammar and readability). We did not use them for any other task requiring disclosure under the ICLR 2027 AI Policy. All AI-suggested edits were reviewed sentence by sentence by the authors, who take full responsibility for the content.

ETHICS STATEMENT

S^2COPE discovers concepts with a pretrained VLLM and inherits the biases of its pre-training data (Appendix A.13). All datasets and benchmarks are publicly available and used under their licenses. The human evaluation (Section 4.2.1) involved volunteers comparing anonymized pairs of concept lists; no personal information was collected.

REPRODUCIBILITY STATEMENT

The method is described in Section 3, with the full procedure in Algorithm 1 (Appendix A.2), and the datasets and evaluation protocol in Section 4.1. All training and evaluation hyperparameters and hardware are listed in Table 7 (Appendix A.1). The ISIC shortcut-mitigation and VLLM-as-a-Judge protocols are detailed in Appendices A.6 and A.7. We will release our code upon acceptance.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Antonio Almudévar, José Miguel Hernández-Lobato, and Alfonso Ortega. There was never a bottleneck in concept bottleneck models. *arXiv preprint arXiv:2506.04877*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

Davide Caffagni, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, Pier Luigi Dovesi, Shaghayegh Roohi, Mark Granroth-Wilding, and Rita Cucchiara. Seeing beyond words: Self-supervised visual learning for multimodal large language models. *arXiv preprint arXiv:2512.15885*, 2025.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 132–149, 2018.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

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

Ruijun Chen, Jiajian Guo, Hongzhan Chen, Fanqi Wan, Qifan Wang, and Xiaojun Quan. Realign: Structured revision for small language model alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 12005–12020, 2025.

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

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2818–2829, 2023.

Mia Chiquier, Utkarsh Mall, and Carl Vondrick. Evolving interpretable visual classifiers with large language models. In *European Conference on Computer Vision*, pp. 183–201. Springer, 2024.

Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pp. 168–172. IEEE, 2018.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Antonio De Santis, Schrasing Tong, Marco Brambilla, and Lalana Kagal. Learning concept bottleneck models from mechanistic explanations. In *The Fourteenth International Conference on Learning Representations*, 2026.

Xinpeng Ding, Kui Zhang, Jianhua Han, Lanqing Hong, Hang Xu, and Xiaomeng Li. Pami-vdpo: Mitigating video hallucinations by prompt-aware multi-instance video preference learning. *arXiv preprint arXiv:2504.05810*, 2025.

Jinlan Fu, Hao Fei, Xiaoyu Shen, Bryan Hooi, Xipeng Qiu, See-Kiong Ng, et al. Chip: Cross-modal hierarchical direct preference optimization for multimodal llms. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting clip's image representation via text-based decomposition. In *The Twelfth International Conference on Learning Representations*, 2024.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14375–14385, 2024.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Raban Iten, Tony Metger, Henrik Wilming, Lídia Del Rio, and Renato Renner. Discovering physical concepts with neural networks. *Physical review letters*, 124(1):010508, 2020.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Hongrui Jia, Chaoya Jiang, Haiyang Xu, Wei Ye, Mengfan Dong, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. Symdpo: Boosting in-context learning of large multimodal models with symbol demonstration direct preference optimization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9361–9371, 2025.

Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1548–1558, 2021.

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

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pp. 5338–5348. PMLR, 2020.

Henry W Leung and Jo Bovy. Deep learning of multi-element abundances from high-resolution spectroscopic data. *Monthly Notices of the Royal Astronomical Society*, 483(3):3255–3277, 2019.

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

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Meike Nauta, Ricky Walsh, Adam Dubowski, and Christin Seifert. Uncovering and correcting shortcut learning in machine learning models for skin cancer diagnosis. *Diagnostics*, 12(1):40, 2021.

Tuomas Oikarinen and Tsui-Wei Weng. Clip-dissect: Automatic description of neuron representations in deep vision networks. In *The Eleventh International Conference on Learning Representations*, 2023.

Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, 2024.

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. *arXiv preprint arXiv:2504.02821*, 2025.

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.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Laura Rieger, Chandan Singh, William Murdoch, and Bin Yu. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *International conference on machine learning*, pp. 8116–8126. PMLR, 2020.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, 2018.

Fawaz Sammani, Jonas Fischer, and Nikos Deligiannis. Clip-free, label-free, zero-shot concept bottleneck models. *arXiv preprint arXiv:2503.10981*, 2025.

Simon Schrodi, Julian Schur, Max Argus, and Thomas Brox. Selective concept bottleneck models without predefined concepts. *Transactions on Machine Learning Research*, 2025.

Vésteinn Snæbjarnarson, Kevin Du, Niklas Stoehr, Serge Belongie, Ryan Cotterell, Nico Lang, and Stella Frank. Taxonomy-aware evaluation of vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9109–9120, 2025.

Divyansh Srivastava, Ge Yan, and Tsui-Wei Weng. Vlg-cbm: Training concept bottleneck models with vision-language guidance. *Advances in Neural Information Processing Systems*, 37:79057–79094, 2024.

Guohao Sun, Can Qin, Yihao Feng, Zeyuan Chen, Ran Xu, Sohail Dianat, Majid Rabbani, Raghuveer Rao, and Zhiqiang Tao. Structured policy optimization: Enhance large vision-language model via self-referenced dialogue. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 741–751, 2025.

Sirnam Swetha, Rui Meng, Shwetha Ram, Tal Neiman, Son Tran, and Mubarak Shah. Smpro: Self-supervised visual preference alignment via differentiable multi-preference multi-group ranking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 37951–37960, 2026.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1): 180161, 2018.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12884–12893, 2021.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

Haochen Wang, Anlin Zheng, Yucheng Zhao, Tiancai Wang, Zheng Ge, Xiangyu Zhang, and Zhaoxiang Zhang. Reconstructive visual instruction tuning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Penghao Wu, Yushan Zhang, Haiwen Diao, Bo Li, Lewei Lu, and Ziwei Liu. Visual jigsaw post-training improves mllms. *arXiv preprint arXiv:2509.25190*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific data*, 10(1):41, 2023a.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19187–19197, 2023b.

Heeji Yoon, Jaewoo Jung, Junwan Kim, Hyungyu Choi, Heeseong Shin, Sangbeom Lim, Honggyu An, Chaehyun Kim, Jisang Han, Donghyun Kim, et al. Visual representation alignment for multimodal large language models. *arXiv preprint arXiv:2509.07979*, 2025.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023.

Michael Zevin, Scott Coughlin, Sara Bahaadini, Emre Besler, Neda Rohani, Sarah Allen, Miriam Cabero, Kevin Crowston, Aggelos K Katsaggelos, Shane L Larson, et al. Gravity spy: integrating advanced ligo detector characterization, machine learning, and citizen science. *Classical and quantum gravity*, 34(6):064003, 2017.

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

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image bert pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022.

Ke Zhu, Liang Zhao, Zheng Ge, and Xiangyu Zhang. Self-supervised visual preference alignment. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 291–300, 2024.

A TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

A.1 IMPLEMENTATION DETAILS

See Table 7 for all implementation details.

Table 7: **Implementation Details.**

Hyperparameter	**Value**
Concept Proposal	
Base VLLM	Qwen3-VL-8B-Instruct
Sampling temperature	2.0
Top-p / Top-k	1.0 / 100
Repetition penalty	1.1
Candidate descriptions per image	16
Contrastive Reward	
Cross-modal encoder	OpenCLIP ViT-H/14 (LAION-2B)
Reward batch size	1,300
Reward temperature τ	0.07
Preference Pairing	
Pairing strategy	Easy Neg.
Candidate pairs per image	$\binom{16}{2} = 120$
Selected pairs per image	16
DPO Training	
DPO β	0.05
DPO loss	Sigmoid
Epochs per iteration	3
Global batch size	512
LLM / Merger learning rate	1×10^{-6}
Vision tower learning rate	5×10^{-6}
LR schedule	Constant with 1% warmup
Weight decay	0.01
Precision	bf16
Pipeline	
Total iterations	10
GPUs	8× RTX 6000 Blackwell Pro
Total training time	22 hours (176 GPU hours)
Evaluation	
Sampling temperature	0.0
Repetition penalty	1.0
Deduplication threshold	0.65
Classifier	Logistic Regression (LBFGS)
Max iterations	5,000
L2 regularization λ	1×10^{-3}

A.2 ALGORITHM

The complete optimization procedure described in Section 3 is detailed in Algorithm 1.

Algorithm 1: S^{2}COPE: Self-Supervised Concept Discovery via Preference Learning

Input: Unlabeled dataset $\mathcal{X}$, Initial VLLM concept generator π_0, Frozen cross-modal encoder $\mathcal{E}$ (e.g., CLIP), Iterations K, Batch size B, Candidate descriptions per image N
Output: Optimized concept generator π_K

// Iterative self-supervised discovery loop
for $k = 0, \ldots, K-1$ **do**
  $\mathcal{D}_k \leftarrow \emptyset$ // Initialize preference dataset for iteration k
  Sample a batch of unlabeled images $\mathcal{B} = \{x_1, \ldots, x_B\} \sim \mathcal{X}$
  // Encode original images with the frozen vision encoder
  $v_x \leftarrow \mathcal{E}(x)$ for each $x \in \mathcal{B}$; $\mathcal{V} \leftarrow \{v_x : x \in \mathcal{B}\}$
  for *each* *anchor image* $x \in \mathcal{B}$ **do**
    // 1. Sample candidate descriptions
    Sample $\mathcal{C}(x) = \{c_1, \ldots, c_N\} \sim \pi_k(\cdot \mid x)$ using high-temperature decoding
    // 2. Cross-Modal Contrastive Scoring
    for *each* *description* $c \in \mathcal{C}(x)$ **do**
      $t_c \leftarrow \mathcal{E}(c)$ // Extract normalized text embedding
      // Score grounding and distinctiveness
      $R(c, x) \leftarrow \frac{t_c^\top v_x}{\tau} - \log \sum_{v \in \mathcal{V}} \exp\left(\frac{t_c^\top v}{\tau}\right)$
    // 3. Easy Negative Preference Selection
    $\mathcal{P}_x \leftarrow$ the 16 pairs (c_i, c_j) in $\mathcal{C}(x)$ with the largest absolute reward gaps $|R(c_i, x) - R(c_j, x)|$
    for *each* $(c_i, c_j) \in \mathcal{P}_x$ **do**
      $c_w \leftarrow \arg\max_{c \in \{c_i, c_j\}} R(c, x)$ // Higher-scoring description
      $c_l \leftarrow \arg\min_{c \in \{c_i, c_j\}} R(c, x)$ // Lower-scoring description
      $\mathcal{D}_k \leftarrow \mathcal{D}_k \cup \{(x, c_w, c_l)\}$
  // 4. Preference Optimization for Autonomous Discovery
  Update π_{k+1} by minimizing the DPO loss $\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_k)$ on dataset $\mathcal{D}_k$
return π_K

A.3 Comparison with Latent Self-Supervised Learning

Table 8: **Comparison with Latent Self-Supervised Learning.** Linear probing accuracy comparing the continuous, black-box representations of SimCLR against our discrete, interpretable concepts. Our approach yields a 20.4-point absolute improvement over the fine-tuned latent baseline.

Method	Top-1 Accuracy (%)
SimCLR (Pretrained ResNet-50)	55.00
SimCLR (Fine-tuned ResNet-50)	63.08
S^{2}COPE (Ours)	**83.46**

To benchmark our interpretable visual lexicons against standard latent self-supervised learning, we compare our framework with SimCLR (Chen et al., 2020). We first evaluate a pretrained SimCLR ResNet-50 backbone by removing its projection head and training a linear probe directly on the frozen representations. We then fine-tune this backbone on our unlabeled dataset by initializing a new projection head before applying the identical linear probing protocol. Both SimCLR baselines strictly follow the data augmentation pipeline detailed in the original implementation.

Table 8 shows that our method outperforms both SimCLR baselines. Fine-tuning improves SimCLR accuracy from 55.00% to 63.08%, while our approach achieves 83.46%, a gain of 20.4 percentage points over the fine-tuned baseline. These results show that S^{2}COPE learns interpretable concepts with strong discriminative performance relative to the evaluated SimCLR representations. Since the methods differ in architecture and pretraining, this comparison does not isolate the effect of the concept bottleneck.

A.4 Cross-Modal Encoder Capacity

To evaluate sensitivity to the cross-modal encoder, we ablate the frozen CLIP architecture used for the contrastive reward (Table 9). Both encoders yield strong gains over the unoptimized VLLM, with ViT-H/14 achieving a modest advantage over ViT-B/16 (83.46% vs. 82.31%).

Table 9: **Impact of Cross-Modal Encoder Capacity.** Ablation over the frozen CLIP encoder used for the contrastive reward.

Encoder	**Unoptimized VLLM**	**VLLM + S^{2}COPE**	
		ViT-B/16	**ViT-H/14**
Top-1 Accuracy (%)	69.23	82.31	**83.46**

A.5 Scaling Unlabeled Data

We vary the number of unlabeled training images from 100 to 1,300 (Table 10). Accuracy rises from 76.54% to 83.46% and peaks at the largest scale, suggesting that S^{2}COPE will continue to benefit from larger unlabeled collections.

Table 10: **Impact of Unlabeled Dataset Size.** Accuracy has not saturated at 1,300 unlabeled images, suggesting that S^{2}COPE scales effectively with more unlabeled data.

Dataset Size	**Unoptimized**	**+ S^{2}COPE**				
		100	400	700	1000	1300
Top-1 Acc. (%)	69.23	76.54	80.38	81.15	81.15	**83.46**

A.6 ISIC Shortcut-Mitigation Protocol

The concept generator remains frozen throughout the ISIC experiment. For downstream filtering, we rank the discovered concept clusters by training-set AUROC against SONIC group membership (SONIC is the ISIC Archive collection of benign lesions photographed with colored patches) and retain the top ten. Each training image receives the maximum activation over these clusters. We set the threshold to match the fraction of training descriptions that mention a patch, remove the 6,463 flagged images, and retrain the linear probe using the same concept bank and probe hyperparameters. Group membership is used for cluster selection and detection evaluation in this application, but not for training the concept generator or generating concepts. The error rate under patch insertion is computed separately for each probe over the malignant test images that it classifies correctly before patch insertion.

A.7 VLLM-as-a-Judge Protocol

We evaluate a fixed set of 39 before-and-after examples spanning natural images, galaxies, and blood-cell microscopy. This set is not a random sample of the eight classification benchmarks. GPT-5.6-Sol (model identifier `gpt-5.6-sol`, accessed through the OpenAI Codex CLI) receives each image and its two concept lists without model identities or class labels. The prompt asks which list describes the main subject more accurately and specifically, penalizing incorrect, unsupported, generic, or repetitive details. It instructs the judge to disregard background, lighting, camera style, and image quality, and not to favor longer descriptions.

Each pair is evaluated in both A/B orders with medium reasoning effort, and the judge is constrained to return a JSON verdict of A or B. Win rates in Table 2 are computed over the resulting 78 judgments (39 images $\times$ 2 orders); e.g., the evaluator prefers S^{2}COPE over the unoptimized VLLM in 67 of them (85.9%). The result measures preference on this example set, not a benchmark-wide quality estimate.

Baseline concept lists. The four baselines do not describe individual images; each constructs a dataset-level concept bank and grounds it with CLIP. To obtain a per-image concept list for a baseline, we take the concept bank it produced for the dataset the image belongs to (iNaturalist, CUB, Galaxy10, or BloodMNIST) and select the k bank entries with the highest OpenCLIP ViT-H/14 image-text cosine similarity, where k equals the number of phrases in the S^{2}COPE description for that image so that list lengths match. LaBo uses its submodularly selected concepts, LF-CBM its filtered concept set, DCLIP the union of its per-class descriptors, and TextSpan the descriptions selected by its spanning procedure (the full general description set for CUB and Galaxy10, whose spanning sets were not retained). These lists are then judged against the S^{2}COPE concepts under the same protocol. Note that LF-CBM and DCLIP concept banks are conditioned on class names, so their concepts occasionally contain taxonomic identifiers (e.g., "Arctostaphylos"), which the judge credits when correct.

A.8 Quantitative Evolution of Concept Diversity

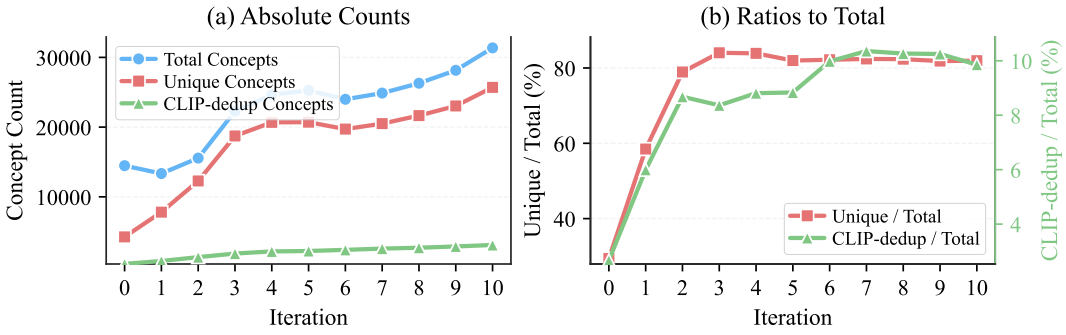


Figure 6: **Evolution of Concept Diversity.** We track the absolute counts (a) and ratios (b) of unique and CLIP-deduplicated concepts generated across optimization iterations. Our optimized concept generator progressively escapes generic mode collapse, expanding its vocabulary to produce a richer, highly distinct, and less redundant visual lexicon.

To demonstrate that S^{2}COPE successfully suppresses repetitive conversational priors, we track the diversity of the generated concept pool across iterations. Figure 6 illustrates the absolute counts and ratios of unique and CLIP-deduplicated concepts. The unoptimized base model (Iteration 0) suffers from generic mode collapse, yielding a highly redundant lexicon with a low uniqueness ratio. As optimization progresses, both the absolute volume and the proportion of distinct concepts increase substantially and stabilize. This suggests that the contrastive reward encourages the VLLM to actively explore the long tail of its vocabulary, expanding its capacity to articulate granular visual details.

A.9 Qualitative Trajectory of Concept Refinement

Figure 7 qualitatively tracks how hypothesized concepts for individual images evolve throughout the S^{2}COPE loop. Initially, the unoptimized base model defaults to broad, non-discriminative linguistic priors. Through successive iterations, the optimization systematically discards these generic labels, refining the lexicon into increasingly precise descriptions of visible structures.

A.10 Extended Visualizations Across Domains

Figure 8 compares concepts from the unoptimized VLLM and the optimized concept generator across our evaluation domains. The examples show a shift from generic descriptions toward more specific visual attributes.

A.11 Human Study Samples

Figure 9 shows two sample questions, which both the human evaluators and the VLLM judge (Appendix A.7) answer.

	Atrovirens 	*Groove Billed Ani* 	*Barred Spiral* 
Base (Iter 0)	*green plant* *dense foliage background* *natural outdoor setting* *fine textured stems* *wild vegetation*	*black bird* *perched bird* *green foliage* *dense vegetation* *leafy branches*	*spiral galaxy* *dark background* *scattered stars* *faint starlight* *cosmic dust*
Iter 3	*green plant* *tall green leaf* *green flower cluster* *bunch of small green buds* *green wetland vegetation*	*black bird* *long tail* *dark glossy feathers* *small fruit buds* *lush vegetation*	*spiral galaxy* *bright central bulge* *glowing disk* *faint halo around galaxy* *dark background space*
Iter 6	*green cyperus plant* *tall green stem* *green flower head cluster* *lush green wetland background* *verdant ferns surrounding the plant*	*black bird* *long tail* *shiny feathers* *dark coat* *lengthy tail feathers*	*spiral galaxy* *bright central bulge* *glowing disk* *barred spiral structure* *smooth halo around galaxy*
Iter 10	*green cyperus* *ball-shaped small green flowers* *tall stem center* *green sugar cane look stem* *vertical reed-like plant*	*black bird* *long black tail* *shiny black feathers* *dark black coat* *smooth dark beak*	*barred spiral galaxy in center* *bright yellow/orange disk* *white glowing spiral arms* *symmetric spiral structure* *faint glow around galactic core*

Figure 7: **Trajectory of Concept Refinement.** For each sample, we trace the evolution of generated concepts from the unoptimized base model (Iter 0) through successive S^{2}COPE iterations. Red marks concepts that are incorrect or uninformative for the category. Our optimized model suppresses these concepts and produces more specific visual attributes.

A.12 Limitations

Our concept discovery quality depends on the capacity of the frozen CLIP encoder and the VLLM's pre-trained vocabulary, both of which can be improved by adopting stronger foundation models as they become available. We train on a single source domain (iNaturalist) and observe strong cross-domain transfer; exploring diverse source domains is a natural next step. As shown in Table 10, accuracy improves overall with the number of unlabeled training images and has not yet saturated at 1,300 images, suggesting that further gains are achievable with larger-scale unlabeled data.

A.13 Broader Impacts

Our work advances interpretable AI by enabling concept discovery without human annotation, which has positive implications for high-stakes domains such as medical imaging and scientific discovery, where transparent reasoning is essential. By removing the reliance on labeled data, our framework also lowers the barrier to deploying interpretable models in data-rich but annotation-scarce settings. On the other hand, the discovered concepts inherit the biases present in the VLLM's pre-training data; practitioners should be aware of this when applying the framework to sensitive domains.

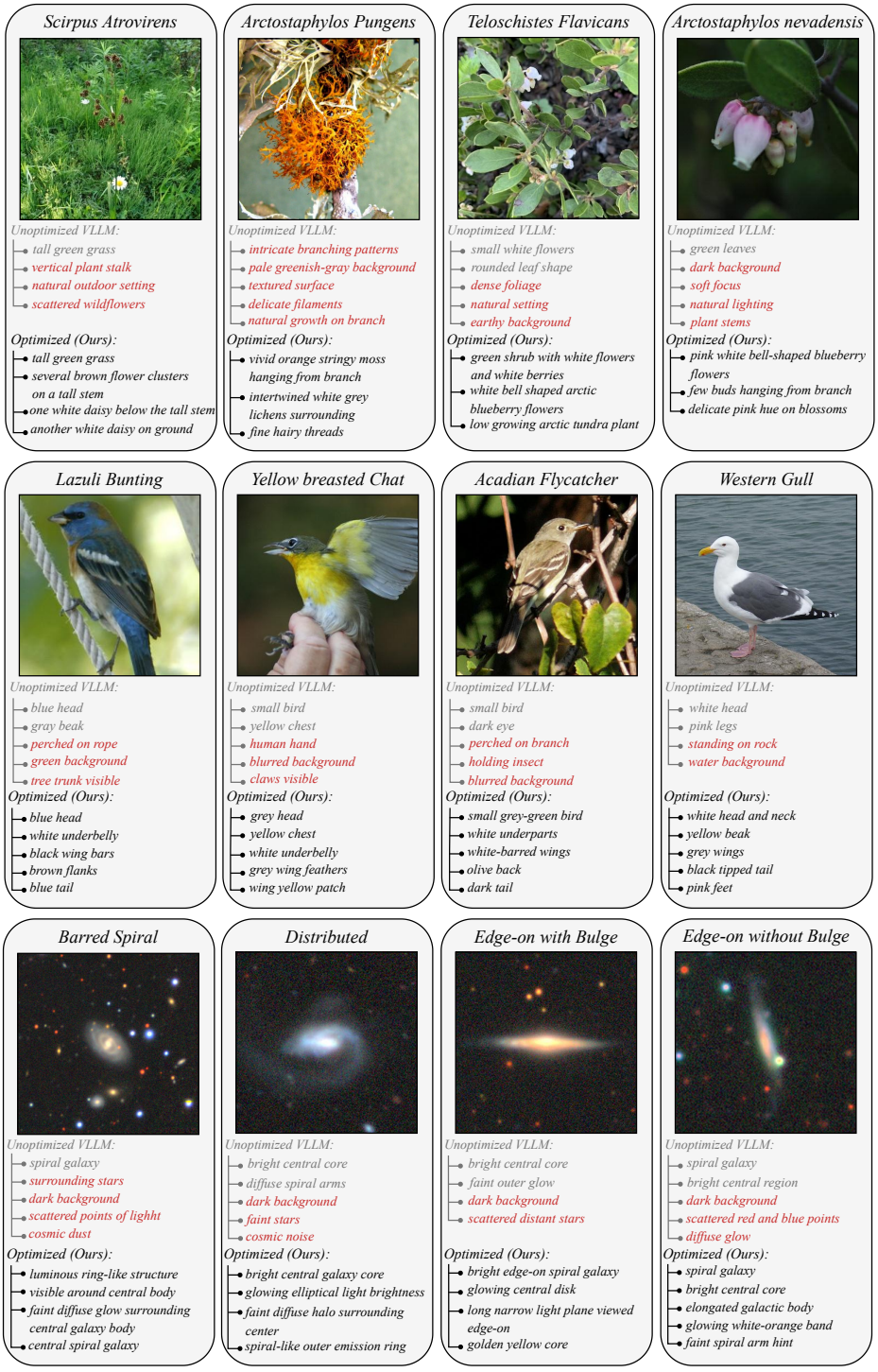


Figure 8: **Visualizing Self-Supervised Concept Discovery.** For each sample, we contrast the top concepts generated by the unoptimized base model (top list) with our optimized concept generator (bottom list). Red marks concepts that are incorrect or uninformative for the category. Our optimized model suppresses these concepts and produces more specific visual attributes.

User Study

Instruction:
Any information about the BACKGROUND is considered low quality.
Focus only on the OBJECT's features when making your choice.



Question:
Please choose the description that is more detailed and accurate about the OBJECT in the image.

Option A:
small bird, green background, perched on thorny branch, olive green body, white underbelly, black eye, dark wing bars, greyish olive wings, short beak, fluffy head ✓

Option B:
small bird, dark beak, black and white wing bars, perched on thorny branch, blurred green background, natural lighting, side profile view



Question:
Please choose the description that is more detailed and accurate about the OBJECT in the image.

Option A:
rocky terrain, small rounded leaves, scattered dry debris, natural outdoor setting, sunlit foliage, textured stones

Option B:
small bush with small green leaves growing between dark rocks, compact shrub, bright green foliage against stone ✓

Figure 9: **User Study Samples.** Two sample questions from our user study, which are also given to the VLLM judge. Participants follow the instruction and select among two anonymized concept lists.