LOTUS: Language-Vision Optimal Transport for Unified Report Synthesis

Xuhui Li*
Key Laboratory of System Software,
Institute of Software,
Chinese Academy of Sciences

Jiachen Liu*
Northwestern Polytechnical University

*Equal contribution Correspondence: lixuhui25@mails.ucas.ac.cn

Abstract

Automated radiology report generation can reduce reporting burden, but existing encoder-decoder systems still rely on soft cross-attention that offers no geometric guarantee that generated tokens consume the correct visual evidence. We present *LOTUS*, an optimal-transport framework that places alignment supervision directly at the decoder evidence interface. LOTUS introduces Cross-Attention OT (CA-OT) to align decoder token states with visual patches, and extends it with Prototype OT (Proto-OT) for structure-level alignment between visual and textual codebooks. Experiments on IU X-Ray and ROCO, complemented by a scalability study on MIMIC-CXR-RRG, show that decoder-side OT improves macro-level clinical-label agreement while maintaining competitive language quality. In multi-seed IU X-Ray evaluation, LOTUS-Deep improves macro clinical F1 from 0.162 to 0.201 over the CE baseline, and on ROCO it yields micro-F1 gains on a thoracic-label proxy (0.395 to 0.600). We also provide a population-level risk-gap bound that justifies expected OT regularization as a principled cross-modal training signal.

1 Introduction

Automated radiology report generation can substantially reduce reporting burden and improve clinical workflow, but accurate grounding remains difficult. Modern encoder-decoder systems often generate fluent yet anatomically inconsistent findings because local next-token supervision does not explicitly penalize cross-modal grounding errors. We propose *LOTUS* (Language-vision Optimal Transport for Unified report Synthesis)[1] to address this problem with optimal-transport supervision at the decoder evidence interface.

Recent progress in Transformer-based report generation (Chen et al., 2020; Wang et al., 2023a; You et al., 2021) has improved cross-modal interaction through attention mechanisms and memory-driven decoding, while prompt-based (Jin et al., 2024) and foundation-model approaches (Liu et al., 2024a,b) further improve fluency and clinical coherence. Nevertheless, these models still rely on *soft cross-attention* as the dominant alignment mechanism. In practice, this leads to two recurring failure modes: attention can diffuse over multiple irrelevant regions, and the decoder can substitute a clinically plausible but spatially incorrect term. Because cross-entropy only supervises token identity, such grounding errors are weakly constrained during training.

We argue that the right place to enforce cross-modal geometry is not a generic encoder-side representation space, but the decoder cross-attention interface where each generated token actively consumes visual evidence. This observation leads to a transport view of generation: token states and visual patches should be matched by a transport plan whose marginals are induced by the decoder's own attention behavior.

To realize this idea, LOTUS introduces **Cross-Attention OT (CA-OT)**, which applies entropic OT directly between decoder token states and visual patches, and **Prototype OT (Proto-OT)**, which refines CA-OT with structure-level alignment between learnable visual and textual codebooks. The resulting DeepOT variant improves local grounding and distributional consistency without imposing heavy encoder-side regularization.

We evaluate LOTUS on IU X-Ray and ROCO with both language metrics and CheXpert-style clinical efficacy, and additionally on the larger MIMIC-CXR-RRG benchmark for scalability. The results show that decoder-side OT improves macro-level clinical-label agreement on structured radiology reports, while prototype refinement recovers

[1] Inspired by the lotus flower "rising unstained from the mud," our model aims to distill reliable clinical semantics from noisy medical imagery.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 1746–1762
November 6-10, 2026 ©2026 Association for Computational Linguistics

part of the high-frequency recall sacrificed by CA-OT alone. This yields a clearer picture of when OT helps, when it creates trade-offs, and why hierarchical alignment is needed.

In summary, the main contributions of this work are three-fold:

- **Decoder-side OT alignment.** We introduce CA-OT at the decoder cross-attention interface, yielding stronger cross-modal alignment and improving multi-seed mean IU X-Ray BLEU-4 from 0.138 to 0.148 over our CE baseline.

- **Hierarchical DeepOT refinement.** We extend CA-OT with Proto-OT over learned codebooks, improving multi-seed mean IU X-Ray macro clinical F1 from 0.162 to 0.201 (+24%) and ROCO thoracic-proxy micro-F1 from 0.395 to 0.600. A scalability study on MIMIC-CXR-RRG confirms that the gains extend to a larger MIMIC-style benchmark.

- **Principled analysis and honest empirical characterization.** We show that expected CA-OT upper-bounds a population-level cross-modal risk gap, and we analyze the important trade-off that CA-OT alone can reduce IU X-Ray micro-F1 on frequent findings.

2 Related works

2.1 Automated Radiology Report Generation

Automated radiology report generation has evolved significantly from early CNN-RNN frameworks to modern Transformer-based architectures. Foundational works like R2Gen (Chen et al., 2020) pioneered the use of memory-driven mechanisms to enhance feature representation during decoding. To refine cross-modal interactions, subsequent studies explored hierarchical alignment strategies (You et al., 2021) and multi-expert token architectures (Wang et al., 2023a), aiming to better capture the correspondence between visual regions and textual semantics. Further improvements were made by optimizing visual feature consistency through graph-guided approaches (Zhang et al., 2024), which addressed the reliability of extracted image features. Recently, paradigm shifts have emerged incorporating prompt learning (Jin et al., 2024) and leveraging foundation models (Liu et al., 2024a,b) to improve clinical coherence and generalization.

More recently, large multimodal language models have been applied to radiology report generation. LLaVA-Rad (Zambrano Chaves et al., 2025) demonstrated that a compact multimodal LLM fine-tuned on clinical data can achieve strong performance on chest X-ray report generation. R2GenGPT (Wang et al., 2023b) integrates GPT-style decoding with visual features for improved report fluency. MedGemma (Sellergren et al., 2025) extends Google's Gemma architecture to the medical domain with broad multimodal capabilities. These foundation-model approaches operate in a fundamentally different resource regime—relying on billion-parameter pretrained LLMs and large-scale clinical pretraining data—compared to our compact, trainable encoder-decoder framework. Our contribution instead targets the setting where a lightweight model is trained directly on standard report-generation benchmarks.

Nevertheless, beneath these architectural innovations, the core alignment mechanism in encoder-decoder models remains largely unchanged. The prevailing reliance on similarity-based attention weights limits the ability to enforce structured matching constraints. Such probabilistic associations do not explicitly minimize the distance between feature distributions, suggesting a need for more mathematically rigorous alignment formulations.

2.2 Cross-Modal Alignment and Medical Vision-Language Pre-training

Effective visual-textual alignment is central to multi-modal medical AI. In the general domain, recent surveys reveal how representational misalignments manifest across object, attribute, and relational levels (Shu et al., 2025), motivating more principled alignment objectives. In the medical domain, contrastive pre-training paradigms have made significant strides: GLoRIA (Huang et al., 2021) demonstrated that contrasting image sub-regions with words yields label-efficient medical representations; MedCLIP (Wang et al., 2022) extended this to unpaired image-text learning; and BiomedCLIP (Zhang et al., 2023) scaled contrastive pre-training to fifteen million biomedical pairs.

However, these VLP frameworks optimize for *representation quality* rather than *generation-time structural correspondence*. Concurrently, OT-based methods have addressed distributional discrepancy in domain adaptation (Zhu and Qi, 2023),

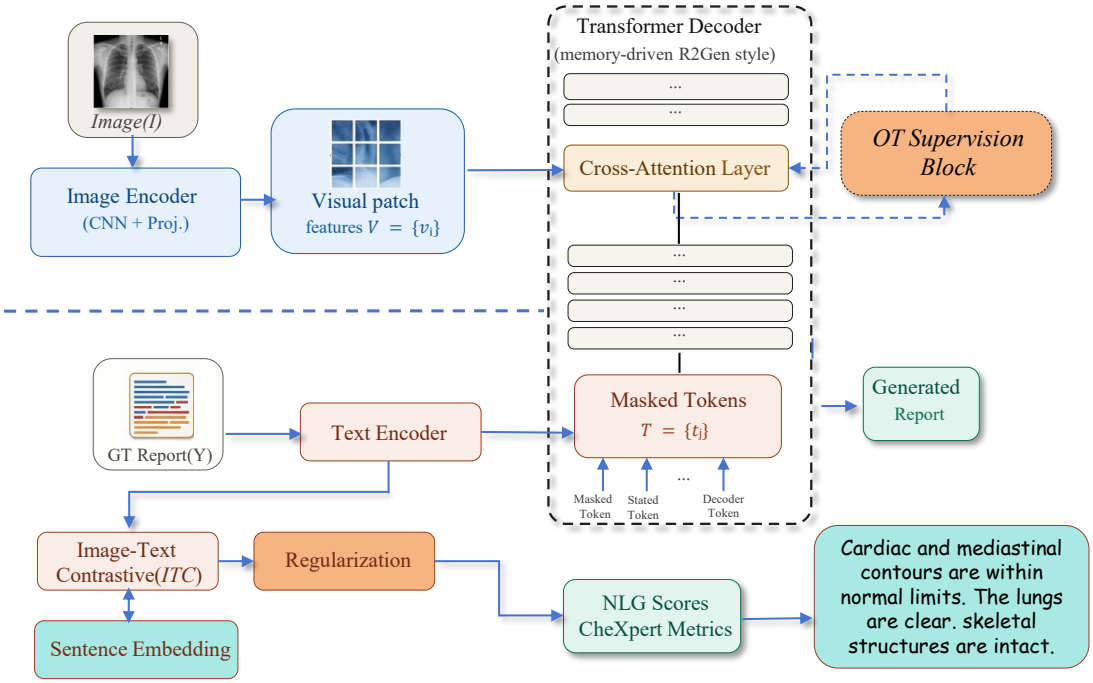


Figure 1: Overview of the LOTUS framework. The system takes medical images as input and generates radiology reports through a vision-language encoder-decoder architecture enhanced with Optimal Transport-based cross-modal alignment.

yet remain confined to post-hoc feature alignment. Our work is the first to integrate OT directly into the sequence generation decoder, enforcing structured visual-textual transport at every generation step.

2.3 Optimal Transport in Deep Learning

Optimal Transport (OT) has emerged as a powerful mathematical tool for comparing probability distributions in machine learning. The seminal work by Cuturi introduced the Sinkhorn algorithm, making OT computationally feasible for deep learning applications (Cuturi, 2013). In computer vision, OT has been applied to domain adaptation (Courty et al., 2017) and generative adversarial networks (e.g., Wasserstein GAN) (Arjovsky et al., 2017). In natural language processing, OT has been used for document similarity and word embedding alignment (Kusner et al., 2015). Recently, there have been attempts to apply OT to vision-language tasks, such as image-text matching (Chen et al., 2023). However, the application of OT to *structured* report generation, particularly for refining embeddings and guiding the Transformer architecture simultaneously, remains under-explored. Our work bridges this gap by integrating OT directly into the generation pipeline to ensure robust semantic alignment.

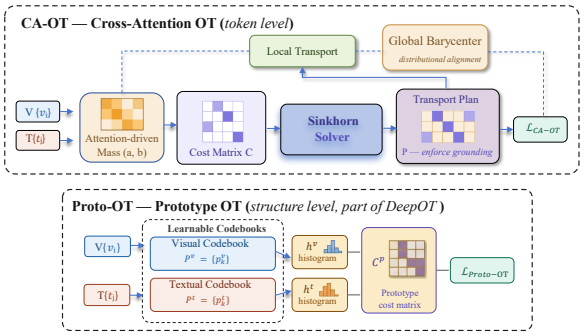


Figure 2: Detailed illustration of the two Optimal Transport modules in LOTUS. Top: Cross-Attention OT (CA-OT) aligns decoder tokens with visual regions at the cross-attention interface. Bottom: Prototype OT (Proto-OT) performs structure-level alignment between visual and textual semantic codebooks.

3 The Proposed Framework

Our framework builds upon the memory-driven R2Gen architecture (Chen et al., 2020) and introduces two complementary OT-based components: (1) **Cross-Attention OT (CA-OT)**, which applies transport supervision at the decoder token–region interaction interface, and (2) **DeepOT** with **Prototype OT (Proto-OT)**, which further enforces structure-level alignment between visual and textual semantic codebooks. Figure 1 and Figure 2 illustrates the overall framework.

Symbol	Meaning
$\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^{N_v}$	Image patch features, with $\mathbf{v}_i \in \mathbb{R}^d$
$\mathcal{T} = \{\mathbf{t}_j\}_{j=1}^{N_t}$	Decoder hidden states for valid report tokens, with $\mathbf{t}_j \in \mathbb{R}^d$
$y_{1:L}$	Ground-truth report token sequence of length L
$\mathbf{a} \in \Delta^{N_v-1}$	Source marginal over image patches for CA-OT
$\mathbf{b} \in \Delta^{N_t-1}$	Target marginal over decoder token states for CA-OT
$\mathbf{C} \in \mathbb{R}^{N_v \times N_t}$	Pairwise transport cost matrix between patches and token states
$\mathbf{P} \in \mathbb{R}^{N_v \times N_t}_{\geq 0}$	Transport plan for CA-OT
ε	Entropic OT regularization coefficient
$\mathcal{P}^v, \mathcal{P}^t$	Visual and textual prototype codebooks in Proto-OT

Table 1: Core notation used in the proposed framework.

3.1 Problem Formulation and Notation

Given a dataset $\mathcal{D} = \{(x_i, y_{i,1:L_i})\}_{i=1}^N$ of images and reports, the model predicts each token autoregressively as $p(y_t \mid y_{<t}, x_i)$. The image encoder produces a set of patch features $\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^{N_v}$ with $\mathbf{v}_i \in \mathbb{R}^d$, and the decoder produces hidden states $\mathcal{T} = \{\mathbf{t}_j\}_{j=1}^{N_t}$ with $\mathbf{t}_j \in \mathbb{R}^d$ for the valid report tokens. We reserve L for report length throughout the paper and use $y_{1:L}$ for token sequences, avoiding any overload between text tokens and decoder states. LOTUS adds OT supervision to this standard encoder-decoder pipeline without introducing a separate pretraining stage.

Table 1 summarizes the main symbols used throughout Section 3.

3.2 Cross-Attention OT (CA-OT)

Motivation. Conventional OT-based report generation aligns visual and text features at a generic representation level. This may over-regularize language modeling, leading to template collapse. We propose moving OT supervision to the decoder cross-attention interface, where each generated token explicitly consumes visual evidence.

Cross-Attention-Induced Marginals. We construct the source marginal $\mathbf{a} \in \Delta^{N_v-1}$ and target marginal $\mathbf{b} \in \Delta^{N_t-1}$ from decoder cross-attention over H heads and the token-validity mask $m_j \in \{0, 1\}$:

$$a_i = \frac{\tilde{a}_i}{\sum_{i'} \tilde{a}_{i'}}, \qquad \tilde{a}_i = \frac{1}{H} \sum_{h=1}^{H} \sum_{j=1}^{N_t} m_j A_{j,i}^{(h)}, \quad b_j = \frac{m_j}{\sum_{j'} m_{j'}}. \tag{1}$$

where $A^{(h)} \in \mathbb{R}^{N_t \times N_v}$ is the cross-attention map at head h with $\sum_i A_{j,i}^{(h)} = 1$ (row-softmax). Intuitively, a_i aggregates the cross-attention mass that patch i receives from all valid (non-padded) decoder tokens; explicit normalization by $\sum_{i'} \tilde{a}_{i'}$ ensures $\mathbf{a} \in \Delta^{N_v-1}$. Symmetrically, b_j is the uniform distribution over valid tokens, giving $\mathbf{b} \in \Delta^{N_t-1}$.

Local Transport. We define a pairwise cost matrix $\mathbf{C} \in \mathbb{R}^{N_v \times N_t}$ where $C_{ij} = c(\mathbf{v}_i, \mathbf{t}_j)$. The cost function $c(\cdot, \cdot)$ is chosen as either normalized squared Euclidean distance or cosine dissimilarity. The entropic regularized OT loss is formulated as:

$$\mathcal{L}_{\text{ca-local}} = \min_{\mathbf{P} \in \Pi(\mathbf{a}, \mathbf{b})} \langle \mathbf{P}, \mathbf{C} \rangle - \varepsilon H(\mathbf{P}), \tag{2}$$

with $H(\mathbf{P}) = -\sum_{i,j} P_{ij} \log P_{ij}$ and $\Pi(\mathbf{a}, \mathbf{b}) = \{\mathbf{P} \geq 0 \mid \mathbf{P}\mathbf{1} = \mathbf{a},\ \mathbf{P}^\top \mathbf{1} = \mathbf{b}\}$, following the standard minimization convention of Cuturi (2013): $\varepsilon > 0$ controls entropic smoothing (a larger ε yields a higher-entropy, smoother transport plan), and as $\varepsilon \to 0$ this recovers the exact 1-Wasserstein distance $W_1(\mu_i, \nu_i)$. We write the objective as $\langle \mathbf{P}, \mathbf{C} \rangle - \varepsilon H(\mathbf{P})$ throughout; the equivalent form $\langle \mathbf{P}, \mathbf{C} \rangle + \varepsilon \sum_{i,j} P_{ij}(\log P_{ij} - 1)$ used by our implementation differs only by the additive constant $-\varepsilon$ and induces the same optimizer. We solve Eq. (2) via Sinkhorn iterations (ε=0.1, 30 iters). We ablate the choice of cost function $c(\cdot, \cdot)$ (squared ℓ_2 vs. cosine) in Appendix B.10.

Global Consistency. To ensure global semantic alignment, we compute barycenters $\bar{\mathbf{v}} = \sum_i a_i \mathbf{v}_i$ and $\bar{\mathbf{t}} = \sum_j b_j \mathbf{t}_j$. The total CA-OT objective is a weighted sum of the local transport loss and the global loss $c(\bar{\mathbf{v}}, \bar{\mathbf{t}})$:

$$\mathcal{L}_{\text{CA-OT}} = \lambda_{\text{ca}}^{\text{loc}} \mathcal{L}_{\text{ca-local}} + \lambda_{\text{ca}}^{\text{glo}} \|\bar{\mathbf{v}} - \bar{\mathbf{t}}\|_2^2, \qquad \lambda_{\text{ca}}^{\text{loc}} + \lambda_{\text{ca}}^{\text{glo}} = 1. \tag{3}$$

This placement enforces token-level grounding at the decoder evidence interface while preserving generation semantics.

3.3 Deep OT Refinement via Prototypes

Motivation. CA-OT improves local grounding but aligns only sample-level token/patch sets. To model higher-order semantic geometry, we introduce prototype-level transport between visual and textual latent distributions.

Learnable Prototypes. We maintain visual and textual codebooks $\mathcal{P}^v = \{\mathbf{p}_k^v\}_{k=1}^K$ and $\mathcal{P}^t = \{\mathbf{p}_k^t\}_{k=1}^K$, where $\mathbf{p}_k \in \mathbb{R}^d$. Soft assignments are

computed with temperature τ_p:

$$q_{ik}^v = \frac{\exp(\mathbf{v}_i^\top \mathbf{p}_k^v/\tau_p)}{\sum_{k'} \exp(\mathbf{v}_i^\top \mathbf{p}_{k'}^v/\tau_p)}, \quad q_{jk}^t = \frac{\exp(\mathbf{t}_j^\top \mathbf{p}_k^t/\tau_p)}{\sum_{k'} \exp(\mathbf{t}_j^\top \mathbf{p}_{k'}^t/\tau_p)}. \tag{4}$$

From assignments and masses $(\mathbf{a}, \mathbf{b})$, we construct histogram distributions $\mathbf{h}^v, \mathbf{h}^t \in \Delta^{K-1}$:

$$h_k^v = \sum_{i=1}^{N_v} a_i q_{ik}^v, \qquad h_k^t = \sum_{j=1}^{N_t} b_j q_{jk}^t. \tag{5}$$

Since $\sum_k q_{ik}^v = 1$ (softmax) and $\sum_i a_i = 1$, we have $\sum_k h_k^v = 1$; similarly $\sum_k h_k^t = 1$. Hence $\mathbf{h}^v, \mathbf{h}^t \in \Delta^{K-1}$ are valid probability vectors over the K prototypes.

Prototype Transport. Defining prototype cost $C_{k\ell}^p = c(\mathbf{p}_k^v, \mathbf{p}_\ell^t)$, the entropic OT on the prototype space is $\mathcal{L}_{\text{p-local}} = \min_{\mathbf{Q} \in \Pi(\mathbf{h}^v, \mathbf{h}^t)} \langle \mathbf{Q}, \mathbf{C}^p \rangle - \varepsilon H(\mathbf{Q})$, where $H(\mathbf{Q}) = -\sum_{k\ell} Q_{k\ell} \log Q_{k\ell}$ (the same convention as Eq. (2)). We enforce global alignment on prototype barycenters $\bar{\mathbf{p}}^v = \sum_k h_k^v \mathbf{p}_k^v$ and $\bar{\mathbf{p}}^t = \sum_k h_k^t \mathbf{p}_k^t$ (distinguished from the CA-OT barycenters $\bar{\mathbf{v}}, \bar{\mathbf{t}}$):

$$\mathcal{L}_{\text{Proto-OT}} = \lambda_{\text{po}}^{\text{loc}} \mathcal{L}_{\text{p-local}} + \lambda_{\text{po}}^{\text{glo}} \|\bar{\mathbf{p}}^v - \bar{\mathbf{p}}^t\|_2^2, \quad \lambda_{\text{po}}^{\text{loc}} + \lambda_{\text{po}}^{\text{glo}} = 1. \tag{6}$$

CA-OT handles instance-level token/patch alignment; Proto-OT enforces structure-level distributional alignment via codebook transport, constituting the DeepOT refinement. In our implementation Proto-OT is not a standalone replacement for CA-OT: it reuses the decoder-side token states and attention-derived masses from CA-OT, so we treat it as a refinement layer built on top of the local alignment mechanism.

3.4 Unified Training Objective

Language Modeling. For a report of length L with tokens $y_{1:L}$, the standard cross-entropy loss is:

$$\mathcal{L}_{\text{LM}} = -\sum_{t=1}^{L} \log p(y_t \mid y_{<t}, \mathbf{V}). \tag{7}$$

Image-Text Contrastive Regularization. Using normalized globals $(\bar{\mathbf{v}}, \bar{\mathbf{t}})$, we compute batch similarity $S_{ab} = \bar{\mathbf{v}}_a^\top \bar{\mathbf{t}}_b / \tau_c$. We apply symmetric InfoNCE:

$$\mathcal{L}_{\text{ITC}} = \frac{1}{2}\big(\text{CE}(\mathbf{S}, \mathbf{y}) + \text{CE}(\mathbf{S}^\top, \mathbf{y})\big), \tag{8}$$

where $\mathbf{y}$ denotes the diagonal matching targets $(y_a = a)$.

Anti-Collapse Regularization. To prevent vocabulary collapse, let $p_{bt}(w)$ be the token distribution at sample b, step t. We define the batch-average distribution $\bar{p}(w)$. Note that $\mathcal{L}_{\text{ent}} = \sum_w \bar{p}(w) \log \bar{p}(w) = -H(\bar{p})$ is the *negative* entropy of $\bar{p}$; minimizing $\mathcal{L}_{\text{ent}}$ is therefore equivalent to *maximizing* $H(\bar{p})$, which penalizes degenerate peaky token distributions:

$$\mathcal{L}_{\text{ent}} = \sum_w \bar{p}(w) \log \bar{p}(w). \tag{9}$$

Additionally, we apply batch repulsion on sentence embeddings to encourage diversity:

$$\mathcal{L}_{\text{rep}} = \frac{1}{B(B-1)} \sum_{a \neq b} \cos^2(\bar{\mathbf{t}}_a, \bar{\mathbf{t}}_b). \tag{10}$$

Final Training Objective. The unified loss combines the above terms, including a diversity term $\mathcal{L}_{\text{div}} = \alpha_{\text{ent}} \mathcal{L}_{\text{ent}} + \alpha_{\text{rep}} \mathcal{L}_{\text{rep}}$:

$$\mathcal{L} = w_{\text{lm}} \mathcal{L}_{\text{LM}} + w_{\text{ca}} \mathcal{L}_{\text{CA-OT}} + w_{\text{po}} \mathcal{L}_{\text{Proto-OT}} + w_{\text{itc}} \mathcal{L}_{\text{ITC}} + w_{\text{div}} \mathcal{L}_{\text{div}}. \tag{11}$$

To prevent OT branches from dominating early training, we replace the nominal weights $w_{\text{ca}}, w_{\text{po}}$ with *effective* weights computed per step:

$$\tilde{w}_{\text{ca}} = \min\left\{ w_{\text{ca}},\ \rho_{\text{ca}} \cdot \frac{\text{sg}(\mathcal{L}_{\text{LM}})}{\mathcal{L}_{\text{CA-OT}} + \delta} \right\}, \tag{12}$$

and analogously for $\tilde{w}_{\text{po}}$, where $\text{sg}(\cdot)$ denotes stop-gradient and $\delta = 10^{-6}$ is a numerical stability term. At each step $\tilde{w}_{\text{ca}}$ is recomputed using the current loss magnitudes, ensuring the CA-OT branch never contributes more than ρ_{ca} times the LM loss. The final objective replaces $w_{\text{ca}} \to \tilde{w}_{\text{ca}}$ and $w_{\text{po}} \to \tilde{w}_{\text{po}}$ in Eq. (11). This yields stable one-stage training while preserving clinically meaningful alignment.

3.5 Theoretical Justification

We provide a formal justification for incorporating OT regularization into the report generation objective.

For a sample $(x_i, y_{i,1:L_i})$, let μ_i and ν_i denote the empirical distributions over visual patches and decoder token states weighted by cross-attention masses, with population-level marginals $P_v = \mathbb{E}[\mu_i]$ and $P_t = \mathbb{E}[\nu_i]$.

Our key result is that minimizing the expected instance-level OT loss directly upper-bounds the

population-level cross-modal risk gap $\Delta R(\theta) = |\mathbb{E}_{z\sim P_t}[\ell(f_\theta(z), y)] - \mathbb{E}_{z\sim P_v}[\ell(f_\theta(z), y)]|$. Under standard Lipschitz assumptions:

$$\begin{aligned}\Delta R(\theta) &\le (K \cdot L) \cdot W_1(P_v, P_t) \\ &\le (K \cdot L)\big(\frac{1}{\lambda_{\text{ca}}^{\text{loc}}}\mathbb{E}[\mathcal{L}_{\text{CA-OT}}] \\ &\quad + \varepsilon \log(N_v N_t)\big).\end{aligned} \tag{13}$$

This justifies CA-OT as a principled cross-modal training signal. The complete derivation, including the entropic approximation bound (Lemma 3.1), instance-to-population bridge (Proposition 3.2), and detailed proof of the main theorem, is provided in Appendix B.

4 Experiments

In this section, we present our experimental setup, datasets, evaluation metrics, implementation details, and results.

4.1 Datasets and Evaluation Metrics

Datasets. We evaluate on two complementary datasets (statistics in Table 2): **IU X-Ray** (Demner-Fushman et al., 2016) (7,470 chest X-ray images, 3,955 reports) emphasizes detailed multi-finding clinical narratives with long-tail pathology coverage. **ROCO** (Pelka et al., 2018) (81,000+ images, 8 modalities) emphasizes concise multi-modal captioning with broad visual domain coverage.

This pairing tests both depth of clinical reasoning (IU X-Ray) and breadth of modality generalization (ROCO), revealing dataset-dependent OT behavior analyzed in Section 5.1.

We evaluate report generation quality using standard natural language generation metrics:

- **BLEU** (Papineni et al., 2002): Bilingual Evaluation Understudy scores (BLEU-1 through BLEU-4) measuring n-gram precision.
- **ROUGE-L** (Lin, 2004): Recall-Oriented Understudy for Gisting Evaluation, focusing on longest common subsequence.
- **CheXpert-style Clinical Efficacy** (Irvin et al., 2019): We map generated/reference reports to 14 thoracic finding labels with an automatic labeler and report multilabel agreement using $\text{F1}_{\text{micro}} = \frac{2\sum_\ell \text{TP}_\ell}{2\sum_\ell \text{TP}_\ell + \sum_\ell \text{FP}_\ell + \sum_\ell \text{FN}_\ell}$ and $\text{F1}_{\text{macro}} = \frac{1}{L}\sum_{\ell=1}^{L} \text{F1}_\ell$, where $L = 14$.

4.2 Experimental Setup

We evaluate three core systems: (1) **CE Baseline (R2Gen)**: the original memory-driven R2Gen trained with cross-entropy only; (2) **LOTUS-Cross (CA-OT)**: decoder-side Cross-Attention OT added to the base generator; and (3) **LOTUS-Deep (CA-OT + Proto-OT)**: the hierarchical variant that augments CA-OT with prototype-level transport. Auxiliary terms such as diversity regularization and ITC are analyzed through ablations rather than introduced as separate headline systems.

Algorithm 1 summarizes the unified optimization pipeline. In each training step, we extract visual patches and decoder states with cross-attention-derived masses, then apply OT alignment at the decoder interface (CrossAttn-OT); for OT-Deep, we add prototype-level transport. The final objective combines language modeling with OT and optional ITC/diversity regularization, while ratio-capped weighting prevents auxiliary OT branches from overwhelming the main generation loss.

Algorithm 1: One-Step Training of CrossAttn/DeepOT

Input: Batch (I, Y), parameters θ
Output: Updated θ

Forward: Extract visual patches $\mathbf{V}$, decoder states $\mathcal{T}$, and masses $(\mathbf{a}, \mathbf{b})$ from cross-attention.;
OT Branches: Compute $\mathcal{L}_{\text{CA-OT}}$; if DeepOT, compute $\mathcal{L}_{\text{Proto-OT}}$.;
Auxiliary Losses: Compute $\mathcal{L}_{\text{LM}}$, $\mathcal{L}_{\text{ITC}}$, and $\mathcal{L}_{\text{div}}$.;
Ratio-Capped Weighting: Cap OT contributions w.r.t. $\mathcal{L}_{\text{LM}}$ using preset ratios.;
Optimize: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{LM}} + \tilde{w}_{\text{ca}}\mathcal{L}_{\text{CA-OT}} + \tilde{w}_{\text{po}}\mathcal{L}_{\text{Proto-OT}} + w_{\text{itc}}\mathcal{L}_{\text{ITC}} + w_{\text{div}}\mathcal{L}_{\text{div}}$.;
  Update $\theta \leftarrow \theta - \eta\nabla_\theta\mathcal{L}_{\text{total}}$.;

Main-table IU X-Ray results are reported as multi-seed mean $\pm$ std; the CE baseline and both LOTUS variants use matched random initializations, architecture, budget, and checkpoint-selection protocol (best validation loss). ROCO LOTUS-Deep results are reported as mean $\pm$ std over completed runs; configurations with only one run are labeled as such.

4.3 Implementation Details

The model is implemented in PyTorch on an NVIDIA RTX A6000 GPU (48GB VRAM). Training requires 14–16 hours for ROCO and 4–6 hours for IU X-Ray. We use a ResNet-101 backbone (ImageNet pretrained) and R2Gen's memory-driven Transformer (3 layers, 8 heads, $d = 512$). Op-

Dataset	IU X-Ray			ROCO		
	Train	Validation	Test	Train	Validation	Test
Image	5.23K	0.75K	1.50K	65.42K	8.18K	8.18K
Report	2.77K	0.40K	0.79K	65.45K	8.18K	8.18K
Patient	2.77K	0.40K	0.79K	65.45K	8.18K	8.18K
Average Length	37.56	36.78	33.62	20.91	20.71	21.14

Table 2: Details of two benchmark datasets used in this work.

timization uses Adam with `lr_ve`$= 5 \times 10^{-5}$, `lr_ed`$= 1 \times 10^{-4}$, weight decay $= 5 \times 10^{-5}$, and gradient clipping (norm=1.0). For OT, we set $\varepsilon = 0.1$, Sinkhorn iterations $= 30$, prototype temperature $\tau_p = 0.8$, and ITC temperature $\tau_c = 0.07$. For LOTUS-Cross we use $w_{\text{ca}} = 0.05$, $(\lambda_{\text{ca}}^{\text{loc}}, \lambda_{\text{ca}}^{\text{glo}}) = (0.7, 0.3)$, and $w_{\text{itc}} = 0.02$. For LOTUS-Deep we add $w_{\text{po}} = 0.02$, $(\lambda_{\text{po}}^{\text{loc}}, \lambda_{\text{po}}^{\text{glo}}) = (0.7, 0.3)$, and $w_{\text{div}} = 0.01$, with ratio-capped terms ($\rho_{\text{ca}} = \rho_{\text{po}} = 0.5$).

Additional OT implementation details: Proto-OT uses K=32 randomly initialized, independently learned visual and textual prototypes; token states are taken from the last decoder layer; the reported results use squared ℓ_2 cost (a cosine-cost sensitivity study is given in Appendix B.10); and all OT quantities are computed in a d=256, ℓ_2-normalized projection space. Dataset preprocessing follows R2Gen, and the training scripts and evaluation code are released for reproducibility.

4.4 Comparison with State-of-the-Art Methods

Table 3 compares LOTUS-Deep against published encoder-decoder methods on IU X-Ray. LOTUS-Deep achieves consistent improvements on all NLG metrics and provides the only CheXpert-style clinical efficacy evaluation among the compared methods. Note that large-scale foundation models (e.g., LLaVA-Rad (Zambrano Chaves et al., 2025), MedGemma (Sellergren et al., 2025)) operate in a fundamentally different resource regime—relying on billion-parameter pretrained LLMs and large-scale clinical pretraining data—and are therefore not directly comparable in terms of architectural complexity and data requirements. We acknowledge this as a limitation and include these approaches in the related work discussion to clarify the scope of our contribution.

Table 5 presents a component-wise ablation on IU X-Ray. Ratio-capped weighting is the most critical component: removing it ($\rho = \infty$) causes a notable drop across all metrics, confirming that unconstrained OT branches destabilize language modeling. Proto-OT provides the largest gain beyond CA-OT, while diversity regularization offers a smaller but consistent improvement. We do not report an "only Proto-OT" system because Proto-OT is implemented as a refinement on top of the CA-OT-derived token/patch masses rather than as an independent alignment branch. Controlled non-OT baselines (CE + diversity only, CE + ITC only, and a non-OT feature-alignment variant under the same architecture, budget, and multi-seed evaluation protocol) that isolate the effect of the transport objective from auxiliary losses are reported in Appendix B.9.

Table 6 analyzes hyperparameter sensitivity. The entropic regularization parameter ε shows moderate sensitivity, with optimal performance at $\varepsilon = 0.1$. Sinkhorn iterations saturate around 30 iterations. Most critically, the ratio cap ρ is highly sensitive: removing it ($\rho = \infty$) causes BLEU-4 to drop from 0.192 to 0.176, confirming its importance for stable training.

4.5 Results and Analysis

LOTUS achieves consistent improvements over baseline methods (Tables 3 and 4). On IU X-Ray, in multi-seed evaluation, OT-Deep obtains the best mean BLEU-4 (0.170 vs. 0.138 for CE), ROUGE-L (0.371), and macro-level clinical score (F1-Ma: 0.201 vs. 0.162), indicating better balanced performance across findings including less frequent classes. The original submission's BLEU-4 of 0.192 came from one run; the multi-seed mean is 0.170 ± 0.020, still a clear and consistent improvement over the OT-Cross mean (0.148 ± 0.034). On ROCO, OT-Cross is competitive on language metrics (ROUGE-L: 0.216) while OT-Deep yields the strongest micro clinical scores (F1-Mi: 0.600 in multi-seed evaluation). Because ROCO is a multi-modality dataset and CheXpert covers only thoracic findings, ROCO CheXpert-style F1 should be in-

Method	B-1	B-4	R-L	F1-Ma	F1-Mi
AlignTrans (You et al., 2021)†	0.484	0.172	0.379	—	—
METransformer (Wang et al., 2023a)†	0.483	0.172	0.380	—	—
PromptMRG (Jin et al., 2024)†	0.401	0.098	0.281	—	—
R2GenGPT(Deep)(Wang et al., 2023b)†	0.488	0.173	0.377	—	—
R2Gen (Chen et al., 2020)†	0.470	0.165	0.371	0.141	0.447
R2GenCMN (Chen et al., 2021)†	0.470	0.170	0.375	—	—
LOTUS-Cross (ours, multi-seed mean)	0.481	0.148	0.365	0.180	0.345
LOTUS-Deep (ours, multi-seed mean)	**0.507**	0.170	0.371	**0.201**	0.405
LOTUS-Deep (ours, best run)	0.507	**0.192**	**0.401**	**0.203**	0.402

Table 3: Comparison with state-of-the-art encoder-decoder methods on IU X-Ray. †: numbers taken from published papers. F1-Ma and F1-Mi are CheXpert-style clinical efficacy metrics (higher is better). Prior papers did not report these metrics, so we only report them for our reproduced CE baseline and LOTUS variants. Our CE baseline nearly matches the published R2Gen scores on BLEU-1/BLEU-4/ROUGE-L (0.470/0.166/0.371 vs. 0.470/0.165/0.371), supporting it as a faithful reproduction for clinical-metric comparison. LOTUS summary rows report multi-seed means; the bottom row shows the individual LOTUS-Deep run with the highest BLEU-4 for comparison with single-run results, including the original submission's BLEU-4 of 0.192. "—" indicates the metric was not reported in the original paper.

Method	**B-4**	**ROUGE-L**	**F1-Ma**	**F1-Mi**	**R-Ma**	**R-Mi**	**P-Ma**
IU X-Ray (multi-seed mean ± std)							
CE Baseline	0.138 ±0.005	0.343 ±0.061	0.162 ±0.040	0.339 ±0.062	0.175 ±0.043	0.338 ±0.063	0.196 ±0.049
OT-Cross	0.148 ±0.034	0.365 ±0.012	0.180 ±0.010	0.345 ±0.015	0.179 ±0.011	0.348 ±0.014	0.192 ±0.011
OT-Deep	**0.170** ±0.020	**0.371** ±0.027	**0.201** ±0.004	**0.405** ±0.005	**0.194** ±0.007	**0.404** ±0.011	**0.221** ±0.009
ROCO							
CE Baseline	0.022	0.213	0.099	0.395	0.138	0.390	**0.114**
OT-Cross	0.022	0.216	0.108	0.520	0.129	0.516	0.105
OT-Deep	**0.032** ±0.000	**0.238** ±0.002	**0.121** ±0.006	**0.600** ±0.001	**0.142** ±0.007	**0.599** ±0.001	0.107 ±0.012

Table 4: Main results on IU X-Ray and ROCO test sets. IU X-Ray results are multi-seed mean ± std; ROCO OT-Deep results are mean ± std over completed runs, and remaining ROCO rows are single-run. Clinical metrics: F1-Ma (macro), F1-Mi (micro), R-Ma/R-Mi (recall), P-Ma/P-Mi (precision).

Config	B-4	R-L	F1-Ma	F1-Mi
LOTUS-Deep (full)	0.192	0.401	0.203	0.402
w/o ρ-capping	0.183	0.389	0.188	0.381
w/o diversity	0.191	0.399	0.201	0.405
w/o Proto-OT	0.179	0.376	0.184	0.352
CE Baseline	0.166	0.371	0.141	0.447

Table 5: Key ablations on IU X-Ray (single round). Full model achieves best balance. Multi-seed controlled ablations are in Appendix B.9.

Parameter	Range tested	Best	BLEU-4 range
ε (ot_epsilon)	0.01–0.5	0.1	0.183–0.192
Sinkhorn iterations (N_{iter})	10–100	30	0.189–0.192
ρ (ratio cap)	0.2–∞	0.5	0.176–0.192

Table 6: Hyperparameter sensitivity analysis on IU X-Ray.

terpreted as a thoracic-subset proxy rather than a complete clinical metric; only 30.1% of ROCO test captions receive at least one meaningful thoracic label, so we additionally report a modality-agnostic entity-overlap metric in Appendix B.8. On IU X-Ray, LOTUS-Cross changes multi-seed F1-Mi only slightly relative to CE (0.345 vs. 0.339), while LOTUS-Deep reaches 0.405. A MIMIC-CXR-RRG scalability study (Appendix B.11) further shows that the generation-quality and clinical-F1 gains extend to a larger MIMIC-style benchmark. Detailed qualitative examples, per-label analysis, and t-SNE visualizations of learned prototypes are provided in Appendix B.6.

5 Discussion

5.1 Dataset-Dependent OT Behavior

Our experiments reveal dataset-dependent OT behavior. IU X-Ray contains longer, clinically structured reports (average 37.56 tokens) with complex multi-finding narratives, benefiting from DeepOT's hierarchical alignment. ROCO contains shorter caption-style texts (average 20.91 tokens) across eight modalities, where simpler CA-OT suffices. The per-label analysis (Fig. 5, Appendix B.5) shows IU X-Ray has more balanced support across findings, while ROCO exhibits extreme class imbalance where micro-metric gains on frequent labels dominate.

This suggests a practical design heuristic: DeepOT is better suited to structured clinical report generation where semantic consistency across multiple findings matters, whereas CA-OT is a lighter choice for shorter multi-modal captioning settings where broad generalization is prioritized.

5.2 Clinical Efficacy Evidence

Beyond text-overlap improvements, our model shows favorable clinical-label agreement. On IU X-Ray, the multi-seed F1-Ma improvement (0.162 $\rightarrow$ 0.201, +24%) indicates substantially better identification of less-frequent but clinically important findings (e.g., pleural effusion, pneumothorax). On ROCO, we observe substantial gains in thoracic-proxy F1-Mi (0.395 $\rightarrow$ 0.600), suggesting more accurate prediction on frequent labels; because only 30.1% of ROCO test captions receive meaningful thoracic labels, we complement this with a modality-agnostic entity-overlap analysis (Appendix B.8), where the OT variants also improve over the CE baseline (0.320 $\rightarrow$ 0.372). However, smaller macro-level gains on ROCO indicate that rare classes remain a bottleneck, motivating future work on rare-finding sensitivity. We emphasize that clinical-label agreement and reduced attention diffuseness (Appendix B.12) do not by themselves establish radiologist-validated token–region grounding, which we leave to future work.

6 Conclusion

We presented LOTUS, an OT-based framework for radiology report generation that places transport supervision at the decoder evidence interface via Cross-Attention OT (CA-OT), strengthened by Prototype OT (Proto-OT) in the DeepOT variant. Experiments on IU X-Ray and ROCO, with a scalability study on MIMIC-CXR-RRG, demonstrate improved macro-level clinical-label agreement with strong language quality, particularly on structured report generation tasks. The gains are metric- and dataset-dependent: DeepOT prioritizes cross-finding consistency for clinical reports, whereas CA-OT favors lightweight generalization for shorter captions, and we explicitly document the lexical–clinical and macro–micro trade-offs between them. Future work will focus on efficient OT approximation, adaptive weighting mechanisms, localization-level grounding evaluation with radiologist review, and validation on the full credentialed MIMIC-CXR benchmark.

7 Limitations

Our approach has several limitations. First, computational overhead from Sinkhorn iterations increases training time compared to standard attention mechanisms, and the model is sensitive to OT hyperparameters which can lead to training instability without careful tuning. The ratio cap mechanism helps prevent collapse but introduces additional complexity. Second, our current evaluation focuses on NLG metrics and multi-label classification F1 scores; we do not report binary classification accuracy for individual findings, which would be important for clinical deployment. Third, while we add a scalability study on MIMIC-CXR-RRG (Appendix B.11), evaluation on the full credentialed MIMIC-CXR benchmark—which requires a lengthy access-approval process—remains future work. Fourth, our claims concern clinical-label agreement and cross-modal alignment: we do not perform localization-level grounding evaluation (e.g., pointing-game or IoU against annotated regions) or radiologist review of generated reports, so improved clinical-label F1 and reduced attention diffuseness should not be read as validated visual grounding. Fifth, our comparisons are scoped to lightweight, trainable encoder-decoder models; foundation-model report generators operate in a different resource regime and are not directly comparable. Finally, the ROCO clinical evaluation uses CheXpert thoracic labels as a subset proxy (only 30.1% of test captions receive meaningful labels), which we complement with entity-overlap metrics but which remains an imperfect clinical measure.

Several directions remain for future work. We plan to validate on the full MIMIC-CXR benchmark, binary classification metrics, and

localization-level grounding evaluation with radiologist review. The current model-only optimization could benefit from incorporating retrieval mechanisms or knowledge bases to provide external clinical context. Additionally, the effectiveness of OT-based supervision on smaller pretrained models (rather than large vision-language backbones) remains unexplored—this could make the approach more accessible for resource-constrained settings. Finally, efficient OT approximation methods could reduce computational overhead for real-time clinical applications.

Ethics Statement

LOTUS is a research prototype intended as a report-drafting aid only. Generated reports must be reviewed and edited by qualified clinicians before any use; the system should not be used for autonomous diagnosis, triage, or clinical decision-making. Several concrete failure modes matter for deployment. **Hallucinated findings**: fluent but incorrect generated findings could trigger unnecessary follow-up imaging or patient anxiety. **Missed abnormalities**: errors of omission could delay care, and aggregate F1 improvements do not preclude per-finding misses. **Over-trust**: fluent language may encourage clinicians to accept drafts even when clinically wrong. **Bias and generalization**: training data reflects the demographic, scanner, and site distribution of retrospective benchmarks; performance may degrade under distribution shift (different scanners, protocols, or populations), and subgroup performance should be audited before deployment. Any clinical use would require prospective multi-site validation, subgroup auditing, and post-deployment monitoring. The datasets used (IU X-Ray, ROCO, MIMIC-CXR-RRG) are publicly available research datasets; no patient-identifying information is presented.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR.

Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. 2023. PLOT: Prompt learning with optimal transport for vision-language models. In *The Eleventh International Conference on Learning Representations*.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5904–5914, Online. Association for Computational Linguistics.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, Online. Association for Computational Linguistics.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. 2017. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. 2021. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3922–3931.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Preprint*, arXiv:1901.07031.

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. 2024. PromptMRG: Diagnosis-driven prompts for medical report generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(3):2607–2615.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966. PMLR.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. 2024a. Bootstrapping large language models for radiology report generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18635–18643.

Zhizhe Liu, Zhenfeng Zhu, Shuai Zheng, Yawei Zhao, Kunlun He, and Yao Zhao. 2024b. From observation to concept: A flexible multi-view paradigm for medical report generation. *IEEE Transactions on Multimedia*, 26:5987–5995.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M. Friedrich. 2018. Radiology objects in COntext (ROCO): A multimodal image dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, volume 11043 of *Lecture Notes in Computer Science*, pages 180–189, Cham. Springer International Publishing.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 60 others. 2025. MedGemma technical report. *Preprint*, arXiv:2507.05201v1.

Dong Shu, Haiyan Zhao, Jingyu Hu, Weiru Liu, Ali Payani, Lu Cheng, and Mengnan Du. 2025. Large vision-language model alignment and misalignment: A survey through the lens of explainability. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1713–1735, Suzhou, China. Association for Computational Linguistics.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023a. METransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11558–11567.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023b. R2GenGPT: Radiology report generation with frozen LLMs. *Meta-Radiology*, 1(3):100033.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. MedCLIP: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3876–3887, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Di You, Fenglin Liu, Shen Ge, Xiaoxia Xie, Jing Zhang, and Xian Wu. 2021. AlignTransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 72–82, Cham. Springer International Publishing.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu Wei, and 8 others. 2025. A clinically accessible small multimodal radiology model and evaluation metric for chest X-ray findings. *Nature Communications*, 16(1):3108.

Ke Zhang, Hanliang Jiang, Jian Zhang, Qingming Huang, Jianping Fan, Jun Yu, and Weidong Han. 2024. Semi-supervised medical report generation via graph-guided hybrid feature consistency. *IEEE Transactions on Multimedia*, 26:904–915.

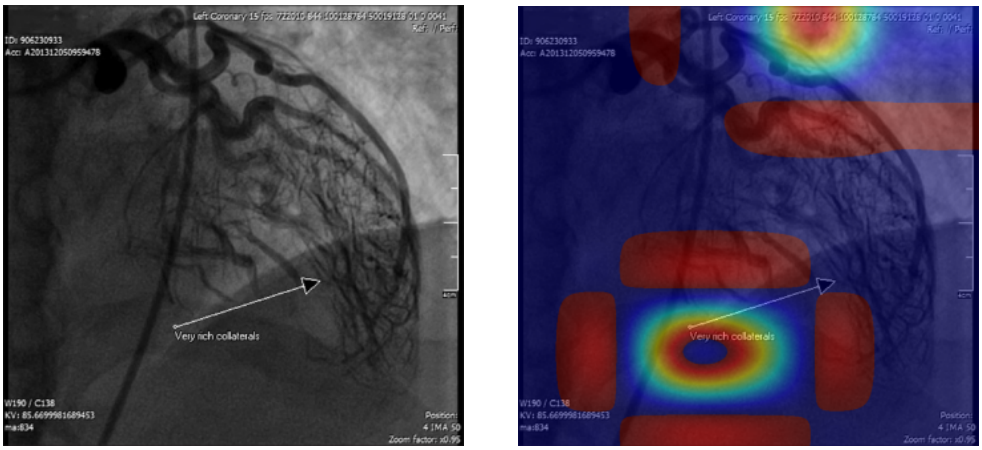
Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Matthew P. Lungren, Tristan Naumann, and Hoifung Poon. 2023. Large-scale domain-specific pretraining for biomedical vision-language processing. *Preprint*, arXiv:2303.00915v1.

Yanan Zhu and Yangyang Qi. 2023. Domain adaptation based on attention-weighted optimal transport and cluster alignment. In *2023 4th International Conference on Computer Vision, Image and Deep Learning (CVIDL)*, pages 514–517.

Appendix

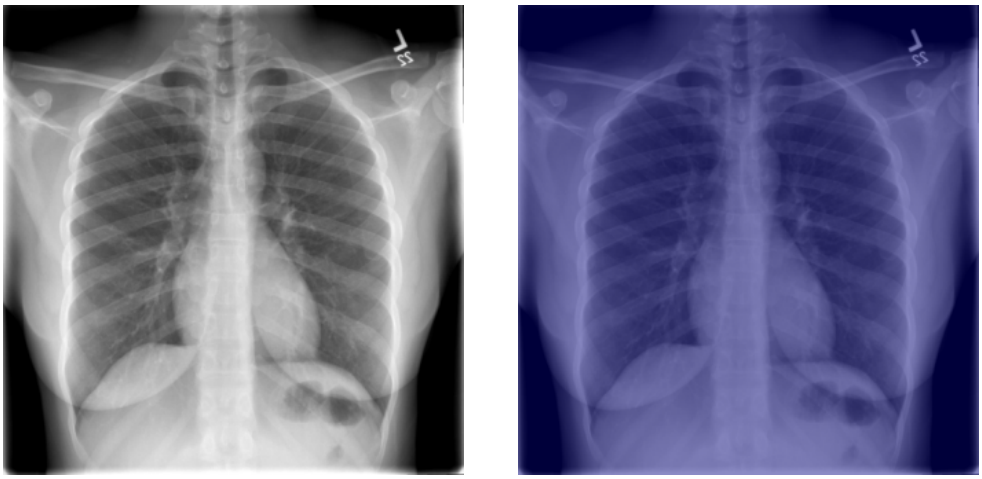
A Qualitative Comparison Examples

This appendix provides detailed qualitative comparison of report generation results on ROCO and IU X-Ray datasets.



GT: left coronary angiography showing a rich network of collateral vessels between obtuse marginal branch and the left anterior descending artery.
CE: right coronary angiography showing no collateral vessels. the right circumflex artery is occluded with left anterior descending artery.
CA-OT: left coronary angiography showing the rich network of collateral vessels between obtuse marginal branch and the right coronary artery.
DeepOT: left coronary angiography showing a rich network of collateral vessels between obtuse marginal branch and the left anterior descending artery.

Figure 3: ROCO qualitative example (ID: 76255): image, attention map, and reports. Blue marks correct content; red marks incorrect content.



GT: cardiac and mediastinal contours are within normal limits. the lungs are clear. bony structures are intact.
CE: the cardiomediastinal silhouette and pulmonary vasculature are within normal limits in size. the lungs are clear of focal airspace disease. there are no acute bony findings.
CA-OT: cardiac and mediastinal contours are within normal limits. the lungs are clear. skeletal structures are intact.
DeepOT: cardiac and mediastinal contours are within normal limits. the lungs are clear. bony structures are intact.

Figure 4: IU X-Ray qualitative example (ID: CXR2785_IM-1220): image, attention map, and reports. Color meaning follows Fig. 3.

B Theoretical Proofs

B.1 Proof of Lemma 3.1 (Entropic Approximation Bound)

Write the entropic objective as $J_\varepsilon(P) = \langle P, C\rangle - \varepsilon H(P)$, where $H(P) = -\sum_{i,j} P_{ij} \log P_{ij}$, and let

$$P^\star = \arg\min_{P\in\Pi(\mu,\nu)} \langle P, C\rangle, \qquad P_\varepsilon^\star = \arg\min_{P\in\Pi(\mu,\nu)} J_\varepsilon(P)$$

denote the exact and entropic optimal plans. Then $W_1(\mu,\nu) = \langle P^\star, C\rangle$ and $W_{1,\varepsilon}(\mu,\nu) = J_\varepsilon(P_\varepsilon^\star)$.

Lower bound. For any feasible coupling $P \in \Pi(\mu,\nu)$, the entropy on an $N_v \times N_t$ grid satisfies $0 \le H(P) \le \log(N_v N_t)$. Therefore

$$J_\varepsilon(P) = \langle P, C\rangle - \varepsilon H(P) \ge \langle P, C\rangle - \varepsilon \log(N_v N_t) \ge W_1(\mu,\nu) - \varepsilon \log(N_v N_t).$$

Taking the minimum over P yields

$$W_{1,\varepsilon}(\mu,\nu) \ge W_1(\mu,\nu) - \varepsilon \log(N_v N_t).$$

Upper bound. Evaluating the entropic objective at the exact optimal plan $P^\star$ gives

$$W_{1,\varepsilon}(\mu,\nu) = J_\varepsilon(P_\varepsilon^\star) \le J_\varepsilon(P^\star) = \langle P^\star, C\rangle - \varepsilon H(P^\star) \le \langle P^\star, C\rangle = W_1(\mu,\nu),$$

since $H(P^\star) \ge 0$.

Combining the two inequalities proves

$$W_1(\mu,\nu) - \varepsilon \log(N_v N_t) \le W_{1,\varepsilon}(\mu,\nu) \le W_1(\mu,\nu).$$

□

B.2 Proof of Proposition 1

Proposition 1. *Let $W_1(\mu,\nu)$ denote the 1-Wasserstein distance. Then:*

$$W_1(P_v, P_t) \;\le\; \mathbb{E}_{(x_i, y_{i,1:L_i})\sim\mathcal{D}}\big[W_1(\mu_i,\nu_i)\big]. \tag{14}$$

Proof. Since all μ_i and ν_i are supported on the common ground space $(\mathbb{R}^d, c)$, each optimal coupling $\gamma_i \in \Pi(\mu_i,\nu_i)$ is a measure on the same product space $\mathbb{R}^d \times \mathbb{R}^d$. Define $\bar\gamma = \mathbb{E}_{i\sim\mathcal{D}}[\gamma_i]$. For any measurable set $A \subseteq \mathbb{R}^d$, we have $\bar\gamma(A\times\mathbb{R}^d) = \mathbb{E}_i[\gamma_i(A\times\mathbb{R}^d)] = \mathbb{E}_i[\mu_i(A)] = P_v(A)$, and symmetrically $\bar\gamma(\mathbb{R}^d \times A) = P_t(A)$, so $\bar\gamma \in \Pi(P_v, P_t)$. Therefore:

$$W_1(P_v, P_t) \;\le\; \int c\, d\bar\gamma \;=\; \mathbb{E}_i\left[\int c\, d\gamma_i\right] \;=\; \mathbb{E}_i[W_1(\mu_i,\nu_i)]. \tag{15}$$

□

B.3 Proof of Theorem 1

Theorem 1. *Let f_θ be K-Lipschitz (gradient clipping) and ℓ be L-Lipschitz. Define $\Delta R(\theta) = |\mathbb{E}_{z\sim P_t}[\ell(f_\theta(z), y)] - \mathbb{E}_{z\sim P_v}[\ell(f_\theta(z), y)]|$. Then:*

$$\begin{aligned}\Delta R(\theta) &\le (K\cdot L)\cdot W_1(P_v, P_t)\\ &\le (K\cdot L)\left(\frac{1}{\lambda_{\text{ca}}^{\text{loc}}}\mathbb{E}[\mathcal{L}_{\text{CA-OT}}] + \varepsilon\log(N_v N_t)\right).\end{aligned} \tag{16}$$

Proof. Define $h = \ell \circ f_\theta$, which is $(K\cdot L)$-Lipschitz. By Kantorovich–Rubinstein duality:

$$|\mathbb{E}_{z\sim P_t}[h(z)] - \mathbb{E}_{z\sim P_v}[h(z)]| \le (K\cdot L)\cdot W_1(P_v, P_t), \tag{17}$$

where the dual form gives:

$$W_1(P_v, P_t) = \sup_{\|h\|_{\text{Lip}} \leq 1} \big(\mathbb{E}_{z\sim P_t}[h(z)] - \mathbb{E}_{z\sim P_v}[h(z)]\big). \tag{18}$$

Substituting $h = \ell \circ f_\theta$ with $\|h\|_{\text{Lip}} \leq K \cdot L$:

$$\Delta R(\theta) \leq (K \cdot L) \cdot W_1(P_v, P_t). \tag{19}$$

Applying the instance-to-population bound (Proposition 3.2) yields

$$\Delta R(\theta) \leq (K \cdot L) \cdot \mathbb{E}[W_1(\mu_i, \nu_i)]. \tag{20}$$

By Lemma 3.1, $W_1(\mu_i, \nu_i) \leq W_{1,\varepsilon}(\mu_i, \nu_i) + \varepsilon \log(N_v N_t)$. Eq. (2) uses the equivalent $(\log P_{ij} - 1)$ form, which differs from $W_{1,\varepsilon}$ only by the additive constant $-\varepsilon$, so it induces the same optimizer and the same bound up to a harmless constant shift. Since $\mathcal{L}_{\text{CA-OT}} \geq \lambda_{\text{ca}}^{\text{loc}} \mathcal{L}_{\text{ca-local}}$, we obtain

$$W_1(\mu_i, \nu_i) \leq \frac{1}{\lambda_{\text{ca}}^{\text{loc}}} \mathcal{L}_{\text{CA-OT}} + \varepsilon \log(N_v N_t). \tag{21}$$

Taking expectation and substituting gives the stated theorem. □

B.4 Implications for Proto-OT

The Proposition-1 argument applies verbatim at the prototype level. Let $\tilde{\mu}_i$ and $\tilde{\nu}_i$ denote prototype histogram distributions. Then:

$$W_1(\tilde{P}_v, \tilde{P}_t) \leq \mathbb{E}[W_1(\tilde{\mu}_i, \tilde{\nu}_i)]. \tag{22}$$

where $\tilde{P}_v = \mathbb{E}[\tilde{\mu}_i]$ and $\tilde{P}_t = \mathbb{E}[\tilde{\nu}_i]$ are population-level prototype marginals. With the same entropic argument as above, $W_1(\tilde{\mu}_i, \tilde{\nu}_i) \leq \mathcal{L}_{\text{p-local}} + \varepsilon \log(K^2)$, so Proto-OT bounds the *structure-level* discrepancy between visual and textual codebooks. It complements, rather than replaces, the token-level risk-gap bound of Theorem 1.

B.5 Per-Label Clinical Efficacy Analysis

To better understand the dataset-dependent performance differences, we analyze per-label F1, recall, and support distributions for DeepOT on both datasets. Figure 5 shows that IU X-Ray exhibits more balanced support across findings, with relatively even distribution across both frequent and rare classes. This explains why OT-Deep improves multi-seed macro-F1 (0.162 $\rightarrow$ 0.201) on IU X-Ray, as the model can leverage hierarchical alignment to improve performance on less frequent but clinically important findings.

In contrast, ROCO shows extreme class imbalance where a small number of high-frequency labels dominate the dataset. This explains why OT-Cross (simpler alignment) suffices for language metrics, while OT-Deep's gains are concentrated in micro-F1 (0.395 $\rightarrow$ 0.600) on frequent patterns. The per-label analysis reveals that rare classes remain a bottleneck on ROCO, motivating future work on rare-finding sensitivity and reweighted optimization.

B.6 Detailed Results Analysis

This appendix provides additional analysis of the main experimental results, including detailed trade-off analysis, t-SNE visualization of learned prototypes, and discussion of dataset-dependent OT behavior.

B.6.1 When does CA-OT help and hurt?

In the multi-seed IU X-Ray results in Table 4, LOTUS-Cross changes F1-Mi only slightly relative to CE (0.345 vs. 0.339), whereas LOTUS-Deep reaches 0.405. This supports a larger gain for the combined CA-OT and Proto-OT system on this metric; the aggregate scores alone do not establish the mechanism behind the difference.

B.6.2 t-SNE Visualization of Learned Prototypes

Figure 6 shows t-SNE visualization of prototype codebooks, demonstrating that learned visual and textual prototypes cluster into semantically coherent groups (e.g., lung findings, cardiac findings, normal anatomy). This directly validates that Proto-OT learns meaningful structure-level alignment.

B.7 Training Stability Analysis

While a formal convergence guarantee for non-convex Transformer optimization is beyond the scope of this work, we make two observations that explain the empirical stability benefit of OT regularization.

First, the entropic OT loss $\mathcal{L}_{\text{CA-OT}}$ is differentiable w.r.t. θ under $\varepsilon > 0$, because the Sinkhorn solution $\mathbf{P}^\star$ is a differentiable function of the input embeddings (Cuturi, 2013). Its gradients are locally Lipschitz, with constants that scale as $O(1/\varepsilon)$ together with the smoothness of the cost network. This smoothness distinguishes OT from hard-assignment losses that produce sparse, discontinuous gradients.

Second, unlike KL divergence, the Wasserstein distance W_1 yields finite, well-behaved gradients even when the visual distribution μ_i and token distribution ν_i have non-overlapping support—a common situation in early training when embeddings are not yet aligned. This property, combined with the ratio-capped weighting (Eq. 12), explains the stable loss curves observed empirically in our ablation study (Table 5).

B.8 ROCO Thoracic-Proxy and Entity-Overlap Evaluation

ROCO is a multi-modality dataset, and the CheXpert labeler covers only 14 thoracic findings designed for chest X-rays. Among the 5,000 ROCO test captions, only 1,507 (30.1%) receive at least one meaningful thoracic label. ROCO CheXpert-style F1 is therefore reported as a *thoracic-subset proxy*. To provide a modality-agnostic complementary metric, we extract medical entities (modality, anatomy, findings, and descriptors) from generated and reference captions and compute an entity-overlap F1.

Model	CheXpert F1-Mi	Entity-overlap F1
CE baseline	0.395	0.320
LOTUS-Cross	0.520	0.366
LOTUS-Deep	0.600	0.372

Table 7: ROCO thoracic-proxy CheXpert F1-Mi and modality-agnostic entity-overlap F1.

Both OT variants improve over the CE baseline under the entity-overlap metric, supporting the conclusion that the ROCO gains are not an artifact of the thoracic label subset. However, we keep ROCO ablation behavior out of the central attribution claims (Section B.9) because it is dataset-dependent; the clean attribution evidence comes from IU X-Ray.

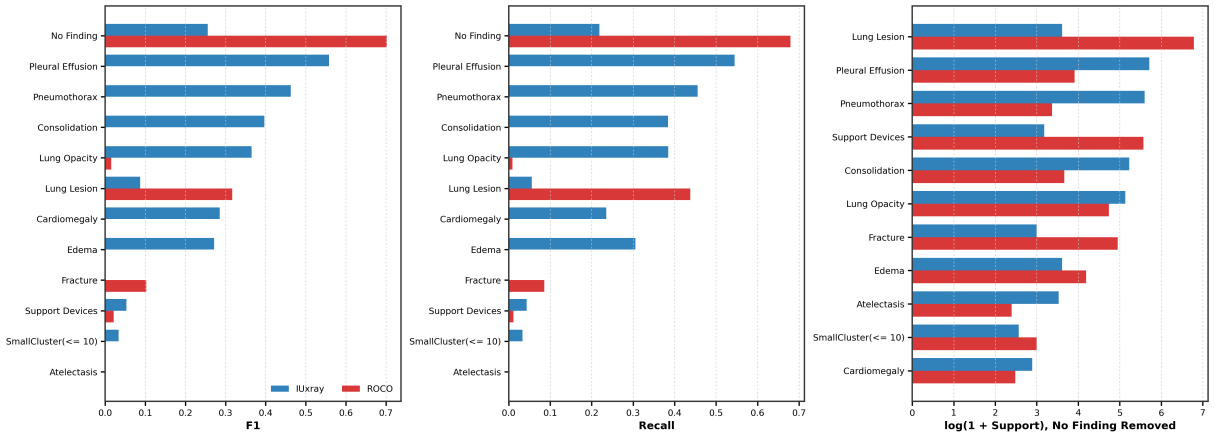


Figure 5: Per-label comparison of DeepOT on IU X-Ray and ROCO: F1, Recall, and Support.

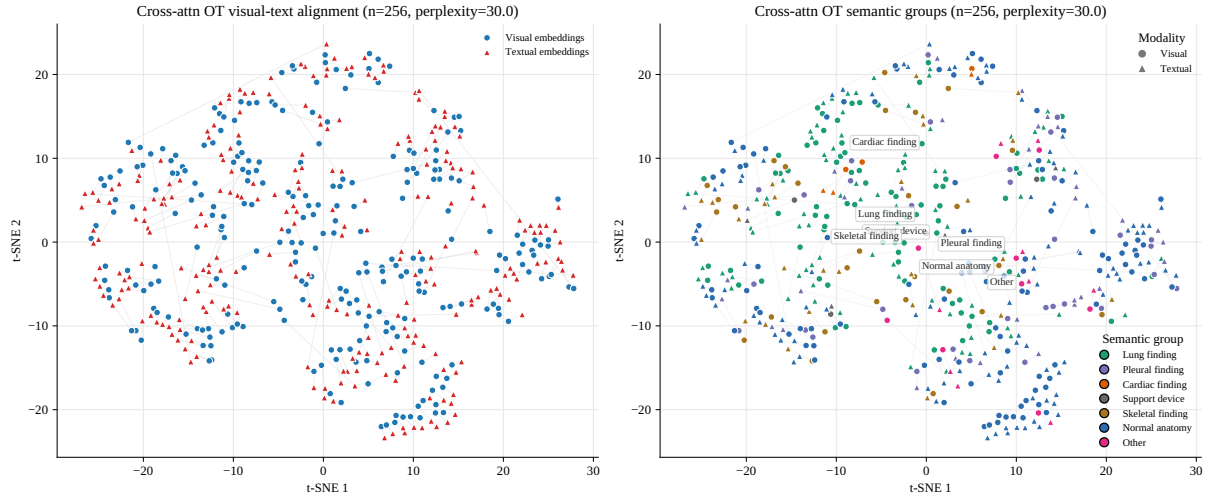


Figure 6: t-SNE visualization of Cross-Attention embeddings showing semantic clustering.

B.9 Controlled Non-OT Ablation

To separate the effect of OT from auxiliary objectives, we add controlled ablations under the same architecture, budget, and multi-seed evaluation protocol (mean $\pm$ std). Auxiliary-only variants and non-OT feature alignment remain below LOTUS-Deep on macro clinical F1, supporting attribution of the clinical macro-F1 gain to the transport objective rather than to ITC, diversity regularization, or generic feature alignment.

B.10 OT Cost-Function Sensitivity

We compared cosine distance and squared ℓ_2 costs in multi-seed IU X-Ray evaluation (macro clinical F1-Ma). The two costs are comparable, with squared ℓ_2 slightly higher in mean; we therefore retain squared ℓ_2 as the reported setting.

B.11 MIMIC-CXR-RRG Scalability Study

To assess external validity at larger scale without the lengthy credentialed-access process of full MIMIC-CXR, we use the public MIMIC-CXR-RRG split, which maintains patient-level train/validation/test separation. CE, Cross-OT, and Deep-OT are trained from scratch for 30 epochs under the same setup; Deep-OT is evaluated across multiple runs. We also compare against the available IAMJB baseline and Qwen-2.5-VL-7B on this benchmark.

Deep-OT achieves the best generation quality on MIMIC-CXR-RRG (BLEU-1, BLEU-4, ROUGE-L) and improves clinical F1 over CE and the external baselines. We present this as additional evidence on a larger MIMIC-style benchmark, while keeping the full credentialed MIMIC-CXR evaluation as future work.

System	F1-Ma	F1-Mi	BLEU-4	ROUGE-L
CE + diversity only	0.168 ± 0.012	0.390 ± 0.039	0.142 ± 0.015	0.346 ± 0.006
non-OT feature alignment	0.171 ± 0.015	0.370 ± 0.021	0.140 ± 0.006	0.346 ± 0.004
CE + ITC only	0.177 ± 0.003	0.351 ± 0.015	0.122 ± 0.001	0.338 ± 0.002
LOTUS-Deep	**0.195** ± 0.010	**0.400** ± 0.012	**0.142** ± 0.009	**0.346** ± 0.009

Table 8: Controlled IU X-Ray ablation (multi-seed mean ± std). Clinical and language metrics are shown for the same systems.

Cost	F1-Ma (multi-seed mean ± std)
cosine	0.193 ±0.005
squared-ℓ_2	**0.195** ±0.010

Table 9: OT cost-function sensitivity on IU X-Ray (F1-Ma).

Model	B-1	B-4	R-L
IAMJB baseline	0.323	0.078	0.244
Qwen-2.5-VL-7B	0.341	0.115	0.277
CE	0.321	0.107	0.266
Cross-OT	0.344	0.121	0.268
Deep-OT (Ours)	**0.351** ± 0.002	**0.134** ± 0.002	**0.279** ± 0.001

Model	F1-Ma	F1-Mi	R-Ma	R-Mi	P-Ma	P-Mi
IAMJB baseline	0.273	0.424	0.258	0.371	0.320	0.495
Qwen-2.5-VL-7B	0.180	0.315	0.292	**0.388**	0.330	0.447
CE	0.282	0.369	0.276	0.342	0.292	0.330
Cross-OT	0.344	0.412	0.347	0.356	0.342	0.489
Deep-OT (Ours)	**0.363** ± 0.004	**0.428** ± 0.012	**0.361** ± 0.001	0.376 ± 0.005	**0.366** ± 0.013	**0.496** ± 0.005

Table 10: MIMIC-CXR-RRG scalability results. Deep-OT is reported as multi-seed mean ± std. The panels show language, F1, recall, and precision metrics for the same systems.

B.12 Attention-Concentration Analysis

To probe whether OT changes decoder-to-patch attention behavior, we compare the CE baseline (OT removed) and LOTUS-Deep under the same architecture and random initialization. We measure the number of effective patches (inverse participation ratio of attention mass) and the normalized entropy of decoder-to-patch attention. Removing OT spreads attention over 77.6/98 effective patches with normalized entropy 0.95, whereas LOTUS-Deep concentrates attention to 43.7/98 patches with entropy 0.82 and higher peak attention. We interpret this conservatively: OT reduces attention diffuseness, but this does not constitute radiologist-validated token–region localization; pointing-game/IoU evaluation against annotated regions and expert review remain future work.

B.13 Prototype-Codebook Utilization

Proto-OT depends on the quality of the learned codebooks, so we check for codebook collapse. Across training, all 32/32 visual and all 32/32 textual prototypes remain active, with normalized usage entropy of 0.98 and 0.99, respectively, indicating no collapse and near-uniform mode coverage.