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.

Supplementary Material for: A Resource-Constrained Evaluation of Lightweight Adaptation for Cross-Dataset Chest X-ray Report Generation

Minji Kim[1 [0009−0003−1514−5210]], Seun Kim[1 [0009−0009−4541−8105]], and Jang-Hwan Choi[1,2 [0000−0001−9273−034X]]

[1] Department of Artificial Intelligence, Ewha Womans University, Seoul, South Korea {2277039, se0102s, choij}@ewha.ac.kr
[2] Department of Artificial Intelligence and Software, Ewha Womans University, Seoul, South Korea

Table S1. Cluster distribution for MetaVIB training.

Cluster	Number of Samples
0	1
1	47
2	95
3	109
4	78
5	111
6	63
7	40

Table S2. Change in qualitative scores relative to baseline (Δ = method − baseline). Positive values indicate improvement over baseline. For omission and normal template bias, positive Δ denotes fewer omissions or less bias relative to baseline.

Model	ΔSemantic	ΔOmission	ΔMajor omis.	ΔNormal bias	ΔSpecificity
Vision–decoder LoRA	+0.06	+0.08	+0.08	+0.10	+0.02
SWAD-inspired LoRA	+0.04	+0.10	+0.06	+0.10	−0.02
KD–ASL LoRA	−0.10	−0.04	−0.02	−0.08	−0.10
Vision-only LoRA	−0.18	−0.10	−0.06	−0.12	−0.06
MoME-TTA	−0.02	−0.02	0.00	0.00	0.00
MetaVIB	−0.02	−0.06	−0.04	+0.08	0.00

Table S3. LLM-assisted qualitative evaluation scores. Mean scores across 50 randomly sampled CheXpert test cases (random seed: 42). All seven models were evaluated on identical reference reports. Omission and normal template bias scores reflect mean Likert-scale ratings; major omission rate, false positive rate, and strong normal template bias rate are proportions (0–1). Higher semantic alignment, abnormality alignment, and finding specificity indicate better performance; higher omission, false positive, and normal template bias indicate worse performance.

Model	Semantic alignment	Omission	Major omission rate	False pos. FP rate		Abnorm. alignment	Normal bias	Strong bias rate	Specificity
Baseline	2.18	2.18	0.46	0.52	0.38	1.02	1.22	0.42	0.90
Vision–decoder LoRA	2.24	2.10	0.38	0.48	0.38	1.06	1.12	0.40	0.92
SWAD-inspired LoRA	2.22	2.08	0.40	0.50	0.36	1.00	1.12	0.42	0.88
KD–ASL LoRA	2.08	2.22	0.48	0.44	0.34	0.98	1.30	0.46	0.80
Vision-only LoRA	2.00	2.28	0.52	0.60	0.48	0.94	1.34	0.50	0.84
MoME-TTA	2.16	2.20	0.46	0.52	0.38	1.02	1.22	0.42	0.90
MetaVIB	2.16	2.24	0.50	0.58	0.40	1.02	1.14	0.38	0.90

Prompt Used for LLM-Assisted Qualitative Evaluation

The following prompt was used for the LLM-assisted qualitative evaluation (ChatGPT, OpenAI, GPT-5.5 Thinking; accessed May 2026). Each candidate report was anonymized and evaluated only against the corresponding reference report.

```
You are performing LLM-assisted semantic evaluation of generated chest X-ray reports.

Important constraints:
- You are NOT evaluating the actual image.
- Use only the reference_text as the ground-truth report.
- Compare each candidate report against the reference_text.
- Focus on clinically relevant semantic agreement.
- Do not reward generic normal templates when the reference contains abnormal findings.
- Penalize clinically important omissions.
- Penalize false-positive disease mentions that are not supported by the reference.
- Do not infer findings from the image path.
- Return valid JSON only.

For each candidate, output:
{
  "case_id": "...",
  "candidate_id": "...",
  "semantic_alignment_score": 1-5,
  "clinically_important_omission": 0-3,
  "false_positive_finding": 0-3,
  "abnormality_alignment": 0-2,
  "normal_template_bias": 0-2,
  "finding_specificity": 0-3,
  "short_reason": "one concise sentence"
}

Scoring definitions:

semantic_alignment_score:
1 = very poor semantic alignment with the reference
2 = poor semantic alignment
3 = partially aligned
4 = mostly aligned
5 = highly aligned with the reference

clinically_important_omission:
0 = no important omission
1 = minor omission
2 = moderate omission
3 = major clinically relevant omission

false_positive_finding:
0 = no false-positive finding
1 = minor false-positive finding
2 = moderate false-positive finding
3 = major false-positive finding

abnormality_alignment:
0 = wrong normal/abnormal status
1 = partially correct or ambiguous
2 = correct normal/abnormal status

normal_template_bias:
0 = no generic normal-template bias
1 = mild generic normal wording
2 = strong normal-template bias, especially if the reference contains abnormal findings

finding_specificity:
0 = no useful finding specificity
1 = vague or generic
```

```
2 = partially specific
3 = specific and clinically informative

Return a JSON array only.
```