This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

A Scientific Supplement

This supplement records the task contract, cohort construction, candidate projection, federated optimization, and supporting results used in the main paper. All tables report frozen experiment artifacts; no additional model selection was performed for the supplement.

A.1 Scope and Limitations

FedPref is evaluated in a controlled six-site simulation derived from one public dataset. The design varies site capacity and disease prevalence while holding the ontology and annotation rule fixed. The evidence therefore does not cover institution-specific reporting styles, annotation policies, clinical workflows, or external domain shift; these questions require genuinely independent institutional cohorts.

The locked test contains eight reports in three normalized-text duplicate groups; one four-report group has inconsistent targets, introducing a small source of evaluation ambiguity.

The Central, FedPref, and Local arms use the same retained preference pairs and matched pair exposure. Their comparison estimates the effect of cross-site adapter aggregation within the chosen-response SFT plus fixed-reference DPO pipeline. The study does not include a matched Central/Federated/Local SFT control on canonical annotations, alternative preference objectives, or systematic sensitivity analyses over teacher composition, scoring weights, pair margin, DPO strength, and round count. The candidate-pool diagnostics establish available response contrast, but their distinct cohorts and candidate inventories preclude attributing downstream gains to teacher heterogeneity alone.

FedPref applies pair-count-weighted, factor-wise averaging to complete LoRA states. Exact low-rank aggregation methods such as FedEx-LoRA are not compared; factor-wise averaging need not equal averaging the clients' effective updates. The protocol keeps reports and annotations local but provides no formal privacy guarantee: communicated updates are not protected by secure aggregation or differential privacy, and leakage attacks were not evaluated. Finally, the six-client, all-participation, three-round study reports 7.81 GiB of cumulative adapter traffic but does not establish runtime or network scaling under larger or partially participating federations. Direct comparisons with other federated preference methods remain future work.

A.2 Task Contract and Cohort Support

Output contract. The model receives a radiology report and returns exactly one JSON object with two keys: `diseases`, a list of supported diseases, and `findings`, a list of disease–location objects. The semantic prompt applied the following rules to every target and teacher model:

1. emit only the two required keys and no prose or Markdown;

2. include positive, explicitly proposed, or hedged findings, while excluding findings that are only negated or absent;
3. represent each localized finding as exactly one disease and one location from the frozen vocabularies in Table 3;
4. include every localized disease in the disease list, while permitting a supported disease without a location when the report provides none;
5. expand each right- or left-sided subregion to its corresponding whole-lung parent, without inferring a subregion from a whole-lung statement;
6. deduplicate and sort diseases and findings in canonical vocabulary order; and
7. emit empty lists when the report supports no target finding.

The system message instructed the model to perform careful radiology extraction, return one JSON object, use only the supplied vocabulary, and avoid unsupported inference.

Table 3. Frozen output vocabularies in canonical order.

Diseases	Locations
lung opacity; pleural effusion; atelectasis; enlarged cardiac silhouette; pulmonary edema/hazy opacity; pneumothorax; consolidation; fluid overload/heart failure; pneumonia	right lung; right apical zone; right upper lung zone; right mid lung zone; right lower lung zone; right hilar structures; right costophrenic angle; left lung; left apical zone; left upper lung zone; left mid lung zone; left lower lung zone; left hilar structures; left costophrenic angle; mediastinum; upper mediastinum; cardiac silhouette; trachea

Cohort integrity. Table 4 reports location support for the patient-disjoint cohorts defined in Sec. 4 of the main paper; gold-standard patients were excluded from the automatically labeled training and development data. Mediastinum, upper mediastinum, and trachea have zero positive support because none of the nine target diseases localizes there in these annotations; primary Location+Disease evaluation is therefore supported over 15 regions.

A.3 Candidate Generation, Projection, and Pair Construction

Generation. Each frozen teacher generated one completion per report through its native chat template. Decoding was greedy with a 768-token output limit, BF16 weights, and no quantization or retrieval augmentation. Thinking was explicitly disabled for Qwen3 models. Exact model revisions are recorded in the public reproduction package[3].

[3] https://github.com/flint-xf-fan/FedPref.

Table 4. Report-level positive support by location. Training and development use automatically derived labels; diagnostic and test use gold-standard labels.

Location	Training ($n = 4{,}800$)	Dev. ($n = 600$)	Diagnostic ($n = 100$)	Test ($n = 400$)
Right lung	2,162	289	53	203
Right apical zone	173	26	3	15
Right upper lung zone	163	26	3	19
Right mid lung zone	416	57	8	29
Right lower lung zone	1,149	148	35	108
Right hilar structures	875	120	23	102
Right costophrenic angle	628	80	15	62
Left lung	2,263	304	54	206
Left apical zone	162	22	2	13
Left upper lung zone	128	14	3	10
Left mid lung zone	421	47	6	35
Left lower lung zone	1,362	168	32	122
Left hilar structures	847	120	26	97
Left costophrenic angle	703	92	14	62
Mediastinum	0	0	0	0
Upper mediastinum	0	0	0	0
Cardiac silhouette	950	111	27	101
Trachea	0	0	0	0

Table 5. Teacher-output projection audit. "Strict" denotes a valid raw JSON object before projection; "eligible" denotes successful canonical projection.

	Training (4,800 outputs per teacher)					Development (600 per teacher)		
Teacher	Strict	Eligible	Discarded	Chosen	Rejected	Strict	Eligible	Discarded
Llama-3.1-8B	3,650	4,771	29	815	1,219	440	598	2
Mistral-7B	3,375	4,672	128	606	1,350	414	583	17
Qwen2.5-14B	4,715	4,799	1	1,061	244	593	600	0
Qwen3-14B	4,697	4,800	0	479	148	590	600	0

Label-independent projection. Projection operated on teacher text without opening the reference annotation. It extracted the first brace-balanced JSON object, rejected duplicate JSON keys and nonstandard constants, and required both top-level lists. Extra top-level fields were dropped and audited. Unknown vocabulary items and malformed finding entries were dropped without synonym or fuzzy matching. Recognized finding diseases were inserted into the disease list; subregions received their whole-lung parents; and the result was deduplicated and canonically sorted. An output with no valid object received the canonical empty fallback and was ineligible for oracle or pair construction. Table 5 reports the resulting audit.

Preference-pair eligibility. For each report, the highest- and lowest-scoring projected teachers formed at most one pair when their canonical structures differed, the composite-score margin was at least 0.10, and the chosen candidate had strictly higher Location+Disease F1. The training corpus retained 2,961 of 4,800

Table 6. Realized case-mix divergence against capacity-matched IID allocations. Values of D and the null 95th percentile are percentage points.

Cohort	N	Observed D	Null 95th	Ratio	p
Training reports	4,800	9.53	1.49	6.41	0.0005
Retained training pairs	2,961	8.83	1.89	4.67	0.0005
Development reports	600	10.14	4.14	2.45	0.0005
Development retained pairs	376	9.53	5.31	1.80	0.0005

reports (61.69%); the development corpus retained 376 of 600 (62.67%). Their mean score margins were 0.426 and 0.414, with medians 0.400 and 0.382.

A.4 Federation Construction and Non-IID Audit

Capacity-constrained multilabel allocation. All reports were assigned to sites before pair scoring. Site capacities were fixed at 28%, 22%, 18%, 14%, 10%, and 8% of each cohort. A multilabel Dirichlet profile with $\alpha = 0.5$ and skew weight 0.7 allocated intact report rows. Development reused the label profiles derived from training, with independent draw seeds 20260713 and 20260714. The realized site support is reported in Table 2 of the main paper.

Heterogeneity statistic. For site k, let n_k be cohort size, $p_{k\ell}$ the prevalence of label ℓ, and p_ℓ its global prevalence. We used nine disease indicators plus a normal-report indicator ($L = 10$) and measured

$$D = \sqrt{\frac{1}{NL} \sum_k n_k \sum_{\ell=1}^{L} (p_{k\ell} - p_\ell)^2}, \qquad N = \sum_k n_k. \tag{4}$$

The null distribution came from 1,999 capacity-preserving permutations of intact multilabel report rows. The reported Monte Carlo probability uses the plus-one correction. This audit was outcome-blind and completed before training. Table 6 reports all four cohorts; retained training pairs and development reports are the training and evaluation cohorts used for the primary study.

Table 7 reports disease prevalence by site among retained training pairs.

A.5 Optimization Details

Adapter and training settings. All arms used LoRA rank 16 with $\alpha = 32$, zero dropout, and no trainable bias, targeting the `q`, `k`, `v`, `o`, `gate`, `up`, and `down` projections. BF16 training used length 2,048, fused AdamW, cosine decay, 0.03 warmup, zero weight decay, gradient checkpointing, no packing or shuffling, and fail-closed truncation. Physical batch size was one, with gradient accumulation 24 for Central and four per client. SFT used one epoch at 2×10^{-4}; each of three DPO rounds used one epoch at 5×10^{-6}, $\beta = 0.1$. Seeds were 20260713, 20260714, and 20260715.

Table 7. Disease prevalence (%) by site among retained training pairs. A report may contain multiple diseases, so columns need not sum to 100.

Label	S0	S1	S2	S3	S4	S5	Global
Normal	34.3	43.1	23.6	7.2	24.2	6.7	25.7
Atelectasis	24.5	35.5	41.2	48.8	26.6	43.0	36.0
Consolidation	4.5	4.1	8.5	10.7	11.1	6.4	7.0
Enlarged cardiac silhouette	13.0	18.3	23.0	35.8	41.4	62.7	27.8
Fluid overload/heart failure	0.8	2.7	2.5	3.1	3.0	5.4	2.6
Lung opacity	61.5	50.8	73.4	89.5	62.6	72.3	67.4
Pleural effusion	15.8	15.6	21.4	35.2	29.0	27.1	22.5
Pneumonia	14.5	9.1	19.1	24.9	19.2	23.2	17.4
Pneumothorax	2.5	1.4	1.6	3.3	1.0	3.8	2.2
Pulmonary edema/hazy opacity	5.2	7.6	16.5	23.3	8.4	17.8	12.4

Matched exposure and checkpoint lineage. Every pair received three DPO exposures in Central, FedPref, and Local. Round 1 continued from the corresponding round-0 SFT root; rounds 2 and 3 continued from the preceding policy; FedPref clients received the previous global policy. The SFT root remained the fixed DPO reference. Central opened the exact six-shard union, while FedPref and Local used the same local shards and differed only by aggregation. The server averaged complete LoRA states in float32 with pair-count weights.

A.6 Supporting Results

Development trajectories. Table 8 gives all prespecified communication rounds. FedPref's client mean passes Local at round 2, while its worst-site advantage widens from 0.36 points at round 0 to 9.10 points at round 3. The monotonic FedPref trajectory supports the interpretation that cross-site evidence accumulates through aggregation.

Table 8. Development trajectories. Values are mean ± sample standard deviation over three seeds; all metrics are percentages.

Round	System	Client-mean $F_1^{\text{loc+dis}}$	Worst-site $F_1^{\text{loc+dis}}$	Preference accuracy
0	Central	57.85 ± 0.14	55.72 ± 0.11	86.70 ± 0.27
	FedPref	54.93 ± 0.47	53.67 ± 0.18	82.00 ± 0.41
	Local	55.73 ± 0.37	53.32 ± 1.30	83.95 ± 0.15
1	Central	69.34 ± 0.07	67.45 ± 0.37	90.51 ± 0.15
	FedPref	61.24 ± 0.30	59.18 ± 0.47	86.44 ± 0.27
	Local	62.29 ± 0.13	53.80 ± 0.06	87.85 ± 0.31
2	Central	71.21 ± 0.11	69.49 ± 0.37	94.15 ± 0.00
	FedPref	66.56 ± 0.52	65.03 ± 0.47	89.63 ± 0.00
	Local	65.55 ± 0.26	54.84 ± 0.35	89.98 ± 0.31
3	Central	71.77 ± 0.65	70.03 ± 1.15	94.95 ± 0.27
	FedPref	69.11 ± 0.17	68.12 ± 0.22	91.49 ± 0.27
	Local	66.62 ± 0.32	59.02 ± 1.45	91.67 ± 0.85