SKIM: Semantic Knowledge Infused Modeling for Medical Report Generation

Hyojeong Lee
Yonsei University
Seoul, Republic of Korea
hyojoy@yonsei.ac.kr

Youngwan Jo
Yonsei University
Seoul, Republic of Korea
jyy1551@yonsei.ac.kr

Inpyo Hong
Yonsei University
Seoul, Republic of Korea
hip9863@yonsei.ac.kr

Sanghyun Park*
Yonsei University
Seoul, Republic of Korea
sanghyun@yonsei.ac.kr

Abstract

The automatic generation of medical reports has become an essential technology for optimizing diagnostic processes and mitigating the increasing workload of radiologists. However, existing methods face a fundamental trade-off: processing dense visual features for completeness is computationally prohibitive, while aggressive compression often sacrifices the fine-grained visual information necessary for accurate report generation. To address this challenge, we propose a framework that infuses report-level textual knowledge into compact visual representations. Our approach first extracts variable-length visual tokens from input images using a BEiT encoder. A Fixed-size Visual Summarization (FVS) module then compresses these tokens into a small set of representative features through learnable attention pooling. Subsequently, a Cross-Modal knowledge infusion (CMKI) module progressively enriches these visual summaries with report-level semantic knowledge via gated blending and residual connections. This process provides the Large Language Model (LLM) with highly distilled, context-aware representations that enable accurate and coherent report generation. We demonstrate our framework's effectiveness on two widely-adopted benchmark datasets, IU-Xray and MIMIC-CXR. Our approach improves BLEU-4 and ROUGE-L over recent baselines. These results indicate that semantic-guided compression can be an efficient alternative to uniform visual processing for medical VLMs.

CCS Concepts

• **Applied computing** → *Health informatics*; • **Computing methodologies** → **Artificial intelligence**.

Keywords

Medical report generation, Vision language models, Multi-modal, Slot summarization, Knowledge infusion

ACM Reference Format:
Hyojeong Lee, Youngwan Jo, Inpyo Hong, and Sanghyun Park. 2026. SKIM: Semantic Knowledge Infused Modeling for Medical Report Generation. In *Proceedings of The 41st ACM/SIGAPP Symposium on Applied Computing (SAC'26).* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Generating radiology reports is an essential yet challenging task where simple image feature extraction is insufficient. The task also requires aligning those features with report-level semantics and clinical terminology. As imaging volumes surge, automated systems can reduce radiologists' workload and support clinical decision-making, complementing advances in disease detection and prognosis [22, 1]. Consequently, it has become a focal topic at the interface of artificial intelligence and clinical practice.

The field has progressed from general image captioning [24] to domain-adapted generators such as R2Gen [5], reflecting a shift toward clinically grounded modeling. Over time, language decoders evolved from recurrent networks to modern large language models (LLMs) with stronger priors, and visual encoders advanced from convolutional neural networks (CNNs) to Transformer backbones. Since R2Gen, most approaches have followed two principal lines: adapting large vision–language models [28, 17, 30, 25] and designing specialized architectures with task-specific modules or knowledge infusion [4, 7, 32, 26, 12, 20, 31].

Building on these advances, we propose a specialized interface that functionally decouples expression from content. The LLM governs linguistic form ("how to say it"), whereas the interface constrains clinical content by distilling and conditioning visual evidence ("what to say"). This approach strengthens the visual-textual connection by reliably capturing shared information without relying on external knowledge bases.

Despite promising results with pretrained Large Vision-Language Models(LVLMs), practical deployment is hampered by two coupled issues. First, visual processing is often inefficient, as uniformly handling all patch tokens in high-resolution studies inflates computation and enforces a flat representation misaligned with expert reading patterns. Indeed, recent work on VLM scaling laws has established that for a fixed inference budget, the optimal strategy is not to shrink the language model but to drastically reduce the number of visual tokens[15]. Second, cross-modal integration is frequently shallow, biasing models toward safe, generic statements

[*] Corresponding author

SAC'26, Thessaloniki, Greece

ACM ISBN 979-X-XXXX-XXXX-X/26/03
https://doi.org/XXXXXXX.XXXXXXX

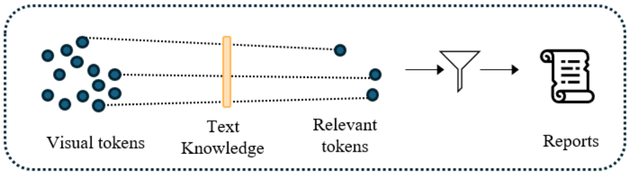


Figure 1: Conceptual overview. Dense visual tokens are compressed into a concise, fixed-size summary and then enriched with report-level semantics, reducing computational load while enhancing semantic fidelity.

and hindering fine-grained, context-dependent semantics. This reveals a gap wherein efficiency and semantic fidelity are typically optimized in isolation, motivating a unified approach that reduces visual complexity while preserving essential clinical meaning.

To implement this approach, we introduce a summarize–then enrich framework as shown in Figure 1. First, a Fixed-size Visual Summarization (FVS) module distills dense patch tokens into a compact, semantically meaningful summary. Second, a Cross-Modal Knowledge Infusion (CMKI) module—inspired by AdaIN-style feature modulation—adjusts the summarized representation using report-level semantic context. This framework provides an efficient and semantically faithful interface for multimodal report generation.

- Fixed-size Visual Summarization (FVS) compresses dense visual tokens into a compact summary that reduces the LLM token budget and approaches linear-in-tokens aggregation. The design reflects radiologists' global-to-focal reading strategy.
- A Cross-modal Knowledge Infusion(CMKI) progressively enriches the summary with semantic context—using AdaIN-style gated residual blending—without external knowledge bases.
- We achieve state-of-the-art results on the IU-Xray and MIMIC-CXR benchmarks with exceptional parameter efficiency.

2 Related Work

2.1 Visual Information Compression

To mitigate this computational burden, one major branch of research has focused on adaptive token selection. These methods dynamically identify and process only a subset of the most informative tokens often termed expert tokens, at various layers of the model. By focusing computational resources on semantically crucial patches while down-weighting or pruning redundant ones, models like DynamicViT[19] and others tailored for specific domains[27] can significantly reduce inference costs. Further exemplifying this trend towards task-specific information processing, specialized fields like medical imaging are also developing methods to guide LLMs with highly structured visual inputs. For instance, Diff-RRG [33] moves beyond simple token representation by explicitly computing longitudinal disease-wise patch differences to guide radiology report generation. However, this approach operates on the principle of selection, carrying an inherent risk of discarding individually less salient tokens that hold vital context for a comprehensive global understanding. This limitation highlights the need for a mechanism that summarizes the entire visual context, rather than selecting a partial subset.

To address this, a more holistic paradigm has emerged that uses a small, fixed set of learnable queries to distill information from the entire patch sequence. This approach, exemplified by Perceiver IO [11]and notably the Q-Former in BLIP-2 [14], shifts the paradigm from token selection to comprehensive information extraction. The Q-Former, specifically, employs a full transformer architecture where a set of learnable queries interacts with visual features through cross-attention, and also with each other through self-attention. However, this intricate, iterative structure introduces substantial architectural overhead. The inclusion of self-attention among queries, while enabling complex inter-query reasoning, creates an indirect and potentially inefficient pathway for visual summarization, as computational resources are spent on query-to-query interactions rather than purely on distilling from the source visual features.

The critical importance of this efficiency-oriented distillation is strongly validated by recent work on VLM scaling laws [15] Their research establishes that the compute-optimal strategy for VLMs involves extreme token compression coupled with the largest possible LLM. Crucially, they also demonstrate that operating in these high-compression regimes achieves strong performance across almost all benchmarks, challenging the conventional notion that a large number of visual tokens is essential for high fidelity. This provides a clear impetus for developing architectures that, unlike Q-Former, are explicitly designed for maximal token reduction.

In contrast to Q-Former's architectural complexity, we introduce our Fixed-size Visual Summarization (FVS) module, designed not merely for compression, but as a high-efficacy semantic interface for LLM. FVS employs a more direct, single-pass refinement process using a sequence of gated cross-attention layers, deliberately omitting query-side self-attention. This design forces the learnable summary slots to exclusively distill the most salient information from the visual patch features provided by BEiT [3] encoder. The core thesis is that the performance of a vision-language model is critically dependent on how visual information is "packaged" for the LLM. By transforming the high-dimensional, spatially redundant visual features into a concise, fixed-length sequence of semantically rich summary tokens, FVS produces a representation that is not only structurally optimal for the LLM's input architecture, but also enables more computationally tractable and effective cross-modal fusion in subsequent layers. This effective interfacing allows the LLM to ground its reasoning in potent, pre-digested visual evidence, significantly enhancing the model's overall performance on complex, multimodal tasks.

2.2 Cross-Modal Fusion and Knowledge Integration

While effective visual information compression is a necessary first step, its utility is fundamentally limited without a sophisticated mechanism to align the resulting representation with the semantic nuances of the textual report. Effectively bridging this semantic gap is a central challenge in medical report generation. Early methods often relied on simple fusion techniques such as concatenation or shallow attention, which lack the capacity to capture complex diagnostic relationships. To address this, more sophisticated strategies have emerged, falling into two main categories.

The first category leverages contrastive learning on large-scale image-text pairs, exemplified by MedCLIP [29]. These models learn powerful, globally-aligned representations. However, their global alignment objective is inherently limited in capturing the fine-grained, localized correspondence required for detailed clinical reports. This approach struggles to differentiate between reports with similar overall topics but subtle, diagnostically critical differences, a gap our work aims to address by directly learning from report-level semantics.

The second category involves the explicit integration of structured medical knowledge, typically through knowledge graphs [8, 9]. By injecting facts about anatomies and diseases, these methods can improve clinical accuracy. Nonetheless, they face significant practical hurdles, as knowledge graphs require extensive, expert-driven manual curation and face challenges in scalability and maintenance.

Our work carves a distinct path that learns to fuse visual information with semantic context derived directly from reports during training, thereby avoiding the limitations of both prior approaches. The architectural inspiration for this mechanism comes from Adaptive Instance Normalization (AdaIN), a technique renowned for feature modulation in style transfer [10]. AdaIN aligns the channel wise mean and variance of a content input x to those of a style input y, as formulated below.

$$\mathrm{AdaIN}(x, y) = \sigma(y)\frac{x - \mu(x)}{\sigma(x)} + \mu(y) \tag{1}$$

While originally used for the intra-modal task of mapping pixel-level statistics between images, its underlying principle is highly compelling for our inter-modal alignment problem. Notably, the AdaIN transformation contains no learnable affine parameters; its modulation is governed purely by the feature statistics of the style input. Adopting this concept allows the semantic context from a given report to directly and dynamically calibrate the visual features. This results in a highly adaptive fusion process where the visual representation is precisely adjusted to align with the specific nuances of the target text, a mechanism we will detail further in the Method section.

3 Method

3.1 Overview

We propose a framework, semantic knowledge infused modeling (SKIM), for medical report generation. The goal of our framework is to generate a clinically accurate medical report from a given set of chest X-ray views. As shown in Figure 2, our framework extracts visual features from each view. These features are then processed by a learnable visual summarization module, which compresses and fuses them into a single, fixed-size representation. Subsequently, a cross-modal knowledge infusion module enriches this visual summary with textual semantics. Finally, a frozen LLM generates the report based on the enriched representation. A key aspect of our approach is that only the intermediate modules are trained, acting as a lightweight and efficient bridge between the large, frozen models.

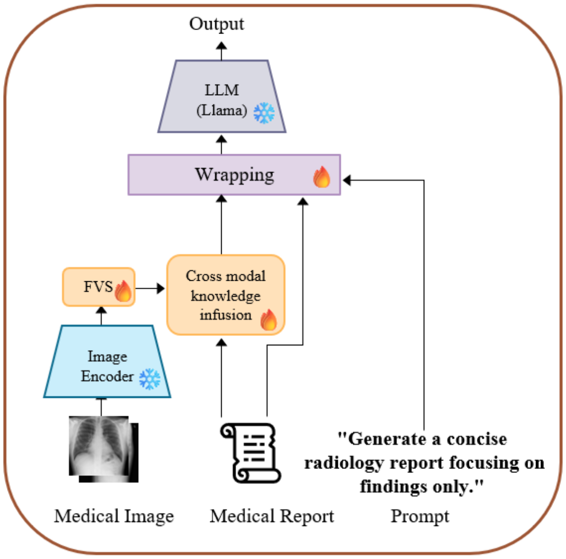


Figure 2: The overall architecture of our proposed model SKIM, which employs frozen foundation models. Fixed-Size Visual Summarization (FVS) and Cross-Modal Knowledge Infusion (CMKI) modules bridge a frozen image encoder and a frozen LLM decoder.

3.2 Visual Feature Extraction

3.2.1 *Visual Encoding and Multi-view Aggregation.*

Our model first extracts patch-level visual features using a frozen BEiT-Large encoder. For a study containing multiple views (e.g., frontal and lateral X-rays), the encoder yields a sequence of patch embeddings $\mathbf{E}_v \in \mathbb{R}^{L \times D_v}$ for each of the V views.

To create a single, holistic representation for the entire study, we aggregate these individual view embeddings into a unified matrix $\mathbf{H}$ by averaging them:

$$\mathbf{H} = \frac{1}{V}\sum_{v=1}^{V}\mathbf{E}_v \tag{2}$$

This aggregation strategy produces a final visual representation $\mathbf{H} \in \mathbb{R}^{L \times D_v}$, where the patch sequence length is $L = 196$ and the feature dimension is $D_v = 1024$. By maintaining a fixed-size representation regardless of the number of views, we ensure a consistent input structure for the subsequent module. This unified representation $\mathbf{H}$ serves as the input to our Fixed-size Visual Summarization (FVS) module.

3.2.2 *Fixed-Size Visual Summarization.*

To distill salient information from $\mathbf{H}$, as shown in Figure 3, we introduce a summarization module with N_s learnable slot embeddings $\mathbf{S}^{(0)} \in \mathbb{R}^{N_s \times D_v}$. The number of slots N_s is a dataset-dependent hyperparameter, set to 32 for IU-Xray and 3 for MIMIC-CXR. These slots are refined over N=2 layers. At each layer n, the slots $\mathbf{S}^{(n-1)}$ attend to the visual information via a standard Multi-Head Attention (MHA) block with 8 heads:

$$\hat{\mathrm{S}}^{(n)} = \mathrm{MHA}(\mathrm{Q} = \mathrm{S}^{(n-1)}, \mathbf{K} = \mathbf{H}^{(n-1)}, \mathbf{V} = \mathbf{H}^{(n-1)}) \tag{3}$$

where $\mathbf{H}^{(0)} = \mathbf{H}$ and $\mathbf{H}^{(n>0)} = \hat{\mathrm{S}}^{(n)}$. For the second layer, we set $\mathbf{H}^{(n)} = \hat{\mathrm{S}}^{(n)}$. This design choice encourages a hierarchical self-refinement of the summary slots, allowing them to focus on the most salient features distilled from the previous layer rather than repeatedly attending to the full, noisy patch token set. The updated slots are then combined with the original slots via a gated residual connection:

$$\mathrm{g}^{(n)} = \sigma(\mathrm{Linear}([\mathrm{S}^{(n-1)}; \hat{\mathrm{S}}^{(n)}])) \tag{4}$$

$$\mathrm{S}^{(n)} = \mathrm{LayerNorm}(\mathrm{g}^{(n)} \odot \hat{\mathrm{S}}^{(n)} + (1 - \mathrm{g}^{(n)}) \odot \mathrm{S}^{(n-1)}) \tag{5}$$

where the Linear block is a single feed-forward layer(FFN). The final visual summary $\mathbf{V}_{summ} = \mathrm{S}^{(N)}$ provides a condensed representation of the input images.

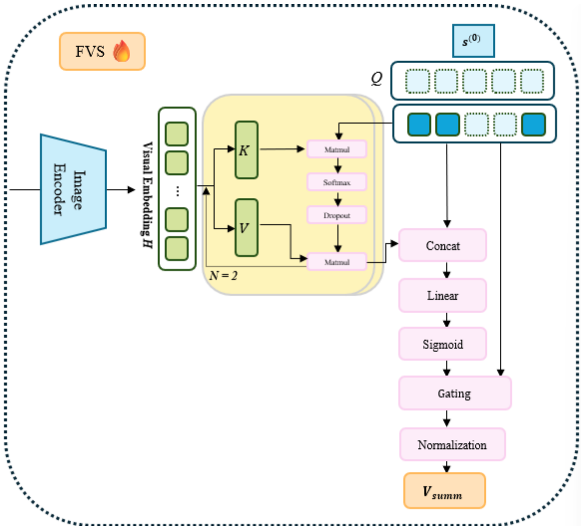


Figure 3: The architecture of Fixed-Size Visual Summarization (FVS) module

3.3 Cross-Modal Knowledge Infusion

To bridge the modality gap, as illustrated in Figure 4, CMKI enriches the condensed visual summary $\mathbf{V}_{summ}$ with high-level semantic knowledge derived from the report text. The design of this module is conceptually inspired by AdaIN, which modulates a content signal using statistics from a style signal. We adapt this principle of dynamic feature modulation for a deep and adaptive alignment between visual and textual modalities.

3.3.1 Knowledge Alignment and Enrichment. First, the visual summary from the previous stage, denoted as $\mathbf{V}_{summ}$, is projected into the LLM's hidden space via a linear layer, resulting in an initial representation $\mathbf{V}_{proj}$. Concurrently, we extract a global semantic vector $\mathbf{T}$ by averaging the embeddings of the report tokens, which is then broadcast to match the shape of $\mathbf{V}_{proj}$.

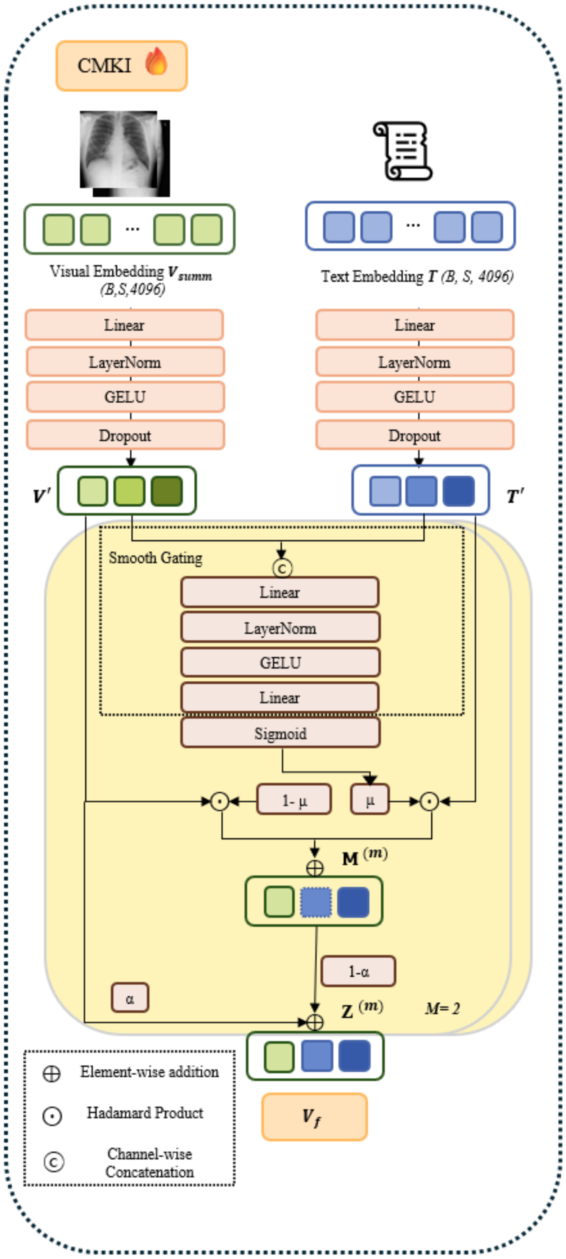


Figure 4: The architecture of Cross-Modal Knowledge Infusion (CMKI) module

This module iteratively enriches the visual features over a multi-layer block (M=2). Let the initial input to this block be $\mathbf{Z}^{(0)} = \mathbf{V}_{proj}$. For each layer m, we first apply transformations to get $\mathbf{V}' = \mathrm{Transform}_v(\mathbf{Z}^{(m-1)})$ and $\mathbf{T}' = \mathrm{Transform}_t(\mathbf{T})$.Each of the transformation blocks is a two-layer MLP with GELU activations. A dynamic gate $\mu^{(m)}$ is then computed from these features using a SmoothGating block, which is a separate two-layer MLP with a final Sigmoid activation.

$$\mu^{(m)} = \text{SmoothGating}([\mathbf{V}'; \mathbf{T}']) \tag{6}$$

$$\mathbf{M}^{(m)} = (1 - \mu^{(m)}) \odot \mathbf{V}' + \mu^{(m)} \odot \mathbf{T}' \tag{7}$$

This operation allows the visual content $\mathbf{V}'$ to be dynamically modulated by the semantic style of the text $\mathbf{T}'$-a principle adapted from AdaIN. The state is then updated via a residual connection controlled by a learnable scalar weight $\alpha^{(m)}$:

$$\mathbf{Z}^{(m)} = (1 - \alpha^{(m)})\mathbf{Z}^{(m-1)} + \alpha^{(m)}\mathbf{M}^{(m)} \tag{8}$$

3.3.2 *Final Representation.* The final representation $\mathbf{V}_f$ is produced by applying a conservative residual connection between the initial projected summary $\mathbf{V}_{proj}$ and the final enriched state $\mathbf{Z}^{(M)}$. Here, λ is a fixed hyperparameter set to 0.2.

$$\mathbf{V}_f = \mathbf{V}_{proj} + \lambda(\mathbf{Z}^{(M)} - \mathbf{V}_{proj}) \tag{9}$$

3.3.3 *Role of CMKI during Inference.* A key design choice in our framework is to apply CMKI only during training and bypass it at test time. During training, CMKI uses the ground-truth report's semantic vector $\mathbf{T}$ to modulate visual features via the gating-and-blending mechanism (Eq. (6–9)), providing deep semantic supervision akin to knowledge distillation. This alignment signal backpropagates to the upstream visual summarizer (FVS) and the linear projection, training them to produce visual tokens that are naturally closer to the textual domain.

At inference, when no ground-truth text is available, CMKI is disabled. The FVS and projection layers—already shaped by the training-time supervision—directly output the semantically aware visual summary $\mathbf{V}_{\text{proj}}$ to the LLM, together with a fixed instruction prompt no ground-truth tokens are ever fed to the LLM. In this design, CMKI provides a strong supervisory signal during training that improves the visual frontend without risking leakage, yielding a lean and robust inference pipeline.

3.4 Training Objective

The final enriched visual representation, $\mathbf{V}_f \in \mathbb{R}^{N_s \times D_{LLM}}$, serves as a soft visual prompt that conditions the frozen LLM. Specifically, these N_s vectors are prepended to the input embedding sequence of the report tokens. This effectively steers the language generation process based on the distilled and semantically enriched visual evidence.

The trainable parameters of our model are optimized via a standard auto-regressive training objective. The model's learning is guided by an instruction-tuning format where an instruction prompt $\mathbf{P}$ "Generate a concise radiology report focusing on findings only." is also provided. For a given report R consisting of a sequence of tokens $\{r_1, ..., r_T\}$, the loss function $\mathcal{L}(\theta)$ for the trainable parameters $\boldsymbol{\theta}$ is the negative log-likelihood of the target tokens:

$$\mathcal{L}(\theta) = -\sum_{t=1}^{T} \log p_\theta(r_t | r_{<t}, \mathbf{V}_f, \mathbf{P}) \tag{10}$$

where p_θ is the distribution defined by the LLM, and $r_{<t}$ denotes the ground-truth prefix before r_t.

4 Experiment

4.1 Experimental setup

4.1.1 *Datasets.* We evaluated our model on two widely used benchmark datasets: IU-Xray dataset [6] and MIMIC-CXR dataset [13], In IU-Xray, we trained the model for a maximum of 15 epochs with a learning rate of 1×10-5. The batch size was set to 8 for training and 12 for validation. In MIMIC-CXR dataset, the model was trained for a maximum of 5 epochs with a learning rate of 1×10-6 and a gradient clipping value of 1. We followed the widely known data split provided by R2GenCNN [4], The final data distributions used for training are shown in Table 1.

Dataset	Training	Validation	Test
IU-Xray	2,070	297	591
MIMIC-CXR	270,746	2,130	3,858

Table 1: Dataset Statistics

4.1.2 *Implementation Details.* While our experiments employ the Llama2-7B chat model [23] as the LLM, it is important to note that the SKIM framework is architecturally decoupled from the specific LLM. The interface modules (FVS, CMKI) are designed as a universal bridge, allowing the frozen LLM decoder to be readily substituted with other foundational models. The experiments were performed on a single NVIDIA A100 GPU. For report generation at the inference stage, we feed only the image tokens to the SKIM model, excluding the ground truth report and any disease classification labels. To assess performance, we utilize widely-used natural language generation (NLG) metrics: BLEU [18], METEOR [2] and ROUGE-L [16].

4.2 Model Analysis

4.2.1 *Comparison with State-of-the-art Methods.* As shown in Table 2, our model achieves significant improvements on each dataset. Specifically, our model significantly surpasses previous methods, achieving top scores in both BLEU-4 and ROUGE-L. This underscores our model's ability to generate reports that are both clinically accurate and structurally coherent. The superior performance stems from our novel two-stage architecture. Whereas conventional methods typically focus on either enhancing visual features through complex architectures or integrating knowledge using external priors, our model effectively bridges the modality gap through a summarization then enrichment process. The FVS module first creates an efficient, structured summary of visual evidence, which allows the subsequent CMKI module to perform a more targeted and effective semantic alignment. This design proves particularly advantageous, facilitating a more meaningful cross-modal fusion than directly feeding raw or coarsely processed visual features into an LLM.

4.2.2 *Clinical Efficacy (CE).* For Clinical Efficacy (CE), we compute precision, recall, and F1-score across 14 clinical labels. These labels are automatically extracted from both generated and reference reports using the CheXbert labeler[21]. The scores shown in Table 3 are reported as CE, where higher values indicate better performance.

Method	Ref.	Category	LLM	IU-Xray						MIMIC-CXR					
				B-1	B-2	B-3	B-4	M	R-L	B-1	B-2	B-3	B-4	M	R-L
Visual Enhancement															
MeTransformer	CVPR'23	Expert Tokens	×	0.483	0.322	0.228	0.172	0.192	0.380	0.386	0.250	0.169	**0.124**	0.152	0.291
Diff-RRG	MICCAI'25	Visual Patches	✓	-	-	-	-	-	-	**0.405**	0.251	0.169	0.120	0.164	0.276
Modal Fusion															
R2Gen	EMNLP'20	Memory-driven Transformer	×	0.470	0.304	0.219	0.165	0.187	0.371	0.353	0.218	0.145	0.103	0.142	0.277
R2GenCMN	ACL'21	Cross-Modal Memory	×	0.475	0.309	0.222	0.170	0.191	0.375	0.353	0.218	0.148	0.106	0.142	0.278
MA	AAAI'24	Adaptive Fusion	×	0.501	0.328	0.230	0.170	0.213	0.386	0.396	0.244	0.162	0.115	0.151	0.274
Knowledge Enhancement															
ASGMD	ESWA'24	auxiliary signal guided	×	0.489	0.326	0.232	0.173	0.206	0.397	0.372	0.233	0.154	0.112	0.152	0.286
PromptMRG	AAAI'24	Diagnosis Prompts	✓	0.401	-	-	0.098	0.160	0.281	0.398	-	-	0.112	0.157	0.268
LLM-based Decoder															
BLIP-2	ICML'23	Q-former	✓	0.476	0.273	0.210	0.168	0.181	0.372	0.377	0.221	0.125	0.088	0.152	0.274
R2GenGPT(base)	Meta-Rad'23	Adapters	✓	0.466	0.301	0.211	0.156	0.202	0.37	0.365	0.237	0.163	0.117	0.136	0.277
(Ours)		Visual Summary + Modal alignment	✓	**0.518**	**0.356**	**0.260**	**0.195**	**0.220**	**0.405**	0.392	**0.255**	**0.175**	**0.124**	**0.168**	**0.293**

Table 2: Performance Comparison of Medical Report Generation Models on IU-Xray and MIMIC-CXR Datasets. The best and second-best results are shown in bold and underlined, respectively. B-n for BLEU-n, M for METEOR and R for ROUGE-L. BLIP-2 values are reported from [7].

Model	**P**	**R**	**F1**
R2Gen	0.333	0.273	0.276
R2GenCMN	0.334	0.275	0.278
MeTransformer	0.364	0.309	0.334
R2GenGPT (base)	0.341	0.312	0.325
(Ours)	**0.411**	**0.399**	**0.353**

Table 3: Results of the CE metrics on the MIMIC-CXR. P is Precision, R is Recall score.

Model	**Trainable Params (M)**
R2GenCMN	59.10
R2Gen	78.50
MeTransformer	152.00
BLIP-2	188.00
PromptMRG	219.92
(Ours)	**132.00**

Table 4: Trainable parameters (M)

4.2.3 Efficiency and Parameter Analysis. Our model's efficiency gains stem from utilizing a fixed, short token path into the LLM, rather than from minimizing the total number of trainable parameters. In our framework, the image encoder and LLM remain frozen, and the trainable parameter count is designed to be essentially insensitive to the number of summary slots, N_s. This is because, apart from the slot embedding matrix itself, all shared projection, attention, and normalization weights are N_s-agnostic.

Table 4 compares the trainable parameter counts of our model against representative methods. The results position our approach as a compelling trade-off, while not the smallest model in absolute terms, it is significantly more parameter-efficient than powerful contemporary baselines such as BLIP-2 and PromptMRG. This demonstrates a key advantage in balancing high representational capacity with the architectural efficiency of the trainable interface.

4.3 Ablation Studies

4.3.1 Effect of Modules. To validate our core components, we conduct an ablation study as shown in Table 5, incrementally adding the FVS and CMKI module to a baseline model (A). As shown in (B), adding the CMKI module substantially boosts performance. For example, BLEU-4 on IU-Xray improves from 0.167 to 0.183. This validates its effectiveness in bridging the semantic gap by pre-aligning visual features with textual semantics, thus providing the LLM with more coherent representations for generation.

The effect of the FVS module (C) is more nuanced and dataset-dependent. We attribute this diminished effect to the multi-view structure of the dataset. The mandatory step of averaging view embeddings acts as a preliminary, coarse summarization. Consequently, applying the FVS module afterward provides limited additional benefit, as the critical patterns might have already been diluted, a problem exacerbated by the dataset's small size. However, these limitations are resolved in our Full Model (D). The results strongly support our central hypothesis: FVS compression, though potentially lossy in isolation, becomes maximally effective when its compact representations are enriched by CMKI's semantic alignment. This synergy validates our two-stage design, where FVS first creates an efficient structure which is then infused with clinically-aware semantics by CMKI.

4.3.2 Qualitative Analysis. To qualitatively assess our model's performance, we conducted a case study comparing our model against

Models	Modules		IU-Xray						MIMIC-CXR					
	FVS	CMKI	B-1	B-2	B-3	B-4	M	R-L	B-1	B-2	B-3	B-4	M	R-L
(A)	-	-	0.480	0.312	0.223	0.167	0.217	0.379	0.383	0.233	0.155	0.108	0.150	0.268
(B)	-	✓	0.501	0.335	0.244	0.183	0.213	0.382	0.400	0.243	0.160	0.112	0.148	0.268
(C)	✓	-	0.477	0.302	0.217	0.165	0.198	0.386	**0.403**	0.250	0.166	0.117	0.147	0.269
(D)	✓	✓	**0.518**	**0.356**	**0.260**	**0.195**	**0.220**	**0.405**	0.392	**0.255**	**0.175**	**0.124**	**0.168**	**0.293**

Table 5: Ablation study of each module on each dataset. The ✓ indicates which modules are added to the baseline.

the baseline. As illustrated in Figure 5, our model demonstrates superior clinical reliability. For this case, the ground truth report indicates a normal heart size. Our model accurately generates a report stating the heart is "within normal limits," correctly reflecting the ground truth. In contrast, the baseline model incorrectly reports that "the heart is enlarged," committing a significant False Positive error. This type of hallucination could lead to unnecessary clinical follow-ups and patient anxiety, highlighting our model's enhanced safety and trustworthiness in a clinical setting. Overall, our model exhibits a lower propensity for hallucinating non-existent pathologies, leading to more consistent and reliable diagnostic report generation.

4.3.3 Effect of Slot numbers. We conducted an ablation study to determine the optimal number of slots N_s for our FVS module, with results presented in Table 6. The study reveals a clear trend where performance improves as N_s increases from a small value, suggesting a larger capacity is needed to effectively summarize complex visual features. Notably, the model achieves its peak performance at N_s=32, reaching the best scores on key metrics like Bleu-4 0.195 and ROUGE-L 0.405. As N_s increases further to 64 and 128, performance begins to decline, likely due to feature redundancy.

Number of Slots (N_s)	B-1	B-2	B-3	B-4	M	R-L
2	0.466	0.322	0.235	0.174	0.215	0.389
16	0.495	0.319	0.219	0.155	0.208	0.390
32	**0.518**	**0.356**	**0.260**	**0.195**	**0.220**	**0.405**
64	0.498	0.315	0.222	0.163	0.195	0.336
128	0.464	0.300	0.212	0.155	0.204	0.338

Table 6: Ablation study on the effect of the number of slots (N_s) on the IU-Xray dataset.

5 Conclusion and Future Work

In this paper, we proposed SKIM, a framework that effectively bridges a frozen visual encoder and an LLM for medical report generation. Our core contribution is a summarized alignment process, composed of a Fixed-Size Visual Summarization (FVS) module and a Cross-Modal Knowledge Infusion (CMKI) module. Experimental results show that our model achieves state-of-the-art performance on both the MIMIC-CXR and IU-Xray datasets.

Despite its strong performance, our model's ability to detect complex abnormal findings is constrained by the inherent class imbalance of the training data. The dataset is predominantly composed of normal cases, which can limit the model's sensitivity for less frequent, yet clinically significant, pathologies. Future work will address this data imbalance using techniques like advanced data augmentation to increase the weight of minority classes during training. This approach aims to enhance the model's sensitivity to a wider range of pathologies while maintaining its high specificity for normal findings, thereby maximizing its overall clinical utility.

However, relying on a frozen, general-domain LLM introduces challenges. A key advantage of the SKIM framework is its LLM-agnostic design, which allows for the future adoption of more advanced models. To address the more fundamental challenge of ensuring clinical validity regardless of the underlying LLM, our key future direction is making the summary slots interactive. This will allow clinicians to directly query, validate, and refine the visual findings captured by the model.

6 Acknowledgments

This research is supported by the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. RS-2023-00229822).

References

[1] Zubair Ahmad, Shabina Rahim, Maha Zubair, and Jamshid Abdul-Ghafar. 2021. Artificial intelligence (ai) in medicine, current applications and future role with special emphasis on its potential and promise in pathology: present and future impact, obstacles including costs and acceptance among pathologists, practical and philosophical considerations. a comprehensive review. *Diagnostic Pathology*, 16, (Mar. 2021). doi:10.1186/s13000-021-01085-4.

[2] Satanjeev Banerjee and Alon Lavie. 2005. Meteor: an automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2021. Beit: bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

[4] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. 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)*. Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, (Eds.) Association for Computational Linguistics, Online, (Aug. 2021), 5904–5914. doi:10.18653/v1/2021.acl-long.459.

[5] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2022. Generating radiology reports via memory-driven transformer. (2022). https://arxiv.org/abs/2010.16056 arXiv: 2010.16056 [cs.CL].

[6] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2015. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23, 2, 304–310.

[7] Ankan Deria, Komal Kumar, Snehashis Chakraborty, Dwarikanath Mahapatra, and Sudipta Roy. 2024. Inverge: intelligent visual encoder for bridging

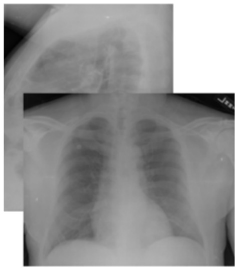
Original Images	Ground truth	Base	Ours
	"heart size and pulmonary vascularity within normal limits . no focal infiltrate pneumothorax or pleural effusion identified ."	"the heart is enlarged . the lungs are clear without focal airspace consolidation pleural effusion or pneumothorax . degenerative changes of the thoracic spine are present . no acute osseous abnormality is identified in the skeletal survey ."	"the heart and mediastinum are within normal limits . the lungs are clear . there is no focal consolidation or pleural effusion . no pneumothorax . bony structures are intact ."

Figure 5: A qualitative comparison from the IU X-ray dataset illustrating our model's superior accuracy. For the given input X-ray, where the ground truth report indicates a normal heart size, our model correctly generates a report describing the heart as 'within normal limits'. In contrast, the baseline model commits a significant False Positive error, incorrectly diagnosing an 'enlarged' heart.

modalities in report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2028–2038.

[8] Tiancheng Gu, Kaicheng Yang, Xiang An, Ziyong Feng, Dongnan Lin, and Weidong Cai. 2025. Orid: organ-regional information driven framework for radiology report generation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 378–387.

[9] Xiaofei Huang, Wenting Chen, Jie Liu, Qisheng Lu, Xiaoling Luo, and Linlin Shen. 2025. Damper: a dual-stage medical report generation framework with coarse-grained mesh alignment and fine-grained hypergraph matching. In *Proceedings of the AAAI Conference on Artificial Intelligence* number 4. Vol. 39, 3769–3778.

[10] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. (2017). https://arxiv.org/abs/1703.06868 arXiv: 1703.06868 [cs.CV].

[11] Andrew Jaegle et al. 2022. Perceiver io: a general architecture for structured inputs outputs. (2022). https://arxiv.org/abs/2107.14795 arXiv: 2107.14795 [cs.LG].

[12] Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. 2024. Promptmrg: diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence* number 3. Vol. 38, 2607–2615.

[13] Alistair E. W. Johnson et al. 2019. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. (2019). https://arxiv.org/abs/1901.07042 arXiv: 1901.07042 [cs.CV].

[14] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning* (Proceedings of Machine Learning Research). Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, (Eds.) Vol. 202. PMLR, (23–29 Jul 2023), 19730–19742. https://proceedings.mlr.press/v202/li23q.html.

[15] Kevin Y Li, Sachin Goyal, Joao D Semedo, and J Zico Kolter. 2024. Inference optimal vlms need fewer visual tokens and more parameters. *arXiv preprint arXiv:2411.03312*.

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

[17] Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. 2024. Bootstrapping large language models for radiology report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence* number 17. Vol. 38, 18635–18643.

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

[19] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. 2021. Dynamicvit: efficient vision transformers with dynamic token sparsification. (2021). https://arxiv.org/abs/2106.02034 arXiv: 2106.02034 [cs.CV].

[20] Hongyu Shen, Mingtao Pei, Juncai Liu, and Zhaoxing Tian. 2024. Automatic radiology reports generation via memory alignment network. In *Proceedings of the AAAI Conference on Artificial Intelligence* number 5. Vol. 38, 4776–4783.

[21] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. 2020. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*.

[22] Stefanie Steinhauser and Sabrina Welsch. 2025. Large language models in radiology workflows: an exploratory study of generative ai for non-visual tasks in the german healthcare system. *Health Policy*, 161, 105444. doi:https://doi.org/10.1016/j.healthpol.2025.105444.

[23] Hugo Touvron et al. 2023. Llama 2: open foundation and fine-tuned chat models. (2023). https://arxiv.org/abs/2307.09288 arXiv: 2307.09288 [cs.CL].

[24] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: a neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3156–3164.

[25] Xiao Wang, Fuling Wang, Yuehang Li, Qingchuan Ma, Shiao Wang, Bo Jiang, Chuanfu Li, and Jin Tang. 2024. Cxpmrg-bench: pre-training and benchmarking for x-ray medical report generation on chexpert plus dataset. (2024). https://arxiv.org/abs/2410.00379 arXiv: 2410.00379 [cs.CV].

[26] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023. Metransformer: radiology report generation by transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11558–11567.

[27] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023. Metransformer: radiology report generation by transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11558–11567.

[28] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023. R2gengpt: radiology report generation with frozen llms. *Meta-Radiology*, 1, 3, 100033. doi:https://doi.org/10.1016/j.metrad.2023.100033.

[29] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. Medclip: contrastive learning from unpaired medical images and text. (2022). https://arxiv.org/abs/2210.10163 arXiv: 2210.10163 [cs.CV].

[30] Qilong Xing, Zikai Song, Youjia Zhang, Na Feng, Junqing Yu, and Wei Yang. 2025. Mca-rg: enhancing llms with medical concept alignment for radiology report generation. *arXiv preprint arXiv:2507.06992*.

[31] Youyuan Xue, Yun Tan, Ling Tan, Jiaohua Qin, and Xuyu Xiang. 2024. Generating radiology reports via auxiliary signal guidance and a memory-driven network. *Expert Systems with Applications*, 237, 121260. doi:https://doi.org/10.1016/j.eswa.2023.121260.

[32] Yan Yang et al. 2024. Token-mixer: bind image and text in one embedding space for medical image reporting. *IEEE Transactions on Medical Imaging*, 43, 11, 4017–4028.

[33] Hannah Yun, Junyeong Maeng, Eunsong Kang, and Heung-Il Suk. 2025. Diff-rrg: longitudinal disease-wise patch difference as guidance for llm-based radiology report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 152–161.