REASON BEFORE YOU GENERATE: LATENT CONSISTENCY CONTROL FOR RADIOLOGY REPORT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

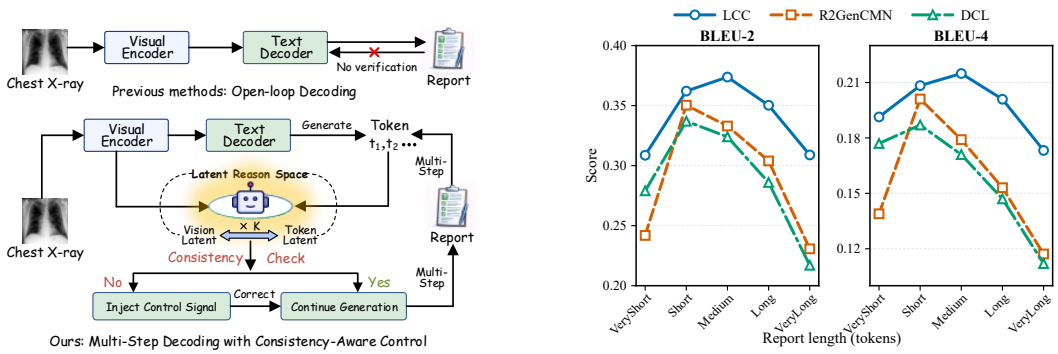
Radiology report generation aims to automatically produce clinically coherent diagnostic reports from medical images. However, most existing RRG methods follow an open-loop autoregressive generation paradigm, where multimodal evidence is primarily used for token prediction, while the evolving report state is rarely re-evaluated against image-grounded semantics during decoding. To address this issue, we propose Latent Consistency Control (LCC), which establishes a shared cross-modal latent space and explicitly regulates semantic consistency throughout autoregressive generation. LCC first condenses visual and textual representations into compact semantic prototypes and performs iterative latent cross-modal reasoning to obtain a unified reasoning representation. Building on this latent space, we introduce a multi-step consistency-aware decoding mechanism that periodically evaluates the generated prefix against image-grounded semantics and conditionally injects a lightweight corrective signal when potential inconsistency is detected. In this way, LCC enables intermediate verification and corrective regulation rather than relying solely on sequence-level generation. Experiments on benchmark RRG datasets demonstrate consistent improvements in both natural language generation quality and clinical factual correctness, outperforming representative state-of-the-art methods on most metrics.

1 INTRODUCTION

Radiology Report Generation (RRG) addresses the automatic synthesis of clinically grounded diagnostic narratives from medical images such as chest X-rays (Hou et al., 2025a;b). As radiographic examinations constitute a cornerstone of routine clinical diagnosis and treatment planning, radiologists are required to carefully analyze both normal anatomical structures and subtle abnormalities across multiple regions before documenting comprehensive findings. This process is time-intensive and cognitively demanding, particularly in high-throughput clinical environments (Jing et al., 2019; Yan & Pei, 2022; Wang et al., 2023). Consequently, developing automated RRG systems has become an influential research direction in medical imaging and multimodal learning, with the potential to alleviate radiologists' workload, reduce reporting latency, and improve diagnostic consistency (Li et al., 2022b; Tanida et al., 2023; Yang et al., 2025).

The evolution of RRG methodologies reflects broader advances in multimodal representation learning. Early approaches (Zhang et al., 2020; Bu et al., 2024) adopt encoder-decoder architectures, leveraging convolutional neural networks for visual encoding and recurrent neural networks for sequential report generation. With the emergence of Transformer-based models (Li et al., 2022a; 2023; Wang et al., 2023; Liu et al., 2024a; Jin et al., 2024), attention mechanisms have become the dominant paradigm, enabling improved long-range dependency modeling and cross-modal alignment. Subsequent studies incorporate memory modules (Chen et al., 2020; 2021; Shen et al., 2024) and structured medical knowledge (Zhang et al., 2020; Wang et al., 2022; Yang et al., 2023), including knowledge graphs and domain-specific priors, to enhance semantic consistency between images and generated text. More recently, large language models (Liu et al., 2024a; Jin et al., 2024) have been introduced as powerful decoders, benefiting from prompt design and domain adaptation strategies to produce fluent and contextually rich reports.

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.



(a) Open-loop vs. consistency-aware decoding.

(b) Performance across report-length groups.

Figure 1: Illustration and analysis of consistency-aware decoding in RRG. (a) Comparison between conventional autoregressive decoding and LCC with periodic consistency verification and correction. (b) Performance on the IU X-Ray across ground-truth report-length groups, where LCC shows larger gains on longer reports.

Despite recent progress in RRG, a fundamental limitation remains in the autoregressive generation process, as illustrated in Figure 1. Existing approaches largely follow an open-loop decoding paradigm, where visual and multimodal evidence is primarily exploited to support subsequent token prediction, while the evolving report is generated without explicit consistency feedback. As shown in Figure 1a, once an inaccurate finding, imprecise anatomical description, or weakly grounded statement is introduced into the generated prefix, it becomes part of the subsequent decoding context and may progressively influence later predictions. This issue becomes more pronounced for longer reports, where the performance gap between LCC and competing methods increases substantially, as shown in Figure 1b. However, the partially generated report is rarely explicitly verified against image-grounded semantics, nor is there a dedicated mechanism to correct potential deviations once they occur. Consequently, local inconsistencies may accumulate and propagate through long-form generation, eventually leading to semantic deviations and clinically inaccurate descriptions.

In this paper, we propose Latent Consistency Control (LCC), a reason-before-generation framework that introduces a latent reasoning space and periodically regulates cross-modal consistency during report generation. First, a dual semantic condensation module compresses visual and textual token sequences into compact semantic representations to stabilize multimodal interaction. We then perform latent cross-modal semantic reasoning through iterative refinement within the shared latent space. At each reasoning step, visual and textual semantics interact to progressively update the latent states, resulting in a global reasoning representation that summarizes aligned multimodal information. To further maintain semantic consistency during generation, we introduce a multi-step decoding with consistency-aware control mechanism. During the decoding process, the partially generated report is periodically evaluated against both visual representations and the latent reasoning state. If semantic deviation is detected, a lightweight control signal is injected into the language model to adjust subsequent predictions and restore alignment with the visual evidence. Experiments on benchmark datasets demonstrate that our method outperforms state-of-the-art approaches across multiple metrics. In summary, the main contributions of this paper are as follows:

- We identify a critical limitation of existing RRG methods: during long-form autoregressive decoding, intermediate report states are rarely explicitly verified and regulated against image-grounded semantics, allowing local inconsistencies to propagate and gradually lead to semantic deviations.
- We propose Latent Consistency Control (LCC), which combines latent cross-modal semantic reasoning with multi-step consistency-aware decoding. LCC periodically evaluates the semantic consistency of the evolving report prefix and conditionally introduces corrective signals to regulate subsequent generation.
- Extensive experiments on benchmark datasets show that LCC improves both linguistic quality and clinical factual consistency, outperforming state-of-the-art methods across multiple NLG and clinical evaluation metrics.

2 Related Work

Cross-modal Alignment and Reasoning for Radiology Report Generation. Accurate image-text alignment is fundamental to RRG, as clinically meaningful descriptions require visual abnormalities to be associated with appropriate textual semantics. Existing studies (Huang et al., 2023; Jin et al., 2024; Huang et al., 2025) have therefore explored contrastive learning, structured medical knowledge, and fine-grained semantic alignment to bridge the modality gap. For example, DCL (Li et al., 2023) incorporates dynamically constructed medical knowledge together with image-report contrastive learning, whereas HSA (Zhu et al., 2024) performs hierarchical semantic alignment at both region and disease levels. Recent methods also exploit large language models (Liu et al., 2024a;b; Wang et al., 2025) and longitudinal multi-view information (Liu et al., 2025; Miao et al., 2026) to strengthen visual–textual representation learning and improve report generation. Despite these advances, most existing approaches focus primarily on representation-level cross-modal alignment, with limited attention to the evolving semantic state of the generated report during decoding. In contrast, our method constructs a compact latent reasoning space in which visual and textual semantics are progressively integrated before being used for consistency-aware generation.

Clinical Consistency and Reliability in Radiology Report Generation. Recent studies have explored report reliability through report refinement, localized abnormality assessment, and clinically grounded supervision. For instance, DART (Park et al., 2025) introduces a self-correcting realignment mechanism to refine an initial report according to the corresponding radiograph. RadSCR (Yan et al., 2026) further incorporates self-critiquing mechanisms to verify localized abnormality hypotheses before producing the complete report. More recently, SAT-RRG (Liu et al., 2026) identifies local semantic conflicts using an LLM referee and applies adaptive token-level optimization to emphasize clinically important errors, while BoxMed-RL (Jing et al., 2025) improves verifiability through spatially grounded reinforcement learning. These studies improve report reliability through report-level refinement, localized hypothesis assessment, or training-time optimization. By comparison, LCC targets the autoregressive decoding process itself by periodically evaluating intermediate report prefixes against image-grounded semantics and conditionally injecting corrective signals when inconsistency is detected.

3 Method

In this paper, we present Latent Consistency Control (LCC), a reason-before-generation framework. The architecture of LCC is as shown in Figure 2.

3.1 Semantic Representation

Textual Encoder. Given a ground-truth radiology report $Y^{\mathrm{gt}}=\{y_t\}_{t=1}^{T}$, we use the LLaMA backbone, shared with the report generator, to extract contextual textual representations under teacher forcing: $H^{\mathrm{gt}}=\{h_t^{\mathrm{gt}}\}_{t=1}^{T}\in\mathbb{R}^{T\times d_t}$, where d_t denotes the LLaMA hidden dimension. The resulting representations provide textual semantic supervision for cross-modal representation learning. At inference, the same backbone extracts textual states from the evolving generated prefix, eliminating the need for a reference report.

Visual Encoder. Given an input image I, we adopt a visual encoder to extract its visual representations. Specifically, the encoder produces a set of visual tokens $V=\{v_i\}_{i=1}^{N_v}\in\mathbb{R}^{N_v\times d_v}$ where N_v denotes the number of visual tokens and d_v is the feature dimension.

3.2 Dual Semantic Condensation

Although visual tokens V and textual representations H^{gt} encode rich local contextual information, they also contain modality-specific redundancy that may interfere with subsequent cross-modal interaction. We therefore introduce a Dual Semantic Condensation (DSC) module to summarize each modality into a compact global semantic representation, providing stable semantic anchors for latent reasoning.

We denote the condensed visual and textual representations as $z_v=\mathrm{Cond}_v(V)$ and $z_t^{\mathrm{gt}}=\mathrm{Cond}_t(H^{\mathrm{gt}})$, respectively. Both condensation operators are implemented using learn-

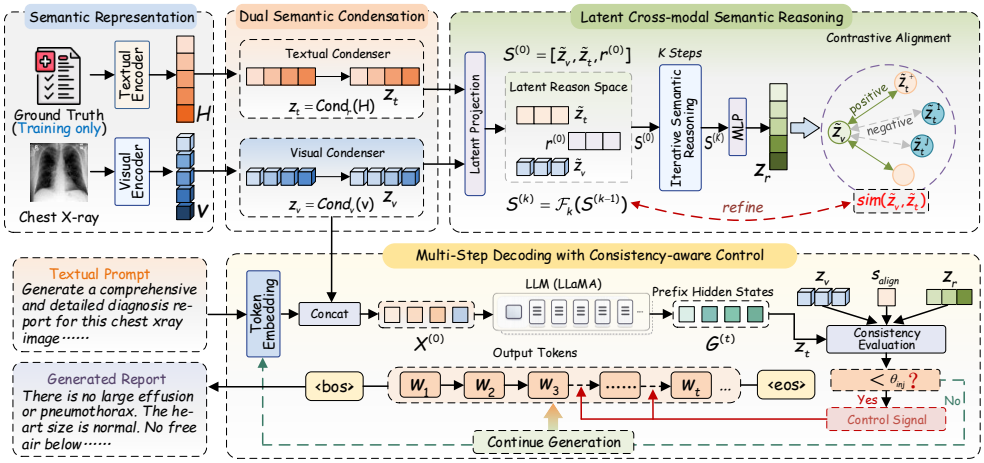


Figure 2: Architecture of LCC. Dual Semantic Condensation produces compact multimodal representations, Latent Cross-modal Semantic Reasoning integrates visual and textual semantics in a shared latent space, and Multi-Step Decoding with Consistency-aware Control regulates intermediate report generation through consistency evaluation and corrective control.

able attention-based aggregation, as follows:

$$z_v = \sum_{i=1}^{N_v} \alpha_i v_i, \qquad \alpha_i = \frac{\exp(w_v^\top v_i)}{\sum_{j=1}^{N_v} \exp(w_v^\top v_j)},$$
$$z_t^{\mathrm{gt}} = \sum_{t=1}^{T} \beta_t h_t^{\mathrm{gt}}, \qquad \beta_t = \frac{\exp(w_t^\top h_t^{\mathrm{gt}})}{\sum_{k=1}^{T} \exp(w_t^\top h_k^{\mathrm{gt}})}. \tag{1}$$

where α_i and β_t denote modality-specific attention weights, and w_v and w_t are learnable parameter vectors. The resulting representations provide compact summaries of the visual and textual semantics for subsequent cross-modal reasoning.

3.3 Latent Cross-modal Semantic Reasoning

Although DSC provides compact visual and textual representations, the two modalities remain encoded in modality-specific spaces. Directly combining them may therefore limit fine-grained semantic interaction. To address this issue, we introduce a Latent Cross-modal Semantic Reasoning (LCSR) module that performs iterative interaction in a shared latent space. A latent reasoning token is jointly updated with the visual and textual representations, allowing cross-modal evidence to be progressively integrated into a unified reasoning state.

Latent Space Construction. We first project the condensed visual and textual representations into a shared d-dimensional latent space and introduce a learnable reasoning token $r^{(0)} \in \mathbb{R}^d$ to aggregate cross-modal evidence:

$$\tilde{z}_v = W_v z_v, \qquad \tilde{z}_t^{\mathrm{gt}} = W_t z_t^{\mathrm{gt}}, \qquad S^{(0)} = [\tilde{z}_v, \tilde{z}_t^{\mathrm{gt}}, r^{(0)}], \tag{2}$$

where W_v and W_t are learnable projection matrices and $S^{(0)} \in \mathbb{R}^{3\times d}$ denotes the initial latent reasoning sequence. $\tilde{z}_v, \tilde{z}_t^{\mathrm{gt}} \in \mathbb{R}^d$ are projected into the same d-dimensional latent space.

Iterative Semantic Reasoning. We then perform iterative latent reasoning to progressively refine the shared latent states. At each reasoning step k, the latent sequence is updated through a Transformer-based reasoning operator $\mathcal{F}_k(\cdot)$:

$$S^{(k)} = \mathcal{F}_k\left(S^{(k-1)}\right), \quad \mathcal{F}_k(S) = \mathrm{FFN}(\mathrm{MHSA}(S) + S) + S, \tag{3}$$

where MHSA and FFN denote multi-head self-attention and a feed-forward network, respectively. After K reasoning steps, the resulting sequence is decomposed as $S^{(K)} = [z_v^{(K)}, z_t^{(K)}, r^{(K)}]$. The

reasoning state $r^{(K)}$ aggregates the cross-modal semantic evidence accumulated through iterative interaction, and the final reasoning representation is obtained as $z_r{=}\psi(r^{(K)})$ through a projection head $\psi(\cdot)$.

Contrastive Alignment. To further regularize the shared semantic space, we apply bidirectional contrastive alignment between the condensed visual and textual representations. The projected features are first L2-normalized, and their pairwise similarities within a mini-batch are computed as:

$$\hat{z}_v^i = \frac{\tilde{z}_v^i}{\|\tilde{z}_v^i\|_2}, \qquad \hat{z}_t^i = \frac{\tilde{z}_t^i}{\|\tilde{z}_t^i\|_2}, \qquad S_{ij} = \frac{(\hat{z}_v^i)^\top \hat{z}_t^j}{\tau}, \tag{4}$$

where τ is the temperature parameter.

Given a batch of B matched image-report pairs, the diagonal elements S_{ii} are treated as positives and the remaining pairs as in-batch negatives. The bidirectional contrastive losses are defined as:

$$\mathcal{L}_{v\to t} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(S_{ii})}{\sum_{j=1}^{B}\exp(S_{ij})}, \qquad \mathcal{L}_{t\to v} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(S_{ii})}{\sum_{j=1}^{B}\exp(S_{ji})}. \tag{5}$$

The symmetric alignment objective and pairwise semantic agreement score are then obtained as:

$$\mathcal{L}_{\text{align}} = \frac{1}{2}\left(\mathcal{L}_{v\to t} + \mathcal{L}_{t\to v}\right), \qquad s_{\text{align}} = \hat{z}_v^\top \hat{z}_t. \tag{6}$$

3.4 Multi-Step Decoding with Consistency-aware Control

Autoregressive report generation may gradually deviate from the visual evidence as previously generated tokens increasingly influence subsequent predictions. We therefore introduce a multi-step consistency-aware decoding mechanism that periodically evaluates the generated prefix and, when necessary, injects a corrective semantic signal into subsequent generation. Different from evaluating complete reports after generation, our consistency module operates on intermediate prefixes, allowing semantic deviations to be detected before they propagate to subsequent decoding steps.

Decoding Initialization. Given the visual representation z_v, we construct an image-conditioned prompt $X^{(0)}{=}[\mathbf{e}_{\text{bos}}, E_{\text{prompt}}(z_v)]$, using the predefined textual prompt template shown in Figure 2. At each decoding step t, the LLM (implemented as LLaMA-7B) produces hidden states $G^{(t)}{=}\text{LLM}(X^{(t)})$, from which the next-token distribution is obtained through the output projection and softmax. The predicted token is appended to the current sequence, and autoregressive decoding continues until an end-of-sequence token is produced or the maximum generation length is reached.

Consistency Evaluation. To maintain semantic consistency during generation, we periodically evaluate the generated prefix against the visual evidence and latent reasoning representation. To avoid unreliable estimation from extremely short prefixes, consistency evaluation starts after a warm-up period of T_{warm} decoding steps and is subsequently performed every T_{check} steps.

At the j-th consistency checkpoint, the hidden states of the generated prefix $G_{1:t_j}$ are first condensed into a prefix-level textual representation $z_t^{\text{pre},j}{=}\text{Cond}_t(G_{1:t_j})$ and subsequently mapped into the shared latent space, yielding $\tilde{z}_t^{\text{pre},j}$. The projected visual and prefix representations, $\tilde{z}_v$ and $\tilde{z}_t^{\text{pre},j}$, are then processed by LCSR to obtain the reasoning representation z_r^j. The consistency feature f_{cons}^j and score s_{cons}^j are computed as:

$$f_{\text{cons}}^j = [\tilde{z}_v \,\|\, \tilde{z}_t^{\text{pre},j} \,\|\, z_r^j \,\|\, s_{\text{align}}^j], \quad s_{\text{cons}}^j = \sigma\big(W_2\phi\big(W_1 f_{\text{cons}}^j\big)\big). \tag{7}$$

where $\|$ denotes vector concatenation, $\phi(\cdot)$ is the ReLU activation function, and $\sigma(\cdot)$ maps the output to $[0,1]$. A higher s_{cons}^j indicates stronger agreement between the current report prefix and the image-grounded latent representation, whereas a lower score indicates potential cross-modal inconsistency. Here, s_{align}^j is computed between the normalized $\tilde{z}_v$ and $\tilde{z}_t^{\text{pre},j}$ following Eq. (6).

Control Signal Injection. When $s_{\text{cons}}^j{<}\theta_{\text{inj}}$, the current prefix is regarded as deviating from the image-grounded semantic trajectory. We first construct a discrepancy-aware control feature:

$$u_{\text{ctrl}}^j = [\tilde{z}_v \,\|\, \tilde{z}_t^{\text{pre},j} \,\|\, z_r^j \,\|\, \Delta\tilde{z}_{vt}^j], \qquad \Delta\tilde{z}_{vt}^j = \tilde{z}_v - \tilde{z}_t^{\text{pre},j}. \tag{8}$$

The feature is then transformed into a gated corrective representation through a learnable control translator:

$$\mathbf{c}^j = f_\theta(\mathbf{u}^j_{\text{ctrl}}) \odot \sigma\Big(g_\theta(\mathbf{u}^j_{\text{ctrl}})\Big), \tag{9}$$

where f_θ and g_θ are learnable transformations and $\odot$ denotes element-wise multiplication. The adaptive gate controls the magnitude of semantic correction according to the current cross-modal discrepancy.

The resulting control representation is injected into the embedding of the latest generated token when the consistency criterion is violated:

$$\tilde{x}_{t_j} = e(w_{t_j}) + \mathbb{I}\left(s^j_{\text{cons}} < \theta_{\text{inj}}\right) c^j, \tag{10}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. The controlled representation updates the current decoding state, thereby guiding subsequent token generation toward image-grounded semantics.

3.5 Training Objective

The model is jointly optimized for report generation, cross-modal alignment, consistency discrimination, and corrective control. Given visual features v and ground-truth report tokens $\{y_t\}_{t=1}^T$, the autoregressive language modeling objective is:

$$\mathcal{L}_{\text{lm}} = -\sum_{t=1}^{T} \log p(y_t \mid y_{<t}, v)\,. \tag{11}$$

Consistency Loss. To enable the consistency network to distinguish aligned from inconsistent image-text states, we construct matched and mismatched prefix pairs during training. For each image, a prefix sampled from its paired report serves as a positive sample, whereas a prefix randomly drawn from another report in the same mini-batch is used as a negative sample while keeping the image fixed. The prefix length is randomly sampled during training to simulate intermediate report states encountered during autoregressive decoding. Let $y_{\text{cons}} \in \{0, 1\}$ denote the consistency label. The consistency objective is:

$$\mathcal{L}_{\text{cons}} = -y_{\text{cons}} \log(s_{\text{cons}}) - (1 - y_{\text{cons}}) \log(1 - s_{\text{cons}}). \tag{12}$$

Alignment Loss. We additionally employ the contrastive objective $\mathcal{L}_{\text{align}}$ defined in Section 3.3 to preserve cross-modal semantic alignment between paired visual and textual representations.

Corrective Control Loss. To train the corrective controller, an inconsistent prefix $\tilde{Y}_{1:t_j}$ is constructed at checkpoint t_j, and the corresponding control representation c^j is injected into the prefix state. The control objective combines ground-truth continuation supervision with magnitude regularization:

$$\mathcal{L}_{\text{ctrl}} = -\sum_{t=t_j+1}^{T} \log p\Big(y_t \mid \tilde{Y}_{1:t_j}, v, c^j\Big) + \gamma \frac{1}{n_{\text{ctrl}} D} \|\mathbf{C}\|_F^2. \tag{13}$$

where $\mathbf{C} \in \mathbb{R}^{n_{\text{ctrl}} \times D}$ contains the control embeddings and γ controls the regularization strength.

Overall Objective. The complete training objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{lm}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{ctrl}}\mathcal{L}_{\text{ctrl}}, \tag{14}$$

where λ_{cons}, λ_{align}, and λ_{ctrl} balance the contributions of the auxiliary training objectives.

4 Experiments and Analysis

4.1 Experiment Settings

Datasets. **IU X-Ray** (Shin et al., 2016) is a publicly available benchmark dataset for RRG. It contains 7,470 chest X-ray images paired with 3,955 reports. **MIMIC-CXR** (Johnson et al., 2019)

Table 1: Comparison with baselines on the IU X-Ray and MIMIC-CXR datasets. B-1/B-2/B-3/B-4 denote BLEU-1/2/3/4, MTR denotes METEOR, RG-L denotes ROUGE-L, P/R/F1 denote Precision, Recall, and F1 score, and RG-F1 denotes RadGraph F1. The best results are denoted in bold. ‡ indicates results reproduced using the official implementations; all other baseline results are taken from the corresponding original papers.

Datasets	Models	NLG Metrics ↑						CE Metrics ↑			RG-F1 ↑
		B-1	B-2	B-3	B-4	MTR	RG-L	P	R	F1	
IU X-Ray	R2GenCMN ‡ [ACL 21']	0.486	0.319	0.232	0.179	0.199	0.380	-	-	-	-
	DCL ‡[CVPR 23']	0.473	0.310	0.223	0.159	0.187	0.371				
	R2-LLM [AAAI 24']	0.499	0.323	0.238	0.184	0.208	0.390	-	-	-	-
	MAN [AAAI 24']	0.501	0.328	0.230	0.107	0.213	0.386	-	-	-	-
	Med-LMM [ACM MM 24']	-	-	-	0.168	0.209	0.381	-	-	-	-
	DCG [ACM MM 24']	0.514	0.330	0.241	0.186	0.211	0.401	-	-	-	-
	DART [CVPR 25']	0.486	0.348	0.265	0.208	0.205	0.411	-	-	-	-
	DADS [AAAI 26']	0.495	0.328	0.242	0.187	-	0.384	-	-	-	-
	SAT-RRG [CVPR 26']	0.504	0.342	0.254	0.196	0.222	0.400	-	-	-	-
	LCC(Ours)	**0.521**	**0.368**	**0.272**	**0.213**	**0.223**	**0.437**	-	-	-	-
MIMIC-CXR	R2GenCMN ‡ [ACL 21']	0.349	0.207	0.136	0.101	0.131	0.268	0.321	0.262	0.289	-
	DCL ‡[CVPR 23']	0.355	0.221	0.142	0.101	0.137	0.274	0.469	0.341	0.395	-
	R2-LLM [AAAI 24']	0.402	0.262	0.180	0.128	0.175	0.291	0.465	0.482	0.473	-
	MAN [AAAI 24']	0.396	0.244	0.162	0.115	0.151	0.274	0.411	0.398	0.389	-
	Med-LMM [ACM MM 24']	-	-	-	0.128	0.161	0.289	0.412	0.373	0.395	-
	DCG [ACM MM 24']	0.397	0.258	0.166	0.126	0.162	0.295	0.441	0.414	0.404	-
	MLRG [CVPR 25']	0.411	0.277	0.204	0.158	0.176	0.320	**0.549**	0.468	0.505	0.291
	DART [CVPR 25']	**0.437**	0.279	0.191	0.137	0.175	0.310	0.520	**0.546**	0.533	-
	DADS [AAAI 26']	0.428	0.272	0.187	0.131	-	0.291	-	-	-	-
	SAT-RRG [CVPR 26']	0.428	0.280	0.196	0.143	0.167	0.303	-	-	-	0.205
	LCC(Ours)	0.432	**0.285**	**0.213**	**0.171**	**0.188**	**0.337**	0.545	0.533	**0.539**	**0.295**

is a large-scale publicly available chest X-ray dataset widely used for radiology report generation. It comprises 377,110 chest X-ray images paired with 227,835 corresponding radiology reports.

Metrics. We evaluate the generated reports using two categories of metrics: natural language generation (NLG) and clinical efficacy (CE) metrics. For NLG evaluation, we report BLEU (Papineni et al., 2002), METEOR (Denkowski & Lavie, 2011), ROUGE-L (Lin, 2004) to measure n-gram overlap and semantic similarity between generated and reference reports. To assess clinical correctness, we adopt CheXpert (Irvin et al., 2019) to extract 14 medical observations from both generated and ground-truth reports. Precision, recall, and F1-score are computed based on the predicted labels to quantify factual consistency at the clinical level. Additionally, we report RadGraph F1 (RG-F1) (Jain et al., 2021) to evaluate fine-grained clinical factual correctness.

Implementation Details. Following prior work (Chen et al., 2021), the IU X-Ray dataset is split into training, validation, and test sets with a 7:1:2 ratio, while MIMIC-CXR follows its official data split. We use the Swin Transformer as the visual encoder and LLaMA-7B as the report generator. The latent reasoning steps K and the consistency checking frequency T_{check} are set to 3 and 16, respectively. We set T_{warm}=32 to ensure sufficient context before consistency evaluation, and θ_{inj}=0.6 to trigger corrective control. Additional implementation details, computational efficiency, and hyperparameter analyses are provided in Appendices A, B.1, and B.2.

4.2 Main Results

To assess the efficacy of the proposed LCC, we conduct extensive comparisons with a range of state-of-the-art radiology report generation approaches on both benchmark datasets, including **R2GenCMN** (Chen et al., 2021), **DCL** (Li et al., 2023), **R2-LLM** (Liu et al., 2024a), **MAN** (Shen et al., 2024), **Med-LMM** (Liu et al., 2024b), **DCG** (Liang et al., 2024), **MLRG** (Liu et al., 2025), **DART** (Park et al., 2025), **DADS** (Wu et al., 2026), and **SAT-RRG** (Liu et al., 2026) .

As shown in Table 1, LCC achieves strong performance across both datasets, obtaining the best results on BLEU-2, BLEU-3, BLEU-4, METEOR, and ROUGE-L. On MIMIC-CXR, LCC improves

Table 2: Ablation results of LCC on IU X-Ray and MIMIC-CXR. Performance is evaluated using NLG metrics, CE metrics, and RadGraph F1 (RG-F1). The best results are shown in bold.

Datasets	Models	NLG Metrics ↑						CE Metrics ↑			RG-F1 ↑
		B-1	B-2	B-3	B-4	MTR	RG-L	P	R	F1	
IU X-Ray	w/o DSC	0.519	0.350	0.261	0.192	0.186	0.416	-	-	-	-
	w/o LCSR	0.507	0.327	0.246	0.182	0.187	0.382	-	-	-	-
	w/o MDCC	0.496	0.321	0.224	0.176	0.185	0.374	-	-	-	-
	LCC(Ours)	**0.521**	**0.368**	**0.272**	**0.213**	**0.223**	**0.437**	-	-	-	-
MIMIC-CXR	w/o DSC	0.417	0.261	0.187	0.156	0.177	0.312	0.502	0.495	0.498	0.261
	w/o LCSR	0.402	0.255	0.180	0.139	0.162	0.306	0.491	0.482	0.487	0.252
	w/o MDCC	0.395	0.248	0.172	0.131	0.155	0.298	0.478	0.469	0.474	0.245
	LCC(Ours)	**0.432**	**0.285**	**0.213**	**0.171**	**0.188**	**0.337**	**0.545**	**0.533**	**0.539**	**0.295**

Table 3: Ablation analysis of MDCC.

Method	B-4	MTR	RG-L
w/o MDCC	0.176	0.185	0.374
w/o Verification	0.192	0.203	0.402
w/o Correction	0.201	0.206	0.411
LCC	**0.213**	**0.223**	**0.437**

Table 4: Ablation analysis of training objectives.

Method	B-4	MTR	RG-L
w/o $\mathcal{L}_{\text{align}}$	0.184	0.195	0.403
w/o $\mathcal{L}_{\text{cons}}$	0.206	0.217	0.416
w/o $\mathcal{L}_{\text{ctrl}}$	0.198	0.209	0.396
LCC	**0.213**	**0.223**	**0.437**

BLEU-4 from 0.158 to 0.171 and ROUGE-L from 0.320 to 0.337 over the strongest previously reported results, while on IU X-Ray it improves ROUGE-L from 0.411 to 0.437. LCC also achieves competitive clinical precision, recall, and F1 scores on MIMIC-CXR, indicating that the gains in linguistic quality are accompanied by reliable clinical factual consistency. Unlike existing approaches that mainly improve representation learning or sequence generation, LCC explicitly regulates intermediate decoding states through latent cross-modal reasoning and consistency-aware control. By periodically evaluating generated prefixes and correcting potential deviations, LCC maintains stronger alignment between visual evidence and generated reports.

4.3 Ablation Study

To quantitatively evaluate the contribution of each component, we conduct systematic ablation studies by removing key modules or objectives from the full LCC model. Specifically, *w/o DSC* replaces the learnable Dual Semantic Condensation (DSC) module with parameter-free mean pooling for both visual and textual representations, while keeping the remaining architecture unchanged. *w/o LCSR* discards the Latent Cross-modal Semantic Reasoning (LCSR) module, preventing latent semantic interaction before generation. *w/o MDCC* removes the Multi-step Decoding with Consistency-aware Control (MDCC) mechanism, where report generation follows the standard autoregressive paradigm without intermediate consistency evaluation and corrective signals.

The results are summarized in Table 2. Removing any major component consistently degrades performance, confirming the complementary roles of semantic condensation, latent reasoning, and consistency-aware decoding. Without DSC, BLEU-4 drops from 0.213 to 0.192 on IU X-Ray, indicating that compact semantic representations provide more stable anchors for subsequent cross-modal reasoning. Removing LCSR further reduces ROUGE-L from 0.437 to 0.382 on IU X-Ray and RG-F1 from 0.295 to 0.252 on MIMIC-CXR, demonstrating the importance of latent cross-modal interaction. The largest degradation occurs without MDCC, where BLEU-4 decreases to 0.176 on IU X-Ray and RG-F1 to 0.245 on MIMIC-CXR. This confirms that intermediate consistency verification and corrective control are critical for mitigating semantic drift during report generation.

Analysis of Consistency-aware Decoding on IU X-Ray. To further analyze the two key operations in MDCC, we separately remove verification and correction. Specifically, w/o Verification disables consistency-based triggering and applies corrective control at each scheduled checkpoint, whereas w/o Correction retains consistency evaluation but removes control-signal injection. As shown in Table 3, removing either component degrades performance, demonstrating the importance of both adaptive inconsistency detection and corrective regulation. Removing verification causes a larger drop, indicating that indiscriminate intervention is less effective than selective control, while remov-

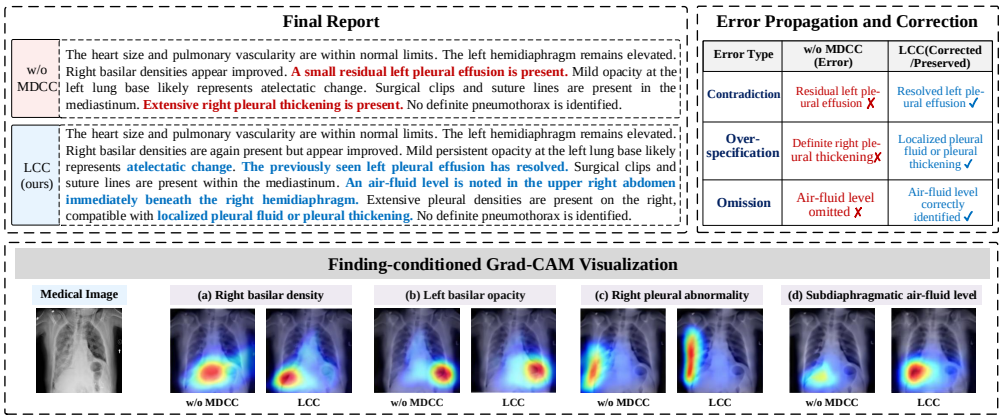


Figure 3: Case study illustrating error propagation and correction during RRG.

ing correction still underperforms the full model, showing that detection alone is insufficient. The complete MDCC achieves the best performance by jointly determining when intervention is needed and applying corrective control accordingly. Further analyses are provided in Appendix B.3.

Analysis of Training Objectives on IU X-Ray. We further investigate the contribution of different training objectives in Table 4. Removing $\mathcal{L}_{align}$ causes a noticeable decline in BLEU-4, indicating that explicit contrastive alignment facilitates the learning of image-report semantic correspondence. Removing $\mathcal{L}_{cons}$ reduces METEOR and ROUGE-L, demonstrating that consistency supervision helps maintain semantic agreement between generated prefixes and image-grounded representations. Removing $\mathcal{L}_{ctrl}$ also leads to performance degradation, demonstrating the importance of explicitly supervising corrective control while constraining the magnitude of corrective representations. Combining all objectives achieves the best results, confirming that these objectives provide complementary supervision for accurate report generation.

4.4 Case Study

Figure 3 illustrates the effect of consistency-aware decoding on a representative case. Without MDCC, the generated report exhibits three clinically relevant errors: a resolved left pleural effusion is described as residual, the right pleural abnormality is over-specified as definite pleural thickening, and the subdiaphragmatic air-fluid level is omitted. LCC corrects these deviations by correctly identifying the pleural effusion as resolved, recovering the omitted air-fluid level, and preserving the uncertainty of the right pleural finding. We further employ Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) to visualize the image regions associated with different clinical findings. The resulting maps exhibit more concentrated responses around relevant anatomical regions, providing qualitative support for improved visual grounding. Additional decoding trajectories and cross-variant comparisons are provided in Appendix C.

5 Conclusion

In this paper, we propose Latent Consistency Control (LCC), a reason-before-generation framework for RRG. LCC addresses the challenge of maintaining cross-modal consistency during long-form autoregressive generation. To mitigate these issues, we introduce dual semantic condensation and latent cross-modal reasoning to establish consistent multimodal semantics before generation, and develop a multi-step decoding with consistency-aware control mechanism to regulate the generation process and correct potential semantic deviations. Extensive experiments on benchmark datasets demonstrate that LCC consistently improves both linguistic quality and clinical factual correctness, achieving superior performance on key NLG and clinical metrics. Ablation studies further confirm that each component contributes complementary gains, with consistency-aware decoding playing a critical role in maintaining stable and image-grounded generation.

AI USE DISCLOSURE

Generative AI tools were used for language polishing and the editing and formatting of tables. All AI-assisted content was reviewed and verified by the authors.

REFERENCES

Shenshen Bu, Taiji Li, Yuedong Yang, and Zhiming Dai. Instance-level expert knowledge and aggregate discriminative attention for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14194–14204, 2024.

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

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

Michael Denkowski and Alon Lavie. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the sixth workshop on statistical machine translation*, pp. 85–91, 2011.

Xiaodi Hou, Xiaobo Li, Zhi Liu, Shengtian Sang, Mingyu Lu, and Yijia Zhang. Recalibrated cross-modal alignment network for radiology report generation with weakly supervised contrastive learning. *Expert Systems with Applications*, pp. 126394, 2025a.

Xiaodi Hou, Xiaobo Li, Simiao Wang, Mingyu Lu, Hongfei Lin, and Yijia Zhang. Mambagen: efficient visual representation learning for automatic radiology report generation. *Expert Systems with Applications*, pp. 129537, 2025b.

Xiyang Huang, Yingjie Han, Runzhi Li, Pengcheng Wu, Kunli Zhang, et al. Cmeaa: Cross-modal enhancement and alignment adapter for radiology report generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 8546–8556, 2025.

Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. Kiut: Knowledge-injected u-transformer for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19809–19818, 2023.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 590–597, 2019.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Q. H. Truong, Du Nguyen Duong, Tan Bui, Pierre J. Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021.

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

Baoyu Jing, Zeya Wang, and Eric Xing. Show, describe and conclude: On exploiting the structure information of chest x-ray reports. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6570–6580. Association for Computational Linguistics, 2019.

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. *Medical Image Analysis*, pp. 103910, 2025.

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

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022a.

Mingjie Li, Wenjia Cai, Karin Verspoor, Shirui Pan, Xiaodan Liang, and Xiaojun Chang. Cross-modal clinical graph transformer for ophthalmic report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20656–20665, 2022b.

Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3334–3343, 2023.

Xiao Liang, Yanlei Zhang, Di Wang, Haodi Zhong, Ronghan Li, and Quan Wang. Divide and conquer: Isolating normal-abnormal attributes in knowledge graph-enhanced radiology report generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 4967–4975, 2024.

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

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

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10348–10359, 2025.

Rui Liu, Mingjie Li, Shen Zhao, Ling Chen, Xiaojun Chang, and Lina Yao. In-context learning for zero-shot medical report generation. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 8721–8730, 2024b.

Yunyi Liu, Yingshu Li, Tong Chen, Lingqiao Liu, Lei Wang, and Luping Zhou. Sat-rrg: Llm-guided self-adaptive training for radiology report generation with token-level push-pull optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 35363–35372, June 2026.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Qiguang Miao, Kang Liu, Zhuoqi Ma, Yunan Li, Xiaolu Kang, Ruixuan Liu, Tianyi Liu, and Kun Xie. Patient-specific multimodal learning with multi-view contrastive alignment for chest x-ray report generation. *Bioinformatics*, 42(8):btag566, 2026.

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

Sang-Jun Park, Keun-Soo Heo, Dong-Hee Shin, Young-Han Son, Ji-Hye Oh, and Tae-Eui Kam. Dart: Disease-aware image-text alignment and self-correcting re-alignment for trustworthy radiology report generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15580–15589, 2025.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Hongyu Shen, Mingtao Pei, Juncai Liu, and Zhaoxing Tian. Automatic radiology reports generation via memory alignment network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 5, pp. 4776–4783, 2024.

Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M Summers. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2497–2506, 2016.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7433–7442, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Song Wang, Liyan Tang, Mingquan Lin, George Shih, Ying Ding, and Yifan Peng. Prior knowledge enhances radiology report generation. In *AMIA Annual Symposium Proceedings*, pp. 486. American Medical Informatics Association, 2022.

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

Zhuhao Wang, Yihua Sun, Zihan Li, Xuan Yang, Fang Chen, and Hongen Liao. Llm-rg4: Flexible and factual radiology report generation across diverse input contexts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 8250–8258, 2025.

Puzhen Wu, Hexin Dong, Yi Lin, Yihao Ding, and Yifan Peng. A disease-aware dual-stage framework for chest x-ray report generation. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 33953–33961, 2026.

Bin Yan and Mingtao Pei. Clinical-BERT: Vision-language pre-training for radiograph diagnosis and reports generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2982–2990, 2022.

Sixing Yan, Ziao Wang, Kejing Yin, William Kwok-wai Cheung, Ka Chun Cheung, and Simon See. Learning self-critiquing mechanisms for region-guided chest x-ray report generation. In *International Conference on Learning Representations*, volume 2026, pp. 78368–78393, 2026.

Longzhen Yang, Zhangkai Ni, Ying Wen, Yihang Liu, Lianghua He, and Heng Tao Shen. Self-supervised anatomical consistency learning for vision-grounded medical report generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2958–2967, 2025.

Shuxin Yang, Xian Wu, Shen Ge, Zhuozhao Zheng, S Kevin Zhou, and Li Xiao. Radiology report generation with a learned knowledge base and multi-modal alignment. *Medical Image Analysis*, 86:102798, 2023.

Xiulong Yi, You Fu, Ruiqing Liu, Hao Zhang, and Rong Hua. Tsget: Two-stage global enhanced transformer for automatic radiology report generation. *IEEE Journal of Biomedical and Health Informatics*, 28(4):2152–2162, 2024.

Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. When radiology report generation meets knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 12910–12917, 2020.

Zhihong Zhu, Xuxin Cheng, Yunyan Zhang, Zhaorun Chen, Qingqing Long, Hongxiang Li, Zhiqi Huang, Xian Wu, and Yefeng Zheng. Multivariate cooperative game for image-report pairs: hierarchical semantic alignment for medical report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 303–313. Springer, 2024.

APPENDIX

A EXPERIMENT SETTINGS

A.1 DATASETS

We evaluate our method on two widely used benchmark datasets for radiology report generation: IU X-Ray and MIMIC-CXR.

IU X-Ray (Shin et al., 2016) contains 7,470 chest X-ray images paired with 3,955 radiology reports. Following prior RRG studies (Chen et al., 2021), we adopt the commonly used 7:1:2 split, with 5,226, 748, and 1,496 images in the training, validation, and test sets, corresponding to 2,770, 395, and 790 reports, respectively.

MIMIC-CXR (Johnson et al., 2019) is a large chest X-ray dataset constructed from clinical records in the MIMIC database, including a much broader collection of studies and disease patterns. Compared with IU X-Ray, it covers more diverse imaging appearances and report descriptions, which increases the difficulty of cross-modal modeling. We adopt the official data split, with 368,960, 2,991, and 5,159 images and 222,758, 1,808, and 3,269 reports in the training, validation, and test sets, respectively. The detailed statistics of the datasets are provided in Table S1.

Table S1: Dataset statistics of IU X-Ray and MIMIC-CXR across training, validation, and test splits.

	IU X-Ray			MIMIC-CXR		
Metric	Train	Val	Test	Train	Val	Test
#Images	5,226	748	1,496	368,960	2,991	5,159
#Reports	2,770	395	790	222,758	1,808	3,269
#Patients	2,770	395	790	64,586	500	293

A.2 METRICS

We evaluate the generated reports using both natural language generation (NLG) metrics and clinical evaluation (CE) metrics.

For linguistic quality, we adopt standard NLG metrics, including BLEU(1–4) (Papineni et al., 2002), METEOR (Denkowski & Lavie, 2011), and ROUGE-L (Lin, 2004). BLEU measures n-gram precision between generated and reference reports, emphasizing exact lexical matching across multiple n-gram orders. METEOR incorporates both precision and recall through explicit alignment strategies, such as stemming and synonym matching, improving sensitivity to semantic similarity. ROUGE-L evaluates the longest common subsequence, capturing sentence-level structural consistency and preserving word order.

For clinical evaluation, we employ CheXpert (Irvin et al., 2019) to extract 14 predefined clinical observations from both generated and reference reports. The extracted labels are aligned between generated and ground-truth reports, and Precision, Recall, and F1-score are computed to quantify clinical factual consistency at the observation level. We further adopt RadGraph F1 (RG-F1) (Jain et al., 2021), which measures the overlap of clinical entities and their relations between generated and reference reports. By jointly evaluating entity recognition and relation extraction, RG-F1 provides a fine-grained assessment of clinical correctness and reflects the coherence of clinically meaningful findings.

A.3 IMPLEMENTATION DETAILS

Following prior work (Chen et al., 2021), IU X-Ray is split into training, validation, and test sets with a 7:1:2 ratio, while MIMIC-CXR follows its official data split. We adopt the Swin Transformer (Liu et al., 2021) as the visual encoder and LLaMA-7B (Touvron et al., 2023) as the report generator. In the Latent Cross-modal Semantic Reasoning (LCSR) module, the number of reasoning steps is set to $K=3$. For the consistency-aware decoding mechanism, the consistency checking interval is

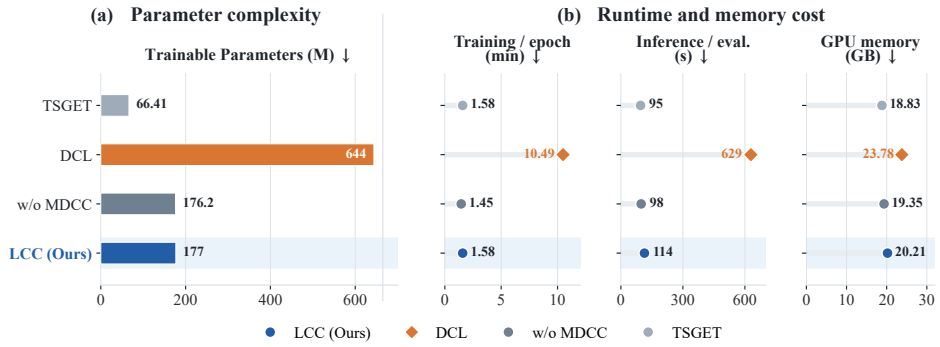


Figure S1: Computational efficiency comparison of different RRG methods. (a) Trainable parameter count. (b) Training time, inference latency, and GPU memory consumption.

set to T_{check}=16, and a warm-up period of T_{warm}=32 is adopted to ensure stable prefix representations before consistency evaluation. The loss weights λ_{cons}, λ_{align}, and λ_{ctrl} are set to 0.2, 0.3, and 1.0, respectively. The magnitude regularization coefficient γ in $\mathcal{L}_{\text{ctrl}}$ is set to 1×10^{-4}. The control injection threshold is set to θ_{inj}=0.6. For cross-modal alignment, we employ a bidirectional contrastive learning objective with a temperature parameter τ=0.07. All models are trained using the AdamW optimizer with an initial learning rate of 5×10^{-5}. For qualitative visualization, we employ Gradient-weighted Class Activation Mapping (Grad-CAM) on the final spatial features of the Swin encoder, using the generation score of each clinical finding as the target for computing the response map. All experiments are conducted on two NVIDIA GeForce RTX 4090 GPUs. We train the model for 50 epochs on IU X-Ray and 30 epochs on MIMIC-CXR, using a batch size of 16 for both datasets. The code is available at https://anonymous.4open.science/r/LCC-DD1F/.

B Experiments and Analysis

B.1 Computational Efficiency Analysis

Since LCC introduces latent cross-modal reasoning and consistency-aware decoding during generation, we further evaluate its computational efficiency to quantify the additional overhead introduced by these components. For a fair comparison, **TSGET** (Yi et al., 2024) and **DCL** (Li et al., 2023) are reproduced using their official implementations and evaluated under the same experimental setting as LCC.

As shown in Figure S1, LCC contains 177M trainable parameters, which is only slightly higher than the 176.2M parameters of the variant without MDCC. This small increase indicates that the proposed consistency-aware control mechanism introduces negligible additional parameter cost. Compared with DCL, which contains 644M trainable parameters, LCC requires substantially fewer trainable parameters, although TSGET remains more lightweight with 66.41M parameters. In terms of runtime and memory consumption, LCC requires 1.58 minutes per training epoch, 114 seconds for inference, and 20.21 GB of GPU memory. Compared with the variant without MDCC, the additional consistency verification and corrective control increase the inference time from 98 to 114 seconds and GPU memory from 19.35 to 20.21 GB, while the training time increases only slightly from 1.45 to 1.58 minutes per epoch. Moreover, LCC remains substantially more efficient than DCL in both runtime and memory consumption. These results indicate that LCC introduces only moderate computational overhead while enabling consistency-aware regulation during autoregressive report generation.

B.2 Hyperparameter Study

We investigate the effects of four key hyperparameters, including the latent reasoning steps K, consistency checking interval T_{check}, warm-up period T_{warm}, and control injection threshold θ_{inj}. The results are shown in Figure S2.

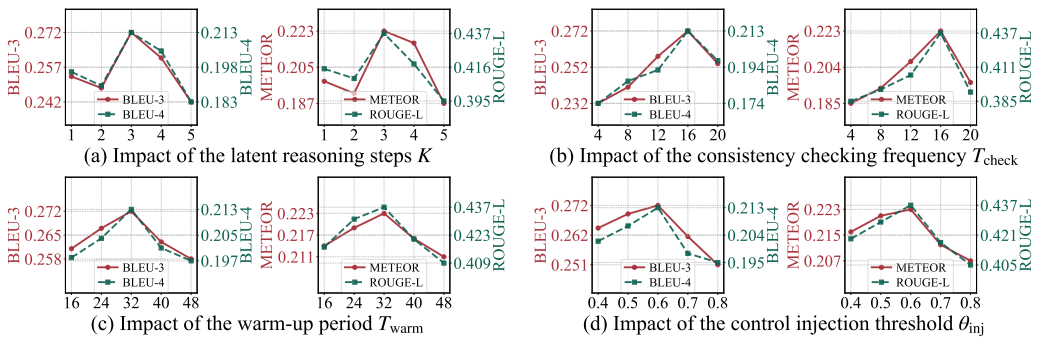


Figure S2: Hyperparameter analysis on IU X-Ray. (a) Impact of the latent reasoning steps K. (b) Impact of the consistency checking interval T_{check}. (c) Impact of the warm-up period T_{warm}. (d) Impact of the control injection threshold θ_{inj}.

Effect of Latent Reasoning Steps K. The parameter K controls the number of iterative reasoning steps in the LCSR module. As shown in Figure S2(a), performance improves as K increases from 1 to 3, indicating that iterative reasoning facilitates progressive interaction between visual and textual semantics in the shared latent space. The best performance is achieved at K=3. Further increasing K to 4 or 5 leads to a moderate decline, suggesting that excessive reasoning may introduce redundant feature updates without providing additional semantic benefits. Therefore, we set K=3 in subsequent experiments.

Effect of Consistency Checking Interval T_{check}. The parameter T_{check} determines the interval between consecutive consistency evaluations during autoregressive decoding. As shown in Figure S2(b), performance gradually improves as the interval increases from 4 to 16 steps and reaches its optimum at T_{check}=16. A small interval results in overly frequent consistency evaluations and corrective interventions, which may interfere with stable generation. In contrast, when the interval is increased to 20, performance decreases because semantic deviations may not be identified sufficiently early. These results indicate that T_{check}=16 provides an effective balance between consistency regulation and decoding stability.

Effect of Warm-up Period T_{warm}. The warm-up period determines when consistency-aware regulation is activated during decoding. As illustrated in Figure S2(c), performance consistently improves when T_{warm} increases from 16 to 32 and reaches the best results at T_{warm}=32. Activating consistency evaluation too early may rely on incomplete prefix semantics, resulting in less reliable consistency estimation. Conversely, delaying the activation to 40 or 48 steps reduces the opportunity to correct potential semantic deviations in a timely manner. Therefore, a moderate warm-up period of T_{warm}=32 provides sufficient semantic context while retaining effective corrective intervention.

Effect of Control Injection Threshold θ_{inj}. The threshold θ_{inj} determines whether corrective control is activated according to the estimated consistency score. As shown in Figure S2(d), performance improves as θ_{inj} increases from 0.4 to 0.6 and achieves the best results at θ_{inj}=0.6. A smaller threshold makes the control mechanism relatively conservative, allowing some semantic deviations to remain uncorrected. In contrast, a larger threshold, such as 0.7 or 0.8, increases the frequency of corrective interventions and may introduce unnecessary perturbations into the decoding process. Overall, θ_{inj}=0.6 achieves a favorable balance between effective correction and generation stability.

B.3 Mechanism Analysis of Consistency-Aware Decoding

We further investigate how the proposed consistency-aware decoding mechanism operates during report generation from three complementary perspectives: consistency discrimination, robustness to report length, and dynamic control behavior, as illustrated in Figure S3.

Consistency discrimination. Figure S3(a) evaluates whether the learned consistency evaluator can identify clinically meaningful semantic deviations. Starting from a clinically coherent report prefix, we construct controlled variants by independently reversing a negation, substituting a clinical finding, or modifying its uncertainty expression. The original and perturbed prefixes are evaluated

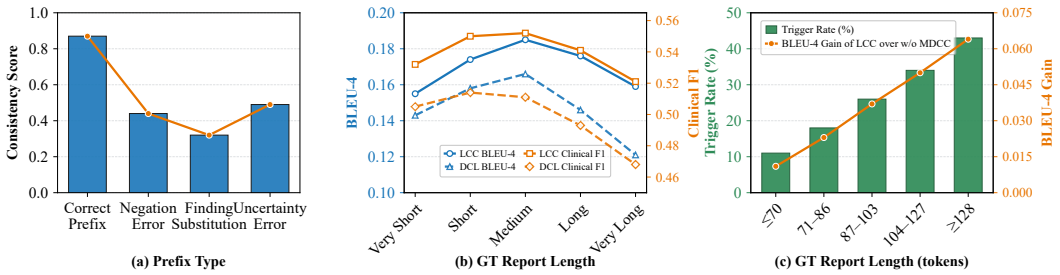


Figure S3: Mechanism analysis of consistency-aware decoding on MIMIC-CXR. (a) Responses of the consistency evaluator to correct and clinically perturbed report prefixes. (b) Length-wise comparison of LCC and DCL in terms of BLEU-4 and clinical F1. (c) MDCC trigger rate and the BLEU-4 improvement of LCC over w/o MDCC across report-length groups. Trigger rate is computed as the proportion of consistency checkpoints at which $s_{\text{cons}} < \theta_{\text{inj}}$.

against the same image-conditioned reasoning representation, thereby isolating the effect of textual semantic corruption. The evaluator produces a substantially stronger response for the coherent prefix than for the perturbed variants. This behavior indicates that the consistency signal is sensitive to clinically relevant discrepancies between the evolving report and the latent visual-semantic evidence, providing a meaningful criterion for determining when corrective control should be activated.

Robustness to report length. Figure S3(b) examines whether the benefit of LCC persists as generation becomes longer. We divide the test reports into five groups according to their ground-truth report lengths and compare LCC with DCL using both BLEU-4 and clinical F1. Both methods become increasingly challenged as the report length grows, reflecting the greater difficulty of maintaining linguistic and clinical consistency over extended autoregressive trajectories. However, LCC consistently maintains stronger performance than DCL, with the difference becoming more evident for longer reports. This observation supports our motivation that semantic deviation can accumulate during prolonged decoding, while latent consistency control helps preserve alignment between generated content and underlying clinical evidence.

Dynamic triggering behavior. To further determine whether MDCC intervenes where consistency control is most needed, Figure S3(c) analyzes its trigger frequency across report-length groups together with the performance gain of LCC over w/o MDCC. The control mechanism is activated more frequently as the report becomes longer, while the advantage over the variant without MDCC follows the same overall trend. This correspondence suggests that the learned consistency criterion does not invoke correction uniformly throughout decoding. Instead, MDCC increasingly intervenes when longer generation trajectories create a higher risk of semantic drift. Together with the results in Figure S3(a) and (b), these observations provide mechanism-level evidence that LCC can detect emerging inconsistency and adaptively regulate generation rather than relying solely on a fixed visual representation established before decoding.

C Qualitative Analysis of Consistency-Aware Decoding

To complement the quantitative evaluation, we provide two qualitative analyses from complementary perspectives. The first traces the evolution of generated prefixes and consistency scores for the representative example in the main text, illustrating when semantic deviations emerge and how corrective control intervenes. The second compares different model variants on an independent case and examines their spatial response patterns, highlighting the respective roles of latent cross-modal reasoning and consistency-aware decoding.

C.1 Generation Trajectory and Consistency Evolution

Figure S4 provides a detailed view of the decoding trajectory for the case analyzed in the main text. Without MDCC, the generation initially remains consistent with the image evidence, but an inaccurate statement regarding residual pleural effusion is introduced at an intermediate checkpoint. This local inconsistency is subsequently incorporated into the autoregressive context, followed by addi-

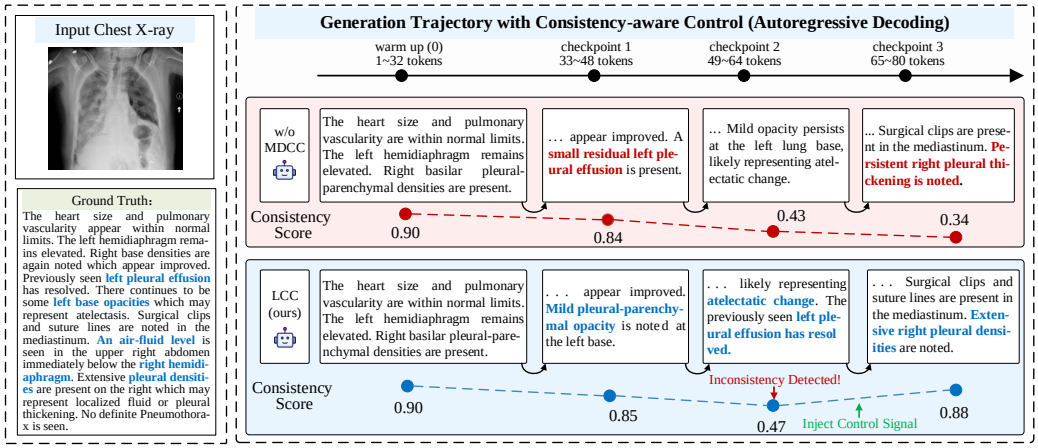


Figure S4: Generation trajectory and consistency-score evolution for a representative case. Without MDCC, intermediate inconsistencies propagate across subsequent decoding steps. LCC detects a low-consistency prefix and injects a corrective signal, leading to improved semantic consistency at the following checkpoint. For w/o MDCC, consistency scores are computed post hoc using the trained consistency evaluator for visualization only and do not participate in generation.

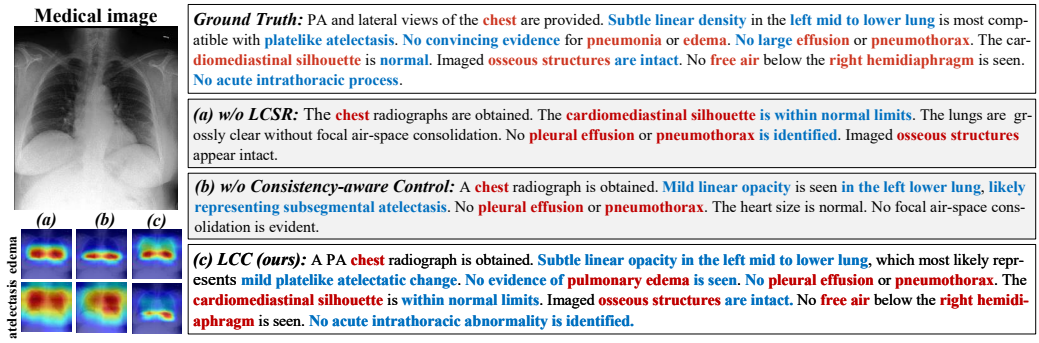


Figure S5: Case study comparing report generation and finding-conditioned Grad-CAM visualizations across different model variants. Red text denotes lesion-related phrases, while blue text indicates the corresponding abnormality descriptions identified in the reports.

tional over-specified descriptions of right pleural abnormalities. Correspondingly, the consistency score decreases from 0.90 to 0.34 over successive checkpoints.

In contrast, LCC identifies a low-consistency prefix at the second checkpoint, where the consistency score falls below the control threshold. The resulting control signal regulates subsequent decoding rather than modifying previously generated tokens. After intervention, the following generation returns to image-supported descriptions, including the resolved pleural effusion and localized pleural findings, while the consistency score recovers to 0.88. This example illustrates how consistency-aware control can limit the propagation of intermediate semantic deviations during autoregressive report generation.

C.2 Cross-Variant Report Comparison and Visual Grounding

Figure S5 presents a case study comparing reports generated by different model variants. From Figure S5, without LCSR, the model fails to capture fine-grained anatomical and pathological cues (e.g., *subtle pulmonary opacity*), resulting in a report that lacks specific abnormality descriptions, reflecting weak alignment between visual patterns and clinical concepts. Without consistency-aware control, the model captures certain abnormal cues (e.g., *mild linear opacity*); however, the description remains coarse, with incomplete localization and reduced diagnostic specificity, due to semantic deviations and misalignment with the image evidence. Conversely, LCC accurately identifies

the subtle opacity (e.g., *linear density in the left mid-to-lower lung*), localizes it to the appropriate anatomical region, and further characterizes it as plate-like atelectasis, providing clinically grounded interpretations that closely align with the ground truth.

We further examine finding-conditioned Grad-CAM maps to assess visual grounding. Without LCSR, the model exhibits diffuse responses across the thoracic region, suggesting weaker correspondence between the target finding and visual evidence. Without consistency-aware control, the responses become more localized but remain relatively broad. In contrast, LCC produces more concentrated responses around the relevant anatomical region, providing qualitative evidence of improved visual grounding.

D Limitations and Future Work

Despite the encouraging results, several limitations remain. First, our experiments are conducted on two publicly available chest X-ray datasets, and the generalizability of LCC to other imaging modalities and clinical environments has not yet been systematically evaluated. Future work will extend the framework to more diverse radiological data and external datasets.

Second, clinical factual consistency is mainly evaluated using automatic metrics such as CheXpert and RadGraph F1. Although these metrics provide useful quantitative assessment, they cannot fully reflect radiologists' judgments regarding diagnostic relevance and clinical correctness. Incorporating expert evaluation will therefore be an important direction for future study.

Finally, the current consistency evaluator is trained using matched and mismatched image–report prefixes, which provides effective supervision for cross-modal consistency discrimination but may not cover all fine-grained semantic deviations encountered during generation. Future work may explore more informative consistency supervision to improve sensitivity to subtle clinical discrepancies.