Attributes Should Come from Images, Not Class Names: Distribution-Conditioned Attribute Selection for Vision-Language Models

Gautam Rajendrakumar Gare, Jia Shi, Zhiqiu Lin, Deepak Pathak, John Galeotti, and Deva Ramanan

Carnegie Mellon University, USA

Abstract. A popular route to interpretable zero-shot classification asks a large language model (LLM) to describe each class name and prompts CLIP with the resulting descriptors. We show that these descriptors carry little visual evidence of their own: removing the class name from the prompt collapses ImageNet accuracy from 59.5% to 15.5%. The diagnosis is that the descriptors are conditioned on the label rather than on the images, so they describe the concept in general and mislead exactly when the data shifts; an LLM insists that strawberries are red, but every strawberry in ImageNet-Sketch is a colorless line drawing. We therefore select attributes from the target image collection instead: we score a large attribute pool against the images in CLIP's joint embedding space and keep the top-scoring attributes per class. Selected this way, class-name-free attribute prompts reach 23.8% on ImageNet (against 15.5% for LLM descriptors), the gain holds on four shifted ImageNet variants, and re-selecting from the LLM's own pool isolates the selection mechanism as the cause. With one image per class, the selected attributes outperform the prompt-tuning method CoOp by 3 points while fitting in under a minute instead of 14 hours, with no learned soft prompt to obscure the decision. Because the attribute set is chosen by the data, it doubles as a readable summary of a dataset, which we use to describe distribution shift in words.

Keywords: Attributes · Vision-language models · Interpretability · Prompting

1 Introduction

Prompting a vision-language model (VLM) such as CLIP [18] with class names turns it into a zero-shot classifier. A popular refinement asks an LLM to describe each class ("a strawberry is red, has seeds, ...") and averages the prompts built from these descriptors [13,17], improving accuracy and offering an interpretation: the image was classified by its attributes. The descriptors, however, are generated from the class name alone; no image is ever consulted.

This paper starts from a direct test of what those descriptors measure. When we score the descriptors of Menon and Vondrick [13] without the class name in

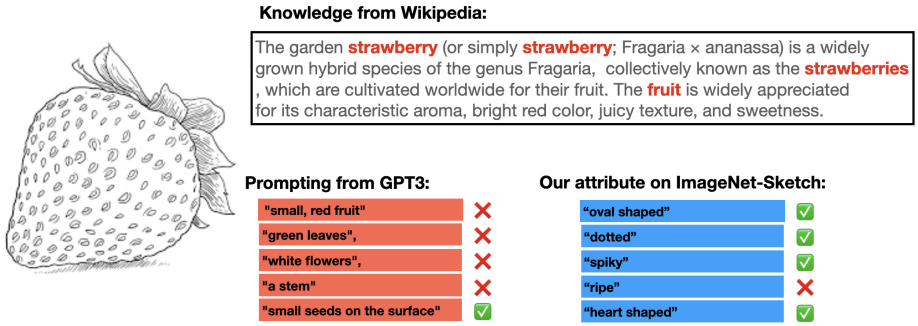

Fig. 1: Class-name-conditioned descriptors fail under distribution shift; image-conditioned attributes remain reliable. For the class *strawberry* on ImageNet-Sketch, an LLM prompted only with the class name [13] or an external knowledge base [22] produces canonical descriptors (e.g., *red*, *green leaves*) because neither ever sees an image. In contrast, our method scores a large pool of candidate attributes against the sketch images themselves, selecting semantic attributes that remain valid in the shifted distribution (e.g., *heart-shaped*, *dotted*) thus aligning them with distribution-specific cues such as *colorless* and *hand-drawn*, rather than relying on source-domain appearance priors. Section 3 quantifies the failure; Section 6.1 turns the same mechanism into a description of the shift.

the prompt, ImageNet accuracy collapses from 59.5% to 15.5% (Table 1). The descriptors themselves supply little visual evidence; the class name does the work. The same conditioning failure is visible qualitatively under distribution shift: the LLM describes strawberries as red and ripe, but every strawberry in ImageNet-Sketch is a colorless line drawing (Figure 1), so the descriptors point the classifier the wrong way precisely when the class name needs help.

Our diagnosis is that descriptors are conditioned on the wrong variable. A class name is a pointer to a concept, and an LLM can only describe the concept in general; the images at hand may look nothing like the general case. The fix follows from the diagnosis: condition attribute selection on the images. We score a large pool of candidate attributes against the target image collection in CLIP's joint embedding space and keep, per class, the attributes the images themselves rank highest. Like Menon and Vondrick [13], we classify with attribute prompts in the joint space of a frozen VLM; unlike them, we select attributes from the images rather than generating them from the class name, which matters because the attribute set then tracks the distribution it describes (Table 1, bottom block).

Conditioning on images pays off beyond the diagnostic. Attribute selection needs so few labeled images that it becomes a strong few-shot learner: with one or two images per class, the selected attributes outperform gradient-based prompt tuning while remaining human-readable and fitting in seconds. And because the selected attributes describe the data rather than the label, the mean attribute profile of a dataset is itself an interpretable object: subtracting the profiles of two datasets describes their distribution shift in words.

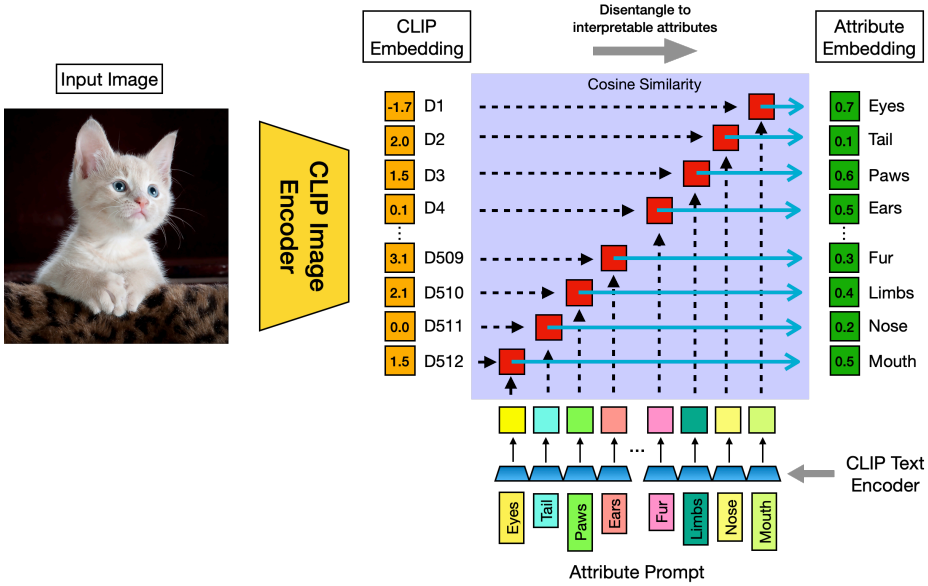

Fig. 2: Distribution-conditioned attribute selection. A frozen CLIP encodes the target images and a large pool of attribute texts; the cosine similarities between them form an attribute feature vector per image. Class-attribute weights are the per-class mean of these features (or the weights of a linear probe trained on them), and the top-k attributes per class define the class's prompt set. Both encoders stay frozen throughout; the only fitted object is the linear pairing.

Our contributions are threefold. First, we establish the class-name confound: LLM-generated descriptors lose most of their accuracy once the class name is removed from the prompt, so their apparent gains cannot be read as attribute-level evidence (Section 3). Second, we propose distribution-conditioned attribute selection, a training-light procedure that scores a fixed attribute pool against the target images with a frozen CLIP and keeps the top-ranked attributes per class; a matched-pool comparison attributes the resulting gains to the selection mechanism rather than the vocabulary (Sections 4 and 5). Finally, we show that the selected attributes act as an interpretable feature space: they support extreme few-shot classification, describe distribution shift in words, extend CLIP to vocabulary its text encoder cannot parse, and localize objects without class names (Sections 5.5 and 6).

2 Related Work

Descriptor prompting with LLMs. Menon and Vondrick [13] prompt GPT-3 for per-class descriptors and average the resulting prompts; CuPL [17] generates full prompt sentences the same way; K-LITE [22] draws external knowledge from WordNet and Wiktionary. All three condition the added text on the class

name alone. Roth et al. [20] showed that replacing such descriptors with random words matches their accuracy once the class name is present, questioning what the descriptors contribute. Our class-name-free evaluation is complementary: we measure the descriptors alone and show they collapse, then repair them by conditioning the selection on images rather than on the label.

Concept bottlenecks and interpretable embeddings. Concept bottleneck models route the prediction through human-named concept scores [10]; LaBo [26] and Label-Free CBM [14] remove the concept-annotation requirement by generating candidate concepts with an LLM and scoring them with CLIP, and SpLiCE [2] decomposes CLIP embeddings into sparse combinations of concept vectors. Architecturally our classifier is a label-free concept bottleneck: attribute scores feed a linear layer. The difference is where the concepts come from; the LLM-generated pools of [14, 26] inherit the class-name conditioning we diagnose in Section 3, whereas we select concepts with the target images and show that this choice, holding the pool fixed, is what moves accuracy.

Attribute datasets and prompt tuning. Attributes have a long history in zero-shot recognition [1, 5, 9, 11]; we use the modern large-scale pools VAW [15] and LSA [16] as our candidate vocabulary. On the adaptation side, CoOp [27] tunes continuous prompt vectors by gradient descent and WiSE-FT [25] ensembles fine-tuned and zero-shot weights; both are strong few-shot baselines but neither yields a human-readable decision. We compare against both in Section 5.5.

3 The Class-Name Confound

Descriptor accuracy rides the class name. The standard protocol of [13] prompts CLIP with "{class name}, which {has/is} {descriptor}" and averages over each class's descriptors. On ImageNet with CLIP-RN50 this reaches 59.5%, ahead of the 55.3% class-name-only baseline (Table 1, top block). The interpretability claim implicit in this protocol is that the descriptors measure attribute evidence in the image. If that were true, the descriptors should retain much of their accuracy when queried alone. They do not: with the template reduced to "{descriptor}", the same descriptor set scores 15.5% (Table 1, bottom block), a 44-point collapse, and the pattern repeats on every shifted ImageNet variant we test. Roth et al. [20] reached a consistent conclusion from the opposite direction: with the class name kept, random words do as well as LLM descriptors.

Why the conditioning is wrong. The descriptors are a function of the class name only, so they can at best describe the concept's typical appearance. Whenever the target distribution departs from typical, label-conditioned descriptors are not merely uninformative but wrong: on ImageNet-Sketch, "red" and "ripe" actively vote against every strawberry in the dataset (Figure 1). Two explanations for the collapse come to mind that do not indict the conditioning: CLIP might simply be unable to score attribute presence without a class anchor, or the GPT-generated pool might be too small and generic to describe images. Section 5.2 tests and rejects both by reselecting from the very same pool.

Finding. LLM-generated descriptors ride the class name: removing it from the prompt collapses ImageNet accuracy from 59.5% to 15.5%. Their gains cannot be read as attribute-level evidence about the image.

4 Method: Selecting Attributes with Images

The diagnosis prescribes the fix: attributes must be conditioned on the images they are meant to describe. We keep the pool of candidate attributes fixed and let the target image collection rank them, using nothing but a frozen CLIP. Figure 2 summarizes the pipeline.

Attribute scoring. Given an image I, the CLIP image encoder yields an embedding $v = f_{\text{img}}(I) \in \mathbb{R}^n$. Given a pool of m attribute strings $\{a_1, \ldots, a_m\}$, the text encoder yields embeddings $t_j = f_{\text{txt}}(a_j)$. The attribute feature of the image is the vector of scaled cosine similarities

$$s_j = \tau \cdot \cos(v, t_j), \qquad j = 1, \ldots, m, \tag{1}$$

with temperature $\tau = 100$ following [18]. We encode each attribute string bare, with no template and no class name. This choice is deliberate: appending the class name would correlate all of a class's attribute scores through the shared name, and it is exactly the confound of Section 3 that we need to avoid.

Class-attribute pairing. Given images with class labels, we estimate a score $w_{c,j}$ for how well attribute j describes class c, in one of two ways. The simplest sets $w_{c,j}$ to the mean of s_j over the images of class c; this needs no training at all and drives the distribution-level applications of Section 6. The classification experiments instead train a linear probe on the attribute features and use its per-class weights (plus bias) as $w_{c,j}$, which sharpens the estimate when labels are available. Either way, class c keeps the k attributes with the largest scores; this ranked word list A_c is the class's prompt set.

Classifying with selected attributes. A test image with embedding v is assigned to the class whose selected attributes it matches best:

$$\hat{c} = \arg\max_c \frac{1}{|A_c|} \sum_{j \in A_c} \omega_{c,j} \cdot \cos(v, t_j), \tag{2}$$

where $\omega_{c,j} = 1$ by default, so every selected attribute votes equally; the *weighted* variant of Section 5.3 sets $\omega_{c,j} = w_{c,j}$ instead, so the attributes most characteristic of a class carry more of its vote. Depending on the protocol under evaluation, each attribute is encoded bare or appended to the class name using the template of [13].

The attribute pool. We pool attributes from VAW [15], LSA [16], and the GPT-3 descriptors of [13], and preprocess the strings to remove explicit class names, duplicates, and punctuation (details in the supplementary material). Everything runs with both CLIP encoders frozen; no image-text pair is ever used for training. This puts the method at the opposite end of the cost spectrum from attribute-supervised VLMs such as STAIR [3], which trains on over a billion pairs to learn an attribute vocabulary.

Prompt	ImageNet	-V2	-Sketch	-A	-R
Class name in the prompt					
Class name only (zero-shot CLIP)	55.31	49.39	31.58	20.92	58.10
+ LLM descriptors [13]	59.47	52.84	33.75	23.51	57.32
+ Ours (top 5)	59.53	53.33	33.12	22.79	56.30
+ Ours (top 10)	60.18	53.40	33.66	23.43	57.49
+ Ours (top 100)	**60.80**	**53.95**	**34.27**	**24.00**	**59.04**
Attribute only (no class name)					
LLM descriptors [13]	15.50	14.00	9.60	8.57	17.91
Ours, same pool as [13] (top 5)	19.40	17.00	10.57	9.80	19.49
Ours, VAW+LSA pool [15, 16] (top 5)	**23.80**	**20.80**	**14.40**	**11.83**	**28.20**

Table 1: Attribute prompts stand on their own only when selected from images. Top-1 accuracy with CLIP-RN50; attributes are selected on ImageNet only and reused unchanged on the four shifted variants. Top block: with the class name in the prompt, all descriptor methods sit within about a point of each other. Bottom block: with the class name removed, LLM descriptors collapse (59.47 to 15.50 on ImageNet) while our selection recovers 3.9 points from the identical pool and 8.3 points from a larger pool. **Bold** marks the best result per column within each block.

5 Results

5.1 Setup

All experiments use CLIP with the ResNet-50 backbone [6, 18] unless stated otherwise. We evaluate top-1 accuracy on ImageNet and four shifted variants: ImageNetV2 [19], ImageNet-Sketch [24], ImageNet-A [8], and ImageNet-R [7]. Attributes are always selected on ImageNet training images only; the shifted variants are never seen during selection. Attribute-pool preprocessing and full training details are in the supplementary material.

5.2 Attribute-only classification

Selection, not vocabulary, closes the gap. We return to Table 1 (bottom block), now with our rows. Reselecting attributes from the exact pool of [13], conditioned on ImageNet training images, lifts class-name-free accuracy from 15.50% to 19.40%. The pool, the model, and the evaluation protocol are identical; only the selection mechanism changed, so the 3.9-point gain is attributable to conditioning on images. This also rejects both alternative explanations from Section 3: CLIP scores attribute presence well enough to reach 19.40% over 1,000 classes with five bare words per class, and the GPT pool does contain useful attributes; label-conditioned generation simply fails to surface them. Widening the

pool to VAW+LSA raises accuracy further to 23.80%, a 53% relative improvement over the LLM descriptors.

The gain survives distribution shift. Attributes selected on ImageNet transfer unchanged to the four shifted variants and preserve the ordering everywhere, with the largest margin on ImageNet-R (28.20 vs 17.91). Absolute numbers remain far below class-name prompting, which is expected; a handful of attributes shared across many classes cannot fully separate 1,000 categories, and we state this boundary in Section 7. The top-5 budget is also deliberately harsh; it buys readability, not accuracy. Uncapping it raises attribute-only ImageNet accuracy to 45.5%, nearly three times the LLM-descriptor baseline (Section 5.3).

5.3 Uncapping the attribute budget

The headline attribute-only results above restrict each class to its top-5 attributes, the harshest and most readable setting. This section removes the cap: each class keeps its full ranked attribute list, selected from a growing number of images per class. We compare the two scoring variants of Eq. (2): unweighted ($\omega_{c,j} = 1$, every selected attribute votes equally) and weighted ($\omega_{c,j} = w_{c,j}$, a class's most characteristic attributes vote more).

Weights help consistently but modestly. Table 2 compares unweighted and weighted variants at 16 selection images per class. Weighting the selected attributes by their class-attribute scores adds roughly 0.3 points on every dataset, for both the LLM descriptor pool and ours. The pool matters far more than the weighting: the full VAW+LSA pool reaches 45.53% on ImageNet against 18.06% for the LLM descriptor pool under identical selection and weighting.

Attribute-only accuracy scales with selection images. Table 3 traces the same weighted and unweighted variants from 1 to 16 selection images per class. Accuracy rises monotonically on all five datasets (ImageNet: 23.17% at 1 shot to 45.53% at 16): with the budget uncapped, the selection keeps improving as more images refine the class-attribute scores. For calibration, the LLM descriptors, which use no images, sit at 15.50% on ImageNet; a single selection image per class already exceeds them under the full weighted pool.

Finding. Conditioning attribute selection on images, with the pool held fixed, lifts class-name-free ImageNet accuracy from 15.5% to 19.4%; a larger pool reaches 23.8% with five attributes per class and 45.5% with the full weighted pool. The selection mechanism, not the vocabulary, is what makes attributes informative.

5.4 Prompting with class names

With the class name restored, our selected attributes give a consistent but small edge, with a monotone dose response in k: 59.53, 60.18, and 60.80 for the top 5, 10, and 100 attributes, against 59.47 for LLM descriptors (Table 1, top block), and the ordering holds on all four shifted variants. We do not lean on this comparison: once the class name is present, much of any descriptor method's gain is prompt ensembling rather than attribute evidence [20], which is precisely

Pool	Selection	Weighted	ImageNet	-V2	-Sketch	-A	-R
LLM descriptors [13]	none (as released)	–	15.50	14.00	9.60	8.57	17.91
LLM descriptor pool	ours, full pool	no	16.67	13.15	6.44	6.44	23.14
		yes	18.06	14.45	7.08	6.68	24.16
VAW+LSA	ours, full pool	no	45.25	37.25	19.14	14.04	36.97
		yes	**45.53**	**37.50**	**19.41**	**14.43**	**37.27**

Table 2: With the attribute budget uncapped, image-conditioned selection triples attribute-only accuracy. Top-1 accuracy, attribute-only prompts (no class name), CLIP-RN50; selection uses 16 images per class. Weighted = each attribute's vote scaled by its class-attribute score (Eq. (2) with $\omega_{c,j} = w_{c,j}$); it adds about 0.3 points everywhere, while switching from the LLM descriptor pool to VAW+LSA under identical selection adds over 27 points on ImageNet. **Bold** marks best result per column.

the confound this paper is about. The class-name-free block carries our claim; the top block shows that image-conditioned selection at least matches label-conditioned generation on the standard protocol too.

5.5 Few-shot classification

One image per class is enough to select good attributes. We select each class's top attributes from only k labeled images and evaluate zero-shot with the selected attribute prompts. At 1 shot this reaches 60.13% against 57.15% for CoOp [27], with the margin persisting at 2 and 4 shots (Table 4), and the fit takes under a minute against CoOp's 14 hours of prompt tuning. Unlike a tuned soft prompt, the fitted object is a list of words per class, so the resulting classifier can be read and audited; the supplementary material discusses editing it by intervention.

Why selection wins the low-shot regime, and where it stops. Selection only has to rank a fixed pool of semantically meaningful candidates, a far smaller hypothesis space than a continuous prompt, so a single image carries enough signal; the attribute features' shared semantic structure (Section 5.6) does the rest. The same property caps the method: the probe's only job is to rank the pool, so extra images refine the ranking but add no capacity, and accuracy plateaus near 61.4% while CoOp and WiSE-FT keep improving and pass us at 8 shots. The method's regime is the extreme low-shot end, and Table 4 shows both sides of the boundary.

Finding. With one image per class, image-conditioned attribute selection beats CoOp by 3 points (60.1 vs 57.2) at 10^{-3} of the fitting cost, while producing a classifier that is a readable list of words; beyond 4 shots, gradient methods pass it.

5.6 Structure of the attribute features

Figure 3 compares pairwise mutual information among CLIP embedding dimensions and among our attribute features on ImageNet. CLIP dimensions are close

Dataset	Weighted	Selection images per class				
		1	2	4	8	16
ImageNet	no	22.92	30.36	36.90	42.06	45.25
	yes	**23.17**	**30.64**	**37.20**	**42.60**	**45.53**
ImageNetV2	no	19.20	26.04	31.32	35.27	37.25
	yes	**19.54**	**26.20**	**31.55**	**35.43**	**37.50**
ImageNet-Sketch	no	10.28	13.19	15.83	17.85	19.14
	yes	**10.41**	**13.41**	**16.08**	**18.08**	**19.41**
ImageNet-A	no	9.93	12.13	12.08	13.38	14.04
	yes	**10.00**	**12.21**	**12.22**	**13.48**	**14.43**
ImageNet-R	no	25.53	29.63	32.00	35.55	36.97
	yes	**25.77**	**29.91**	**32.16**	**35.81**	**37.27**

Table 3: Attribute-only accuracy grows monotonically with selection images when the attribute budget is uncapped. Top-1 accuracy, attribute-only prompts (no class name), full VAW+LSA pool, CLIP-RN50. Weighted is defined as in Table 2; it adds a consistent margin of roughly 0.3 points at every shot count on every dataset. **Bold** marks the better of the two variants per column within each dataset.

to statistically independent; attribute features are visibly correlated, and the top-5 attributes of a class share more information with each other than with other classes' attributes (right panel). We read this honestly: the transformation is not statistical disentanglement, it is a change of basis into named directions whose correlations are semantic (an object that scores high on "ripe" also scores high on "red"). That named, semantically clustered structure is what the few-shot result exploits, and what every application in Section 6 depends on.

6 Applications of Data-Conditioned Attributes

The applications below share one mechanism: the attribute set is a function of an image collection, so any collection (a dataset, a class under shift, an image region) can be summarized in words. Label-conditioned descriptors cannot do this by construction, since they never see images.

6.1 Describing distribution shift in words

Subtracting the mean attribute profile of ImageNet from that of a shifted dataset ranks the attributes that became more and less present. On ImageNet-Sketch the top risers are "colorless", "gray", "cartoon", and "digital" while color attributes fall; on ImageNet-R the risers are "cartoon", "painting", and "tattooed", matching that dataset's stated composition of art, cartoons, and tattoo renditions [7] (Figure 4). We use only unlabeled test images, so no class-specific information leaks. Beyond

Method	Number of shots					Fit time
	1	2	4	8	16	
CLIP linear probe	24.46	35.09	44.60	52.31	57.10	<1 min
CoOp [27]	57.15	57.81	59.99	61.56	**62.95**	14 hr
WiSE-FT [25]	58.30	59.08	60.48	**61.85**	62.84	<1 min
Ours	**60.13**	**60.92**	**61.13**	61.39	61.35	<1 min

Table 4: Attribute selection is a strong extreme-few-shot learner. ImageNet top-1 accuracy with CLIP-RN50; zero-shot CLIP scores 55.31. We linear-probe the attribute features of the k-shot images to select top-k attributes per class, then evaluate zero-shot with the selected attributes; baselines fine-tune or prompt-tune on the same shots. **Bold** marks the best method per column. Our accuracy plateaus beyond 4 shots because the probe only ranks a fixed pool of words; the crossover is stated, not hidden.

inspection, such profiles give a quantitative, interpretable measure of domain gap, relevant to domain adaptation prompting [4] and to monitoring shift in continual settings [12].

Domain attributes as prompt templates. The description is also actionable. We take the top-5 rising attributes of each shifted dataset, computed from unlabeled images only, and build the prompt template "A {attribute} photo of a {class}", ensembled over the five attributes; on ImageNet-R the templates become "A cartoon photo of ...", "A painting photo of ...", and so on. Table 5 evaluates this template against the plain "{class}" and "A photo of a {class}" templates, crossed with our class-specific attribute prompts. The domain template improves plain zero-shot on every dataset, by up to 3.7 points on ImageNet-Sketch, and still adds up to 1.5 points on top of our strongest top-100 attribute prompts. Domain-level and class-level attributes are therefore complementary: one describes how the images look, the other what they contain.

6.2 Case study: color shift breaks CLIP zero-shot

Motivation. CLIP's zero-shot predictions can latch onto color rather than object identity: presented with a yellow eggplant, CLIP drifts toward other yellow concepts such as lemon (Figure 6). To measure this failure mode in isolation we collected a color-variant test set for fruit classification.

Dataset. The source is a public fruit dataset [21] of 37 classes and 4,440 images (100 train, 10 validation, and 10 test images per class). We asked ChatGPT to group the 37 class names by color, yielding 9 colors, formed all 333 color-class phrases ("green apple", "purple lemon", ...), crawled 20 web images per phrase, and manually removed irrelevant results. Merging the color variants

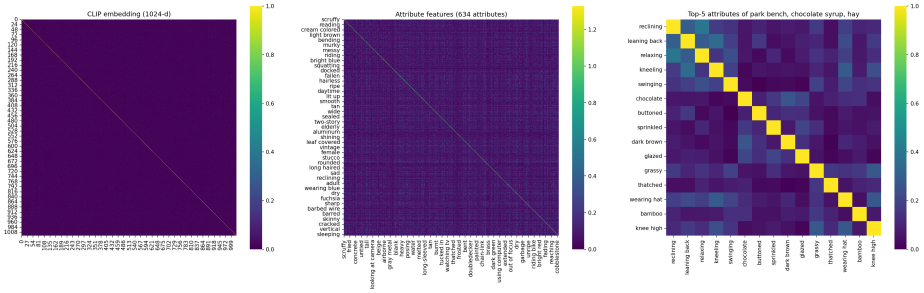

Fig. 3: Attribute features trade statistical independence for named, semantically clustered directions. Pairwise mutual information on ImageNet with CLIP-RN50; all three panels use the sequential viridis scale, dark purple at zero to bright yellow at high mutual information. Left: dimensions of the 1024-d CLIP embedding, near zero off the diagonal. Middle: our 634 VAW attribute features, visibly more correlated. Right: features restricted to the top-5 attributes of three classes (park bench, chocolate syrup, hay); the brighter 5×5 diagonal blocks show that a class's own attributes share information, i.e. the correlations follow semantics.

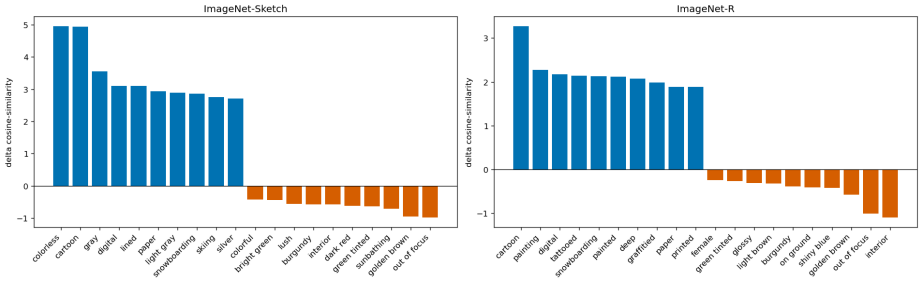

Fig. 4: Distribution shift, described in words. Top-10 rising and falling attributes (change in mean attribute score over unlabeled images, y-axis) of ImageNet-Sketch (left panel) and ImageNet-R (right panel) relative to ImageNet. Sketch gains "colorless", "gray", and "cartoon" and loses color attributes; R gains "cartoon", "painting", and "tattooed", matching its documented composition. Within each panel attributes are ordered by their change, risers first. Blue bars rising above zero become more present under the shift; vermillion bars falling below zero become less present.

by source class gives an additional color-variant test set of the same 37 classes (6,050 images overall; examples in Figure 5).

Result. Zero-shot CLIP drops from 81.02% on the standard test set to 51.25% on the color variants, a 30-point fall from color shift alone (Table 6). Our attribute prompts improve both settings, to 85.63% and 55.09% respectively. The gain is real but the drop is not repaired; robustly classifying counter-stereotypical colors remains open, consistent with our position that attributes complement rather than replace class-name prompting (Section 7).

Dataset	Class-attribute prompts	Prompt template		
		{class}	A photo of a {class}.	A {attribute} photo of a {class}.
ImageNetV2	none	49.39	51.34	**52.10**
	Ours (top 5)	**53.33**	52.74	52.97
	Ours (top 10)	53.40	**53.62**	53.59
	Ours (top 100)	53.95	**54.06**	53.78
ImageNet-Sketch	none	31.57	33.30	**35.31**
	Ours (top 5)	33.12	33.65	**35.28**
	Ours (top 10)	33.65	34.04	**35.75**
	Ours (top 100)	34.27	34.55	**36.04**
ImageNet-A	none	20.92	21.61	**22.58**
	Ours (top 5)	22.78	22.86	**23.65**
	Ours (top 10)	23.43	22.45	**23.76**
	Ours (top 100)	**24.00**	23.16	23.78
ImageNet-R	none	58.18	56.20	**59.29**
	Ours (top 5)	56.29	55.46	**58.63**
	Ours (top 10)	57.49	56.30	**59.66**
	Ours (top 100)	59.05	57.69	**60.44**

Table 5: Domain attributes discovered from unlabeled images make useful prompt templates. Top-1 accuracy, CLIP-RN50. Rows: class-specific attribute prompts (none = plain class name). Columns: prompt templates; the rightmost ensembles the top-5 rising domain attributes of each dataset ("A cartoon photo of a {class}" on ImageNet-R). The domain template wins in most settings, by up to 3.7 points over the plain class name on ImageNet-Sketch, with ImageNet-A top-100 and two ImageNetV2 rows as exceptions. **Bold** marks the best template per row.

Evaluation	Test	Color variant
CLIP zero-shot	81.02	51.25
Ours zero-shot	**85.63**	**55.09**

Table 6: Color shift costs zero-shot CLIP 30 points; attribute prompts recover 4. Top-1 accuracy on the fruit dataset's standard test set and our color-variant test set, CLIP-RN50. **Bold** marks the better method per column.

6.3 Vocabulary the text encoder cannot parse

Some class vocabularies defeat prompt-based classification outright: on iNaturalist [23], whose classes are scientific names such as *Cristidiscoidea*, zero-shot CLIP reaches 3% accuracy because the text encoder has no grounding for the names. Selecting attributes from a few images per class replaces the unparseable name with a set of visual words, raising accuracy to 7.25% with a general-purpose pool of roughly 600 largely human-centric attributes [15]. The absolute number stays low and the pool is poorly matched to fine-grained species; the point is the mechanism, which needs no expert to translate jargon into CLIP's vocabulary.

6.4 Attribute-guided localization

Removing CLIP-ResNet's attention pooling preserves the spatial grid, so attribute similarities can be computed per patch. Querying the patch grid with a

Fig. 5: Examples from the collected color-variant fruit test set. The set spans natural rarities (yellow eggplant), style shifts (sketches and drawings), and artificial colorings (blue and purple lemons), over the 37 classes of the source dataset [21].

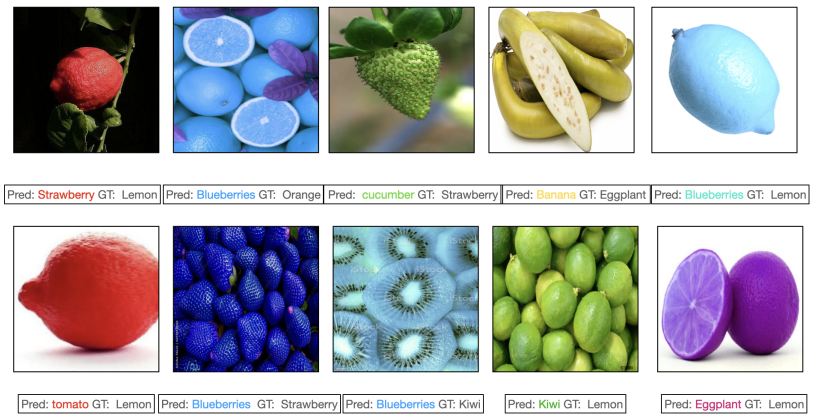

Fig. 6: CLIP zero-shot predictions latch onto color. On color-variant fruit images, zero-shot CLIP often predicts a class that matches the color rather than the object, e.g. a yellow eggplant drifts toward lemon.

class's selected attributes, with no class name in the query, highlights the image regions supporting those attributes and localizes the class (Figure 7). One forward pass yields the heatmaps for any number of attribute queries, and the same mechanism describes individual regions in words, which we leave as qualitative evidence here.

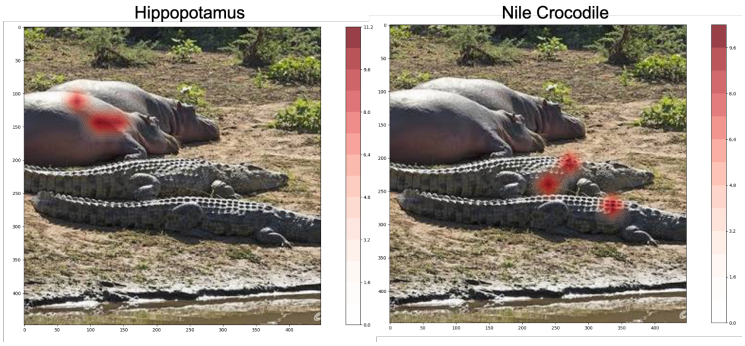

Fig. 7: Class-name-free localization from selected attributes. We remove the attention pooling of CLIP RN50x64, compute cosine similarity between each 14×14 patch embedding and a class's selected attribute texts, and display the resulting heatmap. Querying with the hippopotamus attribute set (left) and the Nile crocodile attribute set (right) localizes each animal; the queries contain no class names. Warmer color denotes higher mean similarity to the attribute set.

7 Limitations

Attribute-only accuracy remains far below class-name prompting (23.80 vs 60.80 on ImageNet with five attributes per class), so when a parseable class name exists, attributes complement it rather than replace it. In few-shot classification the method plateaus at 61.4% while CoOp reaches 62.95% at 16 shots; selection has no capacity to absorb data beyond ranking the pool, and we do not claim the medium-shot regime. The attribute features are more statistically correlated than the CLIP embedding they re-express (Figure 3), which bounds how independently individual attribute scores can be read. Our quantitative results use a single backbone (CLIP-RN50) and a selection pool inherited from general-purpose attribute datasets; the iNaturalist experiment (7.25% absolute) shows what a mismatched pool costs on fine-grained domains.

8 Conclusion

Descriptor methods promise interpretable zero-shot classification, but their evidence rides the class name: removed from the prompt, LLM-generated descriptors collapse from 59.5% to 15.5% on ImageNet. Conditioning attribute selection on the target images repairs this at essentially zero training cost, and a matched-pool comparison shows the selection mechanism itself accounts for the repair. The resulting attribute sets are accurate enough to beat prompt tuning in the extreme few-shot regime and, because they are functions of image collections, they describe datasets, shifts, unparseable vocabularies, and image regions in words. The finding suggests a protocol change beyond our method: any work that claims attribute-level interpretability from descriptor prompts should report class-name-free accuracy, since the standard protocol cannot distinguish attribute evidence from prompt ensembling. For concept-bottleneck pipelines, it argues that concept pools should be selected against the deployment distribution rather than generated from labels.

References

1. Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C.: Label-embedding for attribute-based classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 819–826 (2013) 4
2. Bhalla, U., Oesterling, A., Srinivas, S., Calmon, F.P., Lakkaraju, H.: Interpreting CLIP with Sparse Linear Concept Embeddings (SpLiCE). Advances in Neural Information Processing Systems **37** (2 2024). https://doi.org/10.52202/079017-2678, https://arxiv.org/pdf/2402.10376 4
3. Chen, C., Zhang, B., Cao, L., Shen, J., Gunter, T., Jose, A., Toshev, A., Zheng, Y., Shlens, J., Pang, R., Yang, Y.: STAIR: Learning sparse text and image representation in grounded tokens. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 15079–15094 (2023) 5
4. Dunlap, L., Mohri, C., Guillory, D., Zhang, H., Darrell, T., Gonzalez, J.E., Raghunathan, A., Rohrbach, A.: Using language to extend to unseen domains. In: International Conference on Learning Representations (ICLR) (2023) 10
5. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1778–1785. IEEE (2009) 4
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. vol. 2016-December, pp. 770–778. IEEE Computer Society (12 2016). https://doi.org/10.1109/CVPR.2016.90, http://image-net.org/challenges/LSVRC/2015/ 6
7. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., Gilmer, J.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8340–8349 (2021) 6, 9
8. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15262–15271 (2021) 6
9. Jayaraman, D., Grauman, K.: Zero Shot Recognition with Unreliable Attributes. Advances in Neural Information Processing Systems **4**(January), 3464–3472 (9 2014), https://arxiv.org/pdf/1409.4327 4
10. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: Proceedings of the 37th International Conference on Machine Learning (ICML). vol. 119. PMLR (2020) 4, 19
11. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 951–958. IEEE (2009) 4
12. Lin, Z., Shi, J., Pathak, D., Ramanan, D.: The CLEAR benchmark: Continual learning on real-world imagery. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021) 10
13. Menon, S., Vondrick, C.: Visual classification via description from large language models. In: International Conference on Learning Representations (ICLR) (2023) 1, 2, 3, 4, 5, 6, 8
14. Oikarinen, T., Das, S., Nguyen, L.M., Weng, T.W.: Label-Free Concept Bottleneck Models. 11th International Conference on Learning Representations, ICLR 2023 (4 2023), https://arxiv.org/pdf/2304.06129 4

15. Pham, K., Kafle, K., Lin, Z., Ding, Z., Cohen, S., Tran, Q., Shrivastava, A.: Learning to Predict Visual Attributes in the Wild. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition pp. 13013–13023 (6 2021). https://doi.org/10.48550/arxiv.2106.09707, https://arxiv.org/abs/2106.09707v1 4, 5, 6, 12
16. Pham, K., Kafle, K., Lin, Z., Ding, Z., Cohen, S., Tran, Q., Shrivastava, A.: Improving Closed and Open-Vocabulary Attribute Prediction Using Transformers. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **13685 LNCS**, 201–219 (2022). https://doi.org/10.1007/978-3-031-19806-9{_}12/TABLES/6, https://link.springer.com/chapter/10.1007/978-3-031-19806-9_12 4, 5, 6
17. Pratt, S., Covert, I., Liu, R., Farhadi, A.: What does a platypus look like? Generating customized prompts for zero-shot image classification. Proceedings of the IEEE International Conference on Computer Vision pp. 15645–15655 (9 2022). https://doi.org/10.1109/ICCV51070.2023.01438, https://arxiv.org/pdf/2209.03320 1, 3
18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML). vol. 139, pp. 8748–8763. PMLR (2021) 1, 5, 6
19. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do ImageNet Classifiers Generalize to ImageNet? 36th International Conference on Machine Learning, ICML 2019 **2019-June**, 9413–9424 (2 2019), https://arxiv.org/pdf/1902.10811 6
20. Roth, K., Kim, J.M., Sophia Koepke, A., Vinyals, O., Schmid, C., Akata, Z.: Waffling around for Performance: Visual Classification with Random Words and Broad Concepts. Proceedings of the IEEE International Conference on Computer Vision pp. 15700–15711 (6 2023). https://doi.org/10.1109/ICCV51070.2023.01443, https://arxiv.org/pdf/2306.07282 4, 7
21. Seth, K.: Fruits and vegetables image recognition dataset. https://www.kaggle.com/datasets/kritikseth/fruit-and-vegetable-image-recognition (2020), kaggle 10, 13
22. Shen, S., Li, C., Hu, X., Yang, J., Xie, Y., Zhang, P., Gan, Z., Wang, L., Yuan, L., Liu, C., Keutzer, K., Darrell, T., Rohrbach, A., Gao, J.: K-LITE: Learning Transferable Visual Models with External Knowledge. Advances in Neural Information Processing Systems **35** (4 2022). https://doi.org/10.52202/068431-1132, https://arxiv.org/pdf/2204.09222 2, 3
23. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8769–8778 (2018) 12
24. Wang, H., Ge, S., Xing, E.P., Lipton, Z.C.: Learning Robust Global Representations by Penalizing Local Predictive Power. Advances in Neural Information Processing Systems **32** (5 2019), https://arxiv.org/pdf/1905.13549 6
25. Wortsman, M., Ilharco, G., Kim, J.W., Li, M., Kornblith, S., Roelofs, R., Lopes, R.G., Hajishirzi, H., Farhadi, A., Namkoong, H., Schmidt, L.: Robust fine-tuning of zero-shot models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7959–7971 (2022) 4, 10

26. Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., Yatskar, M.: 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 (CVPR). pp. 19187–19197 (2023) 4
27. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022) 4, 8, 10

Supplementary Material

This supplementary material provides the experimental details referenced from the main paper (Section A), a linear-probing experiment measuring how much of CLIP's representation power the attribute basis preserves (Section B), and a discussion of editability and attribute-pool relevancy (Section C).

A Experimental Setup Details

Models. All quantitative results use CLIP with the ResNet-50 backbone; the localization figure uses CLIP RN50x64 because its larger 14×14 penultimate grid gives a finer heatmap. Both encoders stay frozen in every experiment.

Attribute pool and preprocessing. The candidate pool combines the attribute vocabularies of VAW (634 attributes after processing) and LSA, plus the GPT-3 descriptors released by Menon and Vondrick when the protocol calls for their pool. Preprocessing removes strings that contain explicit class names, exact duplicates, and grammatical punctuation. Each attribute is encoded bare, with no prompt template and no class name, for the reason given in the main paper: a shared template or class name would correlate all of a class's attribute scores.

Scoring and pairing. Attribute features are scaled cosine similarities with temperature $\tau = 100$. Class-attribute weights come from a linear probe trained on the attribute features (classification experiments; the probe's per-class weight rows plus bias rank the pool) or from the per-class mean attribute feature (distribution-level applications, which need no labels or training). Linear probes are trained with AdamW (learning rate 10^{-4}, weight decay 0.01) on a single RTX 3090 GPU; fitting completes in under a minute for every setting in the paper.

Evaluation protocols. The class-name protocol follows Menon and Vondrick exactly, including their grammatical modifier: the prompt is "{class name}, which is/has/{...} {attribute}" chosen by the attribute's leading word. The attribute-only protocol reduces the prompt to the bare attribute string. A test image is scored by the mean similarity to each class's selected attribute prompts, and top-1 accuracy is reported everywhere.

B Representation Power of the Attribute Basis

The main paper's mutual-information analysis shows the attribute features are semantically structured; this section measures how much of CLIP's discriminative power the change of basis preserves. We re-express each image as its attribute-similarity vector, keeping the top-1024 attributes to match the dimensionality of the CLIP-RN50 embedding, and linear-probe both representations under identical settings on full ImageNet.

Table 7 reports three attribute pools: the VAW vocabulary, the joint (VAW + LSA + LLM) pool, and a control pool of random strings. The joint pool lands

Representation	ImageNet	-V2	-Sketch	-A	-R
Attribute basis, random-string pool	65.37	54.16	24.26	13.51	37.26
Attribute basis, VAW pool	69.83	58.19	27.65	13.88	42.61
Attribute basis, joint VAW+LSA+LLM pool	71.05	59.92	30.36	16.09	46.61
CLIP embedding (upper bound)	**73.10**	**61.58**	**33.04**	**18.69**	**52.15**

Table 7: The attribute basis preserves most of CLIP's linear separability, and semantic pools beat random strings. Top-1 accuracy of a linear probe trained on ImageNet, evaluated on ImageNet and its shifted variants; probes on attribute-similarity features (top-1024 attributes, matching the 1024-d CLIP-RN50 embedding) vs the raw CLIP embedding. The joint pool sits 2.0 points below the CLIP upper bound on ImageNet; the random-string control trails the joint pool by 5.7 points, so vocabulary semantics matter for the representation. **Bold** marks the best result per column.

within 2.0 points of the raw CLIP embedding on ImageNet (71.05 vs 73.10), so the interpretable basis costs little linear separability. The pool ordering (random < VAW < joint) shows that the vocabulary's semantics and diversity both matter for representation quality: random strings retain non-trivial accuracy, as arbitrary projections of a strong embedding do, but trail the joint pool by 5.7 points on ImageNet and by more under shift.

C Discussion

Editability and intervention. Because a class's representation is a ranked list of named attributes, the classifier admits test-time intervention in the spirit of concept bottleneck models [10]: a practitioner can inspect the word list, delete or reweight an attribute that should not matter, or substitute one that should (replacing "yellow" with "purple" to describe an unusually colored lemon). We report this as a property of the representation, not a validated capability: manual swap edits in early experiments gave small gains that did not close the color-shift gap of the main paper's fruit case study, and systematic intervention protocols in the style of CBM interventions remain future work.

Attribute-pool relevancy. The interpretability and accuracy of our method hinge on the relevancy of the attribute pool. Fine-grained domains need fine-grained vocabulary: the main paper's iNaturalist experiment improves zero-shot accuracy from 3% to 7.25% with a largely human-centric pool of roughly 600 attributes, and a species-focused vocabulary should close more of the remaining gap. The failure mode is mild because selection is cheap: a mismatched pool costs accuracy but never requires retraining, and enlarging the pool toward a universal vocabulary (in the limit, an English dictionary) lets the data mine whatever is relevant. We leave the universal-pool experiment to future work.