DENSITY-AWARE TRANSLATION OF SPURIOUS CORRELATIONS IN ZERO-SHOT VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language models (VLMs), such as CLIP, achieve powerful zero-shot classification. However, their predictions remain highly sensitive to spurious correlations, where common background or contextual cues dominate predictions over semantic content. Earlier solutions typically rely on fine-tuning, but this undermines the advantages of pre-trained models. Others depend on prompt engineering, which is prone to hallucination issues. In addition, most approaches are limited to a single modality, increasing the risk of misalignment between text and images. In this work, we propose **Density-Aware Translation** (DAT) that refines image-text similarity scores using a local geometric density term derived from group reference sets. Our approach is motivated by the phenomenon that CLIP embeddings exhibit a modality gap and lie on an anisotropic shell in the feature space: common patterns cluster near the mean, while rare patterns are pushed outward. This geometry creates uneven alignment, where spurious correlations are amplified while semantically meaningful but rare cues are marginalised. To address this, we employ a relative measure that rescales similarities based on embedding density, suppressing overconfident scores in diffuse regions while preserving dense, semantically consistent matches. Experimental results on benchmark datasets demonstrate consistent improvements in worst-group and average accuracy, highlighting density-aware translation as a simple and effective calibration mechanism for reliable zero-shot classification using multimodal models.

1 INTRODUCTION

Vision-Language models (VLMs), such as CLIP (Radford et al., 2021), have advanced multimodal learning by aligning image and text embeddings in a shared latent space, enabling strong zero-shot capabilities that generalise to unseen classes without requiring additional training (Chen et al., 2023). This capability makes them widely used for tasks such as classification (Qian & Hu, 2024), retrieval (Sain et al., 2023), and reasoning (Subramanian et al., 2022) in different domains.

Despite their remarkable success, VLMs are still susceptible to spurious correlations (Bommasani, 2021; Chuang et al., 2023; Varma et al., 2024), where predictions rely on frequent but semantically irrelevant cues rather than meaningful content. Levi & Gilboa (2025) show that frequently occurring concepts tend to align more closely with the modality mean vector, exhibiting higher conformity, which reflects semantic blurring. This geometric bias makes the model over-rely on such frequent but uninformative patterns, while down-weighting rarer yet more informative signals, ultimately compromising robustness across groups. For instance, spurious correlations arise in chest X-ray diagnostics, where models trained to detect pneumonia have been shown to rely on hospital-specific markers instead of true lung pathology (Zech et al., 2018; DeGrave et al., 2021).

Existing approaches to mitigating spurious correlations in VLMs fall into three groups. (i) Fine-tuning and adaptors (Zhang & Ré, 2022; Goyal et al., 2023; Varma et al., 2024) rely on bias-aware objectives but require labelled supervision, undermining zero-shot generalisation. Wu et al. (2023) also proposes a supervised, concept-aware correction using gradient-based concept discovery with the white-box model access. (ii) Text-side methods (Chuang et al., 2023; Trager et al., 2023; An et al., 2024) edit prompts or projections, but risk cross-modal misalignment and often depend on domain expertise or LLMs, which can be unreliable and inconsistent (Abbasi et al., 2025; Huang et al., 2025; Molahasani et al., 2025). (iii) Multimodal embedding adjustments (Adila et al., 2024;

Lu et al., 2025) alter image features with text guidance, but linear projections distort geometry and translation-based methods require dataset-specific scaling, calibrated by training data.

In this work, we revisit the geometric limitations of VLMs such as CLIP, by examining how their similarity scores neglect the anisotropic, ellipsoidal structure of the embedding space. Prior studies have shown that CLIP embeddings are unevenly distributed, with dense and sparse regions emerging along different directions (Liang et al., 2022; Levi & Gilboa, 2025). Building on this observation, we propose the **Density-Aware Translation** (DAT) that incorporates *local geometric information* into similarity computation. In particular, we quantify local geometry via a relative density ratio, estimated from reference sets constructed through sampling to capture variations across both labels and spurious attributes. This allows us to down-weight spurious similarities that occur when a sample aligns with the wrong prompt, while preserving similarity for samples that lie in dense, representative regions of their correct group. Importantly, our approach operates fully in the zero-shot regime, without access to model parameters, and only relies on a small, balanced reference set to capture group-level density characteristics.

From a theoretical perspective, we model group distributions using the Kent (Fisher-Bingham) distribution (Kent, 1982; Mardia & Jupp, 2009), and show that raw CLIP scores systematically deviate from Bayes-optimal decision boundaries. Our DAT provably reinstates anisotropy-sensitive log-likelihood terms that cosine similarity ignores, aligning the discriminant with Bayes-optimal scoring. We evaluate our method on standard spurious correlation benchmarks and demonstrate consistent improvements in both worst-group and average accuracy, while preserving the flexibility of zero-shot inference. These results underscore the importance of explicitly modelling anisotropy and local density in multimodal embeddings to achieve robust and fair classification.

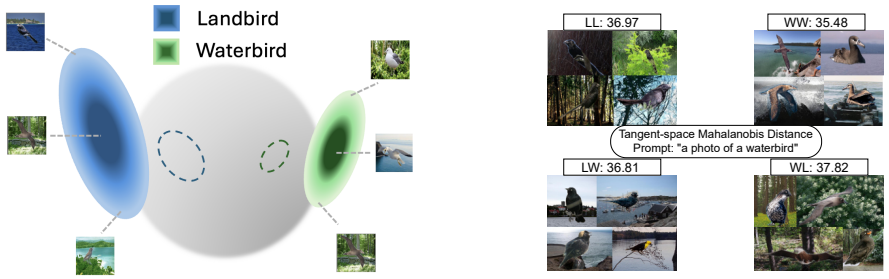
The main contributions can be summarised as follows:

- We introduce DAT, a simple zero-shot mechanism that rescales image–text similarities using group reference densities, requiring no fine-tuning, no prompt engineering, and no access to spurious attribute labels at test time.
- We provide a theoretical analysis showing that DAT corrects cosine's bias under anisotropic embeddings, reinstating missing log-likelihood terms and aligning with Bayes-optimal decision rules.
- Through experiments on different benchmarks and across multiple VLMs, we demonstrate consistent improvements in different metrics.

2 Related Work

Debiasing VLMs via Fine-tuning. Several recent works adapt CLIP and related VLMs through bias-aware fine-tuning objectives. Yang et al. (2023) introduces a multimodal contrastive loss that explicitly separates spurious attributes from class-defining features by encoding spurious cues in language. Varma et al. (2024) identifies spurious correlations at the region level via clustering, and then applies a region-aware loss that suppresses spurious regions while emphasising causal ones. Zhang & Ré (2022) develops contrastive adaptors that not only align sample embeddings with their class prototypes but also bring same-class samples closer together, improving group robustness with minimal parameters. Zhang et al. (2024) propose a prompt tuning method that disentangles causal and spurious features by decoupling alignment into two contrastive phases. Pang et al. (2024) instead leverages vectorised group attributes to explicitly debias image representations under sub-population shifts. Dehdashtian et al. (2024) frame CLIP debiasing in a reproducing kernel Hilbert space and employ a statistical dependence measure to decorrelate representations from spurious attributes.

Zero-shot Debiasing in VLMs. Some methods aim to mitigate spurious correlations through embedding-based debiasing without retraining, thereby preserving the zero-shot capability of VLMs. Ideal-Prompt (Trager et al., 2023) constructs an ideal text representation by combining basis vectors in the embedding space. Perception CLIP (An et al., 2024) is a two-stage procedure that first infers contextual attributes (e.g., background) and then conditions object classification on them. Orth-Cali (Chuang et al., 2023) project text embeddings onto subspaces orthogonal to spurious directions, showing that text-only calibration suffices for robust classifiers and fair generative models. Ge et al. (2023) improves zero-shot accuracy by detecting potentially misclassified images via pre-



(a) Ellipsoidal shape of embedding in CLIP.

(b) Distance inconsistency in CLIP.

Figure 1: Motivation for DAT. (a) CLIP embeddings exhibit anisotropic ellipsoidal structure, where frequent, spuriously correlated samples cluster near the mean, while rare but semantically meaningful samples lie in sparser regions. (b) Tangent-space Mahalanobis Distance for group embeddings shows that spuriously aligned LW has a lower distance, on average, to the waterbird prompt than the true WL group. Abbreviations: *LL*: landbird with land background, *WL*: waterbird with land background, *LW*: landbird with water background, *WW*: waterbird with water background.

diction consistency and augmenting text prompts with hierarchical label information from WordNet. Adila et al. (2024) proposes ROBOSHOT that leverages large language models to extract spurious-related insights from task descriptions and applies linear projection to suppress harmful and enhance beneficial embedding components, though it inherits fragility from LLM-generated cues. Most recently, Lu et al. (2025) presents TIE, a zero-shot framework that translates image embeddings along directions guided by spurious text prompts, reducing reliance on shortcuts while preserving distributional structure. In TIE, it is assumed that the spurious labels of the test images are available. To relax this assumption, TIE* is introduced, which operates without access to spurious labels.

3 Density-Aware Translation

In this section, we begin with the formal problem setup in Section 3.1, followed by the motivation in Section 3.2, the proposed DAT in Section 3.3, and finally the theoretical analysis in Section 3.4.

3.1 Preliminaries

We study the group robustness setting (Sagawa et al., 2020) in the context of zero-shot classification. Let $(x, y, a) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{A}$ denote an input image x, with class label $y \in \mathcal{Y}$, and a spurious attribute $a \in \mathcal{A}$ (e.g., background or context). We denote $|\mathcal{Y}| = K$ for the number of classes and $|\mathcal{A}| = M$ for the number of spurious attributes. Each group is defined as a pair (y, a),

$$g_{y,a} \in \mathcal{G} = \mathcal{Y} \times \mathcal{A}, \quad |\mathcal{G}| = K \cdot M,$$

so that robustness is evaluated at the group level. For the zero-shot model, we denote $\phi_I(.)$ as the frozen image encoder and $\phi_T(.)$ as the frozen text encoder.

Following (Lu et al., 2025), we assume access to text prompts describing both labels and spurious attributes. For each label $y \in \mathcal{Y}$, we define a class-only prompt $t_y \in \mathcal{T}^{\mathcal{Y}}$, e.g., "`a photo of a` y". For each spurious attribute $a \in \mathcal{A}$, we define an attribute prompt $t_a \in \mathcal{T}^{\mathcal{A}}$, e.g., "`a photo of a` a". Finally, to represent groups, we construct concatenated prompts as group prompts $t_{y,a} \in \mathcal{T}^{\mathcal{Y},\mathcal{A}}$, e.g., "`a photo of a` y `with` a". To capture group geometry, we consider group samples from the training or validation set as $\{x_{y,a}^{(h)}\}_{h=1}^{N_{y,a}}$, where $N_{y,a}$ denotes the number of available samples in group (y, a). When the spurious attribute a is not explicitly provided, we employ DAT*, which infers group membership in a zero-shot manner via attribute prompts:

$$\hat{a} \;=\; \arg\max_{a \in \mathcal{A}} \langle \phi_I(x), \phi_T(t_a) \rangle \,. \tag{1}$$

The corresponding image embeddings are then obtained as

$$\text{DAT} : z_{y,a}^{(h)} \;=\; \phi_I\big(x_{y,a}^{(h)}\big), \qquad \text{DAT*} : z_{y,\hat{a}}^{(h)} \;=\; \phi_I\big(x_{y,\hat{a}}^{(h)}\big).$$

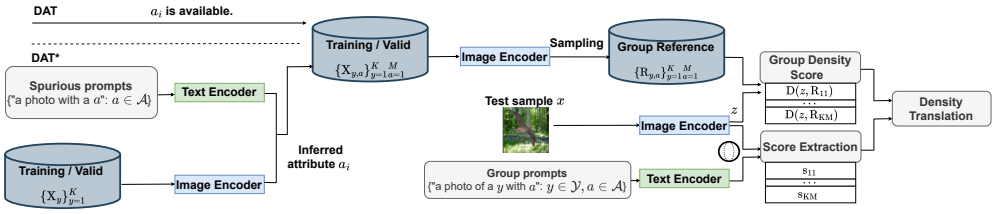


Figure 2: Pipeline of DAT/DAT*: DAT constructs group reference sets, encodes test images and prompts, estimates local density, and rescales similarities. DAT* follows the same pipeline but infers group attributes automatically, without explicit annotations.

3.2 Understanding the Geometric Bias

As illustrated in Figure 1a, CLIP embeddings are distributed anisotropically on ellipsoidal shells: frequent concepts (often correlated with spurious attributes) cluster closer to the mean, while rare but semantically meaningful concepts lie in sparser, peripheral regions (Levi & Gilboa, 2025). For example, in the Waterbirds dataset (Sagawa et al., 2020), "waterbird on water" is much more common than "waterbird on land." CLIP's embedding geometry places these spuriously correlated samples near the centre, while rare but informative cases occupy the periphery.

This geometric bias creates two key challenges for zero-shot classification. First, scoring against a single prompt per class means rare but valid samples can receive lower similarity than spuriously aligned ones, sometimes matching spurious prompts more strongly than their true class. Second, CLIP often overemphasises the largest object and the first prompt token while down-weighting smaller but critical details (Abbasi et al., 2025), letting background or context cues dominate and limiting the value of prompt augmentation alone.

We analyse group misalignment using the Tangent–Space Mahalanobis Distance (TMD) on the unit sphere (Pennec, 2006). Let all embeddings be ℓ_2–normalised, $\hat{\mu}_{y,a}$ be group Fréchet mean on the unit hypersphere $\mathbb{S}^{d-1}$ and $\mathrm{Log}_{\hat{\mu}_{y,a}}$ be the Riemannian logarithm map that projects points from the sphere to the tangent space at $\hat{\mu}_{y,a}$. Let $z_{y,a}^{(h)} \in \mathbb{S}^{d-1}$ be the h-th image embedding in group (y, a), and $w_{y,a} \in \mathbb{S}^{d-1}$ be the corresponding normalised text embedding. Then we define:

$$\delta_{y,a}^{(h)} = \mathrm{Log}_{\hat{\mu}_{y,a}}\big(z_{y,a}^{(h)}\big), \quad \delta_{w,y,a} = \mathrm{Log}_{\hat{\mu}_{y,a}}\big(w_{y,a}\big), \quad \Sigma_{y,a} = \frac{1}{N_{y,a}} \sum_{h=1}^{N_{y,a}} \delta_{y,a}^{(h)}\, \delta_{y,a}^{(h)\top},$$

$$\mathrm{TMD}_{y,a} = \sqrt{\delta_{w,y,a}^{\top} \big(\Sigma_{y,a}\big)^{-1} \delta_{w,y,a}},$$

where $\Sigma_{y,a}$ denotes the empirical covariance of tangent vectors within group (y, a), and $\delta_{w,y,a}$ represents the projection of the text embedding into the tangent space at the group mean. As shown in Figure 1b, using Waterbirds with CLIP ViT-L/14, the spuriously aligned group LW (landbird on water) has a smaller TMD to the "a photo of a waterbird" prompt than the true WL (waterbird on land) group. This indicates that background correlation (water) dominates alignment, biasing CLIP toward spuriously aligned groups and away from semantically correct ones.

3.3 Debiasing via Density-Aware Translation

As shown in Figure 2, we first construct reference sets per group using a sampling procedure with an equal budget n per group. We then apply deterministic feature-space exemplar herding (Rebuffi et al., 2017) (Appendix B.2), which greedily selects embeddings whose running average matches the group mean. This yields a compact set of representative exemplars for each (y, a), providing stable local neighbourhoods for density estimation. Following Levi & Gilboa (2025), frequent (common) samples tend to lie closer to the centre of the embedding distribution, so sampling towards the mean allows us to capture these typical examples and compute a fair density metric for each group. For each group (y, a) we construct a reference set of size n:

$$\{z_{y,a}^{(h)}\}_{h=1}^{N_{y,a}} \xrightarrow{\text{Sampling}} R_{y,a} = \{z_{y,a}^{(h)}\}_{h=1}^{n}$$

Bias annotations (e.g., background or gender) can optionally be used, but are not required. In cases without explicit group labels, we have **DAT***, which employs an auxiliary zero-shot classifier to infer group membership using Equation 1.

Density Ratio Estimation. For each query image $x \in \mathcal{X}_{\text{test}}$, and we then compute its embedding $z = \phi_I(x)$. The local density of z relative to $R_{y,a}$ is estimated with the simplified local outlier factor (SLOF) (Schubert et al., 2014) as our default proxy:

$$D_{y,a}(z) = \mathrm{SLOF}(z; R_{y,a}) = \frac{1}{k} \sum_{z_o \in \mathrm{NN}_k(z)} \frac{\text{k-dist}(z)}{\text{k-dist}(z_o)}.$$

$\mathrm{NN}_k(z)$ denotes the k-nearest neighbours of z in $R_{y,a}$ and $\text{k-dist}(\cdot)$ is the distance to the k-th neighbour. A larger SLOF value indicates that z lies in a sparser region, i.e., it is less representative of the group. We use SLOF due to its simplicity. Other relative density measures are also applicable.

Density Translation. Given group prompts $t_{y,a}$ and class-only prompts t_y, we compute the raw similarities as

$$s_{y,a}(x) = \langle \phi_I(x), \phi_T(t_{y,a}) \rangle, \qquad s_y(x) = \langle \phi_I(x), \phi_T(t_y) \rangle.$$

We then translate group scores using local density:

$$\tilde{s}_{y,a}(x) = \frac{s_{y,a}(x)}{(D_{y,a}(z) + \varepsilon)^{\lambda}}, \tag{2}$$

where $\lambda > 0$ controls the translation strength and small $\varepsilon > 0$ ensures numerical stability. DAT adjusts raw similarities by reducing scores for samples that appear highly similar to a group prompt but lie far from the core of a group's image embeddings, while preserving scores for samples close to their true group.

Aggregation and Prediction. In addition to group-specific scores $\{\tilde{s}_{y,a}(x)\}_{a \in \mathcal{A}}$, we define a class-marginal (attribute-averaged) score:

$$\tilde{s}_{y,\mathrm{Avg}}(x) = \frac{1}{M+1} \Big(\sum_{a \in \mathcal{A}} \tilde{s}_{y,a}(x) + s_y(x) \Big). \tag{3}$$

Final predictions are made by maximising over group-corrected and averaged scores:

$$\hat{y}_q \;=\; \arg\max_{y \in \mathcal{Y}} \Big\{ \max_{a \in \mathcal{A}} \tilde{s}_{y,a}(x),\; \tilde{s}_{y,\mathrm{Avg}}(x) \Big\}. \tag{4}$$

The full procedure of DAT/DAT* is outlined in Algorithm 1 and Algorithm 2 in Appendix B.1. Appendix B.3 further illustrates how density shapes predictions, showing that SLOF separates rare samples from spuriously frequent ones and that DAT enlarges the score margin over the baseline.

3.4 Theoretical Analysis

We first show that raw cosine similarity is systematically biased in anisotropic embedding geometries. We then assume a standard log-density fidelity link for SLOF, and prove that DAT reinstates the anisotropy-sensitive terms missing from pure cosine scoring.

Anisotropy Effect on Cosine Similarity. Due to the anisotropic nature of CLIP embeddings, a query embedding $z \in \mathbb{S}^{d-1}$ can exhibit elliptical concentration, which is naturally captured by the Kent (Fisher–Bingham) distribution (Kent, 1982; Mardia & Jupp, 2009).With parameters (κ, β, Γ) and orthonormal frame $\Gamma = (\gamma_1, \gamma_2, \gamma_3, \ldots)$, its density is

$$p(z \mid \kappa, \beta, \Gamma) = c_d(\kappa, \beta)\, \exp\Big\{ \kappa\, \gamma_1^\top z + \beta \big[(\gamma_2^\top z)^2 - (\gamma_3^\top z)^2 \big] \Big\}, \quad z \in \mathbb{S}^{d-1}.$$

Here, $\kappa \geq 0$ controls axial concentration toward γ_1, and β controls ellipticity (anisotropy) in the (γ_2, γ_3)-plane. The Bayes log-density is therefore

$$\log p(z) = \kappa\, \gamma_1^\top z + \beta \big[(\gamma_2^\top z)^2 - (\gamma_3^\top z)^2 \big] - \log c_d(\kappa, \beta). \tag{5}$$

Cosine scoring with text direction w uses only the axial projection $w^\top z$. In particular, if we take $w = \gamma_1$ (the Kent mean axis), cosine recovers the linear term $\kappa\, \gamma_1^\top z$ but ignores the quadratic anisotropy term $\beta \big[(\gamma_2^\top z)^2 - (\gamma_3^\top z)^2 \big]$.

Proposition 1. *Let p be $\mathrm{Kent}(\kappa, \beta, \Gamma)$ with $\beta \neq 0$ and set the cosine direction $w_y = \gamma_1$. Then there exist $z_+, z_- \in \mathbb{S}^{d-1}$ with $w_y^\top z_+ = w_y^\top z_-$ but $\log p(z_+) > \log p(z_-)$. Hence, ranking by cosine similarity can disagree with Bayes ranking $\log p(z)$.*

This shows that cosine similarity systematically overlooks anisotropy effects, potentially misranking rare but semantically important samples.

Assumption 1 (log-SLOF fidelity)**.** *There exist constants $\alpha > 0$, $\eta \in \mathbb{R}$, and a bounded error $\epsilon_{y,a}(z)$ with $|\epsilon_{y,a}(z)| \leq c$ such that*

$$\log D_{y,a}(z) \;=\; \alpha\big(-\log p_{y,a}(z)\big) \;+\; \eta \;+\; \epsilon_{y,a}(z),$$

where $p_{y,a}(z)$ is the group density.

This assumption formalises the link between kNN distance statistics and negative log-density (Loftsgaarden & Quesenberry, 1965; Biau & Devroye, 2015). LOF/SLOF are widely used density proxies that satisfy such a relation up to bounded error (Breunig et al., 2000; Zimek et al., 2012).

DAT Margin. Given group (y, a), let $s_{y,a}(x) = \langle \phi_I(x), \phi_T(t_{y,a}) \rangle$ denote the raw similarity (cosine between image and text embeddings), and let $D_{y,a}(z)$ be a local sparsity proxy (SLOF). For the purpose of theoretical analysis, we work in the logit space, defining

$$\ell_{y,a}(x) := \tau\, w_{y,a}^\top z,$$

where $w_{y,a}$ is the normalised text embedding and $\tau > 0$ is a temperature. In logit space, Eq. 2 corresponds to subtracting $\lambda \log D$ from the raw logit ℓ. We then define the DAT margin as

$$m_{y,a}(z) \;:=\; \ell_{y,a}(z) \;-\; \lambda\, \log D_{y,a}(z).$$

Theorem 1 (Local Bayes alignment)**.** *Under Assumption 1,*

$$m_{y,a}(z) \;=\; \tau\, w_{y,a}^\top z \;+\; \alpha\lambda\, \log p_{y,a}(z) \;+\; r_{y,a}(z), \qquad r_{y,a}(z) := -\lambda\eta - \lambda\, \epsilon_{y,a}(z),$$

so that $|r_{y,a}(z)| \leq \lambda(|\eta| + c) =: B_0$. Hence, the DAT discriminant equals a logit term plus a (scaled) log-likelihood term, up to a bounded remainder. With equal priors, argmax over (y, a) is Bayes-aligned.

Corollary 1 (DAT reinstates anisotropy under Kent)**.** *Under Assumption 1 and the Kent (Fisher-Bingham) model for the group density, the DAT margin admits the decomposition*

$$m(z) \;\approx\; \underbrace{\big(\tau\, w + \alpha\lambda\, \kappa\, \gamma_1\big)^\top z}_{\textit{linear (axial) part}} \;+\; \underbrace{\alpha\lambda\, \beta\big[(\gamma_2^\top z)^2 - (\gamma_3^\top z)^2\big]}_{\textit{anisotropy correction}} \;-\; \alpha\lambda\, \log c_d(\kappa, \beta),$$

where the constant terms (including $r_{y,a}(z)$) do not affect the argmax. Thus, when $\beta \neq 0$, DAT adds the missing quadratic, anisotropy-sensitive term that pure cosine scoring ignores, correcting its bias in elliptical embeddings.

Proofs of Proposition 1, Theorem 1, and Corollary 1 are detailed in Appendix A.

Beyond Bayes alignment. We also prove that DAT strictly decreases standard groupwise surrogate risks (logistic/hinge) relative to raw cosine similarity under mild conditions; see Appendix A.4 for statements and complete proofs.

4 Experiments

Following (Adila et al., 2024; Lu et al., 2025), we conduct zero-shot classification experiments on benchmark datasets designed to test spurious correlations.

Datasets and Models. We use Waterbirds (Sagawa et al., 2020), CelebA (Liu et al., 2015), COVID-19 (Cohen et al., 2020), and FMoW (Christie et al., 2018). Further details about the datasets are provided in Appendix B.4. Our evaluation covers multiple vision–language models: three CLIP variants (ViT-B/32, ViT-L/14, ResNet-50) (Radford et al., 2021), as well as ALIGN and AltCLIP. For COVID-19, we adopt BiomedCLIP (Zhang et al., 2023), a CLIP variant fine-tuned on biomedical data. Following (Lu et al., 2025), we use CLIP ViT-L/14 for FMoW due to the dataset's complexity. Across all settings, experiments are performed using frozen embeddings from the pre-trained models.

Table 1: Zero-shot classification results on Waterbirds using the Worst-Group accuracy (WG), Average accuracy (Avg), and Gap between Average and Worst-Group accuracy (Gap). Higher WG (↑) and Avg, and lower Gap (↓) values are better.

Method	CLIP (ViT-B/32)			CLIP (ViT-L/14)			CLIP (ResNet50)		
	WG	Avg	Gap	WG	Avg	Gap	WG	Avg	Gap
ZS	41.37	68.48	27.11	31.93	83.72	51.79	35.36	80.64	45.28
Group Prompt	43.46	66.79	23.33	10.44	56.12	45.68	49.84	70.96	21.12
Ideal words	60.28	79.20	18.92	64.17	87.67	23.50	39.90	79.48	40.39
Orth-Cali	54.99	69.19	14.20	58.86	86.31	27.45	60.84	84.47	19.67
Perception CLIP	59.78	**82.50**	22.72	54.12	86.74	32.62	48.21	**91.51**	43.30
ROBOSHOT	54.41	71.92	17.51	45.17	64.43	19.26	<u>69.60</u>	<u>96.05</u>	42.45
TIE	<u>71.35</u>	79.82	<u>8.47</u>	78.82	84.12	**5.30**	52.96	83.62	30.66
TIE*	61.24	76.91	15.67	61.60	78.98	17.38	34.11	81.19	47.08
DAT	**75.08**	80.36	**5.28**	**83.33**	**89.57**	<u>6.42</u>	**75.08**	83.83	**8.75**
DAT*	64.02	<u>82.33</u>	18.31	<u>79.75</u>	<u>87.87</u>	8.12	63.71	82.65	<u>18.94</u>

Metric and Baselines. We use three metrics in our experiments: average accuracy % (Avg), worst-group accuracy % (WG), and the gap between the two % (Gap), which are evaluated in many spurious correlation works (Adila et al., 2024; Lu et al., 2025). A robust model should have a high Avg and WG, with a small Gap between them. In all result tables, the best score is highlighted in bold, and the second-best score is underlined. We benchmark our approach against both simple baselines and the latest methods in robust zero-shot classification. Specifically, the baselines consist of standard zero-shot classification (ZS) and a variant that incorporates group information through prompting (Group prompt). For comparison with prior work, we include recent state-of-the-art approaches such as Ideal words (Trager et al., 2023), Orth-Cali (Chuang et al., 2023), Perception CLIP (An et al., 2024), ROBOSHOT (Adila et al., 2024), and TIE/TIE* (Lu et al., 2025), which are detailed in Section 2.

Prompt Details. In zero-shot classification, we employ three categories of text prompts: label prompts, spurious prompts, and group prompts. For spurious prompts, we utilized those provided by (Lu et al., 2025), and for group prompts, we concatenated the class and attribute templates. To support reproducibility, the full list of all types of prompts used in our experiments is included in Appendix B.5.

Settings. For DAT, the neighbourhood size k is set to 10 on Waterbirds, CelebA, and COVID-19, and to 30 on FMoW. The number of reference samples per group n is 56 for Waterbirds, 128 for CelebA, 40 for COVID-19, and 50 for FMoW. The scaling parameter λ is set to 10 for Waterbirds, COVID-19, and FMoW, and 1 for CelebA. For Waterbirds, COVID-19, and FMoW, we construct the reference sets from the training split. For CelebA, we use the validation split, which provides enough group coverage. We utilized an NVIDIA H100 GPU with frozen weights. Details on implementation efficiency, in comparison to TIE, are provided in Appendix B.5, where we show that DAT achieves higher efficiency.

4.1 Results

Waterbirds and CelebA. As shown in Table 1, DAT achieves the highest WG accuracy across all baselines, surpassing the strongest prior by roughly 4-14%, depending on the backbone. The largest gains are observed with ViT-L/14 and ResNet-50 backbones. DAT* also achieves competitive performance, particularly with ViT-based models, and maintains comparable average accuracy. Table 2 reports zero-shot results on CelebA. Consistent with the trends observed in Table 1, DAT and DAT* achieve the strongest performance across most metrics, particularly in terms of WG accuracy. Moreover, DAT* attains results comparable to DAT, with only a minor difference.

Table 2: Zero-shot classification results on CelebA using the Worst-Group accuracy (WG), Average accuracy (Avg), and Gap between Average and Worst-Group accuracy (Gap). Higher WG (↑) and Avg, and lower Gap (↓) values are better.

Method	CLIP (ViT-B/32)			CLIP (ViT-L/14)			CLIP (ResNet50)		
	WG	Avg	Gap	WG	Avg	Gap	WG	Avg	Gap
ZS	78.89	84.27	5.38	73.35	81.20	7.85	69.69	81.58	11.89
Group Prompt	74.90	80.38	5.48	68.94	77.86	8.92	70.59	79.48	8.89
Ideal words	78.12	80.96	2.84	76.67	**89.15**	12.48	65.65	76.27	10.62
Orth-Cali	77.92	82.31	4.39	77.69	81.39	3.70	69.13	76.47	7.34
Perception CLIP	76.46	80.32	3.86	78.70	81.41	2.71	<u>80.22</u>	85.17	**4.95**
ROBOSHOT	80.52	84.77	4.25	82.61	85.54	2.93	73.96	80.90	6.94
TIE	**82.63**	85.11	**2.48**	84.60	86.17	<u>1.57</u>	75.32	81.71	6.39
TIE*	<u>82.61</u>	85.10	<u>2.49</u>	81.98	84.27	2.29	75.30	81.70	6.40
DAT	78.53	<u>87.09</u>	8.56	**85.35**	<u>86.54</u>	**1.19**	**80.79**	<u>87.09</u>	<u>6.30</u>
DAT*	78.53	**87.11**	8.58	<u>84.93</u>	<u>86.54</u>	1.61	78.53	**88.29**	9.76

Table 3: Zero-shot classification results on COVID-19 (medical dataset) and FMoW (multi-label dataset).

(a) COVID-19 (Biomed-CLIP).

COVID-19			
Method	WG	Avg	Gap
ZS	44.83	61.81	16.98
Group Prompt	27.58	48.27	20.69
Ideal words	23.53	56.84	33.31
Orth-Cali	44.83	51.72	<u>6.89</u>
Perception CLIP	48.84	56.87	8.03
ROBOSHOT	32.75	53.10	20.35
TIE	52.17	62.50	10.33
TIE*	50.22	61.08	10.86
DAT	<u>65.22</u>	**75.69**	10.47
DAT*	**72.41**	<u>74.30</u>	**1.89**

(b) FMoW (ViT-L/14).

FMoW			
Method	WG	Avg	Gap
ZS	18.06	26.02	7.96
Group Prompt	8.75	14.69	5.94
Ideal words	11.14	20.21	9.07
Orth-Cali	19.45	26.11	6.66
Perception CLIP	12.61	17.70	<u>5.09</u>
ROBOSHOT	10.88	19.79	8.91
TIE	<u>20.19</u>	26.62	6.43
TIE*	19.84	26.65	6.81
DAT	**27.75**	**31.19**	**3.44**
DAT*	18.63	<u>29.56</u>	10.93

Medical domain and Multi-Label. We further evaluate DAT on medical imaging using COVID-19 chest X-ray datasets with BioMedCLIP. As shown in Table 3a, DAT and DAT* outperform all baselines, achieving over 20% higher WG than the latest baselines and reducing Gap by 5%. To assess scalability to richer label and attribute structure, we also test on FMoW, which has 62 classes, and 5 spurious attributes. Table 3b shows that DAT attains the highest WG and Avg, and markedly reduces the Gap. The label-free variant, DAT*, is competitive as well, especially on WG and Avg.

Evaluation of Other Models. To evaluate the effectiveness of our method across a broader range of VLMs, we also test DAT/DAT* on ALIGN and AltCLIP. For a fair comparison, baseline results are taken from Adila et al. (2024), which is one of the latest baselines that evaluates these two models using these datasets. As shown in Table 4, on both Waterbirds and CelebA, DAT and DAT* in most cases outperform previous methods, with a remarkable improvement in WG accuracy, highlighting robustness across distinct VLMs.

Ablation Study. We investigate the sensitivity of DAT to the scaling factor (λ), the number of samples per group (n), and the number of neighbours for density estimation (k). Results in Figure 3 show that DAT remains stable across a wide range of settings, consistently outperforming the strongest baseline (TIE), as well as the simplest baseline (zero-shot classification). Additionally, a detailed discussion on the effects of prompt structure and spurious feature specification is provided in Appendix B.8 and C.

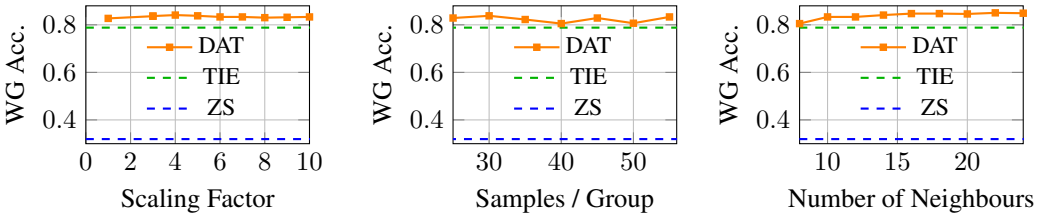


Figure 3: Comparison of WG accuracy under varying (a) scaling factor, (b) number of samples per group, and (c) number of neighbours using Waterbirds and CLIP(ViT-L/14).

Table 4: Generalisation of DAT to other models: zero-shot classification results using ALIGN and AltCLIP on Waterbirds and CelebA datasets.

	Waterbirds						**CelebA**					
	ALIGN			AltCLIP			ALIGN			AltCLIP		
Method	WG	Avg	Gap	WG	Avg	Gap	WG	Avg	Gap	WG	Avg	Gap
ZS	50.3	72.0	21.7	35.8	90.1	54.3	77.2	81.8	4.6	79.7	82.3	**2.6**
Group Prompt	5.8	72.5	66.7	29.4	82.4	53.0	67.4	78.3	10.9	79.0	82.3	3.3
ROBOSHOT	41.0	50.9	9.9	54.8	78.5	23.7	**83.4**	**86.3**	2.9	77.2	86.0	8.8
DAT	**73.36**	**82.90**	**9.54**	**81.31**	**91.54**	**10.23**	82.76	84.82	**2.06**	**83.89**	86.64	2.75
DAT*	63.72	76.16	12.44	56.23	89.76	33.53	82.69	84.80	2.11	83.89	**86.68**	2.79

Robust Text Prompt. Following An et al. (2024); Lu et al. (2025), we expand group descriptions by incorporating multiple semantically related variants of spurious attributes (e.g., alternative ways of describing land or water backgrounds in Waterbirds, as generated in Liang et al. (2022)) and averaging their embeddings. This robustified representation is intended to better capture attribute variability and has been shown to improve both WG and Avg accuracy across backbones. The full details are provided in Appendix B.6. As shown in Table 5, however, group-robust prompts result in only marginal changes, minor improvements in some cases, but overall negligible differences. In contrast, for DAT*, reported in Appendix B.6, robust prompts have a more pronounced effect, as group assignments in DAT* are directly influenced by spurious attributes.

Table 5: Group robustify prompting evaluation using the Waterbirds dataset.

Method	ViT-B/32			ViT-L/14			ResNet-50		
	WG	Avg	Gap	WG	Avg	Gap	WG	Avg	Gap
DAT	75.08	80.36	5.28	83.33	89.57	6.42	75.08	83.83	8.75
DAT Robust	74.99	80.12	**5.13**	83.49	89.09	5.60	75.39	83.65	8.26

5 Conclusion

We introduced the Density-Aware Translation method, a simple zero-shot mechanism that rescales image-text similarities using a geometric density proxy computed from small group references, addressing the bias introduced by the ellipsoidal shape of CLIP embeddings. Across multiple datasets, models, and settings, we demonstrated that DAT consistently improves performance, raising WG and Avg accuracy while reducing the gap between them. Theoretically, we showed that DAT's multiplicative correction leads to an additive discriminant that aligns cosine similarity with a Bayes-style log-likelihood, while reinstating anisotropy-sensitive terms that cosine similarity alone overlooks. Although DAT and its variant DAT* perform strongly compared to prior methods, we believe future work should explore adaptive multimodal approaches for density estimation in the embedding space that are less sensitive to reference selection and prompt design.

Ethics Statement

This work aims to mitigate spurious correlations in vision–language models to enhance robustness across diverse groups. All experiments were performed on publicly available benchmark datasets, in accordance with their respective licenses and usage guidelines. No private, identifiable, or personally sensitive data was used. The research was conducted in full compliance with the ICLR Code of Ethics.

Reproducibility Statement

All theoretical claims are accompanied by complete proofs in Appendix A. Experimental settings and implementation details, including hyperparameters and density estimation choices, are described in Section 4, while dataset details are provided in Appendix B.4. The full set of prompts evaluated in this study can be found in Appendix B.5 and Appendix B.6. Finally, algorithms for DAT and DAT* are explicitly presented in Appendix B.1.

References

Reza Abbasi, Ali Nazari, Aminreza Sefid, Mohammadali Banayeeanzade, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. Clip under the microscope: A fine-grained analysis of multi-object representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9308–9317, 2025.

Dyah Adila, Changho Shin, Linrong Cai, and Frederic Sala. Zero-shot robustification of zero-shot models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Bang An, Sicheng Zhu, Michael-Andrei Panaitescu-Liess, Chaithanya Kumar Mummadi, and Furong Huang. PerceptionCLIP: Visual classification by inferring and conditioning on contexts. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Gérard Biau and Luc Devroye. *Lectures on the nearest neighbor method*, volume 246. Springer, 2015.

Rishi Bommasani. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2000.

Zixiang Chen, Yihe Deng, Yuanzhi Li, and Quanquan Gu. Understanding transferable representation learning and zero-shot transfer in clip. *arXiv preprint arXiv:2310.00927*, 2023.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6172–6180, 2018.

Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023.

Joseph P Cohen, Paul Morrison, and Lan Dao. Covid-19 image data collection: Prospective predictions are the future. arXiv preprint arXiv:2006.11988, 2020.

Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

Sepehr Dehdashtian, Lan Wang, and Vishnu Boddeti. FairerCLIP: Debiasing CLIP's zero-shot predictions using functions in RKHSs. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Christiane Fellbaum. *WordNet: An electronic lexical database*. MIT press, 1998.

Yunhao Ge, Jie Ren, Andrew Gallagher, Yuxiao Wang, Ming-Hsuan Yang, Hartwig Adam, Laurent Itti, Balaji Lakshminarayanan, and Jiaping Zhao. Improving zero-shot generalization and robustness of multi-modal models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19338–19347, 2023.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Wilson. What are bayesian neural network posteriors really like? In *Proceedings of the International Conference on Machine Learning (ICML)*, 2022.

John T Kent. The fisher-bingham distribution on the sphere. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(1):71–80, 1982.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

Meir Yossef Levi and Guy Gilboa. The double-ellipsoid geometry of clip. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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)*, pp. 3730–3738, 2015.

Don O Loftsgaarden and Charles P Quesenberry. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 36(3):1049–1051, 1965.

Shenyu Lu, Junyi Chai, and Xiaoqian Wang. Mitigating spurious correlations in zero-shot multimodal models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Kanti V Mardia and Peter E Jupp. *Directional statistics*. John Wiley & Sons, 2009.

Mahdiyar Molahasani, Azadeh Motamedi, Michael Greenspan, Il-Min Kim, and Ali Etemad. Prism: Reducing spurious implicit biases in vision-language models with llm-guided embedding projection. *arXiv preprint arXiv:2507.08979*, 2025.

OpenAI. ChatGPT. `https://chat.openai.com/`, 2023. Large language model.

Yijiang Pang, Bao Hoang, and Jiayu Zhou. Cross-modality debiasing: using language to mitigate sub-population shifts in imaging. *arXiv preprint arXiv:2403.07888*, 2024.

Xavier Pennec. Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127–154, 2006.

Qi Qian and Juhua Hu. Online zero-shot classification with clip. In *European Conference on Computer Vision (ECCV)*. Springer, 2024.

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 *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2001–2010, 2017.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy S Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. Clip for all things zero-shot sketch-based image retrieval, fine-grained or not. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Erich Schubert, Arthur Zimek, and Hans-Peter Kriegel. Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. *Data Mining and Knowledge Discovery*, 28(1):190–237, 2014.

Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. *arXiv preprint arXiv:2204.05991*, 2022.

Matthew Trager, Pramuditha Perera, Luca Zancato, Alessandro Achille, Parminder Bhatia, and Stefano Soatto. Linear spaces of meanings: compositional structures in vision-language models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 15395–15404, 2023.

Maya Varma, Jean-Benoit Delbrouck, Zhihong Chen, Akshay Chaudhari, and Curtis Langlotz. Ravl: Discovering and mitigating spurious correlations in fine-tuned vision-language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Shirley Wu, Mert Yuksekgonul, Linjun Zhang, and James Zou. Discover and cure: Concept-aware mitigation of spurious correlation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

Yu Yang, Besmira Nushi, Hamid Palangi, and Baharan Mirzasoleiman. Mitigating spurious correlations in multi-modal models during fine-tuning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 39365–39379, 2023.

John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric K. Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Medicine*, 15(11), 2018.

Jie Zhang, Xiaosong Ma, Song Guo, Peng Li, Wenchao Xu, Xueyang Tang, and Zicong Hong. Amend to alignment: Decoupled prompt tuning for mitigating spurious correlation in vision-language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

Michael Zhang and Christopher Ré. Contrastive adapters for foundation model group robustness. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

Arthur Zimek, Erich Schubert, and Hans-Peter Kriegel. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(5):363–387, 2012.

A Theoretical Analysis and Proofs

A.1 Proof of proposition 1

Proof. Fix any $t \in (-1, 1)$ and $r \in \big(0, \sqrt{1-t^2}\big)$. Define

$$\gamma_1^\top z_\pm = t, \qquad \gamma_2^\top z_+ = r,\ \gamma_3^\top z_+ = 0, \qquad \gamma_2^\top z_- = 0,\ \gamma_3^\top z_- = r,$$

and take remaining coordinates 0 so that $z_\pm \in \mathbb{S}^{d-1}$. Then $w^\top z_+ = w^\top z_- = t$ (same cosine), while

$$\log p(z_+) - \log p(z_-) = \beta\Big[(\gamma_2^\top z_+)^2 - (\gamma_3^\top z_+)^2 - \big((\gamma_2^\top z_-)^2 - (\gamma_3^\top z_-)^2\big)\Big] = 2\beta r^2 \neq 0.$$

Thus $\log p$ strictly prefers one of $z_\pm$ (sign set by β), whereas cosine ties them. □

A.2 Proof of Theorem 1

Proof. By the definition of the DAT margin,

$$m_{y,a}(z) \;=\; \tau\, w_{y,a}^\top z \;-\; \lambda\, \log D_{y,a}(z).$$

Based on Assumption 1, which states that for some $\alpha > 0$, $\eta \in \mathbb{R}$ and a bounded error $\epsilon_{y,a}(z)$ with $|\epsilon_{y,a}(z)| \le c$,

$$\log D_{y,a}(z) \;=\; \alpha\big(-\log p_{y,a}(z)\big) \;+\; \eta \;+\; \epsilon_{y,a}(z).$$

Substituting this into the margin gives

$$\begin{aligned} m_{y,a}(z) &= \tau\, w_{y,a}^\top z \;-\; \lambda\Big\{\alpha\big(-\log p_{y,a}(z)\big) + \eta + \epsilon_{y,a}(z)\Big\} \\ &= \tau\, w_{y,a}^\top z \;+\; \alpha\lambda\, \log p_{y,a}(z) \;-\; \lambda\eta \;-\; \lambda\, \epsilon_{y,a}(z). \end{aligned}$$

Define the remainder

$$r_{y,a}(z) \;:=\; -\lambda\eta - \lambda\,\epsilon_{y,a}(z),$$

so that

$$m_{y,a}(z) \;=\; \tau\, w_{y,a}^{\top} z \;+\; \alpha\lambda\,\log p_{y,a}(z) \;+\; r_{y,a}(z).$$

Because $|\epsilon_{y,a}(z)| \le c$, we have the uniform bound

$$|r_{y,a}(z)| \;\le\; \lambda\big(|\eta| + c\big) \;=:\; B_0.$$

This proves the stated decomposition and bound.

Bayes alignment. With equal priors, the Bayes rule ranks groups by $\log p_{y,a}(z)$. For any two groups $g = (y, a)$ and $g' = (y', a')$,

$$m_g(z) - m_{g'}(z) \;=\; \alpha\lambda\Big(\log p_g(z) - \log p_{g'}(z)\Big) \;+\; \tau\big(w_g - w_{g'}\big)^{\top} z \;+\; r_g(z) - r_{g'}(z).$$

Hence m ranks identically to the Bayes score whenever the Bayes gap dominates the bounded perturbations, e.g.

$$\alpha\lambda\,\big|\log p_g(z) - \log p_{g'}(z)\big| \;>\; \big|\tau\,(w_g - w_{g'})^{\top} z\big| \;+\; |r_g(z)| + |r_{g'}(z)|,$$

and in particular when $\tau = 0$ (or the τ-term is treated as a fixed bias) and B_0 is small. Thus, the DAT discriminant equals a similarity term plus a (scaled) log-likelihood term up to a bounded remainder, and is Bayes-aligned in the argmax under equal priors in the sense above. □

A.3 Proof of Corollary 1

Proof sketch. By Theorem 1, $m(z) = \tau\, w^{\top} z + \alpha\lambda\,\log p(z) + r(z)$. Substitute the Kent log-density; collect linear ($\propto \gamma_1^{\top} z$), quadratic ($\propto (\gamma_2^{\top} z)^2 - (\gamma_3^{\top} z)^2$), and constant terms. Absorb bounded $r(z)$ and constants into a bias; these do not change the argmax. □

A.4 Groupwise Surrogate-Risk Improvement

For a group g (e.g., a specific (y, a)), let $Z \in \mathbb{S}^{d-1}$ denote its image embedding and $Y \in \{\pm 1\}$ the group label in a one-vs-rest reduction.[1] Recall from Theorem 1 (main text) the DAT margin decomposition

$$m(Z) \;=\; m_0(Z) \;+\; \alpha\lambda\, L(Z) \;+\; r(Z), \qquad m_0(Z) := \tau\, w_g^{\top} Z, \quad L(Z) := \log p_g(Z),$$

where $\alpha\lambda > 0$ is the density weight and r is a bounded remainder.

For a convex surrogate ℓ, the groupwise surrogate risk is

$$R_{\ell}^{(g)}(m) \;:=\; \mathbb{E}\big[\ell\big(Y\, m(Z)\big)\,\big|\, G = g\big], \qquad \ell \in \{\ell_{\log},\, \ell_{\text{hinge}}\},$$

with $\ell_{\log}(t) = \log(1 + e^{-t})$ and $\ell_{\text{hinge}}(t) = \max\{0,\, 1 - t\}$.

Assumption 2 (Hard-mass nondegeneracy)**.** *There exists $c_g > 0$ such that for the path*

$$m_{\theta}(Z) \;:=\; m_0(Z) \;+\; \theta\big(\alpha\lambda\, L(Z) + r(Z)\big), \qquad \theta \in [0, 1],$$

we have, for all $\theta \in [0, 1]$,

$$\mathbb{E}\big[\sigma\big(-Y\, m_{\theta}(Z)\big)\, Y\, L(Z)\,\big|\, G = g\big] \;\ge\; c_g \quad \textit{and} \quad \mathbb{E}\big[\mathbf{1}\{Y\, m_{\theta}(Z) < 1\}\, Y\, L(Z)\,\big|\, G = g\big] \;\ge\; c_g,$$

where $\sigma(t) = 1/(1 + e^{-t})$ is the logistic sigmoid, and there exists $B_g < \infty$ such that $|r(Z)| \le B_g$ almost surely.

Theorem 2 (Strict groupwise surrogate-risk decrease)**.** *Under Assumptions 2, if $\alpha\lambda\, c_g > B_g$, then for $\ell \in \{$logistic, hinge$\}$,*

$$R_{\ell}^{(g)}(m) \;<\; R_{\ell}^{(g)}(m_0).$$

[1] The multiclass (multi-group) case follows by standard one-vs-rest aggregation; we state the binary reduction for clarity.

Proof. **Logistic.** We have:

$$\frac{d}{d\theta}R^{(g)}_{\log}(m_\theta) = \mathbb{E}\big[\ell'_{\log}\big(Ym_\theta(Z)\big)\,Y\big(\alpha\lambda L(Z)+r(Z)\big)\,\big|\,G=g\big] \;= -\,\mathbb{E}\big[\sigma\big(-Ym_\theta(Z)\big)\,Y\big(\alpha\lambda L(Z)+r(Z)\big)\big]\,.$$

$$\frac{d}{d\theta}R^{(g)}_{\log}(m_\theta) = -\,\alpha\lambda\,\mathbb{E}[\sigma(-Ym_\theta)\,Y\,L] \;-\; \mathbb{E}[\sigma(-Ym_\theta)\,Y\,r]\,.$$

Assumption 2 gives the first expectation $\geq c_g$, while $|\mathbb{E}[\sigma(-Ym_\theta)\,Y\,r]| \leq \mathbb{E}[|r|] \leq B_g$ a.s. Hence $\frac{d}{d\theta}R^{(g)}_{\log}(m_\theta) \leq -\alpha\lambda c_g + B_g < 0$.

Hinge. For the hinge loss:

$$\frac{d}{d\theta}R^{(g)}_{\text{hinge}}(m_\theta) \in -\,\mathbb{E}\big[\mathbf{1}\{Ym_\theta(Z)<1\}\,Y\big(\alpha\lambda L(Z)+r(Z)\big)\,\big|\,G=g\big]\,,$$

where the right-hand side is any measurable subderivative (a.e. well-defined).

$$\frac{d}{d\theta}R^{(g)}_{\text{hinge}}(m_\theta) = -\,\alpha\lambda\,\mathbb{E}[\mathbf{1}\{Ym_\theta<1\}\,Y\,L] \;-\; \mathbb{E}[\mathbf{1}\{Ym_\theta<1\}\,Y\,r]\,.$$

Assumption 2 lower-bounds the first term by c_g, while $|\mathbb{E}[\cdot]| \leq B_g$ for the second. Thus $\frac{d}{d\theta}R^{(g)}_{\text{hinge}}(m_\theta) \leq -\alpha\lambda c_g + B_g < 0$ and integration finishes the proof. □

B Experiments and Datasets Details

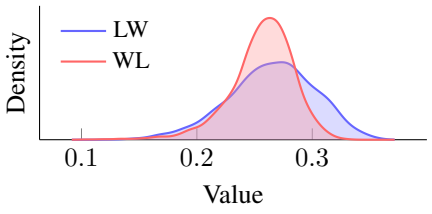
B.1 Algorithm

The algorithms for DAT and DAT* are presented in Algorithms 1 and 2, respectively. Their primary distinction lies in the construction of the reference sets: DAT assumes direct access to the spurious attribute, whereas DAT* infers it through zero-shot classification.

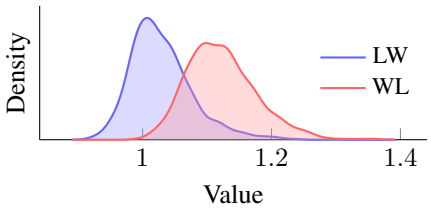
Algorithm 1 DAT

Input: Image x, Image encoder $\phi_I(\cdot)$, Text encoder $\phi_T(\cdot)$, Group prompts $\{t_{y,a}\}$, Class-only prompts $\{t_y\}$, Reference sets $\{R_{y,a}\}$.
Output: Predicted label $\hat{y}$.

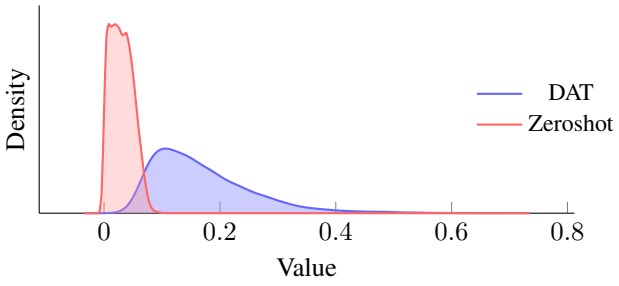
$z \leftarrow \phi_I(x)$ ▷ Compute image embedding
for each group (y, a) **do**
  $s_{y,a}(x) \leftarrow \langle z, \phi_T(t_{y,a})\rangle$ ▷ Compute group-wise similarity
  if $|R_{y,a}| < n$ **then**
    $D_{y,a}(z) \leftarrow \infty$ ▷ Insufficient reference samples
  else
    $D_{y,a}(z) \leftarrow \mathrm{SLOF}(z; R_{y,a})$ ▷ Local density
  end if
  $\tilde{s}_{y,a}(x) \leftarrow \frac{s_{y,a}(x)}{(D_{y,a}(z)+\varepsilon)^\lambda}$ ▷ DAT correction
end for
$\tilde{s}_{y,\mathrm{Avg}}(x) \leftarrow \frac{1}{M+1}\Big(\sum_{a\in\mathcal{A}}\tilde{s}_{y,a}(x) + s_y(x)\Big)$ ▷ Aggregate class-marginal score
$\hat{y} \leftarrow \arg\max_{y\in\mathcal{Y}} \max_{a\in\mathcal{A}}\{\tilde{s}_{y,a}(x), \tilde{s}_{y,\mathrm{Avg}}(x)\}$
return $\hat{y}$



(a) Distribution of zero-shot score.



(b) Distribution of SLOF scores.



(c) Score difference distributions.

Figure 4: Density comparisons on Waterbirds with CLIP ViT-L/14. (a) For WL images (waterbird on land), raw cosine similarity to the WL vs LW prompts shows heavy overlap, and cosine alone cannot reliably separate the true group from its spuriously aligned counterpart. (b) SLOF values for the same WL images, measured against WL vs LW references; WL assigns a higher SLOF (sparser) than LW, revealing density asymmetry. (c) Across the dataset, DAT enlarges the max–min group score gap relative to the baseline, yielding more decisive predictions.

Algorithm 2 DAT*

Input: Image x, Image encoder $\phi_I(\cdot)$, Text encoder $\phi_T(\cdot)$, Group prompts $\{t_{y,a}\}$, Class-only prompts $\{t_y\}$, Spurious prompts $\{t_y\}$, Reference sets $\{R_{y,a}\}$.
Output: Predicted label $\hat{y}$.

$z \leftarrow \phi_I(x)$ ▷ Compute image embedding
$\hat{a} = \arg\max_{a \in \mathcal{A}} \langle z, w_a \rangle$. ▷ Infer spurious attributes
Make the reference sets $\{R_{y,a}\}$
for each group (y, a) **do**
  $s_{y,a}(x) \leftarrow \langle z, \phi_T(t_{y,a}) \rangle$ ▷ Compute group-wise similarity
  if $|R_{y,a}| < n$ **then**
    $D_{y,a}(z) \leftarrow \infty$ ▷ Insufficient reference samples
  else
    $D_{y,a}(z) \leftarrow \mathrm{SLOF}(z; R_{y,a})$ ▷ Local density
  end if
  $\tilde{s}_{y,a}(x) \leftarrow \frac{s_{y,a}(x)}{(D_{y,a}(z)+\varepsilon)^{\lambda}}$ ▷ DAT correction
end for
$\tilde{s}_{y,\mathrm{Avg}}(x) \leftarrow \frac{1}{M+1}\Big(\sum_{a \in \mathcal{A}} \tilde{s}_{y,a}(x) + s_y(x)\Big)$ ▷ Aggregate class-marginal score
$\hat{y} \leftarrow \arg\max_{y \in \mathcal{Y}} \max_{a \in \mathcal{A}} \{\tilde{s}_{y,a}(x), \tilde{s}_{y,\mathrm{Avg}}(x)\}$
return $\hat{y}$

B.2 Reference-set construction

Given a group pool of images $X_{y,a} = \{x_{y,a}^{(h)}\}_{h=1}^{N_{y,a}}$, let the (unit–norm) image embeddings be $z_{y,a}^{(h)} = \phi_I(x_{y,a}^{(h)})/\|\phi_I(x_{y,a}^{(h)})\|_2$ and define $\mathcal{G}_{y,a}^{feat} = \{z_{y,a}^{(h)}\}_{h=1}^{N_{y,a}} \subset \mathbb{R}^d$. The group mean in feature space is

$$\mu_{y,a} = \frac{1}{N_{y,a}} \sum_{h=1}^{N_{y,a}} z_{y,a}^{(h)}.$$

Starting from an empty reference set $R_{y,a}^{(0)} =$ with running sum $S_0 = \mathbf{0} \in \mathbb{R}^d$, we select a uniform budget of n exemplars by greedy feature-space herding:

$$p_k^{(y,a)} \in \arg \min_{z \in \mathcal{G}_{y,a} \setminus R_{y,a}^{(k-1)}} \left\| \mu_{y,a} - \frac{1}{k} \left(z + \sum_{j=1}^{k-1} p_j^{(y,a)} \right) \right\|_2, \qquad k = 1, \ldots, n,$$

where we set $R_{y,a}^{(k)} = R_{y,a}^{(k-1)} \cup \{p_k^{(y,a)}\}$ and $S_{k-1} = \sum_{j=1}^{k-1} p_j^{(y,a)}$. Equivalently, the selection can be written as

$$p_k^{(y,a)} \in \arg \max_{z \in \mathcal{G}_{y,a}^{feat} \setminus R_{y,a}^{(k-1)}} \Big(\langle z_{y,a}, \mu_{y,a} \rangle - \tfrac{1}{k} \langle z_{y,a}, S_{k-1} \rangle \Big).$$

We break ties deterministically (by smallest index) for reproducibility.

B.3 Density Effect Visualization.

To better illustrate the role of SLOF in mitigating spurious correlations, we provide three complementary visualizations on the Waterbirds dataset with CLIP ViT-L/14.

In Figure 5a, we focus on the marginalized group WL (waterbird on land), which is often misclassified as LW(landbird on water). For these WL images, we plot the cosine similarity with respect to both group prompts *"a photo of a waterbird in land"* and *"a photo of a landbird in water"*. The strong overlap in distributions shows that raw cosine similarity based on group prompts alone cannot reliably separate the true group from its spuriously correlated counterpart.

Figure 5b shows the SLOF values for the same WL images, this time measured relative to reference sets from WL and LW. Here, the WL references consistently assign higher SLOF values compared to LW, confirming that SLOF captures the sparsity structure of group embeddings and highlights the marginalization of rare groups.

Finally, Figure 5c visualizes the distribution of score differences (maximum minus minimum group score) across the dataset, comparing DAT with the uncorrected baseline. DAT widens this gap, producing more confident and reliable predictions.

Together, these results demonstrate that SLOF not only exposes the density imbalance between rare and spurious groups but also provides a mechanism for DAT to improve the discriminative margin in practice.

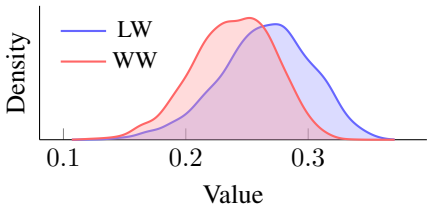
Illustrating the Score and SLOF Distribution across Spurious Correlated Groups. We analyze how DAT influences prediction scores across groups in both the Waterbirds and CelebA datasets. As shown in Figure 5, the raw cosine similarity scores for the true group (Landbird in Water, LW) and the spuriously aligned group (Waterbird in Water, WW) are closely overlapping. This score overlap explains why LW images can be misclassified as WW. However, examining the SLOF distributions reveals that LW samples are consistently denser under their own group reference set and sparser under WW references. Despite the narrow SLOF range (due to low intra-group variance in Waterbirds), this density asymmetry is sufficient for DAT to recalibrate scores effectively.

A similar pattern appears in CelebA, as illustrated in Figure 6. The raw scores for the worst group (Dark-Hair Male, DH-M) and its spuriously correlated counterpart (Blond-Hair Male, BH-M) show substantial overlap, reflecting the dataset's strong hair–gender correlation. In contrast, SLOF scores offer clearer separation, with DH-M images exhibiting lower SLOF values under their correct reference group. The overall SLOF range is wider for CelebA due to higher embedding variance, but DAT's use of relative SLOF values within each group ensures stable and consistent behavior.

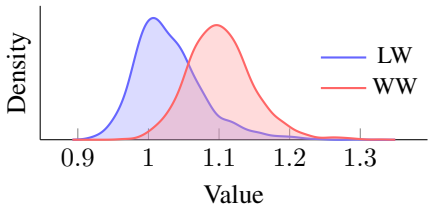
B.4 Datasets

Our method, along with all baselines, is evaluated in the data sets listed below.

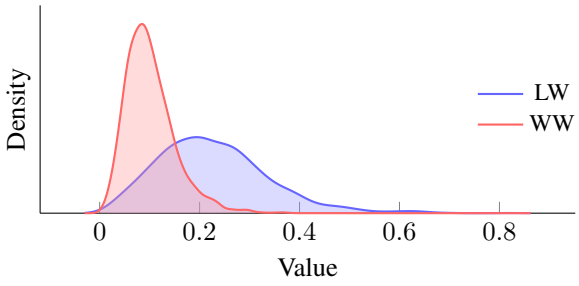
- **Waterbirds** (Koh et al., 2021; Sagawa et al., 2020): A binary bird classification task with labels $y \in \{\text{Landbird}, \text{Waterbird}\}$. The spurious attribute is the background type $a \in \{\text{Land}, \text{Water}\}$, where most landbirds appear on land and most waterbirds over water. This



(a) Distribution of zero-shot score.

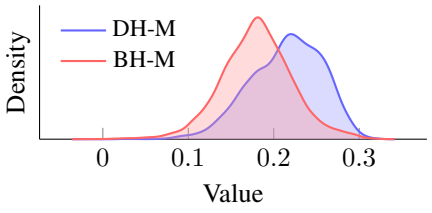


(b) Distribution of SLOF scores.

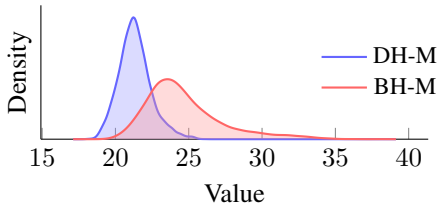


(c) DAT Scores.

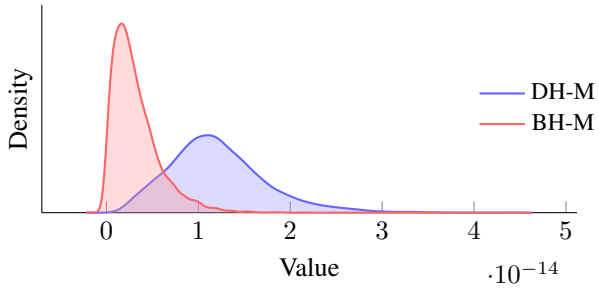
Figure 5: Illustration of score distributions for the worst group in Waterbirds (Landbird in Water, LW) and its spuriously aligned counterpart (Waterbird in Water, WW). (a) Zero-shot cosine scores for both groups largely overlap, making discrimination difficult. (b) SLOF values show clearer separation, with LW having lower SLOF under its own group references. (c) DAT scores incorporate this density difference and sharpen the separation..



(a) Distribution of zero-shot score.



(b) Distribution of SLOF scores.



(c) DAT Scores.

Figure 6: Illustration of score distributions for Dark-Hair Male (DH-M) in CelebA, and its spuriously aligned counterpart (Blond-Hair Male, BH-M). (a) Zero-shot cosine scores for both groups largely overlap, making discrimination difficult. (b) SLOF values show clearer separation, with DH-M having lower SLOF under its own group references. (c) DAT scores incorporate this density difference and sharpen the separation.

produces four groups: Landbird-Land, Landbird-Water, Waterbird-Land, and Waterbird-Water.

- **CelebA** (Liu et al., 2015): A large-scale face dataset (200K+ images) used for binary hair color classification $y \in \{\text{Dark hair}, \text{Blonde hair}\}$. Gender $a \in \{\text{Female}, \text{Male}\}$ is spuriously correlated with hair color, 94% of blond-labeled images are female. Groups are: Female–Dark, Female–Blonde, Male–Dark, and Male–Blonde.
- **COVID-19** (Cohen et al., 2020): An X-ray dataset for pneumonia diagnosis, with task $y \in \{\text{No pneumonia}, \text{Pneumonia}\}$. Gender $a \in \{\text{Male}, \text{Female}\}$ serves as a spurious confounder. Groups include: Male–Pneumonia, Male–No pneumonia, Female–Pneumonia, and Female–No pneumonia.
- **FMoW** (Christie et al., 2018; Izmailov et al., 2022): A large-scale satellite dataset with 62 land-use/building classes. The spurious attribute is the geographical region $a \in \{\text{Africa}, \text{Americas}, \text{Asia}, \text{Europe}, \text{Oceania}\}$. Groups are defined purely by region. FMoW also exhibits a temporal domain shift; training images are collected before 2016, while validation and test images are collected in 2016-2017.

B.5 Implementation and Reproducibility

Details of Prompts. For prompts, we followed (Liang et al., 2022) and utilized their proposed templates. We only created group prompts by simply combining two prompts, without adding any extra information. The prompt details are provided in Table 6.

Implementation Efficiency. We utilized an NVIDIA H100 GPU with frozen weights across all datasets. Table 7 reports the efficiency on the CelebA dataset, one of the large-scale datasets evaluated in this study. As shown, DAT demonstrates higher efficiency than TIE.

Method	CLIP (ViT-B/32)	CLIP (ViT-L/14)	CLIP (ResNet50)
TIE	970	947	951
DAT	203	148	153

Table 7: Comparison of computation time (seconds) efficiency using the CelebA dataset.

B.6 Robust Text Prompt Details - Waterbirds

To evaluate group-robust text prompts, we follow Lu et al. (2025) and adopt their evaluated sentences, generated with GPT-4 (OpenAI, 2023). For Waterbirds, we generate robustified prompts by creating multiple variants of land and water descriptions to represent the spurious attribute (Table 8), along with corresponding group prompts (Table 9). As shown in Table 10, the performance differences between DAT* and DAT* Robust are more pronounced. This aligns with findings from Lu et al. (2025), since DAT* uses the spurious prompts directly to construct group references. Consequently, varying the phrasing of spurious attributes has a greater impact, often improving WG and reducing the GAP, indicating that using multiple semantically aligned prompt variants generally leads to enhanced or comparable performance in terms of WG and GAP.

Table 10: Group robustify prompting evaluation using the Waterbirds dataset.

Method	ViT-B/32			ViT-L/14			ResNet-50		
	WG	Avg	Gap	WG	Avg	Gap	WG	Avg	Gap
DAT*	64.02	82.33	18.31	79.75	87.87	8.12	63.71	82.65	18.94
DAT* Robust	74.68	78.82	4.14	79.64	84.62	4.98	69.00	73.85	4.85

B.7 Robust Text Prompt for CelebA

We further evaluate the robustness of DAT and DAT* to prompt variability on the CelebA dataset. For this setting, we construct group-aligned prompts by generating multiple variants of the spurious attribute using GPT-4 (OpenAI, 2023), as shown in Table 13. We also evaluate group prompts based

on female and male variants (Table 14). These variants are averaged to form robust text embeddings. As reported in Table 11, DAT behaves similarly, with only minor changes in WG and Avg accuracy. On the other hand, Table 12 shows DAT* improvement, as prompt semantics influence the inferred spurious attribute and its alignment to image embeddings.

Table 11: Group robustify prompting evaluation for DAT on the CelebA.

Method	ViT-B/32			ViT-L/14			ResNet-50		
	WG	Avg	Gap	WG	Avg	Gap	WG	Avg	Gap
DAT	78.53	87.09	8.56	85.35	86.54	1.19	80.79	87.09	6.30
DAT Robust	78.53	87.08	8.55	85.05	86.60	1.55	81.36	86.62	5.26

Table 12: Group robustify prompting evaluation for DAT* on the CelebA.

Method	ViT-B/32			ViT-L/14			ResNet-50		
	WG	Avg	Gap	WG	Avg	Gap	WG	Avg	Gap
DAT*	78.53	87.11	8.58	84.93	86.54	1.61	78.53	88.29	9.76
DAT* Robust	79.10	87.20	8.10	85.06	86.61	1.55	81.92	86.85	4.93

B.8 Different Spurious Text Prompt Templates

In addition to the specific wording of the spurious feature itself, the structure of the prompt template can also influence performance. To examine this effect more thoroughly, we evaluated all methods under two template formats, using T1: "{spurious feature label}", and T2: "A photo with a spurious feature, {spurious feature label}" on the Waterbirds dataset and CLIP ViT-B/32. We exclude TIE and DAT from this analysis because they assume access to spurious labels and therefore do not respond to changes in spurious textual formulation. As shown in Table 15, performance remains stable across both formats for DAT*, which achieves the highest worst-group and average accuracy. This suggests that DAT* is robust to prompt phrasing.

Table 15: Effect of prompt template variation on Waterbirds with CLIP ViT-B/32.

Method	T1 Spurious Template			T2 Spurious Template		
	WG	Avg	Gap	WG	Avg	Gap
ZS	41.37	64.48	27.11	41.37	64.48	27.11
Group Prompt	43.46	66.79	23.33	43.46	66.79	23.33
Ideal Words	61.99	78.87	**16.88**	60.28	79.20	18.92
PerceptionCLIP	23.37	61.54	38.17	59.78	**82.50**	22.72
ROBOSHOT	44.35	69.03	24.68	54.41	71.92	<u>17.51</u>
TIE*	56.14	75.00	18.86	61.24	76.91	**15.67**
DAT*	**64.49**	**82.74**	<u>18.25</u>	**64.02**	<u>82.33</u>	18.31

C Discussion on Text Prompts

Identifying strategies to build reliable and generalisable prompts is an open challenge. To examine this further, we conducted a series of experiments evaluating how different prompt formats and levels of object specificity affect DAT performance. To investigate prompt design, we decompose each text prompt into two parts: a *template* and an *object term*, following prior work (Lu et al., 2025). We apply the following templates for prompts:

- T1: A photo with a `[Object]` background

- T2: A photo with a spurious feature, `[Object]`
- T3: `[Object]`

Building on the findings of Ge et al. (2023), which show that object labels exhibit a semantic hierarchy as captured in WordNet (Fellbaum, 1998), we investigate three strategies for selecting object terms in our study: (i) using the immediate parent node in the hierarchy to represent a broader category, (ii) using the spurious feature itself, and (iii) averaging the embeddings of five highly specific sibling terms from the bottom of the hierarchy to capture fine-grained detail. Table 16 lists the candidate terms. This setup allows us to evaluate which level of abstraction yields the most effective prompts. This design allows us to probe how the specificity or abstraction level of the object term impacts model predictions. We conduct these experiments using the DAT* method with CLIP ViT-L/14, which demonstrated strong baseline performance.

Table 16: Object term variants based on WordNet hierarchy.

Granularity	Water Background	Land Background
O1 (Hypernyms)	fluid	ground
O2 (Original)	water	land
O3 (Hyponyms)	sea water, lake water, river water, stream water, creek water	farmland, forest land, arable land, grassland, desert land

Table 17: Prompt structure ablation on Waterbirds with CLIP ViT-L/14 using DAT*.

Prompt	**WG**	**Avg**	**Gap**
T1 + O1	67.44	87.02	19.58
T1 + O2	79.75	87.87	8.12
T1 + O3	80.09	85.76	5.67
T2 + O1	67.60	86.00	18.04
T2 + O2	79.91	86.95	7.04
T2 + O3	80.18	85.47	5.29
T3 + O1	68.69	85.90	17.21
T3 + O2	82.09	87.81	5.72
T3 + O3	80.37	89.06	8.69

From Table 17, we observe that prompt phrasing (T1–T3) has a modest influence, with DAT* demonstrating resilience to syntactic variations. However, the object term plays a larger role: fine-grained and semantically aligned terms (O3) consistently cause higher WG accuracy, while overly broad descriptors like hypernyms (O1) degrade performance. These findings reinforce the importance of context-aware and precise descriptions for zero-shot robustness.

D Usage of Large Language Models

We used OpenAI's GPT-4 and GPT-5 models, with a limited capacity, for language editing and polishing of the manuscript text, as well as for generating synonymous variants of spurious attributes, as described in the paper. The models were not involved in developing ideas, designing methods, and analysing results.

Table 6: Prompts details used for different datasets.

Dataset	Label prompts	Spurious prompts	Group prompts
Waterbirds	a photo of a landbird, a photo of a waterbird	a photo with a water background, a photo with a land background	a photo of a landbird in water, a photo of a waterbird in land, a photo of a landbird in land, a photo of a waterbird in water
CelebA	a photo of a celebrity with dark hair, a photo of a celebrity with blonde hair	a photo of a female, a photo of a male	a photo of a male celebrity with dark hair, a photo of a female celebrity with blonde hair, a photo of a Female celebrity with dark hair, a photo of a Male celebrity with blonde hair
COVID-19	An X-ray image of a chest without Pneumonia, An X-ray image of a chest with Pneumonia	An X-ray image from a female, An X-ray image from a male	an X-ray image of a chest without Pneumonia from a male, an X-ray image of a chest with Pneumonia from a female, an X-ray image of a chest without Pneumonia from a female, an X-ray image of a chest with Pneumonia from a male
FMoW	A satellite image of a/an $y_i{}_{i=0}^{i=61}$	Over Europe, Over Asia, Over Americas, Over Africa, Over Oceania	A satellite image of a/an $y_i{}_{i=0}^{i=61}$ over Europe, [A satellite image of a/an $y_i{}_{i=0}^{i=61}$ over Asia], [A satellite image of a/an $y_i{}_{i=0}^{i=61}$ over Americas], [A satellite image of a/an $y_i{}_{i=0}^{i=61}$ over Africa, [A satellite image of a/an $y_i{}_{i=0}^{i=61}$ over Oceania]

Land	Water
A photo of a land background	A photo of a water background
A photo of a forest background	A photo of an ocean background
A photo of a mountain background	A photo of a sea background
A photo of a terrain background	A photo of a lake background
A photo of a ground background	A photo of a river background

Table 8: Spurious prompts used for robust prompt prompt evaluation of Waterbirds dataset (Lu et al., 2025).

Landbird - Land	Waterbird - Water
A photo of a landbird in land	A photo of a waterbird in water
A photo of a landbird in forest	A photo of waterbird in ocean
A photo of a landbird in mountain	A photo of a waterbird in sea
A photo of a landbird in terrain	A photo of a waterbird in lake
A photo of a landbird in ground	A photo of a waterbird in river
Landbird - Water	**Waterbird - Land**
A photo of a landbird in water	A photo of a waterbird in land
A photo of a landbird in ocean	A photo of waterbird in forest
A photo of a landbird in sea	A photo of a waterbird in mountain
A photo of a landbird in lake	A photo of a waterbird in terrain
A photo of a landbird in river	A photo of a waterbird in ground

Table 9: Group prompts used for robust prompt evaluation of Waterbirds dataset.

Female	Male
A photo of a female	A photo of a male
A photo of a woman	A photo of a man
A photo of a lady	A photo of a gentleman
A photo of a girl	A photo of a boy

Table 13: Spurious prompts used for robust prompt prompt evaluation of the CelebA dataset.

Black Hair - Female	Blonde Hair - Male
a photo of a female celebrity with dark hair	a photo of a male celebrity with blonde hair
a photo of a woman celebrity with dark hair	a photo of a man celebrity with blonde hair
a photo of a lady celebrity with dark hair	a photo of a gentleman celebrity with blonde hair
a photo of a girl celebrity with dark hair	a photo of a boy celebrity with blonde hair
Black Hair - Male	**Blonde Hair - Female**
a photo of a female celebrity with dark hair	a photo of a male celebrity with blonde hair
a photo of a female celebrity with dark hair	a photo of a woman celebrity with blonde hair
a photo of a gentleman celebrity with dark hair	a photo of a lady celebrity with blonde hair
a photo of a boy celebrity with dark hair	a photo of a girl celebrity with blonde hair

Table 14: Group prompts used for robust prompt evaluation of the CelebA dataset.