GROUNDING RADIOLOGY REPORTS THROUGH TEST-TIME LATENT ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating factual radiology reports requires aligning clinical statements with image evidence. A common approach allocates additional test-time computation to sampling and reranking reports from a frozen vision-language model. We identify a *coverage–quality gap* in this approach: larger candidate pools recover more correct findings, yet these gains do not reliably translate into more factual selected reports. This observation motivates using evidence to guide the candidate-generating distribution itself. We introduce *Evidence-grounded Latent Adaptation* (ELA), a test-time framework that optimizes a small, study-specific latent state within a frozen model. Feedback from frozen visual evidence models guides the update by penalizing unsupported assertions and omitted findings. The adapted state then conditions report generation, after which a fixed evidence-based selector chooses the final report. Adaptation requires neither reference reports nor backbone parameter updates, and the latent state is discarded after each study. Experiments across multiple backbones and chest X-ray benchmarks show consistent improvements in report factuality. Controlled analyses isolate gains from latent adaptation beyond evidence-based selection and show that adapted candidate pools can outperform larger pools from the fixed policy. These findings support evidence-guided adaptation of the generation distribution as a complementary direction for test-time scaling in radiology reporting.

1 INTRODUCTION

Radiology report generation requires translating visual observations into clinically meaningful statements. A useful report must describe findings supported by the images while avoiding unsupported claims and clinically relevant omissions. Recent vision-language models (VLMs) provide a flexible foundation for this task, combining image understanding with fluent report generation (Jin et al., 2024; Wang et al., 2023; Bannur et al., 2024). Yet fluency alone does not establish factuality: a plausible report may overlook an abnormality or describe a finding that the images do not support. Improving the alignment between generated statements and image evidence therefore remains a central challenge, particularly when adapting a pretrained model at deployment without access to reference reports.

Test-time computation offers a way to improve reports from a frozen VLM. Instead of relying on a single decode, sampling-and-selection methods generate multiple candidates and rank them using uncertainty, agreement, or external evidence (Wang et al., 2025; Zhang et al., 2025; 2026; Thomas et al., 2025; Yao et al., 2026). Sampling exposes alternative descriptions, while selection attempts to identify the most reliable report among them. The effectiveness of this approach depends on both the quality of the candidate pool and the reliability of the selection signal. Agreement can favor repeated errors, and even an image-informed selector remains restricted to the reports already generated. When the generation policy is fixed, evidence used for selection does not feed back into the distribution that produces those reports.

Our analysis reveals a *coverage–quality gap* in this setting: enlarging the candidate pool recovers more correct findings without reliably improving the factuality of the selected report. Finding-level coverage aggregates information across candidates; different reports may contain different correct observations, alongside unsupported claims or omissions. Greater coverage therefore need not yield a more factual individual report, nor ensure that the selector identifies the best available one. This

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.

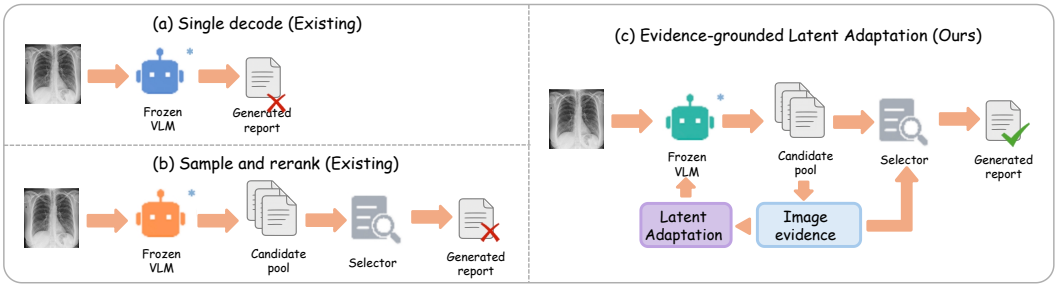

Figure 1: **Using evidence to guide generation and selection.** (a) Direct decoding produces a single report from a frozen VLM. (b) Fixed-policy sampling generates multiple reports and selects among completed candidates. (c) ELA additionally uses image-derived evidence feedback to optimize a study-specific latent state. This state adjusts the distribution used to generate the final candidate pool, which is then evaluated by a fixed evidence-based selector. The backbone remains frozen throughout.

motivates a complementary use of test-time computation: allow evidence to guide the generation distribution before the final candidate pool is formed. Figure 1 illustrates how evidence feedback can extend sampling and selection with study-specific latent adaptation.

We introduce *Evidence-grounded Latent Adaptation* (ELA), a framework that implements this principle through a small, temporary latent state within a frozen VLM. For each study, ELA iteratively samples reports and uses feedback from frozen visual evidence models to update only the latent state. The feedback penalizes unsupported assertions and omissions of evidence-supported findings, encouraging subsequent reports to better reflect the study images. The adapted state then conditions final candidate generation, and a fixed evidence-based selector chooses the output. Adaptation and selection thus act at different stages: the former adjusts which reports are likely to be generated, while the latter evaluates the resulting candidates. The procedure requires neither reference reports nor backbone parameter updates, and the latent state is reset for each study.

Experiments across general-purpose and medical VLMs on multiple chest X-ray benchmarks show consistent improvements in report factuality over greedy decoding. Controlled comparisons establish that image-evidence selection improves over clinical consensus on a shared candidate pool, and that latent adaptation provides further gains over a matched selection-only control. On IU-XRAY, adapted final candidate pools also outperform larger pools generated by the fixed policy. Together with analyses of recovered findings and adaptation components, these results support using visual evidence both to assess completed reports and to guide study-specific changes to the distribution from which they are generated.

2 Related Work

Radiology report generation and factual grounding. Radiology report generation has developed from encoder–decoder architectures to models incorporating memory and structured medical knowledge (Jing et al., 2018; Chen et al., 2020; Zhang et al., 2020; Liu et al., 2021), and more recently to LLM- and VLM-based systems (Jin et al., 2024; Wang et al., 2023; Lee et al., 2024). Within this progression, factual grounding has become an explicit modeling objective. MAIRA-2 associates generated findings with image locations (Bannur et al., 2024), while retrieval-augmented approaches incorporate exemplar reports and domain-specific supervision (Zhang et al., 2025). Preference optimization and reinforcement learning further align generation with clinical objectives (Xiao et al., 2025; Liu et al., 2025; Zhu et al., 2025; Jing et al., 2025). These approaches establish the value of clinical structure and evidence-aware learning during model development. Our focus is adaptation at deployment: given a pretrained generator, we use the current study's image evidence to optimize a temporary latent state without reference reports or backbone updates. The resulting adaptation is specific to that study and is discarded after report generation.

Test-time inference for radiology reporting. Additional inference-time computation can improve reporting through reliability assessment, candidate selection, and structured reasoning. Semantic consistency has been used to estimate report- and sentence-level uncertainty and support abstention

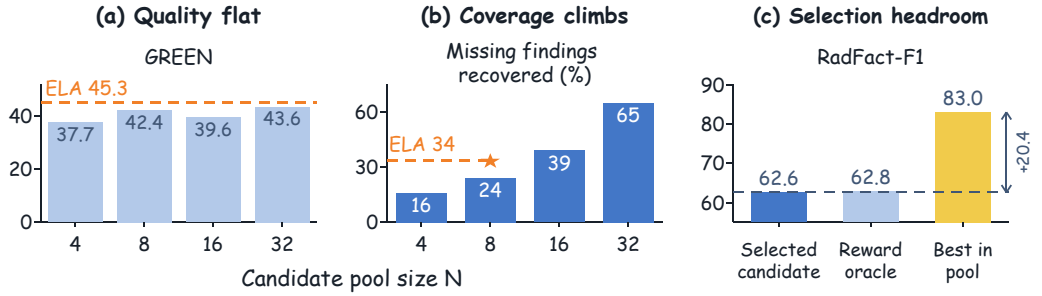

Figure 2: **Finding recovery and selected-report factuality measure different aspects of a candidate pool.** Candidate-pool diagnostics on IU-XRAY with Qwen2.5-VL-3B: (a) selected-report GREEN as the final pool grows; (b) recovery of reference findings absent from a designated no-update pool; (c) the within-pool selection audit, including the report selected by the evaluation metric itself.

(Wang et al., 2025). Clinical Consensus Selection ranks sampled reports by agreement within the candidate pool (Zhang et al., 2026), while process reward models assess sentence-level correctness and support report filtering and best-of-N selection (Thomas et al., 2025). Beyond selecting completed reports, Thought Graph Traversal structures organ-specific reasoning through prompting and extends the generation process using reasoning budget forcing (Yao et al., 2026). These approaches make different uses of inference-time computation; they should not all be reduced to post-hoc reranking. Our work studies a further mechanism: optimizing a continuous latent state using visual-evidence feedback before generating the final candidate pool. This mechanism complements candidate selection by adapting the distribution on which selection operates. Our coverage–quality analysis examines the distinction between recovering correct findings somewhere in a pool and obtaining a factual selected report.

Test-time adaptation and latent optimization. Test-time adaptation learns from unlabeled inputs using objectives such as entropy minimization and consistency across augmentations (Wang et al., 2021; Zhang et al., 2022). Continuous prompt and prefix tuning provide ways to condition frozen models through a small set of learnable vectors (Zhou et al., 2022; Li & Liang, 2021). Test-Time Prompt Tuning extends this principle to individual test images by optimizing prompts through entropy minimization over selected augmented views (Shu et al., 2022). For language-model reasoning, TTRL derives reinforcement-learning rewards from majority voting on unlabeled test data (Zuo et al., 2025). Most closely related, Latent Thought Policy Optimization updates instance-specific latent vectors through policy gradients while keeping the language model frozen, using a confidence-based reward from the model's own output distributions (Ye et al., 2025). Our method shares the principle of instance-specific latent optimization, but grounds its feedback in condition-level support estimated from the study images by frozen evidence models. The reward explicitly accounts for both unsupported assertions and omitted findings, and the adapted generator is paired with a fixed evidence-based selector. This distinction is relevant to radiology reporting because confidence or agreement alone does not establish whether a clinical statement is supported by the patient's images.

3 The Coverage–Quality Gap in Fixed-Policy Sampling

We examine what additional sampling contributes when the report generator and selector are held fixed. For an image set I, let $\mathcal{Y}_N$ denote a pool of N reports produced by a frozen VLM under a fixed generation protocol. The selector returns $\hat{y}_N = \arg\max_{y \in \mathcal{Y}_N} S(y, I)$. Two properties of this pipeline must be distinguished: whether a correct finding appears anywhere in the pool, and whether the selected report is factual as a whole. Our diagnostics use frozen Qwen2.5-VL-3B on IU-XRAY. Reference reports are used only for retrospective analysis, never for generation or selection.

More recovered findings, limited gains in the selected report. Sampling and selection already provide a substantial improvement over greedy decoding: an $N = 8$ pool raises GREEN from 24.54 to 42.38. Increasing the pool to $N = 32$ yields 43.62, with non-monotonic scores at intermediate pool sizes (Fig. 2a). Meanwhile, we ask whether a larger candidate pool recovers more missing findings than a new pool of the original size. Using positive CheXbert labels excluding *No Finding*,

we identify 98 reference findings absent from a designated $N = 8$ no-update pool across 296 studies. We t hen generated new fixed-policy pools, which recover 24% of these findings at $N = 8$ and 65% at $N = 32$ (Fig. 2b). Recovery is computed over the same 98 missing findings across all studies. Appendix A details this calculation and the separate candidate-generation runs used for panels (a) and (b).

Coverage does not determine report quality. Pool-level recovery credits a finding whenever any candidate mentions it; it does not penalize other errors in that candidate or require all recovered findings to appear in a single report. One candidate may mention an effusion and another an opacity, while both contain additional unsupported statements. Consequently, greater recovery need not improve the factuality of the selected report. We call the observed divergence between these readouts the *coverage–quality gap*. This is a finding about the evaluated generation and selection protocol, rather than a general claim that larger pools cannot contain better reports.

Selection leaves quality available in the pool. The within-pool audit of adapted candidate pools further distinguishes candidate availability from selection quality (Fig. 2c). The selected candidates obtain 62.6 RadFact-F1, whereas choosing each study's highest-RadFact-F1 candidate yields 83.0. This evaluation oracle uses information unavailable at inference and therefore measures headroom, not a deployable selection rule. The gap shows that the current selector leaves substantial quality unexploited in the audited pools. It also establishes that improved selection remains a viable direction.

These observations motivate a complementary intervention: use visual feedback to increase the likelihood of evidence-supported reports before forming the final candidate pool, while retaining a fixed selector. This changes what selection has to work with, rather than requiring selection alone to recover quality from an unchanged generator. Whether this intervention helps is an empirical question, which we examine through matched no-update and reward-shuffling controls in Section 5.

4 ELA: Evidence-Grounded Latent Adaptation

ELA uses visual evidence twice: to guide a temporary change to the report-generation distribution, and to select from the resulting candidates. For each study, we initialize a small latent state, sample reports, and update that state using evidence-based feedback. The retained state conditions the final candidate pool and is discarded after selection. The generator, report labeler, and visual evidence models remain frozen throughout. Figure 3 summarizes this procedure.

4.1 A Study-Specific Latent State

For a study with image set I, the frozen VLM defines a report policy $\pi_\theta(y \mid I)$. We condition this policy on a learnable state z to obtain $\pi_{\theta,z}(y \mid I)$. The main configuration uses $m = 2$ latent vectors at reserved positions preceding the image tokens in one early decoder layer. Writing H_ℓ for that layer's hidden states and $\mathcal{P}$ for the reserved positions, the intervention is

$$H_\ell[\mathcal{P}] \leftarrow H_\ell[\mathcal{P}] + z, \qquad z \in \mathbb{R}^{m \times d}, \tag{1}$$

where d is the hidden dimension. We initialize $z = 0$ for every study. The matched no-update policy $\pi_{\theta,0}$ retains the same reserved positions and decoder configuration. This control isolates latent updates from changes introduced by inserting the reserved positions. The multi-layer variants use a separate matrix z_ℓ at each adapted layer.

Why a frozen model can change its predictions. An autoregressive decoder predicts each token from both the images and the preceding context. Freezing its weights fixes the transformation applied to that context; it does not fix the output distribution when the context changes. Because the latent positions precede the report tokens, subsequent attention layers can propagate changes in z to the hidden states and next-token logits used for report generation. Continuous conditioning exploits this property in prefix tuning (Li & Liang, 2021); here, the conditioning state is optimized anew from each study's visual feedback. The latent vectors receive study-specific information through these updates, without requiring them to attend to later image tokens. The objective below determines which changes to this internal context are favored.

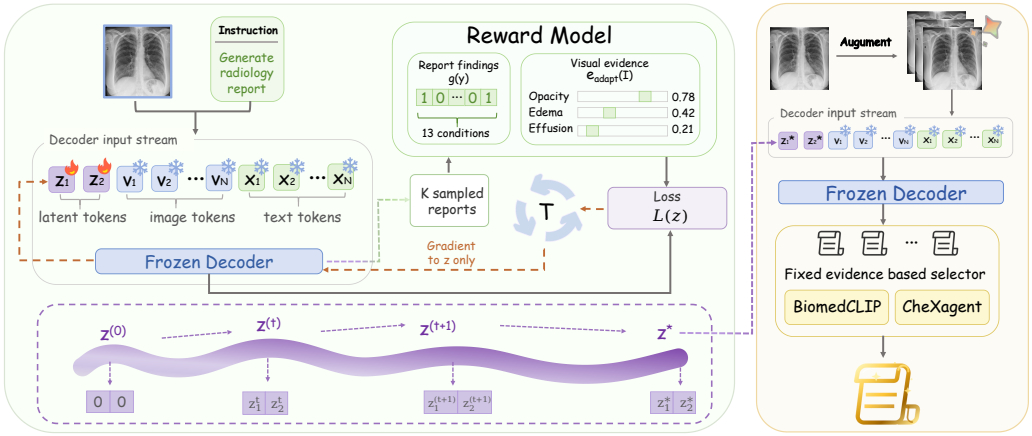

Figure 3: **Study-specific adaptation followed by report selection.** During adaptation, sampled reports are mapped to clinical findings and scored against frozen image-derived evidence. These scalar rewards guide updates to the latent state through the frozen generator's report log-probabilities; the evidence models are not differentiated through. After adaptation, the retained state conditions reports from the original and augmented images. A fixed evidence-based selector chooses the final report. Only the temporary latent vectors are optimized.

4.2 Learning from Visual Evidence

Feedback on assertions and omissions. Let $\mathcal{C}$ contain the 13 clinical conditions excluding *No Finding*. A frozen CheXbert-style labeler maps a report to binary indicators $g_c(y)$ of whether condition c is asserted (Smit et al., 2020). Scores from frozen CheXagent and TorchXRayVision models (Chen et al., 2024b; Cohen et al., 2022) are averaged to provide image-derived support scores $e_c^{\text{adapt}}(I) \in [0,1]$. These scores are model estimates, not reference labels. We compare each assertion with its estimated support using

$$R_{\text{adapt}}(y,I) = \frac{\sum_{c\in\mathcal{C}} w_c\, \tau\big(g_c(y), e_c^{\text{adapt}}(I)\big)}{\sum_{c\in\mathcal{C}} w_c},$$
$$\tau(g,e) = \begin{cases} 0, & g=1,\ e\geq 0.5,\\ 2e-1, & g=1,\ e<0.5,\\ 1-2e, & g=0. \end{cases} \tag{2}$$

The fixed weights $w_c = (\max(q_c, \epsilon_q))^{-\gamma}$ give more weight to conditions with lower fixed priors q_c; their values and the evidence protocol are given in Appendix A.

The default reward is the asymmetric reward $R_{\text{adapt}} = R_{\text{asym}}$. The score distinguishes two kinds of error. Reporting a condition with low support is penalized, while omitting a condition with high support also incurs a penalty. For example, at $e = 0.8$, mentioning the condition scores 0 and omitting it scores -0.6. At $e = 0.2$, the corresponding scores are -0.6 and 0.6. Thus supported assertions need no positive bonus to be favored: correcting an omission already raises the reward. The asymmetry makes suppressing a low-support assertion more valuable than adding an equally strongly supported missing condition. It encodes a particular error tradeoff, whose empirical effect is evaluated against a symmetric reward.

Turning report-level feedback into latent updates. At step t, we sample K complete reports from $\pi_{\theta,z_t}(\cdot \mid I)$ and score them with R_{adapt}. The normalized advantage $A_k = (R_{\text{adapt}}(y_k, I) - \mu_R)/(\sigma_R + \epsilon)$ indicates whether report y_k scores above or below its rollout group, where μ_R and σ_R are the group reward mean and standard deviation. We then teacher-force each sampled report through the generator and compute its mean token log-probability $\bar{\ell}_z(y_k \mid I) = |y_k|^{-1} \log \pi_{\theta,z}(y_k \mid I)$. The reward and labeler need not be differentiable: they provide fixed scalar weights, while gradients flow through the report log-probabilities to z.

We minimize the reward-weighted objective

$$\mathcal{L}(z) = -\frac{1}{K}\sum_{k=1}^{K} A_k \bar{\ell}_z(y_k \mid I) + \beta\, \widehat{\mathrm{KL}}_{\mathrm{k3}}(\pi_{\theta,z} \,\|\, \pi_{\theta,0}) + \lambda_\perp \mathcal{L}_\perp(z). \tag{3}$$

The first term encourages higher-reward reports and discourages lower-reward ones, with length normalization controlling each report's contribution. The KL term anchors the update to the matched zero-state policy. The final term penalizes latent alignment with a direction estimated from the association between rollout hidden states and selector scores. This is a heuristic regularizer, not a guarantee that adaptation and selection become independent. Its definition and the detailed latent-update equations are given in Appendix C. Only z is optimized, with advantages and the estimated direction held fixed during each update.

Local analysis of latent adaptation. To isolate the reward-weighted term, consider a small gradient step with the rollout group and advantages held fixed. Writing $u_t(y) = \nabla_z \bar{\ell}_z(y \mid I)|_{z=z_t}$, this step is $\Delta z = \eta K^{-1} \sum_{k=1}^{K} A_k u_t(y_k)$. Under local smoothness, a first-order expansion for any fixed report y gives

$$\bar{\ell}_{z_t+\Delta z}(y \mid I) - \bar{\ell}_{z_t}(y \mid I) = \frac{\eta}{K}\sum_{k=1}^{K} A_k\, u_t(y)^\top u_t(y_k) + O(\eta^2). \tag{4}$$

For a report absent from the rollout group, the first-order change is positive when its gradient aligns with the aggregate advantage-weighted rollout gradient. Appendix C provides the derivation and assumptions. This calculation isolates the reward term; empirical support for the full regularized AdamW procedure comes from the matched no-update and shuffled-reward controls in Table 3.

4.3 Final Generation and Selection

After adaptation, we retain a latent checkpoint z_I^* using R_{adapt}, without consulting reference reports or evaluation metrics. The retained state conditions the final pool $\mathcal{Y}_{z_I^*}$: one report from the original images and seven from augmented views in the main configuration.

The fixed selector combines condition-level support from BiomedCLIP and CheXagent (Zhang et al., 2023; Chen et al., 2024b). BiomedCLIP scores fixed positive and negative prompts using a two-way softmax over scaled image–text similarities; CheXagent normalizes the next-token logits for Yes and No in condition-specific visual queries. Writing their study-level support estimates as $e_c^{\mathrm{BM}}(I)$ and $e_c^{\mathrm{CX}}(I)$, we score a report by

$$S_{\mathrm{select}}(y, I) = \sum_{c \in \mathcal{C}} g_c(y) \left[(2e_c^{\mathrm{BM}}(I) - 1) + (2e_c^{\mathrm{CX}}(I) - 1) \right]. \tag{5}$$

The protocol aggregates each evidence model's support over the input images by a condition-wise maximum and applies a degeneration penalty. We return the highest-scoring candidate, $\hat{y} = \arg\max_{y \in \mathcal{Y}_{z_I^*}} S_{\mathrm{select}}(y, I)$, and discard z_I^*. Appendix A details checkpoint retention, image augmentation, report-return rules, and the inference budget, including the additional adaptation rollouts.

The two scores have distinct roles. The adaptation reward evaluates both asserted and omitted conditions to guide the latent update; the selector aggregates support for conditions asserted in each completed candidate. These proxy scores are never added together. The selector remains fixed when comparing adapted and no-update pools, so those comparisons isolate changes to candidate generation.

5 Experiments

We first evaluate whether ELA improves factuality across frozen backbones and datasets, then separate the contributions of evidence-based selection and latent adaptation. Additional comparisons examine final-pool size, recovered findings, and the adaptation objective.

Table 1: **Main results.** ELA improves GREEN and RadFact-F1 over greedy decoding across all nine backbone–dataset pairs.

	IU-XRAY (n=296)					CheXpert Plus (n=200)					MIMIC-CXR (n=2,347)					Overall
Method	**G**	**RF**	RaTE	R-L	CIDEr	**G**	**RF**	RaTE	R-L	CIDEr	**G**	**RF**	RaTE	R-L	CIDEr	Avg.
Qwen2.5-VL-3B	24.54	36.45	47.64	21.02	56.90	15.22	47.96	41.79	20.94	50.89	9.49	39.49	46.15	23.32	56.11	35.86
+ **ELA**	**45.26**	**65.10**	**49.45**	**23.62**	**69.66**	**17.48**	**55.97**	**41.93**	**21.08**	**53.73**	**12.79**	**47.90**	**47.71**	**23.37**	**58.56**	**42.24**
Qwen3-VL-4B	59.58	73.08	53.97	24.27	88.86	21.64	62.11	**45.33**	20.21	63.86	16.50	53.42	50.30	24.17	72.11	48.63
+ **ELA**	**63.47**	**76.79**	**54.83**	**24.88**	**91.99**	**22.20**	**64.44**	45.11	**20.52**	**67.78**	**18.10**	**57.96**	**51.16**	**24.28**	**74.70**	**50.55**
HuatuoGPT-7B	16.25	4.97	46.56	21.54	67.59	13.98	32.91	43.46	**21.08**	60.07	8.68	26.51	47.77	23.60	67.96	33.53
+ **ELA**	**22.88**	**10.77**	**47.51**	**22.65**	**71.70**	**17.61**	**45.28**	**45.10**	21.03	**63.72**	**11.82**	**35.69**	**48.91**	**24.11**	**70.91**	**37.31**

Table 2: **Selection and pool-size comparisons.** Image-evidence selection outperforms consensus on shared candidates, while ELA surpasses the larger fixed-policy pool on both factuality metrics.

Method	GREEN	RadFact-F1	RaTE	ROUGE-L	CIDEr	Average
Single sample	30.42	38.15	47.32	21.85	57.95	39.14
CCS consensus	35.17	43.18	48.26	21.88	57.34	41.17
Image-evidence selector (ours)	44.84	57.92	**50.06**	23.66	68.56	49.01
Fixed policy (N=32)	43.62	63.80	49.11	**23.72**	67.76	49.60
ELA (N=8)	**45.26**	**65.10**	49.45	23.62	**69.66**	**50.62**

5.1 Experimental Setup

Datasets and metrics. We evaluate on IU-XRAY (Demner-Fushman et al., 2016), CheXpert Plus (Chambon et al., 2024), and MIMIC-CXR (Johnson et al., 2019), following the MedEvalKit evaluation protocol (LASA Team et al., 2025). Our primary factuality metrics are GREEN and RadFact-F1, with GLM-4V-Flash used for RadFact-F1 evaluation (Ostmeier et al., 2024; Bannur et al., 2024; Zhipu AI, 2024). We additionally report RaTE, ROUGE-L, and CIDEr to assess clinical-entity consistency and reference-based report quality (Zhao et al., 2024; Lin, 2004; Vedantam et al., 2015). Table 1 abbreviate GREEN, RadFact-F1, and ROUGE-L as G, RF, and R-L, respectively. In Table 1, Overall averages the five metrics across all three datasets; bold marks the better result within each Base–ELA pair.

Backbones and comparison protocols. We use three frozen VLMs: Qwen2.5-VL-3B (Bai et al., 2025b), Qwen3-VL-4B (Bai et al., 2025a), and the medical-specific HuatuoGPT-7B (Chen et al., 2024a). Each Base–ELA pair uses the same backbone weights, study IDs, and report prompt. Base denotes greedy decoding; ELA additionally performs latent adaptation and candidate generation and selection. These comparisons evaluate the complete inference procedures and are not compute-matched. The controlled analyses below use Qwen2.5-VL-3B on IU-XRAY and distinguish shared-pool selector comparisons from matched no-update controls. Reference reports, test labels, and evaluation metrics are excluded from adaptation and selection. Full decoding, optimization, and augmentation settings are provided in Appendix A.

5.2 Main Results

Factuality improves across backbones and datasets. Table 1 shows improvements in both primary factuality metrics for all nine backbone–dataset combinations. The largest gains occur for Qwen2.5-VL-3B on IU-XRAY, where GREEN and RadFact-F1 increase by 20.72 and 28.65 points, respectively. Gains also extend to the stronger Qwen3-VL-4B baseline and the medical-specific HuatuoGPT-7B, although their magnitude varies across models and datasets. The broader report-quality metrics show a similar pattern: ELA improves 25 of the 27 RaTE, ROUGE-L, and CIDEr scores. These results establish the benefit of the complete inference procedure; the following analyses identify which components account for that benefit.

Table 3: **Component analysis.** Full ELA achieves the highest GREEN and RadFact-F1 among the evaluated variants.

Condition	GREEN	RadFact-F1	RaTE	ROUGE-L	CIDEr	Average
Greedy decoding	24.54	36.45	47.64	21.02	56.90	37.31
Selection only (z=0)	42.38	59.23	49.12	**23.89**	**70.82**	49.09
Latent update only	40.61	57.05	48.88	22.75	64.91	46.84
Shuffled evidence reward	42.81	62.07	49.23	23.76	68.59	49.29
ELA	**45.26**	**65.10**	**49.45**	23.62	69.66	**50.62**

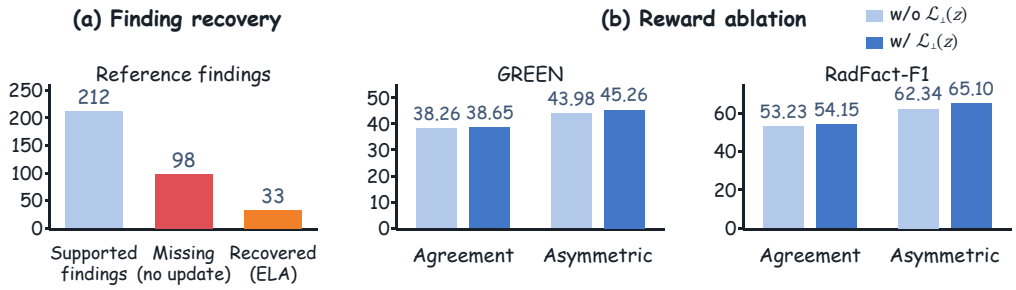

Figure 4: **Candidate-content and adaptation-objective diagnostics.** IU-XRAY with frozen Qwen2.5-VL-3B. (a) Findings present in the adapted pool that were absent from the matched no-update pool, using $N = 8$ final candidates. (b) Factuality under asymmetric and agreement rewards, with and without selector-direction regularization. Reference findings are used only for the retrospective audit.

5.3 Source of Improvements

Evidence-based selection on a shared candidate pool. We first hold candidate generation fixed and vary only the selector. The official CCS `qwen_vl_emb` consensus selector (Zhang et al., 2026) improves over a single sampled report. Applying our frozen image-evidence selector to these same candidates yields a further 9.67-point improvement in GREEN and 14.74 points in RadFact-F1. Because the candidate pool is shared, this comparison attributes the gain to the selection rule rather than to different generated reports. It supports the value of study-specific visual evidence for choosing among candidates in this setting.

Latent adaptation adds to selection. We next isolate the latent update using the component-study protocol in Table 3. All non-greedy variants use a final pool of $N = 8$. The no-update control retains the adaptation rollout schedule while fixing $z = 0$, as specified in Appendix A. With the selector held fixed, ELA improves GREEN by 2.88 points and RadFact-F1 by 5.87 over this control. Latent adaptation without the final selector also performs below the complete method on both factuality metrics, supporting the use of adaptation and selection together. To examine the role of report-specific feedback, we shuffle the report–reward pairing while retaining both components. This lowers GREEN by 2.45 points and RadFact-F1 by 3.03, indicating that correctly associating feedback with the sampled reports contributes to the adaptation gain.

Adaptation versus a larger final candidate pool. Having isolated the update, we compare it with increasing the number of fixed-policy candidates. Table 2 shows that ELA with $N = 8$ final candidates exceeds the fixed-policy $N = 32$ configuration on both GREEN and RadFact-F1. This suggests that adapting the generator can improve the quality available to a small final pool. The comparison concerns final-pool size, not total inference cost: ELA also generates reports during latent optimization.

Changes in candidate content. Figure 4(a) examines which findings appear in the generated pools. Of 212 reference findings identified by CheXbert in the audit, 98 are absent from the matched no-update candidates. The adapted $N = 8$ pool contains 33 of these missing findings. This shows a

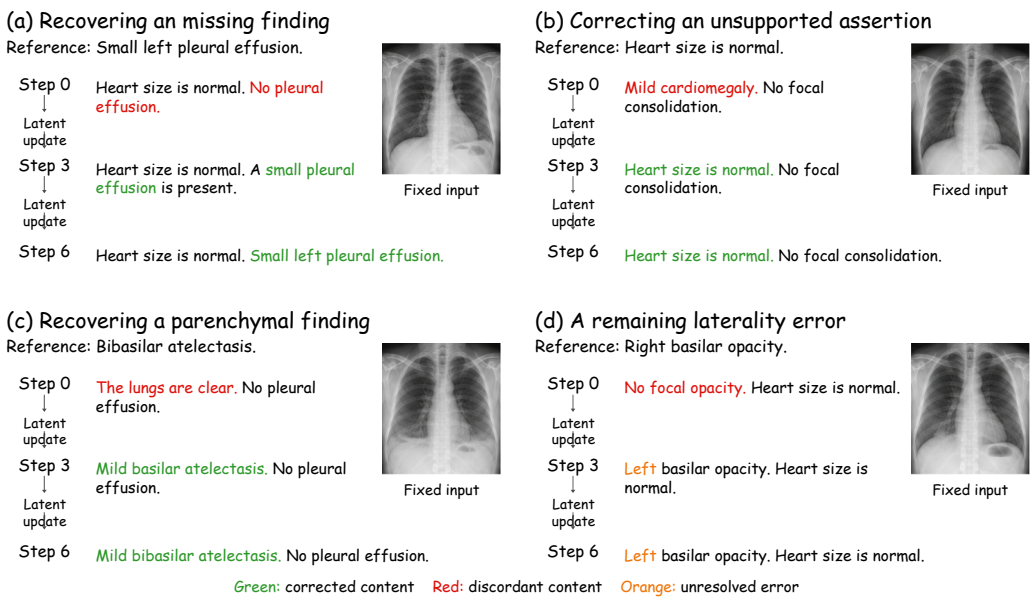

Figure 5: **Schematic examples of report changes under latent adaptation.** Each panel holds the image input fixed, with the reference line specifying an intended finding. The examples contrast corrections in finding presence with residual errors in clinical attributes. The chest images and report excerpts are constructed illustrations; the step labels do not represent recorded optimization trajectories.

change in observed candidate content beyond reranking the original pool. Together with the no-update and shuffled-reward controls, the result supports the role of evidence-guided updates in changing the candidates available for selection.

Reward design and regularization. Finally, we examine the choices within latent adaptation (Fig. 4b). Replacing the asymmetric reward with the symmetric agreement reward lowers both GREEN and RadFact-F1 in this comparison. Adding $\mathcal{L}_{\perp}(z)$ improves RadFact-F1 by 2.76 points under the asymmetric reward, compared with less than one point under the agreement reward. The benefit of the regularizer therefore depends on the reward formulation; the ablation supports its empirical use without establishing independence between adaptation and selection. Additional placement studies in Appendix B show that the tested early-layer configurations outperform the no-update control, whereas the tested late-layer configurations do not.

Scope of condition-level feedback. Figure 5 illustrates the distinction between recovering a finding and specifying its clinical attributes. For a fixed study, the adaptation reward assigns identical scores to reports with the same condition labels, even when their laterality, anatomical extent, or severity differs. Such attributes can change during generation, but their correctness is not explicitly supervised by the current feedback.

6 Conclusion

We identify a *coverage–quality gap* in test-time radiology report generation: larger candidate pools increase finding coverage without reliably improving the selected report. ELA introduces temporary latent tokens for each sample and optimizes them using condition-level visual evidence while keeping the VLM frozen. The optimized state guides final candidate generation, followed by evidence-based report selection. Across three frozen backbones and three chest X-ray datasets, ELA consistently improves factuality over greedy decoding. On IU-XRAY, image-evidence selection outperforms CCS on the same candidate pool, while ELA with $N{=}8$ final candidates outperforms a fixed-policy pool with $N{=}32$. These results show that evidence-grounded latent reasoning improves report generation by reshaping the candidate-generating distribution before final report selection.

AI USE STATEMENT

We used AI tools to polish the writing of this manuscript, including grammar, clarity, and concision. The use of pretrained models for report generation, image-evidence assessment, and automated evaluation is part of our research methodology and is documented in the paper. The authors are responsible for the research design, implementation, experimental analysis, and conclusions. We reviewed all AI-assisted text before incorporating it and remain responsible for the final manuscript.

ETHICS STATEMENT

ELA studies radiology report generation using the existing IU-XRAY, CheXpert Plus, and MIMIC-CXR benchmarks. Pretrained generators, evidence models, and automated evaluators may inherit biases from their training data and can produce incorrect or incomplete assessments. Condition-level visual support does not verify every clinical statement in a generated report. Improvements in automatic factuality metrics do not establish clinical safety, demographic fairness, or uniform benefits across clinical conditions; our analysis also identifies condition-specific performance degradations. The method is intended for research, and any clinical use would require prospective validation and review by qualified clinicians. Access to and use of the underlying images and reports remain subject to the original datasets' access requirements and data-use terms.

REPRODUCIBILITY STATEMENT

Section 4 specifies the latent-state parameterization, evidence-grounded reward, optimization objective, and final report-selection procedure. Section 5 describes the datasets, frozen backbones, evaluation metrics, and comparison protocols. Appendix A provides the optimization schedule, evidence-model configuration, augmentation settings, candidate budgets, and evaluation conventions. Appendix B reports latent-token placement studies and condition-level error analyses. Reference reports, test labels, and downstream evaluation metrics are excluded from latent-state optimization and final report selection.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a. doi: 10.48550/arXiv.2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Perez-Garcia, Valentina Salvatelli, Harshita Sharma, et al. MAIRA-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024. doi: 10.48550/arXiv.2406.04449.

Pierre J. Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven Q. H. 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. *CoRR*, abs/2405.19538, 2024. doi: 10.48550/ARXIV.2405.19538. URL https://doi.org/10.48550/arXiv.2405.19538.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*, 2024a. doi: 10.48550/arXiv.2406.19280.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of EMNLP*, 2020.

Zhihong Chen, Maya Varma, Justin Xu, Magdalini Paschali, Dave Van Veen, Andrew Johnston, Alaa Youssef, Louis Blankemeier, Christian Bluethgen, Stephan Altmayer, et al. A vision-language foundation model to enhance efficiency of chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024b.

Joseph Paul Cohen, Joseph D. Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P. Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, and Hadrien Bertrand. Torchxrayvision: A library of chest x-ray datasets and models. In *Proceedings of the 5th International Conference on Medical Imaging with Deep Learning*, volume 172 of *Proceedings of Machine Learning Research*, pp. 231–249. PMLR, 2022. URL https://proceedings.mlr.press/v172/cohen22a.html.

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

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. Promptmrg: Diagnosis-driven prompts for medical report generation. *AAAI*, 2024.

Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. In *Proceedings of ACL*, 2018.

Peiyuan Jing, Kinhei Lee, Zhenxuan Zhang, Huichi Zhou, Zhengqing Yuan, Zhifan Gao, Lei Zhu, Giorgos Papanastasiou, Yingying Fang, and Guang Yang. Reason like a radiologist: Chain-of-thought and reinforcement learning for verifiable report generation. *arXiv preprint arXiv:2504.18453*, 2025. doi: 10.48550/arXiv.2504.18453.

Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Yifan Peng, et al. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

LASA Team, Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Jie Tan, Deli Zhao, Tingyang Xu, Hao Zhang, and Yu Rong. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025.

Suhyeon Lee, Won Jun Kim, Jinho Chang, and Jong Chul Ye. Llm-cxr: Instruction-finetuned llm for cxr image understanding and generation. In *Proceedings of ICLR*, 2024.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of ACL*, 2021. doi: 10.18653/v1/2021.acl-long.353.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of CVPR*, 2021.

Hong Liu, Dong Wei, Zhe Xu, Xian Wu, Yefeng Zheng, and Liansheng Wang. RRG-DPO: Direct preference optimization for clinically accurate radiology report generation. In *Proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pp. 552–562. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-032-04971-1_52. URL https://papers.miccai.org/miccai-2025/0795-Paper1273.html.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S. Chaudhari, and Jean-Benoit Delbrouck. GREEN: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 374–390. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.21. URL https://aclanthology.org/2024.findings-emnlp.21/.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2209.07511.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In *Proceedings of EMNLP*, 2020.

Alois Thomas, Maya Varma, Jean-Benoit Delbrouck, and Curtis P. Langlotz. Process reward models for sentence-level verification of lvlm radiology reports. *arXiv preprint arXiv:2510.23217*, 2025.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of CVPR*, 2015.

Chenyu Wang, Weichao Zhou, Shantanu Ghosh, Kayhan Batmanghelich, and Wenchao Li. Semantic consistency-based uncertainty quantification for factuality in radiology report generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1739–1754. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.95. URL https://aclanthology.org/2025.findings-naacl.95/.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *Proceedings of ICLR*, 2021. doi: 10.48550/arXiv.2006.10726.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology*, 1(3):100033, 2023.

Ting Xiao, Lei Shi, Peng Liu, Zhe Wang, and Chenjia Bai. Radiology report generation via multi-objective preference optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8664–8672, 2025. doi: 10.1609/aaai.v39i8.32936. URL https://ojs.aaai.org/index.php/AAAI/article/view/32936.

Yue Yao, Zelin Wen, Yan Tong, Xinyu Tian, Xuqing Li, Xiao Ma, Dongliang Xu, and Tom Gedeon. Thought graph traversal for test-time scaling in chest x-ray VLLMs. *Pattern Recognition*, 179:113639, 2026. doi: 10.1016/j.patcog.2026.113639.

Wengao Ye, Yan Liang, and Lianlei Shan. Thinking on the fly: Test-time reasoning enhancement via latent thought policy optimization. *arXiv preprint arXiv:2510.04182*, 2025. doi: 10.48550/arXiv.2510.04182.

Baochang Zhang, Chen Jia, Shuting Liu, Heribert Schunkert, and Nassir Navab. Semantic-aware chest x-ray report generation with domain-specific lexicon and diversity-controlled retrieval. In *Proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pp. 607–616. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-032-04978-0_58. URL https://papers.miccai.org/miccai-2025/0821-Paper3102.html.

Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. In *Proceedings of NeurIPS*, 2022.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

Xi Zhang, Yingshu Li, Zaiqiao Meng, Jake Lever, and Edmond S. L. Ho. Ccs: Clinical consensus selection for radiology report generation. *arXiv preprint arXiv:2605.30131*, 2026.

Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. When radiology report generation meets knowledge graph. In *Proceedings of AAAI*, 2020.

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. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.836. URL https://aclanthology.org/2024.emnlp-main.836/.

Zhipu AI. GLM-4V-Flash. https://docs.bigmodel.cn/cn/guide/models/free/glm-4v-flash, 2024. Accessed July 15, 2026.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. doi: 10.1007/s11263-022-01653-1.

Kangyu Zhu, Peng Xia, Yun Li, Hongtu Zhu, Sheng Wang, and Huaxiu Yao. MMedPO: Aligning medical vision-language models with clinical-aware multimodal preference optimization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 80207–80222. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhu25v.html.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding, and Bowen Zhou. TTRL: Test-time reinforcement learning. *arXiv preprint arXiv:2504.16084*, 2025. doi: 10.48550/arXiv.2504.16084.

APPENDIX

A EXPERIMENTAL PROTOCOL

Evaluation setup. We evaluate Qwen2.5-VL-3B, Qwen3-VL-4B, and HuatuoGPT-7B on the IU-XRAY, CheXpert Plus, and MIMIC-CXR splits used by MedEvalKit. Each Base–ELA pair uses the same study IDs, frozen backbone weights, and report prompt. Base uses greedy decoding, whereas ELA includes latent adaptation, candidate generation, and final selection. These comparisons evaluate the complete inference procedures and are not compute-matched. Unless otherwise stated, component and placement studies use Qwen2.5-VL-3B on IU-XRAY.

Following the MedEvalKit protocol, the reference supplied to the evaluation suite concatenates the findings and impression fields as "Findings: {findings} Impression: {impression}." GREEN and RadFact-F1 are the primary metrics, with GLM-4V-Flash used for RadFact-F1 evaluation; RaTE, ROUGE-L, and CIDEr are secondary metrics. All metric values are reported on a $100\times$ scale. Evaluation runs separately from generation and latent optimization. Reference reports, test labels, and downstream evaluation scores are not available to adaptation, checkpoint retention, or final selection. No hyperparameter is selected using the evaluated test references.

Latent state and optimization. The main configuration uses two zero-initialized latent vectors at reserved positions preceding the image tokens, with injection before the first language-decoder block. Only these vectors are optimized; the backbone, visual evidence models, and report labeler remain frozen. The state is initialized independently for every study and discarded after final selection. No pretrained prompt is loaded.

We schedule $T = 6$ iterations with $K = 8$ rollout reports per iteration. The implementation uses AdamW8bit with learning rate 10^{-2}, $(\beta_1, \beta_2) = (0.9, 0.999)$, optimizer epsilon 10^{-8}, and zero weight decay. We clip the latent-gradient norm at 1.0 before each optimizer step. The KL coefficient is $\beta = 0.1$, and the selector-direction penalty uses $\lambda_\perp = 1.0$. The advantage denominator uses epsilon 10^{-6}; direction and latent-vector normalization use epsilon 10^{-8}. These constants serve different numerical roles.

Decoding and stopping. Adaptation rollouts use the original study images, sampling temperature 1.0, top-$p = 0.95$, and a maximum of 200 new tokens. Checkpoint evaluation and final candidate generation use greedy decoding with a maximum of 512 new tokens. The repetition penalty is 1.0 in all three stages. Sampling inherits top-k from the loaded backbone's generation configuration; the inference wrapper does not override it. Generation stops at a model-defined end-of-sequence token or the stage-specific length limit. Padding is removed before report scoring.

Degenerate reports and group advantages. A report is degenerate if it is empty, contains the placeholder `{}`, or has fewer than four alphanumeric tokens after removing the findings and impression headings. Reports with at least 20 tokens are also rejected when their ratio of unique tokens to total tokens is below 0.25. The reward service assigns these reports -9.0. For optimization, rewards at or below -4.5 identify degenerate rollouts; they are excluded from the group mean and sample standard deviation. With at least two valid rewards and standard deviation above 10^{-6}, valid advantages follow Equation A2 and degenerate rollouts receive advantage -1.5. If valid and degenerate rollouts coexist but the valid rewards have insufficient variation, their advantages are $+1.5$ and -1.5, respectively. Otherwise, an all-degenerate or constant-reward group skips the update. Thus T counts scheduled iterations rather than guaranteed optimizer steps.

KL estimation and selector-direction regularization. The reference forward pass disables latent injection while retaining the same input and reserved token positions. For sampled token $y_{k,r}$, let $\delta_{k,r} = \log \pi_{\theta,0}(y_{k,r} \mid I, y_{k,<r}) - \log \pi_{\theta,z}(y_{k,r} \mid I, y_{k,<r})$. The implemented k3 estimate is

$$\widehat{\mathrm{KL}}_{\mathrm{k3}} = \frac{1}{K} \sum_{k=1}^{K} \sum_{r=1}^{L_k} \left[\exp(\delta_{k,r}) - \delta_{k,r} - 1 \right].$$

The policy-gradient term averages each report's token log-probabilities, whereas the KL term sums over tokens before averaging over reports. The selector direction uses the frozen selection scores and mean-pooled generated-token hidden states at the deepest adapted layer. The penalty sums over all latent vectors at all adapted layers. Both advantages and the estimated direction are detached during optimization; Appendix C gives the equations.

Visual evidence for adaptation and selection. The asymmetric reward uses CheXagent and TorchXRayVision (XRV) condition scores, with equal-weight fusion after aggregating each model's evidence across the original study images (Chen et al., 2024b; Cohen et al., 2022): $e_c^{\text{adapt}}(I) = [e_c^{\text{CX}}(I) + e_c^{\text{XRV}}(I)]/2$. The evidence cache stores a condition-wise maximum across images for each model. CheXagent-2-3b support is the two-way softmax of Yes and No logits at the first answer-token position. The reward service uses the stored support scores directly, without fitting an additional calibration model or subtracting per-condition medians in this asymmetric-reward branch. Evidence comes from the original images and stays fixed across latent updates and augmented-view generation. The IU-XRAY caches contain matching entries for all 296 studies.

Final report selection uses the BiomedCLIP–CheXagent positive-claim score in Equation 5, with a separate maximum over images for each model. BiomedCLIP uses a two-way softmax over scaled similarities to fixed positive and negative condition prompts. The selector excludes *No Finding* and assigns degenerate reports a score of -9.0. Adaptation reward and selection score are evaluated separately and are never added together.

Report labels and condition weights. Reports are labeled by `f1chexbert` in `rrg` mode, using the CheXbert checkpoint supplied through `StanfordAIMI/RRG_scorers` (Smit et al., 2020). Positive and uncertain labels map to $g_c(y) = 1$; negative and blank labels map to zero. Empty reports receive an all-zero label vector and are handled by the degeneration rule above. We exclude *No Finding* from reward and selection scoring.

The inverse-prior weights use $\gamma = 0.5$ and $\epsilon_q = 10^{-3}$. The implementation fixes the following q_c values, using them as condition-prior constants rather than estimating them from the evaluated test references:

Condition	q_c
Enlarged Cardiomediastinum	0.10
Cardiomegaly	0.12
Lung Opacity	0.24
Lung Lesion	0.04
Edema	0.13
Consolidation	0.07
Pneumonia	0.05
Atelectasis	0.16
Pneumothorax	0.04
Pleural Effusion	0.26
Pleural Other	0.02
Fracture	0.04
Support Devices	0.30

Checkpoint retention. For the full method, we evaluate one greedy report after each scheduled iteration and retain the checkpoint with the highest adaptation reward.

Final candidates and image augmentation. The retained state conditions $N = 8$ final reports: seven from augmented versions of the complete image set and one from the original images, all decoded greedily. Each image in each augmented version independently receives a random ordering of the three transforms below. Their application probabilities are 0.7, 0.6, and 0.4 for crop, brightness/contrast, and window-level adjustment, respectively.

For cropping, a scale $s \sim \mathcal{U}(0.82, 0.97)$ multiplies both image height and width, rounded to the nearest integer. The crop location is sampled uniformly among valid positions. Thus the interval specifies retained side lengths; the corresponding area fraction is approximately s^2. Brightness/contrast uses Albumentations `RandomBrightnessContrast` with brightness limit 0.08 and contrast limit 0.10. The contrast multiplier lies in $[0.90, 1.10]$, and the brightness offset lies in $[-0.08, 0.08]$ times the maximum image intensity.

For window-level adjustment, let l and h be the image's second and 98th intensity percentiles, and set $r = \max(h - l, 1)$. The center is $(l + h)/2 + ur$, where $u \sim \mathcal{U}(-0.15, 0.15)$, and the width is vr, where $v \sim \mathcal{U}(0.6, 1.4)$. Intensities are remapped linearly within that window, clipped to $[0, 255]$,

and converted to 8-bit values. The width divisor is bounded below by 1. This preset does not use flips, rotations, geometric distortions, blur, or hue/saturation perturbations.

Final report return. The retained greedy report is returned when all selection scores tie or the selected candidate fails to improve its adaptation reward. For the full method, the initial Base report remains eligible and is returned if the output does not improve its reward. These decisions use image-derived feedback without reference reports or evaluation metrics.

Component controls (Table 3). The matched no-update control retains the reserved latent positions, injection configuration, $T = 6$, $K = 8$ rollout schedule, checkpoint evaluation, decoding settings, evidence interface, and selector, while fixing $z = 0$. It performs no backward pass or optimizer step. This matches rollout and evaluation opportunities rather than FLOPs or wall-clock time. Unlike this control, greedy Base performs no adaptation rollouts or final-pool selection.

The *Latent update only* control retains the optimization and checkpoint-evaluation procedure but omits final-pool selection. It returns the retained checkpoint's greedy report. For *Shuffled evidence reward*, each iteration independently permutes the K rewards within its rollout group before computing advantages. This branch retains the final scheduled checkpoint so reward-based checkpoint selection cannot restore the disrupted pairing. It then generates and selects from a final pool of $N = 8$. The retained-report comparison still applies, but the additional Base-report check is disabled for this control.

Final-pool comparison (Table 2). The fixed-policy $N = 32$ row follows the matched no-update protocol above, with 31 augmented greedy reports and one original-image greedy report. The adapted $N = 8$ row uses the retained latent state to generate seven augmented reports and one original-image report, also greedily. Both use the same BM+CX selector and return rules. The comparison varies final-pool size and latent adaptation.

Reward and regularization ablation (Fig. 4(b)). The symmetric alternative to the asymmetric adaptation reward is $R_{\text{agree}}(y, I) = 13^{-1} \sum_{c \in \mathcal{C}} (2g_c(y) - 1)(2e_c^{\text{CX}}(I) - 1)$. This branch uses CheXagent evidence; the asymmetric branch uses the CX/XRV mean above. For each reward formulation, we compare latent optimization with and without $\mathcal{L}_\perp(z)$. The reward comparison includes its evidence route; final BM+CX selection stays fixed.

Inference budget. Adaptation schedules $TK = 48$ rollout reports, in addition to checkpoint-evaluation decodes and, when selection is enabled, the final candidate pool. Consequently, N describes final candidates rather than the total generation or computation budget.

Reference-finding recovery audit. For Figures 2(b) and 4(a), we label reference reports and candidates using CheXbert in `rrg` mode. We retain the 13 conditions excluding *No Finding*, using the binary mapping specified above. The audit uses reference-label support, without an additional image-evidence filtering step. A finding is counted as a study–condition pair, and is present in a candidate pool if any report asserts it. Across 296 studies, a designated matched no-update pool with $N = 8$ misses 98 of the 212 reference findings. These 98 pairs form the fixed denominator for the recovery analysis. For Figure 4(a), both pools contain $N = 8$ reports for the same 296 studies. The adapted pool contains 33 of the 98 previously missing pairs, leaving 65 unrecovered.

The newly generated $N = 8$ pool provides a comparison at the same candidate budget, allowing recovery from regenerating candidates to be compared with recovery at a larger pool size. Reusing the pool that defines the missing findings would give zero recovery by construction.

We separately generate comparison pools with $N \in \{4, 8, 16, 32\}$ for the same studies, keeping $z = 0$ and using one original-image report and $N - 1$ augmented-view reports, each decoded greedily. These audit runs skip latent-optimization iterations and use the `clinical` augmentation preset. The pools are not nested subsets of a common larger pool. They recover 16, 24, 38, and 64 of the fixed 98 missing pairs, respectively. Recovery is micro-averaged: we sum recovered pairs over all studies and divide by 98, giving 16.33%, 24.49%, 38.78%, and 65.31%. The quality curve in panel (a) comes from separate no-update runs. Reference labels are used only for this retrospective audit and are unavailable during inference.

Adapted-pool selection audit. Figure 2(c) uses a separate archived candidate audit on the same 296 studies, $m = 2$, $T = 6$, $K = 8$, and a final pool of $N = 8$. This audit differs from the fixed-policy diagnostics in panels (a,b). The three bars report RadFact-F1 for the candidate selected before fallback,

the candidate with the highest adaptation reward, and the candidate with the highest RadFact-F1, respectively. The last choice uses reference-based evaluation unavailable during inference.

B LATENT-STATE DIAGNOSTICS

B.1 DECODER-DEPTH PLACEMENT

Table A1 compares latent injection at early and late decoder layers while keeping the backbone frozen. Both early placements improve GREEN and RadFact-F1 over the matched no-update control, whereas both late placements score below it on these two metrics. Extending early injection from one layer to nine leaves lowers GREEN and RadFact-F1. Under this diagnostic protocol, modifying the latent context earlier is therefore more effective than increasing the number of injection layers. The results support the choice of a single early injection site among the tested configurations.

Table A1: **Latent injection depth.** Early-layer injection achieves higher GREEN and RadFact-F1 than late-layer injection in this diagnostic.

Placement	GREEN	RF	RaTE	R-L	CIDEr
Matched no-update (z=0)	42.38	59.23	49.12	23.89	70.82
Early, 1 block (pre-1)	45.26	65.10	49.45	23.62	69.66
Early, 9 blocks (1–9)	45.00	61.74	50.37	24.09	70.11
Late, 1 block (pre-36)	40.28	58.28	49.17	23.96	69.37
Late, 9 blocks (28–36)	38.01	56.69	48.50	23.72	67.11

B.2 LATENT-TOKEN INSERTION POSITION

Table A2 compares two latent-token insertion positions against a selection-only baseline in a separate nine-layer diagnostic. The baseline keeps $z = 0$ and performs no adaptation rollouts. It selects among $N = 8$ greedy reports: one from the original images and seven from augmented views. The Δ column reports the GREEN-point gain over this baseline. Placing the latent tokens before the image tokens yields a small numerical advantage over placing them between the image and text tokens.

Table A2: **Latent-token insertion position.** Both insertion positions improve GREEN over selection only.

Configuration	GREEN	Δ
Selection only (z=0)	39.89	0.00
Before image tokens	44.37	+4.48
Between image and text tokens	43.99	+4.10

B.3 CONDITION-LEVEL ERROR ANALYSIS

The condition-level analysis in Figure A1 shows that aggregate improvements coexist with heterogeneous changes within the clinical label set. Pleural Effusion and Cardiomegaly improve across all three datasets, while Lung Opacity improves on CheXpert Plus and MIMIC-CXR but declines on IU-XRAY. The displayed conditions also include regressions in Lung Lesion and Enlarged Cardiomediastinum on IU-XRAY, Pneumonia on CheXpert Plus, and Fracture on MIMIC-CXR. Condition-level feedback therefore does not imply uniform improvement across its target conditions.

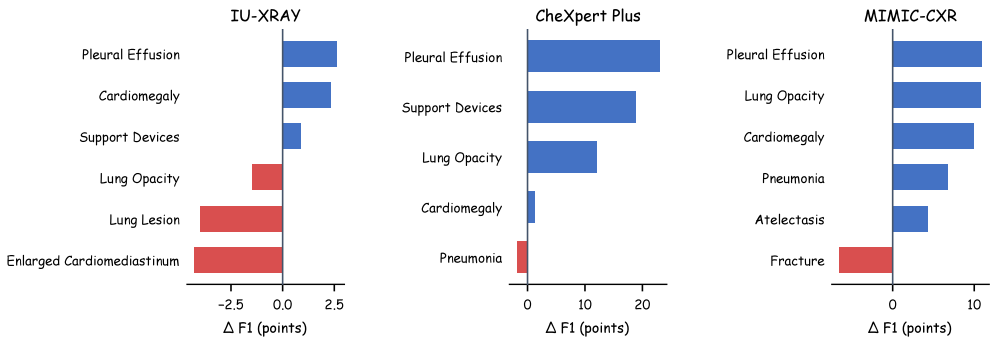

Figure A1: Reported changes in condition-level CheXbert-F1 under ELA on three chest X-ray datasets. Each panel displays a subset of conditions. Blue and red bars indicate positive and negative F1-point changes, respectively; horizontal-axis scales differ across panels.

C Latent Update: Details and Local Interpretation

Effect of reward asymmetry. For a fixed image set, write $e_c = e_c^{\text{adapt}}(I)$ and $W = \sum_c w_c$. The reward can equivalently be expressed as

$$R_{\text{adapt}}(y, I) = C(I) + \frac{1}{W}\sum_c w_c a(e_c) g_c(y), \qquad a(e) = \begin{cases} 4e - 2, & e < 0.5, \\ 2e - 1, & e \geq 0.5, \end{cases} \tag{A1}$$

where $C(I) = W^{-1}\sum_c w_c(1 - 2e_c)$ is independent of the report. This follows by subtracting $\tau(0, e)$ from $\tau(1, e)$. For equally strong support and non-support, asserting the low-support condition carries twice the magnitude of the benefit of asserting the high-support condition, before prevalence weighting. Group centering cancels $C(I)$, leaving these condition-specific assertion preferences to determine the advantages. Consequently, the reward's distinction from the final selector lies in its evidence inputs, weights, and asymmetric coefficients; Neither score models interactions between conditions or verifies attributes outside the label set.

Report-level optimization. For rollouts $y_k \sim \pi_{\theta, z_t}(\cdot \mid I)$, the advantage and length-normalized log-probability used in the main objective are

$$A_k = \frac{R_{\text{adapt}}(y_k, I) - \mu_R}{\sigma_R + \epsilon}, \tag{A2}$$

$$\bar{\ell}_z(y_k \mid I) = \frac{1}{L_k}\sum_{r=1}^{L_k} \log \pi_{\theta, z}(y_{k,r} \mid I, y_{k,<r}). \tag{A3}$$

The sampled tokens, rewards, and advantages are held fixed while differentiating the objective. In particular, the visual evidence models and report labeler supply feedback without requiring a differentiable mapping from report text to reward. The per-report length normalization is part of the optimization surrogate; we do not identify it with an unbiased gradient of unnormalized expected report reward.

Selector-direction regularization. Let h_k be the mean-pooled residual state of the generated tokens in report y_k at the adapted layer, and s_k its score from the frozen scoring function used by the regularizer. We estimate

$$\hat{s}_t = \frac{\sum_k (s_k - \bar{s})(h_k - \bar{h})}{\left\|\sum_k (s_k - \bar{s})(h_k - \bar{h})\right\|_2 + \epsilon}, \qquad \mathcal{L}_\perp(z) = \sum_{j=1}^{m} \left(\frac{z_j^\top \hat{s}_t}{\|z_j\|_2 + \epsilon} \right)^2. \tag{A4}$$

Here $\bar{s}$ and $\bar{h}$ are rollout-group means and z_j is a row of the latent matrix. The direction is detached from the gradient computation during each update. It describes an empirical association between hidden representations and scores, not the gradient of the selector with respect to z. Since the decoder's mapping from latent positions to generated-token states is nonlinear, orthogonality in these

coordinates does not imply unchanged selector scores or statistical independence between adaptation and selection. The effect of this regularizer must therefore be established by ablation rather than by a geometric guarantee.

Local analysis of latent adaptation. The autoregressive policy factorizes as $\pi_{\theta,z}(y \mid I) = \prod_r \pi_{\theta,z}(y_r \mid I, y_{<r})$. Changing z changes the internal context used to compute these conditional probabilities even when θ is fixed. For a local interpretation, flatten z into a vector and define $u_t(y) = \nabla_z \bar{\ell}_z(y \mid I)|_{z=z_t}$. Consider an illustrative small gradient step on the reward-weighted term alone, with the current rollout group and its advantages fixed:

$$\begin{aligned} \Delta z &= \frac{\eta}{K} \sum_{k=1}^{K} A_k u_t(y_k), \\ \bar{\ell}_{z_t+\Delta z}(y \mid I) - \bar{\ell}_{z_t}(y \mid I) &= \frac{\eta}{K} \sum_{k=1}^{K} A_k \, u_t(y)^\top u_t(y_k) + O(\eta^2). \end{aligned} \tag{A5}$$

For any fixed report y, assume that $\bar{\ell}_z(y \mid I)$ is twice continuously differentiable in a neighborhood of z_t, with Hessian operator norm bounded by M_y. Taylor's theorem gives the change $u_t(y)^\top \Delta z + r_y$, where $|r_y| \le \frac{1}{2} M_y \|\Delta z\|_2^2$ for a sufficiently small step. With the rollout group and advantages fixed, $\|\Delta z\|_2 = O(\eta)$. Substituting the first line therefore yields the second, which is Equation 4 in the main text. This argument applies to any fixed candidate y, including one not sampled in the current group. The inner products describe shared sensitivity to the latent state: positive alignment with a positive-advantage rollout contributes an increase in the candidate's normalized log-probability, while positive alignment with a negative-advantage rollout contributes a decrease. Contributions from different rollouts can conflict; no individual report is guaranteed to improve. For a fixed report length, a change in mean token log-probability has the same sign as the change in its sequence log-probability.

This calculation explains how feedback on sampled reports can affect future candidates through shared conditioning, without explicitly combining their text. It is a local interpretation, not a derivation of the full optimizer: the implementation uses AdamW, gradient clipping, a KL anchor, and the direction regularizer. It establishes neither monotonic improvement in expected reward nor improvement in reference-based factuality. Those conclusions require empirical evidence and depend on the quality of the visual reward.

Scope of the visual signal. The reward evaluates assertions over a fixed set of clinical conditions. Reports with identical condition labels receive the same reward, even if they differ in laterality, location, severity, temporal comparison, or wording. Within-group normalization removes reward components shared by all rollouts, but it does not resolve these representational limits. When the valid reports all receive the same reward, the protocol skips the update. Thus the mechanism depends on both useful rollout variation and informative evidence; it is not a verifier of every statement in a report.