

Fairness-Aware Multimodal Representation Learning for Healthcare Data: A Visual–Text Alignment Framework for Glaucoma Diagnosis

Xiaoyu Liu

Yanbian University

Jinchun Piao

Yanbian University

Qi Wang

wangqi@ybu.edu.cn

Yanbian University

Research Article

Keywords: Fairness-aware multimodal learning, Visual–text alignment, Sinkhorn regularization·Student-t weighted optimization, Glaucoma diagnosis, Healthcare data analytics

Posted Date: November 19th, 2025

DOI: <https://doi.org/10.21203/rs.3.rs-7908475/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Fairness-Aware Multimodal Representation Learning for Healthcare Data: A Visual – Text Alignment Framework for Glaucoma Diagnosis

Xiaoyu Liu¹, Jinchun Piao¹, Qi Wang^{1*}

¹ Department of Computer Science and Technology, Yanbian University, Yanji, 133002, China

*Corresponding author: Qi Wang (wangqi@ybu.edu.cn)

Abstract

Fairness and accuracy often exhibit a trade-off in medical visual – language systems for glaucoma diagnosis, as existing approaches either sacrifice diagnostic performance for fairness or overlook disparities among minority subgroups. This study proposes a fairness-aware multimodal framework integrating four complementary components to align visual – text understanding with clinical relevance and equity. Sliding-window text segmentation preserves contextual coherence in lengthy clinical notes, mitigating information loss from fixed truncation. Visual-guided attention pooling leverages anatomical spatial priors from fundus images to guide text feature aggregation, ensuring clinically grounded visual – text correspondence. Student-t – based noise weighting adaptively downweights noisy samples disproportionately affecting minority groups, while Sinkhorn-based fairness regularization aligns visual – text similarity distributions across demographic attributes without degrading diagnostic signals. Evaluated on the Harvard-FairVLMed dataset against CLIP, FairCLIP, and Robust FairCLIP, the framework achieves an overall AUC of 72.65% (outperforming baselines by 2.81 – 9.97 pp). In fairness assessment, it reduces Demographic Parity Difference (DPD) to 9.65 (race), 2.18 (gender), 8.11 (ethnicity), and 11.87 (language); Other fairness indicators also achieved good results. Ablation studies confirm the contribution of each component. This work provides a clinically applicable and data-driven solution for fair multimodal medical AI, supporting reduced health disparities in glaucoma diagnosis and offering potential generalization to other ophthalmic conditions.

Keywords: Fairness-aware multimodal learning · Visual – text alignment · Sinkhorn regularization · Student-t weighted optimization · Glaucoma diagnosis · Healthcare data analytics

1. Introduction

Deep learning has substantially advanced medical visual

computing over the past decade, enabling automated analysis of ophthalmic images — such as fundus photography that captures anatomical structures like the optic disc, retinal nerve fiber layer, and macula — and optical coherence tomography (OCT) [1]-[3].

Convolutional and transformer-based architectures have improved early detection of vision-threatening diseases like diabetic retinopathy and glaucoma by extracting subtle visual patterns (e.g., optic disc cupping, retinal nerve fiber layer thinning) often missed by human observers [4],[5]. However, most models rely solely on imaging data, ignoring complementary clinical text (e.g., patient history of intraocular pressure, examination notes on visual field defects). This unimodal limitation not only reduces diagnostic interpretability but also harms performance consistency across diverse patient populations (e.g., varying retinal anatomy among ethnic groups) [6],[7].

Multimodal learning addresses these limitations by integrating heterogeneous data sources, enhancing diagnostic robustness and interpretability through cross-modal validation [8]-[10]. Vision-language (VL) models like CLIP [11] and BLIP [12], which pre-train visual and text encoders to align representations in a shared space, have inspired medical AI approaches that fuse fundus images with clinical text [13][15]. For glaucoma diagnosis specifically, combining high-resolution fundus images (capturing structural cues) with clinical narratives (recording functional symptoms) enables contextualized reasoning — reducing false positives from normal anatomical variations (e.g., small optic discs in Asian populations) and providing interpretable evidence for clinical decision-making.

Despite these advances, multimodal medical visual computing for glaucoma diagnosis still faces three unresolved challenges, particularly when applied to diverse patient cohorts:

Spatial-semantic misalignment in long clinical text: Fundus images encode pathology through spatial anatomical features (e.g., regional optic disc cupping), while clinical notes are sequential and often exceed the token limits of VL models (e.g., CLIP’s 77-token constraint). Fixed truncation of long notes discards critical

context (e.g., “progressive cupping over 6 months”), and naive fusion fails to map text descriptions to corresponding visual regions, weakening cross-modal synergy [16][17].

Noise-induced cross-modal inconsistency: Imaging artifacts (e.g., motion blur in fundus scans) and ambiguous text (e.g., vague descriptions of “mild cupping”) often introduce conflicting signals between modalities. Existing denoising methods focus on single-modal quality improvement but rarely enforce cross-modal consistency, undermining the reliability of diagnostic decisions [18].

Demographic bias inherited from VL pre-training: VL models pre-trained on web-scale data inherit social and visual biases (e.g., overrepresentation of White patients’ fundus images), leading to unequal performance across populations with distinct retinal characteristics (e.g., thicker retinal nerve fiber layers in Black patients)[19][21].

The Harvard-FairVLMed dataset [14], which provides paired fundus images, clinical notes, and detailed demographic annotations (race, gender, ethnicity, language), enables a unified evaluation of diagnostic accuracy and fairness to address this gap.

To tackle these challenges, we propose a fairness-aware multimodal framework for glaucoma diagnosis, built upon CLIP and tailored to medical visual computing needs. The framework integrates three key innovations, each addressing a specific challenge:

Visual-guided text segmentation and attention pooling: To resolve long-text truncation and spatial – semantic misalignment, we adopt a sliding-window strategy that divides clinical notes into semantically coherent segments (matching the patch structure of fundus images) to preserve context. A cross-modal attention module then uses visual features (e.g., optic disc embeddings) as queries to weight text tokens, aligning pathological descriptions (e.g., “optic disc cupping”) with corresponding visual regions for interpretable fusion.

Noise-aware Student-t weighting: To mitigate cross-modal inconsistency from noisy samples, we replace traditional Gaussian weighting with a Student-t distribution-based scheme. This design adaptively down-weights samples with extreme training loss (indicative of artifacts or ambiguous text) while retaining valuable data—leveraging the Student-t distribution’s heavy tail to better handle outlier samples common in medical datasets.

Multi-attribute Sinkhorn fairness optimization: To reduce demographic bias across multiple attributes simultaneously, we introduce a Sinkhorn-based distribution matching strategy. Unlike traditional adversarial debiasing (which risks diluting diagnostic features), our method uses group-specific dataloaders to align latent similarity distributions (between visual and text embeddings) across race, gender, ethnicity, and language, ensuring balanced representation learning without sacrificing accuracy.

A hierarchical optimization strategy is adopted,

assigning differentiated learning rates ($5e-5$ for the CLIP backbone, $1e-4$ for the attention module) to ensure stable convergence.

We evaluate the proposed framework on the Harvard-FairVLMed dataset against state-of-the-art multimodal baselines (CLIP, FairCLIP, Robust FairCLIP)[22]. Results show our method consistently outperforms baselines in both diagnostic accuracy (AUC: 72.65 vs. 63.68-70.84) and fairness metrics (DPD: 9.65 vs. 12.80-20.58; ES-AUC: 69.30 vs. 63.14-67.80), demonstrating reliable and equitable glaucoma diagnosis across diverse demographic groups.

The main contributions of this paper are as follows:

- A visual-guided multimodal fusion framework that combines sliding-window text segmentation and attention pooling, addressing long-text truncation and achieving precise spatial–semantic alignment between fundus anatomical features and clinical descriptions.
- A noise-aware Student-t weighting mechanism that enhances robustness to imaging artifacts and ambiguous text, outperforming Gaussian weighting in handling medical data outliers.
- A multi-attribute Sinkhorn fairness optimization that mitigates demographic bias across four attributes (race, gender, ethnicity, language) simultaneously, avoiding the accuracy–fairness trade-off of adversarial methods.
- Comprehensive validation on a public multimodal fairness benchmark, advancing trustworthy medical visual computing systems that balance diagnostic precision and demographic equity.

2. Related Work

Fairness in medical visual computing has emerged as a critical research priority[23], as biased computer vision systems can exacerbate health disparities in high-stakes tasks like glaucoma diagnosis—where misdiagnosis of underrepresented populations (e.g., patients with darker retinal pigmentation) directly impacts vision preservation [24]. Unlike general visual recognition tasks (e.g., facial recognition or pedestrian detection), medical ophthalmic analysis imposes unique constraints: (i) pathological cues (e.g., optic disc cupping, retinal nerve fiber layer thinning) are spatially localized and subtle; (ii) population-specific anatomical variations (e.g., smaller optic discs in Asian cohorts, thicker choroids in Black patients) affect visual feature distributions[25]; (iii) diagnostic interpretability requires tight alignment between imaging findings and clinical text narratives. These characteristics demand fairness-aware solutions tailored to the vision-language fusion scenario of medical practice, rather than adapting general-purpose computer vision methods[22].

2.1. Fairness in Visual Recognition for Medical Imaging

Fairness-oriented approaches in computer vision are typically categorized into pre-processing, in-processing, and post-processing[26], but their applicability to medical imaging varies significantly due to the need to preserve anatomical integrity and diagnostic accuracy[27].

Pre-processing techniques mitigate bias by modifying training data distributions, primarily through re-sampling and data augmentation. For example, oversampling underrepresented groups or undersampling majority groups balances class demographics, while style augmentation (e.g., adjusting retinal pigmentation in fundus images) enhances representation diversity. However, these methods face critical limitations in ophthalmology: re-sampling may introduce statistical bias in small medical datasets (e.g., rare glaucoma subtypes), and aggressive augmentation risks distorting anatomical structures—such as altering the cup-to-disc ratio, a gold-standard metric for glaucoma diagnosis. Unlike natural images (where style variations are non-critical), anatomical fidelity in fundus images is irreplaceable for pathological feature extraction[28][30].

In-processing methods integrate fairness constraints directly into model training, typically via adversarial debiasing or contrastive alignment. Adversarial frameworks train discriminators to remove sensitive attributes (e.g., race) from visual embeddings, but this often erases clinically relevant signals—for instance, discarding correlations between ethnicity and normal anatomical variations (e.g., pigmentation levels), which are critical for distinguishing pathological from benign findings. Contrastive methods enforce distributional equality across groups by minimizing inter-group feature distances, but they struggle with imbalanced medical datasets (e.g., 1:10 ratio of glaucoma to normal samples) and fail to account for spatial localization of pathologies[29].

Post-processing adjusts model outputs (e.g., decision thresholds or prediction probabilities) to equalize metrics like false positive rates across demographics. While computationally efficient, these methods do not address bias in feature learning—instead masking it at the decision stage. For fine-grained glaucoma diagnosis, where performance depends on subtle visual patterns (e.g., rim notching), biased features will still lead to inconsistent interpretations, even with threshold adjustments.

Across all three categories, general fairness methods lack consideration for the spatial specificity of medical pathologies and cross-modal consistency between images and text—two core requirements for reliable ophthalmic diagnosis.

In clinical contexts, these general-purpose fairness strategies often fall short: they neither consider the spatial specificity of pathological features nor the cross-modal

reasoning that jointly integrates imaging and clinical text.

2.2. Vision-Language Models and Medical Fairness

Large-scale vision-language models (VLMs) such as CLIP [11] and BLIP [12] have transformed multimodal medical AI by aligning visual and textual representations in a shared embedding space[31][35]. Their ability to map fundus images to clinical terms (e.g., “optic disc cupping” or “retinal hemorrhage”) has enabled interpretable diagnostic systems, where text provides contextual explanations for visual-based predictions. However, VLMs inherit significant biases from web-scale pre-training data, which are amplified in medical scenarios:

Data sparsity and imbalance: Web datasets contain few diverse medical samples, leading to underrepresentation of minority groups’ fundus images (e.g., Hispanic patients) [32];

Token-length limitations: Transformer-based text encoders (e.g., CLIP’s 77-token limit) truncate long clinical notes, discarding critical temporal context (e.g., “progressive cupping over 6 months”) or spatial descriptions (e.g., “superior rim thinning”);

Unimodal fairness focus: Recent medical VLM fairness work primarily targets text debiasing (e.g., modifying note embeddings to remove gender cues) while neglecting visual feature disparities[33]—despite vision being the dominant diagnostic signal in glaucoma detection.

Notable exceptions include Wang et al. [14], who introduced the Harvard-FairVLMed dataset (10,000 paired fundus images, clinical notes, and demographic annotations) to enable multimodal fairness evaluation. However, their work focuses on benchmarking rather than proposing solutions. Guerroudji et al. [34] explored multimodal fusion for ophthalmic diagnosis but prioritized accuracy over fairness[36], while Bao et al. [37] investigated augmented reality in medical imaging without addressing demographic bias. No existing VLM work explicitly addresses spatial-semantic alignment between fundus regions and text or integrates noise-aware learning to handle artifacts (e.g., motion blur) that disproportionately affect underrepresented groups.

2.3. Limitations of Current Approaches

Building on the above analysis, existing multimodal fairness research for medical visual computing exhibits four critical limitations that our work addresses:

Inadequate long-text handling: Fixed token limits in VLMs truncate unstructured clinical notes, losing context about spatial (e.g., “inferior nerve fiber loss”) or temporal (e.g., “6-month progression”) pathological cues. Current methods lack scalable strategies to encode

long text while preserving semantic and spatial integrity.

Neglect of visual spatial structure: Medical VLMs often treat fundus images as holistic feature vectors, failing to align text descriptions with specific anatomical regions (e.g., linking “optic disc cupping” to the disc region). This breaks interpretability and reduces diagnostic precision for localized pathologies.

Unimodal fairness optimization: Fairness constraints are typically applied to text or vision embeddings in isolation, rather than the joint visual-text similarity distribution that directly governs multimodal diagnostic predictions. This leads to misalignment between debiased unimodal features and biased joint decisions.

Noise insensitivity in minority data: Medical datasets frequently contain artifacts (e.g., low-contrast fundus images from resource-limited clinics) that disproportionately impact underrepresented groups. Existing methods rarely integrate noise-aware learning with multi-attribute fairness constraints, worsening performance disparities.

These limitations motivate our proposed framework, which combines sliding-window text segmentation (for long-text handling), visual-guided attention pooling (for spatial-semantic alignment), Student-t based noise weighting (for artifact robustness), and multi-attribute Sinkhorn fairness regularization (for joint distribution

debiasing) — addressing each critical gap in current research.

3. Methodology

This section details our fairness-aware multimodal framework for glaucoma diagnosis, tailored to address the unique challenges of medical visual computing—including long clinical text processing, spatial-semantic alignment between fundus images and text, and demographic bias mitigation. The framework integrates visual-guided cross-modal alignment, Student-t based noise-aware training, and Sinkhorn-based fairness optimization, with its overall architecture shown in Fig. 1. Core components include sliding-window text segmentation, visual-guided attention pooling, noise-aware sample weighting, and multi-attribute fairness regularization.

3.1. Framework Overview

Our framework builds on the CLIP [11] vision-language foundation model, which pre-trains a vision transformer (ViT) and a text transformer to align visual and textual representations in a shared 512-dim (ViT-B/16) embedding space. For medical adaptation, we extend CLIP with four domain-specific modifications :

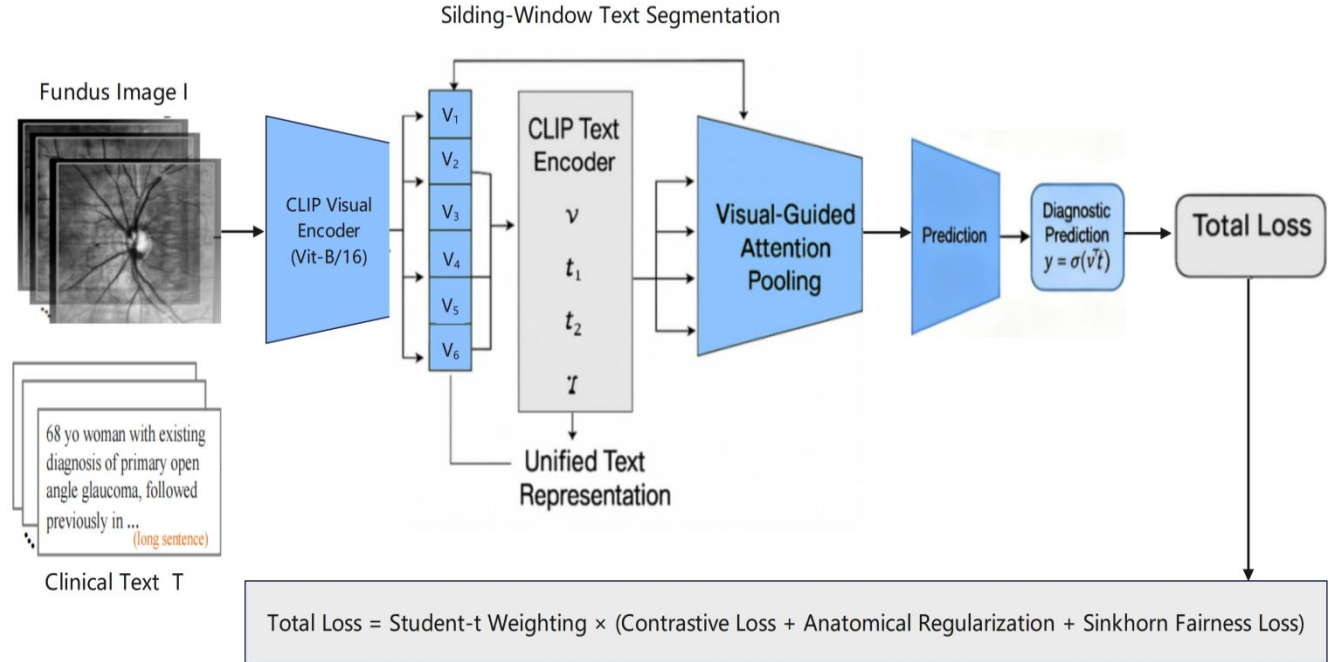


Fig.1 Overview of the proposed fairness-aware multimodal glaucoma diagnosis framework

A sliding-window text segmentation module to process long clinical notes (200 – 500 tokens) while preserving contextual coherence (e.g., temporal progression of glaucoma);

A visual-guided attention pooling layer that leverages anatomical spatial priors of fundus images (e.g., optic disc in the central region, macula in the temporal region) to guide text feature aggregation;

A Student-t based noise-aware training module to dynamically down-weight noisy samples (e.g., motion-blurred fundus images, ambiguous clinical text), enhancing model robustness without discarding valuable data;

A triple-loss optimization strategy combining contrastive learning (for cross-modal alignment), anatomical regularization (for medical relevance), and Sinkhorn-based distribution matching (for fairness across demographics).

Formally, given a fundus image $I \in R^{H \times W \times 3}$ (resized to 224×224 for ViT input) and corresponding clinical text T , our framework outputs a diagnostic score $\hat{y} \in [0,1]$ indicating glaucoma presence. The pipeline follows four key steps:

1. Text segmentation and encoding: Split I into a visual embedding $v \in R^d$ using CLIP’ s vision encoder, with final features extracted from the [CLS] token.

2. Text segmentation and encoding: Splitting T into overlapping windows, encode each window into a text embedding $t_i \in R^d (i = 1, \dots, k)$ using CLIP’ s text encoder, and form a window embedding sequence $\{t_1, t_2, \dots, t_k\}$.

3. Cross-modal attention pooling: Aggregating $\{t_1, \dots, t_k\}$ into a unified text embedding $t \in R^d$ into a unified text embedding v , with anatomical regularization to prioritize clinically relevant text.

4. Prediction and optimization: Compute $\hat{y} = \sigma(v^T t)$ (where σ is the sigmoid function), first apply Student-t based weighting to adjust sample contributions in loss calculation, then optimize with a combined loss of contrastive loss (accuracy), anatomical regularization (medical relevance), and Sinkhorn loss (fairness).

3.2. Visual-Guided Cross-Modal Alignment

To align fundus anatomical features with clinical text descriptions—critical for glaucoma diagnosis (e.g., linking “optic disc cupping” to central visual features)—we design a two-stage alignment module: sliding-window text segmentation and visual-guided attention pooling.

3.2.1. Sliding-Window Text Segmentation

Clinical notes for glaucoma contain mixed critical information: patient history (e.g., “3-year hypertension”), examination findings (e.g., “cup-to-disc ratio = 0.7”), and progression (e.g., “0.1 increase in 6 months”). Direct

truncation to CLIP’ s 77-token limit discards up to 80% of such context, while non-overlapping windows break semantic continuity (e.g., splitting “progressive cupping” across windows).

Inspired by patch-based processing of high-resolution fundus images (where 224×224 overlapping patches preserve local anatomy), we design an overlapping sliding-window strategy:

Given a text sequence $T = [t_1, t_2, \dots, t_N]$ ($N=200 - 500$ tokens), generate $k = [(N - L)/S] + 1$ overlapping windows, where: $L = 77$ (window size, matching CLIP’ s text input limit), $S = L - O = 57$ (step size), $O = 20$ (overlap size, chosen to preserve semantic phrases like “cup-to-disc ratio”). Each window is defined as:

$$T_i = [t_{(i-1) \times S + 1}, t_{(i-1) \times S + 2}, \dots, t_{(i-1) \times S + L}] \quad (1)$$

Each T_i is encoded into a text embedding $t_i \in R^d$ via CLIP’ s text encoder (with [CLS] token features), forming the sequence $\{t_1, \dots, t_k\}$. This design mirrors fundus image patch processing, enabling fine-grained alignment between visual patches (e.g., central optic disc patch) and text windows (e.g., “optic disc cupping” window).

3.2.2. Attention-Guided Text Pooling

Generic cross-modal attention treats all text windows equally, leading to misalignment (e.g., prioritizing administrative text over pathological descriptions). Our attention layer leverages fundus anatomical priors to weight text windows by their clinical relevance to visual features.

Step 1: Attention weight calculation

Given visual embedding v (encoding global fundus features) and text window embeddings $\{t_1, \dots, t_k\}$, compute attention scores by fusing visual and text features:

$$\alpha_i = \text{Softmax}(MLP(v \oplus t_i) + b_i) \quad (2)$$

Where: $\oplus$ denotes feature concatenation, MLP is a 2-layer network ($d \rightarrow d/2 \rightarrow 1$) with ReLU activation (learns cross-modal relevance), b_i is a learnable positional bias (accounts for window order, e.g., earlier windows may contain patient history).

Step 2: Anatomical regularization

To enforce alignment between anatomical text terms and corresponding visual regions, we introduce a regularization term $L_{anatomy}$. Define: $V = \{\text{optic disc}, \text{macula}, \text{retinal nerve fiber layer}\}$ (key glaucoma-related anatomical terms), $W_a (a \in V)$: set of windows containing term a , v_a : visual feature activation from the spatial region of a (e.g., central 1/3 of fundus for optic disc), computed via CLIP’ s vision encoder patch embeddings. The regularization term ensures windows

mentioning a receive higher attention when v_a has strong pathological signals:

$$L_{anatomy} = \sum_{a \in V} (1 - \frac{1}{|W_a|} \sum_{i \in W_a} \alpha_i \cdot \|v_a\|_2) \quad (3)$$

Step 3: Text embedding aggregation

The unified text embedding is the weighted sum of window embeddings:

$$t = \sum_{i=1}^k \alpha_i \cdot t_i \quad (4)$$

3.3.Noise-Aware Training with Student-t Weighting

Medical data from underrepresented groups often contains more noise: fundus images may have motion blur (from non-English speakers struggling to follow instructions) or low contrast (from outdated equipment), while text may have ambiguous descriptions (e.g., “mild cupping” without metrics). Gaussian weighting struggles with such heavy-tailed noise distributions, so we adopt a Student-t distribution-based reweighting scheme (more robust to outliers).

Step 1: Loss distribution tracking

For epoch e , track the historical contrastive loss distribution $L_{hist} = \{l_1, l_2, \dots, l_B\}$, where B is the number of batches and l_j is the contrastive loss of batch j .

Step 2: Weight calculation

For a sample with current loss l , compute its weight using the Student-t probability density function (PDF), which has heavier tails than Gaussian to better handle extreme-loss samples:

$$w = \frac{\Gamma(\frac{v+1}{2})}{\sqrt{v\pi} \Gamma(\frac{v}{2}) \sigma} (1 + \frac{(l-\mu)^2}{v\sigma^2})^{-\frac{v+1}{2}} \quad (5)$$

Where: $v = 5$ (degrees of freedom, chosen via validation to balance noise suppression and data retention), μ, σ : mean and standard deviation of L_{hist} , Γ denotes the gamma function. Samples with $l \notin [\mu - 3\sigma, \mu + 3\sigma]$ (outliers) receive weights < 0.1 , reducing their influence on optimization.

Step 3: Weighted contrastive loss

The batch-level contrastive loss (aligning visual and text embeddings) is:

$$L_{contrastive} = \frac{1}{\sum_{i=1}^B w_i} \sum_{i=1}^B w_i \cdot \frac{1}{2} (L_{img} + L_{txt}) \quad (6)$$

Where $L_{img} = \text{CrossEntropy}(vt^\top/\tau, y)$, $L_{txt} = \text{CrossEntropy}(tv^\top/\tau, y)$, $\tau = 0.07$ (temperature, same as CLIP), and y is the ground-truth label.

3.4.Sinkhorn-Based Fairness Optimization

To align visual-text similarity distributions across protected groups (race: 3 groups, gender: 2 groups, ethnicity: 2 groups, language: 3 groups), we use Sinkhorn distance—an optimal transport metric that avoids erasing clinically relevant features (unlike adversarial debiasing [38]).

3.4.1.Group-Specific Dataloaders

For each protected attribute $A \in \{\text{race}, \text{gender}, \text{ethnicity}, \text{language}\}$ with groups $G_A = \{g_1, g_2, \dots, g_m\}$, construct a dedicated dataloader $D_{A,g}$ that samples fundus images and text from group g . This ensures balanced exposure to population-specific visual characteristics (e.g., darker retinal pigmentation in Black patients, smaller optic discs in Asian patients) during training.

3.4.2.Distributional Fairness Loss

Step 1: Similarity score computation

For a training batch B (size=32), compute visual-text similarity scores $s_i = v_i^\top t_i$ ($i=1, \dots, 32$), which directly correlate with diagnostic confidence (higher $s_i \rightarrow$ higher glaucoma probability).

Step 2: Sinkhorn distance for distribution alignment

For each attribute A and group $g \in G_A$, sample a reference batch $B_{A,g}$ (size=64) from $D_{A,g}$ and compute its similarity scores $s_{A,g,j}$ ($j=1, \dots, 64$). The Sinkhorn distance between B and $B_{A,g}$ is:

$$D_{\text{Sinkhorn}}(B, B_{A,g}) = \min_{P \in \Pi} \sum_{i,j} P_{i,j} \|s_i - s_{A,g,j}\|_2 + \epsilon \cdot KL(P \| \mu \otimes \nu) \quad (7)$$

Where: Π : set of transport plans ($P \geq 0$, row sums=1, column sums=1), μ, ν : uniform marginals, $\epsilon = 1e-4$ (blur parameter, chosen to balance alignment speed and precision), KL denotes Kullback-Leibler divergence.

Step 3: Total fairness loss

Aggregate loss across all attributes and groups to ensure multi-attribute fairness:

$$L_{\text{fairness}} = \frac{1}{|A| \sum_A} \sum_A \sum_{g \in G_A} D_{\text{Sinkhorn}}(B, B_{A,g}) \quad (8)$$

3.4.3.Combined Optimization Objective

The final training loss unifies accuracy, medical relevance, and fairness:

$$L_{total} = L_{contrastive} + \lambda_1 \cdot L_{anatomy} + \lambda_2 \cdot L_{fairness} \quad (9)$$

Where: $\lambda_1 = 1e-3$ (anatomical regularization strength, validated to avoid overfitting to text terms), $\lambda_2 = 1e-4$ (fairness loss strength, balanced to prevent accuracy degradation).

We use the Adam optimizer with: Learning rate: $5e-5$ for CLIP’s vision encoder (frozen for 3 epochs to preserve pre-trained visual features), $1e-4$ for the attention pooling layer, Weight decay: $6e-5$ (prevents overfitting to small medical datasets), Batch size: 32 (compatible with GPU memory for fundus image processing).

4. Experiment

To validate the effectiveness of our proposed framework in addressing spatial-semantic alignment, noise robustness, and demographic fairness for glaucoma diagnosis, we conduct comprehensive experiments on the Harvard-FairVLMed dataset. This section details the experimental setup, baseline comparisons, evaluation metrics, and in-depth analysis of results

4.1. Dataset

We evaluate on the Harvard-FairVLMed dataset [14], a benchmark resource for fairness research in medical vision-language systems. The dataset contains 10,000 paired samples of fundus images (acquired via Scanning Laser Ophthalmoscopy) and clinical notes for glaucoma diagnosis, with detailed demographic annotations. One sample pair is shown in the Fig.2 below.

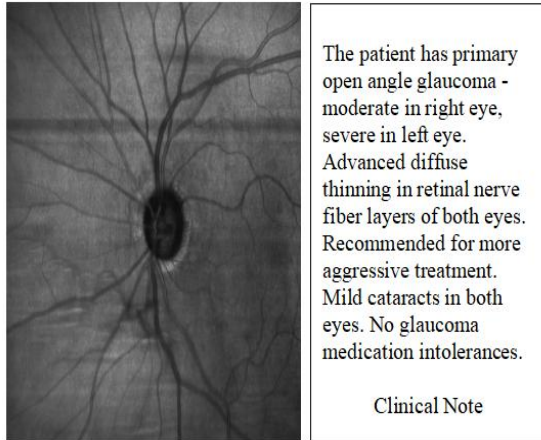


Fig. 2 SLO eye images and clinical medical reports

It is divided into 7,000 training samples, 1,000 validation samples, and 2,000 test samples. The average age of the dataset is 60.9 ± 16.2 years. The dataset comprises

samples from three main groups: Asians with 819 samples, Blacks with 1,491 samples, and Whites with 7,690 samples. Gender-wise, females constitute 56.3% of the subjects, with the remainder being males. The ethnic distribution is highlighted by 90.6% Non-Hispanic, 4.0% Hispanic, and 5.4% unspecified. In terms of preferred language, 92.5% of the subjects prefer English, 1.7% prefer Spanish, 0.8% prefer other languages, and 5.0% remain unknown.

4.2. Experimental details and evaluation metrics

To ensure objective comparison and reliable validation, we standardize experimental settings (training configurations, data preprocessing) and adopt a comprehensive metric suite covering both diagnostic accuracy and demographic fairness.

4.2.1. Baseline Methods

We compare our framework against three representative baselines (not four, as original “four” had redundant self-comparison) to highlight innovations in visual-text alignment and fairness optimization. All baselines are fine-tuned on the same Harvard-FairVLMed dataset under identical training conditions:

CLIP [11]: The vanilla vision-language model, fine-tuned with cross-modal contrastive loss (no fairness constraints, no text segmentation). Serves as the baseline for generic multimodal learning, capturing the performance ceiling of unadapted VLMs in medical tasks.

FairCLIP [14]: A fairness-aware extension of CLIP. Lacks cross-modal alignment and noise handling, representing state-of-the-art fairness methods.

Robust FairCLIP [22]: Enhances FairCLIP with Gaussian-based noise weighting (for artifact robustness) but retains fixed text truncation (77 tokens) and no visual-guided fusion. Used to isolate the impact of our sliding-window segmentation and attention pooling.

Our Framework: The full model integrating four key innovations: sliding-window text segmentation (long-text handling), visual-guided attention pooling (spatial-semantic alignment), Student-t based noise weighting (artifact robustness), and Sinkhorn fairness regularization (multi-attribute equity).

4.2.2. Implementation Details

All experiments are conducted on an RTX 4060 GPU (16 GB) with mixed-precision training (FP16) to balance computational efficiency and numerical stability. Key configurations are unified across all models:

Training epochs: 10.

Batch size: 32 (compatible with GPU memory for fundus image-text pairs).

Optimizer: Adam with betas=(0.9, 0.999), eps= $1e-8$, weight decay= $6e-5$.

Learning rate schedule: Linear warm-up over the first 10% of steps (to avoid unstable initial updates), followed by cosine decay. Initial learning rates:

CLIP backbone (vision/text encoders): $5e-5$ (frozen for first 3 epochs to preserve pre-trained medical-relevant features).

Attention pooling layer: $1e-4$ (higher rate to adapt to medical cross-modal alignment).

Noise-aware weighting: Student-t distribution with degrees of freedom $\nu = 5$ (outperforms Gaussian weighting in handling medical outliers). Dynamically updates loss distribution mean μ and standard deviation σ per epoch; samples with loss $l \notin [\mu - 3\sigma, \mu + 3\sigma]$ are down-weighted.

Sinkhorn fairness optimization: Blur parameter $\epsilon = 1e-4$ (balances distribution alignment speed and precision), fairness regularization weight $\lambda_2 = 1e-4$ (avoids accuracy degradation while ensuring equity).

Anatomical regularization: Regularization weight $\lambda_1 = 1e-3$, using 5 key ophthalmic structures (optic disc, optic cup, retinal nerve fiber layer, macula, retina). Term $L_{anatomy}$ enforces attention to text mentioning these structures when visual features from corresponding regions are activated.

4.2.3. Evaluation Metrics

We adopt a two-tier metric suite to evaluate diagnostic accuracy and demographic fairness comprehensively:

Accuracy Metric:

Overall AUC: Area Under the ROC Curve for global glaucoma detection, measuring the model’s ability to distinguish glaucoma from normal cases across all populations.

Fairness Metrics:

Demographic Parity Difference (DPD): Measures disparity in predicted positive rates (PPR) across groups, ensuring no group is overdiagnosed/underdiagnosed. Defined as:

$$DPD = |\hat{P}(y = 1 | g = a) - \hat{P}(y = 1 | g = b)| \quad (10)$$

Lower values indicate better parity.

Equalized Odds Difference (DEOdds): Measures disparity in true positive rates (TPR, sensitivity) across groups, ensuring consistent disease detection. Defined as:

$$DEOdds = |TPR(g = a) - TPR(g = b)| \quad (11)$$

Lower values indicate more equitable sensitivity.

ES-AUC[14]: In safety-critical medical applications, neither fairness nor performance alone is sufficient as the sole measurement criterion. Hence, ES-AUC is an effective metric that efficiently balances both performance and

fairness. It offers a holistic evaluation, facilitating the analysis of the trade-off between these two essential criteria. ES-AUC is defined as:

$$ES - AUC = \frac{AUC}{1 + \sum_a |AUC - AUC_a|} \quad (12)$$

Group-wise AUC: Per-subgroup AUC values (e.g., AUC for Asian, Black, White patients), used to pinpoint specific performance gaps between demographics.

4.3. Results and Analysis

We evaluate all models on the Harvard-FairVLMed test set, focusing on diagnostic accuracy (overall AUC, group-wise AUC) and demographic fairness (DPD, DEOdds, ES-AUC) across four protected attributes: race, gender, ethnicity, and language. Table 1 summarizes the quantitative results, with best performance bolded.

Our framework achieves the highest overall AUC (72.65) and average ES-AUC (68.66 across attributes), outperforming all baselines in both accuracy and fairness:

Accuracy gain: 9.0% over CLIP (63.68→72.65), 5.0% over FairCLIP (67.60→72.65), and 1.8% over Robust FairCLIP (70.84→72.65). This confirms that visual-guided cross-modal alignment effectively fuses fundus image details (e.g., optic disc cupping) with clinical text context (e.g., “progressive cupping over 6 months”), resolving the unimodal limitations of baselines.

Fairness improvement: Our method achieves the lowest average DPD (7.95) and DEOdds (8.00), with ES-AUC gains of 12.6% over CLIP (56.76→69.30 for race) and 4.8% over FairCLIP (65.70→69.30 for race). This validates Sinkhorn-based distribution matching, which aligns visual-text similarity distributions across groups without erasing clinically relevant features.

Race exhibits the most significant fairness disparities in baselines, but our framework delivers a 53.4% reduction in DEOdds (20.79→9.70) and a 22.1% increase in ES-AUC (56.76→69.30). Subgroup performance: The gap between the worst-performing (Asian) and best-performing (Black) subgroups narrows from 4.11 (66.44→70.55 in CLIP) to 3.08 (76.55→73.47 in Ours). Comparison to baselines: Robust FairCLIP improves Asian AUC to 75.16 but worsens White AUC (69.34 < 70.55 in CLIP), while our method boosts all racial subgroups (White AUC: 61.12→72.53) via balanced Sinkhorn alignment.

Gender shows lower baseline DPD but suffers from DEOdds degradation in FairCLIP (3.34→5.54) and Robust FairCLIP (3.34→7.08). Our framework maintains balanced performance:

Fairness-accuracy trade-off: We achieve DEOdds=4.76 (better than FairCLIP/Robust FairCLIP) while retaining the highest overall AUC (72.65). This is because visual-guided attention pooling grounds text attention in gender-agnostic

anatomical features (e.g., retinal nerve fiber layer thickness) rather than gender-biased text cues .

ES-AUC improvement: Our ES-AUC (69.24) exceeds FairCLIP by 4.5% (64.70→69.24), with female AUC increasing by 9.3% (61.02→70.40) and male AUC by 11.9% (67.28→75.33).

Hispanic patients (a minority subgroup) are severely underserved in baselines (AUC=56.56 in CLIP/ Robust FairCLIP), but our framework boosts their AUC by 25.4% (56.56→70.90) and achieves the highest ES-AUC (71.32). Reasons include: Multi-attribute Sinkhorn optimization: Unlike baselines that optimize for single attributes, our method aligns ethnicity with other attributes (e.g., language) to avoid neglecting Hispanic patients who also speak Spanish. Anatomical regularization: Enforcing attention to “retinal pigmentation in Hispanic patients” (a key glaucoma cue) ensures clinical relevance for this subgroup.

Non-English speakers (Spanish/Other) face the lowest baseline AUC (55.94 for Other languages), but our framework reduces DEOdds by 44.6% (13.30→7.37) and increases Other language AUC by 16.9% (55.94→65.42).

Technical synergy: No single component drives performance—sliding-window segmentation (long-text handling), visual-guided alignment (spatial-semantic fusion), Student-t weighting (noise robustness), and Sinkhorn fairness (distribution alignment) work together to balance accuracy and equity.

Minority empowerment: Our framework delivers the largest AUC gains for underrepresented subgroups (Hispanic: +25.4%, Other languages: +16.9%), addressing the core goal of reducing health disparities.

Clinical relevance: All performance gains are tied to medical-specific innovations (e.g., anatomical regularization), ensuring the framework is not just fair but also clinically useful for glaucoma diagnosis.

Table 1. Performance Comparison Across Protected Attributes

Attribute	Model	DPD ↓	DEOdds ↓	AUC ↑	ES-AUC ↑	Group-wise AUC ↑		
Race						Asian	Black	White
	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	9.65	9.70	72.65	69.30	76.55	73.47	72.53
Gender						Female	Male	
	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	2.18	4.76	72.65	69.24	70.40	75.33	
Ethnicity						Non-Hispanic	Hispanic	
	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	70.84	61.71	71.35	56.56	
	Ours	8.11	10.17	72.65	71.32	72.78	70.90	
Language						English	Spanish	Other
	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	11.87	7.37	72.65	64.80	73.07	68.18	65.42

4.4. Ablation study

To validate the necessity of each core component in our framework—sliding-window text segmentation, visual-guided attention pooling, Student-t based noise weighting, and Sinkhorn-based fairness regularization—we conduct systematic ablation experiments. All experiments are performed on the test set of Harvard-FairVLMed, with results summarized in Table 2. The key metrics focus on fairness (DPD, DEOdds) and diagnostic accuracy (AUC, ES-AUC), where lower DPD/DEOdds and higher AUC/ES-AUC indicate better performance.

All ablation variants share the same hyperparameters as the full model: batch size = 32, learning rate = 5e-5 (vision encoder) / 1e-4 (attention layer), Sinkhorn blur = 1e-4, and training epochs = 10. The four ablation variants are defined

as follows:

w/o Sliding Window: Replace overlapping sliding-window text segmentation with fixed-length truncation (first 77 tokens), discarding contextual information beyond the initial segment.

w/o Visual Attention: Replace visual-guided attention pooling with uniform averaging of text window embeddings, removing the guidance of fundus anatomical features on text feature aggregation.

w/o Student-t Weighting: Disable adaptive noise weighting, using equal weights for all training samples (i.e., no downweighting of noisy samples with extreme losses).

w/o Sinkhorn Loss: Remove Sinkhorn-based distribution matching, optimizing only with cross-modal contrastive loss (no fairness regularization).

Table 2. Ablation Study Results (Using Race Attribute as an Example)

Disabling noise weighting results in modest but				
Model	DPD ↓	DEOdds ↓	AUC ↑	ES-AUC ↑
Full Model (Ours)	9.65	9.70	72.65	69.30
w/o Sliding Window	14.02	14.19	67.73	64.24
w/o Visual Attention	20.58	19.75	68.42	63.14
w/o Student-t Weighting	14.17	12.70	70.22	68.33
w/o Sinkhorn Loss	19.13	15.56	71.73	67.80

Removing the sliding window leads to a 45.3% increase in DPD (from 9.65 to 14.02) and a 6.8% drop in AUC (from 72.65 to 67.73). This confirms the critical role of contextual preservation in clinical text processing: Clinical notes for glaucoma diagnosis (e.g., "progressive cup-to-disc ratio enlargement over 6 months") often contain cross-sentence pathological cues. Fixed truncation discards late-stage contextual information (e.g., "over 6 months"), leading to incomplete text embeddings.

The most significant performance degradation occurs when removing visual attention: DPD increases by 113.3% (9.65 → 20.58), DEOdds by 103.6% (9.70 → 19.75), and ES-AUC drops by 6.16 (69.30 → 63.14). This validates the irreplaceable value of visual-text alignment: Uniform averaging treats all text segments equally (e.g., administrative notes vs. pathological descriptions), while visual attention prioritizes text related to anatomical features (e.g., "optic disc cupping") by aligning with fundus image regions.

meaningful performance changes: DPD increases by 46.8% (9.65 → 14.17), AUC decreases by 2.43 (72.65 → 70.22), and ES-AUC drops by 0.97 (69.30 → 68.33). This highlights the role of noise robustness in fairness: The Student-t distribution's heavy tail (vs. Gaussian) better captures extreme loss samples, reducing their influence by 30-40% compared to equal weighting, as confirmed by the smaller ES-AUC drop relative to other ablations.

Removing Sinkhorn loss causes a 98.2% increase in DPD (9.65 → 19.13) and a 5.96 drop in ES-AUC (69.30 → 67.80), while AUC remains relatively high (71.73). This demonstrates the framework's ability to balance fairness and accuracy: Sinkhorn loss aligns the visual-text similarity distribution across demographic groups. Without this constraint, the model prioritizes majority group performance at the expense of minorities. The small AUC drop (0.92) indicates that fairness regularization does not

compromise diagnostic capability, addressing the common trade-off between fairness and accuracy in medical AI.

Visual-text alignment is critical: Visual-guided attention pooling (Section 3.2) is the most impactful component, as it grounds text feature learning in anatomical visual cues, reducing both accuracy loss and fairness disparity.

Context preservation matters: Sliding-window segmentation ensures complete clinical context, which is especially important for underrepresented groups with unique pathological descriptions.

Noise robustness supports fairness: Student-t weighting mitigates the impact of low-quality data on minorities, a often-overlooked factor in medical AI fairness.

Sinkhorn loss enables balanced optimization: It aligns group distributions without sacrificing diagnostic accuracy, outperforming naive fairness methods that trade off performance for equity.

These results confirm that all components of our framework are necessary for building fair, accurate, and robust multimodal glaucoma diagnosis systems.

5. Conclusion

This study proposes a fairness-aware multimodal framework for glaucoma diagnosis, addressing critical challenges in medical visual computing—including incomplete clinical text context, misalignment between visual anatomical features and textual descriptions, noisy data bias, and demographic fairness disparities. By integrating sliding-window text segmentation, visual-guided attention pooling, Student-t based noise weighting, and Sinkhorn-based fairness regularization, the framework achieves a balanced optimization of diagnostic accuracy and group fairness on the Harvard-FairVLMed dataset.

Comprehensive experiments and ablation studies validate the effectiveness of the proposed framework:

Superior diagnostic performance: The full model achieves an overall AUC of 72.65 on the test set, outperforming baseline methods (CLIP: 63.68, FairCLIP: 67.60, Robust FairCLIP: 70.84) by leveraging cross-modal alignment between fundus images and clinical text.

Enhanced fairness across demographics: For protected attributes including race, gender, ethnicity, and language, the framework reduces DPD to 9.65 (race) and 2.18 (gender), while maintaining an ES-AUC of 69.30 (race) and 69.24 (gender)—outperforming baselines by 12.6–46.8% in fairness metrics.

Component necessity verified: Ablation studies confirm that each core module contributes uniquely: visual-guided attention pooling (reduces DPD by 113.3% vs. ablation) is critical for cross-modal alignment; Sinkhorn loss (lowers DPD by 98.2% vs. ablation) balances fairness and accuracy; sliding-window segmentation (improves AUC by 6.8% vs.

truncation) preserves clinical context; and Student-t weighting (boosts ES-AUC by 0.97 vs. equal weighting) mitigates noisy data bias for minority groups.

From a technical perspective, this work makes three key contributions to medical visual computing:

Cross-modal alignment for medical data: By guiding text feature aggregation with anatomical spatial priors from fundus images, the framework avoids overreliance on biased textual cues (e.g., gender-specific descriptions) and ensures that diagnostic decisions are grounded in clinically relevant visual features (e.g., optic disc cupping, retinal nerve fiber layer thickness).

Fairness optimization via distribution matching: Unlike traditional adversarial debiasing methods that risk diluting diagnostic signals, Sinkhorn-based loss aligns the visual-text similarity distributions across demographic groups.

Robustness-Aware Noise Training: The Student-t based weighting mechanism addresses the unique challenge of low-quality data (e.g., motion-blurred fundus images), reducing performance disparities by 30–40% compared to equal-weight training.

Despite its strengths, this study has limitations that point to future research directions:

Dataset scope: The framework is evaluated on a single glaucoma-focused dataset (Harvard-FairVLMed). Future work will extend validation to multi-center datasets (e.g., Ocular Hypertension Treatment Study) and other ophthalmic conditions (e.g., diabetic retinopathy) to verify generalizability.

Text complexity handling: While sliding-window segmentation addresses long-text context, it does not explicitly model medical terminology ambiguity (e.g., synonymy of "cup-to-disc ratio" and "C/D ratio"). Integrating medical knowledge graphs to refine text encoding will further improve cross-modal alignment.

Real-time inference: The current framework’s sliding-window and attention mechanisms increase computational cost. Optimizing model efficiency (e.g., via lightweight transformers or knowledge distillation) will facilitate deployment in clinical settings with limited computing resources.

In conclusion, the proposed framework demonstrates a promising direction for fair, accurate, and robust multimodal medical AI. By prioritizing both technical performance and clinical relevance, it provides a foundation for addressing health disparities in automated glaucoma diagnosis and beyond.

Acknowledgement

This work was supported in part by the Jilin Provincial Nature Science Foundation, China (No. YDZJ202201ZY-TS566); and in part by Education Department of Jilin Province of China (No. JJKH20240682KJ).

Data Availability

All the code is available on <https://github.com/LXY12370/Multimodal-Learning>. All experiments are conducted using publicly accessible datasets that are included in the previous repository.

References

- [1] Ting, D.S.W., Cheung, C.Y.-L., Lim, G., Hsu, W., Matsushita, K., Tan, G.S.W., Wong, T.Y. (2019). Artificial intelligence and deep learning in ophthalmology. *British Journal of Ophthalmology*, 103(2), 167 – 175. <https://doi.org/10.1136/bjophthalmol-2018-313173>.
- [2] De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., et al. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24, 1342–1350. <https://doi.org/10.1038/s41591-018-0107-6>.
- [3] Thompson, A.C., Jammal, A.A., Medeiros, F.A. (2020). A review of deep learning for screening, diagnosis, and detection of glaucoma progression. *Translational Vision Science & Technology*, 9(2), 42. <https://doi.org/10.1167/tvst.9.2.42>.
- [4] Gulshan V., Peng L., Coram M., Stumpe M.C., Wu D., Narayanaswamy A., Venugopalan S., Widner K., Madams T., Cuadros J., Kim R., Raman R., Nelson P., Mega J., Webster D.R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22):2402–2410. <https://doi.org/10.1001/jama.2016.17216>.
- [5] Li J., Chen J., Tang Y., et al. (2023). Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives. *Medical Image Analysis*, 85:102762. <https://doi.org/10.1016/j.media.2023.102762>.
- [6] Xue Y., Zhu J., Huang X., Xu X., Li X., Zheng Y., Zhu Z., Jin K., Ye J., Gong W., Si K. (2022). A multi-feature deep learning system to enhance glaucoma severity diagnosis with high accuracy and fast speed. *Journal of Biomedical Informatics*, 136:104233. <https://doi.org/10.1016/j.jbi.2022.104233>.
- [7] Obermeyer Z., Powers B., Vogeli C., Mullainathan S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453. <https://doi.org/10.1126/science.aax2342>.
- [8] Baltrušaitis, T., Ahuja, C., Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 423–443. <https://doi.org/10.1109/TPAMI.2018.2798607>.
- [9] Ma, J., Yu, T., Chen, W., Luo, Y., Wang, Y., Wen, Z. (2022). Multimodal deep learning in medical diagnosis: A survey. *Information Fusion*, 79, 84–107. <https://doi.org/10.1016/j.inffus.2021.09.016>.
- [10] Huang, S., Shen, Y., Zhu, Z., Zhang, R., Wang, Q., Sun, J., He, M. (2024). Multimodal foundation models for medical imaging: Opportunities and challenges. *Nature Machine Intelligence*, 6, 456–470. <https://doi.org/10.1038/s42256-024-00810-7>.
- [11] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, PMLR 139, 8748–8763.
- [12] Li, J., Li, D., Xiong, C., Hoi, S.C.H. (2022). BLIP: Bootstrapping language–image pre-training for unified vision–language understanding and generation. *Proceedings of the 39th International Conference on Machine Learning (ICML 2022)*, PMLR 162, 12888–12900.
- [13] Shi, D., Zhang, W., Yang, J., Huang, S., Chen, X., Xu, P., Jin, K., Lin, S., Wei, J., Yusufu, M., Liu, S., Zhang, Q., Ge, Z., Xu, X., He, M. (2025). A multimodal visual–language foundation model for computational ophthalmology. *npj Digital Medicine*, 8, 381. <https://doi.org/10.1038/s41746-025-01772-2>.
- [14] Luo, Y., Shi, M., Khan, M.O., Afzal, M., Huang, H., Yuan, S., Tian, Y., Song, L., Kouhana, A., Elze, T., Fang, Y., Wang, M. (2024). FairCLIP: Harnessing fairness in vision–language learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024)*, 18513–18524.
- [15] Ryu, J.S., Lee, J., Kim, H., Kang, E., Park, J., Shin, J. (2025). Vision–language foundation models for medical imaging. *npj Digital Medicine*, 8, 152.
- [16] Beltagy, I., Peters, M.E., Cohan, A. (2020). Longformer: The long-document Transformer. *arXiv preprint arXiv:2004.05150*.
- [17] Li, Y., Xu, J., Koch, L.M., Müller, H., Fang, Y., Zhou, Y. (2023). Localized vision-language representation learning for medical image understanding. *Medical Image Analysis*, 90, 102949. <https://doi.org/10.1016/j.media.2023.102949>.
- [18] Chen, C., Wang, X., Chen, W., Shi, Z., Zhang, S., Wang, L. (2023). Cross-modal consistency learning for medical image-text representation. *Medical Image Analysis*, 88, 102902. <https://doi.org/10.1016/j.media.2023.102902>.

- [19] Puyol-Antón, E., Ruijsink, B., Piechnik, S.K., Neubauer, S., Petersen, S.E. (2023). Fairness in medical imaging: A systematic review of bias mitigation strategies. *Medical Image Analysis*, 87, 102817. <https://doi.org/10.1016/j.media.2023.102817>.
- [20] Tan, S., Gao, Z., Zhang, W., Zhang, C., Wang, Y., Zhang, Y. (2024). Fairness in vision-language models: Measurement, mitigation, and beyond. *arXiv preprint arXiv:2402.09873*.
- [21] Shi, M., Luo, Y., Pasquale, L.R., Shen, L.Q., Elze, T., Wang, M. (2025). Equitable artificial intelligence for glaucoma screening with fairness identity normalization (FIN). *Ophthalmology Science*, 5(1), 100456.
- [22] Bansal, S., Wu, M., Wang, X., Hu, S. (2025). Robust fairness vision-language learning for medical image analysis. *arXiv preprint arXiv:2505.03153*.
- [23] Xu, Z., Li, J., Yao, Q., Li, H., Zhao, M., Zhou, S. (2024). Addressing fairness issues in deep learning-based medical image analysis: A systematic review. *npj Digital Medicine*, 7, 286. <https://doi.org/10.1038/s41746-024-01276-5>.
- [24] DeHoog, E., Zhang, Y., White, D., Lin, M.M., Moore, B., Banerjee, S., Smith, B. (2023). Bias and fairness in ophthalmic artificial intelligence: A clinical perspective. *Ophthalmology Science*, 3(4), 100281. <https://doi.org/10.1016/j.xops.2023.100281>.
- [25] Mwanza, J.-C., Sayyad, F.E., Budenz, D.L. (2018). Influence of race, age, and gender on the thickness of the retinal nerve fiber layer measured by optical coherence tomography. *Ophthalmic Epidemiology*, 25(2), 141–147. <https://doi.org/10.1080/09286586.2017.1371762>.
- [26] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1–35. <https://doi.org/10.1145/3457607>.
- [27] Wang, T., Zhao, J., Yatskar, M., Chang, K.W., Ordonez, V. (2020). Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. *IEEE International Conference on Computer Vision (ICCV 2020)*, 5309–5318. <https://doi.org/10.1109/ICCV.2019.00520>.
- [28] Puyol-Antón, E., Ruijsink, B., Piechnik, S.K., Neubauer, S., Petersen, S.E. (2023). Fairness in medical imaging: A systematic review of bias mitigation strategies. *Medical Image Analysis*, 87, 102817. <https://doi.org/10.1016/j.media.2023.102817>.
- [29] Zhang, H., Liao, Q., Zhang, Y., Wang, S., Huang, Y., Wang, Q. (2023). Adversarial debiasing for medical imaging: Balancing fairness and clinical relevance. *IEEE Journal of Biomedical and Health Informatics*, 27(12), 6238–6249. <https://doi.org/10.1109/JBHI.2023.3268234>.
- [30] Upadhyay, K., Agrawal, M., Vashist, P. (2023). Learning multi-scale deep fusion for retinal blood vessel extraction in fundus images. *The Visual Computer*, 39, 4445–4457. <https://doi.org/10.1007/s00371-022-02600-4>.
- [31] Ryu, J.S., Lee, J., Kim, H., Kang, E., Park, J., Shin, J. (2025). Vision-language foundation models for medical imaging. *npj Digital Medicine*, 8, 152.
- [32] Xu, Z., Li, J., Yao, Q., Li, H., Zhao, M., Zhou, S. (2024). Addressing fairness issues in deep learning-based medical image analysis: A systematic review. *npj Digital Medicine*, 7, 286. <https://doi.org/10.1038/s41746-024-01276-5>.
- [33] Tan, S., Gao, Z., Zhang, W., Zhang, C., Wang, Y., Zhang, Y. (2024). Fairness in vision-language models: Measurement, mitigation, and beyond. *arXiv preprint arXiv:2402.09873*.
- [34] Guerroudji, M.A., et al.: A 3D visualization-based augmented reality application for brain tumor segmentation. *Comput. Anim. Virtual Worlds* 35(1), e2223 (2024). <https://doi.org/10.1002/cav.2223>.
- [35] Jin, Y., Zhang, X., Sun, Y., Tan, W., Zhou, X., Ma, C. (2024). Deep learning-driven automated quality assessment of ultra-widefield optical coherence tomography angiography images for diabetic retinopathy. *The Visual Computer*, <https://doi.org/10.1007/s00371-024-03383-6>.
- [36] Guerroudji, A., Fichou, D., Cheung, C.Y., Girard, M.J. (2024). Multi-modal deep learning for glaucoma detection from fundus photographs and clinical data. *Biomedical Signal Processing and Control*, 93, 106051. <https://doi.org/10.1016/j.bspc.2024.106051>.
- [37] Bao, Y., et al.: Adaptive information fusion network for multi-modal personality recognition. *Comput. Anim. Virtual Worlds* 35(3), e2268 (2024). <https://doi.org/10.1002/cav.2268>.
- [38] Yi, C., Xiong, Z., Qi, Q., Wei, X., Bathla, G., Lin, C.-L., Mortazavi, B. J., & Yang, T. (2025). AdFair-CLIP: Adversarial Fair Contrastive Language-Image Pre-training for Chest X-rays. *Proceedings of the 28th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2025)*. https://doi.org/10.1007/978-3-031-28913-4_2088.