Analyzing and Enhancing Visual Learning in LLM-based Radiology Report Generation

Supplementary Material

A. Detailed Derivations of Mechanistic Analysis

This appendix provides detailed derivations for the mechanistic analysis in Section 3.2, covering causal attention asymmetry, gradient propagation through stacked decoder layers, and the indirect supervision path for the visual encoder.

A.1. Causal Attention Asymmetry (Eqs. (3)–(4))

Given input hidden states $\mathbf{X} \in \mathbb{R}^{L \times d_L}$, the causal attention operation in an LLM decoder layer is

$$\mathrm{Attn}(\mathbf{X}) = \mathrm{Softmax}_{\mathrm{row}}\Big(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} + \mathbf{M}^c\Big)\mathbf{X}_V = \mathbf{A}\mathbf{X}_V, \quad (11)$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{X}_V = \mathbf{X}\mathbf{W}_Q, \mathbf{X}\mathbf{W}_K, \mathbf{X}\mathbf{W}_V$. The causal mask $\mathbf{M}^c$ is

$$\mathbf{M}^c_{ij} = \begin{cases} 0, & i \geq j, \\ -\infty, & i < j, \end{cases}$$

$\mathbf{A}$ is strictly lower-triangular and thus prohibits each token from attending to any future positions.

In LLM-based RRG, the input sequence concatenates vision and report tokens:

$$\mathbf{X} = [\mathbf{V}, \mathbf{T}], \quad i < j,\ i \in \mathcal{I}_{\text{vision}},\ j \in \mathcal{I}_{\text{report}}.$$

Expanding the attention outputs gives

$$\mathrm{Attn}(\mathbf{T}_j, \mathbf{X}) = \sum_{q \in \mathcal{I}_{\text{vision}}} \mathbf{A}_{jq}\mathbf{V}_q + \sum_{k \in \mathcal{I}_{\text{report}}, k \leq j} \mathbf{A}_{jk}\mathbf{T}_k, \quad (12)$$

$$\mathrm{Attn}(\mathbf{V}_i, \mathbf{X}) = \sum_{q \in \mathcal{I}_{\text{vision}}, q \leq i} \mathbf{A}_{iq}\mathbf{V}_q. \quad (13)$$

Report tokens can aggregate both visual and textual information, whereas vision tokens only aggregate from earlier vision tokens. Because the value projection is position-wise, it does not influence cross-token flow and is omitted for clarity. These expressions correspond to Eqs. (3)–(4) in the main text.

A.2. Gradient Propagation through Causal Attention (Eqs. (6)–(8))

Justification for the dominant V-path approximation. The Jacobians of the $\mathbf{Q}$- and $\mathbf{K}$-paths are row- and column-local: the derivation of the softmax function couples only the entries within the same query row or key column. Thus, for $\mathbf{A}$ defined in Eq. (11), $\partial\mathbf{A}/\partial\mathbf{Q}$ and $\partial\mathbf{A}/\partial\mathbf{K}$ do not transmit gradients across token positions. In contrast, the $\mathbf{V}$-path exhibits a fully dense Jacobian $\mathbf{A}^\top$, which is the only pathway capable of propagating gradients between vision and report tokens. Therefore, cross-token gradient transfer is dominated by the V-path, and ignoring Q/K paths does not affect the structural conclusions.

Let $\mathbf{H}^{(\ell)}$ denote the input to layer ℓ and $\mathbf{H}^{(\ell+1)} = \mathrm{Attn}(\mathbf{H}^{(\ell)})$ denote its output. Then

$$\frac{\partial \mathcal{L}}{\partial \mathbf{H}^{(\ell)}} \approx (\mathbf{A}^{(\ell)})^\top \frac{\partial \mathcal{L}}{\partial \mathbf{H}^{(\ell+1)}}, \quad (14)$$

where $(\mathbf{A}^{(\ell)})^\top$ is strictly upper-triangular due to the causal mask. Thus, gradients propagate only from later (report) positions to earlier (vision) positions.

Example: two stacked decoder layers. For attention matrices $\mathbf{A}^{(0)}$ and $\mathbf{A}^{(1)}$, the combined gradient operator

$$\mathbf{A}^{(1)\top}\mathbf{A}^{(0)\top}$$

is still strictly upper-triangular, so gradients can only flow from a report position j to a vision position i when $i < j$.

A.3. Layer-wise Gradient Accumulation (Eq. (9))

Let $\frac{\partial \mathcal{L}_{\text{CE}}}{\partial \mathbf{H}^{(0)}}$ denote the gradient of the cross-entropy loss with respect to the input hidden states of the first decoder layer, which represents the supervision signal ultimately propagated to the visual encoder. Iterating this relationship over N decoder layers yields

$$\frac{\partial \mathcal{L}_{\text{CE}}}{\partial \mathbf{H}^{(0)}} \approx \Big(\prod_{\ell=N-1}^{0} (\mathbf{A}^{(\ell)})^\top\Big) \frac{\partial \mathcal{L}_{\text{CE}}}{\partial \mathbf{H}^{(N)}}. \quad (15)$$

Because each $(\mathbf{A}^{(\ell)})^\top$ is strictly upper-triangular, their product remains upper-triangular, preserving the causal structure across layers.

Layer-wise closure of causal masking. Since the product of upper-triangular matrices is also upper-triangular, stacking decoder layers cannot introduce new cross-boundary visibility in the forward pass: a vision token never attends to any report token at any layer. In the backward pass, the same causal structure holds in the opposite direction, with gradients flowing along valid causal paths from the later positions to the earlier ones.

Consequently, gradients flow backward solely along the causal paths from report to vision tokens, with cumulative attention coefficients determining the effective supervision strength at each vision position.

A.4. Gradient Path to the Visual Encoder

As defined in the main text, visual embeddings are generated by

$$\mathbf{E}_v = P_v(g_v(\mathbf{I})),$$

where g_v is the visual encoder and P_v the projector. The gradient with respect to the parameters $\boldsymbol{\theta}_v$ of the encoder and projector is

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}_v} = \frac{\partial \mathcal{L}}{\partial \mathbf{E}_v} \frac{\partial \mathbf{E}_v}{\partial \boldsymbol{\theta}_v}, \tag{16}$$

where $\partial \mathcal{L}/\partial \mathbf{E}_v$ originates entirely from the attention-mediated cross-token gradient flow described above. Since no loss is directly applied to vision tokens, the visual encoder receives only *indirect*, linguistically mediated supervision, limiting its ability to learn semantically rich and clinically meaningful visual features.

B. Experimental Details

B.1. Implementation Details

Our implementation follows the R2GenGPT pipeline [39], which employs a Swin Transformer [23] as the visual encoder and a frozen LLaMA2-7B [33] as the LLM decoder. On top of this backbone, we integrate the proposed Semantic Injection (SI) and Semantic Alignment (SA) modules.

Semantic Injection (SI). SI uses a trainable Bio_ClinicalBERT [2] text encoder to extract diagnostic semantics from the ground-truth report. A linear projector maps these textual features into the LLM embedding space for the non-shift MSE objective.

Semantic Alignment (SA). SA aligns visual features with report embeddings, using the LLaMA tokenizer and embedding layer to obtain textual representations. The global (CXR–report) and local (region–sentence) contrastive losses are applied after projecting visual features through a linear projector that matches the dimension of the LLM embedding space.

Training setup. The visual encoder, text encoder, and projectors are trainable, while the LLM is fully frozen. Training is conducted on two NVIDIA A100 (80GB) GPUs for 10 epochs on MIMIC-CXR and 15 epochs on IU-Xray, using AdamW [24] with a learning rate of 1×10^{-4} and a cosine-annealing scheduler. Mixed precision (bf16) is used for efficiency.

The overall training loss is the weighted sum of (i) CE loss on report tokens, (ii) MSE loss for SI, and (iii) contrastive losses for SA, with weights $\lambda_{\text{SI}} = 1.0$ and $\lambda_{\text{SA}} = 0.1$ (see Eq. (10) in the main text).

Inference. At inference, the SI and SA modules are removed and only the standard vision-to-text report generation branch is needed. The LLM decoder remains frozen and uses beam search (beam size = 3), and we set *max_token_output* to 60 for IU X-ray and 100 for MIMIC-CXR. All images are resized and center-cropped to 224×224 for both training and testing.