TELOS: FROM SEMANTIC ACCESS TO REALIZATION VIA EXPRESSION MEMORY AND SUFFIX CONTRAST

Anonymous authors
Paper under double-blind review

ABSTRACT

Radiology report generation requires accurate expression of clinical concepts, including radiographic findings, anatomical sites, negations, and diagnostic uncertainties. While vision–language models (VLMs) generate fluent narratives, overall language fluency does not guarantee the reliability of specialized terminology, particularly when clinical expressions span multiple subword tokens. Standard fine-tuning updates continuous parameters but leaves domain concepts implicit, forcing the decoder to coordinate multi-token realization through unguided subword steps. We investigate whether specialized terminology can be better supported without altering native autoregressive decoding by coupling addressable expression memory with targeted suffix supervision. To this end, we introduce TELOS (**T**erm-level **E**xpression **L**atent memory with **O**nset-to-**S**uffix contrast, an expression-level adaptation framework motivated by classical teleology ($\tau\acute{\epsilon}\lambda o\varsigma$, realization of whole concepts from initial potential). TELOS supports native subword decoding of a frozen VLM through explicit memory and targeted supervision. It maintains an addressable Clinical Term Dictionary indexed by complete expressions. At sparse decoder depths, a Bilateral Evidence Router (BER) combines contextual language cues with visual re-attention to regulate weighted memory access, using an explicit abstention state to suppress current memory reads during non-specialized phrasing. Causal depthwise writeback propagates expression-relevant support across neighboring steps to assist multi-subword realization. To support local realization decisions, Onset-to-Suffix Contrast (OSC) provides targeted supervision at suffix positions against high-scoring non-reference competitors. The resulting interface leaves the pretrained backbone, tokenizer, and vocabulary unchanged. Extensive evaluations on MIMIC-CXR across LLaVA-1.5, Qwen2.5-VL, and LLaVA-Rad demonstrate consistent improvements in report quality and terminology realization: on LLaVA-1.5, TELOS improves RadGraph-F1 from 0.158 to 0.240 and BLEU-4 from 0.080 to 0.158 over decoder LoRA, requiring only 25% trainable parameters compared with matched LoRA.

1 INTRODUCTION

Adapting general vision–language models (VLMs) to specialized biomedical domains introduces profound representational challenges, particularly for clinical radiology report generation. In this clinical task, overall language fluency is insufficient to guarantee reliable diagnostic terminology: autoregressive decoders readily generate fluent narrative templates, yet frequently fail to instantiate precise medical terms. Core radiological findings (e.g., *cardiomegaly*, *pleural effusion*, *pneumothorax*) function as cohesive clinical expressions that convey radiographic observations, anatomical sites, negations, and diagnostic uncertainties. However, modern VLMs universally rely on subword tokenization (Sennrich et al., 2016; Kudo & Richardson, 2018) to maintain an open vocabulary. Although subwords make specialized domain lexemes encodable without expanding the base dictionary, they shatter single semantic concepts into fragmented sequences of subword pieces. Consequently, the autoregressive decoder faces a chain of sequential decisions: after emitting an initial subword piece, it must generate the remaining suffix under evolving prefixes. We therefore focus on supporting reliable specialized expression generation within native subword decoders.

Under evaluated models (LLaVA-1.5 (Liu et al., 2024) and Qwen2.5-VL (Bai et al., 2025)) and reporting protocols, diagnostic probing reveals substantial empirical difficulty in multi-token clinical

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.

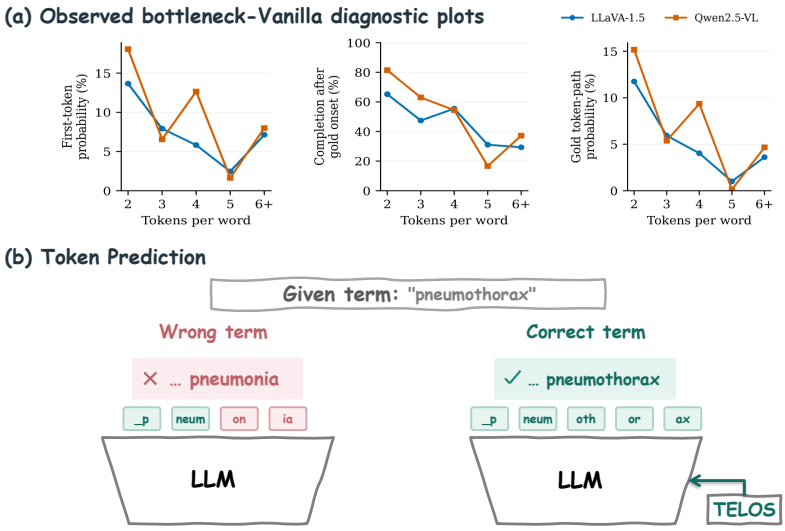


Figure 1: **The onset-to-realization bottleneck and its resolution in clinical report generation.** **(a)** Vanilla diagnostic plots across token lengths (2 to 6+ subwords) on LLaVA-1.5 and Qwen2.5-VL, showing first-token probability, completion after gold onset, and gold token-path likelihood (Appendix A.2). **(b)** Token prediction under target term "*pneumothorax*": unguided decoding lacks explicit concept anchoring and suffers suffix drift after prefix pieces `_p+neum`, collapsing into high-frequency shortcut `on+ia` (*pneumonia*); with TELOS, explicit expression memory and targeted suffix contrast guide the decoder along the full multi-token trajectory `oth+or+ax` (*pneumothorax*).

expression realization (Figure 1(a); Appendix A.2). While predicting the initial onset subword (w_1) already requires specialized domain knowledge (left plot), trailing subwords ($w_{2:m}$) face compounding sequential uncertainty. Crucially, even when the ground-truth onset is teacher-forced, exact word completion exhibits steep degradation as token length scales from 2 to 6+ subwords (middle plot, dropping from $\sim$80% down to 20%–30% for terms ≥ 4 subwords), accompanied by sharp decay in joint token-path likelihood (right plot). This empirical finding demonstrates that terminology challenges extend beyond initial token selection, and subsequent continuation steps face severe realization bottlenecks that warrant dedicated support.

We use "expression" to encompass both multi-token medical words and multi-word clinical terms treated as unified domain units. In radiology report generation, recurring diagnostic expressions must be reliably addressable throughout their generation trajectory. Standard parameter-efficient fine-tuning such as LoRA (Hu et al., 2021) adapts continuous weight matrices across the network, but leaves specialized clinical concepts implicit in activation space without explicit entity representations (Figure 1(b)). Generating one piece at a time from native subword vocabulary without explicit concept anchoring, the unguided decoder easily suffers *suffix drift*: after emitting early prefix tokens (e.g., `_p+neum`), local greedy continuation can be swayed by high-frequency vocabulary shortcuts (e.g., `on+ia`), diverging into an incorrect diagnosis (*pneumonia*) rather than completing the target finding (*pneumothorax*; Figure 1(b), left).

Recognizing that clinical expressions function as cohesive semantic units, our design hypothesis is that **complete expression representation and termwise generation decisions can receive complementary support**: (1) *At the representation level*, explicit, addressable latent representations for complete clinical expressions allow the decoder to retrieve domain concepts during generation; (2) *At the supervision level*, targeted supervision at suffix positions against high-scoring non-reference competitors directly guides local continuation. The former is realized via expression memory, bilateral routing, and causal writeback; the latter is realized through Onset-to-Suffix Contrast (OSC). Both components preserve native subword decoding while supporting specialized domain expressions (Figure 1(b), right).

We operationalize this design in TELOS (**T**erm-level **E**xpression **L**atent memory with **O**nset-to-**S**uffix contrast, conceptually motivated by teleological realization from initial potential; Figure 2).

TELOS freezes pretrained VLM weights, output vocabulary, and autoregressive interface. At sparse decoder depths, lightweight modules query a shared Clinical Term Dictionary indexed by complete expressions. A Bilateral Evidence Router (BER) balances contextual language cues image-grounded compatibility from a visual re-attention module, with an explicit abstention state suppressing current memory reads during general syntax. Causal depthwise writeback propagates expression-relevant support across consecutive steps to sustain multi-subword continuation. Memory is read via weighted combination rather than discrete word selection, allowing the native vocabulary head to govern final generation. At training time, Onset-to-Suffix Contrast (OSC) penalizes top-scoring non-reference competitors at suffix positions, directly connecting continuous expression support to discrete token competition that completes each specialized clinical expression.

In summary, our contributions integrate three foundational dimensions:

Problem and Motivation: We identify the fundamental disconnect between superficial narrative fluency and diagnostic terminology accuracy in VLMs. We show that while subword tokenization enables open-vocabulary encoding, it fragments specialized medical findings into fragile subword chains prone to suffix drift. We formulate multi-token clinical expressions as a concrete entry point and propose providing dedicated, complementary support for whole-expression representation and termwise suffix realization.

Architectural Interface: We develop TELOS, an expression-level interface that leaves the pretrained VLM backbone, tokenizer, and vocabulary completely frozen. TELOS maintains an addressable Clinical Term Dictionary indexed by complete expressions, dynamically combines contextual language priors with visual re-attention via a Bilateral Evidence Router (BER) with an explicit abstention state, carries support forward through causal depthwise writeback, and applies Onset-to-Suffix Contrast (OSC) to penalize non-reference competitors at suffix positions.

Multi-Level Validation: Across three VLM backbones (LLaVA-1.5, Qwen2.5-VL, LLaVA-Rad) and four benchmarks (MIMIC-CXR, CheXpert Plus, IU-Xray, COVCTR), extensive evaluations validate three complementary dimensions: (1) *Report Quality*, where TELOS rivals or outperforms specialized radiology foundation models; (2) *Terminology Behavior*, where TELOS mitigates suffix drift across token lengths 2–6+ under both controlled and unconstrained evaluation; and (3) *Adaptation Efficiency*, requiring only 5.82M trainable parameters (25% of matched LoRA).

2 RELATED WORK

Multimodal foundation models for radiology. Specializing vision–language models for biomedicine and radiology has emerged as a prominent direction. Prior works have introduced dedicated models such as LLaVA-Med (Li et al., 2023a), MedGemma1.5 (Sellergren et al., 2026), CheXagent (Chen et al., 2024), MAIRA-1 (Hyland et al., 2023), and LLaVA-Rad (Zambrano Chaves et al., 2025), typically relying on extensive full-parameter or projector pretraining on paired biomedical datasets. Complementary studies incorporate external knowledge graphs, structured diagnostic prompts, or retrieval-based report aids (Chen et al., 2020; 2021; Yang et al., 2022; Jin et al., 2024; Hou et al., 2025). While domain-specific pretraining yields strong representations, it incurs substantial compute costs and must be repeated for each architecture. We show that off-the-shelf general VLMs, when equipped with a lightweight expression-level latent memory interface, can rival or exceed fully specialized radiology foundation models on clinical report generation.

Parameter adaptation versus memory-augmented interfaces. Parameter-efficient fine-tuning (e.g., LoRA (Hu et al., 2021)) adapts frozen backbones via continuous low-rank updates, but leaves discrete domain entities implicit in activation space. Meanwhile, memory-augmented generation explores external datastores: RAG (Lewis et al., 2020; Borgeaud et al., 2022) retrieves documents at the input layer, kNN-LM (Khandelwal et al., 2019) interpolates nearest-neighbor distributions at the output layer, and Memorizing Transformers (Wu et al., 2022) augment internal attention layers with cached past keys/values at heavy inference cost. Recent methods explore conditional n-gram lookups (Engram (Cheng et al., 2026b)) or memory grafting (Cheng et al., 2026a). Unlike static text n-gram tables, TELOS treats complete clinical expressions as addressable semantic entities, regulates memory access via bilateral context-visual routing with explicit abstention, and guides suffix realization through contrastive ranking (Burges et al., 2005).

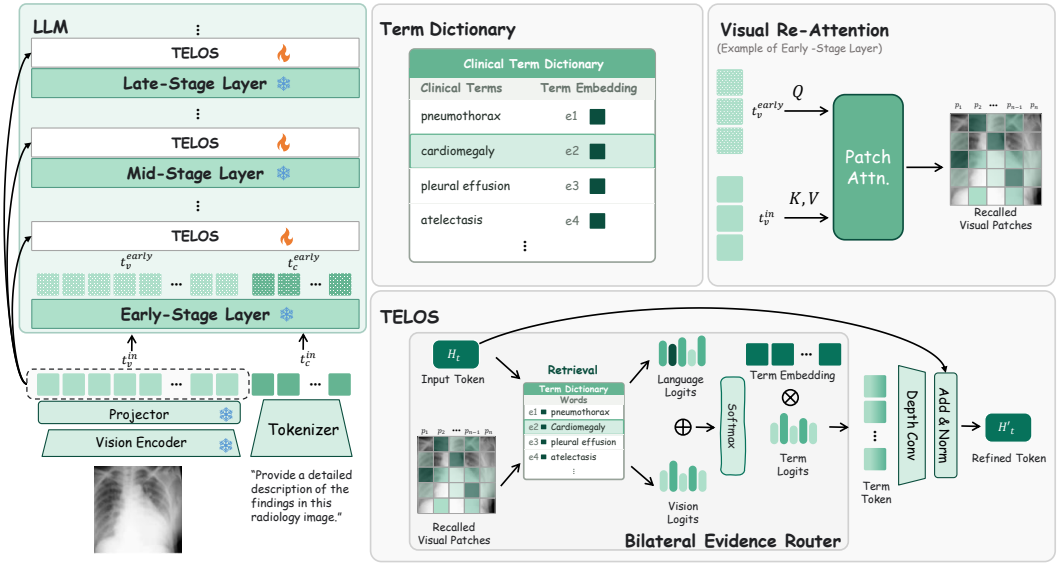


Figure 2: **Overview of the TELOS framework.** A frozen VLM decoder queries explicit clinical expression memory entries at sparse depths via joint contextual and visual re-attention routing with an explicit abstention state, propagates retrieved information via causal writeback, and is trained with Onset-to-Suffix Contrast (OSC). Pretrained backbone parameters remain completely frozen.

3 METHOD

TELOS provides targeted support for clinical terminology while preserving native subword autoregression on frozen VLMs. Leaving pretrained backbone weights frozen, it introduces a modular interface organized into four functional regions (Figure 2): (1) **Task Formulation and Spans** (Section 3.1); (2) an addressable **Clinical Term Dictionary** (Section 3.2) indexed by complete clinical expressions; (3) **Visual Re-attention** (Section 3.3) recalling fine-grained visual patch tokens for expression compatibility; and (4) **TELOS Routing and Causal Writeback** (Section 3.4) balancing contextual and visual evidence with temporal depthwise smoothing. Finally, **Onset-to-Suffix Contrast (OSC)** (Section 3.5) connects expression-level memory to discrete token competition by supervising suffix positions against hard competitors.

3.1 TASK FORMULATION AND TERMINOLOGY SPANS

Autoregressive VLM formulation. Given radiograph I and instruction prompt x, a frozen vision encoder and projector yield visual tokens t_v^{in}, while the tokenizer produces context tokens t_c^{in} (Figure 2, left). The native autoregressive decoder generates report findings $y_{1:T}$ token by token:

$$p_{\theta,\phi}(y_{1:T} \mid I, x) = \prod_{t=1}^{T} p_{\theta,\phi}(y_t \mid I, x, y_{<t}). \tag{1}$$

Multi-token clinical expressions as a concrete entry point. Clinical reports rely heavily on specialized expressions describing radiographic findings, anatomical sites, negations, and diagnostic uncertainties. Under native subword tokenization, each domain expression c decomposes into a sequence of subwords $w_{1:m}$, comprising lexical onset w_1 and trailing suffix $w_{2:m}$. Given report prefix b, the target token-path probability factorizes as:

$$p(w_{1:m} \mid I, b) = p(w_1 \mid I, b) \prod_{j=2}^{m} p(w_j \mid I, b, w_{<j}). \tag{2}$$

While predicting onset w_1 requires domain knowledge, completing suffix $w_{2:m}$ requires coordinating sequential subword decisions. We curate a fixed inventory $\mathcal{D} = \{c_1, \ldots, c_N\}$ of multi-token clinical

expressions from the training split (RadLex/UMLS-aligned, document frequency ≥ 5, span length $2 \leq m \leq 6+$, stop-words removed; Appendix A.2), indexed via a prefix Trie for deterministic longest-match span tagging. For non-dictionary phrasing or out-of-vocabulary expressions, the explicit abstention state ($i = 0$) provides a dedicated routing bypass, allowing the decoder to fall back smoothly to native autoregression rather than forcing spurious memory reads.

3.2 Clinical Term Dictionary: Explicit Expression Memory

Semantic centroids and trainable domain offsets. Let $W_{\text{out}} \in \mathbb{R}^{|\mathcal{V}| \times d}$ be the frozen output embedding matrix of the base language model. For each clinical expression $c_i \in \mathcal{D}$ spanning subwords $w_{i,1:m_i}$, we construct an explicit, addressable memory entry e_i:

$$e_i^{\text{base}} = \frac{1}{m_i} \sum_{j=1}^{m_i} W_{\text{out}}[w_{i,j}], \qquad e_i = e_i^{\text{base}} + \delta_i, \qquad k_i = \text{RMSNorm}(e_i), \quad v_i = e_i. \tag{3}$$

Base vector e_i^{base} provides semantic initialization by averaging constituent subword embeddings in pretrained vocabulary space. Trainable residual $\delta_i \in \mathbb{R}^d$ (initialized to zero) acts as a task-adaptive domain offset, adapting the representation to specialized clinical contexts without distorting general vocabulary geometry. Crucially, each expression string c_i maintains an explicit correspondence with its addressable memory entry e_i. Shared across all intervention depths $\mathcal{L}$, $\{e_i\}_{i=1}^N$ with keys k_i and values v_i require only $O(Nd)$ computation per active position, maintaining ultralight parameters.

3.3 Visual Re-attention and Patch Compatibility

Recalling visual evidence via Patch Attention. To supply visual evidence for term retrieval without losing fine-grained radiographic details in deep layers, image tokens at active layer ℓ (t_v^ℓ) re-attend to initial input visual tokens t_v^{in} via Patch Attention:

$$p^\ell = t_v^\ell + G^\ell \odot \text{PatchAttn}(t_v^\ell, t_v^{\text{in}}), \tag{4}$$

where PatchAttn shares lightweight projection matrices $W_K, W_V \in \mathbb{R}^{d_h \times d_h}$ across $H = 8$ heads ($d_h = d/H$) and intervention depths, and G^ℓ provides layer-dependent residual gating (complete multi-head formulation in Appendix A.1). The updated patch representations $p^\ell = [p_1^\ell, \ldots, p_n^\ell]$ are then matched against expression keys k_i to yield visual compatibility scores (Vision Logits) s_i^ℓ:

$$s_i^\ell = \tau_v \log \left[\frac{1}{n} \sum_{k=1}^{n} \exp\left(\frac{\cos(p_k^\ell, k_i)}{\tau_v} \right) \right], \tag{5}$$

where Log-Mean-Exp serves as a smooth soft-maximum across spatial patches, identifying focal radiographic findings. Cached during the prefill phase, $\{s_i^\ell\}$ incur zero repeated computation during autoregressive generation.

3.4 Bilateral Evidence Router and Causal Writeback

Joint routing with explicit abstention. At each autoregressive generation step t and intervention layer ℓ, decoder hidden state H_t balances two complementary evidence sources: (1) an intrinsic contextual prior (Language Logits) $c_{t,i}^\ell = \gamma \cos(H_t, k_i)$, and (2) extrinsic visual patch evidence (Vision Logits) s_i^ℓ. Radiology reports naturally interleave standard narrative syntax with specialized assertions; an expression is prioritized when anticipated by preceding context and corroborated by focal radiographic features. Crucially, to prevent spurious retrieval during generic prose or unremarkable findings, we introduce an explicit *abstention state* ($i = 0$) with null representation $v_0 = \mathbf{0}$, zero visual score $s_0 = 0$, and learned bias b_0. The **Bilateral Evidence Router (BER)** computes combined logits $a_{t,i}^\ell$ and retrieves memory support via:

$$a_{t,i}^\ell = c_{t,i}^\ell + \beta_\ell s_i^\ell \ (1 \leq i \leq N), \quad a_{t,0}^\ell = b_0; \qquad q_t^\ell = \text{softmax}(a_t^\ell), \quad u_t^\ell = \sum_{i=1}^{N} q_{t,i}^\ell v_i. \tag{6}$$

When generating common syntax, routing probability concentrates on the abstention anchor ($q_{t,0}^\ell \approx 1$), smoothly scaling retrieved memory vector $u_t^\ell \to \mathbf{0}$ so the frozen decoder executes native generation without vocabulary distortion. Memory is read strictly via weighted combination u_t^ℓ rather than hard emission, leaving discrete token selection to the native vocabulary distribution.

Causal depthwise temporal writeback. While the abstention state deactivates retrieval at non-clinical steps, multi-token terms require sustaining continuous semantic representations across time. During the generation of a multi-token clinical term (e.g., *car-dio-meg-a-ly*), isolated per-step memory reads u_t^ℓ can exhibit variance or temporary dips in confidence. A 1D depthwise causal convolution of width 4 mixes memory reads across current and recent positions $\{t-3,\dots,t\}$:

$$r_t^\ell = u_t^\ell + \mathrm{DWConv}_4\big(\mathrm{RMSNorm}(u_{1:t}^\ell)\big)_t, \qquad H_t' = \mathrm{RMSNorm}(H_t + r_t^\ell). \tag{7}$$

Causal writeback acts as a temporal momentum filter: it propagates pre-activated concepts smoothly from onset w_1 across trailing subwords $w_{2:m}$, while remaining inactive when preceding positions were in abstention ($u_{<t}^\ell = \mathbf{0}$). Thus, abstention and causal writeback operate in orthogonal harmony, preserving non-clinical fluency while sustaining multi-subword terminology continuity. Rather than modifying every transformer block, TELOS injects memory modules sparsely at three representative decoder depths ($\mathcal{L} = \{1, 15, 31\}$ in 32-layer backbones; Appendix A.1): the early layer anchors initial domain semantics, the middle layer integrates multimodal evidence, and the late layer refines representations immediately prior to vocabulary projection, requiring only 5.04M trainable parameters (about 25% of matched LoRA, 5.04M vs 20.2M parameters).

3.5 ONSET-TO-SUFFIX CONTRAST

Connecting expression support to token competition. While memory access enriches hidden representations, it does not directly optimize decision boundaries against competing vocabulary tokens at critical suffix choice points. Under standard autoregressive training, cross-entropy optimizes all tokens uniformly across entire report narratives, lacking focused margin separation where subword paths diverge. Onset-to-Suffix Contrast (OSC) bridges expression-level memory to discrete token competition by mining high-scoring non-gold candidates at suffix positions and directly penalizing clinical distractor shortcuts.

For each matched expression occurrence $c \in \mathcal{M}$ spanning subwords $w_{1:m}$, at each suffix position $j \in \{2,\dots,m\}$, OSC mines the top-k non-reference vocabulary tokens $\mathcal{N}_j = \mathrm{TopK}_{v \neq w_j}\, o_{j,v}$ from forward logits $o_j \in \mathbb{R}^{|\mathcal{V}|}$. Let $\mathcal{C}_j = \{w_j\} \cup \mathcal{N}_j$. The localized contrastive loss over this competitive set is:

$$\ell_j = -\log \frac{\exp(o_{j,w_j})}{\sum_{v \in \mathcal{C}_j} \exp(o_{j,v})} = \log \left[1 + \sum_{v \in \mathcal{N}_j} \exp\left(o_{j,v} - o_{j,w_j}\right) \right]. \tag{8}$$

Competitor mining requires zero rollouts or external filtering. The objective normalizes across matched expressions in the report:

$$\mathcal{L}_{\mathrm{OSC}} = \frac{1}{\max(1, |\mathcal{M}|)} \sum_{c \in \mathcal{M}} \frac{1}{m_c - 1} \sum_{j=2}^{m_c} \ell_{c,j}. \tag{9}$$

Targeted supervision against clinical confusion competitors. Standard cross-entropy distributes supervisory loss uniformly across all tokens, optimizing average likelihood over fluent narratives without prioritizing critical decision boundaries where clinical terms fail. In clinical practice, suffix errors concentrate in high-frequency competitors sharing identical morphological prefixes (e.g., continuing `pneu+mo` into `nia` for *pneumonia* instead of completing `thor`→`ax` for *pneumothorax*). By mining the top-k non-reference tokens $\mathcal{N}_j$ directly from forward logits at each suffix step, OSC establishes an auxiliary listwise ranking objective over localized competitive set $\mathcal{C}_j = \{w_j\} \cup \mathcal{N}_j$. This directly penalizes high-scoring clinical distractors and widens the logit margin over deceptive shortcuts without rollout trajectories, external sampling, or filtering overhead.

3.6 TRAINING AND NATIVE INFERENCE

Curriculum joint objective and inference. Training couples standard autoregression with scheduled suffix contrast:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{AR}} + \lambda(e)\mathcal{L}_{\mathrm{OSC}}, \qquad \text{where} \quad \lambda(e) = \lambda_{\mathrm{osc}}\, \mathbf{1}\{e \geq e_{\mathrm{start}}\}. \tag{10}$$

Initial warmup ($\lambda = 0$) stabilizes routing, and subsequent training activates OSC to sharpen suffix margins under learning rate decay. During inference, TELOS decodes natively via greedy or ancestral sampling without rollouts, beam search alterations, or post-hoc reranking. Cached visual compatibilities $\{s_i^\ell\}$ avoid repeated patch scoring, incurring only $O(Nd)$ operations per active layer.

Method	Lexical				BERTScore	Clinical Graph			Clinical Diagnostic Accuracy				
	BLEU-1	BLEU-4	ROUGE-L	METEOR	BS	RG-F1	RG-ER	RadCliQ↓	CheXbert	m-F1-14	M-F1-14	m-F1-5	M-F1-5
General Medical VLMs													
LLaVA-Med	0.256	0.017	0.149	0.158	0.782	0.054	0.066	1.81	0.196	0.292	0.210	0.282	0.219
MedGemma1.5	0.300	0.072	0.207	0.275	0.820	0.214	0.238	1.25	0.388	0.536	0.361	0.596	0.518
General VLMs													
LLaVA-1.5	0.154	0.007	0.090	0.147	0.757	0.029	0.048	2.05	0.168	0.227	0.181	0.220	0.189
+ LoRA	0.274	0.080	0.185	0.264	0.818	0.158	0.192	1.48	0.293	0.343	0.246	0.356	0.314
+ Engram	0.267	0.109	0.279	0.270	0.850	0.189	0.229	1.21	0.302	0.217	0.141	0.218	0.171
+ TELOS (**Ours**)	**0.363**	**0.158**	**0.314**	**0.319**	**0.860**	**0.240**	**0.273**	**1.01**	**0.377**	**0.386**	**0.261**	**0.405**	**0.343**
Qwen2.5-VL-7B	0.246	0.021	0.156	0.212	0.788	0.088	0.111	1.76	0.217	0.319	0.221	0.343	0.302
+ LoRA	0.313	0.116	0.275	0.287	0.821	0.206	0.222	1.41	0.342	0.332	0.228	0.368	0.314
+ Engram	0.311	0.129	0.286	0.283	0.850	0.202	0.240	1.18	0.316	0.241	0.157	0.239	0.195
+ TELOS (**Ours**)	**0.325**	**0.132**	**0.294**	**0.300**	**0.853**	**0.222**	**0.256**	**1.09**	**0.360**	**0.359**	**0.228**	**0.384**	**0.319**
Radiology VLMs													
LLaVA-Rad	0.396	0.154	0.305	0.339	0.863	0.289	0.302	1.18	0.467	0.570	0.447	0.604	0.535
+ Engram	0.431	0.197	0.351	0.381	0.873	0.314	0.336	0.74	0.487	0.576	0.452	0.607	0.533
+ TELOS (**Ours**)	**0.437**	**0.202**	**0.354**	**0.384**	**0.874**	**0.318**	**0.340**	**0.73**	**0.490**	**0.579**	**0.452**	**0.608**	**0.538**

Table 1: **Main performance comparison on MIMIC-CXR across model families.** Progressively shaded green rows highlight our proposed TELOS adapted models across backbones. Bold font indicates superior performance within each backbone. m/M denote micro/macro averaging. Protocol details are in Appendix A.5.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We benchmark in-domain report generation on **MIMIC-CXR** (Johnson et al., 2019) (162,955 train, 1,286 val, 2,463 test frontal examinations). To assess cross-hospital generalization under distinct institutional styles, we benchmark on **CheXpert Plus** (Chambon et al., 2024) under the CXPMRG-Bench partition (Wang et al., 2025) (40,463 train, 5,780 val, 11,562 test) and zero-shot transfer on **IU-Xray** (Demner-Fushman et al., 2016) (3,955 reports with 7,470 radiographs; test split). Cross-modal generalization is evaluated on **COVCTR** (Li et al., 2023b) (728 chest CT scans with paired reports), where 2D axial slices enable evaluation via domain expression inventory adaptation.

VLMs and adaptation. We evaluate LLaVA-1.5-7B, Qwen2.5-VL-7B, and specialized LLaVA-Rad (Appendix A.9). Decoder-LoRA ($r = 8$, $\alpha = 16$, 20.0M parameters) provides the primary universal baseline across all backbones. We also compare against token-level memory grafting (Engram). For domain-adapted LLaVA-Rad, we freeze its weights and assess orthogonal gains. Table 1 additionally benchmarks general medical VLMs LLaVA-Med and MedGemma-1.5 (Appendix A.5).

Training and evaluation. Models are optimized via AdamW (10^{-3} initial learning rate, 5% linear warmup, cosine decay) for 3 epochs on MIMIC-CXR and 10 epochs on CheXpert Plus with per-GPU batch size 16 (Appendix A.9). Report generation follows LLaVA-Rad decoding parameters uniformly (Appendix A.5).

Report evaluation encompasses lexical metrics (BLEU-1/4 Papineni et al., 2002; ROUGE-L Lin, 2004; METEOR Banerjee & Lavie, 2005; BERTScore Zhang et al., 2019), clinical graph overlap (RadGraph-F1, RG-ER Jain et al., 2021), composite error (RadCliQ-v1 Yu et al., 2023), and CheXbert (Smit et al., 2020) diagnostic accuracy (Appendix A.5). Multi-token diagnostics follow in Section 4.4 and Appendix A.3.

4.2 REPORT GENERATION AND DOMAIN ADAPTATION

Adaptation across backbone families on MIMIC-CXR. Table 1 compares TELOS against domain-specialized VLMs, LoRA, and Engram across three backbones. On general LLaVA-1.5, TELOS achieves BLEU-4 0.158 and RadGraph-F1 0.240, substantially surpassing LoRA (0.080 and 0.158) and Engram (0.109 and 0.189) with 25% of matched LoRA parameters (5.82M vs. 20.0M), rivaling specialized LLaVA-Rad (RadCliQ 1.01 vs. 1.18). On Qwen2.5-VL, TELOS likewise outperforms LoRA across clinical metrics (RadCliQ 1.09 vs. 1.41) using 5.04M parameters. On specialized LLaVA-Rad, TELOS provides orthogonal gains, lowering RadCliQ to 0.73 (vs. 1.18 for base and 0.74 for Engram) with high diagnostic fidelity.

Cross-dataset and cross-modal generalization. On CheXpert Plus (Table 4, Appendix A.6), TELOS consistently improves clinical graph and diagnostic metrics over LoRA across backbones,

VLMs	Component	Lexical				BERTScore	Clinical Graph			Clinical Diagnostic Accuracy				
		BLEU-1	BLEU-4	ROUGE-L	METEOR	BS	RG-F1	RG-ER	RadCliQ↓	CheXbert	m-F1-14	M-F1-14	m-F1-5	M-F1-5
LLaVA-1.5 (7B)	Base VLM	0.154	0.007	0.090	0.147	0.757	0.029	0.048	2.05	0.168	0.227	0.181	0.220	0.189
	w/o Visual branch	0.322	0.131	0.277	0.271	0.848	0.194	0.233	1.22	0.303	0.226	0.148	0.222	0.179
	w/o Visual Re-Attention	0.323	0.132	0.278	0.272	0.848	0.194	0.234	1.21	0.307	0.237	0.162	0.239	0.196
	w/o $\mathcal{L}_{\text{OSC}}$	0.362	0.157	0.301	0.307	0.854	0.237	0.266	1.03	0.372	0.384	0.248	0.398	0.331
	TELOS (Full)	**0.363**	**0.158**	**0.314**	**0.319**	**0.860**	**0.240**	**0.273**	**1.01**	**0.377**	**0.386**	**0.261**	**0.405**	**0.343**
Qwen2.5-VL (7B)	Base VLM	0.246	0.021	0.156	0.212	0.788	0.088	0.111	1.76	0.217	0.319	0.221	0.343	0.302
	w/o Visual branch	0.254	0.100	0.265	0.258	0.844	0.181	0.220	1.27	0.288	0.200	0.131	0.177	0.145
	w/o Visual Re-Attention	0.235	0.090	0.257	0.252	0.842	0.173	0.213	1.30	0.286	0.192	0.129	0.167	0.134
	w/o $\mathcal{L}_{\text{OSC}}$	0.315	0.130	0.294	0.300	0.853	0.221	0.252	1.10	0.355	0.352	0.225	0.375	0.318
	TELOS (Full)	**0.325**	**0.132**	**0.294**	**0.300**	**0.853**	**0.222**	**0.256**	**1.09**	**0.360**	**0.359**	**0.228**	**0.384**	**0.319**
LLaVA-Rad (7B)	Base VLM	0.396	0.154	0.305	0.339	0.863	0.289	0.302	1.18	0.467	0.570	0.447	0.604	0.535
	w/o Visual branch	0.433	0.199	0.350	0.381	0.871	0.314	0.334	0.75	0.484	0.575	0.447	0.605	0.538
	w/o Visual Re-Attention	0.434	0.200	0.351	0.383	0.872	0.315	0.338	0.75	0.487	0.575	0.450	0.606	0.540
	w/o $\mathcal{L}_{\text{OSC}}$	0.435	0.200	0.352	0.383	0.873	0.316	0.339	0.75	0.489	0.577	0.451	0.605	0.537
	TELOS (Full)	**0.437**	**0.202**	**0.354**	**0.384**	**0.874**	**0.318**	**0.340**	**0.73**	**0.490**	**0.579**	**0.452**	**0.608**	**0.538**

Table 2: **Component comparisons across three multimodal backbones on MIMIC-CXR.** Shaded rows highlight the full TELOS configuration. Bold indicates best performance per backbone. In Qwen2.5-VL, Log-Mean-Exp pooling depends on updated visual states $\widetilde{V}^{\ell}$, explaining the gap when visual re-attention is ablated. Details in Appendix A.5.

demonstrating institutional robustness. Under zero-shot transfer on IU-Xray without in-domain tuning (Table 5, Appendix A.7), TELOS doubles diagnostic macro-F1 over LoRA on LLaVA-1.5 (0.143 vs. 0.076 across 14 conditions; 0.176 vs. 0.017 on 5 core conditions) and lowers RadCliQ to 1.05 (vs. 1.11), while reaching 0.91 on LLaVA-Rad. On chest CT (COVCTR; Table 6, Appendix A.8) where zero-shot VLMs collapse (BLEU-4 ≤ 0.011), swapping the inventory to CT terminology ($\mathcal{E}_{\text{CT}}$) and training on frozen backbones achieves 0.521 BLEU-4 on LLaVA-1.5 (vs. 0.506 for LoRA) and 0.487 on Qwen2.5-VL (vs. 0.483), establishing modular cross-modal adaptation.

4.3 Ablation Studies

Context, visual branch, and visual re-attention. Table 2 isolates component contributions. On LLaVA-1.5, removing visual branch drops RadGraph-F1 from 0.240 to 0.194 (-4.6 points), confirming reliance on radiographic grounding. On Qwen2.5-VL, visual branch without re-attention drops RadGraph-F1 to 0.173 (vs. 0.181 for context-only), validating that Log-Mean-Exp pooling (Eq. 5) requires updated visual states $\widetilde{V}^{\ell}$.

Suffix competition and its evidence. Table 2 tests whether memory alone suffices for suffix completion. On LLaVA-1.5, removing OSC drops RadGraph-F1 from 0.240 to 0.237 (ROUGE-L: $0.314 \rightarrow 0.301$). Because document-level metrics average across frequent single-token findings, they dilute localized suffix competition. Direct length-stratified diagnostics (Section 4.4, Figure 4) show that removing OSC incurs a 9.0%–10.5% drop in suffix continuation on multi-subword terms (≥ 3 tokens), proving its dedicated role in suppressing suffix competitors.

Intervention depth and hyperparameters. Evaluating intervention depths (Table 9 in Appendix A.9) confirms that multi-layer intervention ($\mathcal{L} = \{1, 15, 31\}$, RadGraph-F1 0.240) substantially outperforms single late-block injection (0.144), validating that expression memory must be maintained across early, middle, and late representations. Candidate sweeps similarly confirm $k = 8$ as optimal (Appendix A.9).

4.4 Expression Access and Realization

Subword token-length distribution agreement. To assess vocabulary length profiles without logarithmic frequency compression, Figure 3 evaluates word-length distribution agreement relative to genuine radiologist reports across 1,000 sampled examinations from MIMIC-CXR and CheXpert Plus (Appendix A.4). This metric measures relative frequency alignment across subword lengths, capturing macroscopic vocabulary profiles. Three empirical patterns emerge: (1) *Common-vocabulary preservation*: On high-frequency single-token words ($\sim$70–80% of vocabulary), TELOS maintains $\geq 97\%$ distribution agreement across backbones, indicating that explicit memory integration preserves standard word proportions. (2) *Intermediate term consistency*: Across intermediate multi-subword terms (2 to 5 tokens), standard fine-tuning degrades sharply (LoRA and Engram drop to 36%–57%), whereas TELOS sustains 80%–100% distribution agreement. (3) *Mitigating long-tail collapse*: On complex expressions requiring 4, 5, and $6+$ subwords, baselines degrade severely; under cross-hospital shift on CheXpert Plus (Figure 3, bottom row), Vanilla and LoRA drop to 0% agreement on

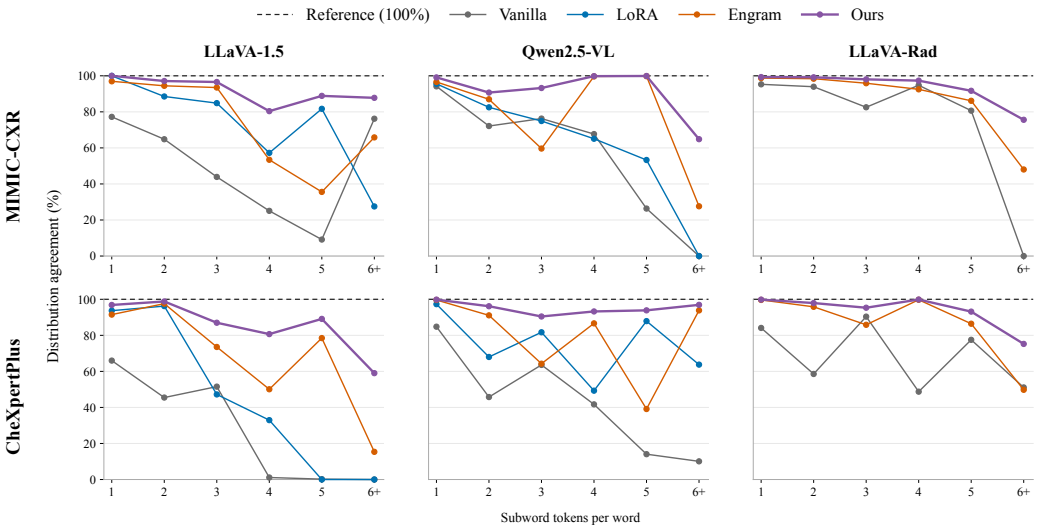


Figure 3: **Agreement with reference subword-length distributions.** Evaluated across subword lengths (1 to 6+ tokens per word) over 1,000 sampled examinations from MIMIC-CXR and CheXpert Plus (dashed line denotes the 100% reference baseline). Scores quantify proportional frequency alignment across token lengths against authentic radiologist text, evaluating distributional representation rather than token-level semantic correctness.

5–6+ token terms (reflecting zero occurrences in the sampled reports; Appendix A.4). In contrast, TELOS sustains 60%–97% distribution agreement at length 6+, stabilizing multi-token generation through explicit memory anchoring and targeted suffix contrast.

Controlled realization and multi-token terminology fidelity. To disentangle concept onset selection from sequential continuation, we evaluate expressions across token lengths (2 to 6+ subwords) under paired diagnostic protocols (Appendix A.3, Figure 4). Comparing full TELOS with its memory-only ablation (*w/o OSC loss*) directly isolates component mechanisms. Expression memory provides autonomous onset access, elevating unprompted generation on LLaVA-1.5 to 29.0%–30.5% on 2-token terms and 15.0%–21.0% at length 4 (where LoRA drops to 11.0% and base models collapse below 1%). However, memory alone remains vulnerable to suffix competition: removing OSC incurs an immediate 9.0%–10.5% drop in prompted completion across lengths 3–6+ (90.5% $\rightarrow$ 80.0% at length 3; 56.5% $\rightarrow$ 46.0% at length 4; Panel a), and lowers unprompted continuation on lengths 4–6+ by 6.0%–7.5% (Panel b). Enforcing suffix contrast resolves this competition, securing 31.8% Term F1 at length 4 (vs. 16.5% for LoRA) and sustaining positive tail F1 on Qwen2.5-VL (3.8% at length 6+ vs. 0.0%). Qualitative cases (Appendix A.10) show finding recovery without template collapse.

5 CONCLUSION

We presented TELOS, an expression-level adaptation framework supporting specialized clinical report generation on frozen vision–language models. By pairing explicit latent expression memory with targeted Onset-to-Suffix Contrast (OSC), TELOS bridges narrative fluency with multi-token terminology realization on frozen decoders, directly addressing the onset access and sequential continuation bottlenecks without full fine-tuning. Across four benchmarks, TELOS achieves consistent gains in clinical report fidelity, terminology realization across subword lengths, and parameter efficiency (5.82M parameters). These findings demonstrate that structured expression-level interfaces provide a modular, parameter-efficient alternative to decoder weight adaptation. Future work includes expanding expression inventories dynamically to open-vocabulary rare pathologies and scaling the routing interface to volumetric 3D imaging sequences.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025. doi: 10.48550/arXiv.2502.13923. URL https://arxiv.org/abs/2502.13923.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pp. 2206–2240. PMLR, 2022.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pp. 89–96, 2005.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P Langlotz. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. *arXiv preprint arXiv:2405.19538*, 2024.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 1439–1449, 2020.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 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)*, pp. 5904–5914, 2021.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. Chexagent: Towards a foundation model for chest x-ray interpretation. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024.

Runxi Cheng, Yuchen Guan, Yongxian Wei, Qianpu Sun, Qixiu Li, Sinan Du, Feng Xiong, Chun Yuan, Yan Lu, and Yeyun Gong. Memory grafting: Scaling language model pre-training via offline conditional memory. *arXiv preprint arXiv:2605.20948*, 2026a.

Xin Cheng, Wangding Zeng, Damai Dai, Qinyu Chen, Bingxuan Wang, Zhenda Xie, Kezhao Huang, Xingkai Yu, Zhewen Hao, Han Zhang, et al. Conditional memory via scalable lookup: A new axis of sparsity for large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4968–4990, 2026b.

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

Wenjun Hou, Yi Cheng, Kaishuai Xu, Heng Li, Yan Hu, Wenjie Li, and Jiang Liu. Radar: Enhancing radiology report generation with supplementary knowledge injection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26366–26381, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. 2021.

Stephanie L Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, et al. Maira-1: A specialised large multimodal model for radiology report generation. *arXiv preprint arXiv:2311.13668*, 2023.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. 2021.

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. Promptmrg: Diagnosis-driven prompts for medical report generation. 38(3):2607–2615, 2024.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. 2019.

Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 conference on empirical methods in natural language processing: System demonstrations*, pp. 66–71, 2018.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. volume 33, pp. 9459–9474, 2020.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. volume 36, pp. 28541–28564, 2023a.

Mingjie Li, Rui Liu, Fuyu Wang, Xiaojun Chang, and Xiaodan Liang. Auxiliary signal-guided knowledge encoder-decoder for medical report generation. *World Wide Web*, 26(1):253–270, 2023b.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, et al. Medgemma 1.5 technical report. *arXiv preprint arXiv:2604.05081*, 2026.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 1715–1725, 2016.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 1500–1519, 2020.

Xiao Wang, Fuling Wang, Yuehang Li, Qingchuan Ma, Shiao Wang, Bo Jiang, and Jin Tang. Cxpmrg-bench: Pre-training and benchmarking for x-ray medical report generation on chexpert plus dataset. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5123–5133. IEEE, 2025.

Yuhuai Wu, Markus N Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. 2022.

Shuxin Yang, Xian Wu, Shen Ge, S Kevin Zhou, and Li Xiao. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical image analysis*, 80:102510, 2022.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, et al. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9), 2023.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, et al. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16(1):3108, 2025.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. 2019.

A Supplementary Analysis and Experimental Details

A.1 Visual Read and Shared Parameters

Let $Z \in \mathbb{R}^{P \times d}$ denote fixed post-projector visual features. For $H = 8$ heads of dimension $d_h = d/H$, current image states V^ℓ query Z immediately before the active decoder block:

$$Q_h^\ell = \mathrm{split}_h(\mathrm{RMSNorm}(V^\ell)), \qquad K_h = \mathrm{RMSNorm}(Z_h W_K), \qquad V_h = Z_h W_V,$$
$$A_h^\ell = \mathrm{softmax}(Q_h^\ell K_h^\top / \sqrt{d_h}) V_h, \qquad G^\ell = \sigma(\mathrm{RMSNorm}(V^\ell) W_G^\ell + b_G^\ell). \tag{11}$$

Here $W_K, W_V \in \mathbb{R}^{d_h \times d_h}$ are shared across heads and layers, and $W_G^\ell \in \mathbb{R}^{d \times H}$ is layer-specific. The updated rows are $\widetilde{V}^\ell = V^\ell + \mathrm{Concat}_{h=1}^{H}(G_h^\ell \odot A_h^\ell)$. Key/value maps start at identity, gate weights at zero, and gate biases at -4. The shared depthwise writer has width four, causal left padding, no affine normalization parameters, and zero-initialized convolution weights. The visual pooling temperature is 0.07 in the experiment configuration. The cosine scale starts at 8 and is clipped to $[1, 64]$; visual coefficients start at 0.1 and are bounded in $(0, 1)$.

Intervention depths and trainable parameter counts. Interventions occur immediately before self-attention at sparse depths: $\{1, 15, 31\}$ for 32-layer architectures (LLaVA-1.5 and LLaVA-Rad) and $\{1, 13, 27\}$ for 28-layer architectures (Qwen2.5-VL). Banks, visual projections, and convolutions are shared across depths; head gates and visual coefficients are layer-specific. Table 3 summarizes the parameter footprint. The frozen VLM includes its vision encoder, multimodal projector, and decoder; only the TELOS module is updated. The explicit zero-valued abstention state introduces no additional parameters.

Backbone	TELOS (M)	LoRA (M)
LLaVA-1.5	5.82	20.0
Qwen2.5-VL	5.04	20.2
LLaVA-Rad	5.82	–[†]

Table 3: Trainable parameter counts for TELOS versus LoRA. All models share the same 1,265-entry common expression inventory. For general backbones (LLaVA-1.5 and Qwen2.5-VL), we report our matched decoder-LoRA baseline ($r = 8$, attention and MLP projections across all decoder layers; counts differ due to layer counts and GQA dimensions). [†]Because LLaVA-Rad is an already domain-adapted foundation model, we evaluate TELOS directly on its frozen weights as an orthogonal plug-in without introducing an additional LoRA baseline.

A.2 Medical-Word Diagnostic Protocol

The diagnostic inventory uses medical single words drawn using the original clinical curation rules and reviewed standalone components. Each selected word has 30 distinct training-report contexts. Selection targets 30 words per length bin, using training document-frequency strata and category diversity before model scoring; scarce bins retain all selected words. Six-token and longer words are pooled before averaging. For bins $2/3/4/5/6+$, LLaVA-1.5 and Rad have $30/30/30/18/10$ words; Qwen has $30/30/29/5/6$. Each model pair uses identical contexts within a backbone, while tokenizer differences mean that bins across backbones can contain different words.

For a word in reference context b, onset probability is $p(w_1 \mid I, b)$. Completion is a binary surface match with a valid lexical boundary under native greedy continuation after supplying w_1. Gold token-path probability is Eq. 2, excluding the following boundary probability. It compounds conditional probabilities with length and is distinct from free-generation recall. We average contexts within words, then equally average words within bins. The diagnostic evaluates expression behavior in training contexts; held-out report quality is evaluated separately.

As shown in Figure 4 and detailed below, the diagnostic protocol compares full TELOS against Vanilla, LoRA, Engram, and the memory-only ablation (*w/o OSC loss*), highlighting how explicit memory enables concept onset initiation while OSC resolves sequential suffix continuation across general (LLaVA-1.5, Qwen2.5-VL) and specialized (LLaVA-Rad) backbones.

A.3 Fine-Grained Term Realization and Generation Across Model Families and Datasets

To provide an exhaustive characterization of multi-token medical terminology realization and report-level generation fidelity, we evaluate all models across five complementary dimensions stratified by term token length ($2, 3, 4, 5, 6+$ subwords). All five metrics are reported as percentages (%), where higher values denote superior performance. Figures 4 and 5 present the complete comparison across three multimodal backbones (LLaVA-1.5, Qwen2.5-VL, and LLaVA-Rad) on both in-domain MIMIC-CXR and cross-institutional CheXpert Plus.

Evaluation protocols across the five dimensions. The five evaluation panels evaluate:

1. **Completion with first token given (Panel a):** Evaluates controlled suffix continuation. Given radiograph I, context $c_{<t}$, and first subword w_1, greedy generation succeeds if emitted tokens match reference suffix $w_{2:m}$ up to the lexical boundary.
2. **Completion without first token (Panel b):** Evaluates unprompted full-term generation. The model receives only the radiograph I and the preceding reference context $c_{<t}$ without the initial token. A trial is marked successful if native greedy generation reproduces the entire multi-token clinical term $w_{1:m}$ from onset to suffix boundary.
3. **Term precision (P, Panel c):** In unconstrained free report generation on held-out test examinations, $P = |\mathcal{T}_{\text{gen}} \cap \mathcal{T}_{\text{ref}}|/|\mathcal{T}_{\text{gen}}|$, measuring the fraction of extracted clinical terms in the candidate report that are present in the ground-truth reference report.
4. **Term recall (R, Panel d):** In unconstrained report generation, $R = |\mathcal{T}_{\text{gen}} \cap \mathcal{T}_{\text{ref}}|/|\mathcal{T}_{\text{ref}}|$, measuring the fraction of reference clinical terms successfully recovered in the generated report.
5. **Term F1 (Panel e):** The harmonic mean of precision and recall ($2PR/(P + R)$), providing a unified measure of clinical entity correctness and coverage.

Empirical findings on MIMIC-CXR. Figure 4 presents comparative results on MIMIC-CXR across all three backbones:

1. **Concept onset initiation via explicit memory:** Comparing panels (a) and (b) exposes the onset access bottleneck. While gold onsets allow LoRA to reach 92.0% completion on 2-token terms, unprompted generation collapses on baselines (LoRA drops to 8.0% at length 3; Engram $< 5\%$). Both *w/o OSC loss* and full TELOS sustain 12.0%–30.5% unprompted generation across lengths 2–6+ (e.g., 29.0%–30.5% at length 2 and 15.0%–21.0% at length 4 on LLaVA-1.5), verifying that explicit memory routing successfully cues whole concepts into decoder representations without prompting.
2. **Isolating OSC: overcoming compounding suffix competition:** Contrasting full TELOS with *w/o OSC loss* isolates the causal role of suffix contrast. On 2-token terms, memory alone matches full TELOS (94.0% completion in panel a), as single-token suffixes face minimal competition. However, on multi-subword terms ($\geq$ 3 tokens), removing OSC degrades prompted continuation by 9.0%–10.5% on LLaVA-1.5 ($90.5\% \rightarrow 80.0\%$ at length 3; $56.5\% \rightarrow 46.0\%$ at length 4; $70.0\% \rightarrow 61.0\%$ at length 6+), and drops unprompted completion on lengths 4–6+ by 6.0%–7.5%. On Qwen2.5-VL, OSC sustains positive tail F1 (3.8%) and precision (9.2%) at length 6+, where *w/o OSC loss* drops to 2.0% and baselines collapse to 0.0%. This demonstrates that OSC's margin separation prevents greedy divergence on trailing subwords.
3. **Orthogonal improvements on specialized foundation models:** On LLaVA-Rad (Figure 4(3)), while domain pretraining yields strong gold-onset completion (62.0%–78.0%), unprompted generation remains constrained (15.0%–30.0%). TELOS boosts unprompted generation to 40.0% at length 6+ (vs. 31.0% for base LLaVA-Rad) and raises Term F1 across lengths (e.g., 26.8% vs. 16.8% at length 4, and 27.0% vs. 24.0% at length 6+ over *w/o OSC loss*), proving that memory routing and suffix contrast provide complementary inductive biases even on specialized weights.

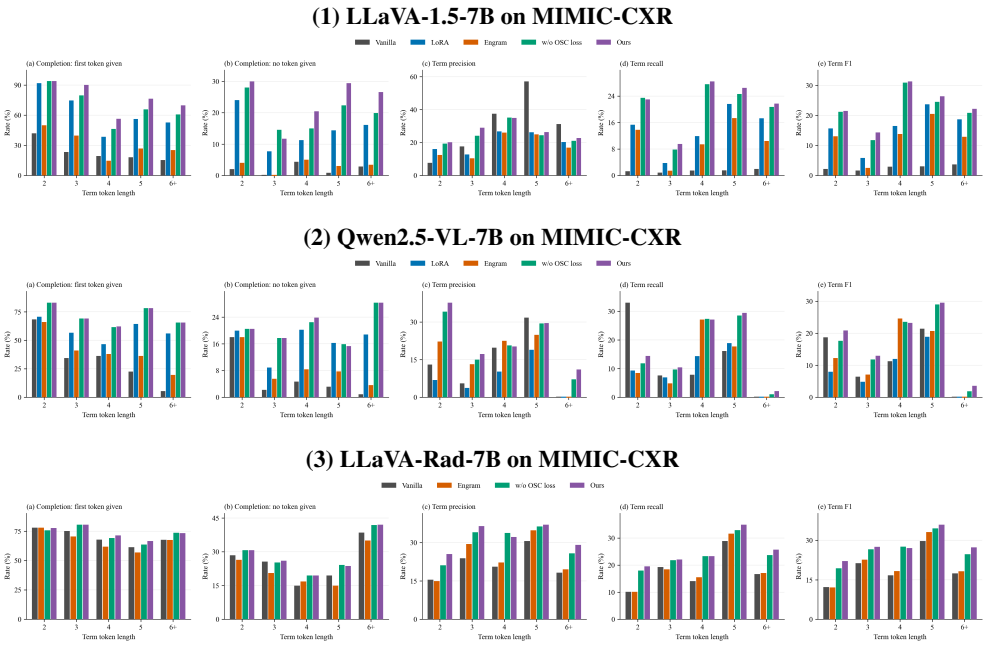


Figure 4: **Clinical term realization and generation across three backbones on MIMIC-CXR.** Evaluated across token lengths under the five protocol dimensions defined above (panels a–e; all metrics in %, higher is better). Each panel compares five configurations: Vanilla, LoRA, Engram, the memory-only ablation (*w/o OSC loss*), and full TELOS (Ours). Contrasting *Ours* with *w/o OSC loss* directly isolates the contribution of Onset-to-Suffix Contrast in overcoming multi-token continuation bottlenecks across LLaVA-1.5 (1), Qwen2.5-VL (2), and LLaVA-Rad (3).

Cross-hospital generalization on CheXpert Plus. Figure 5 presents the results on CheXpert Plus, which evaluates adaptation under institutional distribution shift (Stanford Hospital versus Beth Israel Medical Center). Two key findings emerge:

1. **Vulnerability of standard fine-tuning to institutional shift:** Standard LoRA fine-tuning exhibits severe performance degradation on multi-token clinical expressions under institutional shift. On LLaVA-1.5 (Figure 5(1)), LoRA's completion with gold onset drops to 19.0% on 5-token terms (vs. 52.0% for TELOS), and its free-generation Term F1 collapses to 0.8% at length 5. Engram similarly degrades below 5% F1 across lengths 2–5. Under continuous weight updates, LoRA overfits to dominant institutional phrasing patterns and struggles when target phrasing varies.
2. **Institutional invariance via decoupled expression memory:** Because TELOS stores pathology definitions as invariant latent memory anchors decoupled from decoder weights, it demonstrates superior transferability across hospital styles. On LLaVA-1.5, TELOS maintains 47.0%–72.0% completion under gold onsets (outperforming *w/o OSC loss* by 8.0%–12.0% on multi-subword lengths 3–5, e.g., 61.0% vs. 51.0% at length 3 and 57.0% vs. 45.0% at length 4; panel a) and sustains 13.0%–23.0% unprompted completion. On Qwen2.5-VL and LLaVA-Rad (Figure 5(2, 3)), TELOS likewise dominates completion and Term F1 across token lengths (e.g., 31.8% F1 at length 5 on LLaVA-Rad vs. 24.2% for base), verifying that memory routing and suffix contrast remain mutually reinforcing under cross-hospital shift.

A.4 SUBWORD TOKEN LENGTH DISTRIBUTION OF GENERATED REPORTS

Detailed distribution agreement across 1,000 sampled reports. Following the distribution agreement metric defined in Section 4.4, Figure 3 evaluates vocabulary alignment across 1,000 randomly sampled held-out reports from each dataset (MIMIC-CXR and CheXpert Plus) across all three multimodal backbones (LLaVA-1.5, Qwen2.5-VL, and LLaVA-Rad). Evaluating proportional occurrence deviations relative to authentic radiologist references provides an uncompressed view of vocabulary

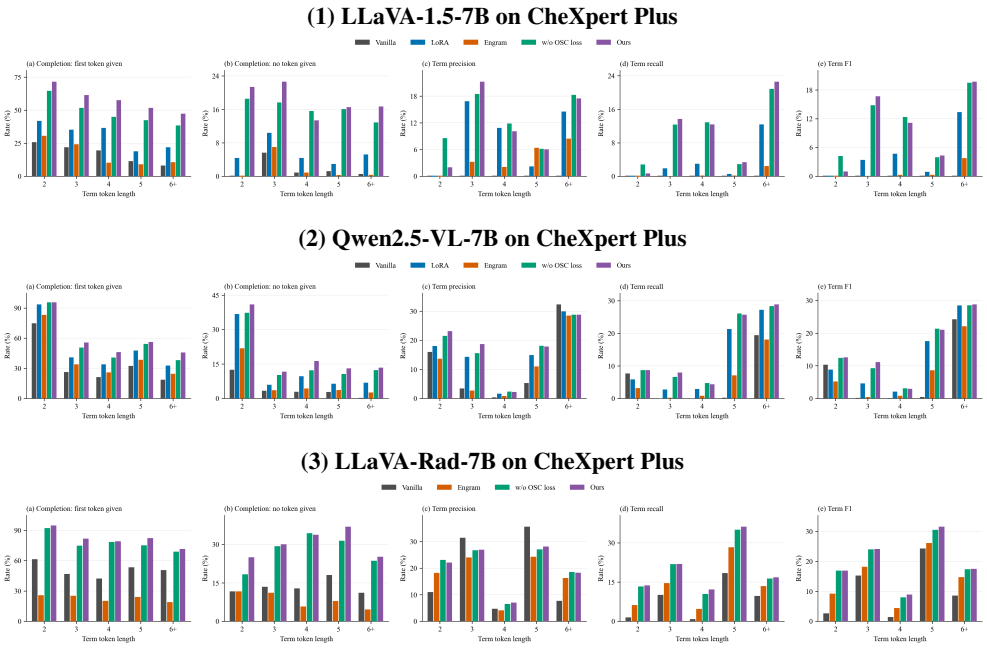


Figure 5: **Clinical term realization and generation on CheXpert Plus under institutional shift.** Multi-backbone evaluation across token lengths following the five protocol dimensions and layout of Figure 4, comparing Vanilla, LoRA, Engram, the memory-only ablation (*w/o OSC loss*), and full TELOS (Ours).

length distribution. By construction, the metric yields 0% when a model produces zero tokens of a given length, or when its relative frequency exceeds twice the reference proportion ($p_{\text{method}} \geq 2p_{\text{ref}}$); in the long-tail evaluations below, 0% scores correspond to complete absence ($p_{\text{method}} = 0$).

Three detailed empirical findings corroborate the main-text evaluation:

1. **Preservation of single-token words ($\geq 97\%$):** Across all architectures on both datasets, TELOS achieves 97.0%–100.0% distribution agreement at length 1 (representing $\sim$ 70–80% of all words in authentic clinical text). This confirms that our explicit expression memory and Bilateral Evidence Router (via its abstention anchor $q_{\perp}$) operate without distorting general connective prose and frequent clinical syntax. In contrast, unadapted base models (Vanilla) drop to 65.0%–85.0% agreement.
2. **Baseline degradation on intermediate multi-token terms:** As subword length scales from 2 to 5 tokens, standard fine-tuning exhibits severe agreement degradation. On LLaVA-1.5 under MIMIC-CXR, LoRA drops from 97% at length 1 down to 57.4% at length 4, while Engram drops from 97% to 36.0% at length 5. On Qwen2.5-VL, LoRA drops to 53.2% at length 5, and Engram drops to 27.8% at length 6+. In contrast, TELOS remains stable, maintaining 80.5%–100.0% distribution agreement across intermediate lengths (2–5 tokens) across all backbones.
3. **Mitigating long-tail collapse under institutional shift:** Under cross-hospital distribution shift on CheXpert Plus (Figure 3, bottom row), baseline failures become acute. On LLaVA-1.5, Vanilla drops to 1.2% at length 4 and collapses to 0.0% at lengths 5 and 6+ (reflecting zero generated instances). LoRA also drops to 0.0% distribution agreement at lengths 5 and 6+, having omitted longer clinical terms. Engram plummets to 15.4% at length 6+. On Qwen2.5-VL and LLaVA-Rad, Vanilla similarly drops to 0.0%–10.0% in the tail. In contrast, TELOS sustains 59.0%–97.0% distribution agreement at length 6+ across backbones, demonstrating robust cross-institutional generalization on complex terminology.

Mechanistic explanation: resolving compounding autoregressive decay. These distributional behaviors elucidate why conventional autoregressive generation struggles with specialized medical vocabulary. Under standard cross-entropy without phrase-level memory, generating an m-token

clinical expression requires traversing an m-step conditional probability path $\prod_{j=1}^{m} p(w_j \mid I, c, w_{<j})$. Because each successive subword token compounds uncertainty, greedy decoding frequently aborts multi-token words entirely or diverts into simpler, single-token generic synonyms (e.g., substituting *"infection"* for *"bronchiectasis"*), causing baselines to produce zero occurrences of long-tail expressions. TELOS overcomes this bottleneck through its dual mechanism: (1) the Bilateral Evidence Router activates whole-concept expression memory anchors that provide persistent, invariant latent guidance throughout the multi-step generation trajectory; and (2) Onset-to-Suffix Contrast enforces a focused logit margin against competing non-gold suffixes, guarding against greedy diversion and premature termination. Consequently, TELOS enables general and specialized VLMs alike to generate reports whose vocabulary depth and complexity authentically mirror genuine clinical documentation without dropping longer clinical terms.

A.5 REPORT METRICS AND EXPERIMENTAL VARIANTS

We use BLEU-1 and BLEU-4, ROUGE-L, METEOR, and BERTScore for text similarity; RadGraph-F1 and RG-ER for graph overlap; raw RadCliQ-v1 for composite predicted error; and CheXbert vector cosine similarity. CheXbert extracts labels for the 14 CheXpert categories, including No Finding. Positive and uncertain labels are mapped to positive, with negative and blank mapped to negative. The five-category subset is Cardiomegaly, Edema, Consolidation, Atelectasis, and Pleural Effusion. Micro precision/recall pool category decisions; macro scores average the per-category quantities. Macro F1 is the mean of category F1 scores, not the harmonic mean of macro precision and recall. Raw RadCliQ is computed from its standardized features and linear coefficients, without a subsequent score normalization.

All primary comparisons use greedy decoding with repetition penalty 1 and a 1,024-token continuation cap. During training, the VLM backbone (vision encoder, multimodal projector, and decoder) remains completely frozen while only TELOS parameters are updated. The component ablation table reports the finalized model-specific inventories and configurations; exploratory variants are excluded from the primary comparison.

A.6 EVALUATION ON CHEXPERT PLUS

To evaluate adaptation on CheXpert Plus (Chambon et al., 2024), Table 4 presents full benchmark results across all three VLM families under the CXPMRG-Bench partition (Wang et al., 2025).

	Lexical				BERTScore	Clinical Graph			Clinical Diagnostic Accuracy				
Method	**BLEU-1**	**BLEU-4**	**ROUGE-L**	**METEOR**	**BS**	**RG-F1**	**RG-ER**	**RadCliQ↓**	**CheXbert**	**m-F1-14**	**M-F1-14**	**m-F1-5**	**M-F1-5**
General Medical VLMs													
LLaVA-Med	0.231	0.017	0.146	0.170	0.781	0.061	0.084	1.76	0.243	0.330	0.149	0.436	0.229
MedGemma1.5	0.172	0.029	0.147	0.267	0.795	0.152	0.174	1.64	0.317	0.539	0.359	0.515	0.457
General VLMs													
LLaVA-1.5	0.200	0.012	0.131	0.169	0.755	0.039	0.047	1.98	0.204	0.105	0.055	0.092	0.062
+ LoRA	0.232	0.057	0.196	0.226	0.816	0.165	0.197	1.42	0.309	0.436	0.254	0.383	0.306
+ Engram	0.182	0.042	0.172	0.174	0.807	0.125	0.170	1.59	0.235	0.251	0.130	0.187	0.126
+ TELOS (**Ours**)	**0.302**	**0.075**	**0.202**	**0.230**	**0.823**	**0.167**	**0.200**	**1.36**	**0.322**	**0.449**	**0.267**	**0.397**	**0.335**
Qwen2.5-VL-7B	0.084	0.009	0.083	0.193	0.755	0.070	0.081	2.28	0.134	0.230	0.142	0.199	0.188
+ LoRA	0.159	0.038	0.180	0.219	0.812	0.158	0.192	1.48	0.290	0.409	0.238	0.385	0.305
+ Engram	0.172	0.037	0.167	0.172	0.808	0.125	0.170	1.61	0.234	0.269	0.153	0.215	0.171
+ TELOS (**Ours**)	**0.303**	**0.074**	**0.199**	**0.227**	**0.821**	**0.162**	**0.195**	**1.39**	**0.313**	**0.438**	**0.256**	**0.395**	**0.315**
Radiology VLMs													
LLaVA-Rad	0.245	0.035	0.179	0.200	0.810	0.128	0.147	1.48	0.321	0.522	0.365	0.521	0.448
+ Engram	0.320	0.088	0.225	0.257	0.830	0.220	0.246	1.17	0.410	0.551	0.367	0.502	0.421
+ TELOS (**Ours**)	**0.345**	**0.100**	**0.240**	**0.273**	**0.836**	**0.241**	**0.258**	**1.09**	**0.427**	**0.577**	**0.399**	**0.531**	**0.464**

Table 4: **Main performance comparison on CheXpert Plus across model families.** Progressively shaded green rows highlight our proposed TELOS adapted models across backbones. Bold font indicates superior performance within each backbone. m/M denote micro/macro averaging. Protocol details are in Appendix A.5.

A.7 ZERO-SHOT EVALUATION ON IU-XRAY

The **IU-Xray** benchmark (Demner-Fushman et al., 2016) comprises 3,955 radiology reports paired with 7,470 frontal and lateral chest radiographs from Indiana University Hospital, partitioned under a standard 7:1:2 train/validation/test split. To evaluate cross-hospital transferability without in-domain adaptation, all models trained on MIMIC-CXR are evaluated directly on the IU-Xray test split without in-domain fine-tuning.

Continuous weight fine-tuning versus decoupled memory grounding. As shown in Table 5, unadapted general VLMs exhibit low clinical alignment under zero-shot transfer (e.g., base LLaVA-1.5 attains only 0.032 RadGraph-F1 and a 2.04 RadCliQ error). While standard LoRA fine-tuning improves surface fluency, it exhibits limited sensitivity on low-prevalence findings, attaining only 0.076 macro-F1 across 14 CheXbert pathologies and 0.017 on the 5 core conditions.

In contrast, TELOS achieves superior zero-shot transferability across all backbones. On LLaVA-1.5, TELOS achieves 0.200 RadGraph-F1 (vs. 0.195 for LoRA) and drives RadCliQ error down to 1.05 (vs. 1.11 for LoRA). On Qwen2.5-VL, TELOS reaches 0.221 RadGraph-F1 and a composite error of 0.96 (outperforming both LoRA's 1.06 and Engram's 1.10). This empirical advantage aligns directly with three architectural properties of our framework:

1. **Decoupled invariant clinical concepts:** Core diagnostic entities (e.g., *cardiomegaly*, *pleural effusion*, *pneumothorax*) represent invariant radiographic definitions across institutions. By storing these entities as discrete latent anchors in external expression memory rather than entangling them into continuous autoregressive decoder weights, TELOS preserves pristine semantic representations that remain unaffected by hospital-specific phrasing.
2. **Abstention-mediated routing agility:** When generating connective text and institutional transitions, the Bilateral Evidence Router activates its explicit abstention state ($q_\perp$), smoothly suppressing memory reads to allow the base VLM's general-language competence to generate natural syntactic bridges without spurious concept retrieval.
3. **Sensitivity on low-prevalence findings:** The macro-averaged diagnostic accuracy (M-F1-14) jumps from 0.076 with LoRA to 0.143 with TELOS on LLaVA-1.5, while 5-disease macro-F1 (M-F1-5) rises from 0.017 to 0.176. This marked increase confirms that visual branch routing directly anchors true underlying pathology, rather than reproducing dominant template frequencies.

Orthogonal benefits for specialized medical backbones. Applied to LLaVA-Rad (an already domain-adapted radiology foundation model), TELOS yields immediate complementary improvements: RadCliQ error drops from 1.14 to 0.91, and CheXbert diagnostic accuracy increases from 0.401 to 0.515. This demonstrates that even when a model has undergone extensive biomedical pretraining, explicit expression-level latent memory provides an orthogonal inductive bias that guards against completion collapse and enhances zero-shot clinical robustness.

	Lexical				BERTScore	Clinical Graph			Clinical Diagnostic Accuracy				
Method	**BLEU-1**	**BLEU-4**	**ROUGE-L**	**METEOR**	**BS**	**RG-F1**	**RG-ER**	**RadCliQ↓**	**CheXbert**	**m-F1-14**	**M-F1-14**	**m-F1-5**	**M-F1-5**
General Medical VLMs													
LLaVA-Med	0.154	0.010	0.112	0.180	0.784	0.050	0.097	1.89	0.097	0.065	0.056	0.099	0.063
MedGemma1.5	0.161	0.038	0.176	0.318	0.812	0.245	0.291	1.13	0.543	0.468	0.304	0.447	0.339
General VLMs													
LLaVA-1.5	0.127	0.014	0.101	0.171	0.732	0.032	0.050	2.04	0.211	0.181	0.042	0.082	0.023
+ LoRA	0.174	0.021	0.213	0.182	0.832	0.195	0.297	1.11	0.439	0.428	0.076	0.034	0.017
+ Engram	0.216	0.024	0.221	0.194	0.836	0.198	0.300	1.09	0.435	0.418	0.055	0.026	0.010
+ TELOS (**Ours**)	**0.223**	**0.032**	**0.224**	**0.206**	**0.834**	**0.200**	0.297	**1.05**	**0.459**	**0.442**	**0.143**	**0.206**	**0.176**
Qwen2.5-VL-7B	0.061	0.011	0.075	0.202	0.759	0.126	0.131	2.02	0.326	0.080	0.056	0.125	0.096
+ LoRA	0.210	0.030	0.224	0.204	0.836	0.196	0.297	1.06	0.461	0.418	0.101	0.140	0.119
+ Engram	0.167	0.020	0.216	0.184	0.832	0.197	0.301	1.10	0.435	0.394	0.064	0.073	0.042
+ TELOS (**Ours**)	**0.225**	**0.046**	**0.242**	**0.228**	**0.843**	**0.221**	**0.317**	**0.96**	**0.483**	**0.422**	**0.106**	**0.147**	**0.131**
Radiology VLMs													
LLaVA-Rad	0.363	**0.072**	0.232	0.277	0.831	0.197	0.238	1.14	0.401	0.517	0.314	0.399	**0.434**
+ Engram	0.354	0.070	0.240	0.265	0.844	0.224	0.294	0.93	0.511	0.514	0.318	0.445	0.382
+ TELOS (**Ours**)	**0.367**	**0.072**	**0.246**	**0.281**	**0.845**	**0.243**	**0.306**	**0.91**	**0.515**	**0.520**	**0.336**	**0.453**	0.389

Table 5: **Zero-shot performance on IU-Xray across model families.** Models trained on MIMIC-CXR are evaluated directly on IU-Xray without fine-tuning. Progressively shaded green rows highlight our proposed TELOS adapted models across backbones. Bold font indicates superior performance within each backbone (m/M denote micro/macro averaging; Appendix A.5).

A.8 CROSS-MODAL EVALUATION ON CHEST CT (COVCTR)

While our primary benchmarks focus on 2D projection chest radiography (MIMIC-CXR, CheXpert Plus, and IU-Xray), clinical report generation often spans diverse imaging modalities. To evaluate whether TELOS transfers across modalities, we benchmark on **COVCTR** (Li et al., 2023b), a clinical report generation dataset consisting of 728 computed tomography (chest CT) examinations with paired clinical reports, partitioned under the identical 7:1:2 train/validation/test split. Crucially, although computed tomography represents a distinct imaging modality from radiography, its data

format consists of 2D axial CT slices, which directly match standard 2D vision backbones without requiring volumetric architectures.

Cross-modal transfer via domain expression inventory. Chest CT scans capture cross-sectional parenchymal patterns described through specialized terminology, including ground-glass opacities (GGO), consolidation, crazy-paving patterns, vascular enlargement, and interlobular septal thickening. Consequently, TELOS achieves cross-modal transfer simply by adapting the expression inventory rather than altering the core model architecture: we extract a dedicated domain-specific expression inventory $\mathcal{E}_{\mathrm{CT}}$ directly from the COVCTR training split using the identical unsupervised clinical phrase mining pipeline (Section 3.1). By substituting the expression bank with target CT terminology and training the lightweight interface on COVCTR while keeping the base VLM frozen, the model is effectively grounded in CT semantics.

Cross-modal findings and mechanistic analysis. Table 6 presents the comparative evaluation on the COVCTR test set across general and medical foundation models. Because clinical graph extractors (RadGraph and CheXbert) were trained strictly on chest X-ray annotations and do not apply to CT terminology, evaluation follows standard NLG benchmarks (BLEU-1–4, ROUGE-L, METEOR, and BERTScore). Beyond raw metric gains, four critical empirical findings emerge:

1. **Failure of off-the-shelf medical VLMs across modalities:** Zero-shot medical foundation models (LLaVA-Med and MedGemma1.5) collapse on chest CT, attaining only 0.006 and 0.011 BLEU-4, respectively. This demonstrates that representations and lexicons acquired from planar 2D thoracic radiographs do not automatically transfer to cross-sectional CT findings without targeted domain adaptation.
2. **Modular adaptation with frozen backbones:** Rather than fine-tuning massive vision–language parameters, TELOS adapts to chest CT by swapping the expression inventory to $\mathcal{E}_{\mathrm{CT}}$ and training the lightweight interface on the target domain. With the general VLM backbone completely frozen, TELOS achieves 0.521 BLEU-4 and 0.671 ROUGE-L on LLaVA-1.5, and 0.487 BLEU-4 on Qwen2.5-VL, outperforming both full decoder-LoRA fine-tuning (0.506 and 0.483 BLEU-4) and Engram (0.510 and 0.477 BLEU-4). This confirms that cross-modal adaptation can be achieved effectively through a modular expression memory interface.
3. **Sustained multi-token realization via OSC on specialized CT terminology:** Compound CT findings span multi-subword sequences. Under standard cross-entropy, models often emit an initial token but derail on subsequent morphological pieces. By contrast, Onset-to-Suffix Contrast (OSC) reinforces continuation probabilities along suffix trajectories, ensuring cohesive realization of multi-word CT findings once an onset is triggered.
4. **Visual compatibility routing on 2D CT slices:** The Bilateral Evidence Router operates directly on 2D CT slice representations through visual compatibility scoring and re-attention, confirming that visual evidence routing generalizes across distinct 2D imaging modalities.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	BERTScore
General Medical VLMs							
LLaVA-Med	0.158	0.068	0.018	0.006	0.130	0.126	0.780
MedGemma1.5	0.223	0.097	0.034	0.011	0.164	0.162	0.793
General VLMs							
LLaVA-1.5	0.252	0.132	0.068	0.044	0.188	0.193	0.762
+ LoRA	0.702	0.623	0.559	0.506	0.641	0.661	0.939
+ Engram	0.696	0.620	0.560	0.510	0.651	0.675	0.941
+ TELOS (**Ours**)	**0.707**	**0.633**	**0.572**	**0.521**	**0.671**	**0.675**	**0.944**
Qwen2.5-VL-7B	0.187	0.085	0.035	0.014	0.149	0.207	0.783
+ LoRA	0.690	0.607	0.539	0.483	0.616	0.630	0.931
+ Engram	0.684	0.603	0.535	0.477	0.617	0.629	0.928
+ TELOS (**Ours**)	**0.697**	**0.612**	**0.543**	**0.487**	**0.622**	**0.636**	**0.936**

Table 6: **Cross-modal performance comparison on chest CT (COVCTR) across model families.** Models evaluated under domain inventory $\mathcal{E}_{\mathrm{CT}}$. Progressively shaded green rows highlight our proposed TELOS adapted models across backbones. Bold font indicates superior performance within each backbone. All scores follow the 0–1 scale consistent with Table 1.

A.9 HYPERPARAMETER SENSITIVITY AND LAYER TOPOLOGY

We evaluate the sensitivity of TELOS to the Onset-to-Suffix Contrast (OSC) hyperparameters on LLaVA-1.5, investigating the loss weight $\lambda \in \{0.01, 0.02, 0.05, 0.10\}$ and candidate pool size $k \in \{1, 4, 8, 16\}$. The loss-weight sweep fixes $k = 8$ (Table 7), while the candidate-count sweep fixes $\lambda = 0.02$ (Table 8). As shown in Tables 7 and 8, performance remains stable across a wide spectrum of configurations, with the default setting ($\lambda = 0.02, k = 8$) achieving the best overall balance across clinical accuracy (RadGraph-F1 0.240, composite RadCliQ 1.01) and lexical fidelity.

We further investigate the architectural placement of TELOS across decoder depths in Table 9. Multi-layer intervention ($\mathcal{L} = \{1, 15, 31\}$) consistently outperforms single-layer and two-layer configurations, demonstrating that maintaining expression memory across early, intermediate, and late representations provides superior semantic grounding compared to isolated intervention.

	Lexical				BERTScore	Clinical Graph			Clinical Diagnostic Accuracy				
Configuration	**BLEU-1**	**BLEU-4**	**ROUGE-L**	**METEOR**	**BS**	**RG-F1**	**RG-ER**	**RadCliQ↓**	**CheXbert**	**m-F1-14**	**M-F1-14**	**m-F1-5**	**M-F1-5**
$\lambda = 0.01$	**0.365**	**0.159**	0.312	0.317	0.859	0.239	0.272	1.03	0.375	0.381	0.253	0.403	0.336
$\lambda = 0.02$ **(Default)**	0.363	0.158	**0.314**	**0.319**	**0.860**	**0.240**	**0.273**	**1.01**	**0.377**	**0.386**	**0.261**	**0.405**	**0.343**
$\lambda = 0.05$	0.363	0.157	0.311	0.315	0.859	0.237	0.271	1.03	0.375	0.380	0.254	0.395	0.333
$\lambda = 0.10$	0.362	0.157	0.311	0.316	0.859	0.236	0.272	1.02	0.373	0.378	0.251	0.396	0.331

Table 7: **OSC loss-weight sensitivity on LLaVA-1.5.** Shaded row highlights the default configuration. Candidate count is fixed at $k = 8$. Bold font indicates the best performance; when the default is tied for best, only the default configuration is highlighted.

	Lexical				BERTScore	Clinical Graph			Clinical Diagnostic Accuracy				
Configuration	**BLEU-1**	**BLEU-4**	**ROUGE-L**	**METEOR**	**BS**	**RG-F1**	**RG-ER**	**RadCliQ↓**	**CheXbert**	**m-F1-14**	**M-F1-14**	**m-F1-5**	**M-F1-5**
$k = 1$	0.363	0.158	0.313	0.317	0.859	0.239	0.273	1.01	0.379	0.385	0.260	0.402	0.338
$k = 4$	0.362	0.157	0.312	0.317	0.859	0.239	0.273	1.01	**0.380**	**0.387**	0.260	0.403	0.339
$k = 8$ **(Default)**	**0.363**	**0.158**	**0.314**	**0.319**	**0.860**	**0.240**	**0.273**	**1.01**	0.377	0.386	**0.261**	0.405	**0.343**
$k = 16$	0.363	0.158	0.313	0.318	0.859	0.240	0.273	1.02	0.379	0.386	0.257	**0.406**	0.342

Table 8: **OSC candidate-count sensitivity on LLaVA-1.5.** Shaded row highlights the default configuration. Loss weight is fixed at $\lambda = 0.02$. Bold font indicates the best performance; when the default is tied for best, only the default configuration is highlighted.

	Lexical				BERTScore	Clinical Graph			Clinical Diagnostic Accuracy				
Configuration	**BLEU-1**	**BLEU-4**	**ROUGE-L**	**METEOR**	**BS**	**RG-F1**	**RG-ER**	**RadCliQ↓**	**CheXbert**	**m-F1-14**	**M-F1-14**	**m-F1-5**	**M-F1-5**
$\mathcal{L} = \{1\}$	0.355	0.145	0.292	0.296	0.852	0.214	0.249	1.12	0.355	0.357	0.230	0.383	0.312
$\mathcal{L} = \{15\}$	0.323	0.134	0.290	0.294	0.851	0.220	0.253	1.11	0.352	0.353	0.226	0.365	0.304
$\mathcal{L} = \{31\}$	0.255	0.085	0.228	0.225	0.826	0.144	0.181	1.42	0.280	0.184	0.105	0.159	0.117
$\mathcal{L} = \{1, 15\}$	**0.368**	0.155	0.304	0.311	0.857	0.231	0.265	1.06	0.363	0.370	0.242	0.386	0.323
$\mathcal{L} = \{1, 31\}$	0.362	0.151	0.298	0.302	0.854	0.222	0.258	1.08	0.360	0.356	0.227	0.388	0.319
$\mathcal{L} = \{15, 31\}$	0.323	0.136	0.299	0.301	0.853	0.226	0.259	1.09	0.362	0.354	0.227	0.363	0.301
$\mathcal{L} = \{1, 15, 31\}$ **(Default)**	0.363	**0.158**	**0.314**	**0.319**	**0.860**	**0.240**	**0.273**	**1.01**	**0.377**	**0.386**	**0.261**	**0.405**	**0.343**

Table 9: **Intervention layer topology on LLaVA-1.5.** Shaded row highlights the default configuration. Single-layer, paired-layer, and three-layer configurations are evaluated. Bold font indicates the best performance within each metric.

A.10 QUALITATIVE ANALYSIS AND CLINICAL CASE STUDIES ACROSS BENCHMARKS

To complement the quantitative benchmark evaluations (Section 4.2) and term-level realization diagnostics (Section 4.4), this section provides fine-grained qualitative case studies of report generation across three diverse clinical scenarios: (1) in-domain projection radiography on MIMIC-CXR (Figure 6), probing post-operative hardware recognition and surgical entity grounding; (2) cross-institutional distribution shift on CheXpert Plus (Figure 7), examining institutional prior collapse under Stanford Hospital protocols; and (3) cross-modal transfer to 2D axial CT on COV-CTR (Figure 8), examining subword derailment during multi-token clinical realization.

Across all qualitative comparisons, text spans are color-coded by reference alignment: **dark green** indicates exact reference matches, **light green** indicates clinically supported descriptions, and **peach** indicates hallucinations or reference inconsistencies.

MIMIC-CXR: Hardware Grounding and Surgical Entity Realization. Figure 6 evaluates an examination presenting midline sternotomy wires and surgical clips from prior CABG surgery with

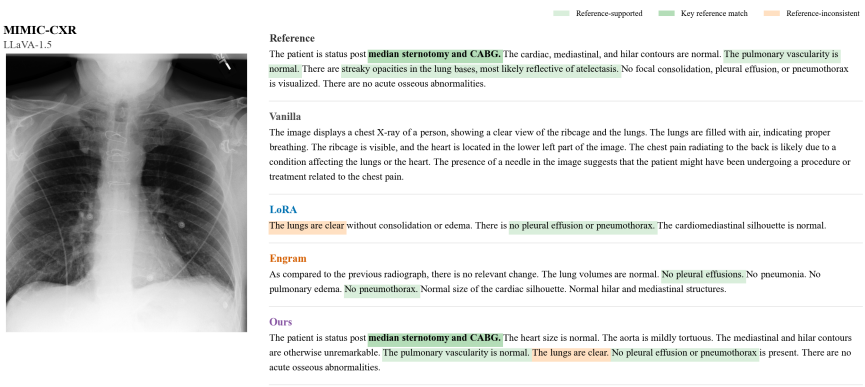


Figure 6: **Qualitative comparison on MIMIC-CXR (LLaVA-1.5).**

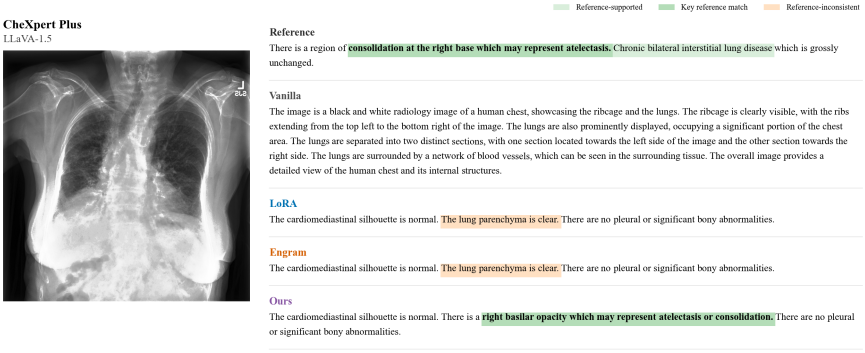


Figure 7: **Qualitative comparison on CheXpert Plus under institutional shift (LLaVA-1.5).**

basilar atelectasis. The zero-shot base VLM misidentifies hardware as a "needle", while standard LoRA fine-tuning overlooks hardware completely, asserting *"The lungs are clear"* in contradiction to reference atelectasis. Engram hallucinates an ungrounded prior radiograph. In contrast, TELOS reliably grounds metallic wire tokens, retrieving and completing ***"median sternotomy and CABG"*** with normal vascularity.

CheXpert Plus: Overcoming False-Negative Template Collapse under Institutional Shift. Figure 7 evaluates institutional shift (Stanford Hospital), where dense right basilar opacity obscures the hemidiaphragm. Both LoRA and Engram suffer false-negative collapse, emitting ***"The lung parenchyma is clear"*** due to source-domain normal priors. In contrast, TELOS's visual compatibility routing remains sensitive to focal density, generating ***"right basilar opacity which may represent atelectasis or consolidation"*** with appropriate differential hedging.

COV-CTR: Eliminating Subword Token-Path Corruption on Chest CT. Figure 8 demonstrates cross-modal transfer to 2D axial CT, where the reference documents peripheral subpleural ground-glass opacity with blurred margins in the right lower lobe. Standard LoRA suffers subword corruption: *pleura* fragments into subwords ('ple' + 'ura'), and unguided cross-entropy derails into device hallucination ***"under the plethysmograph"***. Engram hallucinates left lung involvement. In contrast, TELOS equips target inventory $\mathcal{E}_{\mathrm{CT}}$ with OSC suffix margin separation, achieving verbatim realization: ***"ground glass shadow was seen under the pleura of the lower lobe of the right lung with blurred margins"***.

Direct Comparison of Generated Findings. Table 10 directly contrasts generated report findings against authentic radiologist references across benchmarks. Standard LoRA fine-tuning defaults to

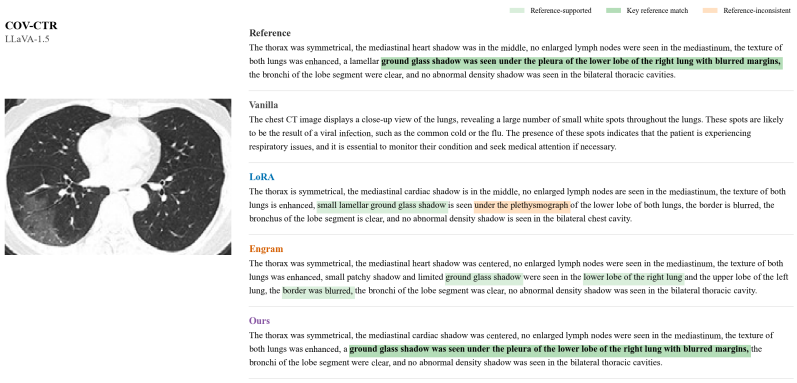


Figure 8: **Qualitative cross-modal transfer on chest CT (COV-CTR) with LLaVA-1.5.**

Benchmark	Reference Finding (Ground Truth)	LoRA Generation (Baseline)	TELOS Generation (Ours)
MIMIC-CXR	Midline sternotomy wires and surgical clips from prior CABG; basilar atelectasis.	*"The lungs are clear. No acute cardiopulmonary abnormality."* [Omits surgical hardware & atelectasis]	*"...median sternotomy and CABG with normal pulmonary vascularity."* [Accurately identifies surgical status]
CheXpert Plus	Right basilar opacity obscuring hemidiaphragm (consolidation vs. atelectasis).	*"The lung parenchyma is clear. No focal consolidation."* [Emits normal template; false negative]	*"...right basilar opacity which may represent atelectasis or consolidation."* [Recovers opacity with clinical differential]
COV-CTR	*"ground glass shadow was seen under the pleura of the lower lobe of the right lung with blurred margins"*	*"...under the plethysmograph..."* [Subword derailment into medical device]	*"ground glass shadow was seen under the pleura of the lower lobe of the right lung with blurred margins"* [Verbatim multi-token realization]

Table 10: **Direct comparison of generated report findings across qualitative case studies.**

generic normal templates ("The lungs are clear"), entirely omitting prominent surgical hardware and dense basilar opacities, or suffers subword fragmentation that derails into spurious medical devices ("plethysmograph"). In contrast, TELOS accurately realizes target clinical expressions, correctly identifying surgical status, hedging dense consolidations, and completing complex technical phrases verbatim without divergence.

A.11 STATEMENT ON GENERATIVE AI

Large Language Models (LLMs) were employed exclusively to assist with manuscript preparation. Their role was limited to refining language and improving readability, specifically contributing to rephrasing sentences, correcting grammar, and smoothing text flow. The LLMs played no part in formulating research questions, designing methodologies, conducting experiments, or analyzing results. All scientific ideas, experimental designs, and analyses were conceived and executed entirely by the authors. The authors accept full responsibility for the manuscript's content, including any text edited with LLM assistance, and have ensured that all text complies with ethical standards and academic integrity.