DELIBRRG: TRUSTWORTHY CROSS-EXAMINATION AND CLINICAL-FIRST DELIBERATION FOR RADIOLOGY REPORT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Radiology report generation (RRG) aims to generate accurate and standardized diagnostic descriptions from medical images. Existing methods improve report quality through stronger representation learning and cross-modal interaction, but typically lack a stable clinical reference and an explicit mechanism for coordinating diagnostic and linguistic objectives. As a result, diagnostic roles may lose functional distinction during semantic interaction, while clinical fact preservation can be weakened by competing generation objectives. Inspired by information structure theory and multi-expert decision-making, we propose Trustworthy Cross-Examination and Clinical-First **Delib**eration for **R**adiology **R**eport **G**eneration (**DelibRRG**), which reformulates RRG as a clinical deliberation process organized by clinical reference establishment, heterogeneous role interaction, and asymmetric coordination. DelibRRG first employs dual-space structural calibration to construct a trustworthy clinical reference by calibrating candidate cross-modal correspondences through neighborhood consensus. It then constructs functionally decoupled clinical and semantic streams that exchange complementary information through controlled bidirectional cross-examination, followed by training-time clinical review of the updated diagnostic representation. Finally, clinical-first asymmetric bargaining assigns unequal bargaining power and a clinical disagreement point to bias joint optimization toward clinical fact preservation. Experiments on MIMIC-CXR and IU-Xray demonstrate strong report generation performance, while the clinical efficacy gains on MIMIC-CXR suggest improved clinical factual accuracy.

1 INTRODUCTION

Radiology report generation (RRG) aims to automatically generate accurate and standardized clinical reports from radiological images. Unlike general image captioning (Vinyals et al., 2015), RRG requires accurate descriptions of anatomical structures, pathological findings, and clinical facts. Balancing language fluency with clinical factual accuracy therefore remains a central challenge.

As illustrated in Fig. 1, early RRG methods mainly focused on cross-modal representation learning and sequence decoding, using attention and shared memory to associate visual and textual information (Chen et al., 2020; 2021). Subsequent studies introduced knowledge enhancement, relational constraints, and fine-grained cross-modal modeling to improve clinical semantics and multimodal interaction (Yang et al., 2023; Chen et al., 2024; Hou et al., 2026). Despite these advances, most methods still emphasize stronger cross-modal associations or richer joint representations for report decoding (Song et al., 2025; Jayas et al., 2026). Without explicit reliability verification, candidate associations may provide an unstable factual basis for subsequent interaction. Moreover, increasingly coupled representations leave the functional distinction between visual diagnosis and semantic expression implicit, allowing diagnostic information to be weakened during deep interaction. Clinical fact preservation and language generation may also compete during joint optimization. Recent gradient-based coordination methods mitigate such conflicts (Zhang et al., 2026), but do not explicitly assign higher decision priority to clinical facts.

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.

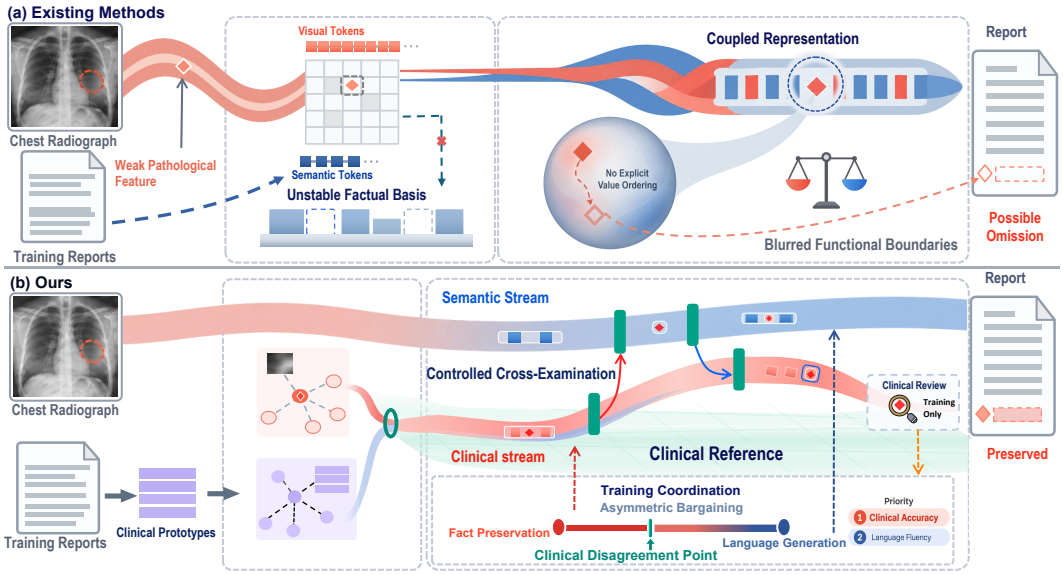


Figure 1: Comparison between existing RRG methods and DelibRRG. Existing methods mainly strengthen cross-modal associations and joint representations, whereas DelibRRG constructs a trustworthy clinical reference through dual-space structural calibration, performs functionally decoupled cross-examination and clinical review, and applies clinical-first asymmetric bargaining.

Information structure theory in games of incomplete information suggests that strategic interaction depends on the information structure available to participants, rather than only on their actions and payoffs (Bergemann & Morris, 2016). Multi-expert decision-making further shows that decision rules can regulate information exchange and the relative influence of experts with different objectives (Zapechelnyuk, 2013). Inspired by these perspectives, we argue that RRG should first establish a trustworthy shared clinical reference, then organize controlled interaction between functionally distinct roles, and finally coordinate competing objectives according to their unequal clinical importance.

Based on this motivation, we propose DelibRRG, which reformulates RRG as a clinical deliberation process organized by clinical reference establishment, heterogeneous role interaction, and clinical-first coordination. DelibRRG first introduces Dual-Space Structural Calibration (DSSC) to calibrate candidate cross-modal correspondences and construct a trustworthy clinical reference. Functionally Decoupled Cross-Examination and Clinical Review (FDCR) then organizes the clinical and semantic streams, enabling controlled bidirectional cross-examination while re-examining the updated diagnostic representation through training-time clinical review. Finally, Clinical-First Asymmetric Bargaining (CFAB) introduces a clinical disagreement point and unequal bargaining power to coordinate clinical fact preservation and language generation, introducing an explicit clinical preference during joint optimization.

In summary, our main contributions are as follows:

1. We reformulate RRG as a clinical deliberation process organized by trustworthy clinical reference construction, functionally decoupled cross-examination, and clinical-first asymmetric bargaining.
2. We introduce DSSC to calibrate candidate cross-modal correspondences through neighborhood consensus and construct a trustworthy clinical reference for subsequent interaction.
3. We develop FDCR to preserve functional role differences during interaction and CFAB to introduce an explicit clinical-first preference during joint optimization.
4. Experiments on MIMIC-CXR and IU-Xray demonstrate strong report generation performance and improved clinical efficacy scores.

2 Related Work

Clinical Reference Learning. Radiology report generation requires a clinically meaningful representation from medical images to support diagnosis and language generation. Early methods strengthened visual-language modeling through relational memory and curriculum learning (Chen et al., 2020; Liu et al., 2021a). PPKED (Liu et al., 2021b), Clinical-BERT (Yan & Pei, 2022), and GSKET (Yang et al., 2022) further introduced abnormal-region cues, disease semantics, and medical-knowledge priors. M2KT (Yang et al., 2023) aligned images, disease labels, and report semantics through a learnable knowledge base, while PromptMRG (Jin et al., 2024) combined diagnosis-aware prompts with similar-report retrieval. DACG (Lang et al., 2025) emphasized abnormal regions, DMVF (Tang et al., 2025) modeled visual information at multiple granularities, and CMCRL (Chen et al., 2025) reduced potential confounding through vision-language causal intervention. Despite these advances, most methods still emphasize stronger individual correspondences, leaving their structural reliability before deep interaction less explored. This motivates constructing a trustworthy clinical reference before subsequent role interaction.

Function-Aware Cross-Modal Interaction. Studies have also explored richer information interaction during report generation. DACG (Lang et al., 2025) and DMVF (Tang et al., 2025) preserve visual contributions and complementary multi-granularity information, while DDaTR (Song et al., 2025) and TSMRG (Cao et al., 2026) model longitudinal changes. CCFN (Hou et al., 2026) further promotes coarse-to-fine visual-textual fusion. With the use of Large language models (LLMs), RRG methods have integrated broader contextual information. R2GenCSR (Wang et al., 2026a) combined current visual features with related cases and residual information, while LlamaRG (Jayas et al., 2026) used multi-view images to guide LLM-based generation. FedMRG (Che et al., 2026) further addressed multi-center heterogeneity through client-aware contrastive learning, diagnosis-driven prompts, and dual adapters. Although these methods strengthen information interaction, they mainly aim to build richer joint representations or generation contexts. Increasingly coupled interaction can blur the functional boundary between diagnosis and semantic expression, allowing clinically important evidence to be weakened during repeated fusion. This motivates functionally decoupled bidirectional cross-examination, followed by clinical review of the updated diagnostic representation under a shared clinical criterion.

Clinical-First Optimization. In addition to representation and interaction design, RRG methods have also used auxiliary clinical objectives during training. Clinical-BERT (Yan & Pei, 2022) strengthened disease semantics through medical-domain pretraining, M2KT (Yang et al., 2023) constrained image-disease-report relations through multimodal alignment, and PromptMRG (Jin et al., 2024) combined disease classification with adaptive logit adjustment. These approaches add clinically relevant supervision, but their objectives are still mainly optimized through joint losses without explicit priority. CAME-Grad (Zhang et al., 2026) more directly addresses conflicts between report generation and clinical supervision through gradient coordination. However, coordinating conflicting objectives does not explicitly determine their relative clinical priority. Because clinical factual accuracy and language fluency do not have equal medical value, we adopt clinical-first coordination that introduces an explicit preference for clinical fact preservation through asymmetric rules.

3 Method

3.1 Overall Framework and Problem Formulation

Given a case with M chest X-ray views, $\mathcal{I} = \{I^{(m)}\}_{m=1}^{M}$, we extract spatial features using the visual encoder $f_v(\cdot)$ and average features at corresponding locations: $V = \frac{1}{M}\sum_{m=1}^{M} f_v(I^{(m)}) \in \mathbb{R}^{N_v \times D_v}$, where N_v and D_v denote the number and dimension of visual tokens, respectively. For a single-view case, this reduces to the standard single-image representation.

Clinical reference construction is performed in the D_v-dimensional visual space. FDCR then maps the representations into a D_b-dimensional bottleneck space for cross-examination before restoring them to the visual dimension. After clinical calibration and role interaction, the generation representation is projected into LLaMA hidden space as $E_{\text{vis}} = \text{LN}(\Pi(V_{\text{gen}}))$, where $\Pi(\cdot)$ is the visual-language projector. Projected features and textual instruction jointly condition autoregressive report generation with LLaMA-2-7B-Chat (Touvron et al., 2023).

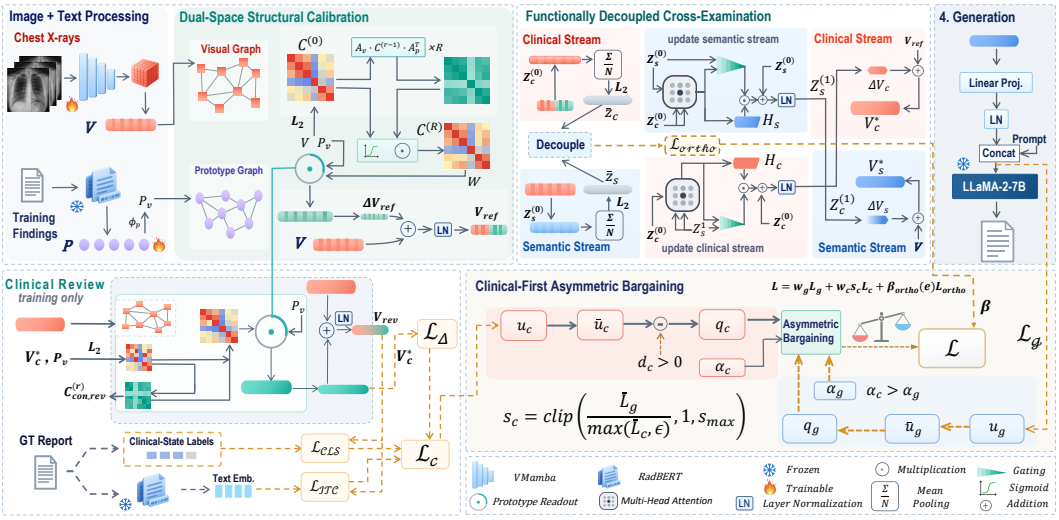


Figure 2: Overview of DelibRRG. Dual-Space Structural Calibration establishes a trustworthy clinical reference. Functionally Decoupled Cross-Examination and Clinical Review enables interaction between distinct clinical and semantic streams, with clinical review used only during training. Clinical-First Asymmetric Bargaining coordinates the clinical and generation objectives, while the semantic stream supports report generation.

As shown in Fig. 2, DSSC first constructs a trustworthy clinical reference, FDCR then organizes functionally decoupled cross-examination and clinical review, and CFAB coordinates clinical fact preservation and report generation through asymmetric bargaining.

During training, each sample provides CheXbert (Smit et al., 2020) labels $Y^{\text{ce}} \in \{0,1,2,3\}^{14}$, while the processed report is encoded by frozen RadBERT (Yan et al., 2022) to obtain $T \in \mathbb{R}^{D_t}$. Neither is used for validation or test generation.

3.2 Trustworthy Clinical Reference via Dual-Space Structural Calibration

Token-level similarity alone does not ensure reliable correspondence. We therefore calibrate cross-space matches using local structures in both spaces.

Clinical prototype initialization. We maintain learnable clinical prototypes $P = \{p_k\}_{k=1}^{K} \in \mathbb{R}^{K \times D_t}$. Training-set *Findings* sentences are encoded by frozen RadBERT, mean-pooled, and clustered by mini-Batch k-means to initialize P. The resulting prototypes are projected into the VMamba space as $P_v = \phi_p(P) \in \mathbb{R}^{K \times D_v}$.

Dual-Space Structural Calibration. We L2-normalize the visual tokens and projected prototypes as $\hat{V}$ and $\hat{P}_v$. For $X \in \{\hat{V}, \hat{P}_v\}$, we construct a top-$k$ neighborhood graph by

$$A_X(i,j) = \frac{\exp(\text{sim}(x_i, x_j))}{\sum_{l \in \mathcal{N}_k(i)} \exp(\text{sim}(x_i, x_l))}, \quad j \in \mathcal{N}_k(i), \tag{1}$$

with zero weight otherwise and self-connections excluded. This yields A_v and A_p for the visual and prototype spaces, respectively.

The initial cross-space correspondence is

$$C^{(0)} = \hat{V}\hat{P}_v^{\top} \in \mathbb{R}^{N_v \times K}. \tag{2}$$

At calibration round $r \in \{1, \ldots, R\}$, neighborhood consensus is

$$C_{\text{con}}^{(r)} = A_v C^{(r-1)} A_p^{\top}. \tag{3}$$

The correspondence is then updated as

$$C^{(r)} = C^{(0)} \odot \sigma\left(\lambda C_{\text{con}}^{(r)}\right). \tag{4}$$

Here, λ controls the calibration strength.

After R rounds, the calibrated correspondence is used for prototype retrieval:

$$W = \mathrm{Softmax}\Big(\tau C^{(R)}\Big), \qquad S = WP_v. \tag{5}$$

Here, $\tau > 0$ is a learnable scale. The retrieved clinical semantics are written back through a controlled residual path:

$$V_{\mathrm{ref}} = \mathrm{LN}(V + \Delta V_{\mathrm{ref}}), \qquad \Delta V_{\mathrm{ref}} = \rho_t \tanh(\eta)\psi(S), \tag{6}$$

where $\psi(\cdot)$ is a semantic mapping network, η is a learnable scalar gate, and $\rho_t = \min(1, (t+1)/T_w)$ gradually increases the residual strength during warm-up.

The resulting V_{ref} serves as the trustworthy clinical reference. Here, "trustworthy" refers to a structurally calibrated reference supported by clinical supervision, rather than a formal guarantee of clinical correctness.

3.3 Functionally Decoupled Cross-Examination and Clinical Review

Because diagnosis and generation have different functional requirements, we initialize functionally decoupled semantic and clinical streams as $Z_s^{(0)} = f_s(V)$ and $Z_c^{(0)} = f_c(V_{\mathrm{ref}})$, respectively, in a D_b-dimensional bottleneck space.

Before interaction, we reduce global directional similarity between the two roles:

$$\mathcal{L}_{\mathrm{ortho}} = \frac{1}{B}\sum_{b=1}^{B}\left(\bar{z}_{s,b}^{\top}\bar{z}_{c,b}\right)^2. \tag{7}$$

This encourages role distinction without token-wise orthogonality.

Bidirectional cross-examination. Let $\mathcal{X} = \{s, c\}$ denote the semantic and clinical streams, with $\bar{s} = c$ and $\bar{c} = s$. We set $\delta_s = 0$ and $\delta_c = 1$, yielding clinical-to-semantic injection followed by semantic-to-clinical verification. For each $i \in \mathcal{X}$,

$$\begin{aligned} H_i &= \mathrm{MHA}\left(Z_i^{(0)}, Z_{\bar{i}}^{(\delta_i)}, Z_{\bar{i}}^{(\delta_i)}\right), \\ G_i &= \sigma\Big(W_i[Z_i^{(0)}; H_i] + b_i\Big), \\ Z_i^{(1)} &= \mathrm{LN}\left(Z_i^{(0)} + G_i \odot H_i\right). \end{aligned} \tag{8}$$

Thus, the semantic stream is updated via $Z_c^{(0)}$, and $Z_s^{(1)}$ returns to the clinical stream for reverse examination, enabling controlled clinical injection followed by semantic-to-clinical verification.

The updated bottleneck representations are restored to the visual dimension:

$$\begin{aligned} V_s^{\star} &= V + \Delta V_s, \qquad \Delta V_s = r_s(Z_s^{(1)}), \\ V_c^{\star} &= V_{\mathrm{ref}} + \Delta V_c, \quad \Delta V_c = r_c(Z_c^{(1)}). \end{aligned} \tag{9}$$

The semantic output is used for generation, $V_{\mathrm{gen}} = V_s^{\star}$, while $V_c^{\star}$ is used for training-time clinical review rather than being written back into the generation stream.

Training-time clinical review. During training, the shared calibration module reprocesses $V_c^{\star}$ while rebuilding its visual graph. The review correspondence starts from $C_{\mathrm{rev}}^{(0)} = \widehat{V_c^{\star}}\widehat{P_v}^{\top}$ and is updated by

$$C_{\mathrm{con,rev}}^{(r)} = A_{\mathrm{rev}} C_{\mathrm{rev}}^{(r-1)} A_p^{\top}, \quad C_{\mathrm{rev}}^{(r)} = C_{\mathrm{rev}}^{(0)} \odot \sigma\Big(\lambda C_{\mathrm{con,rev}}^{(r)}\Big). \tag{10}$$

After R rounds, $W_{\mathrm{rev}} = \mathrm{Softmax}(\tau C_{\mathrm{rev}}^{(R)})$ and $S_{\mathrm{rev}} = W_{\mathrm{rev}} P_v$. The reviewed representation is obtained from $\Delta V_{\mathrm{rev}} = \rho_t \tanh(\eta)\psi(S_{\mathrm{rev}})$ and $V_{\mathrm{rev}} = \mathrm{LN}(V_c^{\star} + \Delta V_{\mathrm{rev}})$.

We supervise the reviewed representation with masked four-state CheXbert classification $\mathcal{L}_{\mathrm{CLS}}$ over 14 observations, ignoring Not Mentioned, and bidirectional InfoNCE $\mathcal{L}_{\mathrm{ITC}}$ (Radford et al.,

2021) between the reviewed global visual representation and the frozen RadBERT report representation. To constrain excessive clinical residuals, we define $r_\Delta = \frac{\mathrm{AvgNorm}(\Delta V_{\mathrm{rev}})}{\max(\mathrm{AvgNorm}(V_c^\star),\epsilon)}$ and $\mathcal{L}_\Delta = [\max(0, r_\Delta - r_0)]^2$. The resulting clinical objective is

$$\mathcal{L}_c = \lambda_{\mathrm{CLS}}\mathcal{L}_{\mathrm{CLS}} + \lambda_{\mathrm{ITC}}\mathcal{L}_{\mathrm{ITC}} + \lambda_\Delta\mathcal{L}_\Delta. \tag{11}$$

Clinical review is training-only and does not replace V_{gen}; validation and test generation use no target reports, report representations, or labels.

3.4 Clinical-First Asymmetric Bargaining

Let Ω denote the set of valid supervised report tokens. The autoregressive generation objective is

$$\mathcal{L}_g = -\frac{1}{|\Omega|}\sum_{t\in\Omega}\log p_\theta(y_t \mid y_{<t}, E_{\mathrm{vis}}). \tag{12}$$

To adapt the balance between clinical preservation and generation, CFAB implements asymmetric bargaining through dynamic utility-based weighting rather than stepwise Nash-product optimization.

For objective $j \in \{c, g\}$, the current utility is

$$u_j^{(t)} = \exp\Big[-\mathrm{sg}\Big(\mathcal{L}_j^{(t)}\Big)\Big], \qquad j \in \{c, g\}, \tag{13}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. Its historical state is maintained by $\bar{u}_j^{(t)} = \gamma\bar{u}_j^{(t-1)} + (1-\gamma)u_j^{(t)}$.

We assign a nonzero disagreement point d_c to the clinical objective and set $d_g = 0$. The corresponding utility margins are $q_c = \max(\bar{u}_c - d_c, \epsilon)$ and $q_g = \max(\bar{u}_g, \epsilon)$. Let α_c and α_g denote their bargaining powers. The normalized dynamic weight is

$$w_j = \frac{2\alpha_j/q_j}{\alpha_c/q_c + \alpha_g/q_g}, \qquad j \in \{c, g\}, \tag{14}$$

Thus, $w_c + w_g = 2$. The settings $\alpha_c > \alpha_g$ and $d_c > 0$ introduce an explicit preference toward the clinical objective, although they do not guarantee $w_c > w_g$ at every optimization step. The final clinical contribution is jointly determined by the dynamic utility weights and the scale compensation term.

Because $\mathcal{L}_c$ and $\mathcal{L}_g$ may have different numerical scales, we maintain exponential moving averages of the raw losses and use

$$s_c = \mathrm{clip}\left(\frac{\bar{\mathcal{L}}_g}{\max(\bar{\mathcal{L}}_c, \epsilon)}, s_{\min}, s_{\max}\right). \tag{15}$$

With $s_{\min} = 1$, this term can increase the effective clinical contribution when $\mathcal{L}_c$ is numerically smaller but does not reduce it. The dynamic weights w_c and w_g are always computed from the original losses, and s_c is applied only in the final combination.

The role-decoupling term is introduced with delayed annealing:

$$\beta_{\mathrm{ortho}}(e) = \begin{cases} 0, & e < 2, \\ \beta_{\mathrm{ortho}}^{\max}\min\left(1, \dfrac{e-2}{4}\right), & e \geq 2, \end{cases} \tag{16}$$

where e is the current epoch.

The complete training objective is

$$\mathcal{L} = w_g\mathcal{L}_g + w_c s_c\mathcal{L}_c + \beta_{\mathrm{ortho}}(e)\mathcal{L}_{\mathrm{ortho}}. \tag{17}$$

Here, $\mathcal{L}_g$ optimizes report generation, $\mathcal{L}_c$ provides clinical fact supervision, and $\mathcal{L}_{\mathrm{ortho}}$ preserves functional differences between the two roles. Together, the dynamic utility weights, asymmetric bargaining power, clinical disagreement point, and scale compensation jointly implement the proposed CFAB mechanism.

4 EXPERIMENTS

4.1 DATASETS

We evaluate DelibRRG on MIMIC-CXR (Johnson et al., 2019) and IU-Xray (Demner-Fushman et al., 2016). For MIMIC-CXR, we use 270,790 training, 2,130 validation, and 3,858 test examples. IU-Xray follows the commonly used RRG split. Reports are cleaned by the dataset preprocessing pipeline before tokenization, and all radiographs are resized to 224×224. The two benchmarks provide complementary evaluation settings across different dataset scales and report distributions.

4.2 EVALUATION METRICS

We evaluate the generated reports using natural language generation (NLG) and clinical efficacy (CE) metrics. On MIMIC-CXR, we report BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE-L, and CIDEr for NLG, together with precision, recall, and F1-score for CE. On IU-Xray, we report BLEU-3, BLEU-4, ROUGE-L, and CIDEr. Combining NLG and CE metrics allows us to assess both textual agreement and the preservation of clinically relevant observations, which directly reflects the goal of DelibRRG.

4.3 IMPLEMENTATION DETAILS

We implement DelibRRG in PyTorch using pretrained VMamba-Base (Liu et al., 2024b) as the visual encoder and frozen LLaMA-2-7B-Chat as the language model, connected by a learnable linear projector. The visual encoder, projector, and clinical deliberation modules are jointly optimized. Clinical prototypes are initialized from training-set *Findings* using frozen RadBERT and remain learnable during training. Multi-view features are aggregated before clinical deliberation.

DSSC is shared between trustworthy clinical reference construction and training-time clinical review. Only the semantic stream is projected to LLaMA for generation, while the clinical stream is reviewed only during training. Validation and test generation use only the input radiographs. We optimize the model with AdamW on a single NVIDIA A40 GPU. Keeping LLaMA frozen ensures that the reported improvements are obtained without adapting the language model, while the visual encoder, projector, and deliberation modules remain trainable.

4.4 EXPERIMENTAL RESULTS

4.4.1 COMPARISON WITH EXISTING METHODS

On MIMIC-CXR, we compare DelibRRG with a broad range of representative RRG methods, including R2Gen (Chen et al., 2020), R2GenCMN (Chen et al., 2021), Clinical-BERT (Yan & Pei, 2022), R2GenGPT (Wang et al., 2023), PromptMRG (Jin et al., 2024), FMVP (Liu et al., 2024c), CMCRL (Chen et al., 2025), RRG-Mamba (Hou et al., 2025), CyclicAligner (Sun et al., 2025), MambaXray-VL (Wang et al., 2025), TSMRG (Cao et al., 2026), FedMRG (Che et al., 2026), and LlamaRG (Jayas et al., 2026). Table 1 reports their NLG performance in terms of BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE-L, and CIDEr.

As shown in Table 1, DelibRRG achieves the best or second-best performance across all seven NLG metrics. Clinical efficacy results are reported in the left part of Table 2. For clinical efficacy evaluation on MIMIC-CXR, we further compare with HERGen (Wang et al., 2024), DMVF (Tang et al., 2025), CCFN (Hou et al., 2026), and AM-MRG (Wang et al., 2026b), together with the representative methods reported above.

DelibRRG achieves the best Precision, Recall, and F1-score among the compared methods, indicating improved preservation of clinically important observations. Together with the NLG results, these gains show that the improvements extend beyond textual similarity to clinical content preservation.

DSSC constructs a trustworthy clinical reference before deep interaction, FDCR enables controlled interaction between functionally decoupled roles, and CFAB prioritizes clinical fact preservation during joint optimization. These complementary mechanisms contribute to the improvements in both NLG and CE performance.

Table 1: Comparison with representative methods on MIMIC-CXR using NLG metrics.

Method	Venue	B-1↑	B-2↑	B-3↑	B-4↑	METEOR↑	ROUGE-L↑	CIDEr↑
R2Gen	EMNLP'20	0.353	0.218	0.145	0.103	0.142	0.277	–
R2GenCMN	IJCNLP'21	0.353	0.218	0.148	0.106	0.142	0.278	–
Clinical-BERT	AAAI'22	0.383	0.230	0.151	0.106	0.144	0.275	0.151
R2GenGPT	Meta-Rad'23	0.408	0.256	0.174	0.125	0.167	0.285	0.244
PromptMRG	AAAI'24	0.398	–	–	0.112	0.157	0.268	–
FMVP	TMM'24	0.389	0.236	0.156	0.108	0.150	0.284	–
CMCRL	TIP'25	0.400	0.245	0.165	0.119	0.150	0.280	–
RRG-Mamba	IJCAI'25	0.406	0.253	0.169	0.121	0.154	0.293	–
CyclicAligner	BIBM'25	0.411	0.254	0.170	0.121	0.157	0.284	0.170
MambaXray-VL	CVPR'25	0.422	0.268	0.184	0.133	0.163	0.289	0.241
TSMRG	TMM'26	0.403	0.244	0.161	0.113	0.158	0.272	–
FedMRG	TMI'26	0.401	0.255	0.175	0.126	–	0.275	–
LlamaRG-SCST	CVPR'26	0.417	**0.270**	0.178	0.130	0.162	**0.302**	0.169
DelibRRG (Ours)	–	**0.423**	0.269	**0.186**	**0.136**	**0.169**	0.294	**0.255**

Table 2: Clinical efficacy results on MIMIC-CXR and report generation results on IU-Xray.

Method	P↑	R↑	F1↑
R2Gen	0.333	0.273	0.276
R2GenCMN	0.334	0.275	0.278
R2GenGPT	0.392	0.387	0.389
HERGen	0.415	0.301	0.317
CMCRL	0.489	0.340	0.401
DMVF	0.451	0.386	0.413
CCFN	0.476	0.358	0.409
AM-MRG	0.532	0.415	0.466
LlamaRG-SCST	0.448	0.431	0.414
DelibRRG (Ours)	**0.540**	**0.435**	**0.482**

Method	B-3↑	B-4↑	R-L↑	CIDEr↑
R2Gen	0.219	0.165	0.371	–
PPKED	0.224	0.168	0.376	0.351
R2GenGPT	0.214	0.161	0.376	0.542
BootstrappingLLM	0.180	0.128	0.291	–
DDaTR	–	0.103	0.307	–
PromptMRG	–	0.098	0.160	–
MambaXray-VL	0.241	0.185	0.371	0.524
DADS	0.242	0.187	0.384	**0.634**
FedMRG	0.173	0.120	0.297	–
DelibRRG (Ours)	**0.249**	**0.194**	**0.393**	0.449

4.4.2 Cross-Dataset Evaluation

On IU-Xray, we compare DelibRRG with R2Gen (Chen et al., 2020), PPKED (Liu et al., 2021b), R2GenGPT (Wang et al., 2023), BootstrappingLLM (Liu et al., 2024a), DDaTR (Song et al., 2025), PromptMRG (Jin et al., 2024), MambaXray-VL (Wang et al., 2025), DADS (Wu et al., 2026), and FedMRG (Che et al., 2026). We further evaluate DelibRRG on IU-Xray to examine its effectiveness under a smaller dataset with a different report distribution. Results are shown in the right part of Table 2.

DelibRRG achieves the best BLEU-3, BLEU-4, and ROUGE-L scores among the compared methods, showing that its improvements in lexical and sequence-level report quality transfer to the smaller IU-Xray benchmark.

4.4.3 Ablation Study

We conduct ablations on MIMIC-CXR to evaluate the contribution of DSSC, FDCR, and CFAB under the same baseline, training settings, and evaluation protocol.

As shown in Table 3, introducing DSSC improves both NLG and CE performance over the baseline, supporting the benefit of DSSC-based clinical reference construction. Adding FDCR on top of DSSC further improves most metrics, supporting its contribution as an integrated module for functionally decoupled interaction and clinical review.

The full configuration achieves the highest BLEU-1, BLEU-4, ROUGE-L, Recall, and F1 among the ablated variants, while DSSC+FDCR retains the highest CIDEr. These results support complementary roles for the three modules within the complete framework.

Table 3: Ablation study on MIMIC-CXR. All configurations use the same VMamba–LLaMA baseline.

	Modules			NLG Metrics				CE Metrics		
Configuration	**DSSC**	**FDCR**	**CFAB**	**B-1↑**	**B-4↑**	**R-L↑**	**CIDEr↑**	**P↑**	**R↑**	**F1↑**
Baseline				0.415	0.131	0.284	0.220	0.498	0.327	0.395
DSSC	✓			0.420	0.133	0.288	0.239	**0.553**	0.377	0.448
DSSC + FDCR	✓	✓		0.421	0.135	0.289	**0.268**	0.535	0.423	0.472
DSSC + CFAB	✓		✓	0.415	0.128	0.288	0.246	0.537	0.403	0.461
DelibRRG	✓	✓	✓	**0.423**	**0.136**	**0.294**	0.255	0.540	**0.435**	**0.482**

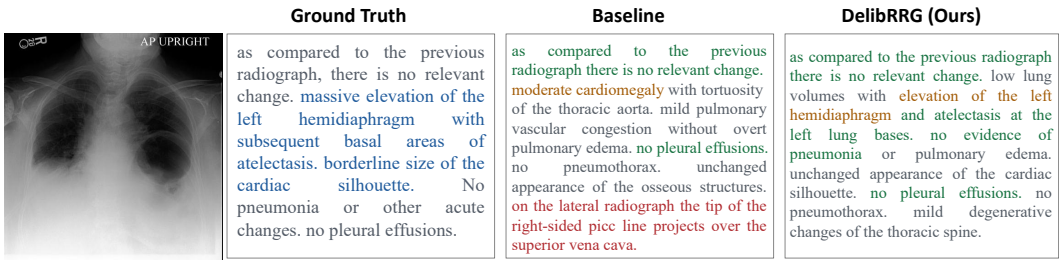


Figure 3: Qualitative comparison on a MIMIC-CXR case. Blue marks findings in the reference report, green marks semantically consistent generated content, orange marks orange marks partial or attribute-level matches, red marks contradictory or unsupported content, and dark gray marks routine or unverifiable descriptions.

4.4.4 Qualitative Results

We compare the reports generated by the baseline and DelibRRG for the same chest radiograph, with particular attention to clinically relevant findings and their anatomical descriptions.

As shown in Fig. 3, the baseline omits the elevated left hemidiaphragm and basal atelectatic findings in the reference report and introduces an unsupported PICC-line description. In contrast, DelibRRG captures both findings while avoiding this unsupported device-related description and maintaining report coherence. This example shows that DelibRRG better preserves clinically relevant findings and reduces unsupported content, consistent with its improved clinical efficacy.

The appendix additionally presents two supplementary chest radiograph cases, comparing reports from DelibRRG and the baseline and illustrating descriptions of bibasilar abnormalities and their possible causes.

5 Conclusion

We propose DelibRRG, which reformulates Radiology report generation as a clinical deliberation process organized by trustworthy clinical reference construction, functionally decoupled interaction, and clinical-first coordination. DSSC constructs a trustworthy clinical reference by calibrating candidate cross-modal correspondences, FDCR enables controlled bidirectional cross-examination with training-time clinical review, and CFAB prioritizes clinical fact preservation through asymmetric bargaining during joint optimization. Experiments on MIMIC-CXR and IU-Xray demonstrate strong report generation performance and improved preservation of clinically important observations. Ablation results further support the complementary contributions of DSSC, FDCR, and CFAB, while qualitative analysis shows better coverage of key pathological findings with coherent report generation. These results suggest that reliable RRG depends not only on stronger multimodal representations, but also on explicitly organizing how clinical evidence is established, examined, and prioritized throughout the generation process.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language polishing, retrieval of potentially relevant literature. All retrieved literature was verified against the original sources, and all AI-assisted outputs were reviewed by the authors. All scientific decisions, experimental results, and conclusions were independently validated by the authors. We take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, Appendix A.1 provides implementation details omitted from the main text, including the complete hyperparameter configurations of DSSC, FDCR, and CFAB, the prototype clustering procedure, optimization and scheduling settings, decoding parameters, and ablation objectives. All corresponding comparisons use the same training pipeline and fixed module-specific hyperparameters unless explicitly stated otherwise. We also report the random seed and sequence-length settings used in our experiments. The released code contains the training, evaluation, and configuration files required to reproduce the reported results.

REFERENCES

Dirk Bergemann and Stephen Morris. Bayes correlated equilibrium and the comparison of information structures in games. *Theoretical Economics*, 11(2):487–522, 2016.

Lili Cao, Haoran Wu, and Bo Xu. TSMRG: Temporal semantic enhancement for medical report generation with longitudinal data. *IEEE Transactions on Multimedia*, 2026.

Haoxuan Che, Haibo Jin, Zhengrui Guo, Yi Lin, Cheng Jin, and Hao Chen. Llm-driven medical report generation via communication-efficient heterogeneous federated learning. *IEEE Transactions on Medical Imaging*, 45(1):28–39, 2026.

Weixing Chen, Yang Liu, Ce Wang, Jiarui Zhu, Guanbin Li, Cheng-Lin Liu, and Liang Lin. Cross-modal causal representation learning for radiology report generation. *IEEE Transactions on Image Processing*, 34:2970–2985, 2025.

Wenting Chen, Linlin Shen, Jingyang Lin, Jiebo Luo, Xiang Li, and Yixuan Yuan. Fine-grained image-text alignment in medical imaging enables explainable cyclic image-report generation. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9494–9509, 2024.

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

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pp. 5904–5914, 2021.

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

Xiaodi Hou, Xiaobo Li, Mingyu Lu, Simiao Wang, and Yijia Zhang. RRG-Mamba: Efficient radiology report generation with state space model. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 7410–7418, 2025. doi: 10.24963/ijcai.2025/824.

Xiaodi Hou, Xiaobo Li, Mingyu Lu, Hai Cui, Hongfei Lin, and Yijia Zhang. Collaborative cross-modal fusion network for medical report generation. *Information Fusion*, 136:104576, 2026.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports, 2021. URL https://arxiv.org/abs/2106.14463.

Tanuja Jayas, Aditya Rastogi, Pavithra Raghavan, Gianluca Brugnara, Kai Schlamp, Martha Foltyn-Dumitru, and Philipp Vollmuth. LlamaRG: A multi-view large language model for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 5745–5754, 2026.

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*, volume 38, pp. 2607–2615, 2024.

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

Wangyu Lang, Zhi Liu, and Yijia Zhang. DACG: Dual attention and context guidance model for radiology report generation. *Medical Image Analysis*, 99:103377, 2025.

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*, volume 38, pp. 18635–18643, 2024a. doi: 10.1609/aaai.v38i17.29826.

Fenglin Liu, Shen Ge, and Xian Wu. Competence-based multimodal curriculum learning for medical 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. 3001–3012, 2021a.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and distilling posterior and prior knowledge for radiology report generation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13753–13762. IEEE, 2021b.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. VMamba: Visual state space model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL https://openreview.net/forum?id=ZgtLQQR1K7.

Zhizhe Liu, Zhenfeng Zhu, Shuai Zheng, Yawei Zhao, Kunlun He, and Yao Zhao. From observation to concept: A flexible multi-view paradigm for medical report generation. *IEEE Transactions on Multimedia*, 26:5987–5995, 2024c. doi: 10.1109/TMM.2023.3342691.

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

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1500–1519, 2020. doi: 10.18653/v1/2020.emnlp-main.117.

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

Jiamei Sun, Pengbo Wang, Ke Zhang, Xiangyu Tan, Zhenqi Fu, and Yan Yang. CyclicAligner: Knowledge-enhanced cyclical alignment for chest X-Ray report generation. In *2025 IEEE International Conference on Bioinformatics and Biomedicine*, pp. 5704–5711, 2025. doi: 10.1109/BIBM66473.2025.11356622.

Quan Tang, Liming Xu, Yongheng Wang, Bochuan Zheng, Jiancheng Lv, Xianhua Zeng, and Weisheng Li. Dual-modality visual feature flow for medical report generation. *Medical Image Analysis*, 101:103413, 2025.

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.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156–3164, 2015.

Fuying Wang, Shenghui Du, and Lequan Yu. HERGen: Elevating radiology report generation with longitudinal data. In *European Conference on Computer Vision*, pp. 183–200, 2024.

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

Xiao Wang, Yuehang Li, Fuling Wang, Shiao Wang, Chuanfu Li, and Bo Jiang. R2GenCSR: Mining contextual and residual information for llms-based radiology report generation. *IEEE Journal of Biomedical and Health Informatics*, 30(9):7667–7680, 2026a. doi: 10.1109/JBHI.2026.3669539.

Xiao Wang, Fuling Wang, Haowen Wang, Bo Jiang, Chuanfu Li, Yaowei Wang, Yonghong Tian, and Jin Tang. Activating associative disease-aware vision token memory for LLM-based X-Ray report generation. *IEEE Transactions on Medical Imaging*, 45(2):583–595, 2026b. doi: 10.1109/TMI.2025.3603416.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. R2GenGPT: Radiology report generation with frozen LLMs. *Meta-Radiology*, 1(3):100033, 2023. doi: 10.1016/j.metrad.2023.100033.

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*, volume 40, pp. 33953–33961, 2026. doi: 10.1609/aaai.v40i40.40688.

An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y. Chang, Amilcare Gentili, and Chun-Nan Hsu. Radbert: Adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, 4(4):e210258, 2022. doi: 10.1148/ryai.210258.

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*, volume 36, pp. 2982–2990, 2022.

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

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.

Andriy Zapechelnyuk. Eliciting information from a committee. *Journal of Economic Theory*, 148 (5):2049–2067, 2013.

Erjian Zhang, Yatong Hao, Liejun Wang, and Zhiqing Guo. The double dilemma in multi-task radiology report generation: A gradient dynamics analysis and solution. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=T7y2wavrFM.

A Appendix

A.1 Reproducibility Details

A.1.1 Module Hyperparameters

Table 4 summarizes the hyperparameters used in the reported experiments.

Table 4: Hyperparameter configuration of DelibRRG.

Component	Hyperparameter	Value
DSSC	Number of clinical prototypes K	128
	Visual neighborhood size k_v	8
	Prototype neighborhood size k_p	8
	Calibration strength λ	4.0
	Calibration rounds R	1
	Retrieval scale τ	learnable; initialized to 10.0
	Semantic warm-up T_w	3000 steps
Clinical Objective	λ_{CLS}	0.02
	λ_{ITC}	0.02
	λ_{Δ}	0.10
	Residual target r_0	0.40
FDCR	Bottleneck dimension D_b	256
	Cross-attention heads	4
	Gate-bias initialization	-3.0
	Cross-examination LR multiplier	$5\times$
	$\beta_{\mathrm{ortho}}^{\max}$	0.015
CFAB	Clinical bargaining power α_c	0.75
	Generation bargaining power α_g	0.25
	Clinical disagreement point d_c	0.10
	Generation disagreement point d_g	0
	EMA coefficient γ	0.95
	Minimum scale $s_{\min}$	1.0
	Maximum scale $s_{\max}$	20.0
	Numerical floor ϵ	10^{-6}

A.1.2 Prototype Initialization

Training-set *Findings* are divided into sentence-level units using periods and semicolons, with fragments shorter than two words removed. Sentence representations are obtained by attention-mask-aware mean pooling. Mini-Batch K-Means uses batch size 8192, five initializations, a maximum of 200 iterations, and random seed 0. No sentence subsampling is applied. The text encoding batch size is 512 with a maximum sequence length of 64 tokens. The resulting 128 prototype centers are represented by 768-dimensional embeddings.

A.1.3 Training and Decoding Configuration

Table 5 summarizes the optimization, scheduling, sequence-length, and decoding settings used in our experiments.

A.2 End-to-End Training Procedure

Algorithm 1 summarizes one training iteration of DelibRRG.

A.2.1 Ablation Configuration

The baseline is trained with $\mathcal{L}_g$ only. For configurations that include DSSC but exclude FDCR, $\mathcal{L}_c$ is computed from the DSSC-calibrated representation. When FDCR is enabled, $\mathcal{L}_c$ is computed from

Table 5: Training and decoding configuration.

Category	Setting	Value
Optimization	Base learning rate	1×10^{-4}
	Minimum learning rate	1×10^{-6}
	Weight decay	0.05
	AdamW (β_1, β_2)	$(0.9, 0.999)$
	LR warm-up	500 steps
	Training epochs	20
	Batch size	16
	Gradient accumulation steps	2
	Precision	`bf16-mixed`
	Random seed	42
Scheduling	DSSC semantic warm-up	3000 steps
	FDCR orthogonality	0 for $e < 2$
	FDCR annealing period	4 epochs
Sequence	Training target length	100 tokens
	RadBERT report length	160 tokens
	Prototype sentence length	64 tokens
Decoding	Beam size	3
	Sampling	disabled
	Minimum new tokens	80
	Maximum new tokens	120
	Repetition penalty	2.0
	Length penalty	2.0

Algorithm 1 Training procedure of DelibRRG

Require: Radiographs I, report tokens Y, clinical labels Y^{ce}, report representation T
Require: trainable parameters θ, including the clinical prototypes P

$V \leftarrow \frac{1}{M}\sum_{m=1}^{M} f_v(I^{(m)})$
$V_{\mathrm{ref}} \leftarrow \mathrm{DSSC}(V, P)$
$V_s^\star, V_c^\star, \mathcal{L}_{\mathrm{ortho}} \leftarrow \mathrm{FDCR}(V, V_{\mathrm{ref}})$
$V_{\mathrm{gen}} \leftarrow V_s^\star$
$\mathcal{L}_g \leftarrow \mathrm{ARLoss}\big(\mathrm{LLaMA}(\Pi(V_{\mathrm{gen}})), Y\big)$
$(\mathcal{L}_{\mathrm{CLS}}, \mathcal{L}_{\mathrm{ITC}}, \mathcal{L}_{\Delta}) \leftarrow \mathrm{ClinicalReview}(V_c^\star, P, Y^{\mathrm{ce}}, T)$
$\mathcal{L}_c \leftarrow \lambda_{\mathrm{CLS}}\mathcal{L}_{\mathrm{CLS}} + \lambda_{\mathrm{ITC}}\mathcal{L}_{\mathrm{ITC}} + \lambda_{\Delta}\mathcal{L}_{\Delta}$
$u_c \leftarrow \exp[-\,\mathrm{sg}(\mathcal{L}_c)]$
$u_g \leftarrow \exp[-\,\mathrm{sg}(\mathcal{L}_g)]$
Update $\bar{u}_c$ and $\bar{u}_g$ with EMA coefficient γ
$q_c \leftarrow \max(\bar{u}_c - d_c, \epsilon)$
$q_g \leftarrow \max(\bar{u}_g, \epsilon)$
$D \leftarrow \alpha_c/q_c + \alpha_g/q_g$
$w_c \leftarrow 2(\alpha_c/q_c)/D$
$w_g \leftarrow 2(\alpha_g/q_g)/D$
Update $\bar{\mathcal{L}}_c$ and $\bar{\mathcal{L}}_g$ with EMA coefficient γ
$s_c \leftarrow \mathrm{clip}\big(\bar{\mathcal{L}}_g / \max(\bar{\mathcal{L}}_c, \epsilon), s_{\min}, s_{\max}\big)$
$\mathcal{L} \leftarrow w_g\mathcal{L}_g + w_c s_c \mathcal{L}_c + \beta_{\mathrm{ortho}}(e)\mathcal{L}_{\mathrm{ortho}}$
Update θ (including P) by backpropagating $\mathcal{L}$

the reviewed clinical representation. For non-baseline configurations in which CFAB is disabled, the active objectives are combined as

$$\mathcal{L} = \mathcal{L}_g + \mathcal{L}_c + \beta_{\mathrm{ortho}}(e)\mathcal{L}_{\mathrm{ortho}}, \tag{18}$$

where the orthogonality term is included only when FDCR is active. The coefficients inside $\mathcal{L}_c$ remain fixed across configurations.

Table 6: Independent clinical factuality evaluation on MIMIC-CXR using RadGraph.

Configuration	$F_1^{(s)}\uparrow$	$F_1^{(p)}\uparrow$	$F_1^{(c)}\uparrow$	$P_{\text{ent}}\uparrow$	$R_{\text{ent}}\uparrow$	$F_1^{\text{ent}}\uparrow$
Baseline	0.2682	0.2400	0.1761	0.2438	0.1840	0.2097
DSSC	0.2768	0.2451	0.1773	0.2548	0.2005	0.2244
DSSC + FDCR	**0.2886**	0.2557	**0.1881**	0.2558	**0.2226**	0.2380
DSSC + CFAB	0.2773	0.2431	0.1761	0.2457	0.2006	0.2209
DelibRRG	0.2882	**0.2563**	0.1868	**0.2599**	0.2204	**0.2386**

Table 7: Clinical finding preservation analysis on the MIMIC-CXR test set. Values are mean findings per report.

Configuration	**Preserved**↑	**Missed**↓	**Omission**↓	**Contradiction**↓	**Uncertain**↓
Baseline	0.700	1.436	1.066	0.267	0.103
DSSC	0.806	1.331	0.975	0.234	0.122
DSSC + FDCR	0.904	1.233	**0.908**	0.231	0.094
DSSC + CFAB	0.863	1.274	0.955	0.216	0.103
DelibRRG	**0.930**	**1.207**	0.910	**0.207**	**0.090**

A.3 Supplementary Experiments

The following analyses reuse the evaluated checkpoints and their saved outputs; no additional model training is performed.

A.3.1 Independent Clinical Factuality Evaluation

We evaluate all 3,858 MIMIC-CXR test reports using the official RadGraph model (Jain et al., 2021) (v0.1.18), which is not used during training.

As shown in Table 6, compared with the baseline, DelibRRG improves Simple F1 from 0.2682 to 0.2882, Partial F1 from 0.2400 to 0.2563, Complete F1 from 0.1761 to 0.1868, and Entity F1 from 0.2097 to 0.2386. The gains therefore extend to an independent relation-based clinical factuality metric that is not used during training.

A.3.2 Clinical Finding Preservation Analysis

We decompose the structured CE labels into preserved findings and target-conditioned clinical errors. Table 7 reports the resulting finding-level preservation and error statistics on the MIMIC-CXR test set.

DelibRRG increases preserved findings from 0.700 to 0.930 ($+32.9\%$), while missed findings decrease from 1.436 to 1.207 (-15.9%) and explicit contradictions decrease from 0.267 to 0.207 (-22.5%).

Figure 4 visualizes the reference-relative clinical finding decomposition. Preserved findings are shown together with missed findings and explicit contradictions; the latter is a diagnostic subset of missed findings.

To further identify which clinical findings benefit from the complete framework, Table 8 reports the false-negative counts for the conditions showing the largest reductions relative to the baseline.

The largest reductions are observed for cardiomegaly and lung opacity, where DelibRRG reduces false negatives from 895 to 487 and from 1315 to 1096, respectively. Improvements are also observed for edema, pleural effusion, pneumonia, pneumothorax, and fracture, while the reduction for atelectasis is smaller than that achieved by some intermediate configurations.

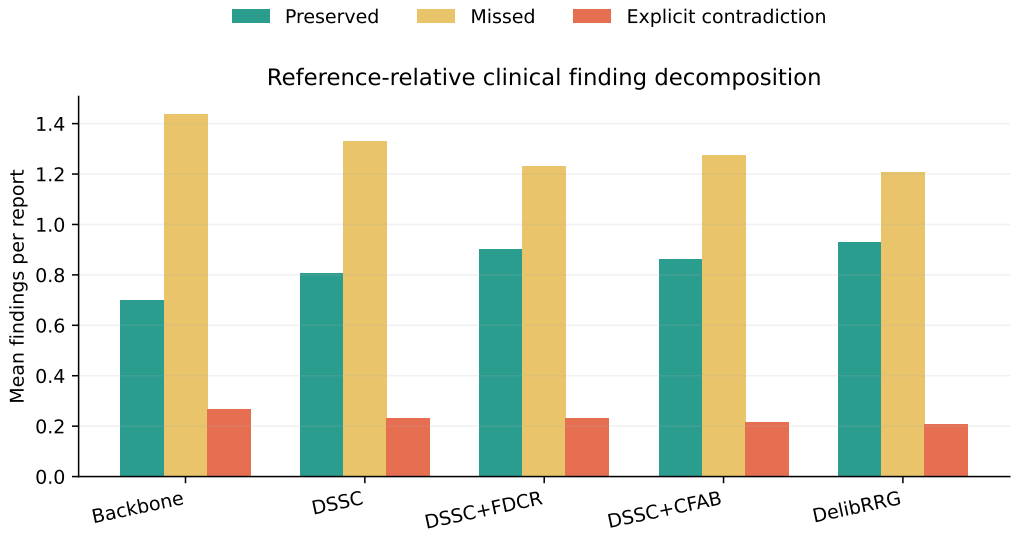


Figure 4: Reference-relative clinical finding decomposition on the MIMIC-CXR test set. Values are mean findings per report.

Table 8: False-negative findings across configurations for the conditions with the largest reduction in the complete model. Abbreviations: Car., Cardiomegaly; Lun. Op., Lung Opacity; Ate., Atelectasis; Ede., Edema; Ple. Eff., Pleural Effusion; Pna., Pneumonia; Ptx., Pneumothorax; Fra., Fracture. FN denotes false negatives; lower values are better.

Configuration	**Car.↓**	**Lun. Op.↓**	**Ate.↓**	**Ede.↓**	**Ple. Eff.↓**	**Pna.↓**	**Ptx.↓**	**Fra.↓**
Baseline FN	895	1315	669	436	424	169	69	164
DSSC FN	640	1167	607	488	444	159	69	164
DSSC + FDCR FN	584	1195	462	366	424	157	60	159
DSSC + CFAB FN	636	1171	**448**	441	448	159	59	158
DelibRRG FN	**487**	**1096**	582	**358**	**366**	**156**	**57**	**156**

A.3.3 CFAB Optimization Dynamics

Figure 5 reports the adaptive objective weights and weighted objective contributions recorded every 500 training batches.

Across the recorded trajectory,

$$\text{mean}(w_c) = 1.052, \qquad \Pr(w_c > w_g) = 87.3\%. \tag{19}$$

The observed training behavior shows that CFAB produces an empirical clinical-side preference while retaining dynamic adaptation between the two objectives.

A.4 Supplementary Case Analyses

Figure 6 presents two additional cases, including the chest radiographs, reference reports, and reports generated by the baseline and DelibRRG. In Case 1, the reference report describes mild bibasilar atelectasis. DelibRRG describes streaky bibasilar opacities and relates them to atelectasis, whereas the baseline report does not mention this finding. In Case 2, the reference report describes bibasilar opacities and lists aspiration or atelectasis as possible explanations. DelibRRG reports the bibasilar opacities and retains both possibilities, while the baseline report does not describe this abnormality. The colors distinguish clinical evidence in the reference, generated content consistent with that evidence, orange marks partial or attribute-level matches, contradictory statements, and routine or unverifiable content.

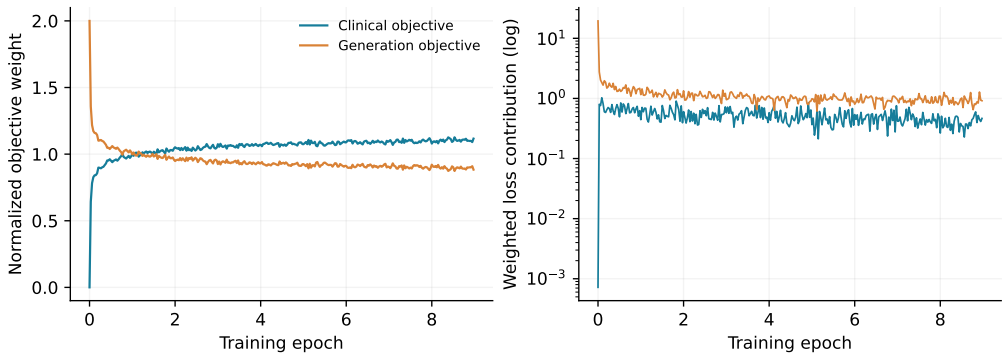


Figure 5: CFAB optimization dynamics during training. The two panels show the adaptive objective weights and the corresponding weighted objective contributions.

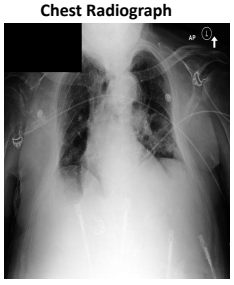
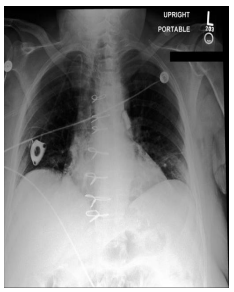
Chest Radiograph	Ground Truth	Baseline	DelibRRG (Ours)
	There are lower lung volumes with secondary mild widening of cardiomediastinal silhouette. There is no pleural effusion, pneumothorax or focal lung consolidation. There are bibasilar opacities which are better seen on the subsequent CT abd, may represent aspiration or atelectasis.	ap portable upright view of the chest . lung volumes are low limiting assessment . there is no focal consolidation pleural effusion or pneumothorax . cardiomediastinal silhouette is within normal limits . atherosclerotic calcifications noted at the aortic arch . degenerative changes seen in the thoracic spine . surgical clips project over the left upper abdomen .	ap portable upright view of the chest was obtained . there are low lung volumes which accentuate the bronchovascular markings . bibasilar opacities may be due to atelectasis or aspiration . no large pleural effusions are seen . the cardiac silhouette is difficult to assess given low lung volumes but appears enlarged . mediastinal contour is unremarkable . bony structures are intact .
	AP portable upright view of the chest. Right chest wall Port-A-Cath again noted with catheter tip extending to the upper SVC region. Midline sternotomy wires are again noted. There is a calcified ovoid structure projecting over the mediastinum likely a calcified lymph node. There is mild basilar atelectasis noted bilaterally. No focal consolidation concerning for pneumonia. No large effusion or pneumothorax is seen. Cardiomediastinal silhouette is stable. Bony structures are intact.	a right-sided port-a-cath terminates at the cavoatrial junction . lung volumes are low . there is no focal consolidation pleural effusion or pneumothorax . the cardiomediastinal silhouette is within normal limits . median sternotomy wires and mediastinal clips are intact and unchanged in appearance .	ap portable upright view of the chest . right-sided port-a-cath is again seen with catheter tip in the region of the mid svc . there are low lung volumes which accentuate bronchovascular markings . streaky bibasilar opacities likely reflect atelectasis . no large pleural effusion or pneumothorax . cardiomediastinal silhouette is stable . median sternotomy wires and mediastinal clips are unchanged .

Figure 6: Chest radiographs and report comparisons for two supplementary cases. Each row shows the chest radiograph, reference report, baseline report, and DelibRRG report. Colors distinguish clinical evidence in the reference (blue), generated content consistent with the evidence (green), orange marks partial or attribute-level matches (orange), contradictory statements (red), and routine or unverifiable content (dark gray).