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

Supplementary Material

A. Appendix

A.1. Choice of K in CCI.

We use $K = 7$ to balance efficiency and interpretability. Figure 1 shows overall CCI runtime grows with K, while deletion curves and AUC remain stable across $K = 3-10$, indicating low sensitivity.

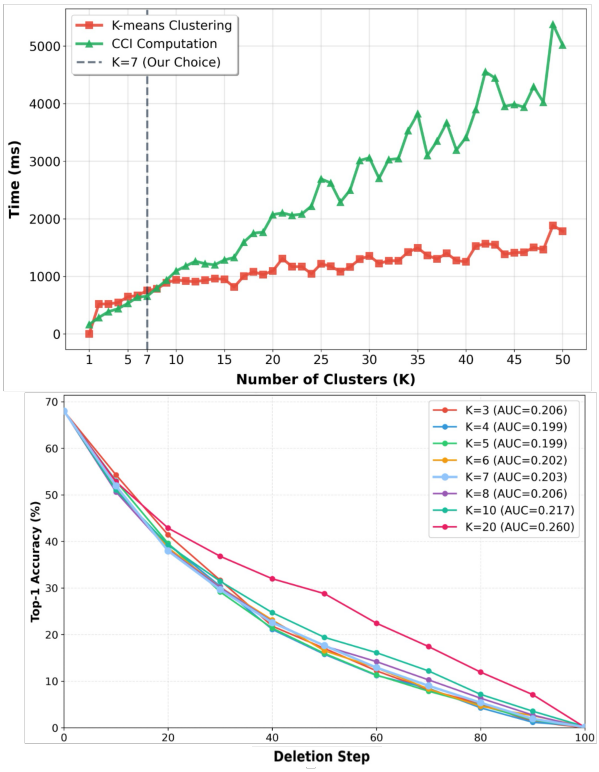

Figure 1. Effect of K.

A.2. Applicability of CCI in medical domain.

We apply CCI to MedCLIP [11] on MIMIC-CXR [4]. In Figure 2, for correct predictions with cardiomegaly (i), CCI focuses on the cardiac silhouette but shifts attention to non-diagnostic lung regions for incorrect predictions (ii), revealing failure source. Further, CCI achieves the lowest deletion and highest insertion AUC (Top-1) compared to Grad-CAM and Grad-ECLIP, as seen in Table 1.

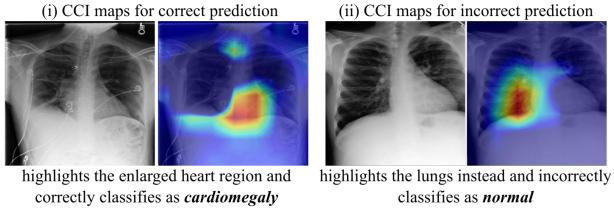

Figure 2. CCI results (medical domain).

Table 1. AUC of Deletion/Insertion curves using Top-1 accuracy (Medical Domain).

Method	GradCAM	Grad-ECLIP	CCI (ours)
Deletion ↓	0.3821	0.3164	0.2361
Insertion ↑	0.2701	0.3218	0.3804

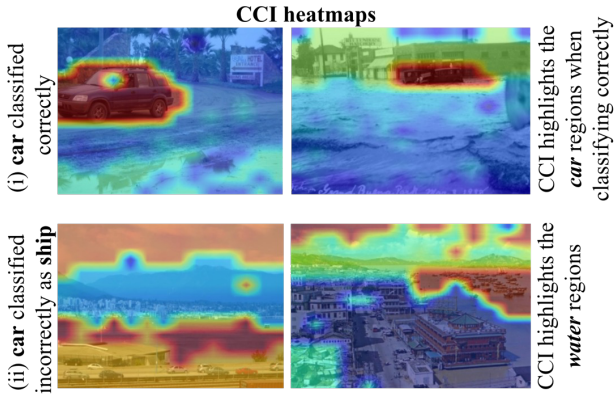

Figure 3. CCI results on NICO dataset.

A.3. Evaluating CCI on additional datasets

NICO. On NICO [3], CLIP's zero-shot accuracy is 88.36%, with notable drops only on a few classes. We pick a few such cases where accuracy $\leq 30\%$ (e.g., *car* in *water*), and note in Figure 3, when classification is correct (i), CCI focuses on the car. When misclassified as *ship* (ii), CCI shifts attention to water and sky, revealing background-driven reliance. This demonstrates that CCI is applicable to real-world datasets like NICO, while COVAR complements them by providing challenging alternatives.

Waterbirds. We compare CCI with TextSpan [1] in Figure 4 on Waterbirds [8]. In correct classifications (i), CCI yields more coherent maps. In misclassifications (ii), TextSpan focuses on the bird while CCI disperses atten-

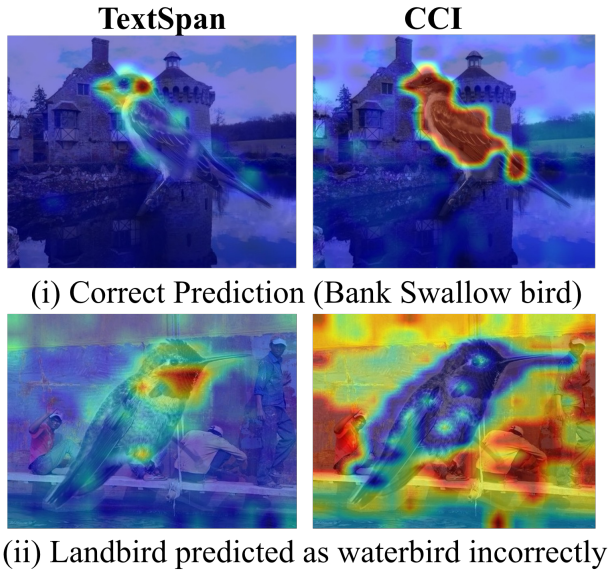

Figure 4. CCI results on Waterbirds dataset.

tion across background. We quantify this with mean bg attention over misclassified images (Grounded-SAM masks): CCI gives 0.1639 v. TextSpan's 0.0423, showing CCI's efficacy for diagnosing spurious bg reliance. We clarify CCI is efficient (3.5x faster than TextSpan) and does not need K+1 forwards as masking is used only at the final transformer block while TextSpan's multi-layer hook-based interceptions makes it slower.

A.4. Human Evaluation.

We conducted a human study with 41 participants to assess both the discriminative quality of attribution maps and the reliability of our editing pipeline. The evaluation consisted of two parts. (i) Class-discriminative map preference. Participants were shown attribution maps generated by CCI, Grad-ECLIP, MaskCLIP, and Grad-CAM, and were asked to select the map that best highlighted features relevant for distinguishing the target class. Across all comparisons, CCI was preferred in 91.2% of cases, indicating that our method more consistently identifies class-discriminative regions. (ii) Object preservation under editing. Participants were asked to rate Emu-edited images on a 1–5 scale based on how well the primary object was preserved after editing. The edited images achieved an average score of 4.86, demonstrating strong object fidelity and confirming the reliability of the editing process.

Together, these results validate both the effectiveness of CCI in producing class-discriminative explanations and the robustness of our editing pipeline in maintaining semantic object integrity.

A.5. Additional Qualitative CCI results

We present further qualitative comparisons of CCI against baseline methods in Figure 5 with a broader set of categories. Across diverse object types, CCI continues to generate heatmaps that are coherent, in contrast to the scattered or noisy activations produced by baselines. For instance, consider the *hair spray* in second row where CCI correctly attends to the spray bottles, while baselines either confuse or focus on small, disconnected fragments. Similarly, in challenging cases where object visibility is poor (see fourth row), CCI isolates the object of interest (*spider*) with sharper boundaries compared to baselines. These additional examples further provide strong evidence on the efficacy of proposed CCI method.

A.6. CCI results with other CLIP variants

We repeat CCI computation with two other pretrained vision–language models, the OpenAI CLIP `ViT-B/32` at `224px` model [5] and a SigLIP variant [12] `ViT-L/16` at `334px`, to evaluate the generality of our methodology. Figures 6 and 7 show results for the two variants repsectively. Findings are consistent with CLIP `ViT-B/16`: CCI generates concept-level, coherent heatmaps, indicating that CCI adjusts effectively to variations in backbone resolution and loss formulation. These findings demonstrate that CCI is independent of model and reliably generates comprehensible visual explanations for encoders from the OpenAI and OpenCLIP families.

A.7. foreground mask computation

We leverage GroundedSAM [6] to generate foreground (FG) and background (BG) masks, which are then used to classify CLIP predictions as foreground-driven (*FG-Er*) or background-driven (*BG-Er*) across various datasets. For each image, we provide GroundedSAM with the prompt `<class name>, foreground objects.` The inclusion of "foreground objects" ensures that any distractor objects such as the *bucket* in Section 3.1 are correctly captured as part of the foreground mask. These masks then serve as a proxy for ground-truth object regions when computing Class-Conditional Importance (CCI) heatmaps. For each prediction, we compute the intersection-over-union (IoU) between the CCI heatmap and the FG/BG masks to determine whether the error is primarily foreground-driven (*FG-Er*) or background-driven (*BG-Er*).

We further validate GroundedSAM on the ImageNet-Segmentation (ImageNet-S) [2] validation set, which contains segmentation annotations on 12,419 images spanning 919 ImageNet categories. We create GroundedSAM masks using the same prompt and compare them against the dataset-provided masks. Figure 9 shows qualitative examples displaying the input image, GroundedSAM predicted mask, and the ground-truth mask. The average IoU between

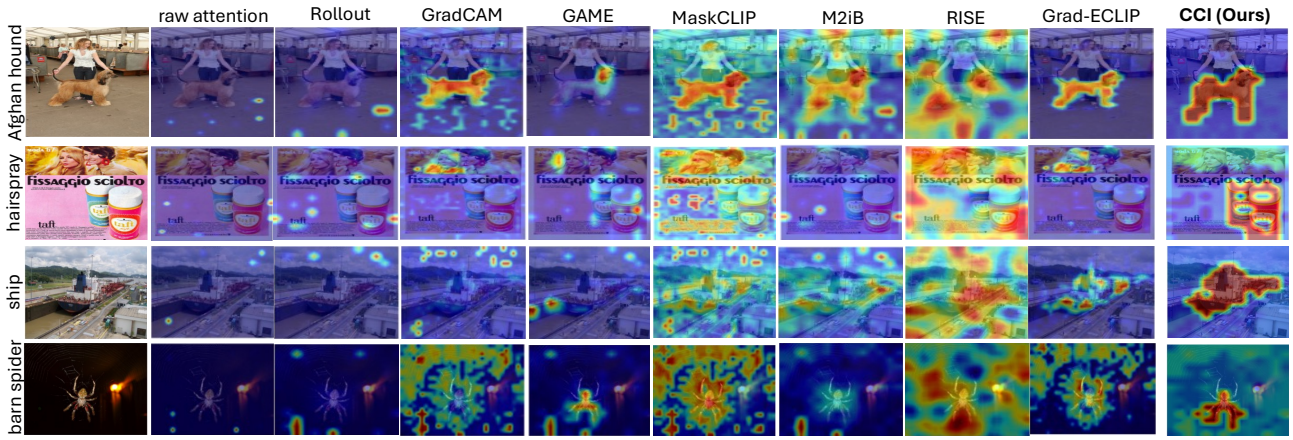

Figure 5. Additional Results comparing CCI against baseline interpretability methods.

predicted masks and ImageNet-S masks is 0.93, demonstrating high alignment.

We show additional qualitative results on the CounterAnimals dataset in Figure 8. One can note that even in cases of heavy occlusion (see fifth row the figure), GroundedSAM correctly captures the foreground region of interest, illustrating that it can reliably captures foreground objects across diverse scenarios.

A.8. Prompting Details

A.8.1. fineG computation

As mentioned in Section 3.1 in the main paper, we used GPT-4o as a vision expert to determine whether the misclassified examples belonging to foreground-driven (*FG-Er*) represent fine-grained visual confusion or egregious failures. Below, we provide the exact prompt:

Example Outputs:

- Ground truth: *siamang*, Predicted: *chimpanzee* → `similar`
- Ground truth: *border collie*, Predicted: *australian shepherd* → `similar`
- Ground truth: *cat*, Predicted: *airplane* → `different`
- Ground truth: *lion*, Predicted: *bicycle* → `different`

By separating errors caused by mild intra-class similarity from more serious classification errors, this automated labelling gives us more information about the nature of model confusion than just accuracy metrics.

A.8.2. class selection

As mentioned in Section 4.1, we curate a representative subset of classes from ImageNet while balancing semantic coverage and visual diversity. The goal is to avoid redundancy (e.g., multiple dog breeds) while still spanning a broad range of living and non-living concepts. To guide this process, we use GPT4o with the prompt shown below:

Table 2 lists all selected classes included in our benchmark, providing the foundation for the subsequent background and structured variant generation.

A.8.3. curating backgrounds

After selecting the classes, we systematically generate diverse background contexts for each image. Our aim is to disentangle model reliance on object appearance from contextual cues by creating multiple, semantically neutral backgrounds. The prompt used is below:

Table 3 lists all backgrounds included in our benchmark. These backgrounds cover natural, urban, and indoor environments, including water, snow, forest, desert, and indoor settings. Each background is applied to all 33 classes, creating systematic variants that allow us to disentangle object-specific recognition from spurious background reliance.

A.9. Classwise results with other CLIP variants

To assess the generality of the background sensitivity patterns observed with CLIP `ViT-B/16`, we evaluated additional CLIP variants on the same 33,000 Bg-varied images. Our goal was to determine which aspects of the per-class accuracy drop are model-specific versus broadly expected across backbones.

Figures 11– 14 present classwise accuracy drops for various CLIP variants, with plots labeled from (a) to (m). For example, in plot (a) (class ID 6), the OpenAI CLIP `ViT-B/32` model exhibits a substantially higher relative accuracy drop compared to the same class under CLIP `ViT-B/16` (80 vs 20%). On the other hand, many other classes, such as IDs 1 and 33, continue to show smaller drops, consistent with the pattern seen in `ViT-B/16`.

Overall, while some classes behave differently across variants, the broader patterns are consistent: strongly background-dependent classes continue to exhibit significant drops, and those that were resilient to background vari-

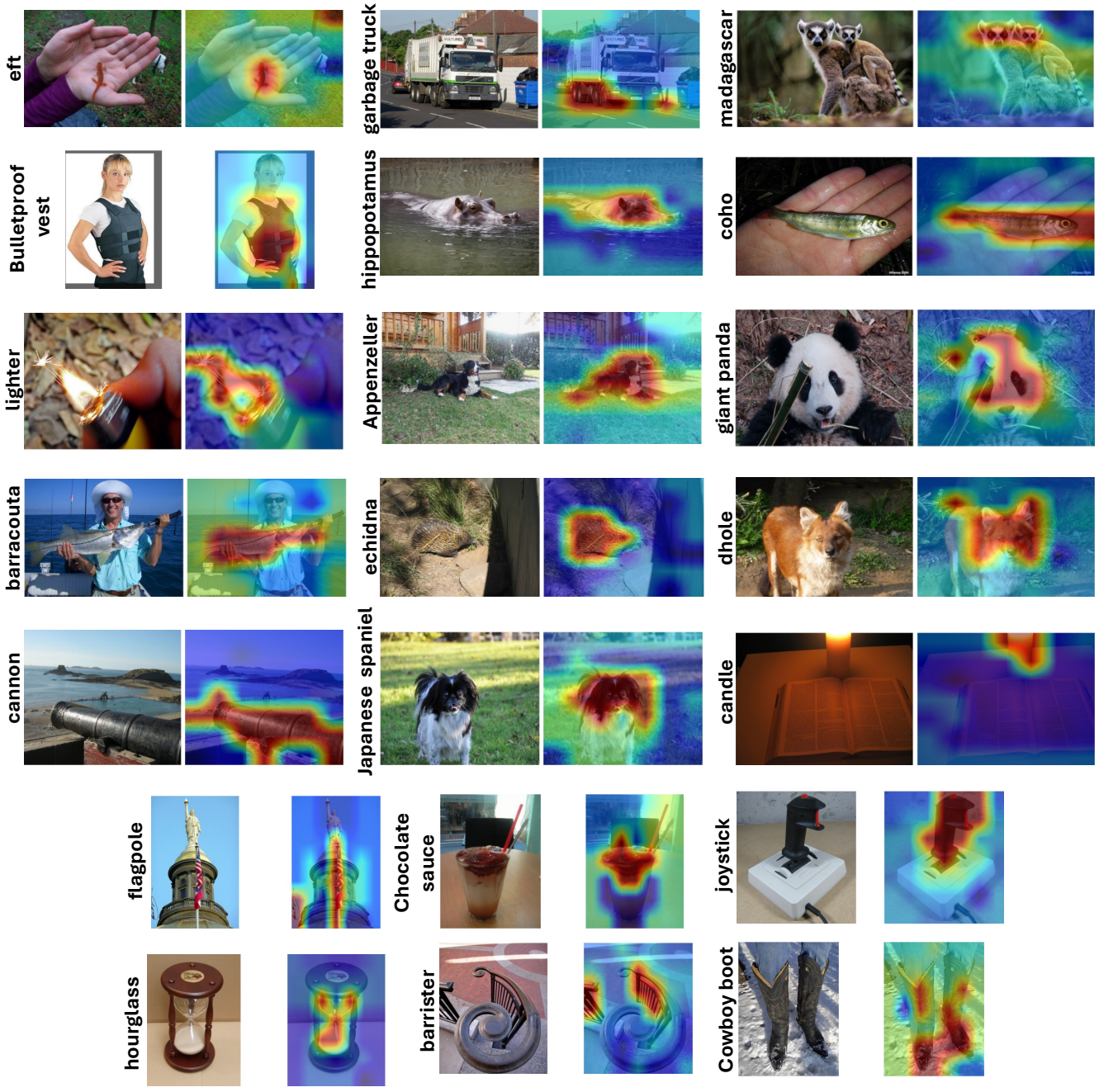

Figure 6. CCI Results with OpenAI CLIP `ViT-B/32-224px` model.

ation with `ViT-B/16` typically remain resilient with other backbones. This highlights the usefulness of our benchmark for examining spurious correlations and implies that the observed background effects are primarily dataset- and class-intrinsic rather than an artifact of a particular model.

A.10. Implementation Details of Variants

Here we provide implementation details for the variants generated in our proposed COVAR dataset.

Background variants. We use Emu2 [10] to create background variants. For each class, we kept the original object in the foreground while changing the background to

Table 2. Final set of 33 classes selected for the benchmark, labeled with Class IDs.

Class ID	Class Name	Class ID	Class Name
1	African elephant	18	park bench
2	Arabian camel	19	prairie chicken
3	Gila monster	20	pretzel
4	airship	21	rain barrel
5	alligator lizard	22	sea anemone
6	barn spider	23	slug
7	black swan	24	stove
8	bulbul	25	street sign
9	bullfrog	26	studio couch
10	cauliflower	27	submarine
11	chimpanzee	28	suspension bridge
12	dishwasher	29	trailer truck
13	electric locomotive	30	vulture
14	great white shark	31	warplane
15	hen	32	water ouzel
16	hermit crab	33	zebra
17	ice bear		

Table 3. Curated backgrounds for the benchmark. Each background is applied to all 33 classes to generate systematic variants.

Bg ID	Description
1	railway track in an outdoor setting
2	colorful garden with flowers and greenery
3	dense tropical forest
4	farmyard
5	hot desert with sand dunes
6	open grassland with tall green grasses
7	swampy area
8	cozy living room
9	savanna
10	calm ocean with clear blue water
11	fluffy white cloud in a bright blue sky
12	highway with empty road stretching behind
13	rocky shore with waves
14	rocky terrain
15	snowy landscape
16	beach
17	forest floor with leaves and sunlight filtering through trees
18	tree branches in a leafy forest
19	crowded marketplace with people and stalls
20	night cityscape with artificial lights

various natural settings. The model received instructions like *"`background description`. Edit only the background and keep the foreground subject intact"*. This ensured that only background pixels were changed, while the object's identity and position remained the same.

Viewpoint variants. To achieve viewpoint diversity, we used Zero123+ [9], a text-to-3D image synthesis method. For a given original image, Zero123+ generates 6 new viewpoints of which we randomly select 2. We created images using 75 inference steps, which were enough to keep fine details in general objects while ensuring consistent viewpoint changes.

fineG computation
System Prompt: You are a vision expert with deep knowledge of object categories and visual characteristics. Your task is to determine whether two categories are visually similar or clearly different based on appearance alone. Consider shape, texture, color, size, and typical visual features that a human would notice.
User Prompt: Ground truth class: `[gt_class]`
Predicted class: `[pred_class]`
Question: Evaluate whether these two categories are visually similar or clearly different. Consider the following:
1. Would a human observer easily confuse the two categories in a standard image?
2. Do they share key visual features (shape, color patterns, textures) that make them look alike?
3. If they are visually distinct and unlikely to be confused, classify them as different.

Respond with a single word only: `similar` if they are visually alike, `different` if they are clearly distinct.

class selection
System Prompt: You are an expert in computer vision and dataset curation. Your task is to select a semantically and visually diverse subset of ImageNet classes for use in understanding spurious correlations in VLMs.
User Prompt: You are given a large set of 1,000 ImageNet classes. Your goal is to propose a smaller subset of about 30–40 classes that are semantically and visually diverse. Follow these guidelines:
1. Ensure coverage across living and non-living categories.
2. Avoid redundancy (e.g., do not include many dog breeds or many bird species). Select only a few representative ones.

curating backgrounds
System Prompt: You are an expert in image editing and dataset creation. Your task is to propose diverse and realistic background settings for synthesizing objects in images.
User Prompt: Generate a list of 20 distinct background types that maximize diversity across scenes. The list should be independent of any specific object class and broadly applicable to placing different kinds of objects. Follow these guidelines:
1. Include both outdoor and indoor settings.
2. Ensure coverage across natural scenes, urban settings, and indoor environments.
3. Avoid repeating backgrounds that are too similar.

Scale variants. We produced scale variants with Stable Diffusion inpainting [7] by outpainting the original image onto larger canvases. Each image was expanded to different scale factors (up to $8\times$), keeping the original object centered (see Figure 15 for an example). The inpainting mask (generated using GroundedSAM as described in A.7) made sure that only the surrounding areas were generated, preserving the original foreground content. We used a standard classifier-free guidance scale of 7.5 and applied 30 diffusion steps for all images.

Others. We applied five standard geometric transformations using OpenCV, designed to preserve semantic content while perturbing pixel-level statistics:
- **Rotation:** images were rotated by a random angle in $[-45°, 45°]$, with borders filled via Stable Diffusion inpainting.
- **Horizontal and Vertical Flips:** standard left–right and top–bottom flips.
- **Crop:** cropping a region from center covering 60–90% of the original image area, followed by resizing.
- **Translation:** shifting the image up to 20% of its width/height in each direction, with borders filled via Stable Diffusion inpainting.

Together, these methods produce a well-rounded set of variants that enable controlled evaluation of background reliance, viewpoint generalization, scale sensitivity, and standard geometric robustness.

A.11. classwise results for all subsets

Figure 16 presents per-class accuracy drops across seven subsets (excluding the bg-varied subset, which is shown in the main paper in Section 4.2). For each class, we compute the accuracy over all images in a subset and compare it to the original ImageNet accuracy for that class.

The plots in Figure 16 show that for all subsets other than scale, accuracy degradation varies significantly across classes: some classes remain consistently robust, while others show notable sensitivity. In contrast, for the scale subset in plot (a), every class exhibits a substantial drop, with a minimum decline of approximately 25%. For example, class 33 shows a modest average drop of 5% across the non-scale subsets, but under the scale subset, the drop is almost 30%, illustrating that scale affects this class much more

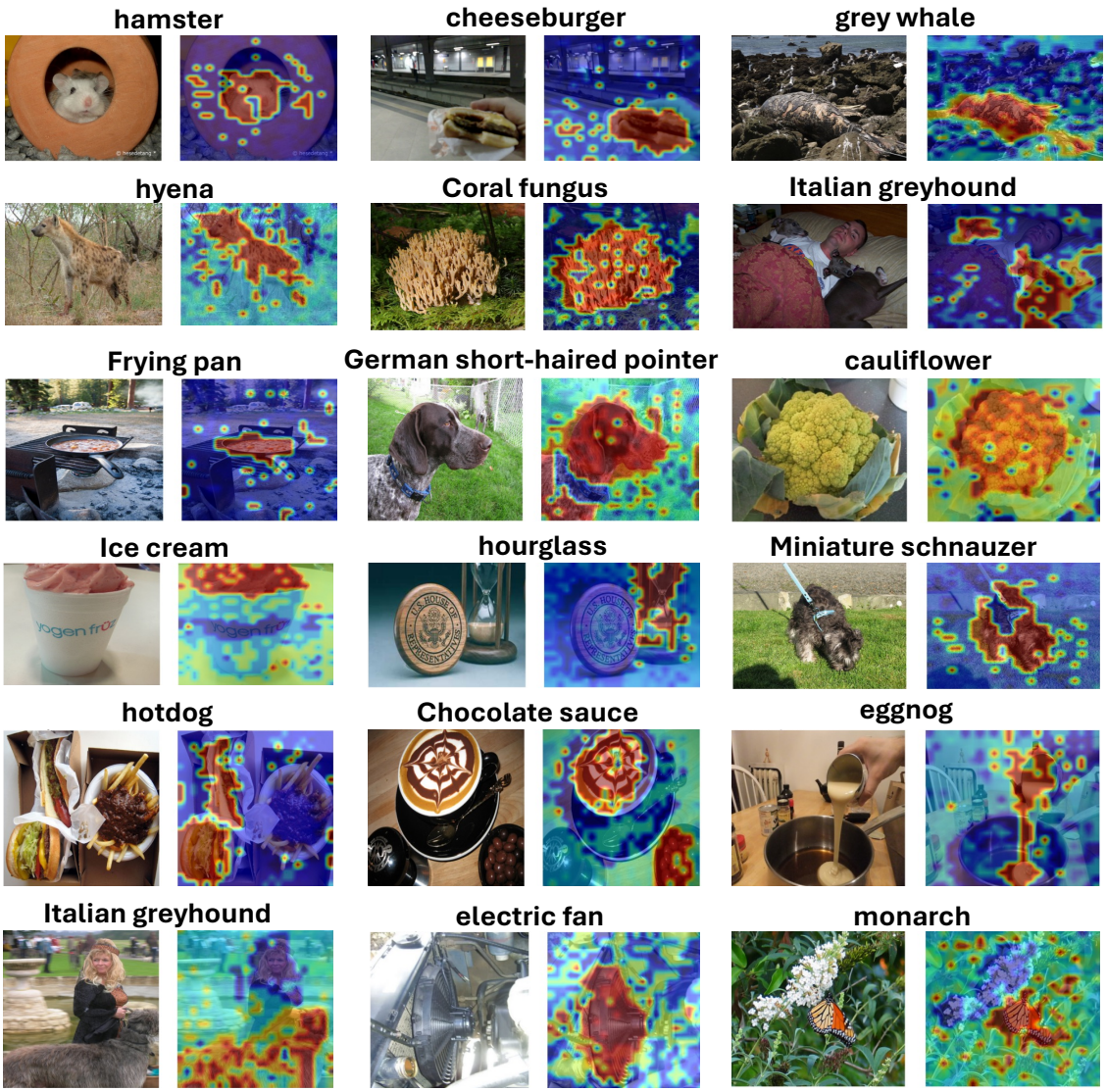

Figure 7. CCI Results with SigLIP `ViT-L/16-334px` model.

strongly than the others. Overall, these results indicate that CLIP's robustness is both class- and subset-dependent, with scale having a uniformly strong impact across all classes.

A.12. CCI Results on COVAR

Figure 17 presents CCI visualizations on samples from COVAR, illustrating how CLIP's focus shifts under different variant conditions. In the first row, we show a pair of images where the right image is a scale-reduced variant of the left. On the original image, CCI correctly focuses on the foreground and predicts the class as *African elephant*. However, in the scale-reduced variant, the model's attention shifts toward the background resulting in a misprediction as a *freight car*, likely due to the railway-track context. The second row depicts a *barn spider* in two different backgrounds: while the model accurately predicts the left image, it attends to the beach background in the right image, erroneously predicting a *crab*. Subsequent rows illustrate additional qualitative patterns, such as v-flip variants where predictions are incorrect yet the model still focuses on the foreground. Notably, in such cases, the misclassified classes remain visually similar to the ground truth, for example predicting a *moving van* instead of a *trailer truck*. These examples complement the main paper's insights, demonstrating that even when accuracy drops for certain variants like v-flip, the fraction of background-driven correlations remains largely unchanged.

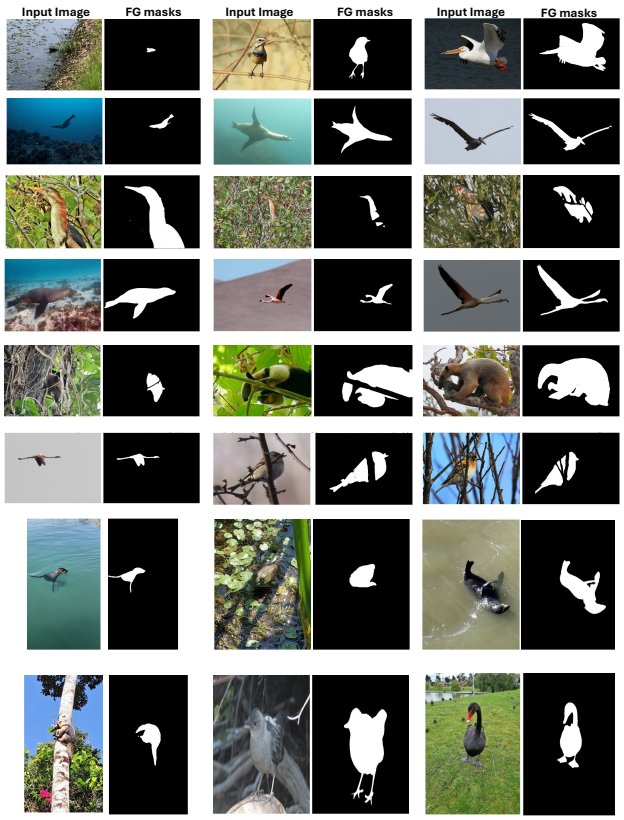

Figure 8. Qualitative results of GroundedSAM on the CounterAnimals dataset.

Figure 9. Comparison of GroundedSAM predicted foreground masks with the ImageNet-S ground-truth segmentation masks. For each image, the first row shows the input image, the second row shows the GroundedSAM predicted mask, and the last row shows the ImageNet-S ground-truth mask.

A.13. Aggregating Predictions by Background Context

To complement the quantitative analysis of robustness and spurious correlations, we also conducted a qualitative examination of CLIP's background biases by aggregating its predictions across images sharing the same background context. Specifically, for each background type, we averaged predictions over all corresponding images and recorded the most frequently predicted classes. Table 4 summarizes representative results for a subset of backgrounds. This view reveals strong, dataset-wide associations between certain backgrounds and particular object categories- for example, railway tracks strongly elicit predictions of *locomotive* or *bullet train*, even when the ground-truth object is unrelated. Such correlations are likely a reflection of CLIP's training distribution, where railway tracks frequently co-occur with trains, leading the model to overweight background context as a cue for object recognition. These observations highlight that CLIP's predictions are often guided more by contextual cues than by the objects themselves, underscoring the importance of explicitly disentangling object and background information in evaluating model behavior.

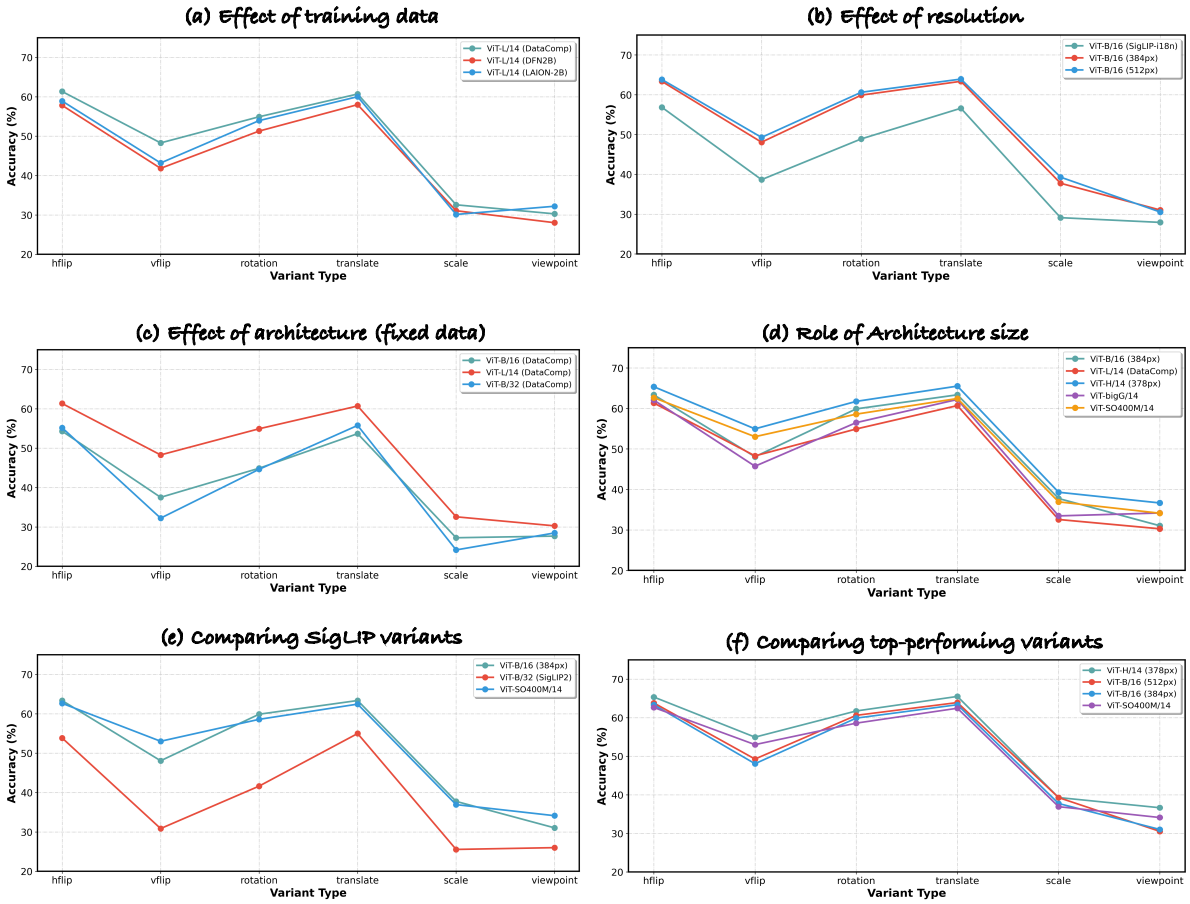

Figure 10. Average CLIP accuracies across dataset subsets.

Table 4. Top Predicted Classes per Background (Averaged across all images)

Background	Classes
railway track	`locomotive`, `bullet train`, `freight car`
rocky shore	`water ouzel`, `hermit crab`
garden	`rain barrel`, `park bench`
tropical forest	`chimpanzee`, `bulbul`
sky	`vulture`, `airship`
road	`zebra`, `trailor truck`, `street sign`
swampy area	`bullfrog`
desert	`Arabian Camel`

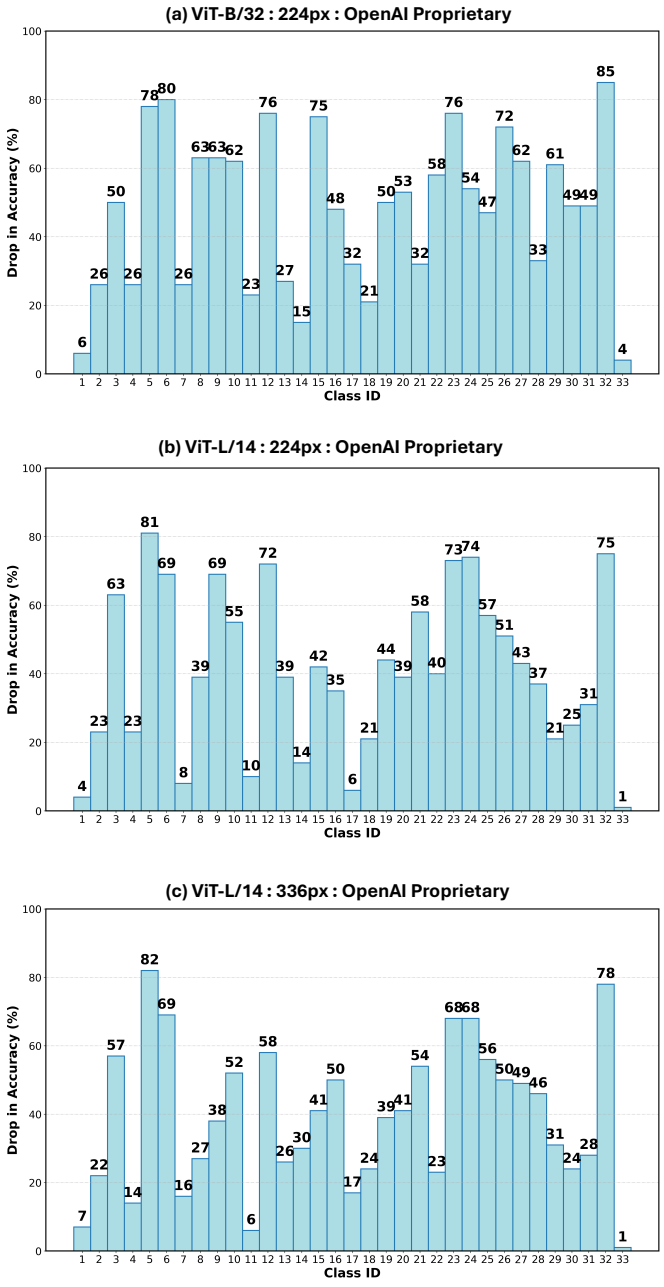

Figure 11. Classwise accuracy drops across OpenAI CLIP variants.

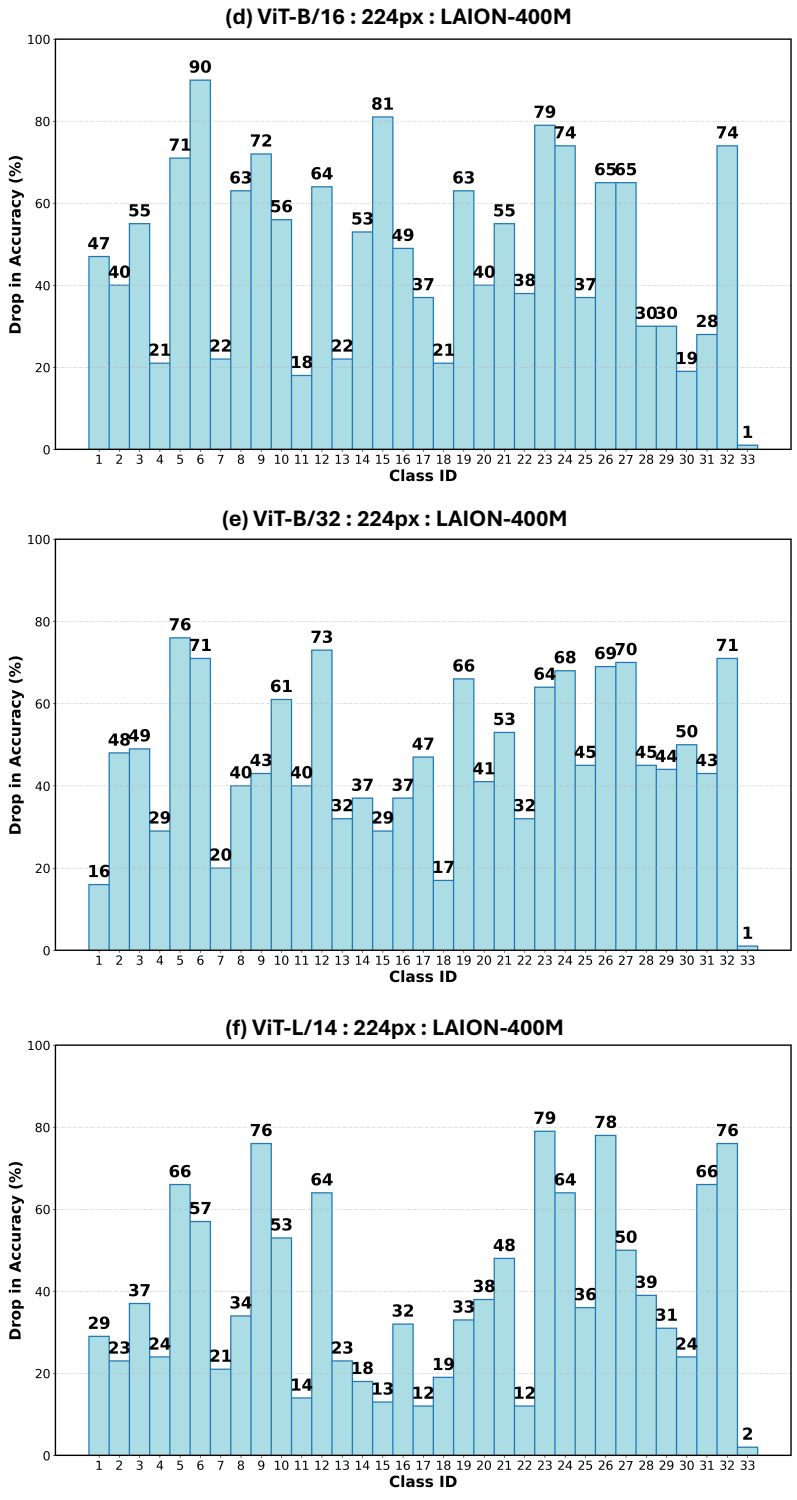

Figure 12. Classwise accuracy drops across OpenCLIP variants pretrained on LAION-400M.

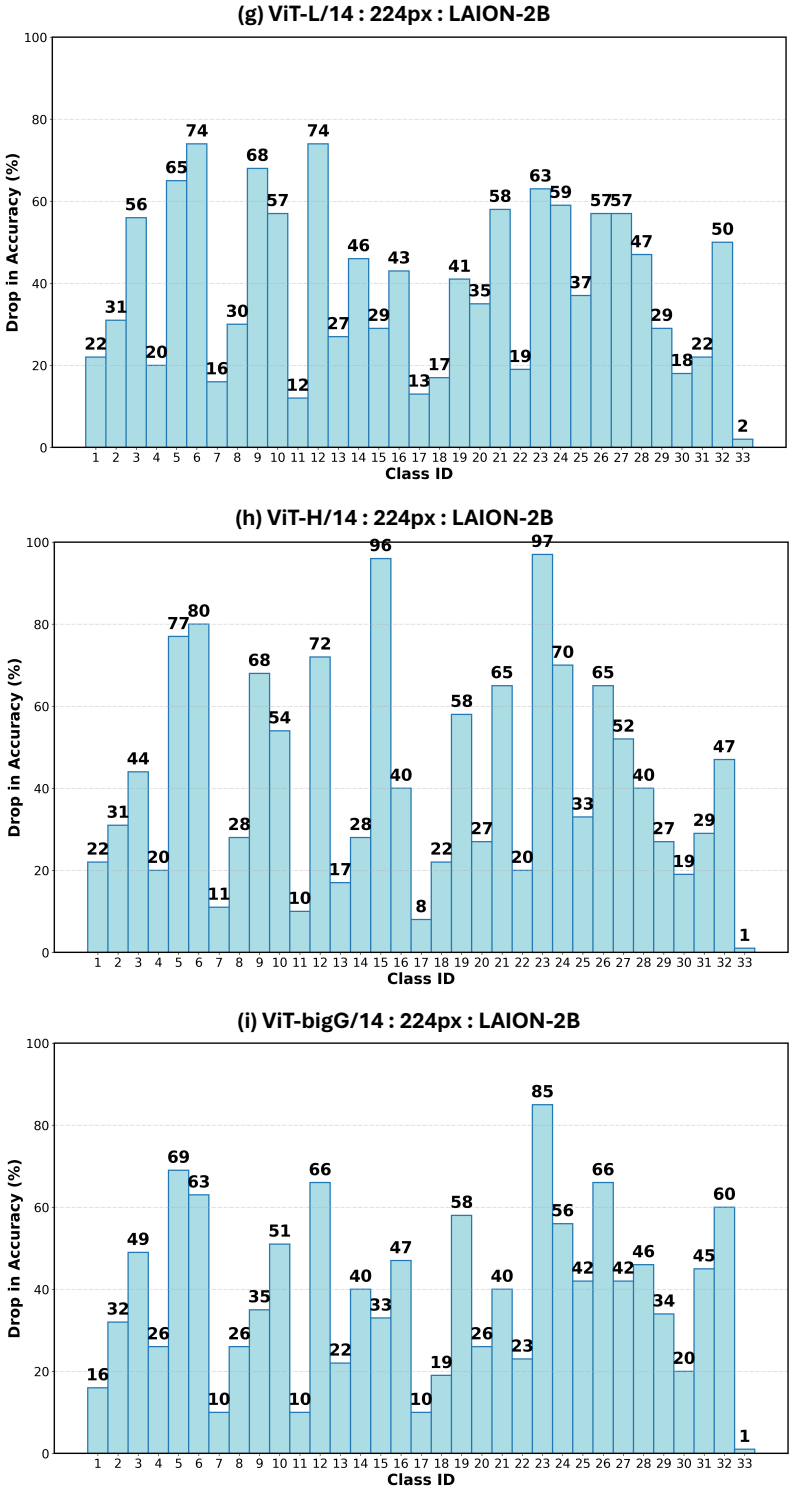

Figure 13. Classwise accuracy drops across OpenCLIP variants pretrained on LAION-2B.

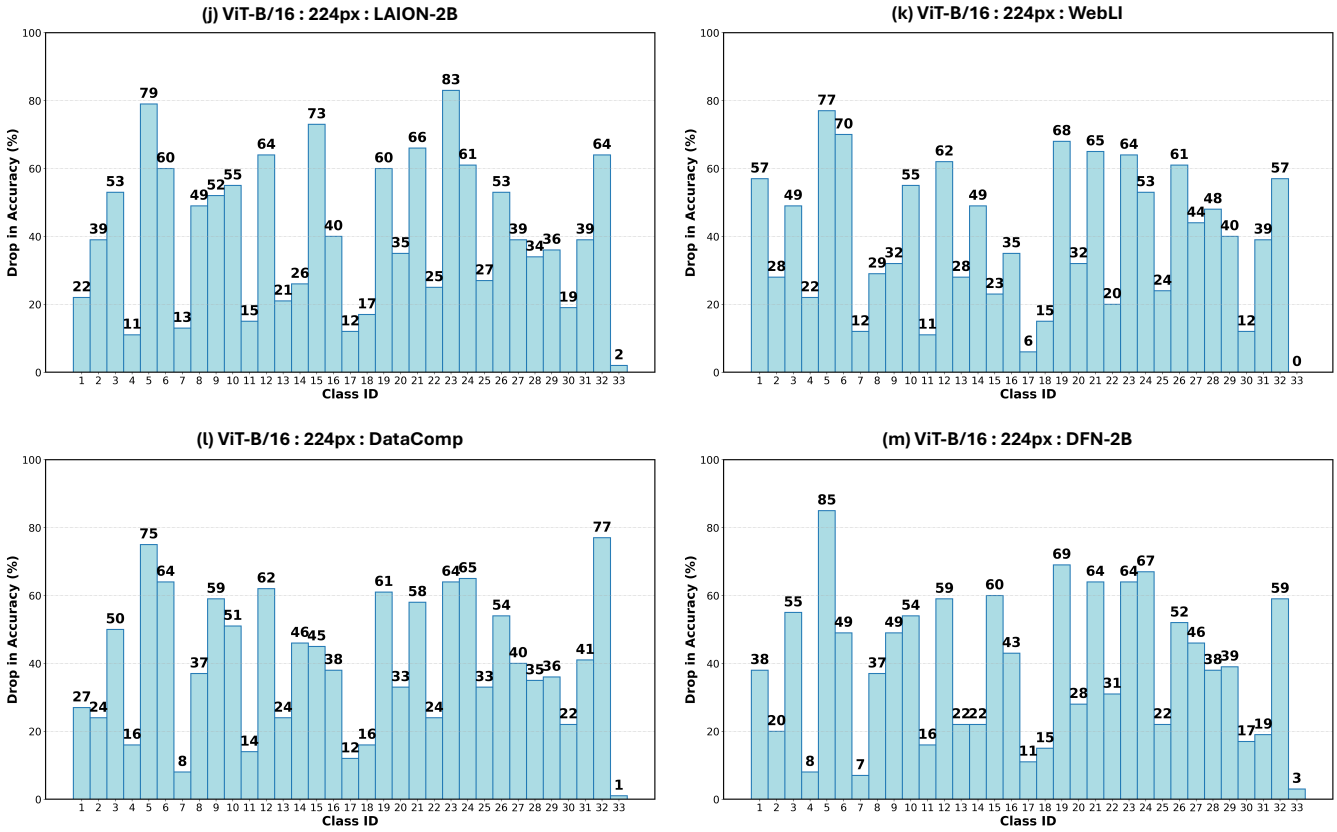

Figure 14. Classwise accuracy drops across `ViT-B/16` OpenCLIP variants on different datasets.

Figure 15. Example demonstrating varying scales for the same input image.

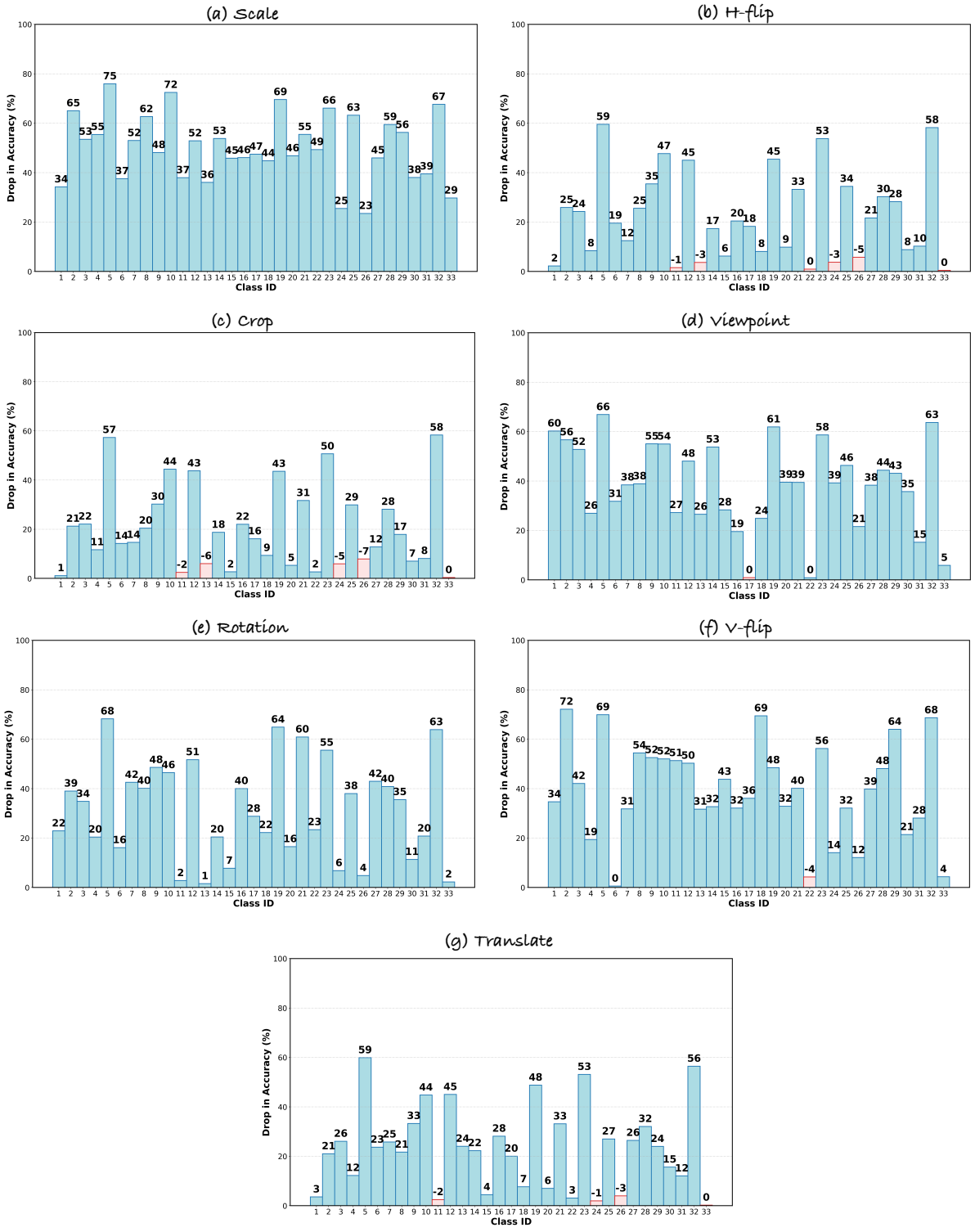

Figure 16. Per-class accuracy drops across various dataset subsets.

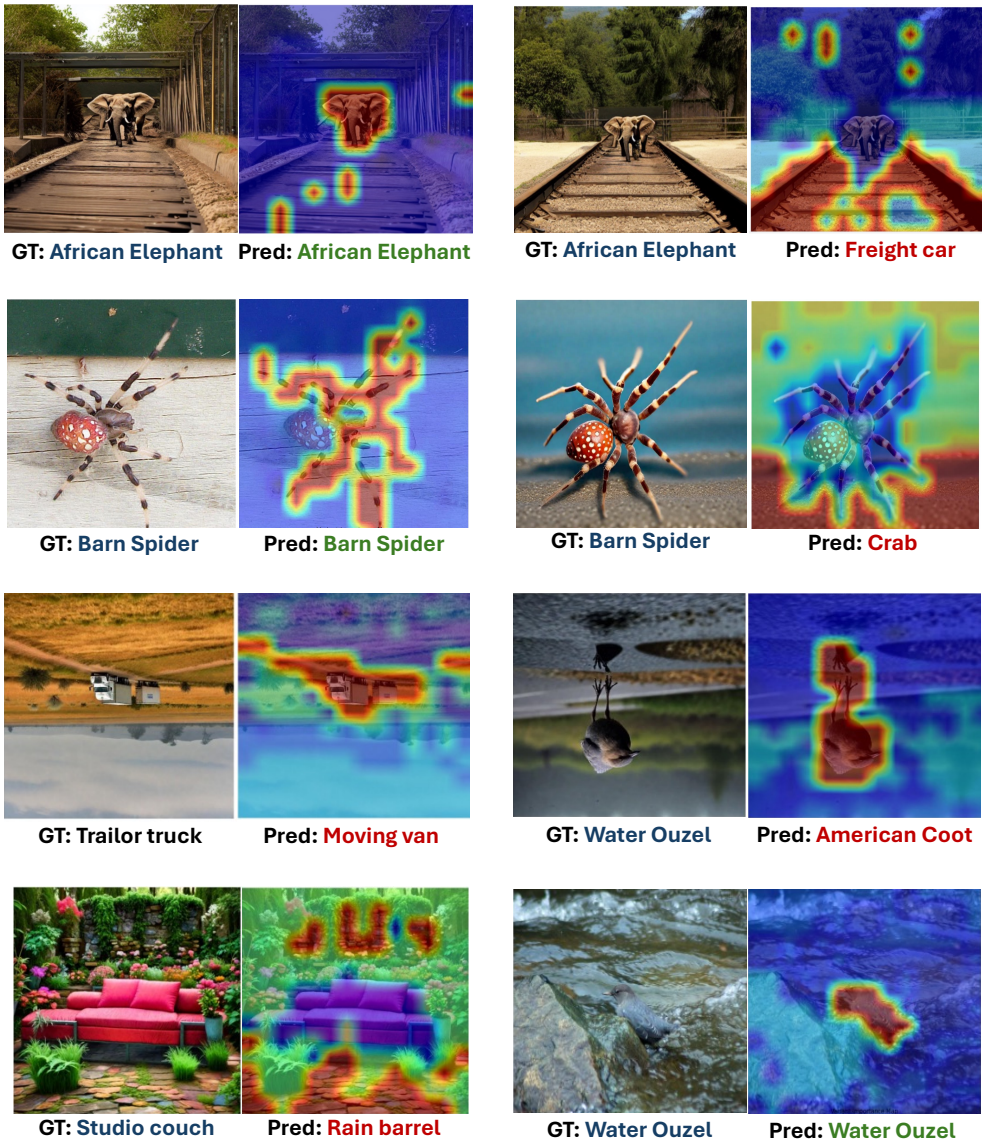

Figure 17. Qualitative CCI Results on COVAR.

References

[1] Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting clip's image representation via text-based decomposition. *arXiv preprint arXiv:2310.05916*, 2023. 2

[2] Shanghua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. Large-scale unsupervised semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 45(6):7457–7476, 2022. 3

[3] Yue He, Zheyan Shen, and Peng Cui. Towards non-iid image classification: A dataset and baselines. *Pattern Recognition*, 110:107383, 2021. 2

[4] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019. 2

[5] 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. 3

[6] 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. 3

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 7

[8] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 2

[9] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 6

[10] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14398–14409, 2024. 5

[11] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3876–3887, 2022. 2

[12] 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