

2026 한국차세대컴퓨팅학회 춘계학술대회

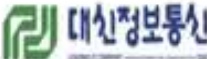
장 소 : 조선대학교 IT융합대학 3층

일 시 : 2026. 5. 7(목) 13:00 ~ 5. 9(토) 11:30

주최:  **한국차세대컴퓨팅학회**
Korean Institute of Next Generation Computing

주관:  **조선대학교 SW중심대학사업단**

후원:

LQB-RGen: 효율적인 방사선 이미지 보고서 생성을 위한 학습 가능한 쿼리 보틀넥

LQB-RGen: Learnable Query Bottleneck for Efficient Radiology Image Report Generation

사오마빈, 다오뒤프엉, 양형정*

Mavin Sao, Duy-Phuong Dao, Hyung-Jeong Yang*

Dept. of Artificial Intelligence Convergence, Chonnam National University

{mavinsao, duyphuongcri, hjyang*}@jnu.ac.kr

Abstract

Radiology report generation aims to automatically translate chest X-ray findings into coherent textual reports, reducing radiologists' workload. However, existing Large Language Model (LLM)-based approaches pass all visual tokens directly to the LLM, resulting in redundant information and increased computational cost. We propose LQB-RGen, a framework introducing a Learnable Query Bottleneck (LQB) that compresses visual features into a compact set of query tokens, trained end-to-end with only the generation loss without multi-stage pretraining. Our contributions are: (1) an LQB module achieving state-of-the-art performance while passing only 16 tokens to the LLM; (2) a query-budget study showing smaller budgets outperform larger ones, suggesting compression acts as implicit regularization; and (3) a decoding sensitivity analysis highlighting an underreported factor in LLM-based R2Gen benchmarking. Experiments on IU-Xray demonstrate state-of-the-art results with a simpler and more efficient pipeline.

Keyword: Learnable Query Bottleneck, Visual Feature Compression, Radiology Report Generation, Large Language Models

1. Introduction

Chest X-ray (CXR) images are essential diagnostic tools, yet their interpretation demands considerable expertise. Automated radiology report generation (R2Gen) has emerged as a promising solution, but current LLM-based systems face a key limitation: visual projectors pass all encoder tokens directly to the LLM without compression, resulting in redundant information and increased computational cost.

While Q-Former in BLIP-2 [7] addresses this

through learnable queries, it requires costly multi-stage pretraining with contrastive and matching objectives. To address these limitations, we propose LQB-RGen, introducing a Learnable Query Bottleneck (LQB) that compresses visual features into a compact token set through cross-attention, trained end-to-end with only the generation loss.

2. Related Work

Radiology Report Generation: R2Gen has evolved

*Corresponding Author

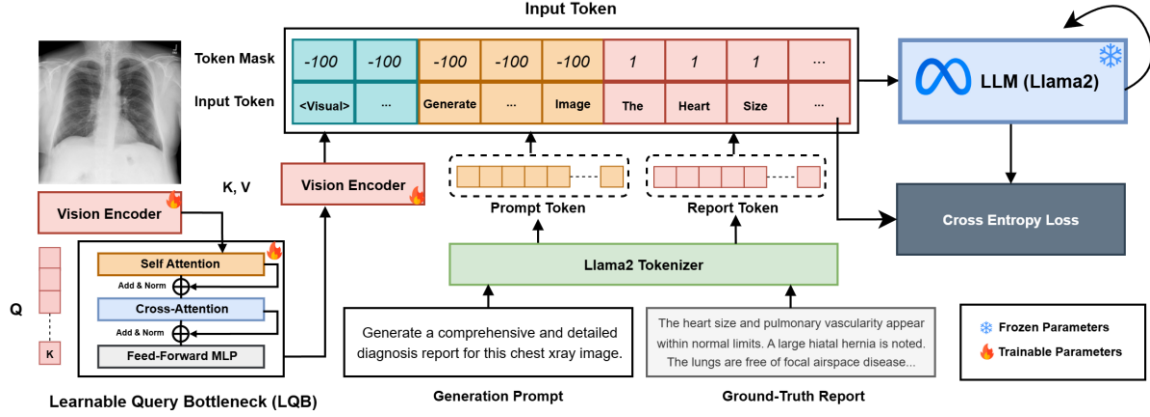


Fig. 1. Overview of LQB-RGen framework

from CNN-LSTM architectures [4] to transformer-based models [1],[2],[5], with several approaches incorporating external knowledge such as knowledge graphs [8] to improve clinical accuracy. However, constructing specialized knowledge bases remains resource-intensive and demands domain expertise.

LLM-based R2Gen Methods: R2GenGPT [6] pioneered frozen LLM usage with a linear projector for visual feature mapping. Other works such as LLM-CXR [9] fine-tune LLMs to unify multiple tasks. However, these methods pass all visual tokens without compression. Our work introduces an information bottleneck that compresses visual features into a minimal token set trained end-to-end.

3. Methodology

As illustrated in Fig. 1, LQB-RGen comprises a visual encoder, LQB module, linear projection, and frozen LLM. We use Swin-Transformer-Base (fully trained) to extract visual features $Z \in \mathbb{R}^{(N_p \times d_v)}$, and LLaMA-2-7B-Chat as the frozen decoder.

The LQB compresses visual features through cross-attention with K learnable query tokens $L \in \mathbb{R}^{(K \times d_v)}$, where $K \ll N_p$. These queries are randomly initialized and optimized during training. Each of the N transformer blocks consists of three sub-layers: (i) multi-head self-attention among the K queries,

enabling inter-query coordination to reduce redundancy; (ii) multi-head cross-attention where queries attend to the full visual features Z (queries as Q , visual features as K and V); and (iii) a feed-forward network. Each sub-layer applies layer normalization and residual connections. The output $Q \in \mathbb{R}^{(K \times d_v)}$ is projected to the LLM's embedding space:

$$H = W_p \cdot Q, \quad \text{where } H \in \mathbb{R}^{K \times d_t} \quad (1)$$

H is concatenated with prompt embeddings and fed to the frozen LLM. Following R2GenGPT [6], we use the prompt: "Generate a comprehensive and detailed diagnosis report for this chest xray image." Unlike BLIP-2's Q-Former, our LQB is trained end-to-end with only:

$$\mathcal{L} = - \sum_{k=1}^L \log p_{\theta}(t_k | l, T_p, T_{<k}) \quad (2)$$

where θ denotes trainable parameters (visual encoder, LQB, projection layer), T_p is the prompt, and $T_{<k}$ denotes preceding tokens.

4. Experiments

4.1. Dataset and Metrics

We evaluate on IU-Xray [10], containing 3,955 reports and 7,470 CXR images. We follow the same data split provided by R2Gen [1] for fair comparison. We adopt standard NLG metrics: BLEU measures n-gram precision; ROUGE-L (RG-L) computes longest

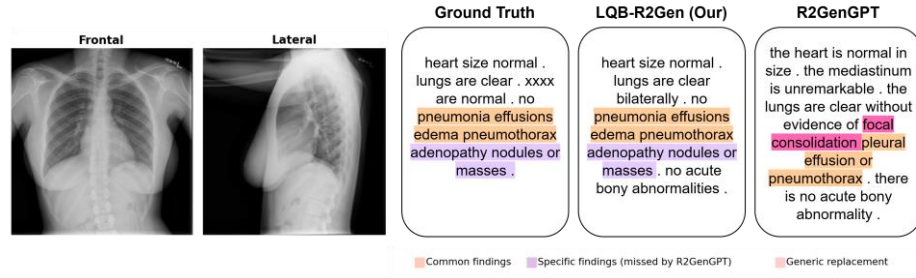


Fig. 2. Qualitative comparison on IU-Xray (Study CXR136). Key medical terms are highlighted: orange for common negative findings, purple for specific findings.

Table 1. Comparison on IU-Xray dataset. Best results are bolded.

Method	BL-4	RG-L	MTR	CDr
R2Gen [1]	0.165	0.371	0.187	–
R2GenCMN [2]	0.170	0.375	0.191	–
MSAT [3]	0.171	0.372	0.190	0.394
METransf. [5]	0.172	0.380	0.192	0.435
R2GenGPT [6]	0.173	0.377	0.211	0.438
LQB-RGen (Ours)	0.199	0.397	0.220	0.640

Table 2. Query-budget efficiency on IU-Xray. K = number of query tokens.

K	BL-4	RG-L	MTR	CDr
16	0.199	0.397	0.220	0.640
24	0.192	0.378	0.205	0.543
32	0.182	0.370	0.198	0.584
48	0.198	0.381	0.206	0.605
64	0.185	0.375	0.202	0.529

Table 3. Decoding sensitivity analysis on IU-Xray (K=16, beam=3). RP = Repetition Penalty, LP = Length Penalty. † = default config.

Config (RP, LP)	BL-4	RG-L	MTR	CDr
rp=1.0, lp=0.8	0.201	0.397	0.217	0.647
rp=1.0, lp=1.0	0.196	0.395	0.218	0.634
rp=1.0, lp=1.2	0.191	0.394	0.218	0.611
rp=1.2, lp=0.8	0.202	0.396	0.217	0.666
rp=1.2, lp=1.0	0.200	0.395	0.218	0.656
rp=1.2, lp=1.2	0.199	0.397	0.220	0.640
†				

common subsequence-based F-score capturing structural similarity; METEOR (MTR) evaluates unigram alignment with synonym and stemming matching; and CIDEr (CDr) computes TF-IDF weighted n-gram cosine similarity designed for descriptive text evaluation.

4.2. Implementation Details

We use Swin-Transformer-Base (fully trained) and frozen LLaMA-2-7B-Chat. The LQB uses 2 transformer layers with 8 attention heads. Training was conducted on an NVIDIA RTX A6000 for 15 epochs with batch size 6, learning rate $1e-4$, AdamW optimizer, and cosine annealing scheduler. For inference, we used beam search with size 3.

4.3. Experiment Results

Table 1 compares LQB-RGen against state-of-the-art methods. With only K=16 query tokens, our method achieves the highest BLEU-4 (0.199) and CIDEr (0.640), outperforming R2GenGPT [6] across all metrics while passing only 16 tokens versus 49.

4.3.1. Qualitative Analysis

Fig. 2 shows a comparison on study CXR136. The ground truth explicitly lists negative findings including adenopathy, nodules, and masses. LQB-RGen reproduces these specific findings, while R2GenGPT omits them in favor of generic phrases like "focal consolidation." This illustrates that the LQB's compressed

representation retains clinically specific information.

4.3.2. Query-Budget Analysis

Table 2 shows that K=16 achieves the best BLEU-4 (0.199) and CIDEr (0.640), outperforming all larger budgets. We attribute this to two factors: first, a smaller query budget forces each query to capture maximally distinct and clinically relevant features, whereas larger K allows queries to encode redundant or noisy visual details (e.g., background texture) that do not contribute to report quality. Second, the frozen LLM has fixed capacity for processing visual context—fewer, higher-quality tokens improve the signal-to-noise ratio in the LLM's attention, allowing it to focus on diagnostically meaningful information. The full token baseline (R2GenGPT, 49 tokens) scores 0.173/0.438, confirming compression substantially benefits performance.

4.3.3. Decoding Sensitivity Analysis

Table 3 presents the effect of decoding hyperparameters on the K=16 model. BLEU-4 and CIDEr are highly sensitive, with gaps of 0.011 and 0.055 respectively, while ROUGE-L and METEOR remain stable ($\Delta \leq 0.003$). Tuning decoding parameters improves performance from the default (0.199 BLEU-4, 0.640 CIDEr) to 0.202 BLEU-4 and 0.666 CIDEr with (rp=1.2, lp=0.8). This highlights that decoding choices disproportionately affect certain metrics and should be disclosed in LLM-based R2Gen benchmarks.

5. Conclusion

We proposed LQB-RGen, introducing a Learnable Query Bottleneck for efficient radiology report generation. Our LQB

compressed visual features end-to-end without multi-stage pretraining, achieving state-of-the-art on IU-Xray (0.199 BLEU-4, 0.640 CIDEr) with only 16 tokens, which further improved to 0.202 BLEU-4 and 0.666 CIDEr with optimized decoding.

Our analysis revealed compression acts as implicit regularization, and decoding hyperparameters significantly impact benchmarking. Future work includes evaluation on MIMIC-CXR and incorporating clinical efficacy metrics.

Acknowledgement

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-지역지능화혁신인재양성사업의 지원을 받아 수행된 연구임(IITP-2026-RS-2022-00156287). 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 인공지능융합혁신인재양성사업 연구 결과로 수행되었음(IITP-2023-RS-2023-00256629). 본 연구 성과물은 2025년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(RS-2025-25425780)

References

- [1] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer," in EMNLP, 2020.
- [2] Z. Chen, Y. Shen, Y. Song, and X. Wan, "Cross-modal memory networks for radiology report generation," in ACL-IJCNLP, 2021.
- [3] Z. Wang, M. Tang, L. Wang, X. Li, and L. Zhou, "A medical semantic-assisted transformer for radiographic report generation," in MICCAI, 2022.

- [4] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," in ACL, 2018.
- [5] Z. Wang, L. Liu, L. Wang, and L. Zhou, "METransformer: Radiology report generation by transformer with multiple learnable expert tokens," in CVPR, 2023.
- [6] Z. Wang, L. Liu, L. Wang, and L. Zhou, "R2GenGPT: Radiology report generation with frozen LLMs," Meta-Radiology, 2023.
- [7] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in ICML, 2023.
- [8] Y. Zhang, X. Wang, Z. Xu, Q. Yu, and D. Xu, "When radiology report generation meets knowledge graph," in AAAI, 2020.
- [9] S. Lee, W.J. Kim, J. Chang, and J.C. Ye, "LLM-CXR: Instruction-finetuned LLM for CXR image understanding and generation," in ICLR, 2024.
- [10] D. Demner-Fushman et al., "Preparing a collection of radiology examinations for distribution and retrieval," JAMIA, 2016.