N-TIDE: Debiasing Unimodal Vision Models via Neutral Text Inversion with CLIP

Chanhee Lee* Jinho Jang* Sarang Han*
Sungkyunkwan University
{leechanhye, jangjinho65, stellahan12}@g.skku.edu

Abstract

*Mitigating bias in vision models is challenging, particularly when semantic attributes subtly influence predictions. While vision-language models like CLIP provide strong debiasing signals, they require text input at inference, limiting their use in image-only settings. We introduce **N-TIDE** (Neutral Text-Inversion for Distillation-based Equilibration), a two-stage framework that distills CLIP's fairness guidance into a unimodal vision model. In the first stage, we propose a novel neutral-text inversion process, which regularizes the model by aligning a trainable neutral-text embedding with CLIP's null-text embedding. This alignment captures semantic debiasing cues without requiring text at test time. In the second stage, we transfer these cues into an image-only encoder via cosine-based feature matching. We further interpret this process through the lens of deterministic diffusion, framing semantic alignment as a guided trajectory.*

Experiments on FairFace show that N-TIDE improves fairness metrics such as Equalized Odds and Representation Bias Difference with minimal accuracy loss. Though the fairness gains are moderate and the diffusion analogy remains conceptual, N-TIDE offers a practical path to integrating multimodal supervision into efficient vision-only models. The code is available at https://github.com/iontail/N-TIDE.git.

1. Introduction

Deep neural networks have shown impressive performance in various computer vision tasks. Nevertheless, concerns about social and representational bias still remain, especially in sensitive areas such as facial recognition and medical imaging. Existing methods for mitigating bias are typically categorized as pre-processing, in-processing, or post-processing [8]. While these approaches can help reduce statistical disparities, they are often limited in addressing deeper semantic biases that originate from the visual content itself.

Recent studies have explored the use of multimodal models such as CLIP, which aligns visual and textual information in a shared semantic space. Several works [3, 5, 15, 22] have demonstrated that CLIP-based techniques, including prompt modification and attribute projection, can effectively suppress biased signals. However, these methods require access to a text encoder during inference, which restricts their applicability in real-world systems where only visual input is available.

Many practical environments, such as embedded platforms, edge devices, and industrial inspection systems, demand models that are both lightweight and fast. In such cases, relying solely on visual inputs is often necessary. This highlights the importance of developing methods that improve fairness without introducing additional complexity during inference.

In this study, we introduce **N-TIDE**, a two-stage framework designed to distill the semantic knowledge of vision-language models into a compact image-only model. Our approach is inspired by Null-Text Inversion [23], which uses a specific text embedding to preserve image semantics during editing. Building on this idea, we define a neutral-text embedding that acts as a bias-corrective signal during training. This embedding is not required once training is complete, allowing the final model to remain free of textual dependencies.

N-TIDE consists of two main components. In the first stage, called *neutral-text inversion*, we construct a connection between CLIP's null-text and a trainable neutral-text embedding. By aligning the corresponding fused representations through a compact fusion module, we extract fairness-related information. In the second stage, we guide the image-only model to match the debiased CLIP features using cosine similarity loss. This can be interpreted as a form of knowledge transfer that conveys semantic structure while correcting for biased representations.

We further offer a conceptual understanding of our method based on the idea of Denoising Diffusion Implicit

*Equal contribution.

Models (DDIM) [29]. The progression from null-text to neutral-text embeddings is analogous to a denoising process, where the model refines its internal representation toward a more neutral and fair outcome. The contributions of this paper are summarized as follows:

- We present N-TIDE, a two-stage debiasing framework that enables fair image-only prediction by transferring semantic supervision from a vision-language model.
- We design a training-time mechanism that leverages the relationship between null and neutral text embeddings to inform fairness in visual representations.
- We provide a conceptual link between our approach and diffusion-based models, offering new insight into the process of learning unbiased features.

2. Related Work

2.1. Mitigating Bias

Bias-mitigation strategies are commonly categorized into three sequential stages: data preprocessing, model-level interventions, and postprocessing of decisions [4,8,9]. Data preprocessing techniques modify the dataset prior to training by applying oversampling, undersampling, or synthetic augmentation, allowing the model to learn a less biased representation. However, these techniques can be time-consuming [19] as each instance must be individually transformed. More importantly, when bias arises from the semantic content of the data or learned information within the model itself rather than class imbalance, simply adjusting sample counts does little to alleviate the underlying unfairness [6].

While statistical biases, such as demographic imbalances, can often be addressed through sampling strategies [18, 33, 34] or post-hoc calibration [13, 24], semantic biases embedded in visual content require deeper interventions at the representation or architectural level. Therefore, we focus on mitigating intrinsic image-based bias—particularly those arising from semantic cues within visual data—rather than statistical artifacts that can be corrected through conventional preprocessing or decision-level adjustment. To evaluate our approach, we conduct experiments on the FairFace dataset [18], which aims to minimize demographic imbalance through careful curation and balanced sampling.

2.2. Debiasing Unimodal Vision Models

In unimodal settings where only image inputs are available, traditional bias mitigation techniques also follow the standard three-stage taxonomy: pre-processing, in-processing, and post-processing [2,4]. In the pre-processing stage, oversampling, undersampling, reweighting, or synthetic data generation are used to rebalance datasets. In-processing approaches include adversarial training, group-wise regularization, or disentangled representation learning. Post-processing methods adjust prediction scores or thresholds after model inference.

Despite their utility, these methods often lack the capacity to incorporate semantic cues beyond pixel-level features, making it difficult to correct deeper, content-driven biases. To address this limitation, recent unimodal works attempt to implicitly suppress biased visual patterns during training. For example, FaceSaliencyAug [20] masks salient facial regions to reduce the model's reliance on biased attributes. DebiAN [21] alternates between discovering biased concepts and training a debiased classifier. DSA [25] modifies attention maps in Vision Transformers to suppress spurious correlations.

Unlike these methods, which rely entirely on image inputs, some recent approaches have attempted to introduce multimodal signals during training. We build upon this idea by incorporating textual semantics at training time, while maintaining unimodal inference.

2.3. CLIP-based Debiasing

CLIP-based debiasing research can be broadly categorized into two main approaches.

(1) **Neutral feature extraction using pre-trained CLIP encoders.** FairCLIP [22] constructs a similarity matrix between image and text embeddings and minimizes group-wise distributional discrepancies using the Sinkhorn distance, thereby leveraging CLIP's pretrained feature space for bias mitigation. FairerCLIP [5] directly operates on frozen image and text embeddings from CLIP, applying debiasing techniques at the representation level. SANER [15] neutralizes protected attribute words (e.g., replacing "woman" with "person"), passes the modified text through the CLIP text encoder, and uses the resulting embedding as a neutral reference. Similarly, Debiasing Vision-Language Models via Biased Prompts [3] and Bend-VLM [10] remove protected-attribute information from text embeddings via orthogonal projection or directional subtraction between attribute-augmented query pairs (e.g., "a photo of a male nurse" vs. "a photo of a female nurse").

(2) **Prompt-based attribute removal.** These methods modify the text prompts themselves to suppress bias-related information. SANER [15] introduces adversarial debiasing losses over modified prompts such as "A photo of a person" to suppress attribute-specific signals in the resulting embeddings. Debiasing Vision-Language Models via Biased Prompts [3] disentangles class and attribute semantics by refining prompts (e.g., "A photo of a doctor," "A photo of a male doctor," and "A photo of a female doctor") so that the embedding of "doctor" captures only class-related

information. BendVLM [10] adopts a similar strategy by generating attribute-augmented queries and removing the protected-attribute direction from the embedding space.

While these methods have demonstrated strong performance in debiasing vision-language models, they suffer from a key limitation: they require the presence of a text encoder at inference time and operate entirely within a multimodal setup. As such, they are not applicable to pure vision-only models that lack textual input or text-processing modules.

Our method, N-TIDE, addresses this limitation by introducing a hybrid training strategy: we leverage CLIP only during training to supervise the learning of a fair image encoder. This is achieved by constructing a pivot between a null-text and a neutral-text embedding, enabling text-informed debiasing without requiring prompts or text encoders at inference time. Unlike prior works that modify prompts or perform attribute subtraction in the embedding space, N-TIDE distills the debiased representation into a standalone image model via feature matching. This allows our model to benefit from multimodal supervision while maintaining full compatibility with unimodal deployment scenarios.

2.4. Text-Inversion Techniques

Null-Text Inversion [23] learns only a single “null-text” embedding—rather than directly optimizing any editing-related text embeddings—in order to preserve the original image’s semantic structure even when an editing prompt is applied. It freezes both the text encoder and the denoising backbone and optimizes only the null-text embedding with an MSE loss so that the reverse diffusion process reconstructs an image as close as possible to the original. By leveraging the deterministic sampling trajectory of DDIM [29], each noise-added intermediate latent is treated as a fixed point along a path between the original and fully noised representations, allowing the model to restore it—and thereby maintain image identity—even under strong conditional prompts.

In contrast, our N-TIDE starts from an unconditional objective and injects text-based bias correction into a unimodal image encoder. Specifically, we train a “neutral text embedding” for bias mitigation and combine it with a lightweight fusion module, applying text guidance only during training so that at inference time the model relies solely on its image encoder to produce fair predictions. Unlike Null-Text Inversion whose goal is to preserve image identity by learning only an uninformative embedding N-TIDE distills the effects of text-based debiasing into a pure image model via targeted fine-tuning.

2.5. Knowledge Distillation for Fairness

Research on fairness via Knowledge Distillation aims to reduce bias by transferring the teacher model’s softened prediction to the student model. For exmaple, in Fairness Without Demographics in Human-Centered Federated Learning [27], the teacher’s softmax outputs are ”softened” by increasing the temparature, then combined with the hard targets (ground-truth labels) to train the student model, while also verifying fairness metrics such as demographic parity and equal opportunity. In Fair Feature Distillation for Visual Recognition [17], MMD-based regularization is used to align each group’s feature distribution in the student model with the teacher’s average distribution, but there is no methodology to define and distill neutral features without language or text information.

3. Methods

3.1. Overall Architecture

The proposed N-TIDE framework consists of two stages: *Neutral-Text Inversion* and *Feature Matching*, as illustrated in Figure 1. In the first stage, we leverage the pretrained CLIP model to construct a pivoting mechanism between its representations under a *null-text embedding* and a trainable *neutral-text embedding*. During this stage, only the neutral-text embedding, a task-specific embedding, and a lightweight classifier including the fusion module are updated, while the CLIP encoders remain frozen. The fusion module combines features from the text and image encoders via element-wise addition, followed by a sequence of linear layers. We adopt addition instead of concatenation to mitigate the model’s tendency to overfit to image features, which we observed when using MLPs over concatenated inputs where the text embedding was often disregarded. This design encourages a more balanced integration of both modalities. The objective of this stage is to guide the model toward a text-informed, yet text-independent, representation of semantic fairness by aligning the fused features from both text conditions.

In the second stage, we train a unimodal image-only model using a feature matching objective. The image encoder is supervised to align its output features with those obtained from CLIP conditioned on the learned neutral-text embedding. This stage serves as a form of knowledge distillation, transferring the debiased representation learned in the first stage to the image-only model. All CLIP-related modules remain frozen, and only the image encoder and classifier are updated. The entire process ensures that the resulting model produces unbiased predictions without requiring a text encoder at inference time.

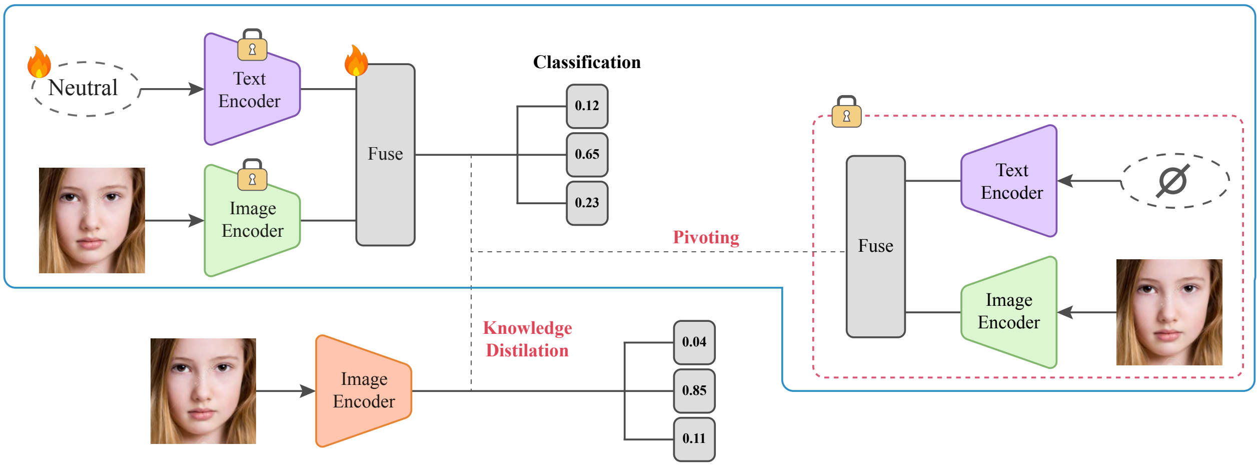

Figure 1. Overall architecture of N-TIDE. The purple and green encoders represent the pretrained CLIP text and image encoders, respectively. The model uses MSE loss between features from a neutral-text embedding and a null-text embedding to align representations. The learned representations are then distilled into an image-only model to encourage unbiased predictions. Components marked with a flame icon indicate trainable parameters, those marked with a lock icon are frozen during training, and ϕ denotes the null-text embedding.

3.2. Neutral-Text Inversion

Null-Text Inversion [23] aims to generate edited images that preserve the semantic content of the original input. To this end, it avoids optimizing text embeddings that are directly related to the editing prompt. Instead, it trains a special *null-text embedding*, representing an uninformative siganl for unconditional representation, while keeping the text encoder and other diffusion model components frozen. This includes the underlying denoising backbone used in models such as Stable Diffusion [26], and score-based generative models [30]. This embedding is optimized so that the denoising process reconstructs an image close to the original, even when conditioned on a prompt during reverse process.

$$\min_{\phi} \left\| \mathbf{z}^*_{t-1}(\phi) - \mathbf{z}_{t-1}(\mathbf{z}_T, \phi, \mathrm{c}) \right\|_2^2 \tag{1}$$

Inspired by this idea, we propose N-TIDE, which takes an opposite approach. While Null-Text Inversion focuses on maintaining image identity in the presence of strong conditional prompts by learning an unconditional text embedding, N-TIDE starts from an unconditional objective and seeks to inject text-based bias correction into a unimodal image encoder. Specifically, we train a *neutral text embedding* that encapsulates the effect of text-based guidance. This embedding is used only during training, enabling the image-only model to produce unbiased predictions without requiring a text encoder at inference time. The training objective, where ϕ and ψ denote the null-text embedding and *neutral-text embedding*, respectively, is defined as follows:

$$\min_{\psi} \|\mathbf{z}^*(\psi) - \mathbf{z}(\phi)\|_2^2 \tag{2}$$

In Null-Text Inversion, a null-text embedding is trained using an auxiliary mean squared error (MSE) loss to align the latent vectors obtained from the forward and backward (denoising) processes at corresponding time steps. In this setting, the forward process maps the original image to a noised latent vector, while the backward process predicts the noise (or score) to reverse this trajectory and reconstruct the original image.

Due to the deterministic nature of DDIM [29], each intermediate latent can be viewed as a point along a fixed trajectory between the original image and its corresponding noised representation. From this perspective, extracting a latent representation and reversing it toward a target label which reflects a subset of the image's semantics can be interpreted as a backward process of the DDIM. Consequently, DDIM inversion techniques can be adapted to enable controllable label prediction through guided manipulation of the latent space. See Section 3.4 for more details.

We construct two pairs: one consisting of an image and a null-text embedding, and another with the same image and a neutral-text embedding. For each pair, we apply an auxiliary loss to align the fused representations obtained from the same CLIP image and text encoders. By pivoting the fused features of the neutral-text pair toward those of the null-text pair, the model learns to emulate a neutral condition. This pivoting encourages the image-only model to make fairer predictions by implicitly incorporating text-guided debiasing [3, 5, 15, 22].

In our implementation, the fusion module is a lightweight MLP that aggregates image and text features via concatenation. During training, only the null-text embedding, the classifier for the specific task, and the fusion module are updated. The fusion module, along with the CLIP image and text encoders, is shared across both pairs during training, as illustrated in Figure 1.

3.3. Feature Matching

Our training process consists of two stages, where only a subset of layers is updated during the first stage. This design choice is motivated by the observation that directly fine-tuning CLIP can distort the semantic space learned from large-scale image-text pairs [16]. To preserve the pretrained semantic structure while adapting to a specific downstream task, we update only the null-text embedding and the classifier in the first stage. Once this training is complete, we proceed with a feature matching step.

For the image encoder, we use ResNet-50 [14], which is the same backbone architecture employed by CLIP. The classifier in the second stage retains the same structure as in the first stage. During this stage, we apply feature matching by aligning the representations produced by the unimodal image encoder with those from CLIP conditioned on the learned neutral-text embedding, using a cosine similarity loss. This contrasts with the MSE loss used in the first stage, where both representations originate from the same CLIP model and inherently share the same semantic space.

In contrast, the unimodal encoder has not been pretrained on large-scale image-text pairs and therefore requires guidance toward CLIP's pretrained semantic space. To achieve this, we apply a cosine similarity loss between features from the unimodal encoder and those from the frozen CLIP model, which are obtained by inputting the image and the learned neutral-text embedding. The null-text embedding used in the first stage is not involved in this step.

We adopt cosine similarity instead of MSE because it provides directional information in addition to magnitude alignment. This property is expected to help the unimodal encoder learn the target semantic space more accurately and efficiently. The resulting feature alignment serves as a regularization mechanism, encouraging the unimodal encoder to acquire a neutralized subspace of CLIP's semantic space. We interpret this feature matching step as a form of knowledge distillation.

3.4. DDIM Perspective on CLIP

We reinterpret our training framework through the lens of deterministic diffusion models, particularly Denoising Diffusion Implicit Models (DDIM) [29], to draw an analogy between CLIP's latent encoding and the trajectory of structured semantic manipulation.

DDIM defines a non-stochastic forward and reverse trajectory between a clean input $\mathbf{x}_0$ and a fully noised representation $\mathbf{x}_T$, where each intermediate latent $\mathbf{x}_t$ lies on a deterministic path defined by the noise prediction model $\boldsymbol{\epsilon}_\theta$. The reverse process at timestep t is given by:

$$\begin{aligned}\mathbf{x}_{t-1} = \frac{\sqrt{\alpha_{t-1}}}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \sqrt{1-\alpha_t}\cdot\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, c)\right) \\ + \sqrt{1-\alpha_{t-1}}\cdot\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, c)\end{aligned} \tag{3}$$

We draw a conceptual analogy: CLIP's transformation from raw images to entangled semantic embeddings parallels the DDIM forward process, while classification from those embeddings resembles a reverse denoising step. In our architecture, we incorporate a neutral-text embedding $\mathbf{t}_{\text{neutral}}$ into the CLIP fusion pipeline, producing fused representations $\mathbf{f} = \mathrm{MLP}(\mathbf{z}_{\text{img}} + \mathbf{z}_{\text{text}})$ that lie in a latent semantic space.

By aligning $\mathbf{f}_{\text{neutral}}$ with $\mathbf{f}_{\text{null}}$ (obtained with null-text embedding), we effectively guide the semantic encoding trajectory in a direction that reduces bias. This operation conceptually mirrors DDIM's guided reverse process, where conditioning steers latent transitions toward desired semantic targets.

Furthermore, while DDIM explicitly manipulates latents through additive noise, CNN-based encoders like CLIP implicitly project images through a sequence of learned nonlinear transformations. Each projection step moves the input toward high-level feature modes in the distribution, which can be interpreted as traversing a hierarchical latent manifold, which is similar to DDIM's forward process.

Our framing thus interprets CLIP's fusion mechanism as a one-step guided sampling, where neutral-text embedding provides directional alignment analogous to DDIM guidance. This perspective offers a theoretical justification for using neutral text as an intermediate anchor for debiasing, enabling controllable semantic transitions in the embedding space without requiring iterative refinement.

4. Experiments

4.1. Experiments Setup

4.1.1 Dataset

We conduct our experiments using the FairFace dataset [18], which contains approximately 108,500 facial images curated to support fairness-aware learning. The images were collected from Flickr's YFCC100M dataset [31] and manually verified to ensure quality and demographic diversity. FairFace provides annotations for race (seven categories: White, Black, Latino/Hispanic, East Asian, Southeast Asian, Indian, and Middle Eastern), gender (male, female), and age (eight age groups ranging from

Attribute	Model	Accuracy ↑	Equal Opportunity Difference ↓	Equalized Odds Difference ↓	Demographic Parity Difference ↓	Representation Bias Difference ↓
Race	ResNet50	0.857	0.024	0.017	0.019	0.079
	N-TIDE (ours)	0.867	0.020	0.015	0.021	0.021
	Improvement Rate	+1.17%	-16.67%	-11.76%	+10.53%	-73.42%
Gender	ResNet50	0.885	0.091	0.091	0.034	0.342
	N-TIDE (ours)	0.892	0.067	0.063	0.042	0.049
	Improvement Rate	+0.79%	-26.37%	-30.77%	+23.53%	-85.67%

Table 1. Results comparison between ResNet50 and N-TIDE. Each attribute (Race and Gender) represents the target label that the corresponding models were trained on. Accuracy shows relative gain (↑), while metrics marked with ↓ are better when lower. The reported Improvement Rate indicates the percentage change of N-TIDE compared to ResNet50. For bias(↓) metrics, more negative values reflect greater improvement.

children to seniors). The dataset is intentionally balanced across these attributes to prevent over-representation of any specific demographic group, thereby enabling unbiased evaluation in facial analysis tasks.

In our experiments, we focus on four race categories—White, Black, East Asian, and Indian—to ensure comparability with UTKFace [35], following the setup in [18]. The age attribute is excluded to concentrate on the core classification task. Our model is trained to predict race and gender from facial images, and FairFace is used as a benchmark to assess the effectiveness of the proposed N-TIDE framework in mitigating social bias within purely image-based vision models.

4.1.2 Metrics

Accuracy Accuracy is the proportion of samples that the model correctly predicts out of all predictions. In other words, it is defined as the number of times the model's predicted values match the true labels across all classes, divided by the total number of samples.

Equal Opportunity Difference Equal Opportunity Difference [13] measures the gap in a model's ability to detect positive cases (True Positive Rate) across different sensitive attribute groups. It computes each group's TPR and then takes the difference between the highest and lowest values, revealing whether some groups are systematically more likely to be correctly classified as positive than others.

$$\Delta_{\mathrm{EOp}} = \max_{g} \mathrm{TPR}_g - \min_{g} \mathrm{TPR}_g$$

Equalized Odds Difference Equalized Odds Difference [13] accounts for disparities in both True Positive Rate and False Positive Rate across groups. For every pair of groups, it computes the absolute difference in TPR and the absolute difference in FPR, averages those two gaps, and then takes the maximum over all pairs. This enforces fairness in both correct positive detections and incorrect positive errors simultaneously.

$$\Delta_{\mathrm{EO}} = \max_{g_1,g_2} \frac{1}{2}\Big(\big|\mathrm{TPR}_{g_1} - \mathrm{TPR}_{g_2}\big| + \big|\mathrm{FPR}_{g_1} - \mathrm{FPR}_{g_2}\big|\Big)$$

Demographic Parity Difference Demographic Parity Difference [12] assesses how differently the model predicts the positive class rate across groups. It calculates the proportion of samples predicted as positive for each group (positive prediction rate, PR [7]) and then takes the difference between the maximum and minimum proportions, indicating whether certain groups receive disproportionately more positive predictions.

$$\Delta_{\mathrm{DP}} = \max_{g} \mathrm{PR}_g - \min_{g} \mathrm{PR}_g$$

While Demographic Parity captures overall balance in positive decisions, it should be interpreted alongside Equal Opportunity Difference and Equalized Odds Difference to obtain a more complete picture of fairness, since those metrics account for group-specific performance gaps in true and false positive rates.

Representation Bias Difference Representation Bias Difference quantifies bias inherent in the data distribution itself. It first computes the mean feature vector for each group, then measures the cosine distance between every pair of group means. The largest distance reflects the greatest imbalance in representation in feature space.

$$\mathrm{RBD} = \max_{g_1,g_2}\Big(1 - \frac{\mu_{g_1} \cdot \mu_{g_2}}{\|\mu_{g_1}\| \, \|\mu_{g_2}\|}\Big)$$

This formulation is precisely the Maximum Mean Discrepancy (MMD) [11] between the two group distributions when using a cosine kernel, making RBD a special case of MMD restricted to comparing mean embeddings via cosine similarity [28].

4.1.3 SetUp

This experiment was conducted under the following common settings. We used an NVIDIA RTX A5000 GPU

Attribute	Model	Accuracy ↑	Equal Opportunity Difference ↓	Equalized Odds Difference ↓	Demographic Parity Difference ↓	Representation Bias Difference ↓
Race	ResNet50	0.857	0.024	0.017	**0.019**	0.079
	CLIP w/o T.E.	**0.867**	0.034	0.021	0.021	0.023
	CLIP w/ T.E.	0.865	0.032	0.020	0.020	0.025
	CLIP†+ random init.	0.865	0.034	0.021	0.020	0.021
	CLIP†+ ‘person’ init.	0.865	0.031	0.019	0.020	**0.019**
	CLIP w/ T.E + distil.	**0.868**	**0.021**	**0.016**	0.020	0.026
	N-TIDE w/o inversion	0.865	0.023	0.017	**0.018**	**0.020**
	N-TIDE (ours)	0.867	**0.020**	**0.015**	0.021	0.021
Gender	ResNet50	0.885	0.091	0.091	**0.034**	0.342
	CLIP w/o T.E.	0.891	0.073	0.069	0.046	0.187
	CLIP w/ T.E.	0.889	0.072	0.070	0.042	0.160
	CLIP†+ random init.	0.888	0.076	0.075	0.046	**0.116**
	CLIP†+ ‘person’ init.	0.889	0.069	**0.066**	0.043	0.119
	CLIP w/ T.E + distil.	**0.896**	0.073	0.073	0.039	0.131
	N-TIDE w/o inversion	**0.896**	**0.069**	0.069	**0.033**	0.139
	N-TIDE (ours)	0.892	**0.067**	**0.063**	0.042	**0.049**

Table 2. Results evaluating the effect of text encoder (T.E.) usage and neutral-text embedding initialization on debiasing performance. We compare baseline models (ResNet50 and CLIP) against N-TIDE under two neutral-text embedding initialization settings: (1) *random init.*, using a random vector; and (2) *‘person’ init.*, using the embedding of the ‘person’ token from the CLIP text encoder. *N-TIDE w/o inversion* excludes the null-text inversion process to isolate its effect, while still using the neutral-text embedding for the prompt. Models marked with † are trained with null-text inversion only, without feature matching. Best and second-best results are highlighted in red and blue, respectively.

and subsampled the FairFace dataset, selecting four classes (White, Black, East Asian, Indian) out of the original seven. The total dataset comprised 60,000 images, split into 45,000 for training, 8,000 for validation, and 7,000 for testing. Input images were augmented with RandomResizedCrop, RandomHorizontalFlip, RandomRotation, and ColorJitter. The batch size was fixed at 64. We trained all models using the AdamW optimizer with a Cosine Decay learning-rate scheduler, and applied label smoothing for regularization.

For the baseline (student) model, we employed a ResNet-50 pretrained on ImageNet and trained it for 15 epochs. The learning rate for backbone parameters was set to 1×10^{-5}, and for the projection and classification heads (MLP) to 1×10^{-4}, decaying to a minimum of 1×10^{-5} via cosine decay without any warmup. Weight decay was fixed at 1×10^{-2}.

For the teacher model, we used the CLIP visual backbone (RN50) and performed knowledge distillation for 10 epochs. The initial learning rate was 1×10^{-4}, decayed to 1×10^{-5} with cosine decay (no warmup), and weight decay was set to 1×10^{-5}.

4.2. Primary Results

Table 2 presents the results of our debiasing experiments on race and gender prediction. It includes evaluations of CLIP-based baselines, both with and without the text encoder, as well as comparisons under different neutral-text initialization schemes. We also investigate the impact of text-encoder distillation and the effect of removing the inversion step in N-TIDE. Finally, we compare the performance of the pure image-based ResNet50 with our complete N-TIDE pipeline.

CLIP-based Baselines Analysis First, comparing *CLIP without T.E.* to *CLIP with T.E.*, adding the text encoder causes a slight drop in accuracy (0.867 → 0.865) but yields consistent improvements across all bias metrics (Equal Opportunity Difference: 0.034 → 0.032; Equalized Odds Difference: 0.021 → 0.020), indicating that simply incorporating T.E. can enhance fairness. In gender prediction, accuracy decreases marginally from 0.891 to 0.889, while Demographic Parity Difference falls from 0.046 to 0.042 and Representation Bias Difference from 0.187 to 0.160.

Next, when comparing *CLIP w/ T.E. + random init.* against *CLIP w/ T.E. + ‘person’ init.*, the model using only random initialization shows negligible change in performance, whereas initializing with the ‘person’ token further improves race prediction metrics (Equal Opportunity Difference: 0.031; Equalized Odds Difference: 0.019; Representation Bias Difference: 0.019), confirming that the initialization strategy critically impacts final performance.

Distillation Analysis The distillation model effectively combines the ImageNet1K pretraining of the student model with CLIP’s text–image alignment knowledge, leading to improved overall performance. Applying *CLIP w/ T.E. +*

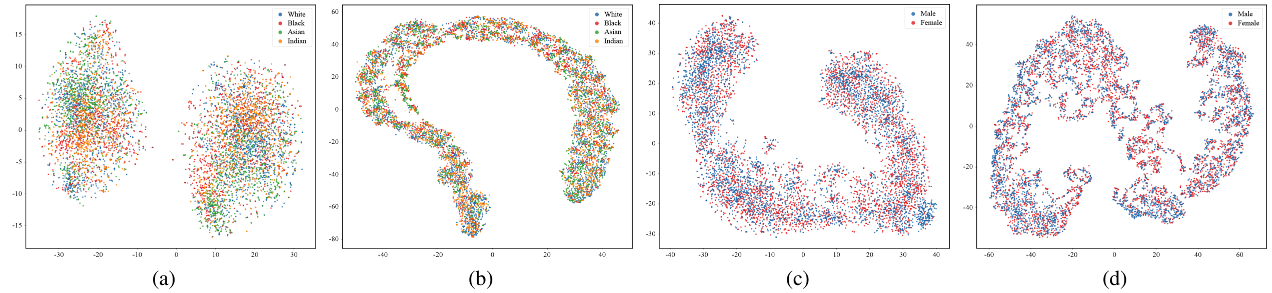

Figure 2. t-SNE [32] visualizations of features extracted from FairFace images after passing through each model's fusion module, with initialization using PCA [1]. Subfigures (a) and (b) show models trained on the **gender** attribute: (a) ResNet50, (b) N-TIDE. In both, colors indicate different **race** groups. Subfigures (c) and (d) show models trained on the **race** attribute: (c) ResNet50, (d) N-TIDE. Here, colors represent different **gender** groups. These visualizations illustrate how each model encodes demographic features and reveal differences in clustering and feature disentanglement between architectures.

distill. raises race prediction accuracy slightly from 0.865 to 0.868 and reduces the Equalized Odds Difference to 0.016 (with an Equal Opportunity Difference of 0.021), but Representation Bias Difference actually increases to 0.026. Although CLIP with T.E. and distillation often ranks second on race, its performance is nearly the same. For gender prediction, accuracy improves from 0.889 to 0.896 and Demographic Parity Difference decreases from 0.043 to 0.039, yet Representation Bias Difference rises to 0.131—demonstrating that its bias-metric performance for gender remains poor despite accuracy gains.

Meanwhile, when compared with the N-TIDE model without the inversion step (*N-TIDE w/o inversion*), both approaches deliver similar gains in Equal Opportunity, Equalized Odds, and Demographic Parity metrics, but Representation Bias Difference remains high. Specifically, N-TIDE w/o inversion lowers race Demographic Parity Difference to 0.018 (accuracy 0.865, RBD 0.020), whereas the full N-TIDE (ours) achieves an Equal Opportunity Difference of 0.020 and an Equalized Odds Difference of 0.015. In gender prediction, N-TIDE w/o inversion reduces Demographic Parity Difference to 0.033, and the complete N-TIDE further drives Representation Bias Difference down to 0.049. These results suggest that while distilling neutral-text embeddings can partially improve fairness metrics, explicitly incorporating our neutral-text inversion process is essential to effectively mitigate bias at the representation level.

N-TIDE without Inversion vs N-TIDE Comparing *N-TIDE w/o inversion* with the full N-TIDE, adding the inversion step reduces race prediction metrics—Equal Opportunity Difference from 0.023 to 0.020, Equalized Odds Difference from 0.017 to 0.015, and Demographic Parity Difference from 0.018 to 0.015. Notably, for gender prediction, the Representation Bias Difference drops significantly from 0.116 to 0.049, clearly demonstrating that the inversion step plays a crucial role in effectively mitigating bias at the model's representation level.

ResNet50 vs N-TIDE As shown in Table 1, N-TIDE's combination of neutral-text embedding and inversion provides clear advantages over the image-only ResNet50. For race, accuracy improves from 0.857 to 0.867; Equal Opportunity Difference from 0.024 to 0.020; Equalized Odds Difference from 0.017 to 0.015; and Representation Bias Difference plummets from 0.079 to 0.021. In gender prediction, ResNet50's accuracy of 0.885 (Demographic Parity Difference 0.034, RBD 0.342) is outperformed by N-TIDE's 0.892 accuracy and an RBD of just 0.049—demonstrating that N-TIDE can substantially reduce distributional bias while maintaining or improving overall accuracy.

Representation Space Analysis To further understand the effect of our proposed debiasing method, we examine the internal feature representations of both the baseline and N-TIDE models. Specifically, we pass input samples through each model and extract the feature embeddings after the fusion module. We then visualize these embeddings using t-SNE [32] with PCA initialization.

Figure 2 presents the resulting visualizations. Each subfigure displays samples grouped by the attribute that is *not* the classification target—i.e., for models trained on gender, we color points by race, and vice versa. In Figure 2a, the ResNet50 model exhibits partial clustering based on the protected attribute (e.g., race), indicating that the model's internal representations still encode demographic information in a structured way. This reveals a potential source of bias leakage.

In contrast, Figure 2b shows that embeddings from the N-TIDE model appear more uniformly distributed, with no clear separation along the protected attribute. This suggests that the learned features are less entangled with non-target demographic signals, providing qualitative evidence that N-TIDE promotes fairer representations by neutralizing attribute-specific structure in the feature space.

Figures 2c and 2d, which correspond to models trained on race and visualized by gender, show minimal differences. In both cases, the feature distributions appear randomly mixed. This result is likely not due to successful debiasing, but rather stems from the inherent characteristics of the gender attribute: it is binary (male, female) and nearly balanced in distribution. As a result, models trained on race naturally exhibit low bias with respect to gender, regardless of debiasing strategy. This setting makes it difficult for bias metrics to capture meaningful disparities, potentially masking subtle representation-level biases.

In summary, our experiments show that simply applying the text encoder and initializing neutral-text embeddings leads to only modest improvements in fairness metrics for CLIP-based models. Incorporating text-encoder distillation further reduces some measures (e.g., Equalized Odds Difference), but at the cost of increased Representation Bias Difference. In contrast, the full N-TIDE pipeline—with neutral-text inversion as a core component—consistently reduces all bias metrics while preserving or slightly improving overall accuracy. As shown in Table 1, N-TIDE reduces Representation Bias for race from 0.079 to 0.021 (over 73% reduction) and for gender from 0.342 to 0.049 (over 85% reduction) compared to ResNet50.

These findings clearly validate that neutral-text embedding combined with inversion-based prompt tuning can substantially mitigate distributional bias in image classification models.

5. Limitations

While N-TIDE introduces a novel distillation-based debiasing framework that transfers CLIP's semantic knowledge into a vision-only model, our study has several limitations.

First, our reinterpretation of CLIP's embedding behavior through the lens of Denoising Diffusion Implicit Models (DDIM) serves as an insightful conceptual analogy. However, this connection remains intuitive rather than theoretically grounded. We do not provide a formal mathematical proof that directly links CLIP's latent encoding trajectory to a deterministic diffusion process. Establishing this connection remains an open direction for future theoretical research.

Second, we limited our experiments to the FairFace dataset, which is relatively balanced across demographic groups. While this setting is suitable for evaluating semantic bias, it does not reflect real-world data imbalances. Due to project constraints, we were unable to test N-TIDE on statistically imbalanced datasets such as UTKFace, which would better assess the method's robustness to distributional skew.

Third, although N-TIDE consistently improves fairness metrics across race and gender classification tasks, the magnitude of improvement is modest in many cases. This indicates that while our inversion-guided distillation adds meaningful fairness signals, additional architectural or training-level interventions may be required to achieve more substantial gains.

Finally, our approach is tailored to mitigating social biases such as race and gender. It remains unclear whether the same neutral-text inversion mechanism can be extended to address other types of biases, such as those arising from texture reliance in CNNs or inappropriate correlations. Broadening the scope of our framework to such biases is an important avenue for future investigation.

6. Conclusion

We presented **N-TIDE**, a two-stage debiasing framework that transfers CLIP's semantic supervision into a standalone image-only model. By learning a trainable neutral-text embedding aligned with CLIP's null-text condition, our method enables fairness-aware training without requiring text input at inference time.

Through fusion-based feature matching, N-TIDE reduces reliance on biased visual cues and promotes more balanced representations. We further offered a diffusion-inspired interpretation of CLIP's latent dynamics to conceptually support our alignment strategy.

Experiments on the FairFace dataset demonstrate that N-TIDE achieves consistent improvements across multiple fairness metrics while preserving classification accuracy. Nonetheless, limitations remain: the diffusion-based perspective lacks formal proof, the method's effectiveness on statistically imbalanced datasets like UTKFace was not tested, and its applicability to non-social biases (e.g., texture bias) is unexplored.

Despite these constraints, N-TIDE provides a promising foundation for integrating multimodal debiasing signals into unimodal models, opening avenues for broader applications in fair vision systems.

References

[1] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010. 8

[2] Feng Chen, Liqin Wang, Julie Hong, Jiaqi Jiang, and Li Zhou. Unmasking bias in ai: A systematic review of bias detection and mitigation strategies in electronic health record-based models, 2024. 2

[3] 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. 1, 2, 4

[4] Sepehr Dehdashtian, Ruozhen He, Yi Li, Guha Balakrishnan, Nuno Vasconcelos, Vicente Ordonez, and Vishnu Naresh Boddeti. Fairness and bias mitigation in computer vision: A survey, 2024. 2

[5] Sepehr Dehdashtian, Lan Wang, and Vishnu Naresh Boddeti. Fairerclip: Debiasing clip's zero-shot predictions using functions in rkhss. *arXiv preprint arXiv:2403.15593*, 2024. 1, 2, 4

[6] Samuel Dooley, Rhea Sukthanker, John Dickerson, Colin White, Frank Hutter, and Micah Goldblum. Rethinking bias mitigation: Fairer architectures make for fairer face recognition. *Advances in Neural Information Processing Systems*, 36:74366–74393, 2023. 2

[7] Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact, 2015. 6

[8] Emilio Ferrara. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1):3, 2023. 1, 2

[9] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 329–338, 2019. 2

[10] Walter Gerych, Haoran Zhang, Kimia Hamidieh, Eileen Pan, Maanas Sharma, Thomas Hartvigsen, and Marzyeh Ghassemi. Bendvlm: Test-time debiasing of vision-language embeddings, 2024. 2, 3

[11] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. 6

[12] Xiaotian Han, Zhimeng Jiang, Hongye Jin, Zirui Liu, Na Zou, Qifan Wang, and Xia Hu. Retiring δdp: New distribution-level metrics for demographic parity, 2023. 6

[13] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016. 2, 6

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[15] Yusuke Hirota, Min-Hung Chen, Chien-Yi Wang, Yuta Nakashima, Yu-Chiang Frank Wang, and Ryo Hachiuma. Saner: Annotation-free societal attribute neutralizer for debiasing clip. *arXiv preprint arXiv:2408.10202*, 2024. 1, 2, 4

[16] Hoin Jung, Taeuk Jang, and Xiaoqian Wang. A unified debiasing approach for vision-language models across modalities and tasks. *Advances in Neural Information Processing Systems*, 37:21034–21058, 2024. 5

[17] Sangwon Jung, Donggyu Lee, Taeeon Park, and Taesup Moon. Fair feature distillation for visual recognition, 2021. 3

[18] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1548–1558, 2021. 2, 5, 6

[19] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9012–9020, 2019. 2

[20] Teerath Kumar, Alessandra Mileo, and Malika Bendechache. Facesaliencyaug: mitigating geographic, gender and stereotypical biases via saliency-based data augmentation. *Signal, Image and Video Processing*, 19(1):1–11, 2025. 2

[21] Zhiheng Li, Anthony Hoogs, and Chenliang Xu. Discover and mitigate unknown biases with debiasing alternate networks. In *European Conference on Computer Vision*, pages 270–288. Springer, 2022. 2

[22] Yan Luo, Min Shi, Muhammad Osama Khan, Muhammad Muneeb Afzal, Hao Huang, Shuaihang Yuan, Yu Tian, Luo Song, Ava Kouhana, Tobias Elze, et al. Fairclip: Harnessing fairness in vision-language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12289–12301, 2024. 1, 2, 4

[23] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6038–6047, 2023. 1, 3, 4

[24] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *Advances in neural information processing systems*, 30, 2017. 2

[25] Yao Qiang, Chengyin Li, Prashant Khanduri, and Dongxiao Zhu. Fairness-aware vision transformer via debiased self-attention. In *European Conference on Computer Vision*, pages 358–376. Springer, 2024. 2

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 4

[27] Shaily Roy, Harshit Sharma, and Asif Salekin. Fairness without demographics in human-centered federated learning, 2024. 3

[28] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and

rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5), Oct. 2013. 6

[29] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 3, 4, 5

[30] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 4

[31] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 5

[32] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 8

[33] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013. 2

[34] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018. 2

[35] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5810–5818, 2017. 6