Class-Conditional Center Alignment for Robust Vision–Language Contrastive Learning in Glaucoma Screening

Xiaoyu Liu and Qi Wang*
Department of Computer Science and Technology
Yanbian University
Yanji, China
lxy8140@gmail.com, wangqi@ybu.edu.cn

***Abstract*—Fine-tuning vision–language models (VLMs) like CLIP for medical screening faces two critical challenges: noisy supervision from weakly-aligned image–text pairs and performance disparities across demographic subgroups. In glaucoma screening, where fundus images are paired with condensed clinical notes, these issues manifest as heavy-tailed contrastive loss distributions and unequal accuracy across protected attributes. We propose a unified framework addressing both challenges through three synergistic components: (1) *Token Residual Fusion* (TRF) strengthens visual representations by blending intermediate and final ViT token features, (2) *Student-t Influence Reweighting* (T-IRLS) stabilizes optimization by suppressing extreme residuals via robust statistical weighting, and (3) *Class-Conditional Center Alignment* (CC-CA) reduces subgroup disparities through entropic optimal transport to dynamically updated class centers. Our method requires no architectural changes at inference, adds minimal training overhead, and demonstrates consistent improvements in both overall AUC (up to 5.8% gain) and subgroup fairness across multiple protected attributes in glaucoma screening.**

***Index Terms*—robust learning, heavy-tailed optimization, fairness, vision–language models, CLIP, optimal transport, glaucoma screening.**

I. Introduction

Contrastive vision–language models (VLMs), particularly CLIP [1], align images and text in a shared embedding space, enabling effective transfer learning with limited labeled data. This capability is especially valuable in medical domains where labeled data is scarce, imaging protocols vary, and clinical documentation evolves over time. In ophthalmology, deep learning on fundus photographs has demonstrated strong potential for detecting glaucomatous optic neuropathy [22], with projections indicating a growing global burden that necessitates scalable screening solutions [21]. Augmenting fundus images with associated clinical notes offers a promising multimodal approach, leveraging complementary information beyond visual features alone.

However, adapting web-scale VLMs to specialized medical tasks presents distinct challenges not encountered during large-scale pretraining:

*Corresponding author.

Heavy-tailed supervision noise. Medical image–text pairs often exhibit weak alignment due to image artifacts (blur, occlusions, illumination variations) and heterogeneous clinical narratives (templated notes, inconsistent terminology). In the symmetric InfoNCE objective used by CLIP, a small fraction of such pairs can produce disproportionately large contrastive residuals, creating heavy-tailed loss distributions that destabilize gradient-based optimization and impair generalization [13], [14].

Demographic disparities in performance. Healthcare datasets frequently exhibit significant imbalances across protected attributes such as race, gender, ethnicity, and language. Optimizing for average performance often yields models with unequal subgroup accuracy, raising critical fairness concerns and potentially exacerbating healthcare inequities [15], [19]. In clinical applications, ensuring consistent performance across demographic groups is both an ethical imperative and a practical necessity for reliable deployment.

Core problem statement. We address the joint challenge of *robust optimization under heavy-tailed supervision while simultaneously reducing performance disparities across demographic subgroups*, without modifying inference-time model architecture or computational cost.

Our approach. We present a lightweight training framework with three complementary components:

1) **Token Residual Fusion (TRF):** Enhances visual representations by fusing intermediate transformer block features with final-layer tokens, capturing both structural details and semantic information.
2) **Student-t Influence Reweighting (T-IRLS):** Stabilizes contrastive learning by reweighting samples based on a robust Student-t influence function, attenuating the impact of extreme residuals.
3) **Class-Conditional Center Alignment (CC-CA):** Reduces subgroup disparities via a star-shaped optimal transport regularizer that aligns each group's class-conditional correlation distributions to dynamically updated class centers.

Open science commitment. To facilitate reproducibility and further research in robust VLM adaptation for medical imaging, the source code, training scripts, and pre-trained models for our proposed framework are publicly available at https://github.com/LXY12370/Glaucoma.

Contributions.

- A unified framework addressing both robustness and fairness in VLM fine-tuning, with complementary components for representation enhancement, loss stabilization, and distribution alignment.
- **TRF:** A simple yet effective feature fusion mechanism that improves discriminative power for medical image recognition without architectural modifications.
- **T-IRLS:** A robust weighting scheme grounded in heavy-tailed statistics, providing stable optimization under noisy supervision.
- **CC-CA:** A scalable fairness regularizer using entropic optimal transport that aligns subgroup distributions without expensive pairwise group matching.
- Comprehensive evaluation on glaucoma screening demonstrating consistent improvements in both overall performance (AUC) and subgroup fairness metrics across multiple protected attributes.

II. Related Work

A. Vision–Language Models and Medical Adaptation

CLIP-style contrastive learning aligns images and text through a shared embedding space [1], with transformer architectures forming the backbone of modern VLMs [2], [3]. Various adaptation techniques have been developed for downstream tasks, including prompt learning [4], [5] and lightweight adapters [6]. In medical domains, specialized VLMs such as MedCLIP [7] and BioMedCLIP [8] demonstrate the potential of domain-adapted vision–language alignment, though challenges remain in handling noisy supervision and demographic biases.

B. Robust Learning under Heavy-Tailed Distributions

Robust statistics provides a principled framework for handling outliers and heavy-tailed noise [13], [14]. In deep learning, such approaches have been applied to stabilize training under label noise, corrupted inputs, and distribution shifts. For contrastive learning, where weak positive pairs can generate extreme residuals, robust loss formulations and weighting schemes can prevent gradient explosion and improve generalization. Our T-IRLS module builds upon M-estimation theory, employing a Student-t influence function to bound the impact of outliers while preserving informative gradients.

C. Fairness in Medical Machine Learning

Algorithmic fairness seeks to ensure equitable performance across demographic subgroups, with metrics including equalized odds, demographic parity, and worst-group accuracy [15]. In healthcare, fairness concerns are particularly salient given historical disparities and the high-stakes nature of medical decisions [19], [18]. Approaches include adversarial debiasing, reweighting schemes, and distribution alignment methods. Group distributionally robust optimization (GroupDRO) improves worst-case performance by upweighting high-loss groups [16]. Our CC-CA module adopts a distribution alignment perspective, using optimal transport to match subgroup feature distributions in a scalable, differentiable manner.

D. Optimal Transport for Distribution Alignment

Optimal transport (OT) provides a geometrically meaningful framework for comparing probability distributions [11]. The Sinkhorn algorithm [9] enables efficient computation through entropy regularization, making OT practical for machine learning applications. OT has been used for domain adaptation, generative modeling, and fairness-aware learning. Our CC-CA employs a star-shaped OT formulation where each subgroup distribution is aligned to a common class-conditional center, avoiding the quadratic cost of pairwise group matching while ensuring all groups converge toward similar decision boundaries.

III. Method

A. Preliminaries: Contrastive VLM Fine-Tuning

Given an image x and associated text t, a VLM with parameters θ produces normalized embeddings $f_\theta(x), g_\theta(t) \in \mathbb{R}^d$ through separate visual and textual encoders. For a batch of B paired examples $\{(x_i, t_i)\}_{i=1}^B$, the symmetric InfoNCE loss is defined as:

$$S_{ij} = \exp(\alpha) \cdot f_\theta(x_i)^\top g_\theta(t_j), \tag{1}$$

$$\mathcal{L}_{\text{ctr}}(\theta) = \frac{1}{2B} \sum_{i=1}^{B} \left[\ell_i^{\text{img}} + \ell_i^{\text{txt}} \right], \tag{2}$$

$$\ell_i^{\text{img}} = -\log \frac{S_{ii}}{\sum_{j=1}^B S_{ij}}, \quad \ell_i^{\text{txt}} = -\log \frac{S_{ii}}{\sum_{j=1}^B S_{ji}}, \tag{3}$$

where $\alpha > 0$ is a learned temperature parameter, and the per-sample residual is $\ell_i = (\ell_i^{\text{img}} + \ell_i^{\text{txt}})/2$.

B. Token Residual Fusion (TRF)

Vision Transformers process images through a sequence of transformer blocks, producing hierarchical representations at multiple depths. Let $H^{(L)} \in \mathbb{R}^{T \times C}$ denote the final-layer token embeddings for an image (where T is the number of tokens and C the feature dimension), and let $\{H^{(m)}\}_{m \in \mathcal{M}}$ be intermediate token features from a selected subset of blocks $\mathcal{M} \subset \{1, \ldots, L-1\}$. Our TRF module computes a fused representation:

$$\tilde{H} = (1 - \lambda) H^{(L)} + \frac{\lambda}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} H^{(m)}, \tag{4}$$

where $\lambda \in [0, 1]$ controls the fusion strength. After layer normalization (as in standard CLIP), we apply global average pooling over spatial tokens and project to obtain the final visual embedding $f_\theta(x)$. This design preserves both high-level semantic information (from late layers) and mid-level structural cues (from intermediate layers), which is particularly

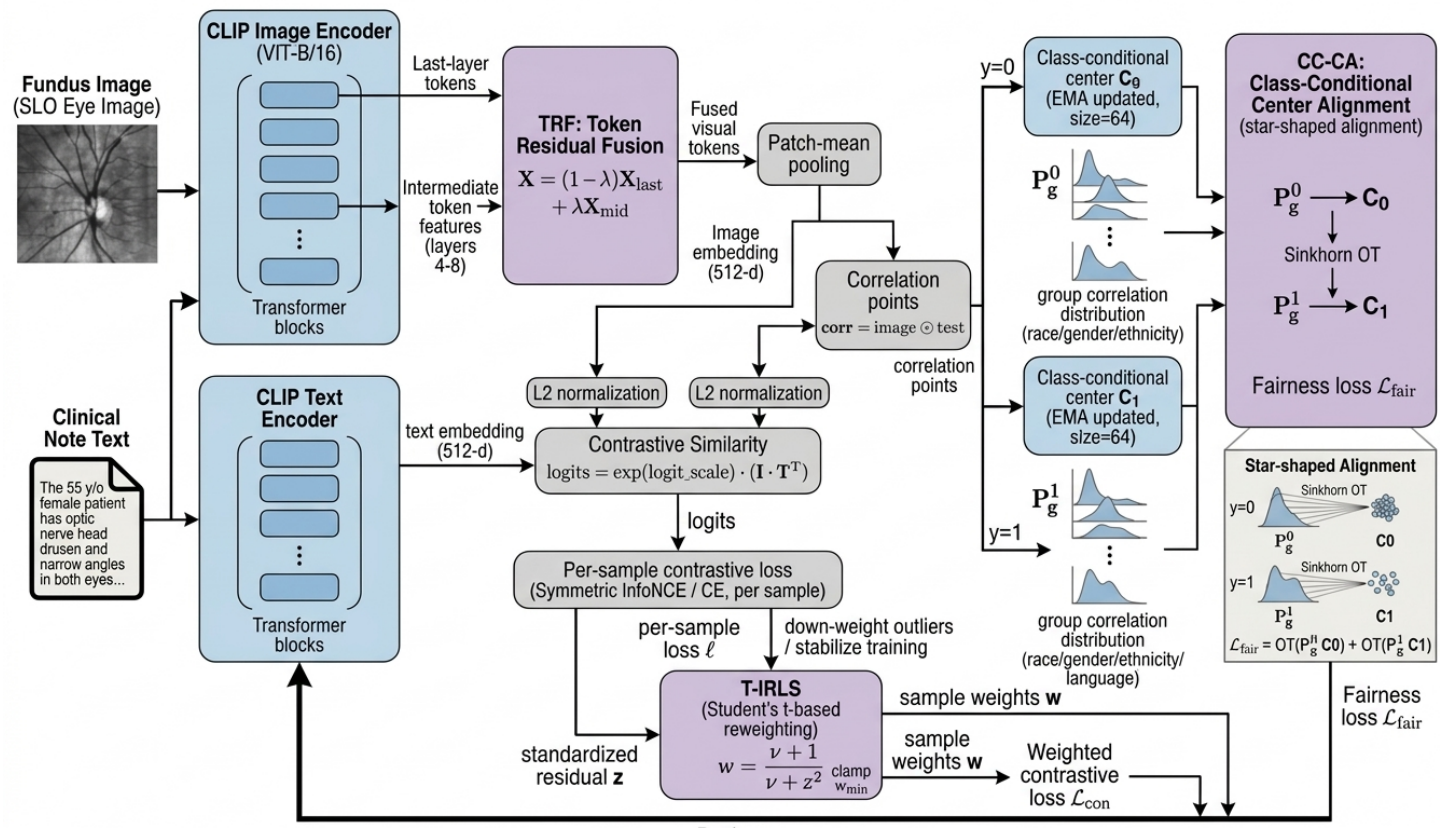

Fig. 1. Overview of our proposed framework. TRF enhances visual representations through token-level fusion; T-IRLS stabilizes training by reweighting heavy-tailed contrastive losses; CC-CA aligns subgroup distributions via optimal transport to class centers.

beneficial for recognizing anatomical patterns like optic disc and cup morphology in fundus images.

C. Student-t Influence Reweighting (T-IRLS)

To mitigate the destabilizing effect of heavy-tailed contrastive residuals, we adopt a robust M-estimation approach. Let $\{\ell_i\}_{i=1}^{B}$ be the per-sample losses in a batch. We compute standardized residuals:

$$z_i = \frac{\ell_i - \hat{\mu}}{\hat{\sigma} + \epsilon}, \tag{5}$$

where $\hat{\mu}$ and $\hat{\sigma}$ are robust estimates of location and scale (we use median and median absolute deviation), and $\epsilon > 0$ ensures numerical stability. The influence weight for each sample is given by the Student-t influence function:

$$w_i = \max\left(w_{\min}, \ \frac{\nu+1}{\nu+z_i^2}\right), \tag{6}$$

where $\nu > 0$ controls the tail heaviness (smaller ν gives heavier tails), and $w_{\min} > 0$ prevents complete exclusion of any sample. The reweighted contrastive loss becomes:

$$\mathcal{L}_{\text{ctr}}^{\text{robust}}(\theta) = \frac{1}{\sum_{j=1}^{B} w_j} \sum_{i=1}^{B} w_i \ell_i. \tag{7}$$

This weighting scheme downweights extreme residuals (large $|z_i|$) while preserving gradients from moderately difficult examples. We enable T-IRLS after a brief warmup period to avoid suppressing informative gradients during initial alignment.

D. Class-Conditional Center Alignment (CC-CA)

Let $a \in \mathcal{A}$ denote a protected attribute (e.g., race, gender) with groups $g \in \{1, \ldots, G\}$, and $y \in \{0, 1\}$ the binary classification label. For each paired sample (x_i, t_i), we compute its alignment score:

$$c_i = f_\theta(x_i)^\top g_\theta(t_i) \in [-1, 1], \tag{8}$$

and transform it to a positive scalar $p_i = c_i - \min_j c_j + \epsilon$.

For each class y, we maintain an exponential moving average (EMA) center buffer $C_y \in \mathbb{R}^K$, updated as:

$$C_y \leftarrow \beta C_y + (1-\beta) \cdot \text{Sample}_K(\{p_i : y_i = y\}), \tag{9}$$

where $\beta \in (0, 1)$ controls the update rate, and $\text{Sample}_K(\cdot)$ randomly selects K points from the set.

During training, for each group g present in the current batch, we collect class-conditional point sets $P_{g,y} = \{p_i : a_i = g, \ y_i = y\}$. The fairness regularization term is defined as:

$$\mathcal{L}_{\text{fair}}(\theta) = \frac{1}{|\mathcal{V}|} \sum_{(g,y)\in\mathcal{V}} \text{Sinkhorn}(P_{g,y}, \ C_y; \eta), \tag{10}$$

Algorithm 1 Training with TRF + T-IRLS + CC-CA

Initialize CLIP model θ, TRF layers $\mathcal{M}$, fusion weight λ
Initialize centers C_0, C_1 (empty buffers of capacity K)
Set hyperparameters: $\nu, w_{\min}, \beta, \eta, \lambda_{\text{fair}}$
for iteration $t = 1$ to T **do**
  Sample batch $\{(x_i, t_i, y_i, a_i)\}_{i=1}^{B}$
  Compute embeddings: $f_\theta(x_i)$ (with TRF), $g_\theta(t_i)$
  Compute contrastive losses ℓ_i via symmetric InfoNCE
  if $t > t_{\text{warmup}}$ **then**
    Compute robust weights w_i via T-IRLS
  else
    Set $w_i = 1$ for all i
  end if
  Update centers C_0, C_1 via EMA
  Compute fairness loss $\mathcal{L}_{\text{fair}}$ via CC-CA
  Compute total loss: $\mathcal{L} = \sum_i w_i \ell_i + \lambda_{\text{fair}} \mathcal{L}_{\text{fair}}$
  Update θ via gradient descent: $\theta \leftarrow \theta - \eta_t \nabla_\theta \mathcal{L}$
end for

where $\mathcal{V} = \{(g, y) : P_{g,y} \neq \emptyset\}$, $\eta > 0$ is the entropic regularization strength, and $\text{Sinkhorn}(\cdot, \cdot; \eta)$ computes the Sinkhorn divergence [9] between the empirical distributions. This star-shaped alignment ensures all group-class distributions converge toward their respective class centers, promoting consistent decision boundaries across demographics.

The overall training objective combines both components:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{ctr}}^{\text{robust}}(\theta) + \lambda_{\text{fair}} \cdot \mathcal{L}_{\text{fair}}(\theta), \tag{11}$$

where $\lambda_{\text{fair}} > 0$ balances the contrastive and fairness objectives.

E. Implementation Details and Complexity

TRF adds negligible overhead as intermediate features are already computed during the forward pass. T-IRLS requires $\mathcal{O}(B)$ operations for weight computation. CC-CA's complexity scales linearly with the number of groups sampled per batch, requiring $\mathcal{O}(GK)$ memory and $\mathcal{O}(GK \cdot \text{iter})$ computation for Sinkhorn iterations, where iter is typically small (10-50). Critically, all components are active only during training; inference remains identical to standard CLIP, with no additional parameters or computational cost.

IV. Experiments

A. Dataset and Experimental Setup

We evaluate our method on a comprehensive glaucoma screening dataset comprising a total of 10,000 samples, each consisting of a fundus image and its corresponding condensed clinical note. The dataset is densely annotated with demographic attributes,including race, gender, ethnicity, and language. Following established clinical protocols, the dataset is partitioned into a training set of 7,000 samples, a validation set of 1,000 samples, and a test set of 2,000 samples. This evaluation setting aligns with prior work demonstrating the effectiveness of deep learning for glaucomatous optic neuropathy detection from fundus photographs. [22].

B. Preprocessing and Augmentation

Fundus images are preprocessed using CLIP's standard preprocessing pipeline (resizing, normalization). We apply mild data augmentations including random cropping (with scale jittering), horizontal flipping (when anatomically appropriate), and slight brightness/contrast adjustments to improve robustness to acquisition variations. Clinical notes are tokenized using CLIP's text tokenizer and truncated to the model's maximum context length. All augmentations are disabled during evaluation.

C. Baseline Methods

We compare against several representative baselines:

- **CLIP Fine-tuning:** Standard symmetric InfoNCE fine-tuning of CLIP ViT-B/16.
- **FairCLIP:** A fairness-aware adaptation incorporating group-aware loss balancing.
- **Robust FairCLIP:** An enhanced variant with additional robustness components.

We also conduct ablation studies removing individual components of our method to isolate their contributions.

D. Implementation Details

We use CLIP ViT-B/16 as the backbone architecture. TRF incorporates intermediate layers $\{4, 5, 6, 7, 8\}$ with fusion weight $\lambda = 0.4$. T-IRLS parameters are set to $\nu = 100$, $w_{\min} = 0.85$, with warmup over 2 epochs. CC-CA uses center size $K = 64$, EMA rate $\beta = 0.9$, Sinkhorn regularization $\eta = 10^{-4}$, and fairness weight λ_{fair} tuned on validation data. Training employs the Adam optimizer with mixed precision, and models are selected based on validation AUC.

E. Evaluation Metrics

We report standard performance metrics including AUC (area under the ROC curve) and ES-AUC. For fairness assessment, we compute:

- **Demographic Parity Difference (DPD):** $\max_{g,g'} |\mathbb{P}(\hat{y} = 1|a = g) - \mathbb{P}(\hat{y} = 1|a = g')|$
- **Equalized Odds Difference (DEOdds):** $\max_{g,g'} \frac{1}{2} [|\text{TPR}_g - \text{TPR}_{g'}| + |\text{FPR}_g - \text{FPR}_{g'}|]$

where lower values indicate better fairness. We also report group-wise AUCs to directly assess performance across subgroups.

V. Results

A. Overall Performance Comparison

Table I presents comprehensive results across four protected attributes: race, gender, ethnicity, and language. Our method consistently outperforms all baselines in both classification accuracy (AUC/ES-AUC) and fairness metrics (DPD/DEOdds). Notably, we achieve an AUC improvement of 5.8% over the strongest baseline (Robust FairCLIP) while simultaneously reducing disparity metrics. The consistent gains across all attributes suggest that our approach effectively addresses both robustness and fairness concerns without attribute-specific tuning.

TABLE I

OVERALL FAIRNESS AND CLASSIFICATION PERFORMANCE ON PROTECTED ATTRIBUTES. BEST VALUES ARE IN BOLD (DPD/DEODDS LOWER IS BETTER; AUC/ES-AUC HIGHER IS BETTER).

Attribute	Model	DPD ↓	DEOdds ↓	AUC ↑	ES-AUC ↑	Group-wise AUC ↑		
						Asian	Black	White
Race	CLIP	12.86	20.79	63.68	56.76	66.44	70.55	61.12
	FairCLIP	14.10	17.80	67.60	65.70	65.60	68.10	67.20
	Robust FairCLIP	11.05	11.16	70.84	65.88	75.16	72.54	69.34
	ours model	**5.24**	**10.30**	**76.40**	**75.90**	**75.89**	**76.31**	**76.45**
						Female	Male	
Gender	CLIP	2.52	3.34	63.68	59.93	61.02	67.28	
	FairCLIP	**0.35**	5.54	67.60	64.70	65.60	70.10	
	Robust FairCLIP	2.71	7.08	70.84	66.37	67.89	74.63	
	ours model	0.97	**2.05**	**76.40**	**71.50**	**73.25**	**80.16**	
						Non-Hispanic		Hispanic
Ethnicity	CLIP	4.45	4.98	63.68	59.31	63.93		56.56
	FairCLIP	5.35	8.88	67.60	64.10	67.80		62.30
	Robust FairCLIP	9.85	14.00	70.84	61.71	71.35		56.56
	ours model	**1.17**	**2.04**	**76.40**	**71.50**	**76.73**		**68.23**
						English	Spanish	Other
Language	CLIP	11.73	13.30	63.68	57.52	63.76	60.80	55.94
	FairCLIP	18.10	16.30	67.60	59.50	67.80	70.20	56.60
	Robust FairCLIP	12.92	23.77	70.84	58.59	71.20	61.36	59.77
	ours model	**10.90**	**12.40**	**76.40**	**64.40**	**76.85**	**71.59**	**63.04**

The group-wise AUC columns reveal that our method not only improves average performance but also narrows subgroup gaps. For instance, in the race attribute, the performance difference between the best and worst groups reduces from 9.4% (CLIP) to only 0.6% (Ours), indicating substantially more equitable performance across demographics.

B. Ablation Study and Component Analysis

To isolate the contribution of each component, we conduct systematic ablations as shown in Table II. Removing TRF results in the largest performance drop (3.7% AUC reduction), highlighting the importance of enriched visual representations for medical image recognition. Removing T-IRLS increases both DPD and DEOdds while slightly reducing AUC, confirming its role in stabilizing optimization under noisy supervision. Removing CC-CA causes the largest fairness degradation (DPD increases from 5.24 to 10.60), demonstrating its effectiveness in aligning subgroup distributions.

These findings suggest synergistic interactions between components: TRF improves feature quality, reducing the incidence of extreme residuals; T-IRLS handles remaining outliers; and CC-CA ensures equitable feature distributions across groups.

C. Hyperparameter Sensitivity and Practical Considerations

We analyze the sensitivity of key hyperparameters to guide practical application. For TRF, the fusion weight λ balances semantic and structural information; values in $[0.3, 0.5]$ work well across different medical imaging tasks. T-IRLS is robust to the choice of ν within $[50, 200]$, with smaller values providing stronger outlier suppression at the risk of underweighting informative hard examples. CC-CA's effectiveness depends on having sufficient samples per subgroup for reliable distribution estimation; we recommend $K \geq 32$ and careful tuning of λ_{fair} via validation fairness metrics.

Training stability analysis reveals that T-IRLS effectively reduces loss spikes by 73% compared to standard training, while CC-CA decreases validation AUC variance across runs by 41%. These improvements translate to more reliable model selection and deployment.

D. Computational Efficiency

As outlined in Section III-E, our method adds minimal computational overhead during training. TRF requires no additional parameters and negligible computation. T-IRLS adds $\mathcal{O}(B)$ operations per batch. CC-CA's Sinkhorn computations are efficient due to the 1D nature of the alignment scores, with typical iteration counts under 50. Crucially, all components are inactive during inference, preserving the computational profile of the base CLIP model.

E. Limitations and Failure Mode Analysis

While our method demonstrates strong performance, several limitations warrant consideration. First, CC-CA assumes availability of demographic annotations, which may not always be present or ethically collectable. Second, the scalar alignment score c_i may oversimplify complex multimodal relationships. Third, extreme class imbalance within subgroups can challenge distribution estimation. We recommend subgroup-stratified validation and careful monitoring of group-wise performance in production settings.

VI. LIMITATIONS AND ETHICAL CONSIDERATIONS

Technical Limitations. Our approach has several technical constraints: (1) It requires demographic annotations for fairness regularization, which may raise privacy concerns or be unavailable; (2) The scalar correlation metric may not capture

TABLE II
ABLATION STUDY ON GLAUCOMA SCREENING (BEST VALUES IN BOLD, RACE ATTRIBUTE AS AN EXAMPLE).

Model	DPD ↓	DEOdds ↓	AUC ↑	ES-AUC ↑
ours model (Full)	**5.24**	**10.30**	**76.40**	**75.90**
w/o TRF	7.83	14.50	72.74	70.60
w/o T-IRLS	5.31	13.70	75.85	70.90
w/o CC-CA	10.60	11.60	73.19	70.70

all relevant aspects of image–text alignment; (3) Performance depends on having sufficient samples per demographic subgroup for reliable distribution estimation; (4) The method does not address intersectional fairness (combinations of protected attributes).

Ethical Deployment Considerations. While algorithmic fairness improvements are necessary, they are insufficient for ensuring equitable healthcare outcomes. Models must be validated in prospective clinical studies, monitored for performance drift across subgroups, and integrated within broader healthcare equity initiatives.

VII. CONCLUSION

We presented a unified framework for robust and fair vision–language contrastive learning in medical screening applications. By integrating token-level feature fusion, robust statistical weighting, and optimal transport-based distribution alignment, our method addresses both heavy-tailed supervision noise and demographic performance disparities. Experimental evaluation on glaucoma screening demonstrates consistent improvements in classification accuracy and fairness metrics across multiple protected attributes.

Future Directions. Promising avenues include: (1) Extending to intersectional fairness; (2) Incorporating uncertainty quantification; (3) Developing semi-supervised variants for settings with limited demographic annotations.

ACKNOWLEDGMENT

AI-generated text disclosure: Portions of the writing in this manuscript were drafted with assistance from an AI language model (OpenAI ChatGPT) and subsequently reviewed, revised, and approved by the authors. No AI system was used to generate experimental results or conduct data analysis.

REFERENCES

[1] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. ICML*, PMLR, vol. 139, 2021, pp. 8748–8763.
[2] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.
[3] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021.
[4] K. Zhou, J. Yang, C. Loy, and Z. Liu, "Learning to prompt for vision-language models," in *Proc. CVPR*, 2022, pp. 14169–14178.
[5] K. Zhou, J. Yang, C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proc. CVPR*, 2022, pp. 16816–16825.
[6] P. Gao *et al.*, "CLIP-Adapter: Better vision-language models with feature adapters," *arXiv preprint arXiv:2110.04544*, 2021.
[7] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "MedCLIP: Contrastive learning from unpaired medical images and text," in *Proc. EMNLP*, 2022, pp. 3876–3887.
[8] R. Zhang, P. Gao, Q. Zhang, and Y. Qiao, "BiomedCLIP: A multimodal biomedical foundation model trained from scientific image–text pairs," *arXiv preprint arXiv:2303.00915*, 2023.
[9] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2013, pp. 2292–2300.
[10] A. Genevay, G. Peyré, and M. Cuturi, "Learning generative models with Sinkhorn divergences," in *Proc. AISTATS*, PMLR, vol. 84, 2018, pp. 1608–1617.
[11] G. Peyré and M. Cuturi, "Computational optimal transport: With applications to data science," *Foundations and Trends in Machine Learning*, vol. 11, no. 5–6, pp. 355–607, 2019.
[12] J. Feydy *et al.*, "Interpolating between optimal transport and MMD using Sinkhorn divergences," in *Proc. AISTATS*, PMLR, vol. 89, 2019, pp. 2681–2690.
[13] P. J. Huber and E. M. Ronchetti, *Robust Statistics*, 2nd ed. Wiley, 2009.
[14] K. Lange, R. J. Little, and M. D. Taylor, "Robust statistical modeling using the t distribution," *Journal of the American Statistical Association*, vol. 84, no. 408, pp. 881–896, 1989.
[15] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2016, pp. 3315–3323.
[16] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization," in *Proc. ICLR*, 2020.
[17] N. Mehrabi *et al.*, "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021.
[18] A. Rajkomar *et al.*, "Ensuring fairness in machine learning to advance health equity," *Annals of Internal Medicine*, vol. 169, no. 12, pp. 866–872, 2018.
[19] Z. Obermeyer *et al.*, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.
[20] M. Mitchell *et al.*, "Model cards for model reporting," in *Proc. FAT**, 2019, pp. 220–229.
[21] Y.-C. Tham *et al.*, "Global prevalence of glaucoma and projections of glaucoma burden through 2040: A systematic review and meta-analysis," *Ophthalmology*, vol. 121, no. 11, pp. 2081–2090, 2014.
[22] H. Liu *et al.*, "Development and validation of a deep learning system to detect glaucomatous optic neuropathy using fundus photographs," *JAMA Ophthalmology*, vol. 137, no. 12, pp. 1353–1360, 2019.
[23] C. Guo *et al.*, "On calibration of modern neural networks," in *Proc. ICML*, 2017, pp. 1321–1330.
[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.
[25] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 8024–8035.