

SegDebias: Test-Time Bias Mitigation for ViT-Based CLIP via Segmentation

Fangyu Wu Yujun Cai

Abstract

Vision language models such as CLIP have shown remarkable performance in zero shot classification, but remain susceptible to spurious correlations, where irrelevant visual features influence predictions. Existing debiasing methods often require access to training data and explicit group labels to perform fine-tuning or adjust embeddings, which limits their practicality in real-world settings. Test-time methods attempt to avoid this constraint, but many still depend on prior knowledge of dataset specific biases, limiting their generalizability in open set settings. In this work, we propose a test-time debiasing method for ViT based CLIP models that requires no additional training or assumptions of bias annotations. Our approach uses a pretrained segmentation model to isolate the target visual attribute, then adjusts the non target regions so that their embeddings are uniformly similar to all class specific text prompts. This procedure removes unintended bias signals from confounding visual regions while preserving the target attribute. Experiments on Waterbirds and CelebA show that our method outperforms existing test-time debiasing approaches in both group robustness metrics and Attention IoU. These results demonstrate the effectiveness of segmentation guided interventions for scalable and annotation free bias mitigation in vision language models.

Introduction

Recent advances in vision-language models (VLMs), such as CLIP [1], have enabled strong o-shot classification performance by aligning large-scale image and text pairs. However, emerging evidence shows that these models are vulnerable to spurious correlations, where unintended visual patterns influence predictions and lead to biased or unreliable behavior in real-world settings [2, 3, 4]. When classifying waterbirds versus landbirds, for example, CLIP often focuses on background elements (water or land) instead of the birds themselves. This misplaced attention leads to poor performance in minority groups that break these common associations, creating significant performance gaps between demographic subgroups in datasets such as CelebA [5], where attributes such as hair color become inappropriately tied to gender-specific facial features.

Existing solutions to this problem face substantial practical limitations. Training-based approaches such as FairCLIP [6] and Contrastive Adapters [7] require group labels and training data, resources frequently unavailable in real-world settings. While test-time methods like DebaisedPrompt [8] and BendVLM [9] avoid retraining costs, they still rely on explicit knowledge of bias-inducing attributes, severely limiting their usefulness in open-set environments where biases may be unknown or complex.

The core challenge lies in identifying and addressing these spurious correlations without prior knowledge of their nature. Through careful analysis of CLIP’s attention patterns, we discovered that these models frequently highlight image regions irrelevant to the classification task, focusing on water backgrounds when identifying birds, or on facial features

when classifying hair color. This attention misalignment directly causes the performance disparities between groups shown in Figure 1.

Based on these insights, we developed SegDebias, a straightforward yet effective test-time approach that addresses the root cause of bias by directly intervening in the visual attention space. Our method operates without any knowledge of dataset-specific biases or protected attributes, making it widely applicable across domains. Instead of trying to identify all possible spurious correlations, SegDebias uses semantic segmentation to isolate the target attribute (such as a bird in waterbird classification), then carefully adjusts the non-target regions to have equal similarity to all candidate text embeddings. This neutralization effectively removes their ability to bias the model’s prediction while preserving the important visual information from the target attribute.

Extensive experiments on Waterbirds [17] and CelebA [18] demonstrate that SegDebias substantially outperforms existing test-time debiasing methods, achieving the best worst-group accuracy while significantly reducing performance gaps between groups. By measuring Attention-IoU [18], we show that our approach successfully redirects CLIP’s attention toward semantically meaningful regions, providing a clear explanation for its improved fairness. To summarize, the primary contributions of our work are:

1. We introduce a segmentation-guided test-time debiasing method for ViT-based CLIP that requires no prior knowledge of dataset biases or protected attributes.
2. We use the Attention-IoU metric to reveal limitations in CLIP’s attention patterns and show how our method improves semantic alignment.
3. We demonstrate that our approach outperforms existing test-time debiasing methods across standard benchmarks in group robustness.

2 Related Work

Representation-Level Debiasing Techniques. Recent efforts to mitigate bias in vision-language models (VLMs) like CLIP [19] have predominantly focused on modifying the model’s output embeddings, either through fine-tuning or test-time adjustments. FairCLIP [24] introduces an optimal-transport-based approach, reducing the Sinkhorn distance between protected-group distributions to balance performance and fairness, but depends on labeled data and extra training. Likewise, FairerCLIP [9] projects CLIP’s image and text embeddings into a reproducing kernel Hilbert space (RKHS) and solves a closed-form optimization to remove bias. Contrastive Adapters [25] train lightweight MLP-based adapters inserted into both the image and text towers of CLIP to adjust their embeddings. In contrast, test-time only methods seek to correct biases at inference: DebaisedPrompt [9] adjusts CLIP’s input prompts to cancel known spurious directions, but requires explicit knowledge of dataset biases whereas Bend-VLM [9] adjusts candidate text embeddings by performing two-stage, per-input nonlinear debiasing: orthogonalizing to local attribute subspaces and equalizing similarity across protected groups. Although these test-time approaches advance CLIP’s fairness, they assume prior access to target datasets with known protected attributes, making them ill-suited to open-set, test-time scenarios where confounders are unknown or complex.

Attention-IoU: Bias Detection through Attention Maps In addition to these debiasing strategies, Attention-IoU [18] introduces an anatomy of bias inside the CLIP decision process by comparing model attention maps to ground truth masks or other attribute maps. Using Grad-CAM, the authors define real-valued Intersection-over-Union between attention

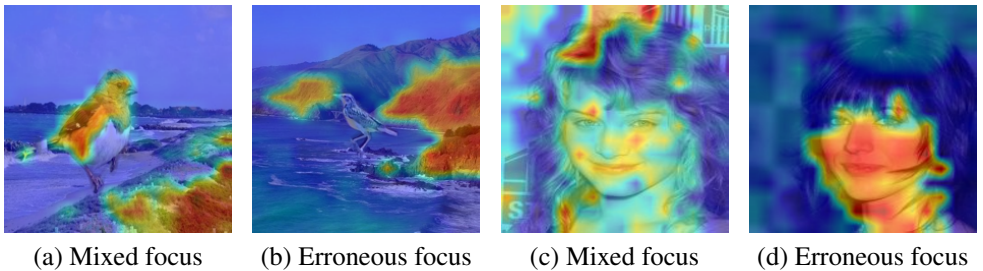


Figure 1: Attention-weight feature maps (Grad-CAM) from a ViT-based CLIP model on two binary classification tasks: waterbird vs. landbird (panels a–b) and blond-hair vs. dark-hair (panels c–d). Panels (a) and (c) show mixed attribution where both target and non-target regions are emphasized; panels (b) and (d) illustrate failure modes where attention is misdirected exclusively to irrelevant features. (highlighted colors from red to green indicate increasing levels of attention map values, while blue areas represent minimal or no focus).

and mask or between two attention maps, yielding the mask score and heatmap score respectively. In Waterbirds [14], the mask score accurately tracks background-induced bias, while in CelebA [13], the framework reveals that classifiers often rely on gendered facial features (e.g., eyes, nose) when predicting attributes like Wearing Lipstick or Blond Hair, even when those features are not directly relevant, highlighting spurious correlations beyond what label statistics alone can reveal. These findings reveal that even without output errors, CLIP’s internal focus can betray biased shortcuts, suggesting that segmentation-guided interventions at test-time could potentially steer attention back to semantically relevant regions.

3 Method

3.1 Motivation

As discussed in Section 1, Figure 1 shows that the ViT encoder in CLIP [5] often overweights spurious, non-target regions such as background textures or contextual cues when computing visual embeddings. In some cases, even when the target object (e.g., a bird) is masked out, the background alone can yield high cosine similarity to an incorrect class prompt. This behavior suggests that confounding factors, although not explicitly modeled, can disproportionately influence CLIP’s zero-shot predictions.

To quantitatively analyze the influence of non-target visual regions on CLIP’s zero-shot predictions, we conducted analysis on two benchmark datasets: Waterbird [14] and CelebA [13] (see Section 4.1 for more details). In Waterbirds, the task is to classify images as “waterbird” or “landbird”; in CelebA, the goal is to classify hair color as either “blond” or “dark.” For each dataset, we segmented images into two regions: the target attribute (bird or hair) and the non-target region (background or face). We then passed the non-target regions through CLIP’s visual encoder to obtain embeddings.

We computed the cosine similarity difference between these non-target embeddings and the class text prompts as:

$$\Delta_{\text{non-target}} = \text{sim}(E_{\text{non-target}}, E_{\text{prompt}_1}) - \text{sim}(E_{\text{non-target}}, E_{\text{prompt}_2}) \quad (1)$$

and similarly for the full image embeddings:

$$\Delta_{\text{full}} = \text{sim}(E_{\text{full-image}}, E_{\text{prompt}_1}) - \text{sim}(E_{\text{full-image}}, E_{\text{prompt}_2}) \quad (2)$$

Here, E_{prompt_1} and E_{prompt_2} correspond to the candidate text embeddings for each class. Plotting $\Delta_{\text{non-target}}$ against Δ_{full} across the dataset reveals a consistent positive correlation (as shown in Figure 2), suggesting that bias in non-target regions significantly influences the final prediction. Specifically, if the non-target region alone aligns more closely with a given class prompt, the full image embedding also tends to exhibit higher cosine similarity with the same class. These results highlight how contextually irrelevant regions can contaminate CLIP’s zero-shot predictions.

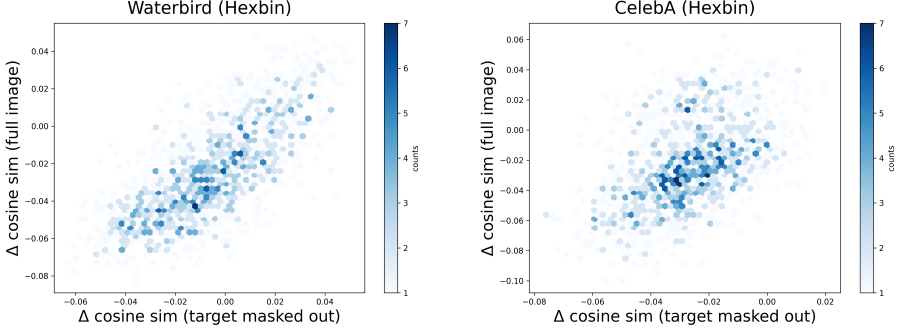


Figure 2: Correlation between cosine similarity differences for non-target regions (x-axis: $\Delta_{\text{non-target}}$) and full images (y-axis: Δ_{full}) with respect to two candidate text embeddings. Each point corresponds to an image randomly sampled from the dataset (1,500 total), illustrating how much the non-target region alone biases the prediction in the same direction as the full image across the Waterbirds [17] and CelebA [13] datasets.

To mitigate the bias introduced by non-target regions, we propose a semantic region-guided strategy that isolates the target attribute (e.g., bird or hair), suppresses irrelevant regions, and recombines the features to produce more robust and semantically grounded predictions.

3.2 SegDebias Framework

SegDebias is a test-time, segmentation-guided debiasing pipeline that operates on any zero-shot classification task with an arbitrary set of text prompts $\{\tau_i\}_{i=1}^N$. Our approach proceeds in four steps:

- 1. Target Attribute Selection** We define the *target visual attribute* to be the semantic region most directly corresponding to the class labels (e.g. the *bird* region in Waterbird vs. Landbird, or the *hair* region for Blond Hair vs. Dark Hair). By focusing on target regions, we avoid the combinatorial complexity of enumerating all potential confounding factors.
- 2. Promptable Segmentation.** To spatially isolate the target attribute identified in Step 1, we employ a two-stage segmentation process guided by vision-language models. First, we use a grounding detector such as Grounding DINO [17] with the prompt “Photo of a $\langle \text{attribute} \rangle$ ” to generate a coarse bounding box around the relevant region. Then, within this region, we apply a promptable segmentation model to extract a fine-grained binary mask M_i that captures the target area. The complement of this mask, M_i^c , represents the non-target or background region and is used in subsequent debiasing steps.

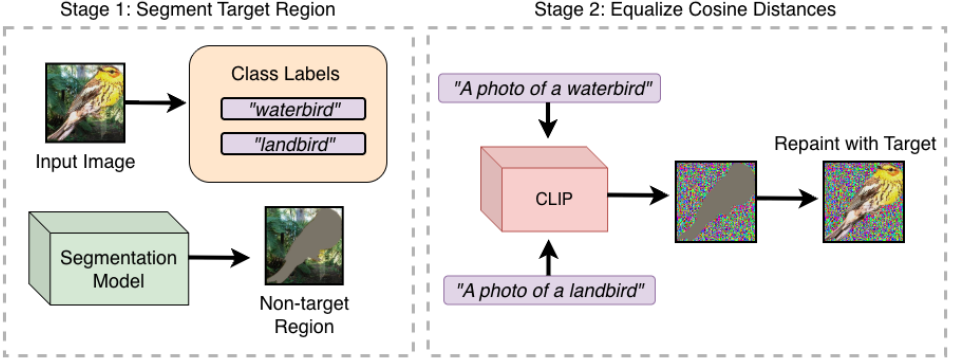


Figure 3: Overview of the proposed debiasing pipeline for zero-shot image classification. Given an input image and associated candidate text embeddings, we identify the target attribute (e.g., bird), segment it from the image, and optimize the background (non-target attributes) to have equal cosine distances with all text embeddings. The target attribute is then repainted into the debiased background and passed to CLIP for final zero-shot prediction.

3. Non-Target Neutralization. To reduce residual bias from the non-target region $M_{\bar{t}}$, we retain the original image but constrain updates to this region only. Specifically, we compute the CLIP visual embedding from the partially masked image, $z_v = f_{\text{CLIP}}(x')$, $x' = x \odot M_{\bar{t}}$, and introduce a small perturbation δ supported only on $M_{\bar{t}}$ to decorrelate the non-target region from any specific class. Let $\{z_{t_i}\}$ denote the text embeddings for each prompt τ_i . We then optimize δ to equalize the cosine similarity between the perturbed visual embedding and all candidate text embeddings:

$$\min_{\delta} \sum_{i < j} [\cos(z_v + \delta, z_{t_i}) - \cos(z_v + \delta, z_{t_j})]^2 \quad \text{s.t. } \text{supp}(\delta) \subseteq M_{\bar{t}}. \quad (3)$$

This procedure ensures that the non-target visual attributes do not favor any particular class, effectively neutralizing spurious associations.

4. Reconstruction and Zero-Shot Classification. Finally, we reconstruct the full image by restoring the original target region and blending it with the debiased background. Let $x' = x \odot M_{\bar{t}}$ denote the background portion of the image and δ the optimized perturbation supported on $M_{\bar{t}}$. The final image is given by

$$\tilde{x} = (x \odot M_t) + (x' + \delta) \odot M_{\bar{t}}. \quad (4)$$

We then compute the CLIP embedding $f_{\text{CLIP}}(\tilde{x})$ and perform zero-shot prediction via $\arg \max_i \cos$

4 Experiments

4.1 Datasets

We evaluate **SegDebias** on two widely used datasets for studying spurious correlations in zero-shot classification: Waterbirds [14] and CelebA [13]. These benchmarks are commonly adopted in recent fairness and debiasing research [4, 22], and we follow the same zero-shot

evaluation setup as prior test-time debiasing work [3, 4].

Waterbirds The Waterbirds dataset [17] is a synthetic benchmark constructed from CUB-200-2011 by overlaying bird images onto either water or land backgrounds. The classification task is to distinguish between *waterbird* and *landbird*, but the background is spuriously correlated with the bird class, for example, most waterbirds appear over water. The smallest “worst-group” (waterbirds on land) comprises only 2% of the training set, making the dataset a standard benchmark for evaluating worst-group robustness.

CelebA CelebA [18] is a large-scale facial attribute dataset with over 200,000 celebrity images annotated with 40 binary attributes. We focus on the binary hair color classification task, distinguishing *blond hair* from *non-blond hair*. Prior work [3, 4, 18] has shown that CLIP often relies on gendered facial features (e.g., eyes or nose) when predicting hair color, leading to performance disparities across subgroups (e.g., blond females vs. males). We follow the standard train/validation/test splits used in FairerCLIP [4] and DebaisedPrompt [3].

4.2 Implementation

We evaluate **SegDebias** under the same test-time zero-shot protocol used by DebaisedPrompt [3] and FairerCLIP [4] on the Waterbirds [17] and CelebA [18] benchmarks. For Waterbirds, the target attribute is *bird*, and we use the prompts “*A photo of a waterbird*” and “*A photo of a landbird*”. For CelebA, the target attribute is *hair*, and we use “*A photo of a celebrity with blond hair*” versus “*... with dark hair*”. These minimal templates are fully compatible with zero-shot inference and focus the model on semantically relevant regions. All images are encoded using OpenAI’s CLIP ViT-L/14. To verify model-agnostic effectiveness, we also validate on ViT-B/16 and ViT-B/32. For spatial localization, we use Grounding DINO [19] with the prompt “Photo of a $\langle$ attribute $\rangle$ ” to obtain a coarse bounding box. To extract fine-grained masks, we use SAM-ViT-B [20] for Waterbirds, and a SegFormer-B5 face parsing model [21] for CelebA.

We iteratively optimize a perturbation δ over the non-target region $M_{\bar{t}}$ to equalize cosine similarity across all text embeddings, stopping when the maximum pairwise difference falls below 0.001. All experiments are conducted in PyTorch on a single NVIDIA A100. Our perturbation-based debiasing procedure converges in under 70 ms on average. We report all performance metrics over three random seeds.

4.3 Group Robustness

We compare SegDebias against approaches in both categories mentioned in Section 1. **Training-based methods (require labels or embedding fine-tuning):** (1) ERM Linear [22]: a linear classifier trained with empirical risk minimization (ERM) on CLIP embeddings. (2) ERM Adapter [3]: a contrastive adapter trained via ERM. (3) WiSE-FT [23]: ensembling a fine-tuned ERM linear head with the original CLIP zero-shot classifier. (4) DFR (Up) [4]: deep feature reweighting using upsampling. (5) CA [24]: a contrastive adapter trained on group-balanced data to improve robustness. (6) FairerCLIP [4]: a projection-based method that debiases CLIP embeddings in RKHS to mitigate the influence of spurious relations. **Test-time only methods (no additional training):** (7) Zero-shot: standard CLIP zero-shot classification. (8) Orth-Proj [3]: removes spurious subspaces via orthogonal projection. (9) Orth-Cali [3]: orthogonal calibration with regularization to align prompt embeddings. (10)

BendVLM [10]: applies two-step, test-time debiasing by orthogonalizing to local attribute subspaces and equalizing similarity to reference images. We use gender and background as debiasing attributes for CelebA [10] and Waterbirds [10], respectively.

Dataset	Waterbirds			CelebA		
	WG (% $\uparrow$)	Avg (% $\uparrow$)	Gap ($\downarrow$)	WG (% $\uparrow$)	Avg (% $\uparrow$)	Gap ($\downarrow$)
<i>Training-based methods</i>						
ERM Adapter	78.4	97.8	19.4	36.7	94.2	57.5
WiSE-FT	65.9	97.6	31.7	80.0	87.4	7.4
DFR (Up)	65.9	96.1	30.2	83.7	91.2	7.5
CA	86.9	96.2	9.3	84.6	90.4	5.8
FairerCLIP	78.1	85.1	7.1	86.1	88.0	1.9
<i>Test-time only methods</i>						
Zero-shot	45.3	84.4	39.1	72.8	87.6	14.9
Orth-Proj.	61.4	<u>86.4</u>	25.0	71.1	<u>87.0</u>	15.9
Orth-Cali.	<u>68.8</u>	84.5	15.7	<u>76.1</u>	86.2	<u>10.1</u>
BendVLM	64.3	82.1	17.8	70.9	84.2	13.3
SegDebias (Ours)	71.6	88.2	<u>16.6</u>	81.6	85.1	3.5

Table 1: Group robustness of VLMs with various debiasing methods using the ViT-L/14 backbone. Best and second-best results are shown in **bold** and underlined, respectively.

Evaluation Metrics. We adopt metrics commonly used in group robustness studies, following prior work [10]: Average Accuracy (Avg), Worst-Group Accuracy (WG), and the Gap. **Avg** reports the mean accuracy across all examples. **WG** denotes the precision of the worst performing group, typically the smallest or most underrepresented subgroup in the data set. **Gap** is defined as the difference between average and worst-group accuracy ($\text{Gap} = \text{Avg} - \text{WG}$), and quantifies disparity in performance across groups.

Empirical Results As shown in Table 1, SegDebias achieves the highest worst-group accuracy on both Waterbirds and CelebA, while also substantially narrowing the Avg–WG gap. In some cases, it even outperforms training-based methods, despite requiring no retraining or prior knowledge of spurious correlations. In CelebA, the overrepresentation of blond-haired female subjects in the test set (49%) contributes to the slight decrease in average accuracy across all test-time debiasing methods, reflecting the removal of spurious shortcuts exploited by the zero-shot model.

Backbone	Method	WG (% $\uparrow$)	Avg (% $\uparrow$)	Gap ($\downarrow$)
ViT-B/16	Zero-shot	25.8	79.8	54.0
	SegDebias	36.4	74.8	38.4
ViT-B/32	Zero-shot	41.5	69.1	27.6
	SegDebias	50.2	61.1	10.9

Table 2: Group robustness metrics of SEGDEBIAS vs. zero-shot CLIP on the Waterbirds dataset using ViT-B/16 and ViT-B/32 backbones.

ViT Backbones To assess the robustness of SegDebias across different patch resolutions, we

evaluated two CLIP backbones, ViT-B/16 and ViT-B/32, on the Waterbirds dataset (Table 2). In both settings, SegDebias consistently improves worst-group accuracy and substantially narrows the accuracy gap, highlighting its effectiveness across backbone variants.

4.4 Attention-IoU Evaluation

Attention Map Extraction and Evaluation. To assess the alignment between CLIP’s ViT encoder attention and semantically relevant regions, we extract attention maps and evaluate them against ground-truth masks using the Attention-IoU metric by Serianni et al. [18].

In ViT[16], the CLS token summarizes global context, and its attention reveals which regions influence the model’s prediction. We extract an attention map $\hat{A} \in [0, 1]^{G \times G}$ by collecting the attention weights from the CLS token to all patch tokens at each transformer block. These weights are averaged across all heads and layers, then reshaped into a spatial $G \times G$ grid and normalized to unit range. The resulting map $\hat{A}$ highlights regions most influential to the final CLS embedding. Given a ground-truth binary mask M_t for the target attribute, we define the predicted attention region as the smallest set of patches that cumulatively capture 50% of the total attention mass. The Attention-IoU score is then computed as:

$$\text{Attention-IoU} = \frac{|\{p : \hat{A}(p) \geq \tau\} \cap M_t|}{|\{p : \hat{A}(p) \geq \tau\} \cup M_t|}, \quad \tau = \min\left\{\tau' : \sum_{p: \hat{A}(p) \geq \tau'} \hat{A}(p) \geq 0.5\right\}. \quad (5)$$

This single-threshold formulation yields a concise measure of spatial overlap between the model’s high-attention regions and the true semantic region of interest.

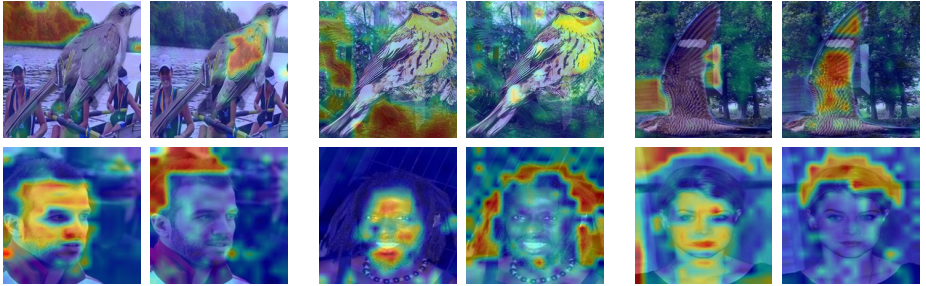


Figure 4: Comparison of attention maps before (left) and after (right) applying SegDebias on two tasks. Top row: three pairs of waterbird examples; bottom row: three pairs of CelebA hair examples. In each pair, vanilla CLIP’s attention (left) tends to scatter over confounding background or depiction biases, whereas SegDebias’s attention (right) aligns more closely with the true semantic region, yielding higher interpretable focus.

Results Table 3 reports Attention-IoU on Waterbirds and CelebA for both zero-shot CLIP and SegDebias. By neutralizing non-target features before recombining with the target region, SegDebias refocuses the model’s attention onto the semantic attribute of interest, yielding a marked increase in Attention-IoU and demonstrating more faithful and interpretable behavior as shown in Figure 4.

	Waterbirds	CelebA
Zero-shot	70.0%	52.9%
SegDebias	79.9%	66.8%

Table 3: Attention-IoU mask score with Clip ViT-L14.

4.5 Ablation Studies

Metric	Target-Only	Noise-Filled	Random Repaint	SegDebias
WG (% $\uparrow$)	56.5	33.9	64.3	71.6
Avg (% $\uparrow$)	88.9	60.2	86.2	88.2
Gap (% $\downarrow$)	32.4	18.3	21.9	16.6

Table 4: Ablation on different variant inputs, experimented with Waterbirds [17] dataset.

To demonstrate the effect of non-target attribute debiasing, we compare four input variants on the Waterbirds dataset in a single pass: (1) **Target-Only Input** retains only the segmented bird region, masking all background; (2) **Noise-Filled Background** restores bird pixels but fills the non-target region with Gaussian noise ($\theta=0.1$) without further optimization; (3) **Random Repaint** repaints the target region onto an unoptimized, randomly-sampled noisy background; and (4) **Full SegDebias**. As shown in Table 4, while both noise-filled and random-repaint variants reduce spurious cues, the unoptimized background still introduces residual bias that can favor one class. In contrast, SegDebias explicitly balances background similarity, leading to the best group fairness metrics in both WG and Gap.

5 Discussion

Effectiveness on Transformer Based Backbones. SegDebias demonstrates strong empirical gains when applied to ViT based CLIP models, aligning with the increasing use of transformers in vision and language representation learning [8, 17]. Unlike convolutional networks, which operate through local receptive fields, transformer models tokenize the image into patches and process them using global self attention. This structure enables ViTs to integrate spatially distant information and attend to object level context across the entire image, making them well suited for masking or region specific interventions. SegDebias leverages this property by neutralizing non target patches while preserving discriminative attributes, which improved group robustness. Our results further support observations that ViTs exhibit a form of semantic grounding through attention that can be modulated at test time [2, 16]. Because each patch embedding contributes globally to the final representation, masking a region does not necessarily break feature continuity. This contrasts with convolutional pipelines, where localized masking can fragment spatial dependencies. As ViTs [5] continue to serve as the foundation for modern vision language systems, segmentation guided strategies like ours offer a promising path for addressing fairness and robustness in a model agnostic and non intrusive way.

6 Conclusion

We propose SegDebias, a segmentation-guided test-time debiasing method for ViT-based CLIP models that requires no representation tuning or prior knowledge of the test distribution. By isolating target attributes and suppressing spurious regions, SegDebias enhances group robustness and attention alignment over existing test-time debiasing methods. Future work could extend SegDebias to abstract attributes and complex biases like multi-attribute entanglement. Integrating self-supervised localization may further enable the model to dynamically verify whether attended regions are semantically relevant during test time.

References

- [1] Kush Bansal, Yujia Xie, Yifan Zhang, and Z. Morley Mao. Cleanclip: Mitigating data poisoning attacks in multimodal contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 112–123, 2023.
- [2] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [3] C.-Y. Chuang, M. S. Li, and J. P. Wang. Debiasing vision-language models via biased prompts. *arXiv:2302.00070v2*, 2023.
- [4] S. Dehdashtian, S. S. Wang, V. Naresh Boddeti, and R. Kumar. Fairerclip: Debiasing clip’s zero-shot predictions using functions in rkhs. *International Conference on Learning Representations (ICLR)*, 2024.
- [5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.
- [6] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision (IJCV)*, 132(2):581–595, 2024.
- [7] W. Gerych, H. Zhang, K. Hamidieh, and M. Sharma. Bend-vlm: Test-time debiasing of vision-language embeddings. *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.
- [8] Jonghwan Kim, Seong Joon Oh, and Bohyung Han. Exposing and mitigating spurious correlations for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023.
- [9] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer reweighting for class imbalance. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2022.
- [10] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023.
- [11] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, 2022.
- [12] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2024.

- [13] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015.
- [14] Yan Luo, Min Shi, Muhammad Osama Khan, Muhammad Muneeb Afzal, Hao Huang, Shuaihang Yuan, Yu Tian, Luo Song, Ava Kouhana, Tobias Elze, Yi Fang, and Mengyu Wang. Fairclip: Harnessing fairness in vision-language learning. 2024.
- [15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.
- [16] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- [17] Shiori Sagawa, Pang Wei Koh, Tatsunori Hashimoto, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.
- [18] Aaron Serianni, Tyler Zhu, Olga Russakovsky, and Vikram V. Ramaswamy. Attention-iou: Examining biases in celeba using attention maps, 2025. arXiv:2503.19846v2.
- [19] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7959–7971, 2022.
- [20] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- [21] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [22] X. Zhang and C. Ré. Contrastive adapter: Enhancing group robustness in vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.