BiPrompt: Bilateral Prompt Optimization for Visual and Textual Debiasing in Vision-Language Models

Sunny Gupta[1], **Shounak Das**[2], **Amit Sethi**[2]

[1]Koita Centre for Digital Health, IIT Bombay, Mumbai
[2]Department of Electrical Engineering, IIT Bombay, Mumbai
sunnygupta@iitb.ac.in, 21D070068@iitb.ac.in, asethi@iitb.ac.in

Abstract

Vision language foundation models such as CLIP exhibit impressive zero-shot generalization yet remain vulnerable to spurious correlations across visual and textual modalities. Existing debiasing approaches often address a single modality either visual or textual leading to partial robustness and unstable adaptation under distribution shifts. We propose **a bilateral prompt optimization framework (BiPrompt)** that simultaneously mitigates non-causal feature reliance in both modalities during test-time adaptation. On the visual side, it employs structured attention-guided erasure to suppress background activations and enforce orthogonal prediction consistency between causal and spurious regions. On the textual side, it introduces balanced prompt normalization, a learnable re-centering mechanism that aligns class embeddings toward an isotropic semantic space. Together, these modules jointly minimize conditional mutual information between spurious cues and predictions, steering the model toward causal, domain invariant reasoning without retraining or domain supervision. Extensive evaluations on real-world and synthetic bias benchmarks demonstrate consistent improvements in both average and worst-group accuracies over prior test-time debiasing methods, establishing a lightweight yet effective path toward trustworthy and causally grounded vision-language adaptation.

Introduction

Vision-Language Models (VLMs), such as CLIP (Radford et al. 2021), demonstrate remarkable zero-shot generalization by learning from massive image-text datasets. However, their reliability in real-world, out-of-distribution (OOD) settings remains a critical concern.

A fundamental weakness of VLMs is their reliance on decision shortcuts such as spurious correlations and background context rather than true causal features (Fan et al. 2023). This severely hinders generalization and trustworthiness, leading to unpredictable failures, such as classifying a spider on a beach as a crab. This reliance on spurious features causes incorrect behavior on novel data combinations.

To improve VLM robustness, region-aware methods guide the model's focus by fine-tuning or altering its architecture (Sun et al. 2024). However, these approaches are often costly and can harm generalization. A more lightweight alternative is test-time prompt tuning, which adapts a model without modifying its weights.

A prominent example is the Spurious Feature Eraser (SEraser) (Ma et al. 2025), which operates on the insight that VLMs possess causal features but are misled by spurious signals. SEraser teaches the model to *ignore* these signals by optimizing a prompt to maximize prediction entropy on auxiliary (spurious) images, forcing the model to rely on causal features.

However, methods like SEraser (Ma et al. 2025) have two key limitations. First, their reliance on random visual erasure or simplistic segmentation can be unstable, as causal features may be inadvertently removed. Second, they focus exclusively on visual bias, ignoring linguistic biases in the static textual prompts, where strong class-name priors can skew predictions.

To overcome these challenges, we propose **Balanced-Prompt SEraser (BiPrompt)**, an enhanced test-time adaptation framework that jointly mitigates both visual and textual biases. Our two primary contributions are:

- **Balanced Prompt Normalization**: Learns an isotropic representation for text embeddings at test-time to reduce linguistic bias.
- **Structured Erasure**: Uses attention maps to disentangle causal (foreground) from spurious (background) features, optimizing a prompt to enforce consistency on causal regions while promoting orthogonality to spurious ones.

By addressing biases from both modalities, BiPrompt achieves more robust OOD performance without model retraining, contributing to more trustworthy foundation models for real-world applications.

Related Work

Efforts to improve the out of distribution reliability of Vision Language Models have largely followed two paths. The first involves region aware methods that explicitly guide visual focus, ranging from simple background masking (Liang et al. 2023) to interactive prompting, such as circling the foreground (Shtedritski, Rupprecht, and Vedaldi 2023) or using mask contours (Yang et al. 2023). More advanced approaches such as Alpha CLIP (Sun et al. 2024) extend CLIP with an additional alpha channel, often derived from

Accepted at the AAAI 2026 Workshop AIR-FM, Assessing and Improving Reliability of Foundation Models in the Real World.

segmentation models such as SAM, to provide pixel wise information about foreground and background regions. While effective, these methods are typically heavyweight, requiring architectural changes or costly fine tuning on large region text datasets, limiting their out of the box applicability for test time adaptation.

A second, more lightweight approach is test-time prompt tuning, which adapts the model to a new task without modifying its weights. A prominent example is Test-time Prompt Tuning (TPT) (Shu et al. 2022), which optimizes a prompt by minimizing the marginal entropy across augmented views of an image, thereby filtering out noisy augmentations based on low confidence. However, this relies on the critical assumption that spurious features will produce low-confidence predictions, which is not always true; a strong spurious feature, like a water background, can erroneously lead to a high-confidence prediction for a land-based object (Ma et al. 2025). Other methods leverage language models to refine embeddings, like ROBOSHOT (Adila et al. 2023), or debias the model by measuring correlations with biased prompts and applying orthogonal projection (Chuang et al. 2023). Our work builds directly on the insights of Spurious Feature Eraser (SEraser) (Ma et al. 2025), which proposed to erase spurious features by maximizing entropy on them. We address these limitations by introducing a structured, attention-guided erasure mechanism and a novel prompt normalization technique to mitigate linguistic bias in both visual and textual modalities.

Methodology

Deep vision–language models (VLMs) such as CLIP have demonstrated remarkable zero-shot generalization across diverse visual domains. However, when deployed in *out-of-distribution (OOD)* or *spurious-bias* settings, these models tend to rely on superficial correlations rather than true causal cues. For instance, a model might associate "water" backgrounds with "boats" or "hospital logos" with particular pathologies, resulting in degraded reliability under distribution shifts.

Formally, given an image x and a textual class prompt t_c, a VLM encodes them into a shared embedding space, where the similarity $s(f_v(x), f_t(t_c))$ determines the prediction. Under distributional shift, the visual encoder f_v becomes entangled with spurious dimensions z_s (e.g., background, texture), while causal dimensions z_c (object or semantic features) remain underrepresented. Our objective is to disentangle and suppress spurious features z_s while preserving causal representations z_c, without requiring retraining or access to domain labels.

Given a pretrained CLIP model, predictions are obtained as:

$$p(y|x) = \mathrm{softmax}(\tau \cdot \mathrm{sim}(f_v(x), f_t(t_y)))\,, \tag{1}$$

where τ is a learnable temperature parameter. An auxiliary sample $\tilde{x}$ is produced via random erasure, and a KL-divergence regularization aligns their predictions:

$$\mathcal{L}_{\text{SEraser}} = D_{\text{KL}}\big(p(y|x) \,\|\, p(y|\tilde{x})\big). \tag{2}$$

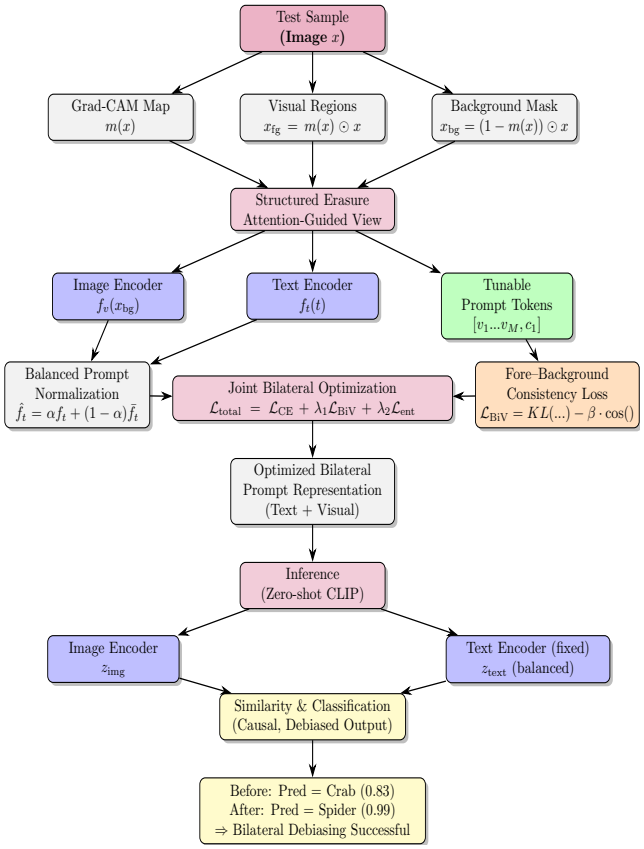

Figure 1: Framework of **BiPrompt**. Given a test sample, BiPrompt performs bilateral debiasing by jointly optimizing visual and textual representations. Structured attention-guided erasure suppresses spurious visual cues, while balanced prompt normalization aligns textual embeddings. The optimized bilateral prompt is then used for zero-shot inference, yielding causal and robust predictions under distribution shift.

To prevent degenerate uniform predictions, SEraser introduces an entropy regularizer:

$$\mathcal{L}_{\text{ent}} = -\frac{1}{C}\sum_{c=1}^{C} p(y=c|x)\log p(y=c|x), \tag{3}$$

yielding the total objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_1 \mathcal{L}_{\text{SEraser}} + \lambda_2 \mathcal{L}_{\text{ent}}. \tag{4}$$

This process enforces invariance across erased views, compelling the model to focus on causal regions.

While effective, SEraser suffers from two core issues: (i) textual prompts remain fixed and thus encode latent linguistic bias, and (ii) random erasure can inadvertently mask causal regions, leading to unstable or suboptimal adaptation.

To address these limitations, we introduce **BiPrompt** a bilateral debiasing framework that simultaneously mitigates visual and textual biases through two complementary mechanisms: (i) *Balanced Prompt Normalization* for text-space regularization, and (ii) *Structured Spurious-Region Erasure*

for spatially guided visual debiasing. Together, they enable fine-grained causal alignment without retraining or additional supervision.

(a) Balanced Prompt Normalization. Standard prompt embeddings $f_t(t_c)$ often exhibit anisotropy in the textual space, favoring dominant or frequent classes. To reduce this imbalance, BiPrompt learns a normalized textual embedding:

$$\hat{f}_t(t_c) = \alpha f_t(t_c) + (1-\alpha)\bar{f}_t, \tag{5}$$

where $\bar{f}_t = \frac{1}{C}\sum_{c=1}^{C} f_t(t_c)$ is the global semantic centroid, and α is a learnable gating parameter. This adaptive interpolation encourages isotropic text embeddings, minimizing linguistic dominance and improving stability under domain shift.

(b) Structured Spurious-Region Erasure. Instead of random masking, BiPrompt employs *attention-guided erasure*. Grad-CAM is used to compute soft attention maps $m(x)$ that highlight causal regions. We construct complementary foreground and background views:

$$x_{\text{fg}} = m(x) \odot x, \qquad x_{\text{bg}} = (1 - m(x)) \odot x. \tag{6}$$

Prediction consistency is enforced between x_{fg} and x, while orthogonality is promoted between x_{bg} and x:

$$\mathcal{L}_{\text{BSE}} = D_{\text{KL}}\big(p(y|x_{\text{fg}}) \,\|\, p(y|x)\big) - \beta \cos\big(p(y|x_{\text{bg}}), p(y|x)\big). \tag{7}$$

This structured erasure selectively suppresses spurious activations while preserving causal semantics.

(c) Overall Objective. The full test-time objective integrates all components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_1 \mathcal{L}_{\text{BSE}} + \lambda_2 \mathcal{L}_{\text{ent}}. \tag{8}$$

Only a few lightweight parametersthe gating α and normalization weightsare updated during adaptation, keeping the optimization efficient and memory-friendly.

Optimization and Inference

During inference, BiPrompt performs the following lightweight test-time adaptation steps:

1. Compute Grad-CAM attention maps to obtain foreground and background views $(x_{\text{fg}}, x_{\text{bg}})$.
2. Extract visual features $f_v(x)$, $f_v(x_{\text{fg}})$, and $f_v(x_{\text{bg}})$, along with normalized text embeddings $\hat{f}_t(t_c)$.
3. Perform one or few gradient updates to minimize $\mathcal{L}_{\text{total}}$.
4. Compute final predictions using similarity $s(f_v(x), \hat{f}_t(t_c))$.

This adaptive process aligns visual and textual embeddings across causal regions, improving generalization to unseen domains. BiPrompt implicitly minimizes the conditional mutual information between spurious features z_s and predictions y:

$$I(z_s; y|z_c) \approx 0. \tag{9}$$

The structured erasure term enforces conditional independence in the visual space, while balanced prompt normalization reduces anisotropy in the textual space. Together, these mechanisms yield a causally disentangled representation that enhances robustness and reliability under distribution shift.

Results

We evaluate the effectiveness of our proposed method, BiPrompt, against several baselines on a diverse set of benchmarks. We first describe the experimental setup, then present the main results on both real-world OOD datasets and simulated spurious-bias scenarios, followed by an ablation study on model generality.

Experimental Setup

Datasets. We evaluate BiPrompt across two experimental settings. For *real-world out-of-distribution (OOD)* data, we use Tiny-ImageNet (Le and Yang 2015), CUB-200 (Wah et al. 2011), and ImageNet-A, which capture naturally occurring domain shifts. For *simulated spurious-bias* data, we adopt the Waterbirds benchmark (Koh et al. 2021) and two datasets generated by the S2E protocolCamelDeer and SpiderCrab (Ma et al. 2025). These datasets explicitly model backgroundobject correlations, allowing us to assess BiPrompt's ability to mitigate shortcut learning and maintain causal generalization under controlled bias.

Baselines and Backbone. We compare BiPrompt with several representative methods, including Vanilla CLIP, TPT (Shu et al. 2022), RoSHOT (Adila et al. 2023), α-CLIP, and SEraser (Ma et al. 2025). All methods, except α-CLIP which uses its own encoder, are built on the pretrained CLIP ViT-B/32 backbone (Radford et al. 2021). For SEraser, we adopt its *Patches* variant using four corner patches from an 8×8 gridand the *Images* variant that leverages OOD reference samples. In simulated bias experiments, all test-time adaptation methods (including BiPrompt) employ SAM (Kirillov et al. 2023) to isolate and erase background regions, ensuring fair and consistent evaluation across frameworks.

Results on Real-World OOD Scenarios

As shown in Table 1, our proposed BiPrompt consistently outperforms all baselines on average across the three real-world datasets. Notably, BiPrompt achieves a Top-1 accuracy of **42.4%**, surpassing the strongest original SEraser strategy (Blocks) by 1.9%. This demonstrates the clear benefit of our dual-pronged approach, which addresses both visual and linguistic biases.

On CUB-200 and ImageNet-A, BiPrompt shows significant gains of 2.1% and 2.5%, respectively, over the next-best SEraser variant. While α-CLIP achieves an outstanding performance on Tiny-ImageNet, this is expected as its checkpoint was retrained on ImageNet (Ma et al. 2025). On the other hand, less-biased benchmarks, such as BiPrompt, prove to be the most robust method.

Results on Simulated Spurious-Bias Scenarios

Table 2 presents the results on datasets explicitly designed to measure robustness against spurious correlations, with a focus on average (AVG) and worst-group (W.G.) accuracy. In these challenging scenarios, BiPrompt demonstrates a substantial improvement over all other methods.

Table 1: Zero-shot classification performance on real-world OOD datasets (Tiny-ImageNet, CUB-200, ImageNet-A). Top-1 accuracy (%) and improvement over Vanilla are shown.

Dataset	Vanilla	TPT	RoSHOT	α-CLIP	Patches	Images	Blocks	**BiPrompt (Ours)**
Tiny-ImageNet	23.2	29.6	49.2	76.0	42.4	41.2	42.8	**44.1** (▲20.9)
CUB-200	12.1	8.7	25.5	44.3	26.2	24.2	28.9	**31.0** (▲18.9)
ImageNet-A	42.1	49.7	38.9	51.5	47.4	45.4	49.7	**52.2** (▲10.1)
Average	25.8	29.3	37.9	57.3	38.7	36.9	40.5	**42.4** (▲16.6)

Table 2: Zero-shot classification performance on simulated spurious-bias scenarios. Average accuracy (AVG.) and worst-group accuracy (W.G.) are reported across Waterbirds, CamelDeer, and SpiderCrab datasets. The symbols ↑ and ↓ denote performance gain or drop relative to the Vanilla baseline for BiPrompt.

Dataset	**Metric**	**Van.**	**MASK**	**TPT**	**RoSHOT**	**α-CLIP**	**SEraser**	**BiPrompt**
Waterbirds	AVG.	67.7	72.0	66.9	68.9	67.6	78.2	**79.9 (↑ 12.2)**
	W.G.	40.0	51.5	34.4	52.3	43.2	65.3	**66.6 (↑ 26.5)**
CamelDeer	AVG.	83.2	93.6	77.7	80.4	92.0	95.7	**97.2 (↑ 14.0)**
	W.G.	66.4	87.2	55.3	60.8	84.4	91.6	**92.8 (↑ 26.4)**
SpiderCrab	AVG.	66.0	91.4	83.5	73.0	86.2	95.3	**97.4 (↑ 31.4)**
	W.G.	42.0	90.4	72.5	50.4	86.0	94.7	**95.4 (↑ 53.4)**
Avg. (3 sets)	AVG.	72.3	85.7	76.0	74.1	81.9	89.8	**91.3 (↑ 19.0)**
	W.G.	49.5	76.4	54.1	54.5	71.2	83.7	**85.0 (↑ 35.5)**

Table 3: Zero-shot classification performance of different VLFMs on the Waterbirds dataset. Average accuracy (AVG.) and worst-group accuracy (W.G.) are reported. The symbol ▲ indicates performance gain over the Vanilla baseline.

Model	**Van.**	**MASK**	**SEraser**	**BiPrompt**
CLIP-L14 (AVG.)	83.7	85.5	87.8	**88.4** (▲4.7)
W.G.	32.9	40.8	58.9	**60.1** (▲27.2)
BLIP-2 (AVG.)	57.7	54.4	55.6	**56.3** (▲-1.4)
W.G.	28.2	35.1	34.7	**35.5** (▲7.3)
Avg. (2 sets) (AVG.)	70.7	69.9	71.7	**72.4** (▲1.7)
W.G.	30.5	37.9	46.8	**47.8** (▲17.3)

Across all three datasets, BiPrompt achieves the highest worst-group accuracy, validating its superior ability to mitigate decision shortcuts. This is particularly evident on the Waterbirds benchmark, where BiPrompt improves the W.G. performance significantly over the Vanilla baseline.

Our method significantly outperforms the original SEraser, which, despite its strong performance, is surpassed by BiPrompt's structured erasure and prompt normalization. In contrast, other methods, such as TPT, perform suboptimally. In contrast, other methods such as TPT assume that high-confidence views capture invariant features, but this often fails when strong spurious cues, such as a desert background, lead to confident yet incorrect predictions. RoSHOT exhibits similarly inconsistent behavior, whereas the consistent and notable gains, particularly in worst-group accuracy, demonstrate the robustness and effectiveness of our method.

Effectiveness Across Different Model Architectures

To evaluate the flexibility and generality of our approach, we tested BiPrompt on other widely-used VLM architectures, specifically **CLIP ViT-L-14** and **BLIP-2**. We conducted this evaluation on the challenging Waterbirds benchmark (Ma et al. 2025), with results shown in Table 3. Our method achieves promising performance on both models, consistently outperforming the baseline.

Conclusion

In this work, we tackled the challenge of unreliable generalization in vision–language models (VLMs) that often depend on spurious correlations when faced with out-of-distribution (OOD) data. While existing test-time methods primarily target visual bias, they often overlook the interplay between visual and linguistic factors. We introduced **BiPrompt**, a bilateral test-time adaptation framework that jointly mitigates both forms of bias through *Balanced Prompt Normalization* for text-space isotropy and *Structured Erasure* for attention-guided visual debiasing. Extensive experiments demonstrate that BiPrompt achieves consistent zero-shot gains, particularly on worst-group accuracy benchmarks. By enabling cross-modal bias correction without retraining, BiPrompt offers a simple yet effective step toward building more reliable and causally grounded foundation models.

References

Adila, D.; Shin, C.; Cai, L.; and Sala, F. 2023. Zero-Shot Robustification of Zero-Shot Models With Foundation Models. *arXiv preprint arXiv:2309.04344*.

Chuang, C.-Y.; Jampani, V.; Li, Y.-T.; Torralba, A.; and Jegelka, S. 2023. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*.

Fan, L.; Krishnan, D.; Isola, P.; Katabi, D.; and Tian, Y. 2023. Improving CLIP Training with Language Rewrites. In *NeurIPS*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Koh, P. W.; Sagawa, S.; Marklund, H.; Xie, S. M.; Zhang, M.; Balsubramani, A.; Hu, W.; Yasunaga, M.; Phillips, R. L.; Gao, I.; et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 5637–5664. PMLR.

Le, Y.; and Yang, X. 2015. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7): 3.

Liang, F.; Wu, B.; Dai, X.; Li, K.; Zhao, Y.; Zhang, H.; Zhang, P.; Vajda, P.; and Marculescu, D. 2023. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7061–7070.

Ma, H.; Zhu, Y.; Zhang, C.; Zhao, P.; Wu, B.; Huang, L.-K.; Hu, Q.; and Wu, B. 2025. Spurious Feature Eraser: Stabilizing Test-Time Adaptation for Vision-Language Foundation Model. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25)*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, 8748–8763. Virtual Conference: PMLR.

Shtedritski, A.; Rupprecht, C.; and Vedaldi, A. 2023. What does clip know about a red circle? visual prompt engineering for vlms. *arXiv preprint arXiv:2304.06712*.

Shu, M.; Nie, W.; Huang, D.-A.; Yu, Z.; Goldstein, T.; Anandkumar, A.; and Xiao, C. 2022. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 14274–14289.

Sun, Z.; Fang, Y.; Wu, T.; Zhang, P.; Zang, Y.; Kong, S.; Xiong, Y.; Lin, D.; and Wang, J. 2024. Alpha-CLIP: A CLIP Model Focusing on Wherever You Want. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15605–15616.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Yang, L.; Wang, Y.; Li, X.; Wang, X.; and Yang, J. 2023. Fine-Grained Visual Prompting. *arXiv preprint arXiv:2306.04356*.