Concept Regions Matter: Benchmarking CLIP with a New Cluster-Importance Approach

Aishwarya Agarwal[1,2*] *Srikrishna Karanam*[2†] *Vineet Gandhi*[1‡]

[1]CVIT, Kohli Centre for Intelligent Systems, IIIT Hyderabad, India

[2]Adobe Research, Bengaluru, India

Abstract

Contrastive vision–language models (VLMs) such as CLIP achieve strong zero-shot recognition yet remain vulnerable to spurious correlations, particularly background over-reliance. We introduce Cluster-based Concept Importance (CCI), a novel interpretability method that uses CLIP's own patch embeddings to group spatial patches into semantically coherent clusters, masking them, and evaluating relative changes in model predictions. CCI sets a new state of the art on faithfulness benchmarks, surpassing prior methods by large margins; for example, it yields more than a twofold improvement on the deletion-AUC metric for MS COCO retrieval. We further propose that CCI when combined with GroundedSAM, automatically categorizes predictions as foreground or background-driven, providing a crucial diagnostic ability. Existing benchmarks such as CounterAnimals, however, rely solely on accuracy and implicitly attribute all performance degradation to background correlations. Our analysis shows this assumption to be incomplete, since many errors arise from viewpoint variation, scale shifts, and fine-grained object confusions. To disentangle these effects, we introduce COVAR, a benchmark that systematically varies object foregrounds and backgrounds. Leveraging CCI with COVAR, we present a comprehensive evaluation of eighteen CLIP variants, offering methodological advances and empirical evidence that chart a path toward more robust VLMs.

1. Introduction

Contrastive vision–language models (VLMs) such as CLIP [10] demonstrate strong generalization in zero-shot recognition, retrieval, and open-vocabulary settings. Despite this success, they remain vulnerable to spurious correlations [16, 18, 19], especially object–background shortcuts learned from large-scale web data. Such shortcuts degrade reliability under distribution shifts and can obscure whether predictions are truly object-centric.

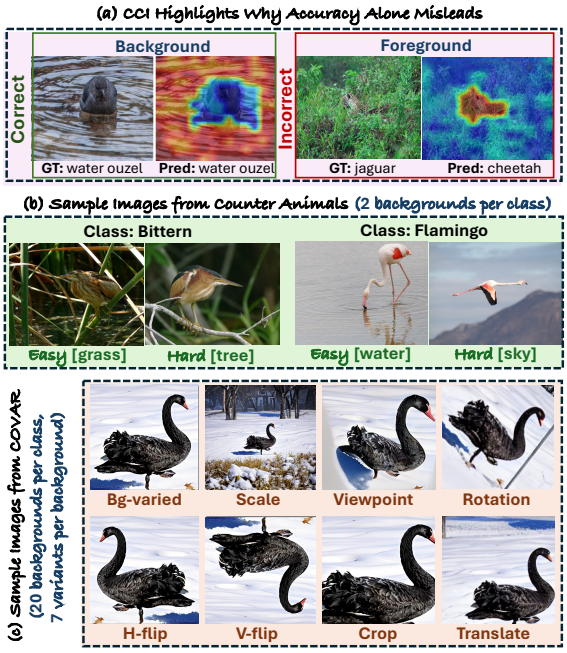

Figure 1. Overview: (a) CCI highlights decision-relevant concept regions, (b) accuracy-only dataset splits can conflate failure causes, and (c) COVAR provides controlled variants for diagnosis.

Recent analyses suggest that background context is a dominant source of these shortcuts, and benchmarks such as CounterAnimals (CA) [16] attempt to quantify this effect by partitioning images into easy and hard sets using accuracy. However, this operationalization is ambiguous: a drop in accuracy does not necessarily indicate background reliance, and correct predictions can still be background-driven. In practice, errors also arise from viewpoint changes, scale mismatches, occlusion, and fine-grained class similarity.

To diagnose these factors explicitly, we propose a unified

*aishwarya.agarwal@research.iiit.ac.in, aishagar@adobe.com
†skaranam@adobe.com
‡vgandhi@iiit.ac.in

approach combining interpretability and controlled evaluation. First, we introduce **Concept Cluster Importance (CCI)**, a training-free attribution method that uses CLIP patch embeddings to form semantically coherent concept regions and measures their contribution through masking-based similarity drops. Second, we introduce **COVAR**, a benchmark with systematic variants that independently manipulate background and geometric factors, enabling fine-grained robustness analysis.

We evaluate CCI against strong gradient-, attention-, and perturbation-based baselines [1, 2, 9, 14, 17, 21, 22], and show consistent gains on faithfulness metrics. Using CCI together with GroundedSAM [11] masks, we categorize errors as foreground-driven or background-driven and find that accuracy-only categorization substantially over-attributes failures to background. Across 18 CLIP variants on COVAR, scale and viewpoint remain major bottlenecks even when absolute accuracy improves, indicating that robust VLM behavior requires more than larger backbones.

In summary, this paper contributes: (i) a concept-level, training-free interpretability method for CLIP; (ii) an attribution-aware robustness benchmark with controlled perturbations; and (iii) a comprehensive empirical study disentangling background reliance from other confounders in modern CLIP variants.

2. Concept Cluster Importance

We present **Concept Cluster Importance (CCI)**, a training-free interpretability method for CLIP models that quantifies the contribution of semantically coherent visual concepts to image–text similarity scores. CCI operates entirely at inference time, requiring no model modification or retraining.

Patch Embedding Clustering. We focus on the patch embeddings $\mathbf{X} = \{\mathbf{z}_i\}_{i=1}^{N}$, which encode localized semantics. To extract coherent visual concepts, we perform K-means clustering over $\mathbf{X}$, yielding $\mathcal{C} = \{C_1, \ldots, C_K\}$, where each cluster aggregates semantically similar patches (See Figure 2 for examples with $K = 7$, where distinct colors denote different clusters).

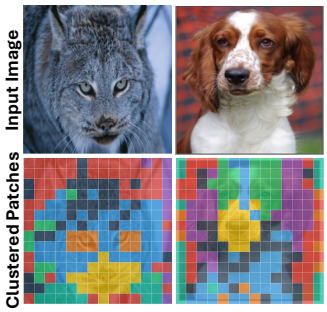

Figure 2. Patch clusters.

Preliminaries. Given an input image I, the CLIP image encoder (e.g., a Vision Transformer) processes I into a sequence of token embeddings $\mathbf{Z} = [\mathbf{z}_{\text{CLS}}, \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N] \in \mathbb{R}^{(N+1)\times d}$ where $\mathbf{z}_{\text{CLS}} \in \mathbb{R}^d$ is the global `[CLS]` embedding used for final image representation, and $\{\mathbf{z}_i\}_{i=1}^{N}$ are patch embeddings corresponding to N fixed-size patches extracted from the image.

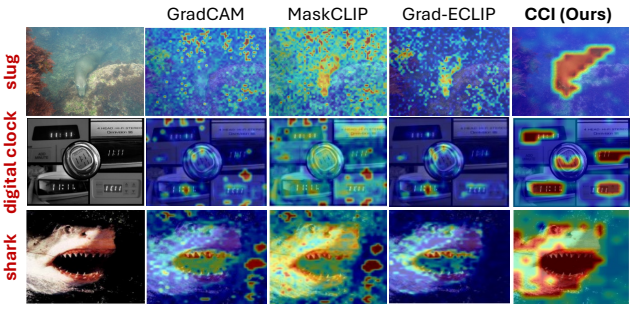

Figure 3. Qualitative comparison of CCI against baselines.

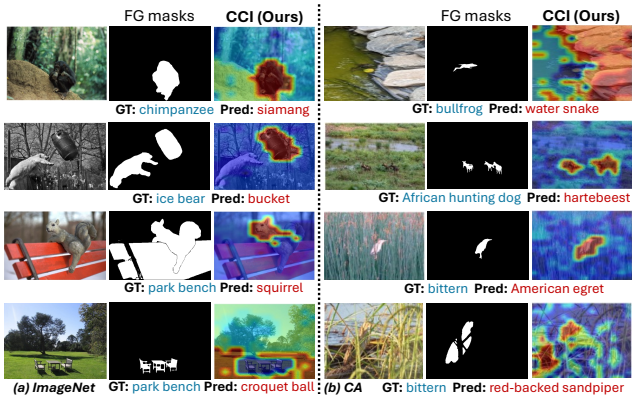

Figure 4. CCI analysis of CLIP failures on (a) ImageNet and (b) Counter Animals. Rows show the image, ground-truth foreground (FG) mask obtained using GroundedSAM, and attribution maps from CCI, with ground-truth (GT) and predicted (Pred) labels.

Attention Attenuation with Cluster Masks. The CLIP image encoder contains L transformer layers, each with self-attention matrices $A^{(l)} \in \mathbb{R}^{(N+1)\times(N+1)}$, where the first token corresponds to the CLS embedding and the rest correspond to patches.

To measure the contribution of cluster C_k, we construct a binary mask $m_k(j)$ that indicates whether patch j belongs to cluster C_k. Let $m_k(j) = \mathbf{1}_{\{j \in C_k\}}$. We modify logits as $\hat{A}_k^{(l)}(i,j) = A^{(l)}(i,j) - m_k(j) \cdot \infty$, applied at every layer and head to block CLS attention to C_k.

Importance Scoring via Similarity Drop. Let $\mathbf{z}_{\text{CLS}}$ be the original final CLS embedding, and $\mathbf{t}$ the corresponding text embedding. The original similarity score is $s = \cos(\mathbf{z}_{\text{CLS}}, \mathbf{t}) = \frac{\mathbf{z}_{\text{CLS}}^\top \mathbf{t}}{\|\mathbf{z}_{\text{CLS}}\|\|\mathbf{t}\|}$. After attention attenuation for cluster k, we obtain a modified CLS embedding $\hat{\mathbf{z}}_{\text{CLS},k}$ and similarity $s_k = \cos(\hat{\mathbf{z}}_{\text{CLS},k}, \mathbf{t})$.

The relative importance of cluster C_k is quantified by the similarity drop $\Delta s_k = s - s_k$. We normalize the importance scores as $w_k = \frac{\Delta s_k}{\sum_{j=1}^{K} \Delta s_j}$, and compute the spatial impor-

tance map $S \in \mathbb{R}^{W \times H}$ as a weighted sum of cluster masks $S = \sum_{k=1}^{K} w_k \cdot m_k$. The map S identifies regions contributing most to the similarity score and is visualized as a heatmap overlay on I.

2.1. CCI Results

We evaluate CCI against a diverse set of baselines spanning gradient-based, attention-based, and perturbation-based techniques: Attention Rollout [1], GradCAM [14], GAME [2], MaskCLIP [22], M2iB [17], RISE [9], and Grad-ECLIP [21]. Unless otherwise specified, all experiments use CLIP with a `ViT-B/16` image encoder, and maps are computed with respect to the ground-truth label.

Qualitative Comparison with Baselines. Figure 3 compares attention maps from CCI and baseline methods. CCI consistently produces coherent, object-aligned heatmaps, whereas baselines yield sparse or noisy patterns. For instance, in the *slug* image (row 1), CCI captures the entire object while baselines highlight scattered regions; in the *digital clock* (row 2), CCI sharply localizes the digits on the clock face, relevant to CLIP's prediction, unlike baselines that miss or misfocus; and in the *shark* example (row 3), CCI emphasizes the teeth-features, whereas baselines diffuse attention across irrelevant regions. Additional examples are provided in supplementary.

Quantitative Comparison. Consistent with prior work, we evaluate the faithfulness of CCI using deletion and insertion metrics [12]. CCI produces a patch-level importance map (14×14 for `ViT-B/16`), which is upsampled to 224×224 to assign pixel-level scores. Pixels are ranked by importance; in *deletion*, top-ranked pixels are iteratively replaced with random noise, while in *insertion*, they are progressively revealed from a blank canvas. At each step, ∼0.5% of pixels are modified, over 100 steps, cumulatively altering about half the image. The model's top-1 and top-5 accuracy is tracked at every step, and the area under the resulting curves (AUC) is used as a summary measure: lower AUC for deletion and higher AUC for insertion indicate causal influence of the highlighted regions. We report these scores on ImageNet-1K classification [4]) and MS COCO cross-modal retrieval [8].

Across both datasets, CCI consistently outperforms all baselines. This is reflected in the AUC scores (Table 1), where CCI achieves state-of-the-art performance across Top-1 and Top-5 metrics. The same holds for image–text retrieval (Table 2), where CCI delivers state-of-the-art results for both image and text retrieval on COCO. Notably, in deletion, CCI attains over two-fold error reduction (0.2670 → 0.1056) in Top-5 IR, over the second-best Grad-ECLIP method. Collectively, these results show that CCI yields faithful, generalizable attributions of CLIP's decisions.

Understanding CLIP's Failure Modes: We use CCI to analyze CLIP's zero-shot predictions by visualizing attention maps with respect to the predicted class (Figure 4, ImageNet-1k in part a and CA in part b), focusing on misclassifications to expose error sources. Several failures are *foreground-driven*: in part a, row 1, a chimpanzee is misclassified as a siamang despite correct focus on the face; in row 2, attention is misdirected to a bucket, leading to error; and in row 3, CLIP fixates on a squirrel rather than the target class. In CA (part b, rows 2–3), CCI likewise reveals attention on the foreground, but errors arise from clutter, occlusion, or subtle visual distinctions.

Other errors are *background-driven*: in part a, row 4, attention to the grassy field results in *croquet ball*, while in part b, row 1, focus on background produces *water snake*. Even under occlusion or partial visibility (part b, rows 2 and 4), CCI consistently highlights the object. Overall, CCI reliably predicts CLIP's attention and offering clear interpretable insights.

3. COVAR: A new benchmark

As discussed in Section 1, the CA benchmark is limited, offering only coarse background variation and no control over viewpoint, scale, flip, or crop. To address these gaps, we introduce COVAR, where each object is placed into multiple backgrounds and, for every such instance, its appearance is systematically varied along several visual factors (horizontal/vertical flip, translation, crop, rotation, viewpoint, and scale). COVAR spans 33 ImageNet classes, with 50 seed images per class, 20 background variants per image, and 11 additional structured transformations. Altogether, this yields 396,000 images for systematic robustness evaluation.

We benchmark CLIP variants from OpenCLIP [7] and OpenAI's original models [10] on COVAR, varying backbone size, patch resolution, and pretraining data (e.g., DataComp [6], LAION [13], DFN [5], WebLI [3]). We also evaluate SigLIP variants [15, 20]. For each model and subset, we report: (i) top-1 accuracy, (ii) *BG-Er* (fraction of background-driven errors from CCI+GroundedSAM), and (iii) *Fine-Er* (fraction of fine-grained confusions among foreground-driven errors). Table 3 present classification accuracy along with *BG-Er* and *Fine-Er*.

Overall performance: Among all models, `ViT-H-qgelu(DFN-5B)` at `378px` achieves the highest average accuracy of 56.8% (Table 3). `ViT-B-SigLIP2` performs strongly at `384px` and `512px`, while `ViT-SO-SigLIP2` maintains competitive accuracy even at `224px`, underscoring its training efficiency. Larger models such as `ViT-bigG` also show reasonable performance.

Performance across eight subsets: On Bg-varied, most models retain accuracies above 55%. Crop does not hurt performance and often helps, while H-flip and Translation cause only minor declines (e.g., `ViT-bigG` remains above 62%). Rotation leads to modest drops, V-flip somewhat

Table 1. Faithfulness evaluation of image explanations on *ImageNet* validation: AUC of Deletion/Insertion curves using Top-1 (@1) or Top-5 (@5) accuracy, with either ground-truth or predicted labels as CLIP text input.

Method	Deletion ↓				Insertion ↑			
	Ground-truth		Prediction		Ground-truth		Prediction	
	@1	@5	@1	@5	@1	@5	@1	@5
raw attention	0.3831	0.6239	-	-	0.2492	0.4195	-	-
Rollout	0.4082	0.6556	-	-	0.2803	0.4665	-	-
Grad-CAM	0.3417	0.5628	0.3518	0.5817	0.2682	0.4454	0.2526	0.4206
GAME	0.3356	0.5734	0.3497	0.5938	0.3611	0.5636	0.3425	0.5384
MaskCLIP	0.2848	0.4885	0.2886	0.4957	0.3335	0.5351	0.3275	0.5267
CLIPSurgery	0.3115	0.5235	0.3217	0.5412	0.3832	0.6021	0.3727	0.5719
M2IB	0.3630	0.5953	0.3633	0.5951	0.3351	0.5411	0.3347	0.5410
Grad-ECLIP w/o λ_i	0.2535	0.4379	0.2634	0.4568	0.3715	0.5831	0.3528	0.5556
Grad-ECLIP	0.2464	0.4272	0.2543	0.4420	0.3838	0.5993	0.3672	0.5749
CCI (Ours)	**0.1809**	**0.3276**	**0.1789**	**0.3318**	**0.4175**	**0.6518**	**0.3893**	**0.6201**

Table 2. Evaluation of **image** explanation faithfulness on *MS COCO* image-text retrieval (*Karpathy's split*) val-set: AUC for Deletion and Insertion curves for performance on image retrieval (IR) and text retrieval (TR) tasks.

Method	Deletion↓				Insertion↑			
	IR		TR		IR		TR	
	@1	@5	@1	@5	@1	@5	@1	@5
raw attention	0.1708	0.3554	0.1923	0.3720	0.1247	0.2552	0.1544	0.2969
Rollout	0.1948	0.3946	0.2268	0.4238	0.1294	0.2932	0.1753	0.3503
Grad-CAM	0.1717	0.3502	0.2161	0.4008	0.1027	0.2216	0.1152	0.2327
GAME	0.1706	0.3552	0.1982	0.3800	0.1537	0.3083	0.2097	0.3735
MaskCLIP	0.1321	0.2841	0.1516	0.2949	0.1423	0.2953	0.1891	0.3514
CLIPSurgery	0.1794	0.3652	0.2381	0.4292	0.1419	0.2941	0.1771	0.3384
M2IB	0.1797	0.3671	0.2057	0.3905	0.1469	0.3004	0.2058	0.3691
Grad-ECLIP w/o λ_i	0.1390	0.2940	0.1827	0.3386	0.1403	0.2895	0.1735	0.3279
Grad-ECLIP	0.1246	0.2670	0.1550	0.2933	0.1576	0.3203	0.2056	0.3761
CCI (Ours)	**0.0650**	**0.1056**	**0.0677**	**0.1184**	**0.1812**	**0.3513**	**0.2224**	**0.3943**

Table 3. CLIP performance on COVAR: accuracy, background-driven error (*BG-Er*), and fine-grained confusion (*Fine-Er*) across subsets.

Model	Bg-varied			H-flip			Translate			Crop			V-flip			Rotation			Viewpoint			Scale			Avg Acc
	Acc	*BG-Er*	*Fine-Er*	Acc	*BG-Er*	*Fine-Er*	Acc	*BG-Er*	*Fine-Er*	Acc	*BG-Er*	*Fine-Er*	Acc	*BG-Er*	*Fine-Er*	Acc	*BG-Er*	*Fine-Er*	Acc	*BG-Er*	*Fine-Er*	Acc	*BG-Er*	*Fine-Er*	
ViT-B/32 (DataComp-1B)	55.8	23.6	40.9	55.2	21.2	44.6	55.8	18.9	46.2	59.4	18.5	47.7	32.2	22.6	33.3	44.6	21.3	39.5	28.5	22.9	29.5	24.2	50.7	16.1	44.5
ViT-B/16 (DataComp-1B)	55.7	14.2	49.3	54.3	11.7	52.7	53.7	11.3	53.4	56.5	11.9	52.2	37.5	13.8	42.6	44.9	12.4	50.4	27.7	15.7	33.5	27.2	30.5	28.4	44.7
ViT-L/14 (DataComp-1B)	62.2	15.7	51.8	61.3	13.5	56.0	60.7	12.4	56.6	62.3	12.2	56.6	48.3	15.2	50.2	54.9	14.5	53.9	30.3	17.1	36.0	32.6	30.2	31.2	51.6
ViT-L/14 (LAION-2B)	59.7	16.9	49.8	58.9	14.0	55.0	60.0	13.8	53.2	60.2	12.4	55.4	43.2	15.4	44.3	53.9	16.2	51.4	32.2	17.1	33.5	30.1	35.5	27.8	49.8
ViT-H/14 (LAION-2B)	60.2	16.4	54.8	59.2	13.6	59.4	59.4	13.0	58.7	60.9	15.1	56.3	45.3	15.9	49.6	54.2	14.8	55.2	30.1	18.2	35.1	31.0	33.8	30.9	50.0
ViT-bigG/14 (LAION-2B)	61.6	17.8	51.2	62.1	16.0	54.3	62.2	16.4	53.6	63.8	17.2	52.3	45.7	15.4	48.3	56.5	17.4	54.0	34.2	19.6	33.0	33.5	32.7	29.7	52.5
ViT-L/14 (DFN-2B)	59.1	15.7	50.2	57.8	13.7	53.2	58.0	14.2	54.1	59.5	15.5	52.3	41.8	15.2	45.6	51.3	13.5	54.4	28.0	17.4	34.7	31.1	29.6	31.8	48.3
ViT-SO-SigLIP2/14	63.3	25.3	47.8	62.7	22.7	52.8	62.4	21.2	53.1	63.6	20.0	53.8	53.0	23.0	51.7	58.6	25.4	51.4	34.1	27.0	32.7	36.9	42.3	29.3	54.3
ViT-B-SigLIP/16	57.5	16.5	49.2	56.8	13.5	53.6	56.6	13.0	53.8	57.4	13.0	54.4	38.7	15.2	44.3	48.9	14.4	52.6	27.9	17.1	33.8	29.1	33.6	26.7	46.6
ViT-B-SigLIP2/16	64.1	19.8	51.5	63.3	17.5	55.2	63.4	15.9	56.7	63.8	16.6	55.2	48.1	19.7	49.9	59.9	20.2	54.5	31.0	25.4	33.1	37.8	36.8	28.3	53.9
ViT-B-SigLIP2/32	54.8	36.0	33.6	53.9	35.8	35.0	55.0	37.2	34.6	58.2	31.2	42.7	30.9	36.1	24.5	41.6	32.3	24.7	26.0	34.7	20.8	25.6	45.7	7.7	43.2
ViT-B-SigLIP2/16 (512)	64.9	20.3	51.8	63.8	18.1	55.2	63.9	16.7	55.8	64.1	17.2	53.1	49.3	20.3	50.1	60.6	20.2	53.8	30.5	24.8	31.9	36.4	39.3	22.3	54.2
ViT-H-qgelu/14 (DFN-5B)	65.2	15.9	54.4	65.4	13.4	57.8	65.5	13.4	57.8	65.7	12.7	58.5	55.0	14.9	54.4	61.7	15.4	57.1	36.7	16.8	36.3	39.3	30.4	33.1	56.8
ViT-H-qgelu/14 (224)	63.2	16.2	51.3	62.9	13.8	55.7	63.3	13.6	54.8	64.1	14.3	54.2	50.9	15.6	49.2	57.4	15.0	52.5	35.9	17.4	34.7	38.2	29.4	23.2	54.5
ViT-B/16 (OpenAI)	52.2	15.6	43.7	52.3	13.8	48.2	51.7	14.0	48.8	54.8	13.1	48.0	35.2	14.2	37.9	42.6	16.0	45.1	26.1	16.8	31.3	25.0	33.9	23.7	42.5
ViT-B/32 (OpenAI)	47.9	14.0	37.3	48.0	12.0	41.5	46.0	13.7	42.0	48.3	12.6	46.3	26.8	13.2	30.2	30.1	15.3	34.4	21.8	14.5	26.4	22.1	38.2	21.4	36.4
ViT-L/14 (OpenAI)	56.5	17.1	44.3	56.7	14.9	48.2	56.0	15.2	48.1	57.4	16.8	45.2	46.9	15.6	44.6	49.2	16.8	46.0	26.7	16.7	34.3	28.3	30.1	12.1	47.2
ViT-L/14-336 (OpenAI)	57.6	15.3	46.4	57.8	12.8	50.8	57.2	13.3	49.7	58.4	11.7	50.9	49.5	13.4	48.1	53.1	15.3	49.2	27.3	14.4	35.4	31.8	28.4	30.4	49.1

larger ones, and Viewpoint and Scale the most severe declines. Across perturbations, larger models consistently outperform smaller ones; for example, under Rotation, `ViT-H-qgelu(378px)` reaches 57.4% versus 30.1% for `ViT-B/32(DataComp-1B)`, highlighting the benefit of greater capacity and training scale.

Among all variants, Scale is the hardest. `ViT-L/14(DataComp-1B)` drops from 62.2% on Bg-varied to 32.6% under Scale, while even `ViT-H-qgelu(378px)` falls from 65.2% to 39.3%. Viewpoint changes also cause substantial drops (e.g., `ViT-L/14` to 30.3%), though unlike Scale they do not coincide with major increases in *BG-Er* (Table 3).

Background reliance. Accuracy drops alone do not explain model failures. Scale not only lowers accuracy but also nearly doubles *BG-Er* relative to Bg-varied (Table 3), reaching 50.7% for `ViT-B/32(DataComp-1B)`. By contrast, under flips, rotations, or viewpoint changes, *BG-Er* remains relatively stable (about 12–17% for `ViT-L/14`). This suggests that these errors mainly reflect broad robustness deficits rather than spurious background correlations.

Model size and data: Larger models do not guarantee robustness. Although `ViT-bigG` and `ViT-H/14(DFN-5B)` achieve higher Bg-varied accuracy (Table 3), they still exhibit substantial *BG-Er* under Scale (≈30–33%). In contrast, models trained on curated data (e.g., DataComp-1B) show lower background reliance, indicating that pretraining data quality shapes shortcut behavior as much as model size.

Backbone and resolution effects. With fixed training data, architecture matters: within DataComp-1B, `ViT-B/16` consistently has lower *BG-Er* than `ViT-B/32` (30.5% vs. 50.7% under Scale, Table 3), suggesting that finer patching helps reduce background reliance. Scaling up models improves raw accuracy, but background-driven errors remain substantial. Likewise, higher input resolution (e.g., `ViT-B-SigLIP2` at `512px`) improves Bg-varied accuracy (64.9%) but only modestly reduces *BG-Er* (39.3% under Scale, Table 3), showing that resolution alone is insufficient.

Fine-grained confusion: Clutter, occlusion, and viewpoint shifts hinder discrimination between fine-grained classes (e.g., *chimpanzee* vs. *siamang*). As shown in Table 3 (*Fine-Er* columns), such errors persist across subsets, reflecting the inherent difficulty of COVAR. At smaller scales, models trade off *Fine-Er* against *BG-Er*; at larger patch sizes (e.g., `ViT-B-SigLIP2/32`), reduced access to detail further obscures fine distinctions and can amplify reliance on spurious background cues.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020. 2, 3

[2] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *International Conference on Computer Vision (ICCV)*, pages 397–406, 2021. 2, 3

[3] Xi Chen, Xiao Wang, Soravit Changpinyo, Anthony J Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 3

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. Ieee, 2009. 3

[5] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023. 3

[6] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:27092–27112, 2023. 3

[7] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 3

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014. 3

[9] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference (BMVC)*, 2018. 2, 3

[10] 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 (ICML)*, pages 8748–8763. PmLR, 2021. 1, 3

[11] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 2

[12] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016. 3

[13] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:25278–25294, 2022. 3

[14] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 2, 3

[15] 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. 3

[16] Qizhou Wang, Yong Lin, Yongqiang Chen, Ludwig Schmidt, Bo Han, and Tong Zhang. A sober look at the robustness of clips to spurious features. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:122484–122523, 2024. 1

[17] Ying Wang, Tim GJ Rudner, and Andrew G Wilson. Visual explanations of image-text representations via multi-modal information bottleneck attribution. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:16009–16027, 2023. 2, 3

[18] Zhuo Xu, Xiang Xiang, and Yifan Liang. Overcoming shortcut problem in vlm for robust out-of-distribution detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15402–15412, 2025. 1

[19] Yu Yang, Besmira Nushi, Hamid Palangi, and Baharan Mirzasoleiman. Mitigating spurious correlations in multimodal models during fine-tuning. In *International Conference on Machine Learning (ICML)*, pages 39365–39379. PMLR, 2023. 1

[20] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *International Conference on Computer Vision (ICCV)*, pages 11975–11986, 2023. 3

[21] Chenyang Zhao, Kun Wang, Janet H Hsiao, and Antoni B Chan. Grad-eclip: Gradient-based visual and textual explanations for clip. *arXiv preprint arXiv:2502.18816*, 2025. 2, 3

[22] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision (ECCV)*, pages 696–712. Springer, 2022. 2, 3