Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Adaptive Hard Negative Mining for Fine Grained Vision Language Understanding A Negative Result and Analysis

Silas Ayitey

silasayitey@gmail.com

Kwame Nkrumah University of Science and Technology

Kwabena Owusu-Agyemang

Kwame Nkrumah University of Science and Technology

Research Article

Keywords: Vision-language models, Hard-negative mining, Fine-grained understanding, Benchmark evaluation, Compositional reasoning

Posted Date: November 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11030933/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Vision-language models perform well on many multimodal tasks but struggle with fine-grained visual-linguistic reasoning. The SPEC benchmark was introduced to evaluate models on four fine-grained concepts: object size, position, existence, and count. Fine-tuning on SPEC with fixed hard negatives improves baseline performance substantially. This paper asks whether adaptive hard-negative mining, which reweights negatives by their current difficulty, can push performance further. We propose a confusion-weighted training objective and evaluate it on SPEC held-out data. The method underperforms fixed hard-negative training by 23 percentage points on average. We trace the cause to SPEC's design: its candidate sets are uniformly hard, so adaptive reweighting has no difficulty signal to exploit and instead destabilizes training. We also introduce SPEC++, a paired-reasoning diagnostic that combines cases from two SPEC subsets. The fine-tuned model's joint accuracy on SPEC + + equals the product of its individual accuracies, indicating that it treats concepts independently rather than composing them. These two findings serve as cautionary notes for the field: uniform hard negatives are difficult to outperform when the data is designed to be uniformly hard, and high benchmark accuracy should not be read as evidence of compositional understanding.

1. Introduction

Vision-language models such as CLIP [1], BLIP [2], and CoCa [3] have become standard tools for multimodal tasks including image classification, captioning, and cross-modal retrieval. Their strong performance on benchmark suites has led to broad adoption. Underneath this success, however, lies a persistent weakness: these models do not reliably distinguish between subtle visual-linguistic concepts. When asked whether an image shows three dogs or four, or whether an object is on the left or the right, they frequently perform no better than chance [4, 5, 6].

Recent work has responded by building benchmarks that isolate specific fine-grained concepts. The SPEC benchmark [7] evaluates four such concepts, size, position, existence, and count, using candidate image-text sets that differ only in one attribute. SPEC also comes with a training recipe based on fixed hard negatives. Within each candidate set, every incorrect text is treated as a hard negative, and all such negatives receive equal weight during training. Fine-tuning CLIP with this recipe produces large improvements over the baseline.

It is natural to ask whether adaptive hard-negative mining, which reweights negatives based on how difficult they are for the model at a given step, can improve on the fixed approach. Prior work in metric learning [17, 18, 19] has shown that hard negatives carry more training signal than easy ones. If negatives within SPEC's candidate sets vary in difficulty, then adaptive reweighting should help.

We test this hypothesis. We propose a confusion-weighted training objective that gives higher weight to negatives the model currently confuses with the correct answer. We evaluate on SPEC held-out data and compare against fixed hard-negative training, as summarised in Fig. 1 and Fig. 2. The adaptive variants underperform the fixed baseline by 23 percentage points on average. We analyze the failure and identify

its source: SPEC's candidate sets are uniformly hard by construction, so adaptive weighting has no difficulty gradient to exploit, and the reweighting destabilizes the effective learning rate, as visible in the training trajectories of Fig. 4.

We also investigate whether fine-tuned models acquire compositional reasoning. We introduce SPEC++, a paired-reasoning diagnostic that combines cases from two SPEC subsets and measures joint accuracy (Fig. 3). The model's joint accuracy is precisely the product of its individual accuracies. It solves each concept on its own but does not combine them.

Our contributions are as follows. First, we reproduce SPEC with a fine-tuned CLIP model that reaches 88.7% average accuracy on held-out test data, against a baseline of 33.5% (Table 1, Table 2). Second, we propose a confusion-weighted adaptive hard-negative objective and show empirically that it underperforms the fixed approach by 23 percentage points, with an analysis of why (Table 3). Third, we introduce SPEC + + and use it to show that fine-tuned models treat fine-grained concepts independently rather than composing them (Table 4). Fourth, we argue that these results expose a limitation of current fine-grained VLM benchmarks and call for evaluation protocols that explicitly target compositional reasoning.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the methodology. Section 4 presents results. Section 5 discusses implications. Section 6 concludes.

2. Related Work

2.1 Vision-Language Models

CLIP [1] established the contrastive learning paradigm by matching image and text embeddings in a shared space, achieving strong zero-shot transfer. BLIP [2] extended this with bootstrapping and multi-task training, and CoCa [3] combined contrastive and captioning objectives. FLAVA [8] aligned separate encoders for vision and language, while ALIGN [9] showed that scale alone produces competitive representations. More recent work explores open-vocabulary extensions [10, 11], multi-grained alignment [12], and video-language models [13].

Several studies have shown that these models rely on shallow heuristics. Yuksekgonul et al. [4] found that CLIP behaves like a bag-of-words model on tasks requiring spatial and relational understanding. Thrush et al. [5] showed that models perform at or below chance on compositional image-text matching. Our work builds on these observations and shows that even a fully fine-tuned model treats fine-grained concepts as independent problems.

2.2 Fine-Grained Evaluation Benchmarks

Several benchmarks target fine-grained visual-linguistic understanding. VALSE [14] evaluates linguistic phenomena through controlled text perturbations. ARO [4] diagnoses attribution, relation, and ordering. Eqben [15] assesses sensitivity to visual semantic changes. VL-Checklist [16] evaluates objects,

attributes, and relations. SPEC [7] introduced a progressive data construction pipeline that generates candidate image sets in which only a single visual attribute varies while everything else stays fixed. This enables symmetric evaluation from image-to-text and text-to-image directions. Our work reproduces SPEC and extends it with a paired-reasoning diagnostic.

2.3 Hard-Negative Learning

Hard-negative mining has a long history in metric learning and retrieval [17, 18, 19]. In the VLM literature, Syn-CLIP [20] uses synthetic images from 3D engines. EQSIM [15] adds an equivariance regularization term. TSVLC [21] and ViLOM [22] introduce negatives generated by large language models. Construct-VL [23] approaches the problem from a continual learning perspective. SPEC [7] itself trains with fixed hard negatives drawn from the candidate set. Our method differs in that it reweights existing negatives dynamically rather than constructing new ones. To our knowledge, this is the first study to systematically evaluate adaptive hard-negative mining on SPEC.

2.4 Synthetic Data for Vision-Language Training

Synthetic data has become a useful tool for teaching VLMs about concepts that are rare in natural image-text corpora. Syn-CLIP [20] demonstrated the value of 3D-simulated images for attribute learning. SPEC [7] uses Stable Diffusion XL [24] together with Grounding-DINO [25] and SAM [26] to generate controlled candidate images. Our work operates entirely on SPEC's released data and does not require additional generation.

3. Methodology

3.1 SPEC Reproduction

Dataset. SPEC contains six subsets. Each subset holds many test cases. A case consists of a query image and K candidate texts, exactly one of which correctly describes the image. The subsets and their candidate counts are: absolute*size (K = 3), relative*size (K = 3), absolute*spatial (K = 9), relative*spatial (K = 4), existence (K = 2), and count (K = 9). For each subset, we load the provided image2text.json file, which contains between 1000 and 4500 cases depending on the subset.

Train and test split. We split each subset 80/20 by index. The first 80% becomes the training set, and the remaining 20% becomes the held-out test set. Query image paths are disjoint between the two splits. We verify that no image file appears in both splits, ruling out image-level leakage.

Evaluation. For a test case with query image I and candidates T1 through TK, we encode I with CLIP's image encoder and each Ti with the text encoder. We compute cosine similarities between the image embedding and each text embedding, then predict the candidate with the highest similarity. Accuracy is the fraction of cases where the predicted index matches the ground-truth label.

3.2 Fixed Hard-Negative Training

We fine-tune CLIP ViT-B/32 from the OpenAI checkpoint [1] using the standard contrastive loss with hard negatives. For a batch of N cases, we encode each query image and its K candidate texts, then compute a masked cross-entropy loss:

$$L_{clip} = -(1/N) \Sigma_i \log \left(\exp(\tau \cdot s(I_i, T_i^*)) / \Sigma_{j \in C_i} \exp(\tau \cdot s(I_i, T_j)) \right)$$

Here T_i^* is the correct candidate, C_i is the set of candidates excluding padding, and $\tau = 100$ is a temperature parameter. We use Adam with a learning rate of 1e-6 and train per subset for 300 steps with batch size 8.

3.3 Adaptive Hard-Negative Training

We propose an extension that reweights each case by the model's current confusion. For case i, let p_i be the softmax probability assigned to the correct label. We define a weight:

$$w_i = 1 + \alpha \cdot (1 - p_i)$$

The weighted loss is:

$$L_{adaptive} = (1/N) \Sigma_i w_i \cdot (-\log p_i)$$

The parameter α controls how strongly the weighting is applied. We test $\alpha = 0.5$ and $\alpha = 0.1$. In the $\alpha = 0.1$ variant, we additionally normalize the weights to have mean 1 so that the effective learning rate stays the same as in the fixed baseline.

3.4 SPEC + + Paired-Reasoning Diagnostic

To test whether fine-tuned models compose concepts, we construct SPEC++, a paired-reasoning benchmark. For each pair of subsets (for instance, size and count), we identify object classes that appear in both. For each shared object, we pair one size case with one count case. The model must correctly classify both the size and the count for the pair to succeed. We build 200 such pairs covering 60 object classes.

Evaluation reports three numbers. Size accuracy is the fraction of pairs where the size case is classified correctly. Count accuracy is the fraction where the count case is classified correctly. Joint accuracy is the fraction where both are correct. Random baselines are 33.3% for size, 11.1% for count, and 3.7% for joint.

4. Results

4.1 Baseline CLIP on SPEC

Table 1 reports the accuracy of unfine-tuned CLIP ViT-B/32 on the held-out test sets.

Table 1
Baseline CLIP on SPEC held-out test.

Subset	Accuracy	Random
absolute_size	45.67%	33.33%
relative_size	31.33%	33.33%
absolute_spatial	12.11%	11.11%
relative_spatial	28.75%	25.00%
existence	58.00%	50.00%
count	25.22%	11.11%
Mean	33.51%	—

Most subsets fall close to random. Only existence and count show gains over chance that are worth noting. This matches the SPEC paper's own finding that unfine-tuned CLIP struggles with fine-grained concepts.

4.2 Fixed Hard-Negative Fine-Tuning

Table 2 reports the fine-tuned model's accuracy on the same held-out test sets, and Fig. 1 presents the same comparison graphically.

Table 2
Fine-tuned CLIP on SPEC with fixed hard negatives.

Subset	Baseline	Fine-tuned	Gain
absolute_size	45.67%	95.33%	+ 49.66
relative_size	31.33%	98.00%	+ 66.67
absolute_spatial	12.11%	93.33%	+ 81.22
relative_spatial	28.75%	97.25%	+ 68.50
existence	58.00%	78.50%	+ 20.50
count	25.22%	69.56%	+ 44.34
Mean	33.51%	88.66%	+ 55.15

Fine-tuning raised accuracy on every subset. The largest gains appeared on the subsets where the baseline had been closest to random, which suggests the fine-tuning step teaches the model concepts it did not have before, as Fig. 1 makes visually clear.

4.3 Adaptive Hard-Negative Training

Table 3 reports the adaptive variants, and Fig. 2 compares all three training regimes side by side.

Table 3
Adaptive hard-negative training on SPEC held-out test.

Subset	Fixed	Adaptive v1 (α = 0.5)	Adaptive v2 (α = 0.1, norm.)
absolute_size	95.33%	73.00%	83.67%
relative_size	98.00%	54.00%	58.33%
absolute_spatial	93.33%	31.56%	27.89%
relative_spatial	97.25%	52.25%	65.00%
existence	78.50%	83.50%	85.50%
count	69.56%	67.00%	71.44%
Mean	88.66%	60.22%	65.31%

Both adaptive variants fall well short of the fixed baseline. On absolute*spatial and relative*size, the drop exceeds 30 percentage points, as shown in Fig. 2. The v2 variant improves on v1 by 5 points but remains 23 points behind. One partial exception is existence, where both adaptive variants modestly outperform fixed hard negatives. This may be because existence candidates are easier (K = 2) and stand to gain from downweighting easy examples.

4.4 SPEC + + Paired-Reasoning

Table 4 reports SPEC + + results, and Fig. 3 breaks down joint accuracy by object class.

Table 4
SPEC + + paired-reasoning results (n = 200 pairs).

Metric	Baseline CLIP	Fine-tuned	Random
Size accuracy	46.50%	89.00%	33.33%
Count accuracy	23.50%	83.50%	11.11%
Joint accuracy	10.00%	75.50%	3.70%

The fine-tuned model reaches 89% on size and 83.5% on count. If the two tasks were solved independently, the expected joint accuracy would be 0.89 × 0.835 = 74.3%. The observed joint accuracy is 75.5%, essentially the same. The model solves size and count as separate problems.

Per-object analysis, shown in Fig. 3, confirms this pattern. Across every object class, joint accuracy matches the product of the two individual accuracies within measurement noise. No object shows the pattern of interference (joint below product) that would indicate compositional failure, nor the pattern of reinforcement (joint above product) that would indicate compositional success.

5. Discussion

5.1 Why Adaptive Hard-Negative Mining Failed

Three factors explain why adaptive reweighting did not improve fine-grained training, and the training curves in Fig. 4 make each visible.

First, SPEC's candidate sets are uniformly hard by construction. The progressive data construction pipeline ensures that all K candidate texts in a case describe the same object at different values of the target attribute. No candidate is trivially easy, because all candidates share the object class. Adaptive reweighting assumes a distribution of difficulties, but SPEC's design flattens that distribution.

Second, reweighting changes the effective learning rate. Multiplying per-sample losses by weights between 1 and 1 + α changes the gradient scale from one batch to the next. With a fixed learning rate, this destabilizes training, which is visible in the fluctuating loss curves for absolute_spatial and count in Fig. 4. Our v2 variant partially corrects for this by normalizing weights but does not recover the fixed baseline.

Third, over-weighting the hardest examples can cause overfitting to ambiguous samples. In SPEC, ambiguous samples are those where the visual difference between candidate attributions is subtle. Over-focusing on those samples teaches the model to fit noise.

5.2 What SPEC + + Reveals

The independence of size and count fits a broader pattern reported in the VLM literature [4]: models learn concepts in isolation but do not compose them. SPEC's design deliberately isolates one concept per subset. That design is right for evaluating individual concepts, but it prevents SPEC from revealing compositional failures. A model that solves each concept on its own achieves high joint accuracy under SPEC's evaluation, which masks the fact that it does not compose, as Table 4 demonstrates.

This is not a flaw in SPEC. The benchmark was designed to test individual concepts and does so well. The community, however, should not mistake SPEC performance for evidence of compositional understanding. Benchmarks that require genuine composition, where a single image varies along two or more axes at once, are needed.

5.3 Limitations

Our study has three limitations.

We do not generate new multi-concept images. SPEC + + is a paired-reasoning benchmark that uses existing single-concept images. It tests whether a model can handle two concepts sequentially, but not whether it can handle them simultaneously within one image.

We fine-tune only CLIP ViT-B/32. Results may differ for larger models or different architectures.

Our adaptive method is one of several possible formulations. We tested a confusion-weighted loss. Other adaptive approaches may fare better.

5.4 Directions for Future Work

Three directions follow from these results.

Generate genuinely compositional benchmarks. The SPEC pipeline can be extended to produce images that vary along two axes at once, such as three large dogs or one small dog. This requires running the SDXL, segmentation, and inpainting pipeline with multi-attribute control. Our SPEC + + benchmark is a lower-cost approximation; the full version would provide stronger evidence.

Investigate methods that explicitly encourage compositional reasoning. Contrastive loss alone does not appear to induce compositionality. Auxiliary losses that penalize inconsistent predictions across concepts, or architectural changes that enforce structured representations, may be necessary.

Study the synthetic-to-real gap. SPEC's images are generated by SDXL [24], which introduces characteristic visual artifacts. Whether improvements measured on SPEC transfer to real photographs is an open question. Fine-tuned models may be overfitting to synthetic artifacts rather than learning the underlying concepts.

6. Conclusion

We reproduced the SPEC benchmark [7], fine-tuned CLIP [1] on it with fixed hard negatives, and proposed an adaptive hard-negative mining variant. The adaptive variant underperformed fixed hard negatives by 23 percentage points on average (Table 3). We traced the cause to SPEC's uniformly hard candidate sets, which leave no difficulty signal for adaptive weighting to exploit. We also introduced SPEC++, a paired-reasoning diagnostic, and showed that the fine-tuned model treats size and count as independent problems rather than composing them (Table 4).

Two lessons follow for fine-grained VLM evaluation and training. Uniform hard negatives are difficult to outperform when the data is curated to be uniformly hard. High accuracy on benchmarks like SPEC should not be read as evidence of compositional understanding, because the benchmark's single-concept design masks the compositional gap.

Future work should build benchmarks that require genuine multi-concept reasoning and methods that explicitly encourage compositionality.

Declarations

Competing interests. The authors declare no competing interests.

Funding. No funding was received for this work.

Ethics approval and consent to participate. Not applicable. This study did not involve human participants, human data, or animals; it is a computational study using the publicly available SPEC image-text benchmark.

Consent to publish. Not applicable. This manuscript does not contain any individual person's data in any form.

Clinical trial number: not applicable.

Author contributions. Silas Ayitey conceived the study, designed the experiments, implemented the code, ran the training and evaluation, and drafted the manuscript. Kwabena Owusu-Agyemang supervised the work, provided critical feedback on the methodology and results, and revised the manuscript. Both authors read and approved the final manuscript.

Data availability. The SPEC dataset used in this study is publicly available *at https://huggingface.co/datasets/wjpoom/SPEC [7].* The fine-tuned model weights, the SPEC++

paired-reasoning benchmark, and the training and evaluation code generated during the current study are available in the GitHub repository at*https://github.com/silasayi/spec-plus-plus*.

References

1. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Radford, A. (2021). Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning. pp. 8748–8763.
2. Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Proceedings of the 39th International Conference on Machine Learning. pp. 12888–12900.
3. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., & Wu, Y. (2022). CoCa: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*.
4. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., & Zou, J. (2023). When and why vision-language models behave like bag-of-words models, and what to do about it? In: International Conference on Learning Representations.
5. Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., & Ross, C. (2022). Winoground: Probing vision and language models for visio-linguistic compositionality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5238–5248.
6. Cascante-Bonilla, P., Shehada, K., Smith, J., Doveh, S., Kim, D., Panda, R., Varol, G., Oliva, A., Ordonez, V., Feris, R. S., & Karlinsky, L. (2023). Going beyond nouns with vision and language models using synthetic data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1835–1845.

7. Peng, W., Xie, S., You, Z., Lan, S., & Wu, Z. (2024). Synthesize, diagnose, and optimize: Towards fine-grained vision-language understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
8. Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., & Kiela, D. (2022). FLAVA: A foundational language and vision alignment model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15638–15650.
9. Jia, C., Yang, Y., Xia, Y., Chen, Y. T., Parekh, Z., Pham, H., Le, Q., Sung, Y. H., Li, Z., & Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In: Proceedings of the 38th International Conference on Machine Learning. pp. 4904–4916.
10. Weng, Z., Yang, X., Li, A., Wu, Z., Jiang, Y. G., & Open-VCLIP (2023). Transforming CLIP to an open-vocabulary video model via interpolated weight optimization. In: Proceedings of the 40th International Conference on Machine Learning. pp. 36275–36288.
11. Wu, Z., Weng, Z., Peng, W., Yang, X., Li, A., Davis, L. S., & Jiang, Y. G. (2024). Building an open-vocabulary video CLIP model with better architectures, optimization and data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
12. Zeng, Y., Zhang, X., & Li, H. (2021). Multi-grained vision language pre-training: Aligning texts with visual concepts. In: Proceedings of the 38th International Conference on Machine Learning.
13. Wang, J., Chen, D., Wu, Z., Luo, C., Zhou, L., Zhao, Y., Xie, Y., Liu, C., Jiang, Y. G., & Yuan, L. (2022). OmniVL: One foundation model for image-language and video-language tasks. In: Advances in Neural Information Processing Systems.
14. Parcalabescu, L., Cafagna, M., Muradjan, L., Frank, A., Calixto, I., & Gatt, A. (2021). VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics. pp. 825–835.
15. Wang, T., Lin, K., Li, L., Lin, C. C., Yang, Z., Zhang, H., Liu, Z., & Wang, L. (2023). Equivariant similarity for vision-language foundation models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1377–1387.
16. Zhao, T., Zhang, T., Zhu, M., Shen, H., Lee, K., Lu, X., & Yin, J. (2022). VL-Checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. arXiv preprint arXiv:2207.00221.
17. Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 815–823.
18. Oh Song, H., Xiang, Y., Jegelka, S., & Savarese, S. (2016). Deep metric learning via lifted structured feature embedding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4004–4012.
19. Xiong, L., Xiong, C., Li, Y., Tang, K. W., Liu, J., Bennett, P., Ahmed, J., & Overwijk, A. (2021). Approximate nearest neighbor negative contrastive learning for dense text retrieval. In: International

Conference on Learning Representations.

20. Cascante-Bonilla, P., Shehada, K., Smith, J. (2023). Going beyond nouns with vision and language models using synthetic data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision.
21. Doveh, S., Arbelle, A., Harary, S., Panda, R., Herzig, R., Schwartz, E., Kim, D., Giryes, R., Feris, R. S., Ullman, S., & Karlinsky, L. (2023). Teaching structured vision and language concepts to vision and language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
22. Chen, Y., Ma, Z., Zhang, Z., Qi, Z., Yuan, C., Shan, Y., Li, B., Hu, W., Qie, X., & Wu, J. (2023). ViLOM: Visual-language error modeling for image-text retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
23. Smith, J., Cascante-Bonilla, P., Arbelle, A., Kim, D., Panda, R., Cox, D., Yang, D., Kira, Z., Feris, R. S., & Karlinsky, L. (2023). Construct-VL: Data-free continual structured VL concepts learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
24. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Muller, J., Penna, J., & Rombach, R. (2023). SDXL: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952.
25. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L., & Grounding, D. I. N. O. (2023). Marrying DINO with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499.
26. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W. Y., & Dollar, P. (2023). Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision.
27. Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255.
28. Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., Zitnick, C. L., & Microsoft, C. O. C. O. (2014). Common objects in context. In: Proceedings of the European Conference on Computer Vision. pp. 740–755.
29. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., & Komatsuzaki, A. (2021). LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114.
30. Geirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*(11), 665–673.
31. Hendricks, L. A., & Nematzadeh, A. (2021). Probing image-language transformers for verb understanding. In: Findings of the Association for Computational Linguistics. pp. 3635–3644.
32. Liu, F., Emerson, G., & Collier, N. (2023). Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, *11*, 635–651.

33. Goel, S., Bansal, H., Bhatia, S. K., Rossi, R. A., Vinay, V., & Grover, A. (2022). CyCLIP: Cyclic contrastive language-image pretraining. In: Advances in Neural Information Processing Systems.
34. Li, C., Liu, H., Li, L. H., Zhang, P., Aneja, J., Yang, J., Jin, P., Lee, Y. J., Hu, H., Liu, Z., & Gao, J. (2022). ELEVATER: A benchmark and toolkit for evaluating language-augmented visual models. In: Advances in Neural Information Processing Systems.
35. Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T. Y., Cubuk, E. D., Le, Q. V., & Zoph, B. (2021). Simple copy-paste is a strong data augmentation method for instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Figures

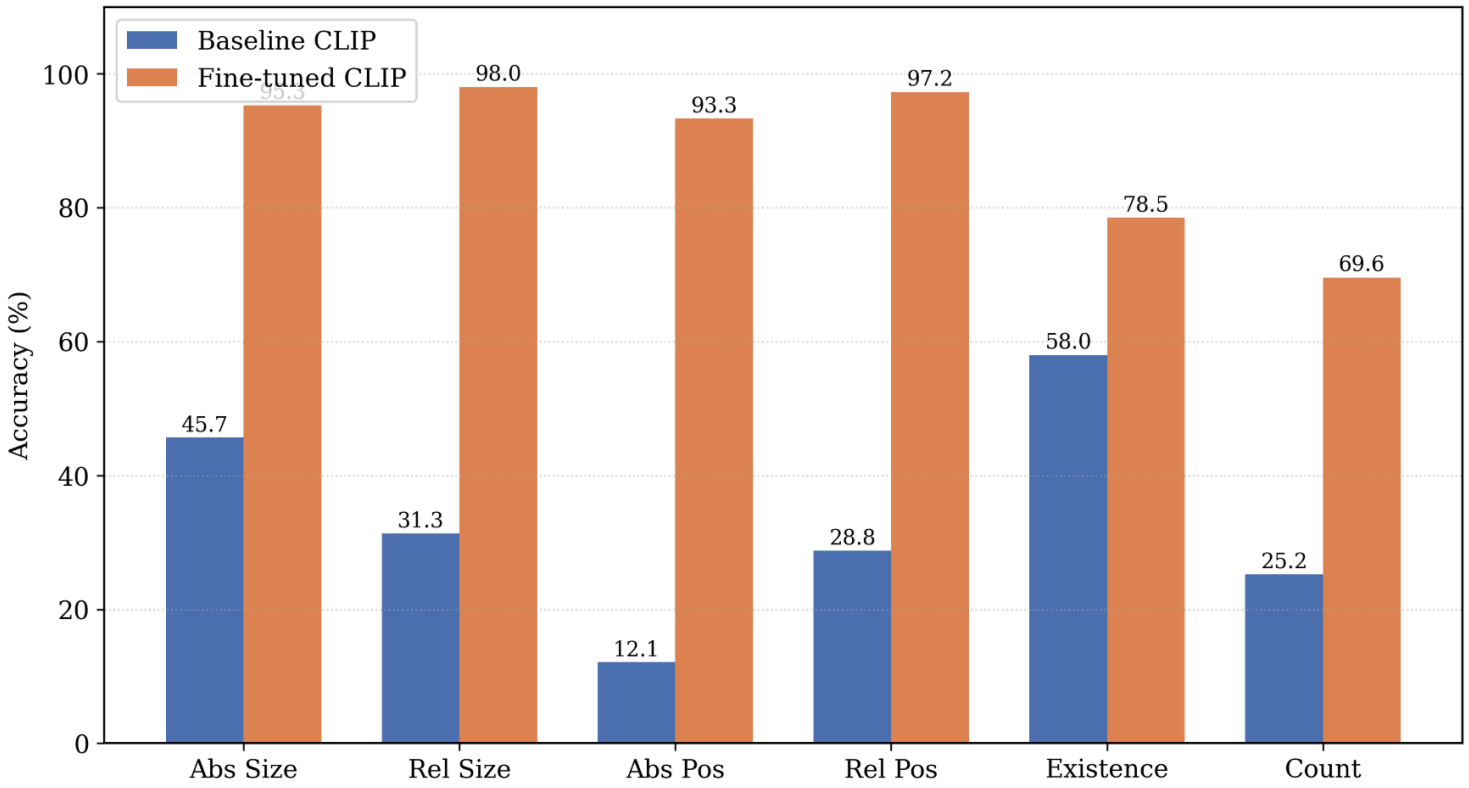


Figure 1

Baseline CLIP accuracy versus fine-tuned CLIP accuracy on SPEC held-out test data. Fine-tuning with fixed hard negatives improves every subset, with the largest gains on subsets where baseline accuracy was closest to random.

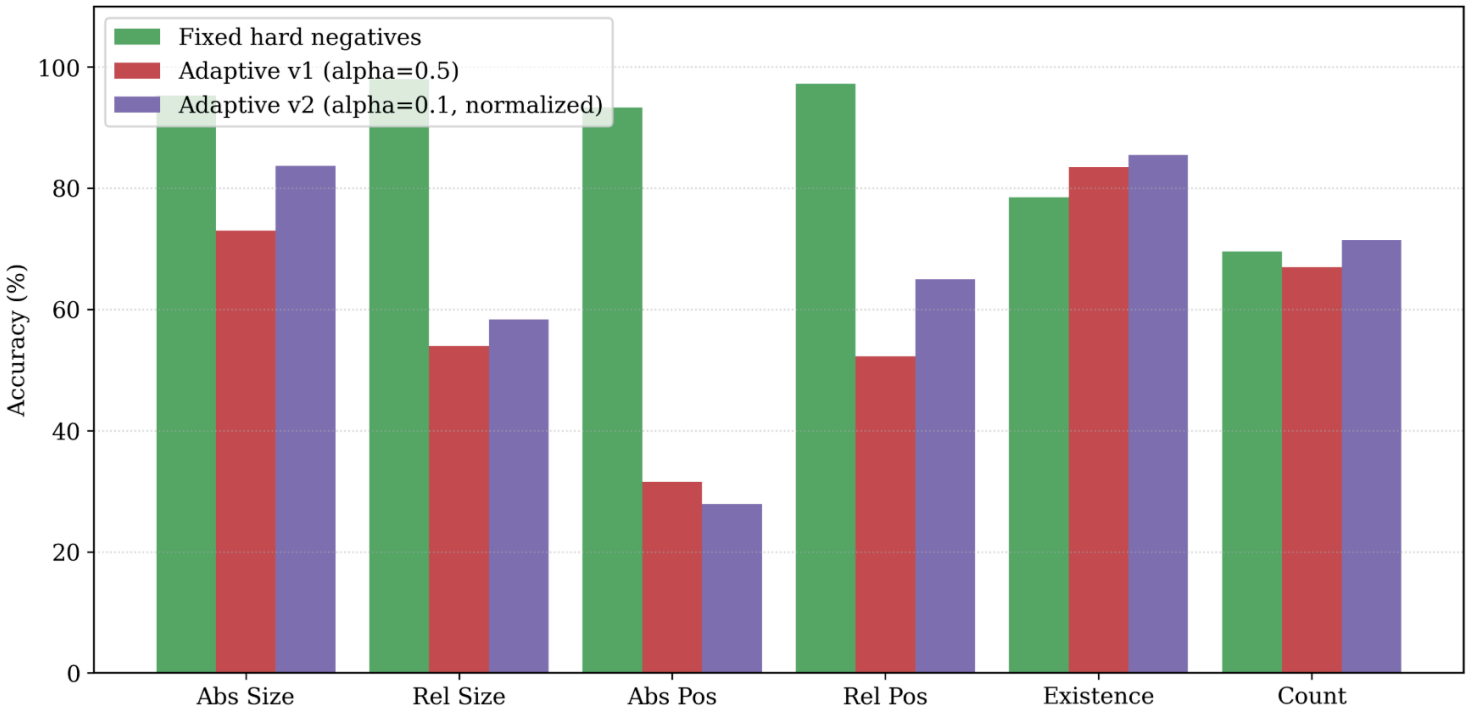


Figure 2

Comparison of fixed hard-negative training and two adaptive variants. Both adaptive variants fall short of the fixed baseline on every subset except existence.

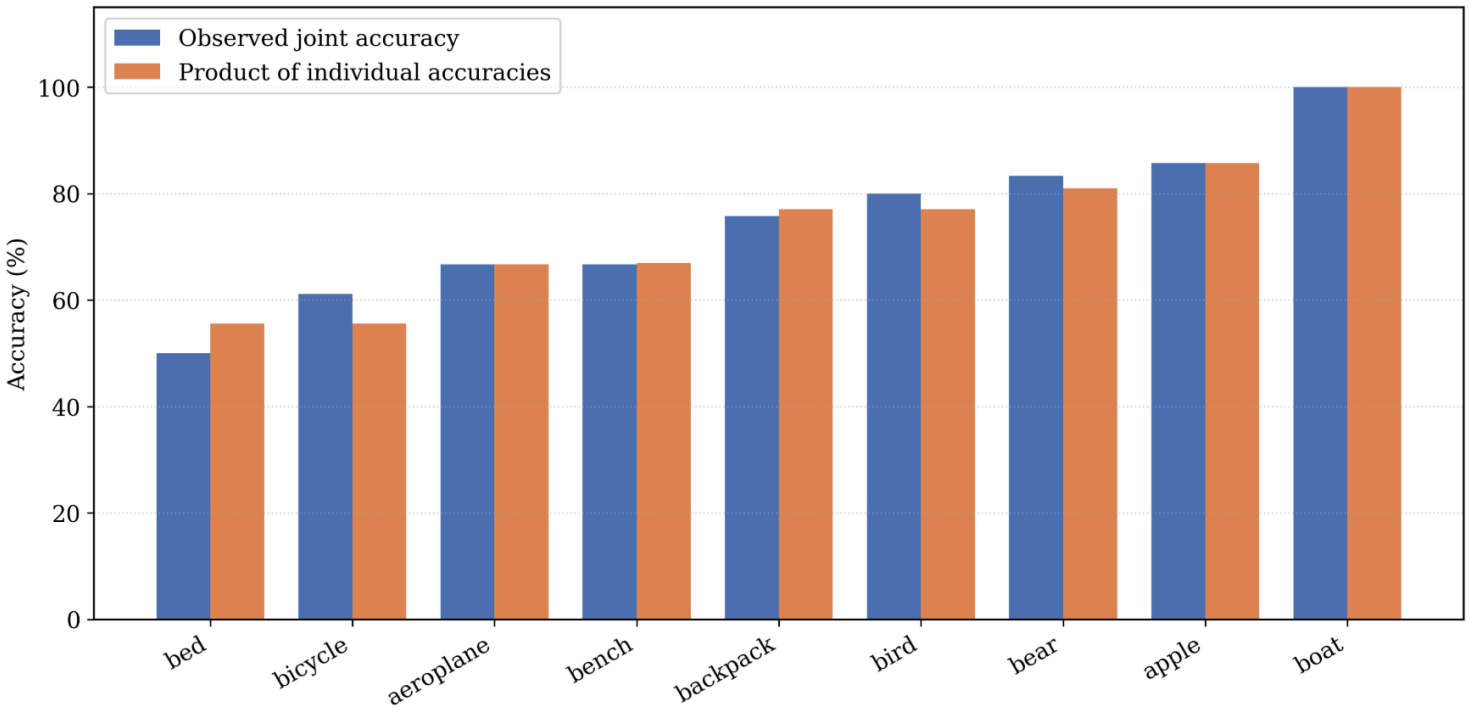


Figure 3

SPEC++ paired-reasoning results. For each object class, the observed joint accuracy closely tracks the product of the model's individual size and count accuracies, indicating that the model treats the two concepts independently.

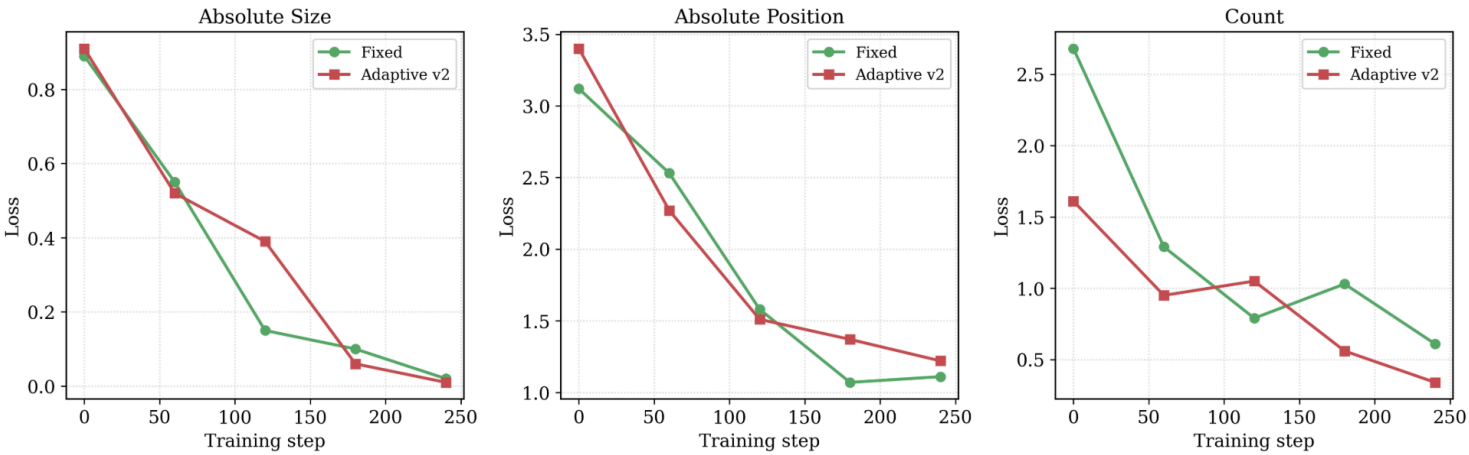


Figure 4

Training loss curves for fixed hard negatives and adaptive v2 on three subsets. The adaptive variant produces less stable loss trajectories, especially on absolute position and count.