Supplementary Material (Appendices)

A. Dataset

This study utilizes the Harvard-FairSeg dataset [31], a first-of-its-kind large-scale benchmark for evaluating fairness in medical image segmentation. It consists of 10,000 scanning laser ophthalmoscopy (SLO) fundus images, each corresponding to a unique patient collected between 2010 and 2021 at a major academic eye hospital. Each image includes high-quality optic disc and cup segmentation masks, initially derived from co-registered 3D optical coherence tomography (OCT) data and further refined via expert manual annotation.

Additionally, each image includes two textual annotations: one from a clinical curator and another generated via ChatGPT. The dataset offers rich demographic annotations across five attributes: gender (female, male), race (Asian, Black, White), ethnicity (Non-Hispanic, Hispanic), preferred language (English, Spanish, Other), and marital status (married, single). In this study, we excluded marital status.

The dataset provides imbalanced subgroup representation: Asian (919), Black (1,473), and White (7,608), along with consistently annotated pixel-wise segmentation masks. This combination of imaging data, expert annotations, and fine-grained demographic metadata forms a strong foundation for fairness-aware algorithm development in glaucoma screening.

B. Evaluation Metrics

We evaluate model performance using both standard segmentation metrics and fairness-specific extensions.

B.1. Standard Segmentation Metrics

We compute the Dice coefficient and Intersection over Union (IoU) between the predicted segmentation mask $\hat{y}$ and the ground truth y, as defined in Equation 12.

$$\text{Dice} = \frac{2|\hat{y} \cap y|}{|\hat{y}| + |y|}, \quad \text{IoU} = \frac{|\hat{y} \cap y|}{|\hat{y} \cup y|} \tag{12}$$

B.2. Equity-Scaled Performance Metrics

To incorporate fairness, we adopt Equity-Scaled (ES) metrics that account for group disparities. Let M_g be the metric value for demographic group g, with G total groups. The average metric is computed in Equation 13.

$$\bar{M} = \frac{1}{|G|} \sum_{g \in G} M_g \tag{13}$$

The equity-scaled version is defined in Equation 14.

$$\text{ES-Metric} = \frac{\bar{M}}{1 + \sum_{g \in G} |\bar{M} - M_g|} \tag{14}$$

This formulation penalizes disparities by scaling the mean performance $\bar{M}$ based on the total deviation across groups. When all group metrics M_g are equal, the penalty term disappears and the ES-Metric equals the mean. In contrast, greater divergence across groups results in a lower ES-Metric, thereby capturing both accuracy and fairness in a single measure.

B.3. Fairness Disparity Metrics

We compute three disparity measures:

- **Standard Deviation (STD)**: measures variance across groups, as shown in Equation 15.

$$\text{STD} = \sqrt{\frac{1}{|G|} \sum_{g \in G} (M_g - \bar{M})^2} \tag{15}$$

- **Disparity Index (DI)**: is the mean absolute deviation from the average, defined in Equation 16.

$$\text{DI} = \frac{1}{|G|} \sum_{g \in G} |M_g - \bar{M}| \tag{16}$$

- **Relative Performance Gap (RPG)**: is the gap between the best and the worst attribute performances, computed according to Equation 17.

$$\text{RPG} = \frac{\max_{g \in G} M_g - \min_{g \in G} M_g}{\max_{g \in G} M_g} \times 100\% \tag{17}$$

These metrics help quantify not only overall performance but a lso how equitably the model performs across demographic groups.

C. Baseline Models

We adopt LViT [18] and SAMed [41] as our primary baselines, representing recent state-of-the-art vision-language models for medical image segmentation. LViT utilizes large-scale visual-language pre-training, while SAMed adapts the Segment Anything Model (SAM) [13] for the medical domain. FairVLM retains the core strengths of these baselines but reduces performance disparities across subgroups. It maintains robust segmentation under varied textual prompts, addressing a key limitation of existing prompt-sensitive vision-language models.

D. Comparison with Existing Fairness and Prompt-Robustness Approaches

Table 7 provides a comprehensive comparison of FairVLM against SOTA fairness-aware frameworks and prompt-robust vision-language models. Methods such as GroupDRO [27] and FEBS [31] improve demographic fairness but lack robustness to prompt variation, making them insufficient for real-world VLM deployment where linguistic variability is common. Conversely, prompt optimization

Table 7. Comparison of FairVLM with related approaches on fairness, prompt robustness, and multimodal support.

Method	Demographic Fairness	Prompt Robustness	Multimodal Support	Limitations
GroupDRO [27] / FEBS [31]	✓	✗	✓	Ignores linguistic variation in prompts
CoOp [44] / CoCoOp [43]	✗	✓	✓	Not fairness-aware; may amplify subgroup bias
Adversarial De-biasing [21]	✓ (Unimodal)	✗	✓ (Limited)	Fragile in VLMs; often reduces accuracy
PromptSmooth [8] / MVP [17]	✗	✓	✓	No fairness calibration across demographic groups
FairVLM (Proposed)	✓	✓	✓	Jointly addresses fairness and robustness

Table 8. Evaluation of SRCP-generated counterfactual (CF) prompts on similarity, diversity, validity, score rank, and selection outcome.

CF #	Prompt	Similarity (S)	Diversity (D)	Valid	Score	Rank	Selected (Top 3)
–	*Original:* The 56 y/o female patient has optic nerve head drusen and narrow angles in both eyes, and a history of basilar artery aneurysms. No evidence of glaucoma mentioned.	–	–	–	–	–	Reference
1	A 76-year-old male presents with bilateral narrow anterior chamber angles and optic disc drusen. He has a known history of basilar artery aneurysms. Glaucoma is not observed.	92.24	47.20	✓	0.6522	3	✓
2	A 65-year-old individual with optic disc calcifications and reduced iridocorneal angles bilaterally, along with a previous diagnosis of basilar artery aneurysms. Glaucoma findings are absent.	91.86	49.05	✓	0.6617	1	✓
3	The patient is a 43-year-old woman with optic nerve head deposits and angle narrowing in both eyes. Past medical records note basilar artery aneurysms. There are no indications of glaucoma.	92.12	48.31	✓	0.6583	2	✓
4	The 57-year-old female patient has optic nerve head drusen and shallow angles in both eyes. She has a known history of basilar artery aneurysms. Glaucoma is not detected.	91.94	38.10	✓	0.5964	4	×
5	A 54 y/o woman presents with optic disc drusen and bilateral narrow angles. She has been previously diagnosed with basilar artery aneurysms. Glaucoma is not present.	91.63	37.50	✓	0.5915	5	×
Invalid	The 56-year-old female patient has optic nerve swelling and narrow angles in both eyes, along with a history of basilar artery aneurysms. Glaucoma is not currently diagnosed.	81.19	32.54	×	–	–	×

techniques like CoOp [44], CoCoOp [43], PromptSmooth [8], and MVP [17] enhance robustness but are not fairness-aware, and in some cases may even amplify subgroup disparities. Adversarial de-biasing [21] has been explored in unimodal or limited multimodal settings, but such methods tend to degrade performance and are less stable in complex VLM architectures. In contrast, FairVLM is a method that jointly addresses both fairness and prompt robustness in a unified framework, while fully supporting multimodal inputs. This balance makes FairVLM uniquely suited for clinical applications where equity, reliability, and linguistic variability are critical.

E. Complete Training Algorithm

Algorithm 1 outlines the entire training procedure of FairVLM. The training begins with the generation of multiple semantically relevant counterfactual (CF) prompts using GPT-based perturbations of both demographic and clinical attributes (Module 1). For each generated prompt, the vision-language model predicts segmentation masks, which are then processed for additional loss computations. To reduce demographic bias in feature representations, Demographic-Aware Feature Normalization (DAFN) is applied (Module 2), where encoded features are normalized using attribute-wise statistics maintained through exponential moving averages across training iterations. Next, the Counterfactual Prompt Regularization (CPR) loss is computed (Module 3) to enforce consistency between original and counterfactual segmentation outputs. Fairness-Calibrated Loss (FCL) is calculated (Module 4) by adaptively penalizing attribute-specific disparities in segmentation performance, using attribute-level Dice scores, fairness gaps, and entropy-based importance weights. Finally, all loss components, such as segmentation loss, CPR loss, and fairness loss, are combined into a total objective function, and model parameters are updated accordingly through gradient descent (Module 5). This iterative process jointly optimizes segmentation accuracy, counterfactual robustness, and demographic fairness throughout training.

F. Illustration of SRCP Mechanism

Table 8 demonstrates an example of how the Semantically Relevant Counterfactual Prompting (SRCP) module operates. Given an original clinical prompt describing a patient's condition, SRCP employs a large language model to generate multiple counterfactual prompts by perturbing both demographic and clinical (terminology, phrasing) attributes. The generated prompts are evaluated based on (i) *semantic similarity* to ensure clinical coherence, (ii) *diversity* to capture meaningful variations, and (iii) *clinical validity* to ensure plausibility for downstream training.

For each original prompt, the SRCP module executes at least m times until it generates $m = 5$ valid counterfactuals. Each candidate counterfactual is considered valid only if it

Algorithm 1 FairVLM: Fair Vision-Language Model

Input: Image x, label y, prompt $p = (\mathcal{G}, C)$, demographic groups $\mathcal{G}$, clinical attributes C, language model $GPT-4o$
Hyperparameters: k, m, ϵ_1, ϵ_2, τ, λ, α
Module 1: Semantically Relevant Counterfactual Prompts (SRCP)
Initialize valid set $P^{'} \leftarrow \emptyset$
while $|P^{'}| < m$ **do**
  Generate candidate counterfactual prompt $p_i^{'}$ from LLM by perturbing $\mathcal{G}$ and/or C
  Compute Jaccard distance $D_i = \delta(p, p_i^{'})$
  Compute cosine similarity $S_i = \cos(\phi(p), \phi(p_i^{'}))$
  if $D_i \in [\epsilon_1, \epsilon_2]$ **and** $S_i \geq \tau$ **then**
    Compute Score$(p_i^{'}) = \lambda \cdot D_i + (1 - \lambda) \cdot S_i$
    Add $(p_i^{'}, \text{Score}(p_i^{'}))$ to $P^{'}$
  end if
end while
Sort $P^{'}$ by descending $Score$
Select top k counterfactuals $\{p_1^{'}, \ldots, p_k^{'}\}$ from $P^{'}$
Define augmented prompt set: $P_a = \{p, p_1^{'}, \ldots, p_k^{'}\}$
Module 2: Segmentation and DAFN (Demographic-Aware Feature Normalization)
for each prompt $p \in P_a$ **do**
  Encode features $z = E_\theta(x, p)$
  for each $g \in G$ **do**
    Compute μ_g, σ_g from group samples D_g
    Update EMA: $\hat{\mu}_g^{(t)} = \alpha \mu_g^{(t)} + (1 - \alpha)\hat{\mu}_g^{(t-1)}$
    Update EMA: $\hat{\sigma}_g^{(t)} = \alpha \sigma_g^{(t)} + (1 - \alpha)\hat{\sigma}_g^{(t-1)}$
  end for
  Aggregate: $\mu_{\text{avg}} = \frac{1}{|G|} \sum_{g \in G} \hat{\mu}_g, \quad \sigma_{\text{avg}} = \frac{1}{|G|} \sum_{g \in G} \hat{\sigma}_g$
  Normalize: $\hat{z} = \frac{z - \mu_{\text{avg}}}{\sigma_{\text{avg}}}$
  Normalize encoded text features similarly
  Decode segmentation: $\hat{y}_p = D_\theta(\hat{z}_I, \hat{z}_p)$
end for
Module 3: Counterfactual Prompt Regularization (CPR)
Compute $\mathcal{L}_{\text{CPR}} = \sum_{i=1}^{k} \left(Dice(y, \hat{y}_{p_i'}) + BCE(y, \hat{y}_{p_i'}) \right)$
Module 4: Fairness-Calibrated Loss (FCL)
for each $g \in \mathcal{G}$ **do**
  Track EMA Dice: $\hat{\text{Dice}}_g^{(t)} = (1 - \alpha)\hat{\text{Dice}}_g^{(t-1)} + \alpha \text{Dice}_g^{(t)}$
end for
Compute fairness gap: $\Delta_{\text{gap}} = \max_g \hat{\text{Dice}}_g - \min_g \hat{\text{Dice}}_g$
Compute mean Dice: $\bar{\text{Dice}} = \frac{1}{|\mathcal{G}|} \sum_g \hat{\text{Dice}}_g$
Compute entropy weights: $\pi_g = \frac{-n_g \log(n_g)}{\sum_g -n_g' \log(n_g')}, \; n_g' = \frac{1}{\log(1+n_g)}$
Compute fairness loss:

$$\mathcal{L}_{\text{FCL}} = \sum_g \pi_g \log \left(\frac{\hat{\text{Dice}}_g}{\bar{\text{Dice}} + \epsilon} \right) + \Delta_{\text{gap}}$$

Module 5: Final Loss and Optimization
Compute base loss: $\mathcal{L}_{\text{base}} = Dice(y, \hat{y}_p) + BCE(y, \hat{y}_p)$
Combine all losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{base}} + \mathcal{L}_{\text{CPR}} + \mathcal{L}_{\text{FCL}}$$

Update model parameters θ via gradient descent using $\mathcal{L}_{\text{total}}$

Table 9. Group-wise Dice scores and fairness metrics on the Harvard-FairSeg dataset. FairVLM achieves the highest ES-Dice and lowest STD, DI, and RPG, indicating improved fairness across sex, ethnicity, race, and language subgroups for both Cup and Rim regions. We evaluate the relative performance of the models across demographic groups, including Male (M), Female (F), Hispanic (Hisp.), Non-Hispanic (NonH.), Asian (Asi.), Black (Bla.), White (Whi.), English (Eng.), Spanish (Span.), and Others (Oth.).

Region	Method	ES-Dice	Dice	M	F	Hisp.	NonH.	Asi.	Bla.	Whi.	Eng.	Span.	Oth.	STD	DI	RPG
Cup	SAMed	84.53	86.71	86.47	87.03	86.53	**89.04**	85.68	87.30	86.70	86.52	90.77	88.38	1.53	1.17	5.61
	SAMed +ADV	84.83	86.82	86.61	86.75	86.61	88.83	85.90	87.05	87.08	86.68	91.31	88.20	1.58	1.17	5.92
	SAMed+GroupDRO	85.13	86.92	86.72	86.70	86.82	88.55	85.83	87.04	87.06	86.84	**90.85**	88.49	1.44	1.08	5.53
	SAMed+FEBS	85.12	86.86	87.18	87.56	87.04	88.24	85.87	87.08	86.72	86.70	90.34	87.94	1.21	0.84	4.95
	SAMed+FairVLM	**86.42**	**87.25**	**87.24**	**87.61**	**87.27**	88.53	**86.91**	**87.31**	**87.37**	**87.18**	88.58	**88.76**	**0.70**	**0.60**	**2.08**
	LViT	85.63	87.37	83.22	87.45	87.38	90.62	85.02	87.36	87.46	87.24	90.71	88.87	2.27	1.52	8.26
	+ADV	85.74	87.46	**88.35**	87.58	82.48	88.75	87.13	87.40	85.57	89.37	90.84	88.96	2.31	1.61	9.20
	LViT+GroupDRO	85.88	87.53	87.41	83.65	87.56	**90.83**	83.26	87.48	87.64	87.46	90.91	89.08	2.54	1.68	8.41
	LViT+FEBS	85.91	87.58	87.53	87.72	87.64	88.91	87.28	87.53	**87.72**	87.49	**91.02**	89.18	1.18	0.90	4.11
	LViT+FairVLM	**87.08**	**87.87**	87.61	**87.74**	**87.66**	89.03	**87.61**	**87.62**	87.65	**87.62**	87.23	**89.32**	**0.69**	**0.51**	**2.34**
Rim	SAMed	79.41	82.91	83.19	82.52	82.77	83.97	78.90	77.58	**84.44**	83.05	85.34	79.89	2.53	2.03	9.09
	SAMed+ADV	79.38	82.91	83.42	82.63	83.08	**84.16**	78.01	76.91	83.95	83.07	85.28	80.15	2.77	2.23	9.81
	SAMed+GroupDRO	79.84	83.08	**83.53**	82.74	82.84	83.90	79.52	77.48	84.54	83.22	84.93	80.65	2.38	1.87	8.77
	SAMed+FEBS	79.88	83.21	83.38	82.89	83.49	84.08	79.52	77.89	84.73	83.28	85.11	79.92	2.44	1.99	8.48
	SAMed+FairVLM	**81.03**	**83.52**	83.22	**83.34**	**83.51**	83.97	**83.23**	**83.28**	83.42	**83.33**	**85.39**	**82.98**	**0.70**	**0.45**	**2.82**
	LViT	80.19	83.45	85.36	83.47	80.34	84.08	81.36	85.43	81.55	83.45	85.52	81.16	1.95	1.66	6.06
	LViT+ADV	80.26	83.52	83.49	80.58	83.46	84.19	86.42	83.51	82.66	83.53	82.63	81.30	1.59	1.11	6.76
	LViT+GroupDRO	80.48	83.61	83.57	83.67	83.54	86.27	**86.49**	83.62	83.74	83.61	85.70	80.41	1.76	1.25	7.03
	LViT+FEBS	80.61	83.67	83.66	83.75	83.63	84.36	83.57	83.69	83.82	83.69	**87.82**	81.50	1.55	0.86	7.20
	LViT+FairVLM	**81.82**	**83.81**	**83.79**	**83.82**	**83.76**	**84.44**	83.81	**83.80**	**83.87**	**83.81**	85.91	**83.62**	**0.68**	**0.44**	**2.67**

Table 10. Group-wise IoU scores and fairness metrics on Harvard-FairSeg. FairVLM achieves the highest ES-IoU and lowest STD, DI, and RPG, indicating improved fairness across sex, ethnicity, race, and language subgroups for both Cup and Rim regions. We evaluate the relative performance of the models across demographic groups including Male (M), Female (F), Hispanic (Hisp.), Non-Hispanic (NonH.), Asian (Asi.), Black (Bla.), White (Whi.), English (Eng.), Spanish (Span.), and Others (Oth.).

Region	Method	ES-IoU	IoU	M	F	Hisp.	NonH.	Asi.	Bla.	Whi.	Eng.	Span.	Oth.	STD	DI	RPG
Cup	SAMed	75.64	78.13	77.83	78.55	77.90	**81.00**	76.88	79.05	78.08	77.91	83.38	80.01	1.93	1.44	7.80
	SAMed+ADV	75.86	78.22	77.91	78.20	77.91	80.80	77.09	78.82	78.46	78.08	**84.32**	79.82	2.10	1.50	8.57
	SAMed+GroupDRO	76.24	78.33	78.04	78.14	78.19	80.44	77.11	78.86	78.42	78.25	83.60	80.19	1.86	1.37	7.76
	SAMed+FEBS	76.16	78.26	78.51	78.79	79.04	80.70	77.08	78.82	78.04	78.12	82.68	79.37	1.57	1.08	6.77
	SAMed+FairVLM	**77.86**	**79.12**	**78.71**	**79.97**	**79.16**	78.88	**78.93**	**79.17**	**78.82**	**78.71**	78.99	**80.71**	**0.64**	**0.45**	**2.48**
	LViT	76.34	78.19	78.46	76.19	**79.09**	78.49	74.67	79.48	79.31	77.87	79.64	74.96	1.87	1.53	6.24
	LViT+ADV	76.63	78.82	75.04	**79.68**	78.03	77.57	79.21	79.16	78.03	75.16	79.24	74.69	1.93	1.57	6.26
	LViT+GroupDRO	76.74	78.81	78.29	78.73	73.13	77.98	78.12	79.08	78.24	78.58	79.12	73.51	2.23	1.66	7.57
	LViT+FEBS	76.71	78.66	78.26	74.06	78.17	74.91	78.18	75.07	78.90	78.55	79.08	74.22	2.08	1.90	6.35
	LViT+FairVLM	**77.73**	**79.21**	**78.73**	78.80	78.96	**78.75**	**79.89**	**79.74**	**79.77**	**78.86**	**79.77**	**77.93**	**0.59**	**0.44**	**2.31**
Rim	SAMed	68.74	72.17	72.52	71.69	72.03	73.07	67.43	65.87	73.99	72.34	**74.68**	68.71	2.91	2.34	11.80
	SAMed+ADV	68.69	72.12	72.76	71.81	72.41	73.42	66.35	64.98	73.25	72.31	74.63	69.24	3.20	2.56	12.93
	SAMed+GroupDRO	69.21	72.37	72.92	71.98	72.08	72.98	68.22	65.68	74.10	72.53	74.12	69.54	2.74	2.16	11.37
	SAMed+FEBS	69.15	72.50	72.74	72.23	72.78	73.29	68.25	66.20	**74.39**	72.63	74.36	68.65	2.82	2.31	11.01
	SAMed+FairVLM	**71.19**	**73.50**	**73.35**	**74.32**	**73.22**	**73.49**	**74.78**	**73.51**	74.11	**73.42**	73.65	**74.68**	**0.57**	**0.50**	**2.09**
	LViT	70.33	73.16	72.44	72.95	66.42	70.32	69.03	70.01	72.48	72.88	73.03	65.84	2.72	2.22	9.85
	LViT+ADV	70.26	73.22	72.74	72.48	73.43	69.65	65.18	71.76	71.01	68.19	72.77	70.01	2.55	1.97	11.24
	LViT+GroupDRO	70.19	73.32	70.19	72.84	70.13	73.41	71.35	73.23	66.06	70.22	**73.62**	68.71	2.42	1.91	10.27
	LViT+FEBS	70.28	73.29	73.17	73.71	72.73	73.71	68.02	73.14	**73.67**	69.04	72.36	72.87	2.02	1.48	7.72
	LViT+FairVLM	**71.43**	**73.53**	**73.56**	**74.39**	**73.48**	**74.01**	**73.51**	**73.35**	73.38	**73.24**	73.21	**73.27**	**0.38**	**0.27**	**1.59**

satisfies the similarity threshold (cosine similarity $\geq 90\%$) and diversity constraint (Jaccard distance within $[0.3, 0.5]$). When it generates m number valid counterfactuals, it ranks them based on the following equation to get top k relevant counterfactual prompts:

$$\text{Score}(p_i^{'}) = \lambda \cdot D_i + (1 - \lambda) \cdot S_i$$

where $\lambda = 0.4$, D_i denotes Jaccard distance, and S_i is the cosine similarity between sentence embeddings. This score balances lexical diversity and semantic consistency. Prompts with higher scores are considered more informative and robust for model training.

As shown in Table 8, Counterfactuals 1–5 pass the validity criteria and are scored accordingly, while the last counterfactual of the table is discarded as it does not satisfy both conditions and is considered invalid. The top $k = 3$ coun-

Table 11. Out-of-distribution (OOD) generalization performance of FairVLM with SAMed and LViT backbones. Each model is trained on either manual or GPT-generated prompts and evaluated on the other. Metrics are reported as (Cup, Rim), demonstrating FairVLM's robustness to prompt domain changes.

Model	Train → Test	ES-Dice	Dice	ES-IoU	IoU	DI Dice	DI IoU	RPG Dice	RPG IoU	STD Dice	STD IoU
FairVLM (SAMed)	Manually Written → GPT-Generated	(86.53, 80.41)	(87.21, 83.39)	(77.64, 70.74)	(78.83, 72.97)	(0.87, 0.73)	(0.64, 0.75)	(2.61, 3.09)	(3.80, 3.06)	(0.52, 0.65)	(0.92, 2.90)
	GPT-Generated → Manually Written	(86.12, 80.24)	(87.01, 83.19)	(77.34, 70.62)	(78.52, 72.75)	(0.88, 0.76)	(0.67, 0.78)	(2.72, 3.15)	(3.92, 3.14)	(0.54, 0.68)	(0.95, 0.87)
FairVLM (LViT)	Manually Written → GPT-Generated	(86.47, 80.52)	(87.29, 83.22)	(77.43, 70.55)	(78.61, 72.68)	(0.89, 0.80)	(0.70, 0.83)	(2.69, 3.11)	(3.88, 3.11)	(0.53, 0.66)	(0.79, 0.91)
	GPT-Generated → Manually Written	(86.34, 80.33)	(87.16, 83.10)	(77.18, 70.48)	(78.44, 72.58)	(0.90, 0.77)	(0.72, 0.79)	(3.75, 2.05)	(3.95, 3.08)	(0.55, 0.67)	(0.58, 0.61)

Table 12. Cross-dataset generalization performance of FairVLM with SAMed and LViT backbones. Models trained on the Harvard-FairSeg dataset are evaluated on external test sets (MosMedData+, QaTa-COV19). Metrics include segmentation accuracy (ES-Dice, Dice, ES-IoU, IoU) and fairness indicators (DI, RPG, STD), showing FairVLM's strong generalizability and equitable performance across domains.

Model	Train → Test	ES-Dice	Dice	ES-IoU	IoU	DI Dice	DI IoU	RPG Dice	RPG IoU	STD Dice	STD IoU
SAMed	MosMedData+ → MosMedData+	71.52	75.46	58.41	62.30	2.95	3.16	6.49	7.56	3.38	4.88
	QaTa-COV19 → QaTa-COV19	79.91	83.21	72.12	75.11	2.90	3.11	5.25	6.31	2.34	3.84
FairVLM (SAMed)	Harvard-FairSeg → MosMedData+	73.12	73.30	60.12	60.55	1.25	1.44	3.72	4.10	0.92	1.25
	Harvard-FairSeg → QaTa-COV19	81.12	81.79	73.64	73.91	0.87	0.76	2.41	2.88	0.69	1.03
LViT	MosMedData+ → MosMedData+	71.41	75.25	57.33	61.58	3.98	3.20	6.52	6.61	2.41	3.91
	QaTa-COV19 → QaTa-COV19	79.84	83.95	72.01	75.28	2.92	3.14	6.29	6.37	2.37	2.87
FairVLM (LViT)	Harvard-FairSeg → MosMedData+	73.11	73.89	59.63	60.01	1.17	1.28	3.45	3.72	0.88	1.19
	Harvard-FairSeg → QaTa-COV19	81.56	82.31	73.01	73.67	0.79	0.71	2.16	2.63	0.66	0.97

terfactual prompts based on the SRCP score are selected and sent through the model one by one as the original prompt during model training. This mechanism ensures that the model is exposed to diverse, clinically meaningful, and linguistically varied descriptions during training, thereby improving its generalization to real-world prompt variability.

G. More Fairness and Ablation Studies

G.1. Comparative Demographic Fairness Evaluation of FairVLM

We provide full attribute-wise fairness results for both Dice and IoU-based metrics in Tables 9 and 10. We couldn't include these detailed results in the main paper due to page limitations. The analysis includes attribute-wise performance across sex (Male, Female), ethnicity (Hispanic, Non-Hispanic, Asian, Black, White), and language (English, Spanish, Others), alongside summary fairness indicators: standard deviation (STD), disparity index (DI), and range performance gap (RPG). FairVLM consistently outperforms both standard VLM baselines (SAMed, LViT) and fairness-aware baselines (ADV, GroupDRO, FEBS) in reducing attribute-wise performance gaps across both anatomical regions (Cup and Rim). Notably, FairVLM achieves the lowest STD, DI, and RPG scores, indicating improved fairness with strong segmentation accuracy across diverse subgroups. These results further demonstrate that FairVLM effectively mitigates demographic disparities while preserving or improving overall model performance.

G.2. Performance of FairVLM on Out-of-Distribution Data in Detail

Because of space limitations in the main paper, we couldn't include the full evaluation of FairVLM's performance on out-of-distribution (OOD) data. Tables 11 and 12 show a complete look at FairVLM's robustness across two OOD scenarios: (i) variations in prompt style, and (ii) generalization across datasets.

In Table 11, FairVLM is evaluated by training on one prompt style (either manually written or GPT-generated) and testing on the other. Across both SAMed and LViT backbones, performance remains notably consistent. FairVLM (SAMed) achieves an ES-Dice of (86.53, 80.41) when trained on manual prompts and tested on GPT-generated ones, compared to (86.12, 80.24) for the inverse setting. Likewise, the LViT backbone achieves (86.47, 80.52) and (86.34, 80.33) in the two respective contexts. Variations in ES-IoU, DI, and RPG metrics are similarly minimal in these experiments. These findings demonstrate that FairVLM is highly resilient to changes in prompt formation.

Table 12 presents the cross-dataset generalization results of FairVLM with SAMed and LViT backbones, trained on Harvard-FairSeg and evaluated on MosMedData+ and QaTa-COV19. While baseline models (SAMed and LViT) trained and tested on the same dataset achieve higher overall segmentation performance (Dice and IoU), FairVLM demonstrates superior equitable segmentation with higher ES-Dice and ES-IoU in most cases.

For instance, on MosMedData+, FairVLM (SAMed) achieves an ES-Dice of 73.12 and ES-IoU of 60.12, outperforming the baseline SAMed (71.52 and 58.41, respec-

Table 13. Effect of DAFN during inference on segmentation performance and fairness. Metrics are reported as (Cup, Rim).

Model	DAFN		ES-Dice	ES-IoU	DI Dice	DI IoU	RPG Dice	RPG IoU	STD Dice	STD IoU
	Train	Test								
FairVLM (SAMed)	×	×	(84.97, 79.58)	(75.94, 69.02)	(1.01, 1.68)	(1.21, 1.98)	(4.13, 7.36)	(5.97, 9.31)	(1.82, 1.86)	(1.66, 1.52)
FairVLM (SAMed)	✓	✓	(86.42, 81.03)	(77.86, 71.19)	(0.60, 0.45)	(0.46, 0.50)	(2.08, 2.82)	(2.48, 2.09)	(0.70, 0.71)	(0.64, 0.57)
FairVLM (SAMed)	✓	×	(86.23, 80.87)	(77.74, 71.08)	(0.68, 0.52)	(0.53, 0.56)	(2.19, 2.93)	(2.55, 2.21)	(0.77, 0.74)	(0.72, 0.66)
FairVLM (LViT)	×	×	(85.91, 80.67)	(76.63, 70.72)	(1.17, 1.21)	(1.24, 1.54)	(6.38, 5.29)	(4.38, 7.64)	(1.75, 1.89)	(1.57, 1.63)
FairVLM (LViT)	✓	✓	(87.08, 81.82)	(77.73, 71.43)	(0.51, 0.44)	(0.43, 0.27)	(2.34, 2.67)	(2.31, 1.59)	(0.69, 0.68)	(0.59, 0.38)
FairVLM (LViT)	✓	×	(86.83, 81.71)	(77.62, 71.30)	(0.59, 0.51)	(0.51, 0.38)	(2.42, 2.73)	(2.44, 1.68)	(0.72, 0.79)	(0.71, 0.52)

tively) despite the baseline having higher Dice (75.46 vs. 73.30) and IoU (62.30 vs. 60.55). A similar pattern holds for the LViT backbone, where FairVLM (LViT) achieves higher ES-Dice (73.11 vs. 71.41) and ES-IoU (59.63 vs. 57.33), whereas the baseline LViT yields better Dice (75.25) and IoU (61.58).

On the QaTa-COV19 dataset, FairVLM also maintains better equitable segmentation. FairVLM (SAMed) achieves an ES-Dice of 81.12 and ES-IoU of 73.64, surpassing SAMed's baseline ES-Dice of 79.91 and ES-IoU of 72.12. Similarly, FairVLM (LViT) reports an ES-Dice of 81.56 and ES-IoU of 73.01, outperforming baseline LViT (79.84 and 72.01). However, baseline models still show slightly higher Dice and IoU scores (83.95 vs. 82.31 for LViT).

These results confirm that although baseline models trained and tested on the same dataset benefit from better overall segmentation metrics, FairVLM consistently provides more balanced and equitable performance across demographic subgroups. Similar trends are observed in disparity indicators (DI, RPG, and STD), where FairVLM maintains significantly lower values, highlighting its fairness-aware design and robust generalization across unseen clinical datasets.

These findings collectively show that FairVLM demonstrates strong out-of-distribution generalization, maintaining high segmentation accuracy and fairness across changes in prompt domain and dataset distribution.

G.3. Effect of DAFN During Inference

To evaluate the role of DAFN during inference, we conducted an ablation study comparing models trained with DAFN but evaluated with and without it at test time. DAFN uses Exponential Moving Averages (EMA) of the feature-wise mean and standard deviation, calculated across demographic groups during training, for normalization during inference. As shown in Table 13, a × sign indicates that DAFN is deactivated at a specific stage. The results show that using DAFN during both training and inference achieves the best segmentation and fairness outcomes. Removing DAFN at inference leads to only marginal performance drops. For example, with FairVLM (SAMed), ES-Dice decreases slightly from 86.42 to 86.23 (Cup) and from 81.03 to 80.87 (Rim); with FairVLM (LViT), the drop is from 87.08 to 86.83 (Cup) and from 81.82 to 81.71 (Rim). Fairness metrics such as DI, RPG, and STD also remain largely unaffected, indicating robustness to the absence of DAFN at test time. In contrast, models trained without DAFN show considerably worse fairness outcomes, including larger disparity and higher group-level variance. For instance, RPG Dice increases from 2.08 to 4.13 for SAMed and from 2.34 to 6.38 for LViT, when DAFN is not used during training. These results confirm that incorporating DAFN during training is essential for promoting equitable representation, while its use during inference further enhances fairness without being strictly required for maintaining robust performance.

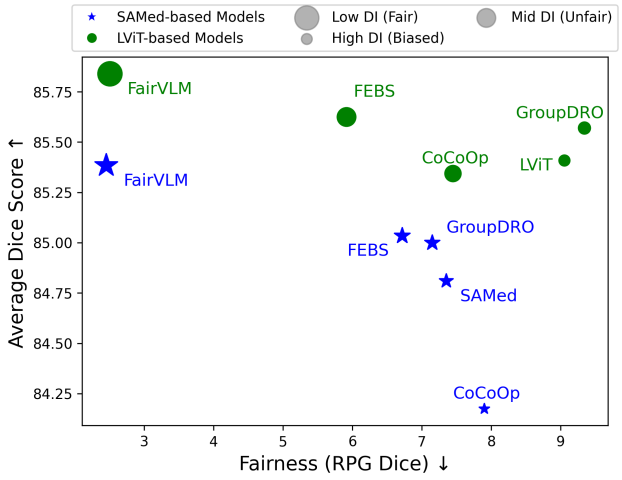

Figure 6. Illustration of the accuracy-fairness tradeoff of compared models. Reported results represent the average of Cup and Rim scores for each metric, providing a concise summary of performance.

G.4. Fairness-Accuracy Tradeoff Analysis

To assess the tradeoff between segmentation accuracy and fairness, we compare SAMed- and LViT-based models using the average of Cup and Rim metrics. Specifically, we use the combined Dice score to represent accuracy, and RPG Dice and DI Dice as fairness metrics, where lower values indicate better subgroup fairness. As shown in Figure 6, FairVLM (SAMed) achieves a combined Dice of 85.39, with the lowest RPG Dice (2.45) and DI Dice (0.53) among

Table 14. Intersectional subgroup performance on ES-Dice and ES-IoU metrics for Cup and Rim regions. Results are shown for SAMed and LViT backbones with and without FairVLM. FairVLM consistently improves performance across sex, ethnicity, and their intersectional subgroups.

Attributes	**SAMed**		**FairVLM (SAMed)**		**LViT**		**FairVLM (LViT)**	
	ES-Dice	ES-IoU	ES-Dice	ES-IoU	ES-Dice	ES-IoU	ES-Dice	ES-IoU
Male	(86.47, 83.19)	(77.83, 72.52)	(87.14, 83.22)	(78.71, 73.35)	(83.22, 85.36)	(78.46, 72.44)	(87.61, 83.79)	(78.73, 73.56)
Female	(87.03, 82.52)	(78.55, 71.69)	(87.32, 83.34)	(79.97, 74.32)	(87.45, 83.47)	(77.19, 72.95)	(87.68, 83.82)	(78.80, 74.39)
Hispanic	(86.53, 82.77)	(77.90, 72.03)	(87.27, 83.29)	(79.16, 73.22)	(87.38, 82.34)	(79.09, 72.42)	(87.63, 83.76)	(78.96, 73.48)
Hispanic Male	(84.40, 81.78)	(74.77, 68.08)	(87.10, 83.05)	(78.84, 73.08)	(84.20, 80.65)	(75.68, 66.23)	(87.52, 83.58)	(78.74, 73.32)
Hispanic Female	(83.63, 80.97)	(74.01, 68.23)	(87.39, 83.51)	(79.66, 73.97)	(85.18, 79.54)	(75.19, 66.62)	(87.76, 83.99)	(78.98, 74.14)
Asian Male	(81.58, 78.80)	(73.78, 67.33)	(86.81, 83.13)	(78.83, 74.68)	(84.92, 80.26)	(75.57, 67.93)	(87.51, 83.71)	(78.63, 73.41)
Asian Female	(82.71, 79.16)	(74.64, 67.52)	(87.67, 83.21)	(79.64, 74.56)	(85.11, 81.33)	(75.48, 67.06)	(87.44, 84.02)	(78.46, 73.47)

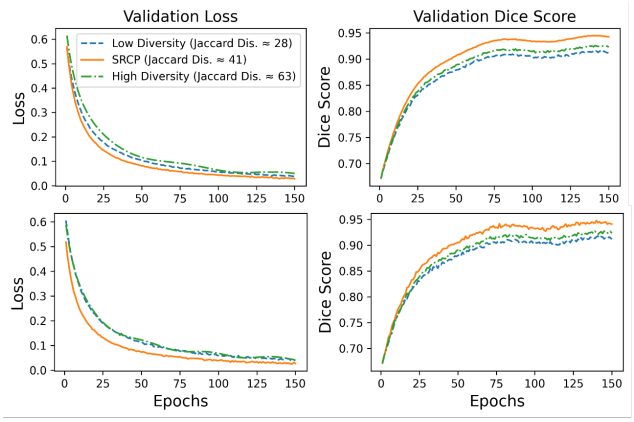

Figure 7. Training convergence under varying prompt diversity levels. Excessive and little variation hinder convergence, while moderate diversity (the default SRCP setting) promotes stable learning and enhances robustness.

all SAMed-based methods, indicating strong fairness without compromising accuracy. Similarly, FairVLM (LViT) outperforms its baseline and other fairness-enhancing counterparts, reaching a combined Dice of 85.84, while maintaining RPG Dice of 2.51 and DI Dice of 0.48. These results demonstrate that FairVLM successfully balances the tradeoff between segmentation accuracy and fairness, consistently outperforming alternative approaches across different backbone architectures.

G.5. Prompt-Diversity Trade-off Analysis

To understand how prompt diversity influences the learning of models, we compare the convergence behavior of FairVLM using SAMed and LViT backbones across three prompt-diversity scenarios: low diversity (10–30), medium diversity (30–50), and high diversity (50–70). As shown in Figure 7, the medium range (30–50), employed in our SRCP (Semantic-Retained Counterfactual Prompting) strategy, consistently led to faster and more stable convergence for FairVLM. As shown in the loss and Dice curves, SRCP achieves sharper validation loss reduction and higher Dice scores over training epochs compared to both lower and higher diversity settings. This indicates that moderate prompt diversity provides an optimal balance, encouraging generalization without introducing excessive variability, thus improving both training efficiency and final segmentation performance across different backbone architectures.

G.6. Intersectional Fairness Performance Comparison

Individuals often belong to multiple demographic subgroups, like Asian and Male, where compounding biases can degrade model performance. Table 14 presents an intersectional fairness evaluation showing how the quality of the segmentation changes between the combined sex and ethnic groups. Notably, the baseline model SAMed shows clear performance degradation when demographic attributes are combined. For instance, SAMed's ES-Dice drops from (86.47, 83.19) in the Male group to (84.40, 81.78) for Hispanic Males and further to (81.58, 78.80) for Asian Males. A similar downward trend is seen in ES-IoU: from (77.83, 72.52) to (74.77, 68.08) and (73.78, 67.33), respectively. This highlights the sensitivity of the baseline to intersectional identities.

In contrast, FairVLM (SAMed) either improves or maintains performance across all such subgroups without degradation. For Hispanic Males, FairVLM (SAMed) increases ES-Dice to (87.10, 83.05) and ES-IoU to (78.84, 73.08), outperforming SAMed. Similarly, in the Asian Male subgroup, it improves ES-Dice from (81.58, 78.80) to (86.81, 83.13) and ES-IoU from (73.78, 67.33) to (78.83, 74.68). Even in groups where baseline performance is relatively high, FairVLM either matches or exceeds it, demonstrating no performance trade-offs.

A similar pattern is observed for the LViT backbone. While LViT shows noticeable drops in some intersectional groups, FairVLM (LViT) improves or maintains these scores. Across all examined combinations, FairVLM consistently avoids performance degradation, providing a fairness-aware framework that ensures equitable and stable performance across demographic intersections.

Table 15. Sensitivity of FairVLM (SAMed and LViT) to prompt similarity (τ) and diversity (δ) thresholds. Model performance remains stable across variations. Reported values represent the average of Cup and Rim scores for each metric, providing a concise summary of performance.

Threshold Setting	Avg Dice	ES-Dice	DI Dice	RPG (%)
FairVLM (SAMed)				
$\tau = 0.85,\ \delta = [0.2,\ 0.4]$	84.95	82.30	0.63	2.96
$\tau = 0.90,\ \delta = [0.3,\ 0.5]$ *(default)*	85.39	83.73	0.53	2.45
$\tau = 0.95,\ \delta = [0.4,\ 0.6]$	84.11	82.35	0.58	2.51
FairVLM (LViT)				
$\tau = 0.85,\ \delta = [0.2,\ 0.4]$	84.45	82.15	0.59	2.89
$\tau = 0.90,\ \delta = [0.3,\ 0.5]$ *(default)*	85.84	84.45	0.48	2.51
$\tau = 0.95,\ \delta = [0.4,\ 0.6]$	84.72	83.17	0.55	2.67

Table 16. Sensitivity of FairVLM (SAMed and LViT) to the number of generated (m) and selected (k) prompts. Reported values reflect standard deviation across Cup and Rim regions and total training time. The default setting ($m = 5, k = 3$) yields the best balance of stability and efficiency.

Prompt Setting (m, k)	STD (Dice)	STD (IoU)	Train Time (h)
FairVLM (SAMed)			
$(3, 2)$	1.35	1.27	5.1
$(5, 2)$	1.31	1.22	5.2
$(5, 3)$ *(default)*	**0.88**	**0.79**	**5.9**
$(5, 4)$	0.87	0.77	6.8
$(7, 5)$	0.87	0.76	7.5
FairVLM (LViT)			
$(3, 2)$	1.29	1.32	6.2
$(5, 2)$	1.28	1.32	6.4
$(5, 3)$ *(default)*	**0.83**	**0.75**	**7.6**
$(5, 4)$	0.82	0.73	8.8
$(7, 5)$	0.81	0.72	10.1

G.7. Sensitivity Analysis of FairVLM to Prompt Thresholds

To assess the robustness of FairVLM to prompt design parameters, we analyze its performance under varying similarity (τ) and diversity (δ) thresholds using both SAMed and LViT backbones on the validation set (Table 15).

Across both backbones, the default threshold setting $\tau = 0.90,\ \delta = [0.3,\ 0.5]$ yields the optimal performance, achieving the highest accuracy and fairness metrics. Specifically, for FairVLM (SAMed), this setting leads to the best trade-off: an average Dice of 85.39, ES-Dice of 83.73, lowest disparity index (DI) of 0.53, and minimal relative performance gap (RPG) of 2.45%. Deviations from this setting (either more relaxed or stricter thresholds) result in small but consistent reductions in performance.

A similar trend is observed for FairVLM (LViT), where the default setting again results in the highest average Dice (85.84), highest ES-Dice (84.45), and lowest DI (0.48) and RPG (2.51%). The performance under non-default thresholds remains relatively stable, with average Dice scores varying within a narrow range (± 0.7), and fairness metrics fluctuating only marginally.

These findings demonstrate that while FairVLM is generally robust to a range of prompt thresholds, the default configuration provides the most balanced and reliable outcome in terms of both segmentation accuracy and subgroup fairness.

G.8. Sensitivity to Prompt Count

We assess the sensitivity of FairVLM (SAMed and LViT) to the number of selected prompts k, across different numbers of generated prompts m. As shown in Table 16, for each m, we present the standard deviation of segmentation performance (Dice and IoU) across the various k values tested, along with the corresponding training time. The default setting $m = 5$, $k = 3$ achieves the best balance, producing the lowest variation in Dice (0.88/0.83) and IoU (0.79/0.75) for SAMed and LViT, respectively, while keeping training times reasonable (5.9h/7.6h). Increasing m and k beyond this point results in minor improvements in stability but significantly higher computational costs. These results confirm that a moderate number of diverse yet semantically consistent counterfactuals ensure stable performance without excessive overhead. The trend remains consistent across both model architectures.