DUETMOE: COUPLING INTER- AND INTRA-SUBGROUP ROBUSTNESS FOR FAIR MEDICAL IMAGE ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

As medical AI expands across diverse healthcare settings worldwide, equitable performance across patient populations is becoming essential to trustworthy clinical use. Fairness in medical image analysis is often evaluated through average performance across predefined subgroups, yet similar subgroup averages can conceal substantial variation among individual patients. Therefore, a reliable medical AI requires addressing two complementary objectives: *inter-subgroup fairness*, which reduces performance disparities across groups, and *intra-subgroup robustness*, which protects poorly served patients within each group. To jointly address these objectives, we propose **DuetMoE**, a subgroup-aware mixture-of-experts framework that couples group-level adaptation with patient-specific clinical guidance, enabling more reliable medical image analysis for individual patients. For settings without linked clinical records, we further introduce **DuetMoE+**, which retains subgroup-routed experts and incorporates a KL-constrained distributionally robust objective to address intra-subgroup robustness. We evaluate our methods on PI-CAI, radiotherapy, and Harvard-FairSeg. Across four evaluation settings, our methods lead in overall or equity-scaled performance, reduce the inter-subgroup mean Dice gap by up to 30.7%, and raise intra-subgroup 25th-percentile Dice by up to 11.5 points over the strongest reported baselines.

1 INTRODUCTION

Medical image analysis supports disease detection and the delineation of structures used in assessment and treatment planning, from prostate MRI to optic structures in glaucoma evaluation and targets in radiotherapy (Saha et al., 2024; Tian et al., 2024a; Choi et al., 2020). Missed lesions or inaccurate boundaries can change the assessment of disease or the extent of a planned treatment target. Reliable predictions therefore matter at the level of the patient whose care they inform. Yet patients within the same institution, tumor-stage group, or demographic group can differ in anatomy, disease extent, and clinical history, and a strong subgroup average may still conceal poor predictions for some of them. This motivates two complementary goals: inter-subgroup fairness, addressing performance disparities across groups, and intra-subgroup fairness, addressing patient-level reliability within each group. Progress on one does not establish progress on the other. Addressing both requires learning subgroup commonalities while accounting for individual differences (Liu et al., 2023).

Clinical records available at the relevant care stage may include age, laboratory measurements such as organ-specific antigen levels, tumor stage, and surgical history (Saha et al., 2024). Multimodal artificial intelligence (AI) integrates this context with imaging to support patient-specific segmentation and downstream decisions that better reflect clinical reasoning. Recent approaches use large language models (LLMs) to encode heterogeneous clinical information alongside visual representations (Xiao et al., 2026). Segmentation models align LLM-derived clinical representations with image features (Oh et al., 2024a; Rajendran et al., 2025; Huemann et al., 2026), while MoME adds center-specific expert routing to address institutional bias (Oh et al., 2024b). Although these methods personalize predictions using clinical context, their architectures do not explicitly tailor computational pathways or clinical guidance to patient-level differences within the same subgroup.

Patients within the same subgroup can differ in both their clinical circumstances and the image features that reach a shared expert. The useful clinical guidance can therefore differ even when the router and expert are shared. To address both levels, we retain subgroup-specific routing while letting

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

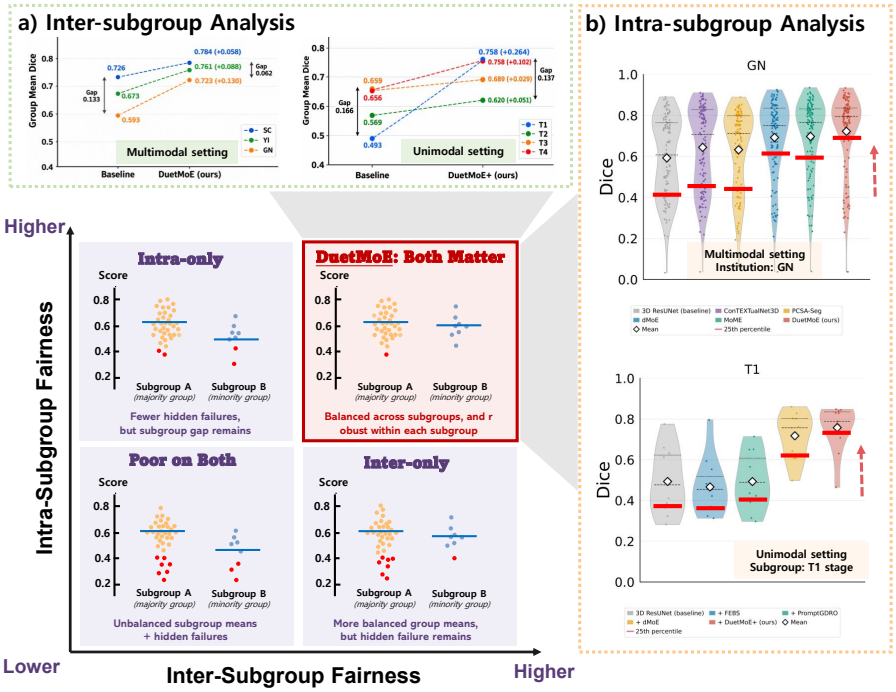


Figure 1: **Overview**. The central schematic depicts inter-subgroup balance and intra-subgroup patient outcomes as complementary goals. The surrounding panels show results from our proposed methods: (a) at the *inter-subgroup level*, subgroup mean performance improves and becomes more balanced; and (b) at the *intra-subgroup level*, the 25th percentile of patient-level Dice increases.

each patient's clinical context shape expert selection and the guidance each expert receives after routing. This motivates us to propose **DuetMoE**. After routing, each expert retrieves *expert-specific clinical guidance* using the features it actually receives. By matching this guidance to each patient's circumstances and assigned features, the design targets intra-subgroup robustness.

Beyond these information-rich settings, some clinical workflows provide only a single task modality without linked clinical records. However, reliable predictions are still needed both across subgroups and among patients within each group. When subgroup labels are available but linked clinical records are not, we introduce **DuetMoE+**, a unimodal variant that retains subgroup-routed experts and uses within-subgroup distributionally robust optimization (DRO) to emphasize high-loss patients. Figure 1 summarizes the inter- and intra-subgroup perspective behind **DuetMoE** and illustrates our methods' performance on both axes. Our contributions are threefold:

- **DuetMoE: inter-subgroup routing and intra-subgroup clinical guidance.** We use subgroup-specific routers to adapt across groups and embed patient-specific clinical memory before and after routing, tailoring expert selection and matching each selected expert's guidance to its assigned image features.
- **DuetMoE+: inter-subgroup routing and intra-subgroup robust training.** Without linked clinical records, the same subgroup label selects the router and defines a within-subgroup KL-robust objective that gives high-loss patients greater training influence.
- **Computational results on both fairness axes.** We conduct extensive comparisons on two public benchmarks, PI-CAI and Harvard-FairSeg, and an in-house radiotherapy dataset, evaluating both subgroup performance and patient-level outcomes. Across four settings, our methods lead in overall or equity-scaled performance. For inter-subgroup balance, our methods reduce the inter-institution mean Dice gap by up to 30.7% relative to the best prior model in overall Dice. For intra-subgroup robustness, our methods improve lowest-10% mean cup Dice by up to 8.4 percentage points over the subgroup-routing baseline.

2 Related Work

Fairness across and within subgroups. Fairness methods in medical segmentation act on data, representations, and training objectives. FairSeg combines demographic subgroup evaluation with Fair Error-Bound Scaling motivated by difficult cases within groups (Tian et al., 2024a); FairDiff improves group representation through synthetic data (Li et al., 2024), and FairDomain incorporates demographic information into cross-domain feature learning (Tian et al., 2024b). Hidden stratification can leave clinically important failures obscured by aggregate performance (Oakden-Rayner et al., 2020). Even when a method addresses difficult cases during training, subgroup means alone do not reveal how outcomes vary among patients in that subgroup. This motivates considering inter-subgroup disparities and intra-subgroup patient variation as complementary evaluation axes.

Clinical context in multimodal medical image analysis. Multimodal segmentation uses clinical information to contextualize image features. In routine clinical workflows, disease-stage information and patient-level variables, such as demographics, biomarkers, and organ-specific measurements, are often available before treatment planning or medical image interpretation. For radiotherapy planning, LLMSeg and ConTEXTualNet3D combine language-model-based clinical encoding with layer-wise image–text alignment (Oh et al., 2024a; Huemann et al., 2026), while Medformer employs vision–language attention and language gating across multiple visual stages (Rajendran et al., 2025). PCSA-Seg maintains language guidance throughout hierarchical decoding using clinical text available at image analysis (Xue et al., 2026). Together, these methods show how clinical information can guide feature processing across scales and stages to better align AI predictions with clinical practice.

Subgroup-aware and context-conditioned experts. In dMoE, a subgroup attribute selects a router that directs image features to shared experts (Oh et al., 2025), while LTSE uses image features and text describing a requested tissue target to select expert decoders for pathology referring segmentation (Tang et al., 2025). Relatedly, the Mixture of Multicenter Experts (MoME) combines clinical fusion with center-specific expert routing (Oh et al., 2024b). In MoME, however, clinical alignment is applied after the expert mixture, rather than retrieving clinical guidance conditioned on the features assigned to each expert. DuetMoE instead integrates clinical context directly into subgroup-aware expert computation.

Intra-subgroup robust risk. When auxiliary clinical context is unavailable but subgroup labels are known, loss-based objectives can give greater training weight to high-loss patients within each group. GroupDRO targets the worst group-average loss (Sagawa et al., 2019), while risk-averse adaptive sampling and positive loss tilting give more influence to difficult examples (Curi et al., 2020; Li et al., 2023). Hierarchical tilted risk and hierarchical ambiguity sets also account for variation both across and within groups (Li et al., 2023; Jo et al., 2026). These general-purpose objectives provide useful tools for intra-subgroup risk control.

3 Method

To address both inter- and intra-subgroup fairness, we propose **DuetMoE**, which combines subgroup-selected routing over shared experts with patient- and region-specific clinical context to guide routing and expert computation. When linked clinical records are unavailable, its unimodal variant **DuetMoE+** retains subgroup routing and adds a within-subgroup robust loss that emphasizes higher-loss patients.

Figure 2 summarizes both settings. For each of N patients, let x_j denote the task input, y_j the segmentation target, $g_j \in \mathcal{G}$ the subgroup label, and c_j the optional clinical context. We write $L_j(\theta)$ for the patient segmentation loss under model parameters θ.

3.1 DuetMoE: adapting experts to groups and patients

A subgroup label alone cannot tell the model which clinical information matters for a particular patient's image. The same clinical variable may matter differently when the model uses a summary of the current feature map or features at individual locations; a shared expert may also receive different feature patterns from patients in that subgroup. Rather than assigning each variable a fixed role, **DuetMoE** lets both image views query the patient's clinical memory before routing and each

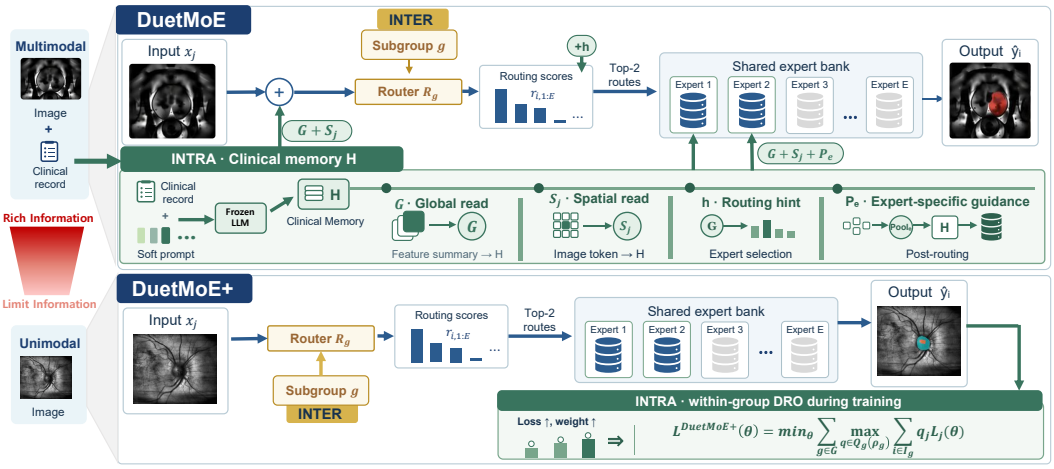


Figure 2: Summary of **DuetMoE** and **DuetMoE+**. Inter-subgroup adaptation: both variants use subgroup-specific routers to select shared experts. Intra-subgroup robustness: **DuetMoE** (top) embeds patient-specific clinical memory before and after routing, using global and spatial reads to guide expert selection and assigned image features to retrieve expert-specific guidance. Without linked clinical records, **DuetMoE+** (bottom) uses within-subgroup KL-DRO to give higher-loss patients greater influence on training their subgroup router and the shared experts.

expert's assigned features query it afterward. This aims to preserve inter-subgroup differences via subgroup-specific routing, while achieving intra-subgroup robustness through patient-conditioned reads that guide expert selection and computation.

Global and spatial clinical context. For patient j, a frozen language model with learnable soft prompts produces clinical memory $H_j = \{h_{jm}\}_{m=1}^{M_j}$ with M_j valid vectors (Lester et al., 2021). At an encoder stage (index omitted), let v_{jt}, $t = 1, \ldots, T$, be the tokens at its T image locations. Global and spatial queries read H_j using a summary of the current feature map and features at individual locations. Define $\hat{v}_{jt} = \mathrm{LN}(v_{jt})$ and $\bar{v}_j = T^{-1} \sum_t \hat{v}_{jt}$ as the normalized tokens and their mean, with learned layer normalization LN over token channels. Let Q and O be the query and output projections. The multi-head attention output $\mathcal{A}(q, H_j)$ uses projected query q and keys and values from H_j, before projection by O. Writing A_j^G for the global attention output before projection, the global context G_j and spatial context S_{jt} are

$$A_j^G = \mathcal{A}(Q(\bar{v}_j), H_j), \quad G_j = O(A_j^G), \quad S_{jt} = O(\mathcal{A}(Q(\hat{v}_{jt}), H_j)). \tag{1}$$

Routing with clinical context. The subgroup label g_j selects the router R_{g_j}, which scores E shared experts using image tokens augmented with global and spatial context. A separate linear projection B, shared across subgroups, maps the global attention output A_j^G to a routing hint for each expert: u_{jt} is the augmented router input, $r_{jt} \in \mathbb{R}^E$ the expert-score vector, and $\pi_{jt} \in \mathbb{R}^E$ the resulting sparse weights.

$$u_{jt} = v_{jt} + \alpha G_j + \beta S_{jt}, \quad r_{jt} = R_{g_j}(u_{jt}) + \lambda B(A_j^G) + \xi_{jt}, \quad \pi_{jt} = \mathrm{SoftmaxTopK}(r_{jt}). \tag{2}$$

During training and evaluation, $\xi_{jt} \in \mathbb{R}^E$ is the score-noise vector in noisy top-K gating (Shazeer et al., 2017). SoftmaxTopK softmax-normalizes the K largest scores (expert set $\mathcal{K}_{jt}$) and sets the rest to zero. The component $\pi_{jt,e}$ is the weight of expert e.

Expert-specific clinical guidance after routing. After routing, we summarize the normalized image features assigned to each expert and use them to read the clinical memory again. Here $z_{j,e}$ is the routing-weighted summary and $P_{j,e}$ the guidance read for patient j and expert e. For each patient–expert pair satisfying $\sum_t \pi_{jt,e} > 0$, we define

$$z_{j,e} = \frac{\sum_t \pi_{jt,e} \hat{v}_{jt}}{\sum_t \pi_{jt,e}}, \quad P_{j,e} = O(\mathcal{A}(Q(z_{j,e}) + b_e, H_j)), \tag{3}$$

where b_e is a learnable query offset for expert e. The summary $z_{j,e}$ is computed separately for each patient, so the clinical guidance $P_{j,e}$ can differ between patients who share a router and expert. The selected experts E_e process tokens augmented with G_j, S_{jt}, and $P_{j,e}$; $\widetilde{v}_{jt,e}$ denotes that expert's input and v_{jt}^{out} the updated token. Their outputs form a weighted residual update:

$$\widetilde{v}_{jt,e} = v_{jt} + \gamma G_j + \delta S_{jt} + \varepsilon P_{j,e}, \quad v_{jt}^{\text{out}} = v_{jt} + \sum_{e \in \mathcal{K}_{jt}} \pi_{jt,e} E_e(\widetilde{v}_{jt,e}) \,. \tag{4}$$

Architecture and training. Only the last four of five encoder stages use clinical context. The six coefficients $(\alpha, \beta, \lambda, \gamma, \delta, \varepsilon)$ are learned separately at each clinical-context stage and shared across patients. More implementation details are provided in Appendix A.1.

3.2 DuetMoE+: inter-subgroup routing and intra-subgroup robust risk

Clinical records may be unavailable or unlinked even when images and subgroup labels are available, as in fundus datasets with demographic labels or radiotherapy archives with planning CT scans and tumor-stage labels. These settings lack the patient-specific clinical memory needed for **DuetMoE**'s context-guided routing and expert updates. To address both subgroup disparities and unreliable predictions within groups, we introduce **DuetMoE+**, a variant that retains subgroup-selected expert routing and targets intra-subgroup fairness through a robust intra-subgroup loss.

Architecture-level inter-subgroup adaptation. We construct **DuetMoE+** from **DuetMoE** by first removing the clinical pathways used before and after routing: the global and spatial reads, the clinical routing hint, and the expert-specific read. We keep the subgroup-selected routers and shared experts, so the available subgroup label can still shape how image features are processed. In such unimodal setting, the inter-subgroup architecture of **DuetMoE+** reduces to dMoE, as shown in Proposition 1, with detailed proof in Appendix A.2.

Proposition 1 (Inter-subgroup architecture reduces to dMoE (Oh et al., 2025))**.** *At an encoder stage, disabling all clinical-memory reads and the memory-derived routing hint in DuetMoE recovers the dMoE update for every patient and image token, with the subgroup routers, shared experts, and gating rule held fixed.*

Loss-level intra-subgroup robust risk. A subgroup mean can hide high-loss patients. **DuetMoE+** uses within-subgroup KL-DRO to give these patients greater training influence on the subgroup routers and shared experts. Let $I_g = \{j : g_j = g\}$ and $\widehat{P}_g$ be the uniform empirical distribution on I_g. For $\rho_g \geq 0$, define $\mathcal{U}_g^{\text{KL}}(\widehat{P}_g) = \{Q_g : D_{\text{KL}}(Q_g \| \widehat{P}_g) \leq \rho_g\}$, where Q_g ranges over distributions on I_g. We train by

$$\min_{\theta} \; \mathcal{L}_{\text{DuetMoE+}}(\theta) = \min_{\theta} \sum_{g \in \mathcal{G}} \widehat{R}_g^{\text{rob}}(\theta), \quad \widehat{R}_g^{\text{rob}}(\theta) = \sup_{Q_g \in \mathcal{U}_g^{\text{KL}}(\widehat{P}_g)} \mathbb{E}_{j \sim Q_g}[L_j(\theta)] \,. \tag{5}$$

The KL radius controls departures from uniform patient weights, allowing greater emphasis on high-loss observed patients. When $\rho_g = 0$ for all groups, the objective reduces to the sum of subgroup mean losses. The next proposition relates this robust risk to within-subgroup loss variation. The following result shows how the robust risk responds to variation among patients within a subgroup.

Proposition 2 (Sensitivity to intra-subgroup variation)**.** *Fix a nonempty subgroup g with finite patient losses. Let $\widehat{R}_g^{\text{mean}}(\theta) = |I_g|^{-1} \sum_{j \in I_g} L_j(\theta)$ and $\sigma_g(\theta)$ be their mean and standard deviation. Replacing each robust risk in Eq.* (5) *by its mean defines the mean-loss (ERM) version, recovered when all $\rho_g = 0$. For fixed θ, as $\rho_g \downarrow 0$,*

$$\widehat{R}_g^{\text{rob}}(\theta) = \underbrace{\widehat{R}_g^{\text{mean}}(\theta)}_{\textit{mean-loss version}} + \underbrace{\sqrt{2\rho_g}\,\sigma_g(\theta)}_{\textit{intra-subgroup variation penalty}} + O(\rho_g). \tag{6}$$

If $\sigma_g(\theta) = 0$, robust and mean risks coincide for every $\rho_g \geq 0$.

A low subgroup mean loss can hide large errors on individual patients. Proposition 2 shows that, for small ρ_g, the robust loss approximately equals the mean loss plus a penalty proportional to the standard deviation of patient losses, making intra-subgroup variation explicit in the training objective. The KL weights give high-loss patients greater training influence within their subgroup; Appendix A.2 provides the proof and weight derivation.

4 EXPERIMENTS

Our evaluation follows the information available to the predictor. We evaluate **DuetMoE** on PI-CAI and institution-grouped radiotherapy (RT), where images are linked to clinical records, and its unimodal variant **DuetMoE+** on T-stage-grouped RT and Harvard-FairSeg, where images and subgroup labels are available without additional clinical context. Subgroup results assess inter-subgroup performance, while available patient-level results examine intra-subgroup reliability. The two RT protocols use different cohorts and subgroup definitions and are analyzed separately.

4.1 DATASETS, BASELINES, AND EVALUATION METRICS

PI-CAI: multimodal prostate imaging. PI-CAI (Saha et al., 2024) pairs prostate MRI with patient information, including age, prostate-specific antigen (PSA), PSA density, and prostate volume, for clinically significant prostate cancer (csPCa) detection. We use institution (RUMC, ZGT, and PCNN) as the subgroup label. We hold out one fold of a patient-level five-fold split of the $1{,}500$ public cases, stratified by institution and csPCa status, as the test set (301 cases: RUMC 160, ZGT 70, PCNN 71). We split the remaining $1{,}199$ cases at the patient level into 960 training and 239 validation cases.

Radiotherapy: multimodal & unimodal prostate target volume delineation. The RT data consist of pelvic CT scans and clinician-drawn clinical target volume (CTV) contours from three institutions, SC, YI, and GN. In the multimodal protocol, institution selects the router and linked records provide patient context, including age, PSA, grade, T/N stage, treatment information, and a de-identified physician code; the held-out set contains 420 cases (SC 189, YI 112, GN 119). The unimodal protocol uses CT and T stage for routing without a linked clinical-memory stream; its held-out set contains 275 cases (T1 11, T2 129, T3 114, T4 21).

Harvard-FairSeg: unimodal retinal imaging. Harvard-FairSeg contains $10{,}000$ scanning laser ophthalmoscopy (SLO) fundus images with optic-cup and neuroretinal-rim masks used in glaucoma assessment. Among its six annotated demographic attributes, we use race ($\{\mathrm{Asian}, \mathrm{Black}, \mathrm{White}\}$) as the subgroup label, motivated by the benchmark's reported racial segmentation disparities and elevated glaucoma risk for Black patients (Tian et al., 2024a). We evaluate both targets on the official $2{,}000$-image test split.

Baseline methods. For multimodal PI-CAI and RT, we compare **DuetMoE** with 3D ResUNet, dMoE (Oh et al., 2025), MoME (Oh et al., 2024b), PCSA-Seg (Xue et al., 2026), and a ConTEXTualNet3D-based baseline adapted to a 3D residual U-Net backbone (Huemann et al., 2026). For unimodal RT and FairSeg, dMoE isolates subgroup routing; MoE, FEBS, GDRO, and Prompt-GDRO provide further controls (Shazeer et al., 2017; Tian et al., 2024a; Sagawa et al., 2019; Panaganti et al., 2026). FairSeg also includes published TransUNet, ADV, FEBS, and FairDiff results (Chen et al., 2021; Madras et al., 2018; Tian et al., 2024a; Li et al., 2024). Complete implementation details and dataset cohorts are detailed in Appendix B.

Evaluation metrics. For PI-CAI, the PI score averages case-level AUROC and lesion-level average precision (Saha et al., 2024). For RT and Harvard-FairSeg, Dice measures overlap between predicted and reference masks. We also report equity-scaled scores (ES-PI and ES-Dice), which adjust the overall scores for differences in subgroup performance (Tian et al., 2024a); metric definitions are in Appendix B.2. For intra-subgroup evaluation, we report sensitivity at 90% specificity on PI-CAI, patient-level Dice distributions and 25th-percentile Dice on RT, and the lowest-10% mean Dice within each racial subgroup on FairSeg.

Statistical analysis. To account for small and imbalanced subgroups, we estimate 95% confidence intervals (CIs) from $1{,}000$ within-subgroup bootstrap resamples for methods with case-level predictions, including reproduced baselines.

4.2 PI-CAI: MULTIMODAL PROSTATE IMAGING

Table 1 compares the institution-grouped multimodal models on the PI-CAI test split. For inter-subgroup performance, **DuetMoE** raises ES-PI from 0.453 with dMoE to 0.500 and improves the

lowest-scoring institution, PCNN, from 0.423 to 0.457. It also has the highest overall PI score (0.623) and PI score in RUMC (0.688). For intra-subgroup detection reliability, **DuetMoE** increases sensitivity at 90% specificity within RUMC from 0.553 with MoME to 0.681 (+12.8 percentage points). It also achieves the highest normalized partial AUROC over false-positive rates up to 0.2 in RUMC (0.564) and ZGT (0.670), supporting stronger detection performance at high specificity within these institutions. More institution-specific results are provided in Tables 5 and 6 in Appendix C.1.

Table 1: **PI-CAI.** Comparison on **multimodal** csPCa detection using institution as the subgroup attribute.

Method	All ($n = 301$)		RUMC ($n = 160$)	ZGT ($n = 70$)	PCNN ($n = 71$)
	ES-PI (CIs)	PI score (CIs)	PI score (CIs)	PI score (CIs)	PI score (CIs)
3D ResUNet	0.466 (0.379–0.535)	0.585 (0.505–0.674)	0.618 (0.523–0.732)	<u>0.667</u> (0.473–0.867)	0.445 (0.269–0.626)
ConTEXTualNet3D-based	0.421 (0.342–0.455)	0.486 (0.411–0.563)	0.533 (0.437–0.658)	0.496 (0.357–0.677)	0.390 (0.277–0.532)
PCSA-Seg	0.432 (0.366–0.484)	0.556 (0.487–0.643)	<u>0.670</u> (0.580–0.760)	0.537 (0.348–0.746)	0.401 (0.302–0.553)
dMoE	0.453 (0.371–0.526)	0.590 (0.511–0.682)	0.647 (0.554–0.768)	**0.671** (0.472–0.870)	0.423 (0.271–0.596)
MoME	**0.510** (0.414–0.564)	<u>0.617</u> (0.542–0.692)	0.669 (0.575–0.777)	0.641 (0.454–0.827)	**0.484** (0.331–0.635)
DuetMoE	<u>0.500</u> (0.408–0.551)	**0.623** (0.543–0.705)	**0.688** (0.588–0.787)	0.637 (0.451–0.839)	<u>0.457</u> (0.310–0.616)

Bold and underlined point estimates indicate the highest and second-highest values in each column, respectively.

4.3 RADIOTHERAPY: MULTIMODAL & UNIMODAL PROSTATE TARGET VOLUME DELINEATION

Multimodal RT: institution. Table 2 shows improved inter-subgroup performance for **DuetMoE**, with the highest ES-Dice (0.716, versus 0.686 for MoME) and overall Dice (0.761). The gain is particularly relevant at GN, the lowest-scoring institution for every method: its Dice rises to 0.723, from 0.693 with dMoE and 0.698 with MoME. The gap between the highest- and lowest-scoring institutions narrows from 0.088 for MoME to 0.061 for **DuetMoE**.

Table 2: **Radiotherapy.** Comparison on **multimodal** 3D radiotherapy target segmentation using institution as the subgroup attribute.

Method	All ($n = 420$)		SC ($n = 189$)	YI ($n = 112$)	GN ($n = 119$)
	ES-Dice (CIs)	Dice (CIs)	Dice (CIs)	Dice (CIs)	Dice (CIs)
3D ResUNet	0.594 (0.566–0.614)	0.674 (0.657–0.691)	0.726 (0.701–0.751)	0.673 (0.645–0.697)	0.593 (0.557–0.624)
ConTEXTualNet3D-based	0.640 (0.612–0.659)	0.729 (0.712–0.746)	0.781 (0.759–0.801)	0.730 (0.702–0.756)	0.644 (0.607–0.680)
PCSA-Seg	0.627 (0.593–0.654)	0.682 (0.665–0.699)	0.701 (0.674–0.729)	0.702 (0.670–0.733)	0.633 (0.596–0.666)
dMoE	0.680 (0.647–0.704)	0.747 (0.731–0.762)	<u>0.784</u> (0.762–0.804)	0.740 (0.710–0.767)	0.693 (0.656–0.724)
MoME	<u>0.686</u> (0.659–0.712)	<u>0.757</u> (0.743–0.771)	**0.786** (0.765–0.804)	**0.772** (0.744–0.799)	<u>0.698</u> (0.663–0.731)
DuetMoE	**0.716** (0.679–0.737)	**0.761** (0.745–0.776)	<u>0.784</u> (0.761–0.804)	<u>0.761</u> (0.731–0.787)	**0.723** (0.687–0.755)

Bold and underlined point estimates indicate the highest and second-highest values in each column, respectively; the tied second-highest SC scores are both underlined.

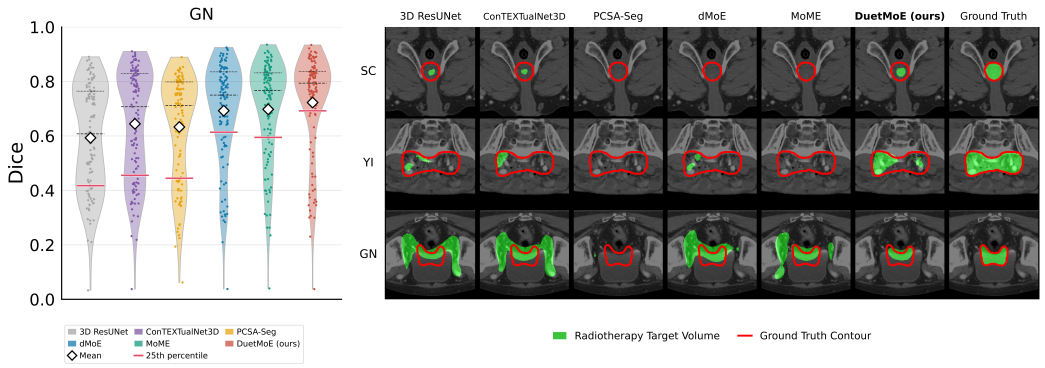


Figure 3: **Multimodal radiotherapy: intra-subgroup performance. Left:** patient-level Dice distributions in GN; violin width reflects estimated density, points represent patients, and red horizontal lines mark the 25th percentiles. **Right:** selected target delineations from SC, YI, and GN for methods; green masks show target volumes and red contours indicate ground-truth boundaries.

Figure 3 (left) shows improved intra-subgroup performance in GN: **DuetMoE** raises 25th-percentile Dice from 0.614 with dMoE to 0.692 (+7.8 percentage points), improving the lower end of the patient distribution. Violin width reflects the estimated density of patient Dice; **DuetMoE**'s thinner low-Dice tail suggests a lower concentration of poor predictions. The selected contours (right) show **DuetMoE**'s more complete target coverage in SC and YI and reduced extension into surrounding tissue in GN.

Unimodal RT: tumor stage. Table 3 reports the image-only RT comparison across T1–T4. **DuetMoE+** obtains the highest overall Dice (0.665) and ES-Dice (0.530). Compared with dMoE (Oh et al., 2025), it improves mean Dice in T2, the lowest-scoring stage, from 0.586 to 0.621 (+3.5 percentage points). ES-Dice rises from 0.499 to 0.530, supporting improved inter-subgroup performance. Additional evaluation metrics show a similar pattern, see Appendix C.2 for details.

Table 3: Comparison on **unimodal** 3D radiotherapy target segmentation using tumor stage as the subgroup attribute.

Method	All ($n = 275$)		T1 ($n = 11$)	T2 ($n = 129$)	T3 ($n = 114$)	T4 ($n = 21$)
	ES-Dice (95% CI)	Dice (95% CI)	Dice (95% CI)	Dice (95% CI)	Dice (95% CI)	Dice (95% CI)
3D ResUNet	0.487 (0.447–0.529)	0.610 (0.589–0.630)	0.492 (0.401–0.579)	0.570 (0.539–0.602)	0.659 (0.626–0.688)	0.656 (0.586–0.720)
+ FEBS	0.434 (0.406–0.467)	0.586 (0.567–0.604)	0.466 (0.387–0.544)	0.522 (0.495–0.549)	0.652 (0.621–0.677)	0.688 (0.651–0.721)
+ MoE	0.452 (0.415–0.492)	0.608 (0.586–0.628)	0.492 (0.395–0.590)	0.543 (0.512–0.573)	0.674 (0.640–0.704)	0.708 (0.663–0.750)
+ GDRO	0.454 (0.421–0.491)	0.607 (0.586–0.627)	0.481 (0.400–0.560)	0.546 (0.516–0.576)	0.671 (0.639–0.700)	0.694 (0.637–0.745)
+ Prompt-GDRO	0.476 (0.440–0.517)	0.612 (0.592–0.631)	0.492 (0.409–0.568)	0.562 (0.532–0.593)	0.669 (0.636–0.697)	0.673 (0.616–0.726)
+ dMoE	0.499 (0.469–0.531)	0.650 (0.628–0.671)	0.718 (0.654–0.772)	0.586 (0.556–0.618)	**0.693 (0.657–0.723)**	**0.778 (0.747–0.803)**
+ **DuetMoE+**	**0.530 (0.499–0.569)**	**0.665 (0.643–0.683)**	**0.757 (0.687–0.812)**	**0.621 (0.590–0.654)**	0.688 (0.654–0.717)	0.758 (0.729–0.784)

Bold marks the highest score, and underlining marks the second-highest score in each column.

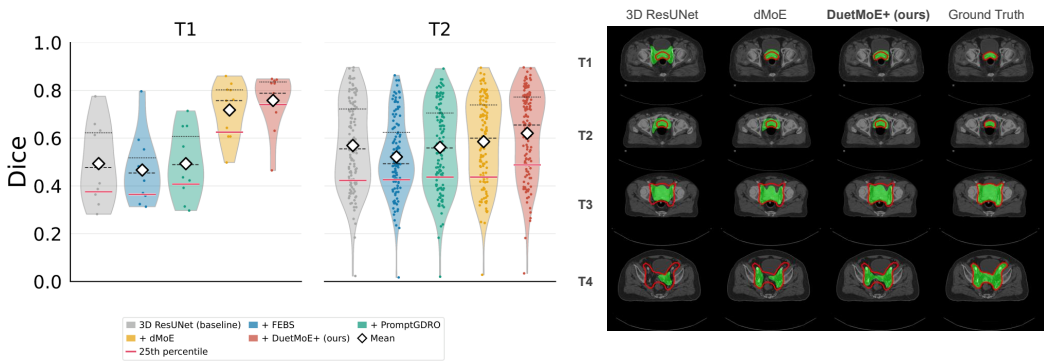


Figure 4: **Unimodal radiotherapy: intra-subgroup performance. Left:** patient-level Dice distributions within T1 and T2. Violin width reflects estimated density, points represent patients, and red horizontal lines mark the 25th percentiles. **Right:** representative target delineations for T1–T4, comparing **DuetMoE+** with the baselines and ground truth. Green masks show target volumes, and red contours indicate ground-truth boundaries.

Figure 4 provides the complementary intra-subgroup view. Relative to dMoE, the T1 and T2 Dice distributions for **DuetMoE+** shift toward higher values, with 25th-percentile gains of 11.5 percentage points (0.625 to 0.740) and 5.1 points (0.437 to 0.488), respectively. In T2, the thinner lower tail also indicates less estimated density at low Dice. Since dMoE already uses subgroup-selected routing, these lower-tail gains support the contribution of robust training within each stage. The selected T1 and T2 contours show predictions that follow the annotated targets more closely.

4.4 HARVARD-FAIRSEG: UNIMODAL RETINAL IMAGING

Table 4 reports cup-segmentation results on Harvard-FairSeg with race as the subgroup attribute. For inter-subgroup performance, **DuetMoE+** obtains the highest ES-Dice (0.843), up from 0.832 with dMoE. It also raises mean Dice in the weakest subgroup, Asian, from 0.844 to 0.852 (+0.8 percentage points), with an overall Dice of 0.865. For intra-subgroup performance, Figure 7 shows

that **DuetMoE+** increases mean cup Dice among the lowest-scoring 10% of patients by 6.5, 8.4, and 1.4 percentage points over dMoE in the Asian, Black, and White subgroups, respectively. Rim segmentation also shows lower-tail gains in the Asian and Black subgroups; more results and additional evaluation metrics are reported in Appendix C.3.

Table 4: Cup Segmentation Comparison on the **unimodal** 2D Harvard-FairSeg dataset using race as the subgroup attribute.

Method	All ($n = 2000$)		Asian ($n = 169$)	Black ($n = 299$)	White ($n = 1532$)
	ES-Dice (CIs)	Dice (CIs)	Dice	Dice	Dice
TransUNet	0.828	0.848	0.827	0.849	0.850
+ ADV	0.826	0.841	0.825	0.842	0.843
+ FEBS	0.825	0.846	0.825	0.848	0.848
+ FairDiff	0.833	0.848	0.832	0.848	0.850
+ MoE	0.830 (0.809–0.847)	0.854 (0.849–0.860)	0.845	0.842	0.857
+ GDRO	0.840 (0.818–0.857)	0.860 (0.855–0.865)	<u>0.850</u>	0.851	0.863
+ Prompt-GDRO	<u>0.842</u> (0.819–0.861)	**0.866** (0.861–0.871)	0.849	**0.859**	**0.870**
+ dMoE	0.832 (0.810–0.853)	0.862 (0.856–0.867)	0.844	0.851	0.866
+ **DuetMoE+**	**0.843** (0.822–0.861)	<u>0.865</u> (0.860–0.869)	**0.852**	<u>0.856</u>	<u>0.868</u>

Bold and underlined point estimates indicate the highest and second-highest values within each target and column, respectively.

5 Discussion & Conclusion

Our results support treating inter-subgroup balance and intra-subgroup robustness as complementary goals for medical image analysis. Across four evaluation settings, **DuetMoE** and **DuetMoE+** achieve leading overall or equity-scaled performance. On PI-CAI, **DuetMoE** achieves the highest overall PI score (0.623) and improves sensitivity at 90% specificity within RUMC by 10.7 percentage points over the strongest baseline for this metric. In multimodal RT, **DuetMoE** narrows the inter-institution mean group Dice gap by 30.7% relative to the best prior model in overall Dice. For unimodal setting, in image-only RT, **DuetMoE+** improves worst-subgroup mean Dice by 3.5 percentage points and within-subgroup 25th-percentile Dice by up to 11.5 points over the subgroup-routing baseline. On Harvard-FairSeg, it also improves lowest-10% mean cup Dice by up to 8.4 percentage points over the same baseline. These within-subgroup gains demonstrate the value of evaluating patients within each subgroup alongside aggregate performance.

These results support embedding patient-level adaptation within subgroup-aware learning. **DuetMoE** embeds clinical memory before and after routing to guide expert selection and feature processing. Without linked clinical records, **DuetMoE+** aligns within-subgroup robust weighting with the subgroup routers and shared experts that process each patient's images. The PI-CAI clinical-context comparisons favor designs with expert-specific guidance (Appendix D.1), while the unimodal component ablation shows that the full **DuetMoE+** achieves the highest equity-scaled and worst-subgroup scores (Appendix D.2). The radius study in Appendix D.3 further shows that settings with similar overall Dice can yield different subgroup lower-tail outcomes. For **DuetMoE+**, radius selection should therefore assess equity-scaled performance and subgroup lower tails alongside overall Dice on a held-out validation set.

In particular, both methods require subgroup labels at inference, and **DuetMoE** additionally relies on linked clinical records. Their sensitivity to missing or inaccurate subgroup metadata and incomplete clinical context remains to be evaluated. These observations motivate extending the framework to incomplete information and developing validation-based radius selection that considers both subgroup performance and patient-level lower tails. A further direction is to combine loss-level robustness training with clinically guided **DuetMoE**, allowing patient-specific computation and loss-based emphasis to act together. Since both mechanisms influence learning within a subgroup, their interaction could either reinforce useful adaptation or place excessive emphasis on noisy or persistently difficult cases. A matched multimodal ablation is needed to test whether the two mechanisms complement or conflict with each other. Overall, our current findings support models that adapt across subgroups while remaining responsive to patient differences within each group.

AI USE STATEMENT

Generative AI tools, including OpenAI Codex and Claude, were used to assist with literature search, refinement of mathematical formulations and experimental protocols, implementation and debugging of research code, figure preparation, and manuscript drafting and editing. AI-assisted suggestions were assessed against cited sources, mathematical derivations, and available code and experiment records; implementation changes were checked with targeted tests where applicable. The authors are responsible for verifying the final text, claims, proofs, code, and reported results, and take full responsibility for all AI-assisted content. The frozen language model used to construct clinical memory is a component of the proposed method, described in Section 3.1 and Appendix A.1.

ETHICS STATEMENT

This study analyzes retrospective medical imaging and associated clinical information. Acquisition of the in-house radiotherapy cohort was approved by the institutional review boards of the three contributing hospitals (SC: 4-2023-0179; YI: 9-2023-0161; GN: 3-2023-0396). Patient records are de-identified, and the in-house cohort is not publicly released because of institutional and IRB restrictions. Data access and permitted use for PI-CAI and Harvard-FairSeg follow the benchmark providers' requirements. Institution, tumor stage, and race labels are used to evaluate performance disparities. The reported subgroup and lower-tail metrics do not establish clinical benefit or deployment safety; clinical use would require further validation and clinician oversight.

REPRODUCIBILITY STATEMENT

Section 3 specifies the models and training objectives. Appendices A.1 and A.2 provide clinical-memory details and robust-loss derivations. Section 4.1 and Appendix B document dataset cohorts, split information, evaluation metrics, training settings, and checkpoint rules. Additional evaluation and ablation protocols are described alongside the corresponding analyses. PI-CAI and Harvard-FairSeg are public benchmarks, while access to the in-house radiotherapy data remains subject to institutional approval.

REFERENCES

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

Min Seo Choi, Byeong Su Choi, Seung Yeun Chung, Nalee Kim, Jaehee Chun, Yong Bae Kim, Jee Suk Chang, and Jin Sung Kim. Clinical evaluation of atlas- and deep learning-based automatic segmentation of multiple organs and clinical target volumes for breast cancer. *Radiotherapy and Oncology*, 153:139–145, 2020.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention*, pp. 424–432. Springer, 2016.

Sebastian Curi, Kfir Y Levy, Stefanie Jegelka, and Andreas Krause. Adaptive sampling for stochastic risk-averse learning. *Advances in Neural Information Processing Systems*, 33:1036–1047, 2020. doi: 10.48550/arXiv.1910.12511.

Zachary Huemann, Samuel Church, Joshua D Warner, Daniel Tran, Xin Tie, Alan B McMillan, Junjie Hu, Steve Y Cho, Meghan G Lubner, and Tyler J Bradshaw. Contextual net 3d: Vision-language modeling in pet/ct for visual grounding of positive findings. *Journal of Imaging Informatics in Medicine*, 2026. doi: 10.1007/s10278-026-01879-2.

Sungho Jo, Seonghwi Kim, and Minwoo Chae. Mitigating spurious correlation via distributionally robust learning with hierarchical ambiguity sets. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 121167–121191, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/c4f85da01632a2211c785f482cb3e043-Paper-Conference.pdf.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.243.

Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. On tilted losses in machine learning: Theory and applications. *Journal of Machine Learning Research*, 24(142):1–79, 2023. doi: 10.48550/arXiv.2109.06141. URL https://jmlr.org/papers/v24/21-1095.html.

Wenyi Li, Haoran Xu, Guiyu Zhang, Huan-ang Gao, Mingju Gao, Mengyu Wang, and Hao Zhao. Fairdiff: Fair segmentation with point-image diffusion. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 617–628. Springer, 2024. doi: 10.1007/978-3-031-72384-1_58.

Mingxuan Liu, Yilin Ning, Salinelat Teixayavong, Mayli Mertens, Jie Xu, Daniel Shu Wei Ting, Lionel Tim-Ee Cheng, Jasmine Chiat Ling Ong, Zhen Ling Teo, Ting Fang Tan, et al. A translational perspective towards clinical ai fairness. *NPJ digital medicine*, 6(1):172, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pp. 3384–3393. PMLR, 2018.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Re. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, 2020. doi: 10.1145/3368555.3384468.

Yujin Oh, Sangjoon Park, Hwa Kyung Byun, Yeona Cho, Ik Jae Lee, Jin Sung Kim, and Jong Chul Ye. Llm-driven multimodal target volume contouring in radiation oncology. *Nature Communications*, 15(1):9186, 2024a. doi: 10.1038/s41467-024-53387-y.

Yujin Oh, Sangjoon Park, Xiang Li, Pengfei Jin, Yi Wang, Jonathan Paly, Jason Efstathiou, Annie Chan, Jun Won Kim, Hwa Kyung Byun, Ik Jae Lee, Jaeho Cho, Chan Woo Wee, Peng Shu, Peilong Wang, Caiwen Jiang, Nathan Yu, Jason Holmes, Jong Chul Ye, Quanzheng Li, Wei Liu, Woong Sub Koom, Jin Sung Kim, and Kyungsang Kim. Mixture of multicenter experts in multimodal ai for debiased radiotherapy target delineation. *arXiv preprint arXiv:2410.00046*, 2024b. doi: 10.48550/arXiv.2410.00046.

Yujin Oh, Pengfei Jin, Sangjoon Park, Sekeun Kim, Siyeop Yoon, Jin Sung Kim, Kyungsang Kim, Xiang Li, and Quanzheng Li. Distribution-aware fairness learning in medical image segmentation from a control-theoretic perspective. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 47003–47016. PMLR, 2025. doi: 10.48550/arXiv.2502.00619. URL https://proceedings.mlr.press/v267/oh25c.html.

Kishan Panaganti, Zhenwen Liang, Wenhao Yu, Haitao Mi, and Dong Yu. Group distributionally robust optimization-driven reinforcement learning for llm reasoning, 2026. URL https://arxiv.org/abs/2601.19280.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Praveenbalaji Rajendran, Yizheng Chen, Liang Qiu, Thomas Niedermayr, Wu Liu, Mark Buyyounouski, Hilary Bagshaw, Bin Han, Yong Yang, Nataliya Kovalchuk, Xuejun Gu, Steven Hancock, Lei Xing, and Xianjin Dai. Auto-delineation of treatment target volume for radiation therapy using large language model-aided multimodal learning. *International Journal of Radiation Oncology, Biology, Physics*, 2025. doi: 10.1016/j.ijrobp.2024.07.2149.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. doi: 10.48550/arXiv.1911.08731.

Anindo Saha, Joeran S. Bosma, Jasper J. Twilt, Bram van Ginneken, Anders Bjartell, Anwar R. Padhani, et al. Artificial intelligence and radiologists in prostate cancer detection on mri (pi-cai): an international, paired, non-inferiority, confirmatory study. *The Lancet Oncology*, 25(7):879–887, 2024. doi: 10.1016/S1470-2045(24)00220-1.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Jiao Tang, Bo Qian, Peng Wan, Wei Shao, and Daoqiang Zhang. LTSE: Language-guided tissue referring segmentation in pathology images with adaptive expert mixture. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15965 of *Lecture Notes in Computer Science*, pp. 405–415, 2025. doi: 10.1007/978-3-032-04978-0_39.

Yu Tian, Min Shi, Yan Luo, Ava Kouhana, Tobias Elze, and Mengyu Wang. Fairseg: A large-scale medical image segmentation dataset for fairness learning using segment anything model with fair error-bound scaling. In *The Twelfth International Conference on Learning Representations*, 2024a. doi: 10.48550/arXiv.2311.02189.

Yu Tian, Congcong Wen, Min Shi, Muhammad Muneeb Afzal, Hao Huang, Muhammad Osama Khan, Yan Luo, Yi Fang, and Mengyu Wang. Fairdomain: Achieving fairness in cross-domain medical image segmentation and classification. In *European Conference on Computer Vision*, pp. 251–271. Springer, 2024b. doi: 10.1007/978-3-031-73116-7_15.

Hanguang Xiao, Ningzhi Hui, Yong Xu, Zhipeng Li, and Jincheng Peng. Medical multimodal large language models: A survey. *Information Fusion*, 134:104386, 2026. doi: 10.1016/j.inffus.2026.104386.

Hengzhi Xue, Yin Dai, Qingyong Li, Yudong Yao, and Yueyang Teng. Progressive cross-scale semantic alignment for language-guided medical image segmentation. *Knowledge-Based Systems*, 340:115745, 2026. doi: 10.1016/j.knosys.2026.115745.

A Additional Methodological Details

A.1 DuetMoE: Architecture, Clinical Memory, and Computation

The following diagrams show the DuetMoE block and the segmentation architecture.

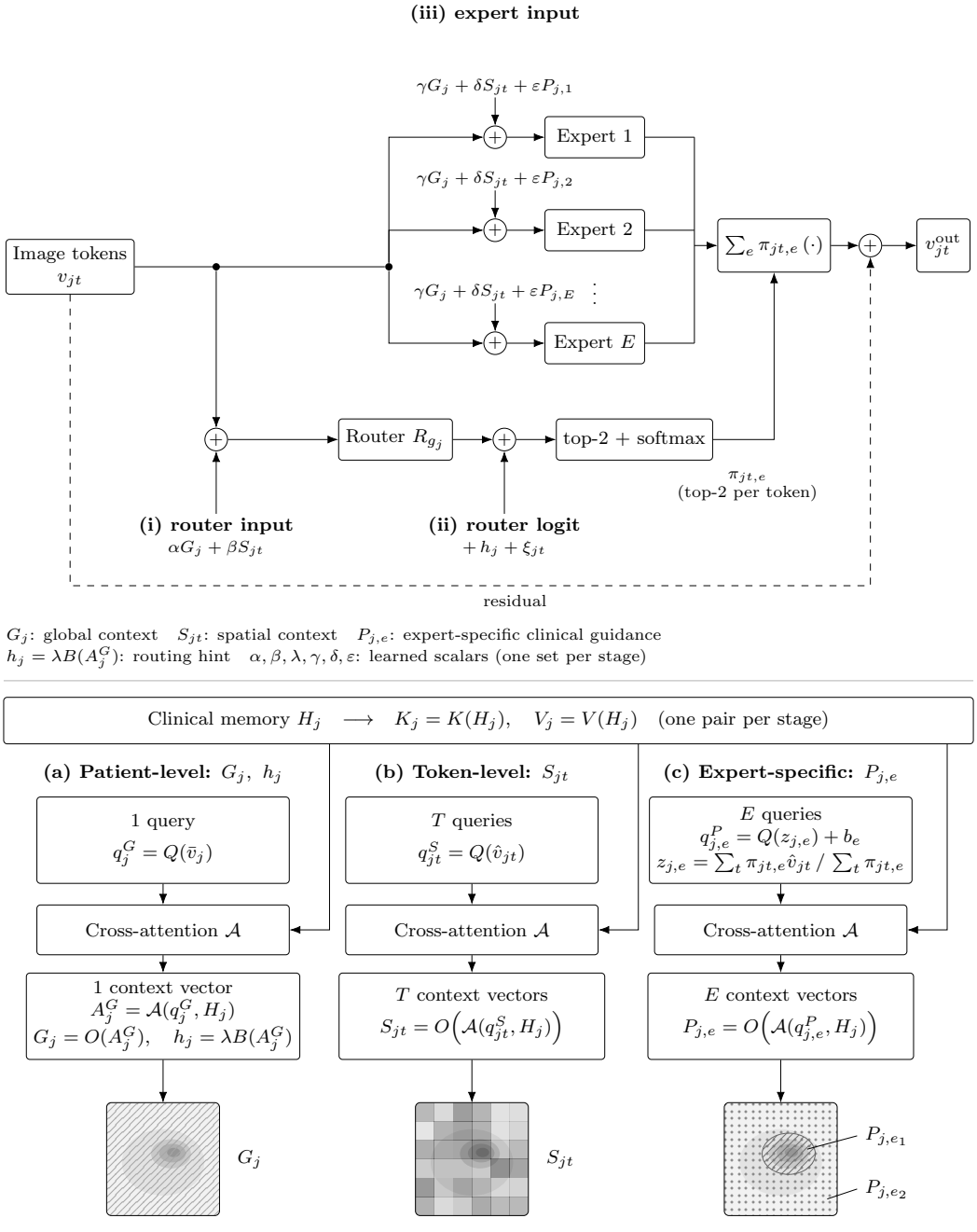


Figure 5: DuetMoE block and clinical-memory reads. Expert-specific queries are evaluated only for patient–expert pairs with positive routing mass, giving at most E queries per patient and stage.

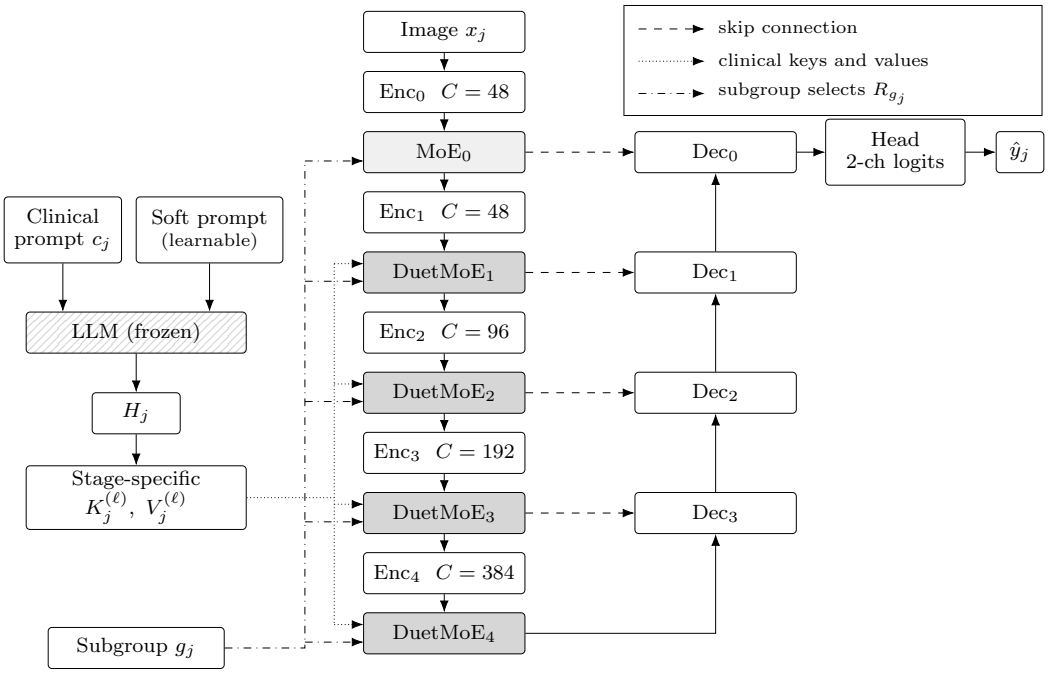


Figure 6: DuetMoE segmentation architecture.

Clinical-memory construction. In the PI-CAI implementation, the clinical record c_j is serialized as a prompt with tags delimiting its fields. A shared learnable soft-prompt sequence is inserted after the beginning-of-sequence token, and a frozen language model encodes the resulting sequence. Hidden states from the tagged conditioning spans form the memory vectors h_{jm}. For numerical fields, an encoding of the value and an observation indicator is appended to the corresponding tagged representation. Padding is masked, so M_j counts only valid memory vectors. The same H_j supplies keys and values to all clinical reads for patient j.

Attention and routing. At each clinical-context stage, LN normalizes each image token over its channels. The global query uses the mean $\bar{v}_j$, while each spatial query uses its own $\hat{v}_{jt}$. Stage-specific linear projections Q and O and the attention operator $\mathcal{A}$ produce Eq. (1). The global attention output before O supplies the expert-wise hint $B(A_j^G)$. For the selected set $\mathcal{K}_{jt}$ of K experts, the sparse weight in Eq. (2) is

$$\pi_{jt,e} = \begin{cases} \dfrac{\exp(r_{jt,e})}{\sum_{k \in \mathcal{K}_{jt}} \exp(r_{jt,k})}, & e \in \mathcal{K}_{jt}, \\ 0, & e \notin \mathcal{K}_{jt}. \end{cases} \tag{7}$$

The noise ξ_{jt} perturbs scores during both training and evaluation, with a fixed random seed at evaluation. The subgroup label selects R_{g_j}, while the experts are shared across subgroups.

Coefficients and initialization. The router-input coefficients satisfy $\alpha, \beta \in [0, 1]$, the hint coefficient $\lambda \geq 0$, and the expert-input coefficients $\gamma, \delta, \varepsilon \geq 0$ with $\gamma + \delta + \varepsilon \leq 1$; the constraints are enforced by projection after every update. They are initialized at $\alpha = \beta = \lambda = 1/7$ and $\gamma = \delta = \varepsilon = 2/7$, held fixed for the first ten epochs, and then learned per stage. The output projection O and the hint projection B are initialized to zero, so every block starts as an image-only routed MoE layer and the clinical reads enter gradually during training.

Blocks, experts, and memory. One MoE block follows each of the five encoder stages; the first block has no clinical reads, and the decoder contains no MoE blocks. Each block has $E = 8$ experts with top-2 routing, one router per subgroup, and experts shared across subgroups. An expert is a two-layer MLP with hidden width $4C$, ReLU, and dropout 0.1, where C is the stage width; the

routing noise ξ_{jt} is Gaussian with a learned, softplus-scaled amplitude (Shazeer et al., 2017). The attention operator $\mathcal{A}$ uses four heads, and each stage has one pair of key and value projections of memory vectors, shared by the global, spatial, and expert-specific reads.

Expert-specific read and update. For each patient–expert pair with positive routing mass, Eq. (3) averages the normalized tokens $\hat{v}_{jt}$ using their routing weights $\pi_{jt,e}$. The resulting summary $z_{j,e}$, together with the expert-specific offset b_e, queries the same patient memory to obtain $P_{j,e}$. Expert pairs with no selected tokens make no contribution. Equation (4) then adds the weighted expert outputs to the original token. All six clinical coefficients are learned per stage, and only the last four of the five encoder stages use these reads.

Language model. The clinical memory is produced by Meta-Llama-3.1-8B-Instruct, kept frozen, with 32 learnable soft-prompt tokens. In radiotherapy, the encoded sequence has 96 positions (the beginning-of-sequence token, the 32 soft-prompt tokens, and 63 report tokens), and all of them are retained as memory vectors of width $4{,}096$. In PI-CAI, the tokenized record has at most 384 positions before insertion of the 32 soft-prompt tokens; only the tagged conditioning spans are retained in memory. There, each memory vector carries an 11-dimensional numeric tail that can be nonzero only at the first position of a numeric field's tagged span (width $4{,}107$).

A.2 DuetMoE+: reweighting, KL reformulation, and sensitivity to intra-subgroup variation

We prove the architectural reduction in Proposition 1, derive the finite-sample representation of the within-subgroup robust risk in Eq. (5), and prove Proposition 2. Throughout the robust-risk analysis, $I_g = \{j : g_j = g\}$ is a nonempty subgroup, $n_g = |I_g|$, and $L_j(\theta)$ is the finite loss of patient j. The parameter vector θ includes all trainable model parameters. The nominal distribution is uniform over the observed patients in each subgroup, and all logarithms are natural.

A.2.1 Proof of Proposition 1

Proof. Fix an encoder stage, its input tokens, subgroup routers, shared experts, and gating rule. Disabling the clinical pathways removes their additive contributions in Eqs. (2) and (4). Algebraically, this is achieved by setting the six clinical coefficients $\alpha, \beta, \lambda, \gamma, \delta, \varepsilon$ to zero. The router and expert inputs consequently become

$$u_{jt} = v_{jt}, \qquad r_{jt} = R_{g_j}(v_{jt}) + \xi_{jt}, \qquad \widetilde{v}_{jt,e} = v_{jt}. \tag{8}$$

Thus, with the same top-K selection and tie-breaking rule,

$$\begin{aligned} \pi_{jt}^{\mathrm{dMoE}} &= \mathrm{SoftmaxTopK}\big(R_{g_j}(v_{jt}) + \xi_{jt}\big)\,, \\ v_{jt}^{\mathrm{out}} &= v_{jt} + \sum_{e \in \mathcal{K}_{jt}^{\mathrm{dMoE}}} \pi_{jt,e}^{\mathrm{dMoE}} E_e(v_{jt}), \end{aligned} \tag{9}$$

where $\mathcal{K}_{jt}^{\mathrm{dMoE}}$ is the selected expert set. This is the subgroup-routed dMoE residual update with the stipulated routers, experts, and gating rule. It holds for every patient and image token. When gating or expert computation is stochastic, equality holds for the same realization of the random draws in both architectures; under the same randomization law, the updates therefore also have the same distribution. No clinical-memory read remains in the update, proving the proposition. □

This is an architectural identity for fixed model components. It does not assert that models trained with different loss objectives attain the same parameters.

A.2.2 Two-level reweighting and composite sample weights

Consider a weighted empirical objective with subgroup weights $\alpha_g \geq 0$, $\sum_g \alpha_g = 1$, and within-subgroup patient weights $\beta_j^{(g)} \geq 0$, $\sum_{j \in I_g} \beta_j^{(g)} = 1$:

$$\min_{\theta} \sum_{g \in \mathcal{G}} \alpha_g \sum_{j \in I_g} \beta_j^{(g)} L_j(\theta). \tag{10}$$

Here, α_g controls the total mass assigned to subgroup g, whereas $\beta_j^{(g)}$ distributes that mass among its patients. For fixed weights and differentiable patient losses,

$$\nabla_\theta \left(\sum_{g \in \mathcal{G}} \alpha_g \sum_{j \in I_g} \beta_j^{(g)} L_j(\theta) \right) = \sum_{g \in \mathcal{G}} \sum_{j \in I_g} \alpha_g \beta_j^{(g)} \nabla_\theta L_j(\theta). \tag{11}$$

Thus, the coefficient of a patient's loss gradient is the product $\alpha_g \beta_j^{(g)}$. An individual coefficient can reflect a larger subgroup mass, a larger within-subgroup weight, or both. This is a statement about an individual coefficient: given the subgroup partition and the complete normalized weight vector $\omega_j = \alpha_g \beta_j^{(g)}$, one can recover $\alpha_g = \sum_{j \in I_g} \omega_j$ and, when $\alpha_g > 0$, $\beta_j^{(g)} = \omega_j / \alpha_g$.

In DuetMoE+, subgroup labels also select the router that processes the image features. This architectural role is distinct from assigning subgroup mass in a loss aggregation. The objective in Eq. (5) uses a sum of subgroup risks, with each subgroup coefficient fixed at one; its inner KL optimization changes patient weights within each subgroup. Averaging this sum over $|\mathcal{G}|$ would multiply the objective and its gradient by $1/|\mathcal{G}|$. The fixed-weight calculation above is therefore a comparison of loss coefficients, rather than an additional outer adversarial optimization in DuetMoE+.

A.2.3 KL REFORMULATION AND PATIENT-WEIGHT INTERPRETATION

Let $u_g = (1/n_g, \ldots, 1/n_g)$ and

$$\Delta_{n_g} = \left\{ q \in \mathbb{R}_+^{n_g} : \sum_{j \in I_g} q_j = 1 \right\}, \qquad D_{\mathrm{KL}}(q \| u_g) = \sum_{j \in I_g} q_j \log(n_g q_j), \tag{12}$$

with $0 \log 0 = 0$. A distribution on the observed patients is represented by $q \in \Delta_{n_g}$. The robust subgroup risk is

$$\widehat{R}_g^{\mathrm{rob}}(\theta) = \max_{q \in \Delta_{n_g}} \left\{ \sum_{j \in I_g} q_j L_j(\theta) : D_{\mathrm{KL}}(q \| u_g) \le \rho_g \right\}. \tag{13}$$

The feasible set is nonempty and compact, since it contains u_g and KL divergence is continuous on the finite simplex under the convention above. The maximum is therefore attained. This formulation keeps the observed patients fixed and varies their probabilities within a KL neighborhood of the uniform empirical distribution.

For every $\rho_g \ge 0$, the exact dual representation is

$$\widehat{R}_g^{\mathrm{rob}}(\theta) = \inf_{\eta_g > 0} \left\{ \eta_g \rho_g + \eta_g \log \left(\frac{1}{n_g} \sum_{j \in I_g} \exp\left(\frac{L_j(\theta)}{\eta_g} \right) \right) \right\}. \tag{14}$$

The use of an infimum includes boundary cases in which no finite positive dual optimizer exists. For $\eta_g > 0$, define

$$q_j^{(g)}(\eta_g) = \frac{\exp(L_j(\theta)/\eta_g)}{\sum_{k \in I_g} \exp(L_k(\theta)/\eta_g)}, \qquad j \in I_g. \tag{15}$$

Derivation and boundary cases. Fix g and θ, and temporarily suppress the subgroup index. Write $Z(\eta) = n^{-1} \sum_j \exp(L_j/\eta)$ and let $q(\eta)$ be the weights in Eq. (15). Direct substitution gives, for every $q \in \Delta_n$,

$$\sum_j q_j L_j - \eta D_{\mathrm{KL}}(q \| u) = \eta \log Z(\eta) - \eta D_{\mathrm{KL}}(q \| q(\eta)). \tag{16}$$

Since KL divergence is nonnegative, maximizing the left-hand side over the simplex gives $\eta \log Z(\eta)$, uniquely attained at $q(\eta)$. Every feasible q consequently satisfies $\sum_j q_j L_j \le \eta\rho + \eta \log Z(\eta)$, proving the upper bound in Eq. (14).

To establish equality and characterize attainment, define

$$K(t) = \log \left(\frac{1}{n} \sum_j e^{t L_j} \right), \qquad p_j(t) = \frac{e^{t L_j}}{\sum_k e^{t L_k}}, \qquad d(t) = t K'(t) - K(t), \quad t \ge 0. \tag{17}$$

These quantities satisfy $d(t) = D_{\mathrm{KL}}(p(t)\|u)$ and

$$K'(t) = \sum_j p_j(t) L_j, \qquad K''(t) = \sum_j p_j(t)\big(L_j - K'(t)\big)^2, \qquad d'(t) = tK''(t). \tag{18}$$

Let $M = \{j : L_j = \max_k L_k\}$, $m = |M|$, and $\rho_{\mathrm{sat}} = \log(n/m)$. If the losses are not all equal, $K''(t) > 0$ for finite t, because every $p_j(t)$ is positive. Thus d is strictly increasing on $(0, \infty)$, starts at $d(0) = 0$, and converges to ρ_{sat}: the weights converge to the uniform distribution on M, whose divergence from u is $\log(n/m)$.

For $0 < \rho < \rho_{\mathrm{sat}}$, there is a unique $t_\rho > 0$ satisfying $d(t_\rho) = \rho$. Set $\eta^* = 1/t_\rho$. Then $p(t_\rho) = q(\eta^*)$ is feasible and attains equality in Eq. (16), so the primal maximum equals the dual value at η^*. The same identity shows that this primal maximizer is unique. Its KL constraint is active.

For $\rho = 0$, the only feasible distribution is u, and

$$\widehat{R}_g^{\mathrm{rob}}(\theta) = \frac{1}{n_g} \sum_{j \in I_g} L_j(\theta). \tag{19}$$

Taking $\eta \to \infty$ in the dual objective recovers this mean. For $\rho \geq \rho_{\mathrm{sat}}$, the uniform distribution on M is feasible and attains $\max_j L_j$. The dual objective has the same limiting value as $\eta \downarrow 0$, because

$$\lim_{\eta \downarrow 0} \eta \log \left(\frac{1}{n} \sum_j e^{L_j/\eta} \right) = \max_j L_j. \tag{20}$$

If all losses equal a constant c, every feasible distribution has expected loss c, and the dual expression is $c + \eta\rho$; its infimum is also c. This includes $n_g = 1$. These cases complete the derivation of Eq. (14) for every nonnegative radius.

For an interior radius and its finite optimizer, higher-loss patients have larger weights: $q_j^{(g)}(\eta_g^*)/q_k^{(g)}(\eta_g^*) = \exp((L_j - L_k)/\eta_g^*)$. More generally, any maximizing vector satisfies $L_j > L_k \Rightarrow q_j^* \geq q_k^*$. Otherwise, swapping those two weights would preserve their KL divergence under the uniform nominal distribution and strictly increase expected loss. At a saturated radius, distinct nonmaximal losses can both receive zero weight, so the general inequality is non-strict. Uniform weights are feasible for every radius, giving

$$\widehat{R}_g^{\mathrm{mean}}(\theta) \leq \widehat{R}_g^{\mathrm{rob}}(\theta) \leq \max_{j \in I_g} L_j(\theta), \tag{21}$$

and the robust risk is nondecreasing in ρ_g because its feasible sets are nested.

Mean-loss baseline and subgroup weighting. At zero radii, Eq. (5) becomes the sum of subgroup means,

$$\mathcal{L}_{\mathrm{mean}}(\theta) = \sum_{g \in \mathcal{G}} \widehat{R}_g^{\mathrm{mean}}(\theta) = \sum_{g \in \mathcal{G}} \frac{1}{n_g} \sum_{j \in I_g} L_j(\theta). \tag{22}$$

Dividing this objective by $|\mathcal{G}|$ gives the group-balanced empirical risk. In contrast, with $N = \sum_g n_g$, the pooled empirical risk is

$$\widehat{R}_{\mathrm{pool}}(\theta) = \frac{1}{N} \sum_j L_j(\theta) = \sum_{g \in \mathcal{G}} \frac{n_g}{N} \widehat{R}_g^{\mathrm{mean}}(\theta). \tag{23}$$

The group-balanced and pooled objectives assign different relative subgroup weights when subgroup sizes differ. Accordingly, the mean-loss version in Proposition 2 preserves the subgroup aggregation of Eq. (5); identifying an empirical baseline with this zero-radius case requires the same aggregation and patient-loss definition.

Gradient interpretation. Assume the patient losses are continuously differentiable near θ. For a realized stochastic forward pass, hold the underlying random draws fixed. With top-K routing and ReLU activations, the following statement applies on regions where the stated differentiability assumption holds. Write $\mathcal{C}_g = \{q \in \Delta_{n_g} : D_{\mathrm{KL}}(q\|u_g) \leq \rho_g\}$ and $\mathcal{S}_g(\theta) = \arg\max_{q \in \mathcal{C}_g} \sum_j q_j L_j(\theta)$. For a direction v, the one-sided directional derivative is

$$\big(\widehat{R}_g^{\mathrm{rob}}\big)'(\theta; v) = \max_{q \in \mathcal{S}_g(\theta)} \sum_{j \in I_g} q_j \nabla_\theta L_j(\theta)^\top v. \tag{24}$$

To see this, set $F_g(\theta, q) = \sum_j q_j L_j(\theta)$ and choose an optimizer q_h at $\theta + hv$, $h > 0$. For any $q \in \mathcal{S}_g(\theta)$, optimality gives

$$\begin{aligned} F_g(\theta + hv, q) - F_g(\theta, q) &\le \widehat{R}_g^{\mathrm{rob}}(\theta + hv) - \widehat{R}_g^{\mathrm{rob}}(\theta) \\ &\le F_g(\theta + hv, q_h) - F_g(\theta, q_h). \end{aligned} \tag{25}$$

Compactness and continuity imply that every subsequential limit of q_h belongs to $\mathcal{S}_g(\theta)$. The first-order Taylor remainder is uniform over $\mathcal{C}_g$, since finitely many patient losses are combined with simplex weights. Dividing these bounds by h gives Eq. (24). If $\mathcal{S}_g(\theta) = \{q^{(g),*}\}$, the same argument for arbitrary parameter perturbations establishes differentiability and

$$\nabla_\theta \widehat{R}_g^{\mathrm{rob}}(\theta) = \sum_{j \in I_g} q_j^{(g),*} \nabla_\theta L_j(\theta). \tag{26}$$

When every subgroup maximizer is unique, the gradient of the sum in Eq. (5) is the sum of these weighted subgroup gradients. This envelope calculation accounts for the loss dependence of the optimal weights; it does not require convexity of the neural-network losses in θ. At a nonunique maximizer, Eq. (24) applies without asserting an ordinary gradient.

A.2.4 Proof of Proposition 2

Proof. Fix θ and a nonempty subgroup g. Define the moments under its uniform empirical distribution by

$$\mu_g = \frac{1}{n_g} \sum_{j \in I_g} L_j(\theta), \qquad z_j = L_j(\theta) - \mu_g, \qquad \sigma_g^2 = \frac{1}{n_g} \sum_{j \in I_g} z_j^2. \tag{27}$$

Thus, the standard deviation uses the empirical-distribution denominator n_g, rather than the unbiased-estimation denominator $n_g - 1$. The recovery of the mean-loss objective when all radii are zero follows from Eq. (19).

If $\sigma_g = 0$, all z_j equal zero. Every feasible distribution then has expected loss μ_g, so $\widehat{R}_g^{\mathrm{rob}}(\theta) = \mu_g$ for every $\rho_g \ge 0$. This also proves the claimed expansion in this case, with zero remainder.

Suppose $\sigma_g > 0$, and temporarily omit g from n, μ, σ, and ρ. Define the centered log moment-generating function

$$\psi(t) = \log\left(\frac{1}{n} \sum_j e^{t z_j} \right). \tag{28}$$

The finite loss vector makes ψ smooth near zero, with $\psi(0) = \psi'(0) = 0$ and $\psi''(0) = \sigma^2$. Taylor expansion therefore gives

$$\psi(t) = \frac{\sigma^2}{2} t^2 + O(t^3), \qquad \psi'(t) = \sigma^2 t + O(t^2). \tag{29}$$

For all sufficiently small positive ρ, the preceding KL derivation provides a unique positive tilt t_ρ satisfying

$$\rho = t_\rho \psi'(t_\rho) - \psi(t_\rho) = \frac{\sigma^2}{2} t_\rho^2 + O(t_\rho^3), \qquad \widehat{R}_g^{\mathrm{rob}}(\theta) = \mu + \psi'(t_\rho). \tag{30}$$

Centering does not change the tilt weights or their KL divergence. Moreover, $t_\rho \to 0$ as $\rho \downarrow 0$, since the divergence is continuous and strictly increasing from zero.

The positive quadratic coefficient in Eq. (30) implies $t_\rho = O(\sqrt{\rho})$. Hence $t_\rho^2 - 2\rho/\sigma^2 = O(\rho^{3/2})$. Dividing this difference of squares by $t_\rho + \sqrt{2\rho}/\sigma \ge \sqrt{2\rho}/\sigma$ yields

$$t_\rho = \frac{\sqrt{2\rho}}{\sigma} + O(\rho). \tag{31}$$

Substituting into the risk identity in Eq. (30) gives

$$\widehat{R}_g^{\mathrm{rob}}(\theta) = \mu + \sigma^2 t_\rho + O(t_\rho^2) = \mu + \sqrt{2\rho}\,\sigma + O(\rho). \tag{32}$$

Restoring the subgroup index and using $\mu_g = \widehat{R}_g^{\mathrm{mean}}(\theta)$ proves Proposition 2. The remainder is pointwise for fixed g and θ; its constant may depend on their finite loss vector. □

The expansion identifies a local standard-deviation term in the empirical robust risk. It describes the training objective, rather than guaranteeing reduced prediction variability or improved performance for every patient.

B Implementation Details

B.1 Dataset cohorts and splits

We report each protocol separately. The two RT evaluations use different cohorts and subgroup definitions, so their results are compared within each protocol.

PI-CAI (institution-grouped multimodal protocol). The cohort comprises the 1,500 public PI-CAI cases (RUMC 800, ZGT 350, PCNN 350), each with T2-weighted, apparent diffusion coefficient, and high-b-value MRI, a case-level csPCa label, and the clinical fields listed in Section 4.1. We hold out one fold of a patient-level five-fold split, stratified by institution and csPCa status, as the test set (301 cases from 295 patients: RUMC 160, ZGT 70, PCNN 71) and split the remaining 1,199 cases at the patient level into 960 training and 239 validation cases, so that no patient appears in more than one split. Institution is the subgroup label for routing, and the linked clinical fields form the clinical memory of the multimodal models.

Harvard-FairSeg (race-grouped unimodal protocol). The official test split contains 2,000 images: Asian 169, Black 299, and White 1,532. The main and paired lower-tail comparisons use a shared KL radius $\rho_g = 0.1$ across racial subgroups; Appendix D.3 examines radius sensitivity.

RT (institution-grouped multimodal protocol). The fixed split contains 709 training cases (SC 659, YI 24, GN 26), 105 validation cases (SC 95, YI 5, GN 5), and 420 test cases (SC 189, YI 112, GN 119). Institution selects the router, and the linked clinical fields listed in Section 4.1 form the clinical memory. These fields include a de-identified physician code, because target delineation can differ between physicians. The code is assigned per institution from the training split: physicians with at least five training cases receive integer codes in order of frequency, and all other physicians, including those absent from the training split or missing from the record, receive the reserved code 0. At inference, an unknown or unrecorded physician is therefore represented by the same code as infrequent physicians during training, and no physician names are used.

RT (T-stage-grouped unimodal protocol). The image-only evaluation contains 275 held-out cases: T1 11, T2 129, T3 114, and T4 21. The 275 test cases are not treated as a subset of the 420-case test split above.

B.2 Fairness metrics

Segmentation overlap. Tables 7 and 8 supplement the main Dice comparisons with intersection over union (IoU). For predicted and reference foreground masks P and Y, with nonzero denominators,

$$\mathrm{Dice}(P,Y) = \frac{2|P \cap Y|}{|P| + |Y|}, \qquad \mathrm{IoU}(P,Y) = \frac{|P \cap Y|}{|P \cup Y|}. \tag{33}$$

Both are higher-is-better scores. We compute them per case and average over the full evaluation set or each subgroup; FairSeg rim and cup are evaluated separately.

Equity-scaled performance. Let M_{all} and M_g denote performance on the full evaluation set and within subgroup $g \in \mathcal{G}$. We report subgroup values, the worst-subgroup value $\min_g M_g$, and the equity-scaled score (Tian et al., 2024a)

$$\mathrm{ES}(M) = \frac{M_{\mathrm{all}}}{1 + \sum_{g \in \mathcal{G}} |M_{\mathrm{all}} - M_g|}. \tag{34}$$

Mean Dice and IoU give ES-Dice and ES-IoU. The PI score averages case-level AUROC and lesion-level average precision (Saha et al., 2024); applying the same scaling gives ES-PI. These scores assess inter-subgroup performance and are complemented by the within-subgroup outcomes below.

Intra-subgroup outcomes. For PI-CAI, institution-level case-score ROC curves give sensitivity at 90% specificity and partial AUROC over false-positive rates up to 0.2 (Tables 5 and 6). For RT, we report patient-level Dice distributions and the 25th percentile within each T-stage subgroup. For FairSeg, the lowest-10% mean Dice averages the lowest $\lceil 0.1 n_g \rceil$ scores in subgroup g, ranked separately by method and target. Tail membership may therefore differ across methods. The radius analysis also reports the minimum of these subgroup lower-tail means.

Uncertainty. For methods with case-level predictions, 95% CIs follow the $1{,}000$-resample subgroup bootstrap in Section 4.1; aggregate-only published results are shown as reported. CIs condition on the evaluated checkpoints and do not capture variation across training seeds or checkpoint selection.

B.3 Experimental setup

Main 2D protocol: Harvard-FairSeg. The main 2D setup uses TransUNet (Chen et al., 2021) with a ViT-B encoder, following the published fairness-baseline protocol. Images are center-cropped and resized to 224×224 pixels. Training uses a batch size of 42, an initial learning rate of 1×10^{-2}, and 100 epochs; the reported checkpoint is selected among those saved every 10 epochs. The lower-tail comparison in Figure 7 and radius sweep in Appendix D.3 use a separate 80-epoch schedule with the fixed final checkpoint (epoch 79).

Main 3D protocol: image-only RT. The image-only RT setup describes a 3D residual U-Net backbone based on 3D U-Net (Çiçek et al., 2016), as used for radiotherapy target delineation (Oh et al., 2024a). Training samples $384 \times 384 \times 128$-voxel patches with a batch size of 4, and evaluation applies sliding-window inference to full CT volumes. The training schedule runs for up to 100 epochs, starting at 5×10^{-5}, with validation-driven early stopping.

Software and hardware. The 2D and 3D implementations report PyTorch (Paszke et al., 2019) on CUDA 11.8, AdamW (Loshchilov & Hutter, 2017) with exponential learning-rate decay, a single NVIDIA A100 80 GB GPU for 2D, and a single NVIDIA RTX A6000 48 GB GPU for 3D.

Multimodal protocols. The segmentation networks in the main multimodal comparisons are trained from scratch, while text encoders are frozen pretrained models; training uses AdamW (Loshchilov & Hutter, 2017) for 101 epochs, with 10 warm-up epochs followed by cosine decay and weight decay 10^{-5}; the validation split is used only for monitoring, and these comparisons use the checkpoint after the final epoch. The base learning rate is 5×10^{-5} for radiotherapy (batches of four 3D crops) and 8.66×10^{-5} for PI-CAI (batches of 12 cases, scaled by the square root of the batch-size ratio), and DuetMoE's six pathway coefficients at each clinical-context stage use $16\times$ the base learning rate. Radiotherapy uses cross-entropy plus foreground soft Dice loss. PI-CAI uses cross-entropy plus softmax focal loss with equal weights and deep supervision with weights $(8, 4, 2, 1)/15$; for PI-CAI, each epoch consists of 80 optimizer steps, and each batch draws two cancer-positive and two cancer-negative cases from each of the three institutions, giving equal case counts per institution at each update. The main multimodal runs use a single NVIDIA RTX PRO 6000 Blackwell Server Edition GPU.

C Additional Experimental Results

C.1 PI-CAI: intra-subgroup detection performance

Tables 5 and 6 complement the PI scores in Table 1 with detection performance at high specificity within institutions. **DuetMoE** raises RUMC sensitivity from 0.574 with the strongest baseline for this metric to 0.681 ($+10.7$ percentage points) and achieves the highest normalized partial AUROC in RUMC and ZGT.

C.2 Unimodal radiotherapy: additional evaluation metrics

Table 7 extends the Dice comparison in Table 3 with IoU and ES-IoU. **DuetMoE+** also achieves the highest ES-IoU (0.411) and raises IoU in T2, the lowest-scoring stage for dMoE, from 0.435 to 0.473, supporting improved inter-subgroup performance.

Table 5: **PI-CAI.** Sensitivity at 90% specificity for **multimodal** csPCa detection, evaluated overall and separately within each institution from case scores. Here n_+ is the number of cancer-positive cases.

Method	All ($n_+ = 85$)	RUMC ($n_+ = 47$)	ZGT ($n_+ = 16$)	PCNN ($n_+ = 22$)
3D ResUNet	0.459	0.468	<u>0.625</u>	**0.364**
ConTEXTualNet3D-based	0.435	0.532	0.438	0.136
PCSA-Seg	0.424	0.532	0.375	0.273
dMoE	<u>0.518</u>	<u>0.574</u>	**0.688**	<u>0.318</u>
MoME	0.506	0.553	<u>0.625</u>	<u>0.318</u>
DuetMoE	**0.565**	**0.681**	**0.688**	0.182

Bold and underlined point estimates indicate the highest and second-highest values in each column, respectively; ties share the same marking.

Table 6: **PI-CAI.** Normalized partial AUROC over false-positive rates up to 0.2 for **multimodal** csPCa detection, evaluated overall and separately within each institution from case scores.

Method	All ($n = 301$)	RUMC ($n = 160$)	ZGT ($n = 70$)	PCNN ($n = 71$)
3D ResUNet	0.447	0.409	0.605	**0.353**
ConTEXTualNet3D-based	0.394	0.473	0.416	0.202
PCSA-Seg	0.407	0.494	0.433	0.252
dMoE	0.459	0.500	<u>0.635</u>	0.238
MoME	**0.500**	<u>0.546</u>	0.605	<u>0.311</u>
DuetMoE	<u>0.488</u>	**0.564**	**0.670**	0.185

Bold and underlined point estimates indicate the highest and second-highest values in each column, respectively; ties share the same marking.

Table 7: **Radiotherapy.** Extended comparison on **unimodal** 3D target segmentation using tumor stage as the subgroup attribute, supplementing Table 3 with IoU and ES-IoU.

Method	All ($n = 275$)				T1 ($n = 11$)		T2 ($n = 129$)		T3 ($n = 114$)		T4 ($n = 21$)	
	ES-Dice (95% CI)	Dice (95% CI)	ES-IoU (95% CI)	IoU (95% CI)	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
3D ResUNet Çiçek et al. (2016)	0.487 (0.447–0.529)	0.610 (0.589–0.630)	0.367 (0.336–0.399)	0.462 (0.440–0.482)	0.492	0.341	0.570	0.420	0.659	0.511	0.656	0.506
+ FEBS Tian et al. (2024a)	0.434 (0.406–0.467)	0.586 (0.567–0.604)	0.322 (0.302–0.346)	0.433 (0.414–0.452)	0.466	0.288	0.522	0.374	0.652	0.501	0.688	0.527
+ MoE Shazeer et al. (2017)	0.452 (0.415–0.492)	0.608 (0.586–0.628)	0.342 (0.314–0.372)	0.461 (0.439–0.482)	0.492	0.345	0.543	0.393	0.674	0.532	0.708	0.557
+ GDRO Sagawa et al. (2019)	0.454 (0.421–0.491)	0.607 (0.586–0.627)	0.340 (0.314–0.372)	0.459 (0.436–0.481)	0.481	0.329	0.546	0.395	0.671	0.528	0.694	0.547
+ Prompt-GDRO Panaganti et al. (2026)	0.476 (0.440–0.517)	0.612 (0.592–0.631)	0.357 (0.329–0.392)	0.462 (0.441–0.482)	0.492	0.338	0.562	0.411	0.669	0.523	0.673	0.520
+ dMoE Oh et al. (2025)	<u>0.499</u> (0.469–0.531)	<u>0.650</u> (0.628–0.671)	<u>0.384</u> (0.351–0.419)	<u>0.507</u> (0.484–0.528)	<u>0.718</u>	<u>0.571</u>	<u>0.586</u>	<u>0.435</u>	**0.693**	**0.556**	**0.778**	**0.641**
+ **DuetMoE+**	**0.530** (0.499–0.569)	**0.665** (0.643–0.683)	**0.411** (0.384–0.444)	**0.521** (0.498–0.540)	**0.757**	**0.621**	**0.621**	**0.473**	<u>0.688</u>	<u>0.547</u>	<u>0.758</u>	<u>0.615</u>

Bold and underlined point estimates indicate the highest and second-highest values in each column, respectively; ties share the same marking. Parentheses show 95% confidence intervals.

C.3 HARVARD-FAIRSEG: SUBGROUP PERFORMANCE AND PATIENT LOWER TAILS

Table 8 extends the cup Dice comparison in Table 4 to both segmentation targets and additional overlap metrics. **DuetMoE+** achieves the highest ES-Dice for both neuroretinal rim (0.755) and optic cup (0.843), providing further evidence on inter-subgroup performance. Figure 7 complements these subgroup scores with the lowest-10% patient Dice within each racial subgroup.

D ABLATION

D.1 CLINICAL-CONTEXT CONFIGURATIONS

We compare clinical-context configurations built from the reads defined in Section 3.1. Global-context configurations inject G_j before routing or at the expert inputs; expert-specific clinical guidance injects $P_{j,e}$ at the expert inputs, optionally with the routing hint $B(A_j^G)$ added to the routing scores. **DuetMoE** embeds global and spatial clinical reads before routing and expert-specific guidance afterward, with a memory-derived routing hint and learned pathway coefficients. Image-only MoE and multimodal dMoE provide reference configurations. These configurations examine where clinical memory enters routing and expert computation.

Table 8: **Harvard-FairSeg.** Extended comparison on **unimodal** 2D neuroretinal-rim and optic-cup segmentation using race as the subgroup attribute. Dice, IoU, and their equity-scaled scores supplement the cup Dice results in Table 4.

Method	All ($n = 2000$)				Asian ($n = 169$)		Black ($n = 299$)		White ($n = 1532$)	
	ES-Dice (95% CI)	Dice (95% CI)	ES-IoU (95% CI)	IoU (95% CI)	Dice	IoU	Dice	IoU	Dice	IoU
Neuroretinal rim										
TransUNet† Chen et al. (2021)	0.703	0.793	0.585	0.671	0.746	0.616	0.731	0.599	0.811	0.691
+ ADV† Madras et al. (2018)	0.700	0.791	0.583	0.668	0.741	0.612	0.729	0.598	0.809	0.689
+ FEBS† Tian et al. (2024a)	0.705	0.795	0.587	0.673	0.748	0.619	0.733	0.602	0.813	0.694
+ FairDiff‡ Li et al. (2024)	0.716	0.800	0.596	0.680	0.757	0.628	0.743	0.611	0.817	0.699
+ MoE Shazeer et al. (2017)	0.733 (0.713–0.752)	0.804 (0.799–0.809)	0.614 (0.596–0.633)	0.685 (0.680–0.691)	0.760	0.635	0.763	0.635	0.817	0.701
+ GDRO Sagawa et al. (2019)	0.742 (0.722–0.763)	0.809 (0.804–0.814)	0.625 (0.607–0.645)	0.691 (0.686–0.697)	0.764	0.641	0.775	0.649	0.820	0.705
+ Prompt-GDRO Panaganti et al. (2026)	0.747 (0.726–0.767)	**0.817** (0.812–0.822)	0.631 (0.612–0.649)	**0.703** (0.697–0.708)	0.767	0.643	**0.784**	**0.663**	**0.828**	**0.717**
+ dMoE Oh et al. (2025)	0.743 (0.723–0.763)	0.813 (0.808–0.818)	0.627 (0.608–0.645)	0.698 (0.692–0.704)	0.769	0.645	0.776	0.652	0.825	0.713
+ **DuetMoE+**	**0.755** (0.735–0.775)	0.816 (0.811–0.821)	**0.639** (0.619–0.657)	0.701 (0.696–0.707)	**0.780**	**0.658**	0.782	0.659	0.827	0.714
Optic cup										
TransUNet† Chen et al. (2021)	0.828	0.848	0.730	0.753	0.827	0.728	0.849	0.758	0.850	0.755
+ ADV† Madras et al. (2018)	0.826	0.841	0.727	0.743	0.825	0.726	0.842	0.748	0.843	0.744
+ FEBS† Tian et al. (2024a)	0.825	0.846	0.727	0.750	0.825	0.725	0.848	0.755	0.848	0.751
+ FairDiff‡ Li et al. (2024)	0.833	0.848	0.736	0.753	0.832	0.735	0.848	0.757	0.850	0.754
+ MoE Shazeer et al. (2017)	0.830 (0.809–0.847)	0.854 (0.849–0.860)	0.739 (0.720–0.754)	0.762 (0.755–0.768)	0.845	0.757	0.842	0.748	0.857	0.765
+ GDRO Sagawa et al. (2019)	0.840 (0.818–0.857)	0.860 (0.855–0.865)	0.750 (0.728–0.766)	0.769 (0.763–0.775)	0.850	0.760	0.851	0.759	0.863	0.772
+ Prompt-GDRO Panaganti et al. (2026)	0.842 (0.819–0.861)	**0.866** (0.861–0.871)	**0.755** (0.732–0.775)	**0.778** (0.772–0.785)	0.849	0.761	**0.859**	**0.770**	**0.870**	**0.782**
+ dMoE Oh et al. (2025)	0.832 (0.810–0.853)	0.862 (0.856–0.867)	0.745 (0.722–0.765)	0.773 (0.766–0.779)	0.844	0.755	0.851	0.761	0.866	0.777
+ **DuetMoE+**	**0.843** (0.822–0.861)	0.865 (0.860–0.869)	**0.755** (0.733–0.773)	0.775 (0.769–0.781)	**0.852**	**0.763**	0.856	0.767	0.868	0.778

Bold and underlined point estimates indicate the highest and second-highest values within each target and column, respectively; ties share the same marking. Parentheses show 95% confidence intervals. † Metrics reported in Tian et al. (2024a). ‡ ES metrics recalculated from Li et al. (2024) using Eq. (34).

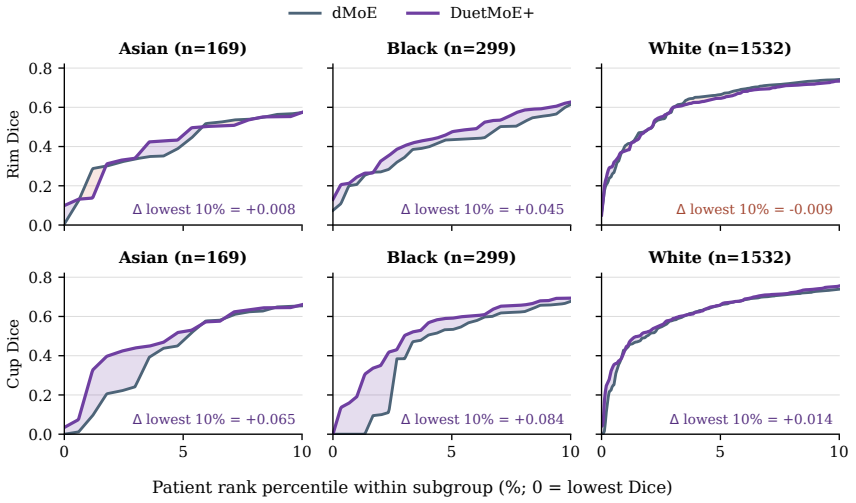


Figure 7: **Harvard-FairSeg: intra-subgroup performance.** Dice curves for the lowest-scoring 10% of patients within each racial subgroup, comparing dMoE and **DuetMoE+**. **Top:** neuroretinal rim; **bottom:** optic cup. The horizontal axis is patient rank percentile (0 denotes the lowest Dice); patients are ranked separately by method, subgroup, and target. Each Δ is the difference in lowest-10% mean Dice (**DuetMoE+** minus dMoE).

The radiotherapy configurations in Table 9 use the main cohort split but an earlier protocol with a reduced clinical prompt and different memory construction and evaluation seeding. Their rows are compared within that protocol, so the DuetMoE score there (0.742) is not comparable with the main-table value (0.761 in Table 2). Table 10 evaluates PI-CAI configurations on the main test split.

DuetMoE achieves the highest overall Dice and ES-Dice in Table 9. On PI-CAI (Table 10), it shares the highest overall PI score (0.623) with the expert-specific guidance configuration and achieves the highest ES-PI (0.500). These comparisons favor the full clinical-memory design for subgroup-aware learning.

Table 9: **Radiotherapy.** Clinical-context configurations for **multimodal** 3D target segmentation using institution as the subgroup attribute, evaluated under the earlier input and evaluation protocol described in the text.

Configuration	All ($n = 420$)		SC ($n = 189$)	YI ($n = 112$)	GN ($n = 119$)
	Dice	ES-Dice	Dice	Dice	Dice
Image-only MoE	0.680	0.593	0.727	0.694	0.594
dMoE	0.736	0.651	0.755	0.779	0.667
Global clinical context at expert input	0.739	0.660	0.757	0.777	0.675
Expert-specific clinical guidance	0.734	0.644	0.757	0.775	0.659
Expert-specific clinical guidance + routing hint	0.739	0.646	**0.761**	**0.784**	0.662
DuetMoE	**0.742**	**0.662**	0.760	0.782	**0.678**

Bold and underlining mark the highest and second-highest distinct values per column; ties share the marking.

Table 10: **PI-CAI.** Clinical-context configurations for **multimodal** csPCa detection using institution as the subgroup attribute, evaluated on the main test split.

Configuration	All ($n = 301$)		RUMC ($n = 160$)	ZGT ($n = 70$)	PCNN ($n = 71$)
	PI score	ES-PI	PI score	PI score	PI score
Image-only MoE	0.595	0.469	0.651	0.648	0.436
dMoE	0.590	0.453	0.647	**0.671**	0.423
Global clinical context before routing	0.577	0.445	0.659	0.607	0.390
Global clinical context at expert input	0.592	0.479	0.649	0.620	0.442
Expert-specific clinical guidance	**0.623**	0.487	0.686	0.634	0.418
Expert-specific clinical guidance + routing hint	0.610	0.465	0.677	0.660	0.414
DuetMoE	**0.623**	**0.500**	**0.688**	0.637	**0.457**

Bold and underlining mark the highest and second-highest distinct values per column; ties share the marking.

D.2 Inter- and intra-subgroup components of DuetMoE+

We ablate inter-subgroup routing and intra-subgroup robust training in the unimodal **DuetMoE+**. **dMoE-ERM** retains subgroup-selected expert routing with ERM training. **Subgroup-DRO (SG-DRO)** applies the within-subgroup robust objective to a shared backbone without subgroup-selected routing. **DuetMoE+** embeds this objective in subgroup-routed expert learning.

Table 11: **Harvard-FairSeg and radiotherapy.** Ablation of subgroup-selected routing and within-subgroup robust training in **unimodal** segmentation, using race and tumor stage as the respective subgroup attributes.

Setting	Method	Overall performance				Worst subgroup		Component
		ES-Dice	Dice	ES-IoU	IoU	Dice	IoU	
Harvard-FairSeg / Race (Rim)	dMoE-ERM	0.743	0.813	0.627	0.698	0.769	0.645	Inter
	SG-DRO	0.749	0.814	0.633	0.699	0.774	0.651	Intra
	DuetMoE+	**0.755**	**0.816**	**0.639**	**0.701**	**0.780**	**0.658**	Inter + Intra
Harvard-FairSeg / Race (Cup)	dMoE-ERM	0.832	0.862	0.745	0.773	0.844	0.755	Inter
	SG-DRO	0.841	0.860	0.752	0.769	0.850	0.759	Intra
	DuetMoE+	**0.843**	**0.865**	**0.755**	**0.775**	**0.852**	**0.763**	Inter + Intra
Radiotherapy / Tumor Stage	dMoE-ERM	0.499	0.650	0.384	0.507	0.586	0.435	Inter
	SG-DRO	0.490	0.616	0.370	0.468	0.515	0.359	Intra
	DuetMoE+	**0.530**	**0.665**	**0.411**	**0.521**	**0.621**	**0.473**	Inter + Intra

Worst-subgroup scores are the minimum subgroup mean Dice or IoU across races (FairSeg) or tumor stages (RT). Component labels indicate the inter-subgroup routing and intra-subgroup robust-training mechanisms. Bold and underlined point estimates indicate the highest and second-highest distinct values within each setting and column, respectively; ties share the same marking.

Table 11 shows that the full **DuetMoE+** achieves the highest equity-scaled scores and worst-subgroup Dice and IoU in all three comparisons. These results support embedding within-subgroup robust training in subgroup-routed expert learning. The patient-level lower-tail analyses in Figures 4 and 7 provide complementary evidence on intra-subgroup outcomes.

D.3 KL-RADIUS SENSITIVITY AND SELECTION IN DUETMOE+

We examine sensitivity to the shared KL radius $\rho_g = \rho$ in **DuetMoE+** through a dedicated Harvard-FairSeg sweep. The main comparison uses $\rho = 0.1$ (Appendix B.1); the sweep varies the radius under a common training setup and fixed final-checkpoint rule. Table 12 reports overall Dice, inter-subgroup metrics, and the minimum across subgroups of the lowest-10% mean Dice.

Table 12: **Harvard-FairSeg.** KL-radius sensitivity of **DuetMoE+** on **unimodal** 2D neuroretinal-rim and optic-cup segmentation using race as the subgroup attribute. A shared radius $\rho_g = \rho$ is evaluated at the fixed final checkpoint.

	Neuroretinal rim				Optic cup			
ρ	Dice	ES-Dice	Worst mean	Worst tail 10%	Dice	ES-Dice	Worst mean	Worst tail 10%
0.01	**0.818**	**0.753**	**0.777**	0.392	**0.863**	0.840	**0.851**	0.481
0.05	0.817	0.750	**0.777**	0.416	0.859	0.841	0.848	0.509
0.10	0.815	0.743	0.766	0.363	0.861	0.836	0.847	0.486
0.15	0.815	0.751	**0.777**	**0.425**	0.861	0.843	**0.851**	0.512
0.20	0.814	0.749	0.774	0.413	0.860	0.840	0.849	0.496
0.25	0.814	0.749	0.775	0.410	0.861	**0.845**	0.849	**0.524**
0.50	0.059	0.034	0.000	0.000	0.068	0.036	0.000	0.000

Results use a single training seed, an 80-epoch schedule, and the fixed final checkpoint on 2,000 test images. Bold and underlining mark the highest and second-highest distinct values per target and column; ties share the marking. Worst mean is the minimum race-subgroup mean Dice. Worst tail 10% is the minimum, across race subgroups, of the mean Dice among the $\lceil 0.1|I_g| \rceil$ lowest-scoring patients in subgroup g. Patients are ranked separately for each radius, subgroup, and target.

Across the tested radii from 0.01 to 0.25, overall Dice varies little (0.814–0.818 for rim; 0.859–0.863 for cup), while subgroup scores and lower tails vary more. Increasing ρ from 0.10 to 0.15 raises the worst-subgroup lowest-10% mean Dice by 6.2 and 2.6 percentage points, respectively. This pattern motivates assessing radius choices with subgroup performance and patient lower tails alongside overall Dice.