MARE: Multimodal Analogical Reasoning for Disease Evolution-Aware Radiology Report Generation

Qingqing Gao[1], Tengfei Liu[1], Xiaoyan Li[1], Xiaodan Zhang[2], Zhongfan Sun[1], Boyue Wang[1*], Baocai Yin[1], Zhaohui Liu[3*]

[1]School of Information Science and Technology, Beijing University of Technology, Beijing, China
[2]College of Computer Science, Beijing University of Technology, Beijing, China
[3]Beijing Tongren Hospital, Beijing, China
{gaoqq, sunzf}@emails.bjut.edu.cn, {tfliu, xiaoyan.li, zhangxiaodan, wby, ybc}@bjut.edu.cn, lzhtrhos@163.com

Abstract

Radiology report generation from longitudinal medical data is critical for assessing disease progression and automating diagnostic workflows. While recent methods incorporate longitudinal information, they primarily rely on multimodal feature fusion, with limited capacity for explicit disease evolution modeling and temporal reasoning. To address this, we propose MARE, an end-to-end framework that formulates longitudinal radiology report generation as a multimodal analogical reasoning task. Inspired by the Abduction–Mapping–Induction paradigm, MARE models latent relational structures underlying disease evolution by aligning lesion-level visual features across time and mapping them to the textual domain for temporally coherent and clinically meaningful report generation. To mitigate the spatial misalignment caused by patient positioning or imaging variation, we introduce an Adaptive Region Alignment (ARA) module for robust temporal correspondence. Additionally, we design Dual Evolution Consistency (DEC) losses to regularize analogical reasoning by enforcing temporal coherence in both visual and textual evolution paths. Extensive experiments on the Longitudinal-MIMIC dataset demonstrate that MARE significantly outperforms state-of-the-art baselines across both natural language generation and clinical effectiveness metrics, highlighting the value of structured analogical reasoning for disease evolution-aware report generation.

Code — https://github.com/gaoqingqing77/MARE

Introduction

Radiology report generation has emerged as a critical task for automating diagnostic workflows and alleviating the reporting burden on radiologists. Existing methods (Chen et al. 2020; Liu et al. 2021a; Chen et al. 2021; Li et al. 2023a,b; Huang, Zhang, and Zhang 2023; Wang et al. 2023; Jin et al. 2024; Liu et al. 2024; Shen et al. 2024; Li et al. 2024; Xiao et al. 2025; Huang et al. 2025; Wang et al. 2025; Zhang et al. 2025) predominantly focus on single-timepoint image-to-text generation, often overlooking the longitudinal nature of clinical data that is crucial for assessing disease progression. Recent advances (Bannur et al. 2023; Zhu et al. 2023; Serra et al. 2023; Hou et al. 2023; Sanjeev et al. 2024; Wang, Du, and Yu 2024; Liu et al. 2025) have begun to incorporate multimodal longitudinal data, combining current and historical radiographs and reports to enhance generation quality.

Despite recent advances, many longitudinal radiology report generation (LRRG) methods still rely on general multimodal fusion strategies, which implicitly expect the model to capture temporal relations, align heterogeneous visual-textual patterns, and reason about disease evolution. This general reasoning paradigm imposes substantial cognitive loads on the model and often results in clinically inconsistent or incoherent narratives. Fundamentally, the temporal progression of pathological findings in chest X-rays often corresponds to changes in how diseases are described in radiology reports. This observation motivates a shift from general reasoning to multimodal analogical reasoning (Zhang et al. 2023), which explicitly models relational structures in a source domain (image modality) and uses them to analogically infer specific missing information in a target domain (report modality). Prior works (Holyoak and Thagard 1996; Reed et al. 2015) have shown that analogical reasoning, by leveraging relational similarities between analogy pairs, can reduce reasoning complexity and enhance generalization.

In this context, we propose to frame the task of LRRG as a multimodal analogical reasoning problem (Mar-LRRG). As illustrated in Figure 1(a), unlike general reasoning that rely on multimodal feature fusion, analogical reasoning fundamentally focuses on recognizing and mapping relational structures across domains (Gentner 1983), with its effectiveness hinging on accurately modeling the relations within the source domain (Reed et al. 2015). From this perspective, the core challenge of LRRG can be distilled into a central research question: how can we accurately model the disease evolution relations embedded in longitudinal medical images to effectively guide report generation?

To answer this question, we analyze the distinctions between Mar-LRRG (Figure 1(c)) and classical multimodal analogical reasoning over knowledge graphs (Mar-KG) (Zhang et al. 2023) (Figure 1(b)), emphasizing the additional complexity in modeling longitudinal medical data. First, unlike Mar-KG tasks where the relations between input pairs are typically predefined, explicit, and single-labeled, medical images in LRRG often contain multiple lesion regions with distinct temporal trajectories. This leads to

*Corresponding author
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

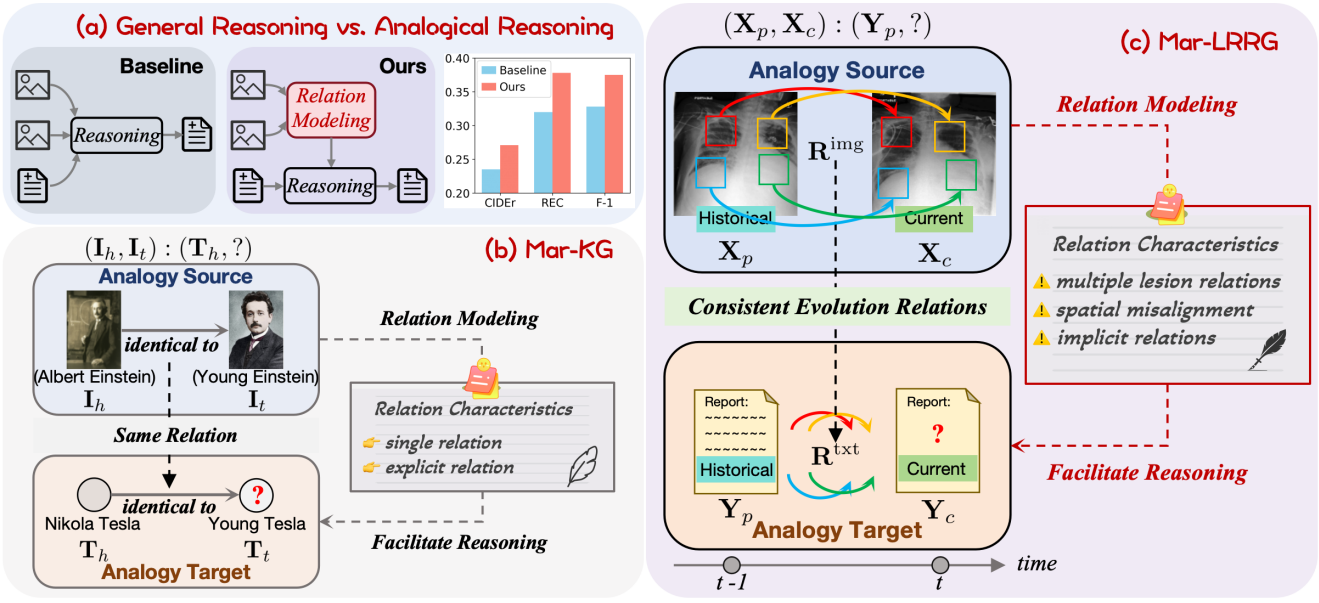

Figure 1: Illustration of (a) general reasoning vs. analogical reasoning, where the bar chart shows that relation modeling improves reasoning effectiveness; and (b) Mar-KG vs. (c) Mar-LRRG, highlighting the unique challenges in modeling disease evolution relations, including multiple lesion relations, spatial misalignment, and implicit relational patterns.

complex and implicit relational patterns that cannot be adequately captured by a single predefined relation. To tackle this, we introduce structured relational modeling components that weakly explicate the implicit disease evolution patterns. Specifically, we design dual evolution consistency losses that constrain relational changes across modalities and enforce temporal coherence in both vision and language streams. Second, longitudinal medical images frequently suffer from spatial misalignment caused by variations in patient positioning, imaging conditions, or anatomical shifts, making it difficult to track lesion-level changes consistently across time. To overcome this, we introduce an adaptive alignment mechanism based on the perceiver resampler (Alayrac et al. 2022), which dynamically extracts and aligns fine-grained features from key pathological regions across historical and current images.

Building on the above analysis, we propose MARE, a complete Mar-LRRG pipeline that operationalizes multimodal analogical reasoning under the Abduction–Mapping–Induction paradigm (Minnameier 2010). In the Abduction stage, MARE first hypothesizes latent disease evolution relations by analyzing cross-time lesion-level changes in medical images. To ensure reliable correspondence under spatial variation, we propose an Adaptive Region Alignment (ARA) module that explicitly aligns salient pathological regions across time, providing a structurally grounded basis for analogical inference. In the Mapping stage, these inferred visual evolution patterns are transferred to the language modality via a cross-modal mapping mechanism. To enforce relational consistency, we introduce Dual Evolution Consistency (DEC) losses, which simultaneously supervise intra-modal progression paths in both the image and report domains. This dual-view constraint enables MARE to learn temporally coherent and clinically faithful analogical mappings. In the Induction stage, a Large Language Model (LLM) integrates the aligned relational patterns and historical report context to generate fluent and diagnostically relevant radiology reports. The entire reasoning pipeline is implemented in an end-to-end manner, allowing joint optimization of all components. Extensive experiments on the Longitudinal-MIMIC dataset show that MARE consistently outperforms state-of-the-art methods in both language generation and clinical metrics. Our main contributions are summarized as follows:

- We innovatively frame LRRG as a multimodal analogical reasoning task and propose MARE, an end-to-end pipeline following the Abduction–Mapping–Induction paradigm.
- We address key challenges in Mar-LRRG by introducing an Adaptive Region Alignment (ARA) module to handle lesion-level spatial misalignment, and Dual Evolution Consistency (DEC) losses to enhance relational modeling across modalities.
- Extensive experiments demonstrate that our analogical reasoning framework with disease evolution modeling strategies achieves substantial improvements over state-of-the-art methods.

Related Work

Analogical Reasoning

Analogical reasoning facilitates knowledge transfer by mapping relational structures from known to novel domains, enabling deeper inference beyond surface similarities (Gentner 1983). Minnameier (Minnameier 2010) formalizes this

process into three stages: Abduction, Mapping, and Induction—outlining how relational knowledge is abstracted, aligned, and generalized. Recent works in deep learning have explored analogical reasoning to equip models with human-like inference capabilities (Hu and Clune 2023). In computer vision, analogical reasoning integrates visual data with relational structures (Prade and Richard 2021; Hu et al. 2021; Hayes and Kanan 2021; Małkiński and Mańdziuk 2025), while in NLP, it is applied to word and sentence-level semantic understanding (Mikolov et al. 2013; Pennington, Socher, and Manning 2014; Chen et al. 2022). Multimodal approaches further combine structured knowledge and pre-trained transformers for enhanced reasoning (Zhang et al. 2023; Cai et al. 2025). LLMs have shown emerging analogical abilities (Yasunaga et al. 2023; Webb, Holyoak, and Lu 2023), but applications in medical AI are scarce. In longitudinal radiology, analogical reasoning aids case comparison, disease modeling, and consistent reporting. Motivated by this, we leverage analogical reasoning with LLMs to enhance cross-time multimodal understanding in radiology report generation.

Longitudinal Radiology Report Generation

Traditional report generation methods (Chen et al. 2020; Liu et al. 2021a; Chen et al. 2021; Li et al. 2023a,b; Huang, Zhang, and Zhang 2023; Wang et al. 2023; Shen et al. 2024; Li et al. 2024; Jin et al. 2024; Liu et al. 2024; Xiao et al. 2025; Huang et al. 2025; Wang et al. 2025; Zhang et al. 2025), mainly focus on single-time data, overlooking longitudinal sequences. Recent works (Bannur et al. 2023; Zhu et al. 2023; Serra et al. 2023; Hou et al. 2023; Sanjeev et al. 2024; Wang, Du, and Yu 2024) incorporate multimodal longitudinal data, which is essential for assessing disease progression. Notably, Zhu et al. (Zhu et al. 2023) use longitudinal data to pre-fill report findings; RECAP (Hou et al. 2023) integrates disease progression graphs; HERGen (Wang, Du, and Yu 2024) applies group causal transformers with contrastive learning; HC-LLM (Liu et al. 2025) captures time-shared and time-specific features to enhance consistency and accuracy. Despite these advances, existing methods insufficiently explore explicit modeling of disease evolution, and we address this gap by framing LRRG as a structured multimodal analogical reasoning task.

Methodology

Problem Definition

Given a radiograph-report pair $\mathbf{D}_c = \{\mathbf{X}_c, \mathbf{Y}_c\}$, where $\mathbf{X}_c$ is the current radiograph and $\mathbf{Y}_c$ denotes the corresponding ground-truth report, along with the prior visit record $\mathbf{D}_p = \{\mathbf{X}_p, \mathbf{Y}_p\}$, the objective of LRRG is to produce a diagnostic report $\hat{\mathbf{Y}}_c$ for $\mathbf{X}_c$ that closely approximates $\mathbf{Y}_c$. This is formulated as maximizing the conditional probability $p(\mathbf{Y}_c|\mathbf{X}_c, \mathbf{D}_p)$ to capture dependencies between current and prior visits for coherent and accurate report generation.

Owing to the fact that the relational structure between radiographs and reports remains implicit yet consistent over time, we formalize the LRRG task as a multimodal analogical reasoning problem. Specifically, we consider the historical and current radiographs $(\mathbf{X}_p, \mathbf{X}_c)$ as the analogy source, and the corresponding historical and current reports $(\mathbf{Y}_p, \mathbf{Y}_c)$ as the analogy target. The objective of Mar-LRRG is to model the consistent relational patterns across time to infer the missing report $\mathbf{Y}_c$, thereby enhancing the effectiveness of the reasoning process. Formally, the task can be expressed as: $(\mathbf{X}_p, \mathbf{X}_c) : (\mathbf{Y}_p, \ ?)$.

Following the Abduction–Mapping–Induction paradigm of cognitive analogical reasoning (Minnameier 2010), we propose the MARE framework, as shown in Figure 2. The three stages form a progressive reasoning flow, with relation modeling as a core component rather than direct multimodal fusion or unconstrained LLM inference. Together with the DEC losses, this design enforces cross-modal semantic consistency and reasoning continuity in an end-to-end manner. The following sections detail the three-stage process and the DEC losses.

Abduction: Disease Evolution Relation Modeling

The **Abduction** stage in analogical reasoning focuses on identifying the disease evolution relations between the historical and current radiographs, laying the foundation for subsequent mapping and induction. Given a historical–current radiograph pair $(\mathbf{X}_p, \mathbf{X}_c)$, each image is encoded by a shared-weight vision encoder $\mathcal{E}_{\text{img}}$:

$$\mathbf{F}^{[\text{cls}]}_{\mathbf{X}_p}, \mathbf{F}_{\mathbf{X}_p} = \mathcal{E}_{\text{img}}(\mathbf{X}_p), \quad \mathbf{F}^{[\text{cls}]}_{\mathbf{X}_c}, \mathbf{F}_{\mathbf{X}_c} = \mathcal{E}_{\text{img}}(\mathbf{X}_c), \quad (1)$$

where $\mathbf{F}^{[\text{cls}]}_{\mathbf{X}_*} \in \mathbb{R}^{d_v}$ and $\mathbf{F}_{\mathbf{X}_*} \in \mathbb{R}^{N \times d_v}$ denote global and local features, with N visual tokens of dimension d_v.

To align spatial regions of disease relevance, we introduce the Adaptive Region Alignment (ARA) mechanism that adaptively samples semantically aligned tokens from the local features of historical and current radiographs. ARA is based on the design of the perceiver resampler (Alayrac et al. 2022), which maps the input visual features to a fixed number of output tokens. First, we predefine the learnable latent input queries $\mathbf{Q}^0 \in \mathbb{R}^{M \times d_v}$. Then, we apply the following iterative process:

$$\begin{aligned} \tilde{\mathbf{Q}}^l &= \mathcal{F}_{\text{Att}}(\mathbf{Q}^l, \mathbf{F}_{\mathbf{X}_*}) + \mathbf{Q}^l, \\ \mathbf{Q}^{l+1} &= \mathcal{F}_{\text{FFN}}(\tilde{\mathbf{Q}}^l) + \tilde{\mathbf{Q}}^l, \end{aligned} \quad (2)$$

where $\mathcal{F}_{\text{Att}}$ is the cross-attention module, and $\mathcal{F}_{\text{FFN}}$ is the feed-forward network. This process is repeated L times, where $l \in \{1, 2, ..., L\}$ denotes the index of the current interaction layer. Finally, the resampled features are obtained as $\tilde{\mathbf{F}}_{\mathbf{X}_*} = \mathbf{Q}^L \in \mathbb{R}^{M \times d_v}$, where M is the number of output tokens, which equals the number of learned latent queries. This ARA mechanism ensures that corresponding regions of interest across different time points are effectively aligned, providing a more coherent representation for downstream disease evolution modeling.

We then concatenate the aligned features and feed them to a relation learner $\mathcal{F}_{\text{evo}}$ to model the implicit visual disease evolution relations:

$$\mathbf{R}^{\text{img}} = \mathcal{F}_{\text{evo}}([\tilde{\mathbf{F}}_{\mathbf{X}_p}; \tilde{\mathbf{F}}_{\mathbf{X}_c}]), \quad (3)$$

yielding $\mathbf{R}^{\text{img}} \in \mathbb{R}^{M \times 2d_v}$ as the learned inter-image disease evolution relations. These relations, further regularized by

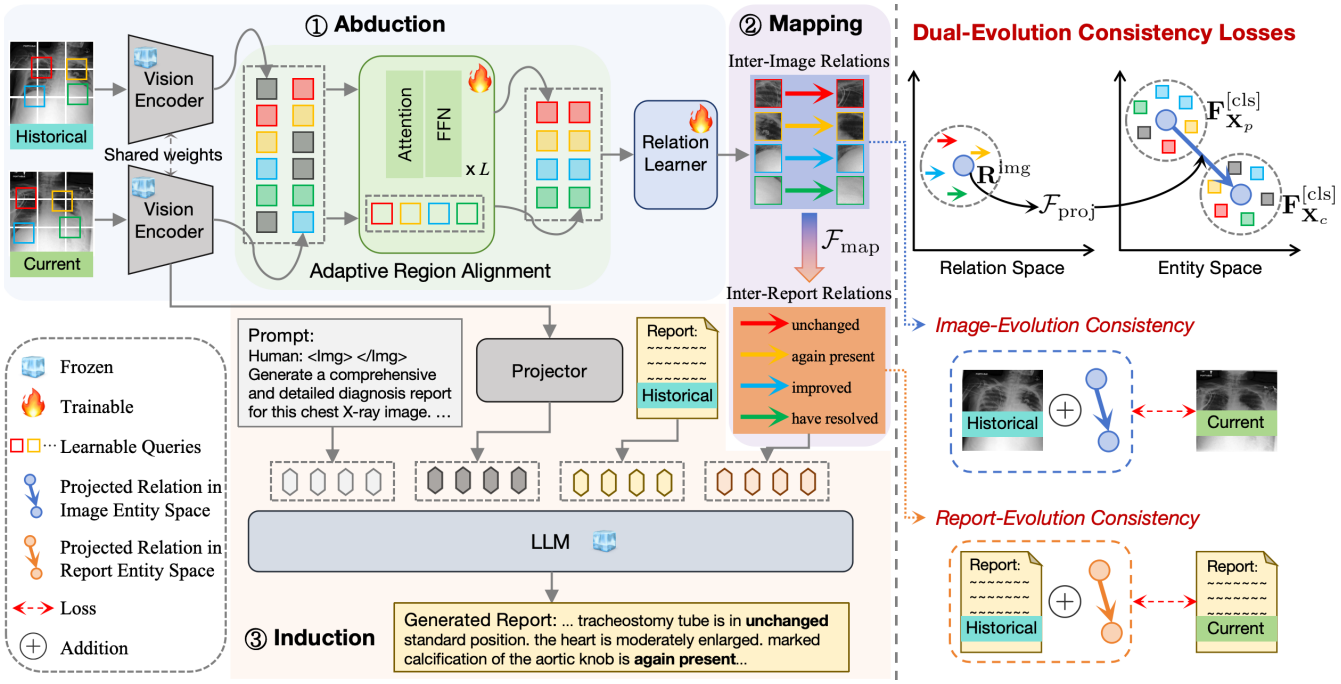

Figure 2: Overview of the MARE framework: ① Abduction stage for disease evolution relation modeling; ② Mapping stage for cross-modal relation projection; ③ Induction stage for disease evolution-aware report generation. The Dual-Evolution Consistency Losses, shown on the right, include Image-Evolution Consistency and Report-Evolution Consistency, which constrain the learning of Inter-Image Relations and Inter-Report Relations, respectively. The four inter-report relations are illustrative examples; the learned relation space is continuous with implicit features beyond these categories.

the image-evolution component of our DEC losses, capture implicit yet clinically meaningful progression patterns that direct multimodal fusion would struggle to disentangle.

Mapping: Cross-modal Disease Evolution Mapping

The **Mapping** stage in analogical reasoning involves projecting the relations discovered in the source domain (visual) to the target domain (text).

We employ a mapping function $\mathcal{F}_{\text{map}}$, implemented as a fully connected layer, to project $\mathbf{R}^{\text{img}} \in \mathbb{R}^{M \times 2d_v}$ into the report relation space $\mathbf{R}^{\text{txt}} \in \mathbb{R}^{M \times d_t}$:

$$\mathbf{R}^{\text{txt}} = \mathcal{F}_{\text{map}}(\mathbf{R}^{\text{img}}). \tag{4}$$

The mapped relations are further constrained by the report-evolution component of our DEC losses, grounding them in the textual evolution embedding space and providing evolution-aware cues for the subsequent induction stage.

Induction: Disease Evolution-Aware Generation

In the **Induction** stage, we generate the current radiology report $\hat{\mathbf{Y}}_c$ based on the analogically transferred disease evolution relations $\mathbf{R}^{\text{txt}}$ and the historical report $\mathbf{Y}_p$. Rather than relying on direct multimodal fusion, this process is explicitly guided by relational patterns derived from visual progression, ensuring temporally coherent and clinically consistent outputs.

Unlike typical Mar-KG tasks that focus on single-entity inference and can be addressed through basic classification or matching, the Mar-LRRG task requires generating comprehensive and detailed radiology reports. This requires not only effective modeling and integration of disease evolution patterns but also the production of logically coherent and clinically accurate narratives.

To meet these requirements, we employ an LLM (Touvron et al. 2023) as the reasoning engine, leveraging its capacity for knowledge integration and contextual inference. The inputs to the LLM include a predefined prompt, the mapped relations $\mathbf{R}^{\text{txt}}$, the historical report $\mathbf{Y}_p$, and optionally other auxiliary information (e.g., $\mathbf{F}_{\mathbf{X}_*}$):

$$\hat{\mathbf{Y}}_c = \mathcal{F}_{\text{LLM}}(Prompt, \mathbf{F}_{\mathbf{X}_c}, \mathbf{Y}_p, \mathbf{R}^{\text{txt}}). \tag{5}$$

During training, we adopt teacher forcing and optimize an autoregressive language modeling loss:

$$\mathcal{L}_{\text{gen}} = -\sum_{t=1}^{T} \log p(\mathbf{y}_c^t \mid \mathbf{y}_c^{<t}, \mathbf{X}_c, \mathbf{D}_p), \tag{6}$$

where $\mathbf{y}_c^t$ is the t-th token of the ground-truth report and $\mathbf{y}_c^{<t}$ denotes all preceding tokens. As the LLM is adopted as an analogical reasoning engine guided by disease evolution relations, this process reduces reasoning complexity, improves semantic grounding, and effectively induces consistent and coherent information in the analogy target domain.

Dual-Evolution Consistency Losses

To further regularize the analogical reasoning process and ensure that disease evolution patterns remain consistent

Model	Year	Inputs	NLG metrics							CE metrics		
			BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	PREC	REC	F-1
Transformer	2017	Single	0.294	0.178	0.119	0.085	0.123	0.256	-	-	-	-
AoANet	2019	Single	0.272	0.168	0.112	0.080	0.115	0.249	-	-	-	-
CNN+Trans	2019	Single	0.299	0.186	0.124	0.088	0.120	0.263	-	-	-	-
R2Gen	2020	Single	0.302	0.183	0.122	0.087	0.124	0.259	-	-	-	-
R2CMN	2021	Single	0.305	0.184	0.122	0.085	0.126	0.265	-	-	-	-
CvT2DistilGPT2	2023	Single	0.365	0.226	0.151	0.107	0.143	0.275	-	0.367	0.258	0.261
Prefilling	2023	Longitudinal	0.343	0.210	0.140	0.099	0.137	0.271	-	-	-	-
HERGen	2024	Longitudinal	0.389	0.242	0.163	0.117	0.155	0.282	-	0.421	0.289	0.295
HC-LLM	2025	Longitudinal	0.404	0.260	0.178	0.128	0.160	0.287	-	0.417	0.357	0.357
R2GenGPT	2023	Single	0.365	0.230	0.156	0.111	0.139	0.271	0.180	0.297	0.240	0.247
+ $\mathbf{Y}_p$		Longitudinal	0.393	0.252	0.174	0.125	0.153	0.284	0.243	0.390	0.336	0.336
+ $\mathbf{X}_p$		Longitudinal	0.364	0.229	0.155	0.110	0.140	0.270	0.178	0.293	0.248	0.250
+ $\mathbf{Y}_p$ & $\mathbf{X}_p$		Longitudinal	0.396	0.256	0.176	0.126	0.154	0.285	0.235	0.389	0.320	0.328
MARE (Ours)	-	Longitudinal	**0.409**	**0.265**	**0.184**	**0.133**	**0.161**	**0.291**	**0.271**	**0.433**	**0.378**	**0.375**

Table 1: Performance comparison of state-of-the-art baselines across NLG and CE metrics. Upper-part results are cited from HERGen, as our Longitudinal-MIMIC curation follows their setup for direct comparability. HC-LLM results are cited from the original publication. Lower-part results for R2GenGPT and its variants are reproduced using official code on Longitudinal-MIMIC with the same backbone. $\mathbf{X}_p$ and $\mathbf{Y}_p$ denote historical image and report, respectively.

across modalities, we propose DEC losses, comprising both image-level and report-level constraints. Acting as a weakly-supervised self-consistency constraint, DEC operates entirely in the feature space, leveraging inherent cross-time correspondences within the data, without requiring manual annotations of evolution relations.

Specifically, reliable disease evolution relations should transform the historical state into the current state within the same modality. However, these learned relations may deviate from the corresponding entity feature space. To address this, as shown in Figure 2, we map the relations into the respective modality feature space and define the two consistency losses therein.

For the **Image-Evolution Consistency** loss, we first project the pooled inter-image relations into the image feature space:

$$\tilde{\mathbf{R}}^{\text{img}} = \mathcal{F}^{\text{img}}_{\text{proj}}(\text{Pool}(\mathbf{R}^{\text{img}})), \tag{7}$$

and define the following loss to ensure that the modeled implicit relations $\mathbf{R}^{\text{img}}$ correctly bridge the global feature evolution from $\mathbf{X}_p$ to $\mathbf{X}_c$:

$$\mathcal{L}_{\text{img}} = \left\| \mathbf{F}^{[\text{cls}]}_{\mathbf{X}_p} + \tilde{\mathbf{R}}^{\text{img}} - \mathbf{F}^{[\text{cls}]}_{\mathbf{X}_c} \right\|_2^2. \tag{8}$$

For **Report-Evolution Consistency**, we first encode reports using a text encoder $\mathcal{E}_{\text{txt}}$ to obtain global features:

$$\mathbf{F}^{[\text{cls}]}_{\mathbf{Y}_p} = \mathcal{E}_{\text{txt}}(\mathbf{Y}_p), \quad \mathbf{F}^{[\text{cls}]}_{\mathbf{Y}_c} = \mathcal{E}_{\text{txt}}(\mathbf{Y}_c). \tag{9}$$

Then the loss for report evolution consistency is defined similarly as follows:

$$\tilde{\mathbf{R}}^{\text{txt}} = \mathcal{F}^{\text{txt}}_{\text{proj}}(\text{Pool}(\mathbf{R}^{\text{txt}})), \tag{10}$$

$$\mathcal{L}_{\text{txt}} = \left\| \mathbf{F}^{[\text{cls}]}_{\mathbf{Y}_p} + \tilde{\mathbf{R}}^{\text{txt}} - \mathbf{F}^{[\text{cls}]}_{\mathbf{Y}_c} \right\|_2^2. \tag{11}$$

By constraining both the visual and textual evolution spaces, DEC losses ensure that the analogically modeled relations are aligned across modalities, improving semantic grounding and temporal coherence.

Overall Objective

The total loss function for training the MARE model is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{gen}} + \lambda_1 \mathcal{L}_{\text{img}} + \lambda_2 \mathcal{L}_{\text{txt}}, \tag{12}$$

where λ_1 and λ_2 are hyperparameters controlling the strength of dual-evolution regularization. The entire process is implemented in an end-to-end manner.

Experiments

Experimental Setting

Dataset. We conduct experiments on the Longitudinal-MIMIC dataset, derived from MIMIC-CXR (Johnson et al. 2019), which contains 377,110 chest X-rays and 227,835 reports from 65,379 patients. Following Zhu et al. (Zhu et al. 2023), a subset of 26,625 patients with $\geq$2 visits was extracted, yielding 94,169 visit pairs with CXR images and reports from consecutive visits, covering 14 common abnormalities. The dataset is split into 92,374 training, 737 validation, and 2,058 test samples, ensuring temporal consistency via patient-wise, time-ordered splits based on 'StudyDate'.

Evaluation Metrics. We evaluate report quality using both natural language generation (NLG) and clinical effectiveness (CE) metrics. NLG metrics include BLEU (Papineni et al. 2002), METEOR (Denkowski and Lavie 2011), ROUGE-L (Lin 2004), and CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015), assessing generation fluency, and relevance. For clinical validity, we compute precision, recall, and F1-score over 14 CheXpert disease categories, using the CheXbert framework (Smit et al. 2020) to automatically map generated text to disease labels, enabling systematic evaluation of clinical consistency and accuracy.

Implementation Details. We implement our model in PyTorch and perform experiments on a single NVIDIA A800 GPU. The visual encoder uses Swin Transformer (Liu et al. 2021b), the text encoder (Eq. 9) adopts CLIP (Radford et al.

Model	BLEU-4	METEOR	ROUGE-L	CIDEr	F-1
w/o ARA	0.129	0.155	0.286	0.243	0.353
w/ ARA	**0.133**	**0.160**	**0.291**	**0.271**	**0.375**

Table 2: Ablation study on the effect of ARA module.

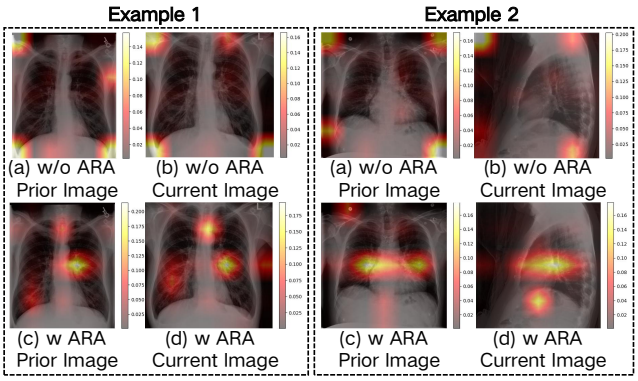

Figure 3: Visualization of attention heatmaps before and after applying ARA on prior and current chest X-ray images for two examples.

2021), and LLaMA2-7B (Touvron et al. 2023) is applied during the induction stage. ARA is configured with $L = 2$ layers and $M = 32$ queries, with DEC weights $\lambda_1 = \lambda_2 = 1$. We use AdamW with an initial learning rate of 1×10^{-4}, a CosineAnnealing scheduler, batch size 8, and 5 epochs. Inference uses beam search with size 3 to balance diversity and quality.

Comparison with State-of-the-Art Methods

We compare our method with a range of state-of-the-art baselines for radiology report generation, including single time-point approaches (e.g., Transformer, AoANet (Huang et al. 2019), CNN+Trans, R2Gen (Chen et al. 2020), R2CMN (Chen et al. 2021), CvT2DistilGPT2 (Nicolson, Dowling, and Koopman 2023)) and longitudinal models (e.g., Prefilling (Zhu et al. 2023), HERGen (Wang, Du, and Yu 2024), HC-LLM (Liu et al. 2025), R2GenGPT (Wang et al. 2023)). As shown in Table 1, our method consistently outperforms all baselines. Longitudinal models generally surpass single-time methods, highlighting the benefit of incorporating historical context; however, Prefilling underperforms CvT2DistilGPT2 due to its conventional architecture. Notably, our method outperforms strong longitudinal baselines such as HERGen and HC-LLM. In particular, when compared to R2GenGPT, which shares the same backbone as ours, we achieve relative improvements of 15.3% in CIDEr and 14.3% in F1, owing to our explicit modeling of disease evolution rather than merely stacking multimodal inputs. These results underscore the effectiveness of multimodal analogical reasoning for capturing temporal dynamics in longitudinal report generation.

$\mathcal{L}_{\text{img}}$	$\mathcal{L}_{\text{txt}}$	BLEU-4	METEOR	ROUGE-L	CIDEr	F-1
✗	✗	0.118	0.151	0.278	0.251	0.330
✗	✓	0.125	0.153	0.283	0.253	0.345
✓	✗	0.129	0.158	0.290	0.265	0.362
✓	✓	**0.133**	**0.160**	**0.291**	**0.271**	**0.375**

Table 3: Ablation results on DEC losses. $\mathcal{L}_{\text{img}}$ and $\mathcal{L}_{\text{txt}}$ denote the image-evolution and report-evolution consistency losses, respectively.

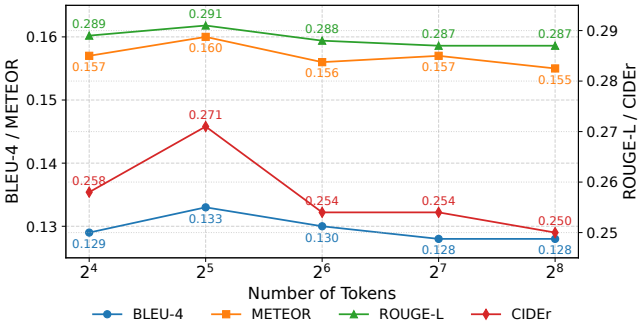

Figure 4: Ablation results on the number of learnable queries in the ARA module.

Ablation Study

Effect of Adaptive Region Alignment. We evaluate the contribution of the ARA module by comparing the model with and without it. As shown in Table 2, removing ARA consistently degrades performance, highlighting the importance of spatial alignment. Figure 3 further illustrates this effect through attention heatmaps on prior and current chest X-ray images. Before applying ARA (Figures (a)–(b)), which depict the attention maps from the last layer of the vision encoder, the attention is scattered and often covers irrelevant regions, reflecting weak spatial alignment and noisy cross-time focus. After applying ARA (Figures (c)–(d)), which show the attention from the ARA latent queries to all image patches, the attention concentrates on clinically meaningful areas (e.g., lesion regions, lung fields) and maintains better consistency across longitudinal image pairs. Notably, in Example 2, effective alignment is achieved even between images from different views, demonstrating ARA's robustness. These results confirm that ARA effectively guides the model to align visual focus across time, thereby enhancing relation modeling and report generation.

Effect of Dual Evolution Consistency Losses. To assess the effectiveness of the proposed DEC losses, we perform an ablation study on its two components: $\mathcal{L}_{\text{img}}$ (image-evolution consistency) and $\mathcal{L}_{\text{txt}}$ (report-evolution consistency). As shown in Table 3, removing both yields the weakest performance, highlighting the importance of evolution-aware relation learning. Adding either loss improves results, with $\mathcal{L}_{\text{img}}$ outperforming $\mathcal{L}_{\text{txt}}$. This is likely because visual evolution captures changes more completely, whereas follow-up reports in MIMIC-CXR often omit unchanged findings and describe only variations, making textual evo-

<table>
<tr><td>Prior Image
  </td><td colspan="2">Prior Report: impression : mild central pulmonar vascular engorgement. findings : no focal consolidation pleural effusion or pneumothorax is detected. heart and mediastinal contours are unchanged compared to prior with mild central pulmonary vascular engorgement. elevation of the right hemidiaphragm is again noted. single-lead pacer is seen in similar position.</td></tr>
<tr><td>Current Image
  </td><td colspan="2">Current Report (Ground Truth): as compared to the previous radiograph the patient has been intubated. the tip of the endotracheal tube projects 3 cm above the carina. a left pectoral pacemaker is in unchanged position. in the interval lung volumes have substantially decreased there are signs indicative of mild-to-moderate pulmonary edema and atelectasis at both lung bases. no evidence of pneumonia. short-term followup with chest radiographs is required.</td></tr>
<tr><th>Method</th><th>Generated Report</th><th>Key Observations & Differences</th></tr>
<tr><td>R2GenGPT (Baseline)</td><td>impression : mild central pulmonary vascular engorgement. no focal consolidation pleural effusion or pneumothorax is detected. findings : a left-sided pacemaker device is seen in unchanged position with leads terminating in the right atrium and right ventricle. heart size is normal . mediastinal and hilar contours are unchanged. there is mild central pulmonary vascular engorgement. no focal consolidation pleural effusion or pneumothorax is detected.</td><td>❌ No mention of intubation, tube, pulmonary edema or atelectasis.
✅ Captures pacemaker and stable findings.
⚠️ Fails to reflect the recent progressions.</td></tr>
<tr><td>Ours</td><td>as compared to the previous radiograph from the patient has been intubated. the tip of the endotracheal tube projects approximately 3 cm above the carina. there is no evidence of pneumothorax or pleural effusion. the cardiomediastinal contours are unchanged. mild pulmonary vascular engorgement is again seen.</td><td>✅ Correctly identifies intubation, tube position.
✅ Captures cross-temporal progressions.
⚠️ No mention of pulmonary edema or atelectasis.</td></tr>
</table>

Figure 5: An illustration of reports generated by different methods. We feed longitudinal chest X-rays and historical reports into R2GenGPT as a baseline. Bold fonts indicate important symptoms mentioned in the ground truth report. Green fonts indicate correctly generated content. Purple fonts indicate common errors across different methods. Red fonts indicate important recent progressions that were not correctly generated.

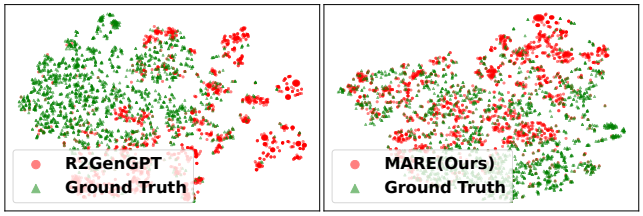

Figure 6: Visualization of feature distributions using t-SNE for the R2GenGPT and MARE (Ours) models.

lution a less complete indicator. Combining both achieves the best performance, indicating that enforcing consistency in both visual and textual spaces can complementarily yield semantically grounded and robust evolution relations, leading to more coherent reports.

Effect of the Number of Learnable Queries in ARA. We analyze the impact of the number of learnable queries, which determines the tokens retained after ARA. As shown in Figure 4, increasing the token number from 16 to 32 yields consistent gains across all metrics, with CIDEr showing the largest improvement. Further increasing the number beyond 32 offers no substantial benefit and slightly reduces performance, suggesting that too few tokens may discard essential lesion cues, while too many may reintroduce noise or misaligned background. Empirically, we set the number of learnable queries to 32 to achieve a favorable trade-off between retaining rich lesion details and minimizing irrelevant spatial noise.

Qualitative Analysis

We visualize predicted reports generated by different methods using the same longitudinal inputs for comparative analysis. As shown in Figure 5, R2GenGPT (Wang et al. 2023) captures stable findings (e.g., pacemaker positions) but fails to reflect disease progression and key changes such as intubation or pulmonary edema. In contrast, our method, with disease evolution modeling, accurately identifies core findings like intubation and tube positions. Nonetheless, both methods fail to identify pulmonary edema and atelectasis, likely due to the subtle nature of these pathologies and the visual encoder's difficulty in capturing them. To further intuitively demonstrate the impact of our framework, we visualize latent feature distributions using t-SNE (Figure 6). Compared to R2GenGPT, our MARE model produces features more closely aligned with ground truth clusters. This demonstrates that our analogical reasoning framework improves not only textual coherence but also semantic alignment in latent space.

Conclusions

In this paper, we introduced MARE, which offers a new perspective on LRRG by reformulating it as a multimodal analogical reasoning task within a unified Abduction–Mapping–Induction framework. MARE performs abduction to infer disease-evolution relations from longitudinal images, maps these relations into the report feature space, and induces the current report with an LLM reasoning engine, conditioned on the mapped relations and the historical report. To more accurately capture disease evolution relations, we designed an ARA mechanism to address spatial misalignment and introduced DEC losses to explicitly supervise the learning of inter-image and inter-report relations. Extensive experiments demonstrate that MARE effectively models disease evolution and achieves state-of-the-art performance in generating clinically meaningful reports.

Acknowledgments

This work was supported by National Natural Science Foundation of China (62441232, 62476179), Capital's Funds for Health Improvement and Research (CFH2024-2-2054), and Beijing Natural Science Foundation (L258053).

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Bannur, S.; Hyland, S.; Liu, Q.; Perez-Garcia, F.; Ilse, M.; Castro, D. C.; Boecking, B.; Sharma, H.; Bouzid, K.; Thieme, A.; et al. 2023. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15016–15027.

Cai, H.; Shen, X.; Li, S.; Shen, W.; and Xu, Q. 2025. Enhancing Multimodal Analogical Reasoning Through Triplet Interaction. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Chen, J.; Xu, R.; Fu, Z.; Shi, W.; Li, Z.; Zhang, X.; Sun, C.; Li, L.; Xiao, Y.; and Zhou, H. 2022. E-KAR: A Benchmark for Rationalizing Natural Language Analogical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, 3941–3955.

Chen, Z.; Shen, Y.; Song, Y.; and Wan, X. 2021. Cross-modal memory networks for radiology report generation. In *Proceedings of Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, 5904–5914.

Chen, Z.; Song, Y.; Chang, T.-H.; and Wan, X. 2020. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1439–1449.

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

Gentner, D. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2): 155–170.

Hayes, T. L.; and Kanan, C. 2021. Selective replay enhances learning in online continual analogical reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3502–3512.

Holyoak, K. J.; and Thagard, P. 1996. *Mental leaps: Analogy in creative thought*. MIT press.

Hou, W.; Cheng, Y.; Xu, K.; Li, W.; and Liu, J. 2023. RECAP: Towards Precise Radiology Report Generation via Dynamic Disease Progression Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2134–2147. Singapore: Association for Computational Linguistics.

Hu, S.; and Clune, J. 2023. Thought cloning: Learning to think while acting by imitating human thinking. *Advances in Neural Information Processing Systems*, 36: 44451–44469.

Hu, S.; Ma, Y.; Liu, X.; Wei, Y.; and Bai, S. 2021. Stratified rule-aware network for abstract visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1567–1574.

Huang, L.; Wang, W.; Chen, J.; and Wei, X.-Y. 2019. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4634–4643.

Huang, X.; Chen, W.; Liu, J.; Lu, Q.; Luo, X.; and Shen, L. 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*, volume 39, 3769–3778.

Huang, Z.; Zhang, X.; and Zhang, S. 2023. Kiut: Knowledge-injected u-transformer for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19809–19818.

Jin, H.; Che, H.; Lin, Y.; and Chen, H. 2024. Promptmrg: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, 2607–2615.

Johnson, A. E.; Pollard, T. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Peng, Y.; Lu, Z.; Mark, R. G.; Berkowitz, S. J.; and Horng, S. 2019. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Li, M.; Lin, B.; Chen, Z.; Lin, H.; Liang, X.; and Chang, X. 2023a. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3334–3343.

Li, M.; Lin, H.; Qiu, L.; Liang, X.; Chen, L.; Elsaddik, A.; and Chang, X. 2024. Contrastive Learning with Counterfactual Explanations for Radiology Report Generation. *arXiv preprint arXiv:2407.14474*.

Li, Y.; Yang, B.; Cheng, X.; Zhu, Z.; Li, H.; and Zou, Y. 2023b. Unify, align and refine: Multi-level semantic alignment for radiology report generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2863–2874.

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

Liu, C.; Tian, Y.; Chen, W.; Song, Y.; and Zhang, Y. 2024. Bootstrapping Large Language Models for Radiology Report Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, 18635–18643.

Liu, F.; Wu, X.; Ge, S.; Fan, W.; and Zou, Y. 2021a. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 13753–13762.

Liu, T.; Wang, J.; Hu, Y.; Li, M.; Yi, J.; Chang, X.; Gao, J.; and Yin, B. 2025. HC-LLM: Historical-constrained large language models for radiology report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 5595–5603.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022.

Małkiński, M.; and Mańdziuk, J. 2025. Deep learning methods for abstract visual reasoning: A survey on raven's progressive matrices. *ACM Computing Surveys*, 57(7): 1–36.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Minnameier, G. 2010. Abduction, induction, and analogy: on the compound character of analogical inferences. In *Model-based reasoning in science and technology: Abduction, logic, and computational discovery*, 107–119. Springer.

Nicolson, A.; Dowling, J.; and Koopman, B. 2023. Improving chest X-ray report generation by leveraging warm starting. *Artificial intelligence in medicine*, 144: 102633.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL)*, 311–318.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Prade, H.; and Richard, G. 2021. Analogical Proportions: Why They Are Useful in AI. In *IJCAI*, volume 2021, 4568–4576.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Reed, S. E.; Zhang, Y.; Zhang, Y.; and Lee, H. 2015. Deep Visual Analogy-Making. In Cortes, C.; Lawrence, N.; Lee, D.; Sugiyama, M.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Sanjeev, S.; Maani, F. A.; Abzhanov, A.; Papineni, V. R.; Almakky, I.; Papież, B. W.; and Yaqub, M. 2024. TiBiX: Leveraging Temporal Information for Bidirectional X-ray and Report Generation. In *MICCAI Workshop on Deep Generative Models*, 169–179. Springer.

Serra, F. D.; Wang, C.; Deligianni, F.; Dalton, J.; and O'Neil, A. Q. 2023. Controllable chest x-ray report generation from longitudinal representations. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Shen, H.; Pei, M.; Liu, J.; and Tian, Z. 2024. Automatic Radiology Reports Generation via Memory Alignment Network. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, 4776–4783.

Smit, A.; Jain, S.; Rajpurkar, P.; Pareek, A.; Ng, A. Y.; and Lungren, M. P. 2020. CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Wang, F.; Du, S.; and Yu, L. 2024. Hergen: Elevating radiology report generation with longitudinal data. In *European Conference on Computer Vision*, 183–200. Springer.

Wang, Z.; Liu, L.; Wang, L.; and Zhou, L. 2023. R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology*, 1(3): 100033.

Wang, Z.; Sun, Y.; Li, Z.; Yang, X.; Chen, F.; and Liao, H. 2025. Llm-rg4: Flexible and factual radiology report generation across diverse input contexts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 8250–8258.

Webb, T.; Holyoak, K. J.; and Lu, H. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9): 1526–1541.

Xiao, T.; Shi, L.; Liu, P.; Wang, Z.; and Bai, C. 2025. Radiology report generation via multi-objective preference optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 8664–8672.

Yasunaga, M.; Chen, X.; Li, Y.; Pasupat, P.; Leskovec, J.; Liang, P.; Chi, E. H.; and Zhou, D. 2023. Large language models as analogical reasoners. *arXiv preprint arXiv:2310.01714*.

Zhang, N.; Li, L.; Chen, X.; Liang, X.; Deng, S.; and Chen, H. 2023. Multimodal Analogical Reasoning over Knowledge Graphs. In *The Eleventh International Conference on Learning Representations*.

Zhang, X.; Shi, Y.; Ji, J.; Zheng, C.; and Qu, L. 2025. MEP-Net: Medical Entity-Balanced Prompting Network for Brain CT Report Generation. In Walsh, T.; Shah, J.; and Kolter, Z., eds., *AAAI*, 25940–25948. AAAI Press.

Zhu, Q.; Mathai, T. S.; Mukherjee, P.; Peng, Y.; Summers, R. M.; and Lu, Z. 2023. Utilizing longitudinal chest x-rays and reports to pre-fill radiology reports. In *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 189–198.