LEARNING TEMPORALLY EQUIVARIANT MULTIMODAL LARGE LANGUAGE MODELS FOR LONGITUDINAL MEDICAL IMAGE COMPARISON

Anonymous authors
Paper under double-blind review

ABSTRACT

Longitudinal medical image comparison requires models to reason about how findings change over time. A reliable model should therefore behave consistently under temporal reversal: if a finding worsens from a historical study to a current study, reversing the image order should yield the corresponding inverse relation. We refer to this property as *temporal equivariance*. A controlled study on CheXTemporal shows that current general and medical multi-modal large language models (MLLMs) often violate this property: similar forward and reverse accuracies do not imply reliable bidirectional correctness. We propose **TEMA**, a *Temporally Equivariant MLLM Adaptation* framework for longitudinal medical image comparison. TEMA couples supervision on original and temporally reversed image pairs with a *visual evidence consistency* objective: temporal semantics should transform under reversal, while their supporting visual evidence should remain consistent. To account for different output structures, TEMA operates at different weighting granularities: sample-level weighting for closed-ended temporal VQA and temporal-token weighting for open-ended report generation. We evaluate our TEMA on temporal VQA and open-ended difference report generation across multiple MLLM backbones. On temporal VQA, TEMA consistently improves both balanced accuracy and strict temporal equivariance over standard LoRA fine-tuning and temporal-inversion baselines across all four evaluated backbones. On difference report generation, TEMA also improves clinically oriented generation metrics across backbones and remains competitive with specialized longitudinal report-generation baseline.

1 INTRODUCTION

Longitudinal medical imaging is inherently comparative (Zhou et al., 2026; Holste et al., 2024; Rokuss et al., 2025). In clinical practice, physicians routinely compare a current examination with prior studies to determine whether a finding has newly appeared, resolved, worsened, improved, or remained stable. Recent medical multi-modal large language models (MLLMs) have shown strong capabilities in image understanding and report generation (Xiao et al., 2025; Li et al., 2025), and several longitudinal methods further incorporate prior images, temporal alignment, or explicit difference representations (Liu et al., 2025b; 2026a; Song et al., 2025; Liu et al., 2025a). However, most existing approaches primarily optimize prediction or generation in the canonical prior-to-current direction, leaving the behavior of generative MLLMs under temporal reversal largely underexplored.

A reliable longitudinal model should not only recognize temporal change, but also represent its *direction* consistently. Consider a model that predicts that a pleural effusion has worsened from a historical study to the current study. If the two images are swapped, the corresponding relation should become improved rather than remain unchanged or transform arbitrarily. We refer to this property as *temporal equivariance*. Let S denote the temporal swap operator and T the corresponding semantic inversion operator. A temporally equivariant model f should satisfy $f(S(X)) = T(f(X))$, where, for example, Worse $\leftrightarrow$ Improved, New $\leftrightarrow$ Resolved, Stable $\leftrightarrow$ Stable.

We first examine whether current MLLMs satisfy this basic structural property. On the CheXTemporal gold set (Prakash et al., 2026), which provides paired chest X-rays annotated with five temporal

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.

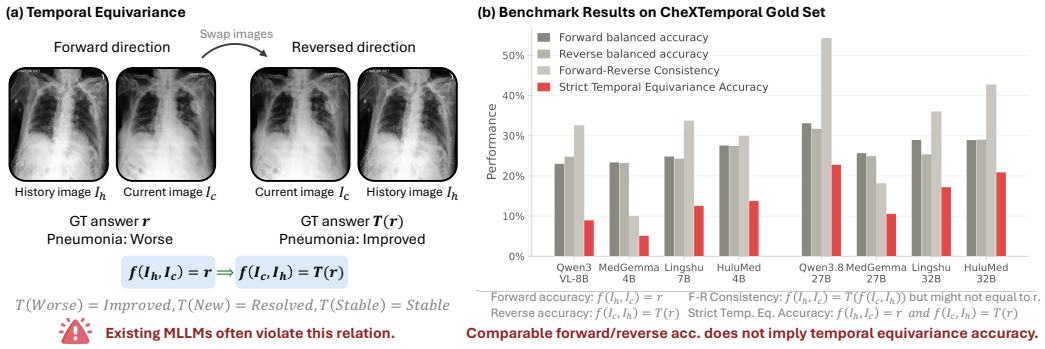

Figure 1: Motivation and empirical evidence of temporal non-equivariance in current MLLMs. (a) A temporally equivariant model should invert the predicted progression relation when the historical and current images are swapped. (b) On the CheXTemporal gold set, existing general and medical MLLMs achieve comparable forward and reverse accuracies, yet almost all below 50% forward-reverse consistency and substantially lower strict temporal equivariance accuracy.

progression states for specific findings, we evaluate a diverse set of general-purpose and medical MLLMs in both the original and reversed temporal directions. As shown in Fig. 1, forward and reverse accuracies are often comparable, yet the predictions are far less reliable when both directions are considered jointly. The best model reaches only 54.22% forward–reverse consistency and 22.83% strict temporal equivariance accuracy. This gap suggests that single-direction progression recognition does not naturally translate into a consistent understanding of temporal transformation.

Recent work has begun to recognize the importance of directional temporal structure. In particular, TILA (Ko et al., 2026) introduces temporal inversion for CLIP-style (Radford et al., 2021) temporal vision-language representation learning, using reversed image pairs to improve fixed-label progression classification. This motivates a broader question: *how should temporal equivariance be learned in autoregressive MLLMs?* Beyond fixed-label classification, autoregressive longitudinal comparison are extended to multiple findings and free-form descriptions. Meanwhile, temporal equivariance should not be limited to output-wise: reversing temporal order should change the temporal relation while preserving the visual evidence underlying that relation.

To this end, we propose **TEMA**, a *Temporally Equivariant MLLM Adaptation* framework for longitudinal medical image comparison. TEMA is built on a simple principle: *When the input image pair is reversed, temporal semantics should transform, while supporting visual evidence should remain consistent*. Accordingly, we introduce reverse supervision by swapping the temporal order and training the model with the corresponding inverse target. We further propose a *visual evidence consistency* objective that regularizes the spatial pattern of visual evidence used across temporal directions and between the two examinations. Rather than interpreting decoder attention as faithful grounding, we use normalized visual attention as a lightweight proxy for the visual evidence consumed during generation. Both tasks share these semantic and visual constraints, but differ in how informative temporal supervision is emphasized: sample-level weighting suppresses the Stable shortcut in closed-ended VQA, whereas temporal-token weighting highlights change expressions in open-ended report generation. We evaluate TEMA on temporal VQA and difference report generation across multiple MLLM backbones. On VQA, it consistently improves balanced accuracy and strict temporal equivariance, with particularly strong gains on directional non-Stable examples. On the standard forward report-generation task, it improves all three macro-averaged automated factual correctness metrics over standard LoRA and other baseline across the four backbones. Ablation studies further verify the contributions of reverse supervision, visual evidence consistency and task-adaptive weighting. Our contributions are threefold:

- We identify *temporal non-equivariance* as a systematic failure mode of current MLLMs on longitudinal chest x-ray images comparison benckmark, and formulate temporal reversal as a controlled way to evaluate whether model predictions obey the expected directional transformation.

- We propose **TEMA**, a *Temporally Equivariant MLLM Adaptation* framework that combines reverse supervision with visual evidence consistency across closed-ended and generative tasks, and adapts the weighting granularity to their output structures.
- We evaluate TEMA across multiple MLLM backbones on both temporal VQA and difference report generation. VQA experiments show simultaneous gains in task accuracy and strict temporal equivariance, while forward report-generation experiments demonstrate greater factual correctness and stronger longitudinal generation performance.

2 Related Work

2.1 Longitudinal Medical Vision-Language Modeling

Longitudinal medical image understanding requires modeling not only the content of individual examinations, but also how clinical findings evolve across time. Early work such as BioViL-T (Bannur et al., 2023) incorporates prior and current chest X-rays into temporal vision-language pretraining, demonstrating that explicitly modeling temporal structure benefits progression classification and report generation. Subsequent methods increasingly introduce dedicated mechanisms for integrating longitudinal information (Zhang et al., 2025; Wang et al., 2024; Dalla Serra et al., 2023). HERGen (Wang et al., 2024) aggregates multiple visits with a group causal transformer, while HC-LLM (Liu et al., 2025b) decomposes longitudinal representations into time-shared and time-specific components. Libra (Zhang et al., 2025) introduces a temporal alignment module to connect prior and current image representations before language generation.

More recent approaches move beyond generic temporal fusion toward explicit difference or disease-evolution modeling (Liu et al., 2026a; Yun et al., 2025; Lu et al., 2024). Diff-RRG (Yun et al., 2025) constructs disease-wise visual difference representations to guide report generation, whereas PriorRG (Liu et al., 2026a) incorporates prior-specific representation learning and decoding strategies for longitudinal reporting. MARE (Gao et al., 2026) formulates longitudinal report generation as multimodal analogical reasoning and explicitly regularizes disease evolution. TIM (Dong et al., 2026) further separates static pathological information from temporal progression and iteratively refines reports across time points, while BiOTPrompt (Liu et al., 2026b) uses bidirectional optimal transport to capture asymmetric disease evolution and identify newly appeared or resolved regions. Despite their differences, these methods share a common goal of improving how temporal differences are represented, aligned, or fused. Most, however, optimize prediction in the canonical prior-to-current direction and rely on task-specific temporal modeling components. In contrast, TEMA adapts existing MLLMs to enforce how predictions should transform under temporal reversal.

2.2 Temporal Directionality and Equivariance

Temporal change is inherently directional, and recent work has begun to exploit this structure explicitly. Most closely related to ours, TILA (Ko et al., 2026) introduces temporal inversion into CLIP-style temporal vision-language learning, using reversed image pairs and progression labels to improve direction-sensitive learning and fixed-label progression prediction. Our TEMA instead focuses on temporal equivariance in autoregressive MLLMs. Rather than restricting inversion to a fixed prediction space, TEMA applies temporal transformation to both closed-ended VQA and free-form difference report generation. It further constrains the visual evidence supporting predictions in both temporal directions and introduces task-adaptive weighting at the sample or token level.

3 Method

3.1 Problem Formulation and Overview

Let I_h and I_c denote the historical and current images, respectively. The forward and reverse inputs are $X^F = (I_h, I_c)$ and $X^R = S(X^F) = (I_c, I_h)$, with targets y^F and $y^R = T(y^F)$. For temporal VQA, T is the fixed permutation Worse $\leftrightarrow$ Improved, New $\leftrightarrow$ Resolved, and Stable $\leftrightarrow$ Stable. For report generation, $y^F = R_c$ describes the current examination relative to the historical one, whereas $y^R = R_h$ reverses this reference direction. Because this transformation is not generally reducible to word substitution, we construct R_h through constrained LLM rewriting that inverts

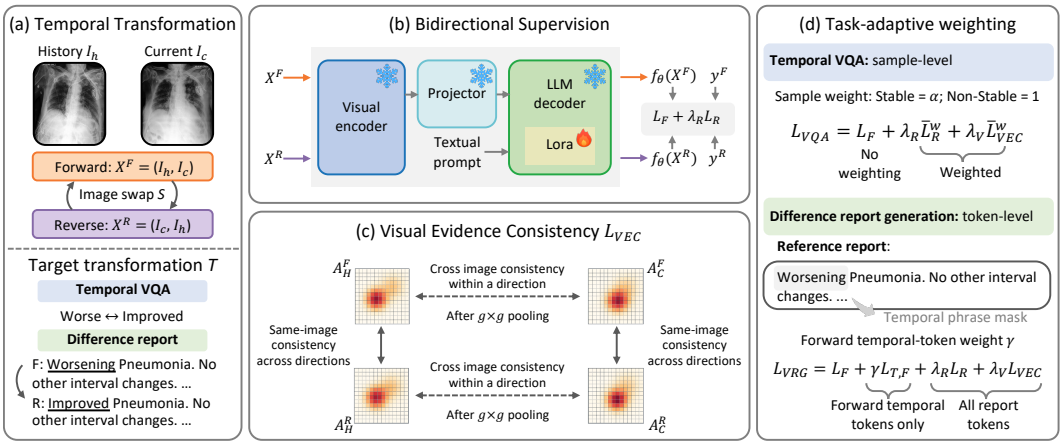

Figure 2: Overview of TEMA. TEMA combines (a) temporal transformation, (b) bidirectional supervision, (c) visual evidence consistency, and (d) task-adaptive weighting. For VQA, weighting is at the sample level, while for report generation, temporal tokens receive additional supervision.

directional changes while preserving stable and other non-directional content; Appendix A provides full construction and quality-control details.

Our TEMA adapts an autoregressive MLLM f_θ toward $f_\theta(S(X)) \simeq T(f_\theta(X))$. As shown in Fig. 2, it enforces this property at two complementary levels. *Bidirectional temporal supervision* teaches how output semantics should transform under temporal reversal, while *visual evidence consistency* encourages the model to preserve the image regions supporting those semantics. Both components are shared by temporal VQA and report generation. Their supervision is weighted at the granularity appropriate to each output space: samples for VQA and temporal tokens for report generation.

3.2 Bidirectional Temporal Supervision

For direction $d \in \{F, R\}$, let $y_i^d = (y_{i,1}^d, \ldots, y_{i,|y_i^d|}^d)$ denote the supervised target answer or report for sample i, where $y_{i,t}^d$ is its target token at position t. For report generation, the reverse target is the constrained LLM-generated inverse defined above. Let Ω_i^d be the supervised output-token positions, and define

$$\ell_{i,t}^d = -\log p_\theta\left(y_{i,t}^d \mid X_i^d, y_{i,<t}^d\right) \tag{1}$$

as the standard token-level cross-entropy, equivalently the causal language-modeling negative log-likelihood under teacher forcing, at position t. The ordinary directional loss is

$$\mathcal{L}_d = \frac{\sum_i \sum_{t \in \Omega_i^d} \ell_{i,t}^d}{\sum_i |\Omega_i^d|}. \tag{2}$$

We construct the two directions synchronously for each image pair and optimize them in the same paired batch. The weight factor λ_R controls the contribution of reverse supervision, and the basic bidirectional objective is

$$\mathcal{L}_{\text{bi}} = \mathcal{L}_F + \lambda_R \mathcal{L}_R. \tag{3}$$

3.3 Visual Evidence Consistency

Reverse supervision constrains the output relation but does not specify whether the two predictions are supported by compatible image evidence. We therefore introduce Visual Evidence Consistency (VEC). VEC uses decoder attention as a lightweight proxy for the visual evidence consumed during generation; we do not interpret it as a post-hoc explanation or a faithful clinical grounding.

Visual evidence maps. From the final decoder layer's position-encoded query and key states, we reconstruct causal attention at positions predicting supervised content tokens, excluding special tokens and tokens containing no alphanumeric characters. For each head, we average across

these queries, retain each image-token span, and normalize over its P_r visual tokens. This gives $A^{d,h}_{i,r} \in \Delta^{P_r-1}$ for direction $d \in \{F, R\}$, physical image $r \in \{H, C\}$, and head h. To filter diffuse or weakly visual teachers, we measure each head's mean image attention mass m_h and normalized spatial entropy $\bar{H}_h \in [0, 1]$. Eligible heads are ranked by $c_h = m_h(1 - \bar{H}_h)$, with at most the top eight retained and detached; comparisons with no eligible head are omitted. A 100-step language-only warm-up sets the mass and entropy gates to their weighted 25th and 75th percentiles, respectively, after which VEC is enabled. Let $\mathcal{H}_{\text{sel}}$ denote the selected teacher heads. For teacher and student visual-evidence maps A^T and A^S, respectively, we define the confidence-weighted head-wise divergence as

$$D\big(\text{sg}(A^T) \| A^S\big) = \frac{\sum_{h \in \mathcal{H}_{\text{sel}}} c_h \ \text{KL}\big(\text{sg}(A^T_h) \| A^S_h\big)}{\sum_{h \in \mathcal{H}_{\text{sel}}} c_h}. \tag{4}$$

Same-image consistency across directions. Temporal swapping changes the role of each image but not its physical identity. Using the divergence D defined above, we define

$$\mathcal{L}_{\text{same}} = \frac{1}{2}\big[D\big(\text{sg}(A^R_H) \| A^F_H\big) + D\big(\text{sg}(A^F_C) \| A^R_C\big)\big]. \tag{5}$$

Thus, the historical image is aligned between its historical role in the forward direction and current role in the reverse direction, and conversely for the physical current image. These comparisons use the native visual-token grid because they refer to the same physical image.

Cross-image consistency within a direction. The two longitudinal images are not pixel registered, but evidence for a temporal relation should concern corresponding coarse anatomical regions. Let $P_g(\cdot)$ denote adaptive pooling to a $g \times g$ spatial grid followed by normalization. We define

$$\mathcal{L}_{\text{cross}} = \frac{1}{2}\big[D\big(\text{sg}(P_g(A^F_C)) \| P_g(A^F_H)\big) + D\big(\text{sg}(P_g(A^R_H)) \| P_g(A^R_C)\big)\big]. \tag{6}$$

The combined visual objective is

$$\mathcal{L}_{\text{VEC}} = \frac{1}{2}\left(\mathcal{L}_{\text{same}} + \mathcal{L}_{\text{cross}}\right). \tag{7}$$

3.4 Task-Adaptive Temporal Weighting

The same temporal principle manifests differently in closed-ended and generative outputs. In VQA, class imbalance and the reversal-invariant Stable label create a sample-level shortcut. In report generation, temporal evidence is sparse within a much longer output sequence. TEMA therefore adapts the granularity of weighting while retaining the same reverse-supervision and VEC framework.

Sample-Level Weighting for Temporal VQA. For a VQA sample with target class c_i, we define its weight as w_i. If $c_i = \textit{stable}$, then $w_i = \alpha$, where $0 \leq \alpha \leq 1$; otherwise, $w_i = 1$. Predicting Stable in both directions can simultaneously reduce the forward and reverse losses without learning a directional transformation. We therefore apply w_i to the reverse supervision and VEC objectives, while retaining every sample at full weight in the forward objective. For a token-decomposable loss, the normalized weighted mean is

$$\overline{\mathcal{L}}^w = \frac{\sum_i w_i \sum_{t \in \Omega_i} \ell_{i,t}}{\sum_i w_i |\Omega_i|}. \tag{8}$$

The VQA objective is

$$\mathcal{L}_{\text{VQA}} = \mathcal{L}_F + \lambda_R \overline{\mathcal{L}}^w_R + \lambda_V \overline{\mathcal{L}}^w_{\text{VEC}}. \tag{9}$$

This asymmetric weighting preserves learning of Stable findings in the canonical task while preventing reversal-invariant samples from dominating the auxiliary temporal objectives.

Token-Level Weighting for Report Generation. A difference report contains both temporal expressions and non-temporal clinical content. To emphasize the sparse tokens that explicitly encode temporal status, we construct a high-precision phrase lexicon covering directional changes such as new/resolved, increased/decreased, and improved/worsened. The lexicon is generated with LLM

assistance and subsequently reviewed and revised by the authors. We use longest-match phrase detection and exclude matches inside negated expressions such as "no new" or "no significant increase." Full construction details and the resulting lexicon are provided in Appendix B.

We apply temporal-token weighting only to the forward report, which is the task evaluated at inference time. Reverse reports remain an auxiliary source of directional supervision and are optimized with the ordinary report loss; additionally upweighting automatically generated reverse phrases can amplify inversion noise. Let $m^F_{i,t} \in \{0, 1\}$ indicate whether forward target token t belongs to a matched true-change phrase. We define an independently normalized forward temporal-token loss

$$\mathcal{L}_{T,F} = \frac{\sum_i \sum_{t \in \Omega^F_i} m^F_{i,t} \ell^F_{i,t}}{\sum_i \sum_{t \in \Omega^F_i} m^F_{i,t}}. \tag{10}$$

When a batch contains no matched forward temporal token, $\mathcal{L}_{T,F}$ is set to zero. The coefficient γ controls temporal-token emphasis. Our primary formulation matches true-change phrases only.

Unlike the token-weighted language loss, report-generation VEC uses all content-bearing output queries. Restricting the evidence map to explicit change words would omit reports that express stability and would discard finding and anatomy tokens that identify where a change occurs. The resulting report-generation objective is

$$\mathcal{L}_{\text{VRG}} = \mathcal{L}_F + \gamma \mathcal{L}_{T,F} + \lambda_R \mathcal{L}_R + \lambda_V \mathcal{L}_{\text{VEC}}. \tag{11}$$

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our method in two complementary settings: closed-ended temporal VQA and open-ended difference report generation. **Temporal VQA.** We use CheXTemporal (Prakash et al., 2026), which provides paired prior–current chest X-rays with five progression labels for particular findings: {`New`, `Resolved`, `Worse`, `Improved`, `Stable`}. Due to the large number of `Stable` samples, we randomly selected a subset of `Stable` samples from the original silver set, along with samples from the other classes, to form the training set. It contains 53,489 pair–finding samples from 19,613 patients, including 19,180 `Stable`, 9,041 `New`, 12,826 `Worse`, 9,628 `Improved`, and 2,814 `Resolved` samples. Evaluation is performed exclusively on the expert-annotated gold set of 1,787 samples, including 154 `New`, 115 `Resolved`, 440 `Worse`, 424 `Improved`, and 654 `Stable` samples. The silver training set and gold test set have no patient-level overlap. **Difference report generation.** We use Task 2 of TemMed-Bench (Zhang et al., 2026), which evaluates difference-report generation from a historical and a current chest radiograph. The Knowledge Corpus contains 17,144 image pairs and is divided into 90% training and 10% validation data at the patient level. The held-out test set contains 1,000 image pairs. Following the benchmark partition, we additionally report results on the *Stable*, *Mixed*, and *Changed* subsets. *Stable* means that all claims in the report are stable, *Changed* means that all claims have changed, and *Mixed* means that the report contains both stable and changed claims.

Evaluation Metrics. Temporal VQA. We evaluate both the original prior-to-current direction and the temporally reversed direction. We report forward balanced accuracy (FB Acc) and reverse balanced accuracy (RB Acc), computed as macro-averaged per-class accuracy. To evaluate temporal equivariance while avoiding inflated scores caused by the self-inverse `Stable` class, we additionally report the same balanced accuracies on the four non-stable classes (FB Acc w/o Stable and RB Acc w/o Stable). We further report Strict Equivariance Accuracy w/o Stable, which requires both the forward and reversed predictions to be correct. **Difference report generation.** We evaluate generated reports using RaTE (Zhao et al., 2024), GREEN (Ostmeier et al., 2024), and ALFA-(m) (Liao et al., 2025), which capture complementary aspects of temporal description and clinical correctness. For each metric, we report the average over the complete test set and over the *Stable*, *Mixed*, and *Changed* subsets. Our quantitative temporal-equivariance evaluation is restricted to VQA, where the inverse label is deterministic. For free-form report generation, we evaluate only the standard historical-to-current output; the generated reverse reports are used as auxiliary training targets.

Implementation Details. We conduct all experiments on four MLLM backbones: Qwen3-VL-8B-Instruct (Yang et al., 2025), HuluMed-4B (Jiang et al., 2025), MedGemma-4B Sellergren et al.

Table 1: Comparison on the CheXTemporal gold set. Balanced accuracies are reported for both forward and reversed directions. Metrics "w/o Stable" exclude the self-inverse `Stable` class. TILA* denotes the autoregressive MLLM adaptation of TILA's temporal-inversion supervision. Best results for each backbone are shown in bold.

Model	Training	FB Acc w/o S.	RB Acc w/o S.	FB Acc	RB Acc	Strict Eq. w/o S.
Qwen3-VL-8B	Zero-shot	27.91	30.22	23.06	24.82	13.24
	LoRA	22.40	20.25	29.33	27.70	11.21
	TILA*	19.86	19.05	28.79	28.17	15.45
	Ours	**33.54**	**34.00**	**30.31**	**30.38**	**33.01**
HuluMed-4B	Zero-shot	32.35	32.26	27.62	27.49	18.36
	LoRA	27.86	25.02	34.28	32.10	18.89
	TILA*	27.27	27.62	34.87	34.82	22.86
	Ours	**40.20**	**40.26**	**36.20**	**36.15**	**38.39**
MedGemma-4B	Zero-shot	26.31	26.36	23.40	23.26	2.74
	LoRA	27.70	23.19	35.77	32.13	17.03
	TILA*	25.58	25.48	34.44	34.58	21.18
	Ours	**40.02**	**39.70**	**37.19**	**36.62**	**39.19**
Lingshu-7B	Zero-shot	29.13	28.36	24.89	24.34	17.12
	LoRA	20.67	18.08	30.72	28.01	11.12
	TILA*	16.75	18.77	28.57	30.06	14.83
	Ours	**35.36**	**36.54**	**31.99**	**33.03**	**34.60**

(2025), and Lingshu-7B (Xu et al., 2026). Images are resized to 224×224. For both tasks, the visual encoder and projector are kept frozen, and LoRA (Hu et al., 2022) is applied to the `q_proj` and `v_proj` modules of the LLM decoder. We use LoRA rank 8, scaling factor 16, and dropout 0.05. All trained models are optimized in bfloat16 with AdamW, a learning rate of 5×10^{-5}, weight decay 0.01, and a cosine schedule with a 3% warm-up. VEC begins after a 100-step language-objective warm-up without VEC, during which we calibrate model-specific visual-mass and entropy thresholds. For all comparisons, Standard LoRA is trained with matched optimization steps to our method. The construction details of reverse targets and temporal phrase lexicon are provided in Appendix A and B. We conduct a human audit of 150 automatically generated training reverse reports; 97.1% of directional temporal-phrase occurrences are correctly inverted and 95.3% of reports are fully correct. **Temporal VQA.** Bidirectional methods are trained for one epoch with a per-device batch of 12 image pairs, expanded to 24 directional inputs per step. The forward-only LoRA baseline is trained for two epochs with batch size 24, matching both the number of optimizer updates and processed directional inputs. We tune the hyperparameters of our method only on Lingshu-7B using the CheXTemporal validation split, and directly apply the selected configuration to all other MLLMs without further tuning. The selected setting uses $\lambda_R = 1$, $\lambda_V = 1$, $\alpha = 0.0$, and a 14×14 spatial grid for visual evidence consistency. **Difference Report Generation.** Bidirectional methods are trained for one epoch with a per-device batch of four image pairs, expanded to eight directional inputs per step. The forward-only LoRA baseline uses two epochs and batch size eight under the same matching scheme. The final configuration is $\lambda_R = 0.1$, $\lambda_V = 1.0$, $\gamma = 0.05$, and $g = 14$.

4.2 Comparison Results

We compare against zero-shot inference, Standard LoRA fine-tuning, and TILA*. TILA (Ko et al., 2026) is originally designed for temporal vision-language representation learning with a fixed progression classification space. Its prediction-level consistency loss is therefore not directly applicable to autoregressive MLLMs without introducing an additional classification head. We construct TILA* by adapting its core temporal-inversion principle to our generative setting.

Temporal VQA. As shown in Table 1, standard LoRA improves overall forward and reverse accuracy, but the gains are largely driven by the self-inverse `Stable` class. After excluding `Stable`, FB/RB accuracy and strict equivariance often decrease relative to zero-shot models, indicating limited improvement on directional temporal changes. TILA* improves strict equivariance through temporal-inversion supervision, but yields limited or inconsistent gains in non-stable FB/RB accu-

Table 2: Difference report generation comparison results. We report RaTE, GREEN, and ALFA(-m) on the full test set (All), the Changed subset, and the macro average across three subsets.

Model	Training	RaTE			GREEN			ALFA(-m)		
		All	Changed	Avg.	All	Changed	Avg.	All	Changed	Avg.
PriorRG	Zero-shot	47.15	42.81	45.95	20.83	13.04	18.69	21.70	9.38	18.29
	Fine-tuned	51.27	46.81	50.04	29.61	13.85	25.29	36.56	15.01	30.62
HuluMed-4B	Zero-shot	46.74	44.14	46.02	14.00	13.59	13.90	21.50	16.40	20.08
	LoRA	51.00	47.48	50.03	26.58	16.16	23.72	34.32	14.88	28.95
	TILA*	51.27	47.34	50.18	26.55	16.98	23.92	33.61	14.73	28.40
	TEMA	**52.40**	**51.98**	**52.20**	**27.38**	**17.12**	**24.29**	**34.69**	**16.52**	**29.67**
MedGemma-4B	Zero-shot	51.64	46.68	50.26	22.79	**21.81**	22.54	21.72	11.55	18.91
	LoRA	51.62	48.62	50.79	27.04	13.60	23.34	36.54	14.71	30.51
	TILA*	51.31	48.54	50.54	26.41	16.24	23.61	35.22	16.46	30.05
	TEMA	**52.17**	**51.58**	**52.00**	**28.67**	18.96	**26.00**	**37.57**	**17.35**	**31.09**
Lingshu-7B	Zero-shot	40.34	40.21	40.30	5.38	3.08	4.76	**32.11**	2.82	24.14
	LoRA	49.46	46.77	48.71	22.34	14.27	20.12	30.03	**14.57**	25.77
	TILA*	49.59	45.03	48.33	21.79	9.12	18.31	29.76	10.80	24.53
	TEMA	**50.43**	**49.67**	**50.22**	**23.43**	**16.91**	**22.36**	31.74	13.20	**26.1**7
Qwen3-VL-8B-Instruct	Zero-shot	41.82	39.46	41.16	8.10	12.33	9.27	9.69	6.51	8.80
	LoRA	47.51	42.15	46.03	**18.23**	6.66	15.04	28.42	7.31	22.60
	TILA*	47.62	43.36	46.44	17.01	8.81	14.75	26.07	9.01	21.37
	TEMA	**48.76**	**45.64**	**47.90**	17.74	**14.39**	**16.10**	**28.61**	**10.86**	**23.05**

racy. In contrast, our method consistently improves both non-stable balanced accuracy and strict equivariance across all four MLLM backbones. On HuluMed-4B, for example, FB Acc w/o Stable increases from 27.27% to 40.20%, while Strict Eq. Acc w/o Stable improves from 22.86% to 38.39%. These results show that our TEMA achieves higher joint forward–reverse correctness alongside higher directional accuracy.

Difference Report Generation. Table 2 compares TEMA with zero-shot inference, standard LoRA, and TILA* under the same MLLM backbone and training setup, and additionally includes PriorRG as a specialized longitudinal report-generation reference. We report RaTE, GREEN, and ALFA(-m) on the full test set, the Changed subset, and the macro average across three temporal subsets. Across all four MLLM backbones, TEMA consistently improves the macro-averaged RaTE, GREEN, and ALFA(-m) over standard LoRA. Improvements are also pronounced on the Changed subset, where TEMA consistently increases RaTE and GREEN across all backbones. We further compare against PriorRG under the same image-only evaluation setting. Since PriorRG uses a dedicated longitudinal architecture and training pipeline, this comparison is system-level rather than backbone-controlled. Notably, TEMA with MedGemma-4B surpasses the fine-tuned PriorRG baseline on all three macro-averaged metrics, while adapting an existing MLLM without introducing a task-specific architecture.

4.3 Ablation Study

We ablate the three components of our objective: reverse supervision, task-adaptive weighting, and visual evidence consistency. Starting from Standard LoRA with only $\mathcal{L}_F$, we sequentially add $\mathcal{L}_R$, task-adaptive weighting, and $\mathcal{L}_{\text{VEC}}$.

Temporal VQA. Table 3 reports the ablation results on Lingshu-7B. Reverse supervision raises Strict Eq. Acc w/o Stable from 11.12% to 14.83%, but by itself reduces forward non-Stable accuracy from 20.67% to 16.75%. Sample weighting removes the `Stable` shortcut from the auxiliary objectives and produces the largest gain, increasing the two metrics to 33.19% and 32.33%, respectively. VEC further improves strict equivariance to 34.60% and forward non-Stable accuracy to 35.36%. Thus, reverse supervision establishes the transformation constraint, weighting focuses it on directional cases, and VEC provides an additional gain by constraining visual attention patterns. Additionally, to separate temporal-equivariance learning from simple class rebalancing, we add a forward-only control that sets the Stable weight to zero without using reversed samples. This

Table 3: Ablation study on CheXTemporal using Lingshu-7B.

Reverse Loss	Task-adaptive weighting	VEC	Strict Eq. w/o S.	FB Acc w/o S.	FB Acc
✗	✗	✗	11.12	20.67	30.72
✓	✗	✗	14.83	16.75	28.57
✓	✓	✗	33.19	32.33	29.93
✓	✓	✓	**34.60**	**35.36**	**31.99**

Table 4: Ablation study on TemMedBench using HuluMed-4B.

Reverse Loss	Task-adaptive weighting	VEC	RaTE		
			All	Changed	Avg.
✗	✗	✗	51.00	47.48	50.03
✓	✗	✗	51.52	49.53	50.97
✓	✓	✗	51.74	51.87	51.77
✓	✓	✓	**52.40**	**51.98**	**52.20**

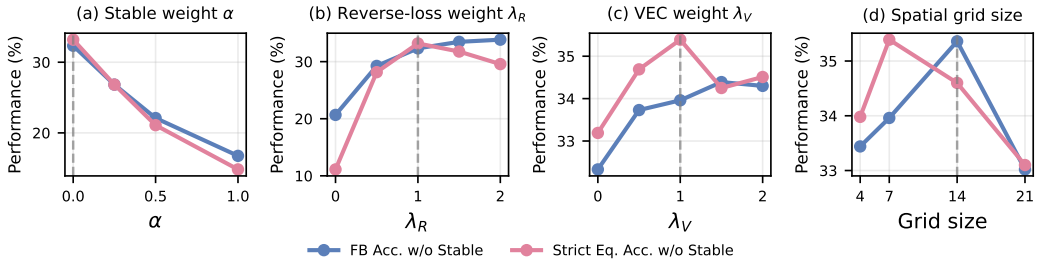

Figure 3: Hyperparameter sensitivity on CheXTemporal using Lingshu-7B. We vary the Stable sample weight α, reverse-loss coefficient λ_R, VEC coefficient λ_V, and spatial grid size. Dashed lines indicate the configurations used in the main experiments.

raises Strict Eq. Acc w/o Stable from 11.12% to 24.45%, while adding reverse supervision and task-adaptive weighting further improves it to 33.19%, showing that the equivariance gains arise not only from class rebalancing but also from reverse supervision.

Difference Report Generation. Table 4 gives the HuluMed-4B ablation using RaTE, a clinically oriented measure of report factual accuracy. Reverse-report supervision ($\lambda_R = 0.1$) improves the Changed subset by 2.05% and the subset macro average by 0.94% over Standard LoRA. Upweighting forward true-change phrases ($\gamma = 0.05$) yields a further 2.34% gain on Changed reports, where temporal language is most informative. Finally, VEC ($\lambda_V = 1$, $g = 14$) achieves the best overall and macro-average scores, improving them by 1.40% and 2.17% over Standard LoRA, respectively, while retaining the strong Changed-subset gain. These results demonstrate the contributions of bidirectional supervision, temporal-language emphasis, and visual-evidence regularization.

Hyperparameter Sensitivity. We study the sensitivity of the sample weight α, reverse-loss coefficient λ_R, VEC coefficient λ_V, and the spatial pooling resolution used by VEC. As shown in Figure 3, performance remains relatively stable around the selected values of λ_R and λ_V, while the Stable weight and spatial grid resolution have a more pronounced effect. For the main experiments, we use $\lambda_R = 1.0$, $\lambda_V = 1.0$, $\alpha = 0.0$, and a 14×14 pooled attention grid.

5 Conclusion

In this paper, we identify temporal non-equivariance as a systematic failure mode of longitudinal medical MLLMs: when the image order is reversed, models often fail to produce the corresponding semantically inverted prediction. Based on the principle that temporal semantics should transform under reveral while supporting visual evidence remains consistent, we propose TEMA, combining bidirectional supervision, visual evidence consistency, and task-adaptive weighting. Across multiple backbones, TEMA improves directional accuracy and strict equivariance on temporal VQA, as well as factual accuracy on forward difference-report generation; ablations confirm the complementary contributions of its components. As with temporal-inversion approaches more broadly, TEMA is most directly applicable to pairwise longitudinal findings whose reverse relations are well-defined. Extending this principle to more complex temporal narratives and sequences with more than two examinations is an important direction for future work.

AI USE STATEMENT

AI tools were used for language assistance, including grammar correction, phrasing, and improvements to readability. In addition, we used DeepSeek-V4-Flash as part of the research workflow for two specific purposes: (i) constructing the reverse-report auxiliary training targets described in Appendix A, for which a sampled subset of the generated targets was manually reviewed for quality; and (ii) proposing candidate lexical variants during the construction of the temporal phrase lexicon described in Appendix B, for which all proposed variants were manually inspected before inclusion. The authors take full responsibility for the methodology, experiments, results, and complete content of this work.

ETHICS STATEMENT

This study uses only publicly available, de-identified medical imaging datasets and involves no new data collection or interaction with human participants. The datasets were used in accordance with their respective licenses and access requirements. Our method is proposed for research purposes and is not intended for direct clinical decision-making. As with other data-driven medical imaging methods, its performance may be affected by biases and distribution shifts across patient populations, institutions, and imaging protocols; appropriate external and clinical validation would therefore be required before deployment.

REPRODUCIBILITY STATEMENT

Section 4 specifies the datasets and splits, evaluation metrics, model backbones, training protocol, and hyperparameter selection procedure used in our experiments. The method section gives the complete training objectives. Appendix A provides the reverse-report generator, decoding configuration, full prompt, examples, sampling protocol, annotation criteria, and audit results, while Appendix B documents construction and use of the temporal phrase lexicons. These details, together with the accompanying and configurations, are provided to support reproduction of the reported experiments.

REFERENCES

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15016–15027. IEEE, 2023.

Francesco Dalla Serra, Chaoyang Wang, Fani Deligianni, Jeff Dalton, and Alison O'Neil. Controllable chest x-ray report generation from longitudinal representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 4891–4904, 2023.

Yiheng Dong, Yi Lin, Shilong Huang, Xiyan Yang, and Xin Yang. Tim: Temporal decoupling with iterative mutual-refinement model for longitudinal radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6951–6961, 2026.

Qingqing Gao, Tengfei Liu, Xiaoyan Li, Xiaodan Zhang, Zhongfan Sun, Boyue Wang, Baocai Yin, and Zhaohui Liu. Mare: Multimodal analogical reasoning for disease evolution-aware radiology report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21180–21188, 2026.

Gregory Holste, Mingquan Lin, Ruiwen Zhou, Fei Wang, Lei Liu, Qi Yan, Sarah H Van Tassel, Kyle Kovacs, Emily Y Chew, Zhiyong Lu, et al. Harnessing the power of longitudinal medical imaging for eye disease prognosis using transformer-based sequence modeling. *NPJ digital medicine*, 7 (1):216, 2024.

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. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025.

Hanbin Ko, Kyeongmin Jeon, Doowoong Choi, and Chang Min Park. Temporal inversion for learning interval change in chest x-rays. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28338–28347, 2026.

Cheng-Yi Li, Kao-Jung Chang, Cheng-Fu Yang, Hsin-Yu Wu, Wenting Chen, Hritik Bansal, Ling Chen, Yi-Ping Yang, Yu-Chun Chen, Shih-Pin Chen, et al. Towards a holistic framework for multimodal llm in 3d brain ct radiology report generation. *Nature Communications*, 16(1):2258, 2025.

Zehui Liao, Shishuai Hu, Ke Zou, Mengyuan Jin, Yanning Zhang, Huazhu Fu, Liangli Zhen, and Yong Xia. Univrse: Unified vision-conditioned response semantic entropy for hallucination detection in medical vision-language models. *arXiv preprint arXiv:2503.20504*, 2025.

Kang Liu, Zhuoqi Ma, Xiaolu Kang, Yunan Li, Kun Xie, Zhicheng Jiao, and Qiguang Miao. Enhanced contrastive learning with multi-view longitudinal data for chest x-ray report generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10348–10359. IEEE, 2025a.

Kang Liu, Zhuoqi Ma, Zikang Fang, Yunan Li, Kun Xie, and Qiguang Miao. Priorrg: Prior-guided contrastive pre-training and coarse-to-fine decoding for chest x-ray report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7206–7214, 2026a.

Tengfei Liu, Jiapu Wang, Yongli Hu, Mingjie Li, Junfei Yi, Xiaojun Chang, Junbin Gao, and Baocai Yin. Hc-llm: Historical-constrained large language models for radiology report generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 5595–5603, 2025b.

Tengfei Liu, Yijian Fan, Boyue Wang, Yongli Hu, Mingjie Li, Jinghua Li, Junbin Gao, Xiaojun Chang, Zhihui Li, and Baocai Yin. Biotprompt: Bidirectional optimal transport guided prompting for disease evolution-aware radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13755–13765, 2026b.

Zilin Lu, Yutong Xie, Qingjie Zeng, Mengkang Lu, Qi Wu, and Yong Xia. Spot the difference: Difference visual question answering with residual alignment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 649–658. Springer, 2024.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson Md, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, et al. Green: Generative radiology report evaluation and error notation. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 374–390, 2024.

Eva Prakash, Yunhe Gao, Chong Wang, Justin Xu, Neal Prakash, Arne Michalson, Seena Dehkharghani, Eun Kyoung Hong, Julie Bauml, Roger Boodoo, et al. Chextemporal: A dataset for temporally-grounded reasoning in chest radiography. *arXiv preprint arXiv:2605.11304*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Maximilian Rokuss, Yannick Kirchhoff, Seval Akbal, Balint Kovacs, Saikat Roy, Constantin Ulrich, Tassilo Wald, Lukas T Rotkopf, Heinz-Peter Schlemmer, and Klaus Maier-Hein. Lesionlocator: Zero-shot universal tumor segmentation and tracking in 3d whole-body imaging. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30872–30885. IEEE, 2025.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Shanshan Song, Hui Tang, Honglong Yang, and Xiaomeng Li. Ddatr: Dynamic difference-aware temporal residual network for longitudinal radiology report generation. *IEEE Transactions on Medical Imaging*, 44(12):5345–5357, 2025.

Fuying Wang, Shenghui Du, and Lequan Yu. Hergen: Elevating radiology report generation with longitudinal data. In *European conference on computer vision*, pp. 183–200. Springer, 2024.

Hanguang Xiao, Feizhong Zhou, Xingyue Liu, Tianqi Liu, Zhipeng Li, Xin Liu, and Xiaoxuan Huang. A comprehensive survey of large language models and multimodal large language models in medicine. *Information Fusion*, 117:102888, 2025.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: Generalist foundation model for unified multimodal medical understanding and reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Hannah Yun, Junyeong Maeng, Eunsong Kang, and Heung-Il Suk. Diff-rrg: Longitudinal disease-wise patch difference as guidance for llm-based radiology report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 152–161. Springer, 2025.

Junyi Zhang, Jia-Chen Gu, Wenbo Hu, Yu Zhou, Robinson Piramuthu, and Nanyun Peng. Temmedbench: Evaluating temporal medical image reasoning in vision-language models. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=kyByX4OnSJ.

Xi Zhang, Zaiqiao Meng, Jake Lever, and Edmond SL Ho. Libra: Leveraging temporal images for biomedical radiology analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 17275–17303, 2025.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Ratescore: A metric for radiology report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15004–15019, 2024.

Shaoyang Zhou, Yingshu Li, Yunyi Liu, Lingqiao Liu, Lei Wang, and Luping Zhou. A review of longitudinal radiology report generation: Dataset composition, methods, and performance evaluation. *Medical Image Analysis*, pp. 104304, 2026.

A Reverse Report Construction

Construction protocol. The original difference report describes the historical-to-current direction. We construct its reverse target by asking an external language model to rewrite the report in the current-to-historical direction. We use `DeepSeek-V4-Flash` with deterministic decoding (`temperature=0`) and disable the model's thinking mode. We do not explicitly set `top_p`, a maximum output length, or a random seed, and leave the remaining decoding parameters at their API defaults. The model is instructed to return only the rewritten report, with no explanation. The complete system prompt is:

```
You reverse longitudinal chest X-ray difference reports.
The input describes changes from the historical image to
the current image.  Rewrite it to describe changes from the
current image to the historical image.
Rules:
1.  Reverse each independent temporal change.
2.  Preserve all clinical details unless they are part of
the change being reversed.
3.  Preserve uncertainty and keep stable or unchanged
findings unchanged.
4.  Do not add, omit, explain, or infer information.
5.  Use the clinically natural inverse for the specific
finding; do not mechanically replace words with generic
antonyms.
6.  Return only the reversed report.
```

For each source report R_c, the user message is `Reverse this report:` $\{R_c\}$. Empty responses and API failures are retried up to five times, and each successful response is stored together with its source report, generator name, and prompt version. This procedure changes the temporal reference direction only; it does not use the radiographs or add new image-derived information.

Representative transformations. Table A2 illustrates the intended behavior. A change is inverted in a finding-specific manner, whereas stable observations, anatomy, laterality, severity, and uncertainty are retained.

Human audit. Using seed 20260925, we sample 150 training pairs without replacement, with 50 each from the Stable-only, Change-only, and Mixed strata. The annotators comprised two PhD holders, both with training in medical knowledge. Each annotator independently compared every source–reverse pair while blinded to identifiers and strata, consulting standard medical references as needed; disagreements were subsequently resolved by consensus. The task was limited to judging whether temporal phrases were correctly inverted or preserved in context, and did not require image interpretation, clinical diagnosis, or assessment of the source report's radiographic validity. Accordingly, images were not shown during annotation.

Criteria. To avoid subjective claim decomposition, the lexicon matcher precomputes all directional and stable phrase occurrences. Annotators count a phrase as correct when it is appropriately transformed in context: directional relations are inverted over their complete finding scope, whereas stable or non-directional contextual uses are preserved. Anatomy, laterality, severity, and uncertainty must remain unchanged. They additionally judge whether all other clinical content is preserved without addition, omission, or alteration. A report is fully correct only when every phrase is handled correctly and all other content is preserved.

Audit results. Only eight of 150 reports require consensus review. Before consensus, exact agreement on the derived fully-correct label is 96.0% (Cohen's $\kappa = 0.55$). Table A1 reports consensus

Table A1: Human audit of 150 automatically constructed training reverse reports.

Criterion	Denominator	Result (95% CI)
Directional phrases correct	136 phrases	97.1% (92.7–98.9)
Stable phrases preserved	249 phrases	99.2% (97.1–99.8)
All directional phrases correct (report)	100 reports	97.0% (91.5–99.0)
All stable phrases preserved (report)	100 reports	98.0% (93.0–99.4)
Other content preserved (report)	150 reports	98.7% (95.3–99.6)
Fully correct reports	150 reports	95.3% (90.7–97.7)

Table A2: Representative forward-to-reverse report transformations.

Type	Historical-to-current report	Current-to-historical target
Change + stable	Slight interval increase in left pleural effusion. Otherwise, no significant interval change.	Slight interval decrease in left pleural effusion. Otherwise, no significant interval change.
Mixed findings	Interval decrease in left pleural effusion and slightly improved aeration at the left lung base. Supportive equipment is unchanged.	Interval increase in left pleural effusion and slightly worsened aeration at the left lung base. Supportive equipment is unchanged.
Uncertainty	Bilateral pleural effusions and pulmonary edema may be slightly decreased from the previous study.	Bilateral pleural effusions and pulmonary edema may be slightly increased from the previous study.
Stable only	Lines and tubes are unchanged. Low lung volumes and left basilar atelectasis or consolidation are also unchanged.	Lines and tubes are unchanged. Low lung volumes and left basilar atelectasis or consolidation are also unchanged.

estimates with Wilson 95% confidence intervals. The 95.3% fully-correct rate indicates that constrained rewriting produces high-quality auxiliary targets, retaining a small amount of label noise.

B Temporal Phrase Lexicon Construction

Purpose and construction. The temporal phrase lexicon identifies the small subset of report tokens that explicitly encode longitudinal status. We use it to construct the temporal-token mask for difference report generation; it is not used to alter the evaluation references or scores. We first defined semantic categories from the temporal relations central to longitudinal radiology reports and used an LLM to propose common lexical variants. We then scanned the TemMed-Bench training reports for matched phrases and their surrounding contexts. The authors manually reviewed the candidates, added recurrent dataset-specific variants, and removed expressions whose temporal interpretation depended too strongly on context. This procedure prioritizes precision over exhaustive coverage so that the auxiliary loss is not applied to ordinary descriptive tokens.

True-change phrases. The lexicon used by the main temporal-token objective contains directional categories for increase, decrease, improvement, worsening, resolution, development, progression, regression, and interval device change. It includes inflectional and multiword variants such as *interval increase*, *more prominent*, *improved*, *near resolution*, *newly placed*, and *extubation*. We exclude superficially temporal but ambiguous terms such as *appear*, *enlarged*, *removal*, and *advanced/retracted* when used without an unambiguous interval construction. Likewise, *more* and *less* are retained only as part of specific directional expressions.

Stable-status phrases. We maintain a separate stable-status lexicon for data stratification and the stable-token ablation. Its categories cover explicit stability (e.g., *stable*, *unchanged*, and *similar to prior*), absence of interval change, absence of a new finding, persistence, and redemonstration. Ambiguous words such as *remain*, *again*, *similar*, or *continue* are not used alone because they do not reliably establish temporal stability. The main report-generation model upweights true-change phrases only; stable phrases are not included in the temporal-token loss unless explicitly enabled in the corresponding ablation.

Matching and token alignment. Matching is case-insensitive and requires word boundaries. Candidate phrases are processed from longest to shortest, and overlapping matches retain only the

longest span. Change-looking words inside explicitly negated expressions, including *no new*, *no significant increase*, and *no interval change*, are excluded from the true-change mask. After phrase spans are located in the report string, tokenizer offset mappings are used to mark all supervised output tokens overlapping those spans. The resulting mask therefore covers the complete matched expression rather than only a single keyword.