HEMP: Hierarchical Entity-aligned Memory Prompting for CT Report Generation

Abstract

Automated CT report generation remains challenging because it requires both fine-grained grounding of pathological entities in multi-view volumetric evidence and clinically coherent language that reflects diagnostic priors from historical cases. Existing methods often rely on noisy cross-modal alignment or underutilize retrieved report knowledge during decoding. We propose HEMP, which is a Hierarchical Entity-aligned Memory Prompting framework for LLM-based CT report generation. HEMP first performs two-stage entity-visual alignment where graph-constrained entity-conditioned attention retrieves initial evidence and an optimal transport refinement enforces principled coupling between entity nodes and visual tokens under marginal constraints. To inject case-conditioned linguistic priors, we design a Conditional Dual-Hopfield Memory module that retrieves case-conditioned report priors and integrates them through both implicit feature-level fusion into entity representations and explicit memory tokens appended to the prompt. Finally, we enforce clinical consistency via hierarchical severity supervision at the entity, region, and sign-group levels. Comprehensive Experiments demonstrate that our method outperforms previous methods and achieves SoTA performance.

1 Introduction

Radiology reports are the primary interface between medical imaging and clinical decision-making. Automating report generation can reduce workload and improve consistency, but remains difficult because models must (i) ground fine-grained pathology entities in complex visual evidence and (ii) generate clinically coherent language that follows reporting conventions. These challenges are amplified for CT, where abnormalities can be subtle, distributed across regions, and often require integrating multi-view volumetric cues with medical priors (Shi et al., 2023; Zhang et al., 2023b).

Existing MRG methods improve parts of this pipeline but leave key gaps. Alignment-centric approaches enhance entity awareness yet can produce noisy or unstable entity-token correspondences on dense volumetric inputs (Yan et al., 2021; Shi et al., 2023). Memory-augmented and RL-based models can improve completeness, but retrieved content is often weakly conditioned on the current case and underused by the decoder (Chen et al., 2021; Qin and Song, 2022). LLM adaptations strengthen fluency, but without reliable grounding and clinically structured supervision they may hallucinate findings or yield inconsistent severity statements (Chen et al., 2024; Yan et al., 2023; Zhou et al., 2024).

We propose HEMP, a Hierarchical Entity-aligned Memory Prompting framework designed to address the complexities of CT report generation through the integration of three specialized components. The architecture utilizes a two-stage entity-visual alignment that performs graph-constrained entity-conditioned association followed by an optimal transport refinement to enforce principled coupling under marginal constraints. To enrich the linguistic output, a Conditional Dual-Hopfield Memory module retrieves case-conditioned report priors and injects them into the prompt via both implicit feature fusion and explicit memory tokens. Furthermore, the framework incorporates hierarchical severity supervision to regularize clinical consistency across the entity, region, and sign-group levels.

Our contributions are:

- A unified hierarchical entity-aligned memory prompting framework for CT report generation that jointly improves grounding, case-conditioned priors, and clinical consistency.
- A robust two-stage entity-visual alignment strategy combining graph-constrained attention with OT-based coupling for stable entity-grounded representations.
- A dual-form memory injection mechanism

and hierarchical severity objective that strengthen LLM conditioning and reduce clinically inconsistent generations.

2 Related Work

The rapid development of intelligent healthcare has established Medical Report Generation (MRG) as a critical yet challenging task (Zhang et al., 2023a; Shen et al., 2024; Zheng et al., 2024a). Earlier research in this field primarily focused on contrastive learning to refine visual-textual alignment, aiming to bridge the semantic gap between medical images and their corresponding descriptions. For instance, Yan et al. (Yan et al., 2021) introduced weakly supervised contrastive learning to enhance clinical correctness. To achieve more precise alignment, subsequent works such as those by Shi et al. (Shi et al., 2023) and Li et al. (Li et al., 2023b) incorporated pathological or dynamic graphs to align fine-grained visual features with textual representations. Further building on these graph-driven approaches, Shi et al. (Shi et al., 2024) utilized prior tissue knowledge to ensure anatomically consistent cross-modal feature alignment. While these methods significantly improve the correlation between modalities, they often treat image-text pairs as static entities, struggling to capture the long-term clinical dependencies and diagnostic logic.

To address the limitations of static alignment, a subsequent paradigm optimizes feature integration through memory mechanisms and reinforcement learning, focusing on the dynamic storage of clinical patterns and the direct optimization of generation quality. Chen et al. () proposed a Cross-Modal Memory Network to record and retrieve historical alignment patterns for more consistent reports. Building upon this, Qin et al. (Qin and Song, 2022) combined reinforcement learning with cross-modal memory to align visual and textual features while optimizing for clinical metrics. XProNet (Wang et al., 2022) further evolved this concept by utilizing prototype-driven learning to capture complex cross-modal patterns. Despite these architectural refinements, these specialized models are often constrained by their limited parameter space, making it difficult to generalize to complex cases that require extensive medical knowledge.

More recently, the research focus has shifted toward large language model based systems, which leverage massive pre-trained knowledge to streamline report synthesis and reasoning. Frameworks such as Dia-LLaMA (Chen et al., 2024) fine-tune LLaMA2 (Touvron et al., 2023) using disease labels as knowledge prompts to guide the generative process. To further improve flexibility and address data imbalance, other LLM-driven approaches (Yan et al., 2023; Zhou et al., 2024) have integrated disease-specific prompts with clinical reinforcement learning. However, these LLM-based methods are predominantly designed for 2D images and often fail to model the fine-grained entity relationships required for volumetric data explicitly.

In contrast to these existing methods, our HEMP introduces a hierarchical entity-aligned memory prompting paradigm specifically for CT scans. By utilizing optimal transport for fine-grained alignment and dual-hopfield memories to inject clinical priors, HEMP activates the inner power of LLMs to bridge the gap between volumetric visual evidence and complex diagnostic reasoning.

3 Methods

An overview of HEMP is shown in Fig. 1, which comprises three modules: (i) Entity-Visual Alignment grounds pathology entities in multi-view CT evidence to produce aligned entity features (see Section 3.1); (ii) Conditional Dual-Hopfield Memory retrieves case-conditioned report priors and injects them as implicit feature fusion and explicit prompt tokens (see Section 3.2); and (iii) Hierarchical Severity Modeling adds multi-level supervision to enforce clinically consistent representations and generation (see Section 3.3).

3.1 Entity-Visual Alignment

To address fine-grained alignment between pathology entities and case specific imaging features in report generation, we adopt a two-stage alignment framework to capture entity relevant visual evidence from multi-view image representations, where medical semantic relations among entities are incorporated via graph constraints. The aligned entity representations are then used for prompt construction.

Attention-Based Entity-Visual Association As the first stage of entity–visual alignment, we follow (Zhang et al., 2023b) and utilize a visual encoder (fine-tuned ResNet101 (Zhang et al., 2025) for brain CT and fine-tuned CT-CLIP (Hamamci et al., 2026) for chest CT) to obtain scan features $V_f \in \mathbb{R}^{N\times P\times d_s}$ for each sample, where N is the number of views, P is the number of pooled patches per view, and d_s is the feature-channel dimension. We then use a visual adaptor to map V_f

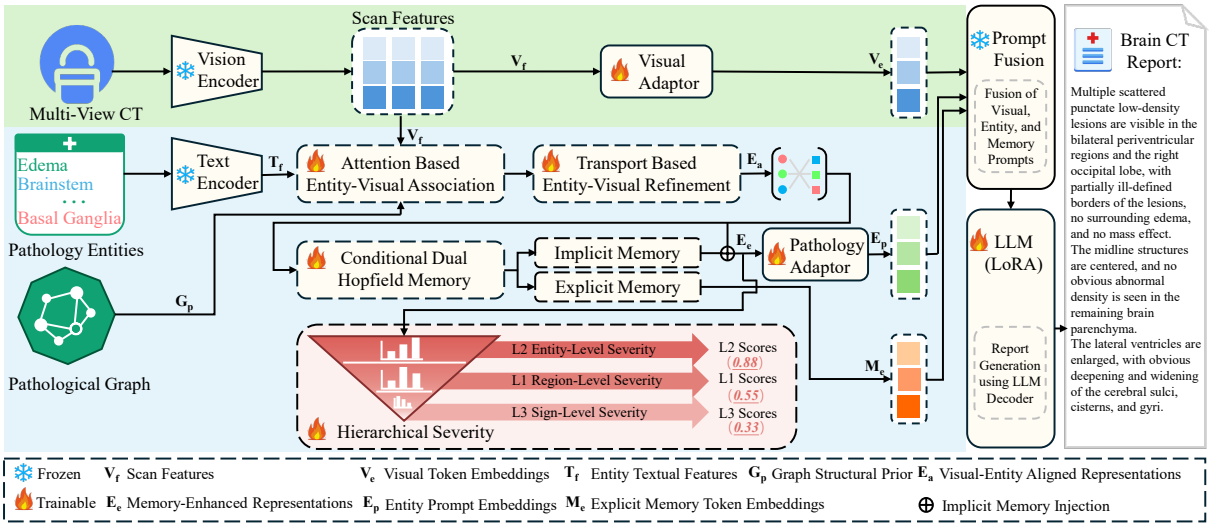

Figure 1: Architecture of HEMP with entity-visual alignment, memory-augmented prompting, and hierarchical severity modeling.

into the LLM word-embedding space of dimension d_w, yielding view-level visual prompt embeddings:

$$V_e = (V_f W_{v_1}^\top + b_{v_1}) W_{v_2}^\top + b_{v_2}, \quad (1)$$

where W_{v_1} and W_{v_2} are learnable projection matrices, and b_{v_1} and b_{v_2} are learnable biases. The resulting visual embeddings V_e are inserted into predefined visual slots in a prompt template. Next, We adopt an attention-based association mechanism to establish entity conditioned visual representations. Specifically, to provide an initial correspondence between pathology entities and visual evidence prior to transport-based coupling, we follow the entity-conditioned prompting routine (Zhang et al., 2025), which combines entity-to-visual cross-attention with adjacency-constrained graph propagation over entity nodes. Given multi-view visual features $V_f \in \mathbb{R}^{B\times N\times P\times d_s}$, they are reshaped into $M = N \cdot P$ visual tokens and mapped to the shared hidden size:

$$X_v = \psi(V_f^{\text{flat}}) \in \mathbb{R}^{B\times M\times d_h}, \quad (2)$$

where $V_f^{\text{flat}} \in \mathbb{R}^{B\times M\times d_s}$ and $\psi : \mathbb{R}^{d_s} \to \mathbb{R}^{d_h}$ denotes a learnable projection applied to each token, d_s denotes the backbone channel dimension and d_h is the shared hidden size used by the prompting module. A pathology entity set $E = \{e_0, e_1, \ldots, e_K\}$ is predefined (Shi et al., 2023), where e_0 denotes a global entity token. Entity textual features $T_f \in \mathbb{R}^{B\times(K+1)\times d_h}$ are obtained from the text encoder, with pooling applied for multi-token entity names. To incorporate structural priors, an adjacency mask $A = \text{Adj}(G_p) \in \mathbb{R}^{(K+1)\times(K+1)}$ is induced from the pathological graph G_p (Shi et al., 2023). Then, the entity-conditioned visual evidence E_0 and graph-constrained embeddings E_1 are computed as:

$$E_0 = \text{MHA}(T_f, X_v, X_v) \in \mathbb{R}^{B\times(K+1)\times d_h}, \quad (3)$$

$$E_1 = \text{MHA}(E_0, E_0, E_0; A) \in \mathbb{R}^{B\times(K+1)\times d_h}, \quad (4)$$

In practice, we employ L_g stacked graph-attention layers equipped with position-wise feed-forward networks and residual connections. The resulting graph-conditioned entity embeddings are then passed to the transport-based refinement stage.

Transport-Based Entity-Visual Refinement As the second alignment stage, we introduce an optimal-transport (OT) formulation (Monge, 1781) to explicitly model the coupling between entity representations and visual tokens under marginal constraints. Let $E_{1,c} \in \mathbb{R}^{B\times K\times d_h}$ denote the entity embeddings after removing the global node e_0 from E_1, and let $V \in \mathbb{R}^{B\times M\times d_h}$ denote the corresponding visual tokens. Both modalities are projected into a shared OT space of dimension d_o and ℓ_2-normalized $\tilde{E} = \text{Norm}(E_{1,c}W_e), \tilde{V} = \text{Norm}(VW_v)$,, where $W_e, W_v \in \mathbb{R}^{d_h\times d_o}$ are learnable linear projections. A pairwise cost matrix $C \in \mathbb{R}^{B\times K\times M}$ is defined as $C_{b,i,j} = \left\|\tilde{E}_{b,i} - \tilde{V}_{b,j}\right\|_2$, Given marginal constraints $\mu \in \mathbb{R}^K$ and $\nu \in \mathbb{R}^M$, the transport plan $T \in \mathbb{R}^{B\times K\times M}$ is obtained by solving the entropy-regularized OT problem:

$$T = \underset{T\in\mathcal{U}(\mu,\nu)}{\text{argmin}} \langle T, C\rangle - \varepsilon H(T), \quad (5)$$

with the feasible set $\mathcal{U}(\mu,\nu) = \{T \geq 0 \mid T1 = \mu,\ T^\top 1 = \nu\}$, ε is the hyper-parameter and 1 denotes an all-ones vector of compatible dimension. The resulting objective is solved by iterative Sinkhorn-Knopp iterations (Sinkhorn and Knopp, 1967),which yields an efficient scheme for entropy

regularized transport (Zhao et al., 2025). Following this formulation, we instantiate the OT objective as a training loss by computing the transport cost together with an entropy regularizer:

$$\mathcal{L}_{\mathrm{OT}} = \langle T, C \rangle - \varepsilon \sum_{i=1}^{K} H(P_{i,:}), \tag{6}$$

where $P_{i,j} = T_{i,j} / \sum_{m=1}^{M} T_{i,m}$ denotes the row-normalized transport plan. This formulation discourages overly diffuse entity-to-visual assignments during optimization. OT-aligned entity representations are derived via barycentric projection:

$$E_{b,i}^{\mathrm{OT}} = \frac{\sum_{j=1}^{M} T_{b,i,j} V_{b,j}}{\sum_{j=1}^{M} T_{b,i,j}} \in \mathbb{R}^{d_h}, \tag{7}$$

Let $E_1^{(0)} \in \mathbb{R}^{B \times 1 \times d_h}$ denote the global node embedding. The final visual-entity aligned representations are formed as:

$$E_a = \left[E_1^{(0)};\ E^{\mathrm{OT}} \right] \in \mathbb{R}^{B \times (K+1) \times d_h}, \tag{8}$$

3.2 Conditional Dual-Hopfield Memory

We introduce a Conditional Dual-Hopfield Memory module to inject case-conditioned linguistic priors into report generation by retrieving relevant historical reports based on the current case representation. Following Modern Hopfield Networks (Ramsauer et al., 2020), we pool aligned entity embeddings E_a to form a case-level query and perform case-conditioned retrieval from a key-value memory (Miller et al., 2016). Retrieval occurs in the case-representation space, while the retrieved report embeddings are aggregated in the linguistic space to produce memory-enhanced entity features E_e and explicit memory tokens M_e for prompt construction.

Implicit Memory Injection We inject retrieved report-level memory into aligned entity features as continuous auxiliary context, capturing reusable description patterns and reducing over-reliance on current visual evidence. Let $E_{a,\mathrm{c}} \in \mathbb{R}^{B \times K \times d_h}$ denote the entity representations after removing the global node e_0 from E_a, where B is the batch size, K is the number of entity nodes removing the global node e_0, and d_h is the hidden dimension. A case-level query is constructed by mean pooling:

$$q = \mathrm{Pool}(E_{a,\mathrm{c}}) \in \mathbb{R}^{B \times d_h}, \tag{9}$$

where $\mathrm{Pool}(\cdot)$ denotes mean pooling over the entity dimension. Historical cases are stored in a dual memory bank as key-value pairs $(k_n, r_n)_{n=1}^{N_{\mathrm{mem}}}$, where $k_n \in \mathbb{R}^{d_h}$ is a case key and $r_n \in \mathbb{R}^{d_h}$ is the associated report embedding. Given q, Top-k keys are retrieved by similarity and aggregated as:

$$\alpha_i = \mathrm{Softmax}\left(\frac{\mathrm{sim}(q, k_i)}{\tau} \right), \tag{10}$$

$$m_{\mathrm{imp}} = \sum_{i=1}^{k} \alpha_i r_i \in \mathbb{R}^{B \times d_h}, \tag{11}$$

where $\mathrm{sim}(\cdot, \cdot)$ denotes cosine similarity and τ is a temperature parameter. The resulting m_{imp} represents a case-conditioned report-level prior that is implicitly injected into the entity representations. A gating vector is computed as:

$$g = \sigma(m_{\mathrm{imp}} W_g + b_g) \in \mathbb{R}^{B \times d_h}, \tag{12}$$

and implicit memory injection is realized via gated residual fusion:

$$E_e = E_a + \mathrm{Expand}(g \odot m_{\mathrm{imp}}), \tag{13}$$

where $\mathrm{Expand}(\cdot)$ denotes broadcasting along the entity dimension and $\odot$ represents element-wise multiplication.

Explicit Memory Injection We introduce explicit memory prompting to make case-conditioned linguistic priors visible to the decoder during generation by injecting retrieved report-level memory into the decoding input. Given retrieved memory $m_{\mathrm{imp}} \in \mathbb{R}^{B \times d_h}$, we construct memory token embeddings via linear projection and reshaping:

$$M_e = \mathcal{R}(m_{\mathrm{imp}} W_m + b_m) \in \mathbb{R}^{B \times L_{\mathrm{mem}} \times d_w}, \tag{14}$$

where $\mathcal{R}(\cdot)$ denotes the reshaping operation, $W_m \in \mathbb{R}^{d_h \times (L_{\mathrm{mem}} d_w)}$ and $b_m \in \mathbb{R}^{L_{\mathrm{mem}} d_w}$ are learnable parameters, d_w denotes the word-embedding dimension of the language model, and L_{mem} is the number of explicit memory tokens. The resulting M_e is integrated with visual prompts and entity prompts to form a unified multimodal input embedding sequence, enabling the decoder to directly attend to retrieved report-level priors during generation.

3.3 Hierarchical Severity Modeling

To provide structured supervision for the generation of pathology-related information, we follow the insights of multi-granularity hierarchical modeling (Wang et al., 2025) to introduce a hierarchical severity modeling module that represents disease severity across multiple semantic granularities. Concretely, we adopt the anatomy-pathology hierarchy of (Shi et al., 2023) and model severity at three complementary levels: entity-level (L2) for localized findings, region-level (L1) by aggregating entity severities within anatomical regions, and sign-group-level (L3) for case-level patterns.

Together, these levels form a unified hierarchical representation to guide report generation.

Hierarchical Score Given $E_e \in \mathbb{R}^{B\times(K+1)\times d_h}$, we exclude the global node to form $E_{\text{e,c}} \in \mathbb{R}^{B\times K\times d_h}$. L2 severity scores are predicted as:

$$H_2 = E_{\text{e,c}} W_2 + b_2, \quad (15)$$

$$s^{\text{L2}} = \sigma(H_2) \in \mathbb{R}^{B\times K}, \quad (16)$$

where $W_2 \in \mathbb{R}^{d_h\times 1}$ and $b_2 \in \mathbb{R}$. L1 severity scores are deterministically derived by aggregating entity-level scores within predefined anatomical groups. Let $\mathcal{I}_g \subseteq \{1,\dots,K\}$ denote the entity index set corresponding to the g-th anatomical region. The aggregation is defined as:

$$s^{\text{L1}}_{b,g} = \frac{1}{|\mathcal{I}_g|}\sum_{i\in\mathcal{I}_g} s^{\text{L2}}_{b,i}, \quad (17)$$

where $s^{\text{L1}} \in \mathbb{R}^{B\times G}$ and G is the number of anatomical regions. L3 severity scores are predicted from pooled case-level representations. Specifically, a pooled feature is obtained as:

$$e_{\text{p}} = \text{Pool}(E_{\text{e,c}}) \in \mathbb{R}^{B\times d_h}, \quad (18)$$

which is then projected to sign-group severity scores by:

$$s^{\text{L3}} = \sigma(e_{\text{p}} W_3 + b_3) \in \mathbb{R}^{B\times S}, \quad (19)$$

where $W_3 \in \mathbb{R}^{d_h\times S}$ and $b_3 \in \mathbb{R}^S$.

Joint Supervision Objective The hierarchical severity scores are optimized jointly during model training. Among the three levels, L2 and L3 are produced by learnable prediction heads, while L1 is obtained through a deterministic aggregation operator. Let the supervision signals be denoted as $y^{\text{L2}} \in [0,1]^{B\times K}$, $y^{\text{L1}} \in [0,1]^{B\times G}$, and $y^{\text{L3}} \in [0,1]^{B\times S}$, whose construction depends on available annotations or pseudo-labeling strategies. The severity-related training objective is formulated as:

$$\mathcal{L}_{\text{sev}} = z_2 + \lambda_{L1} z_1 + \lambda_{L3} z_3. \quad (20)$$

where $z_2 = \mathcal{L}_{\text{L2}}\left(s^{\text{L2}}, y^{\text{L2}}\right)$, $z_1 = \mathcal{L}_{\text{L1}}\left(s^{\text{L1}}, y^{\text{L1}}\right)$, and $z_3 = \mathcal{L}_{\text{L3}}\left(s^{\text{L3}}, y^{\text{L3}}\right)$, with $\mathcal{L}_{\text{L2}}$ and $\mathcal{L}_{\text{L3}}$ implemented using BCE loss, $\mathcal{L}_{\text{L1}}$ instantiated as MSE loss, λ_{L1} and λ_{L3} are weighting coefficients.

4 Experiments

4.1 Dataset

We validate our model on the public CT report datasets (Tang et al., 2024), comprising CTRG-Brain (160,336 images, 6,000 reports) and CTRG-Chest (548,696 images, 1,804 reports). To ensure a fair comparison with baseline models, for CTRG-Brain, we follow (Shi et al., 2023) and partition the data into train, validation, and test sets with a $7:1:2$ ratio, for CTRG-Chest, we adhere to the partition defined in (Deng et al., 2025). Regarding preprocessing, we implement a standardized data-cleaning pipeline (Zhang et al., 2023b) for both datasets, this process removes redundant scans to ensure each sample contains exactly 24 scans, thereby aligning the data with clinical standards.

4.2 Evaluation Metrics

We use both Natural Language Generation (NLG) metrics and Clinical Evaluation (CE) metrics to assess model performance, include BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015), denoted as B1-B4, M, RG, and C, respectively. To evaluate pathological accuracy, We follow (Shi et al., 2023) and employ 24 curated brain CT clinical keywords (e.g., "basal ganglia region", "cerebral sulcus", and "stagnant blood") to compute the F1 score, defined as the harmonic mean of precision and recall. Furthermore, we incorporate the BertScore (Zhang et al., 2019), as well as specialized radiology report evaluation metrics including RadGraph and RadCliQ (Yu et al., 2023), denoted as BS, RG and RC, respectively.

4.3 Implementation Details

Our baseline follows encoder-decoder architecture. For brain CT, we employ a ResNet101 (Zhang et al., 2025) fine-tuned for hemorrhage classification on the CQ500 dataset (Chilamkurthy et al., 2018). For chest CT, we utilize CT-CLIP pre-trained on its associated large-scale CT-RATE dataset (Hamamci et al., 2026). For text generation, we employ MedGemma-4B (Sellergren et al., 2025) as the autoregressive decoder with 4-bit quantization. We further apply LoRA (Hu et al., 2022) with rank 64 for parameter-efficient fine-tuning, resulting in 0.59% trainable parameters in the LLM (25.8M). We set the feature dimensions d_h, d_s, d_w to 512, 2048, 2560. For optimal transport (OT), the entropic regularization coefficient is $\varepsilon = 0.1$, and the number of OT iterations is 100. For memory augmentation, we retrieve Top-$k = 5$ historical samples and use $L_{\text{mem}} = 4$ explicit memory tokens. The loss weights λ_{ot}, λ_{sev}, λ_{L1}, λ_{L3} are set to 0.05, 0.2, 0.05, 0.05, respectively. We train the model for 10 epochs on two RTX 4090 GPUs using AdamW with a learning rate of 8×10^{-5} and a batch size of 4.

Dataset	Methods	Publication	B1	B2	B3	B4	M	RG	C	F1
CTRG-Brain	PGCA (Shi et al., 2023)†	EMNLP 2023	50.2	37.8	30.7	26.5	32.5	43.0	34.0	69.2
	LLaVA-med (Li et al., 2023a)†	NIPS 2023	50.0	39.3	32.0	26.3	31.4	46.7	38.6	64.3
	PromptMRG (Jin et al., 2024)†	AAAI 2024	48.1	38.3	31.6	26.5	31.0	47.4	50.3	69.2
	PCRL (Zheng et al., 2024b)†	EMNLP 2024	62.0	54.7	49.4	45.3	33.1	57.7	96.4	70.6
	HILT (Liu et al., 2025)†	ACL 2025	50.7	39.9	33.1	27.9	30.8	46.1	43.7	68.4
	MEPNet (Zhang et al., 2025)†	AAAI 2025	56.7	46.9	40.7	34.4	34.6	50.4	86.6	67.8
	HEMP	Ours	**65.1**	**57.4**	**52.1**	**47.8**	**35.6**	**61.1**	**97.9**	**71.9**
CTRG-Chest	CAMANet (Wang et al., 2024)†	IJBHI 2024	54.2	45.2	39.4	35.3	26.4	54.2	-	-
	CT2Rep (Hamamci et al., 2024)†	MICCAI 2024	49.8	40.7	35.1	31.2	24.8	51.1	-	-
	Dia-LLAMA (Chen et al., 2026)†	MICCAI 2025	51.1	-	-	29.6	26.2	42.1	-	-
	LHR-RFL (Yi et al., 2025)†	TMI 2025	55.5	44.7	37.6	32.6	25.4	47.7	-	-
	Reg2RG (Chen et al., 2025)†	TMI 2025	49.6	41.4	35.9	32.0	**49.7**	47.7	-	-
	MvKeTR (Deng et al., 2025)†	IJBHI 2025	58.3	48.7	42.4	37.8	28.3	54.2	-	-
	HEMP	Ours	**58.6**	**52.5**	**48.5**	**45.3**	33.8	**65.0**	-	-

Table 1: The performance of our model compared with previous state-of-the-art models on the CT report generation dataset. The best result is highlighted in bold, and the second-best result is underlined. † marks the results that are cited from reference paper.

Model	Dataset	BS-F1	RG-F1	RC
HEMP	CTRG-Brain	0.867	0.625	0.552
	CTRG-Chest	0.839	0.620	0.537

Table 2: Semantic and clinical evaluation of HEMP.

4.4 Comparison Studies

We compare our proposed model with state-of-the-art methods for CT report generation. These methods are categorized into four groups based on their technical emphases: contrastive learning (LHR-RFL (Yi et al., 2025),), cross-modal integration (PGCA (Shi et al., 2023), PCRL (Zheng et al., 2024b), CAMANet (Wang et al., 2024), CT2Rep (Hamamci et al., 2024), HILT (Liu et al., 2025), MvKeTR (Deng et al., 2025)), entity-focused optimization (MEPNet (Zhang et al., 2025)), and large language model adaptation (LLaVA-Med (Li et al., 2023a), PromptMRG (Jin et al., 2024), Dia-LLAMA (Chen et al., 2026), Reg2RG (Chen et al., 2025)).

As illustrated in Table 1, our model achieves the best results on CTRG-Brain across all metrics, indicating higher report quality and stronger clinically relevant content coverage. Specifically, improvements in BLEU and ROUGE-L reflect better lexical fidelity, while gains in CIDEr and F1 scores suggest more accurate and comprehensive reporting of salient findings. Furthermore, HEMP demonstrates consistent performance advantages on the CTRG-Chest dataset. Although Reg2RG remains competitive in the METEOR metric, HEMP maintains its leadership in capturing structured diagnostic information, as evidenced by its superior BLEU and ROUGE-L scores. These advantages are attributed to refined entity representations, case-conditioned report-memory prompting, and hierarchical severity supervision. Compared with baselines such as contrastive learning, cross-modal integration, entity-focused, and LLM adaptation methods, our design provides stronger case-specific conditioning and improved long-range coherence. As shown in Table 2, we further evaluate HEMP's performance regarding semantic alignment and clinical consistency. The results show that HEMP achieves promising scores across BertScore-F1, RadGraph-F1, and RadCliQ metrics for both datasets, with BertScore-F1 exceeding 0.83 and RadGraph-F1 surpassing 0.62.

4.5 Ablation Studies

Table 3 presents a macro-to-micro ablation analysis of our framework, isolating the contributions of entity-visual alignment, hopfield retrieval memory, and hierarchical severity supervision. Starting from a plain encoder-decoder baseline, each component yields consistent gains, and the full model achieves the best overall performance across both NLG metrics and clinical fidelity, indicating modular complementarity.

Our systematic ablation experiments validate that each proposed component addresses specific modes including visual-semantic misalignment, narrative impoverishment, and clinical logical inconsistency. Concretely, The Entity-Visual Alignment module establishes the fundamental mapping between visual features and clinical entities. (a) initiates evidence grounding to yield a 2.0% im-

Method	Entity-Visual Alignment		Hopfield Memory		Hierarchical Severity		B4	M	RG	C	F1
	Attention	Transport	Implicit	Explicit	L2	L1 L3					
Baseline							39.0	30.2	55.5	78.8	62.2
(a)	✓						40.5	31.5	56.8	80.8	64.2
(b)	✓	✓					41.6	32.1	57.1	81.4	64.5
(c)	✓	✓	✓				45.8	34.4	57.2	88.0	66.2
(d)	✓	✓		✓			45.9	34.2	57.1	87.7	66.6
(e)	✓	✓	✓	✓			46.3	34.5	57.3	90.2	66.8
(f)	✓	✓			✓		44.0	33.3	57.0	83.2	67.1
(g)	✓	✓			✓	✓	44.2	33.5	57.0	83.5	67.3
(h)	✓	✓	✓		✓	✓	46.5	35.1	57.8	94.8	69.1
(i)	✓	✓		✓	✓	✓	46.6	34.8	57.6	94.6	69.7
Ours	✓	✓	✓	✓	✓	✓	**47.8**	**35.6**	**61.1**	**97.9**	**71.9**

Table 3: Ablation studies of our proposed method on the CTRG-Brain. The Baseline model is an encoder-decoder framework.

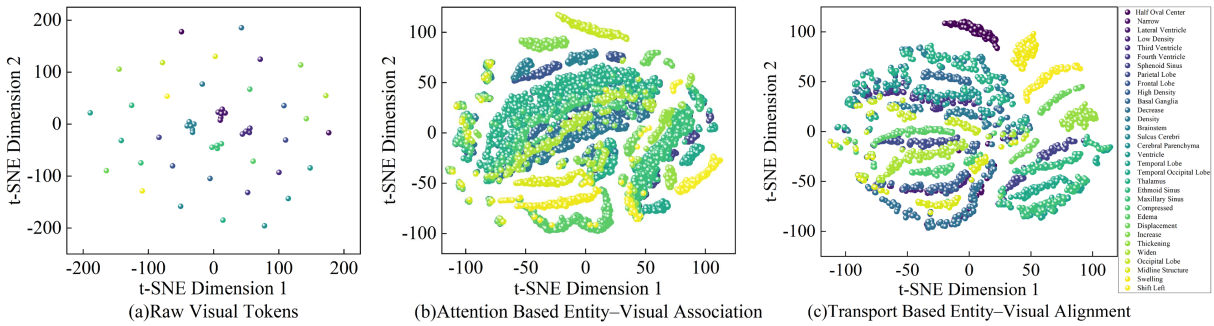

Figure 2: Entity–Visual Alignment Process Across Progressive Stages on the CTRG-Brain.

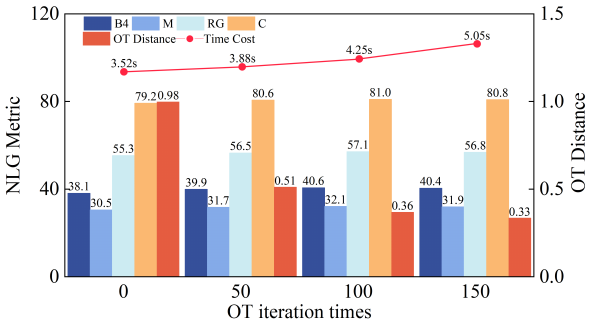

Figure 3: The Impact of Iteration Counts on the OT Module.

provement in F1, (b) enforces structural constraints to minimize spatial ambiguity, thereby enhancing BLEU-4 and CIDEr metrics through more alignment. The Hopfield Memory module significantly enriches the professional quality and content of the generated reports. (c) augments latent semantic representations to drive notable gains in METEOR, ROUGE-L, and CIDEr. Conversely, (d) serves as a structural prompt that refines linguistic quality while providing a specialized boost to the F1, indicating that exemplar-based prompting helps the model recover more accurate clinical labels in the text. The Hierarchical Severity module remains the primary contributor to diagnostic fidelity. (f) achieves a substantial leap of 2.6% in F1, demonstrating that explicit severity classification effectively calibrates the model's diagnostic judgment. Furthermore, (g) incorporates anatomical priors to ensure logical coherence between localized findings and global regional status, further stabilizing the clinical correctness.

4.6 Qualitative Studies

We further assess our two-stage Entity–Visual Alignment in Fig. 2. Raw visual tokens are highly scattered and heavily mixed, indicating that the vision encoder alone does not produce entity-discriminative structure. After Attention-Based Entity-Visual Association, embeddings become more organized with local aggregation, suggesting that cross-attention retrieves entity-conditioned evidence and graph propagation injects medical semantic relations. With transport-based entity-visual refinement, clusters become more compact with clearer inter-class separation, implying that OT coupling reallocates entity–token correspondences under marginal constraints and anchors each entity embedding to relevant visual evidence via barycentric projection.

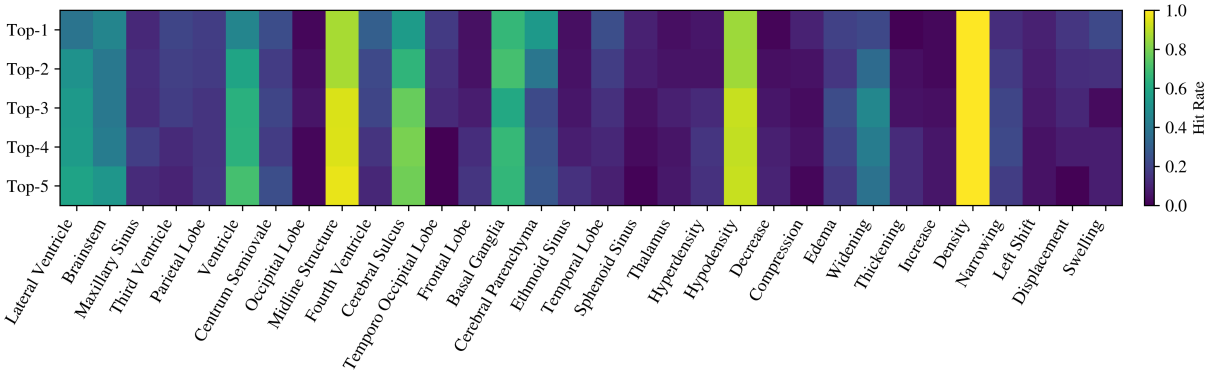

Figure 4: Entity-level Semantic Coverage of Top-k Memory Retrieval on the CTRG-Brain.

We also study the effect of OT iteration count. As shown in Fig. 3, increasing iterations consistently reduces the transport distance, indicating improved alignment. NLG metrics improve markedly from 0 to 100 iterations, but then saturate, while computation increases steadily and rises sharply beyond 100. We therefore adopt 100 iterations as a favorable accuracy–efficiency trade-off.

We evaluate memory retrieval using an entity-level semantic coverage metric, which measures entity overlap between retrieved reports and the target case. Entities are extracted from the reference report, and Top-k hit rates are computed to form an entity heatmap. As shown in Fig 4, coverage increases with k, indicating that retrieval returns topic-relevant reports that support generation. A stricter entity-level faithfulness analysis is provided in the supplementary material.

As illustrated in Fig. 5, the cosine similarity exhibits a robust monotonic decrease from Top-1 to Top-5, confirming the retrieval mechanism's ability to precisely rank cases according to their relevance. Crucially, the fusion weights are distributed uniformly and decay smoothly rather than collapsing into a single-memory dominance. Simultaneously, evaluation results in Fig. 6 reveal that both Micro-F1 and Macro-F1 scores increase significantly as k grows from 1 to 5, reaching a plateau near $k = 5$. The consistent margin of Micro-F1 over Macro-F1 reflects that the retrieval-driven gains are primarily propelled by high-frequency clinical entities, whereas the enhancement for long-tail entities remains relatively constrained. Consequently, $k = 5$ is adopted as the retrieval size for subsequent experiments to strike an optimal balance between information richness and computational efficiency.

Additional technical details including model code, hierarchical severity modeling analysis, and extensive parameter sensitivity studies (e.g., loss weights and memory hyperparameters) are provided in the supplementary material.

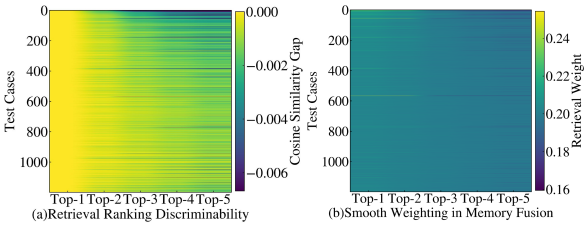

Figure 5: Reliability and Stability of Memory Retrieval and Fusion on the CTRG-Brain.

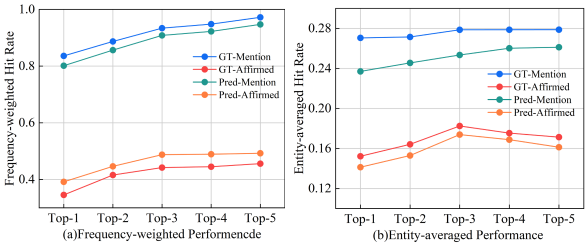

Figure 6: Top-k Memory Retrieval Performance on the CTRG-Brain.

5 Limitations

While HEMP demonstrates promising performance in brain and chest CT report generation, its entity-visual alignment module relies on pre-defined clinical graph structures, which partially limits the model's generalization capability when encountering unseen anatomical structures.

6 Conclusion

We present HEMP, a hierarchical entity-aligned memory prompting framework for LLM-based CT report generation. HEMP integrates two-stage cross-modal alignment, a case-conditioned Conditional Dual-Hopfield Memory module, and hierarchical severity supervision to resolve key failure modes of prior MRG systems. Extensive evaluations on the CTRG-Brain and CTRG-Chest benchmarks demonstrate that HEMP outperforms state-of-the-art methods, while ablations validate the complementary benefits of its core components.

References

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)*, pages 5904–5914, Online. Association for Computational Linguistics.

Zhixuan Chen, Yequan Bie, Haibo Jin, and Hao Chen. 2025. Large language model with region-guided referring and grounding for ct report generation. *IEEE Transactions on Medical Imaging*.

Zhixuan Chen, Luyang Luo, Yequan Bie, and Hao Chen. 2024. Dia-llama: Towards large language model-driven CT report generation. *CoRR*, abs/2403.16386.

Zhixuan Chen, Luyang Luo, Yequan Bie, and Hao Chen. 2026. Dia-llama: Towards large language model-driven ct report generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pages 141–151, Cham. Springer Nature Switzerland.

Sasank Chilamkurthy, Rohit Ghosh, Swetha Tanamala, Mustafa Biviji, Norbert G. Campeau, Vasantha Kumar Venugopal, Vidur Mahajan, Pooja Rao, and Prashant Warier. 2018. Development and validation of deep learning algorithms for detection of critical findings in head ct scans. *arXiv preprint arXiv:1803.05854*.

Xiwei Deng, Xianchun He, Jianfeng Bao, Yudan Zhou, Shuhui Cai, Congbo Cai, and Zhong Chen. 2025. Mvketr: Chest ct report generation with multi-view perception and knowledge enhancement. *IEEE Journal of Biomedical and Health Informatics*, 29(11):8279–8292.

Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. 2024. CT2Rep: Automated Radiology Report Generation for 3D Medical Imaging . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15012. Springer Nature Switzerland.

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, Weicheng Dai, Murong Xu, Hadrien Reynaud, Muhammed Furkan Dasdelen, Bastian Wittmann, Tamaz Amiranashvili, Enis Simsar, Mehmet Simsar, Emine Bensu Erdemir, and 11 others. 2026. Generalist foundation models from a multimodal dataset for 3d computed tomography. *Nature Biomedical Engineering*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR 2022*.

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. 2024. Promptmrg: Diagnosis-driven prompts for medical report generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:2607–2615.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*.

Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. 2023b. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3334–3343.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Che Liu, Zhongwei Wan, Yuqi Wang, Hui Shen, Haozhe Wang, Kangyu Zheng, Mi Zhang, and Rossella Arcucci. 2025. Argus: Benchmarking and enhancing vision-language models for 3D radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16448–16460, Vienna, Austria. Association for Computational Linguistics.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1409, Austin, Texas. Association for Computational Linguistics.

G. Monge. 1781. *Mémoire sur la théorie des déblais et des remblais*. Imprimerie royale.

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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Han Qin and Yan Song. 2022. Reinforced cross-modal alignment for radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 448–458.

Hubert Ramsauer, Bernhard Schafl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber,

Markus Holzleitner, Milena Pavlović, Geir Kjetil Ferkingstad Sandve, Victor Greiff, David P. Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. 2020. Hopfield networks is all you need. *ArXiv*, abs/2008.02217.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Qingya Shen, Yanzhao Shi, Xiaodan Zhang, Junzhong Ji, Ying Liu, and Huimin Xu. 2024. GHCL: gaussian heuristic curriculum learning for brain CT report generation. *Multim. Syst.*, 30(2):69.

Yanzhao Shi, Junzhong Ji, Xiaodan Zhang, Ying Liu, Zheng Wang, and Huimin Xu. 2024. Prior tissue knowledge-driven contrastive learning for brain CT report generation. *Multim. Syst.*, 30(2):98.

Yanzhao Shi, Junzhong Ji, Xiaodan Zhang, Liangqiong Qu, and Ying Liu. 2023. Granularity matters: Pathological graph-driven cross-modal alignment for brain CT report generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6617–6630, Singapore. Association for Computational Linguistics.

Richard Sinkhorn and Paul Knopp. 1967. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21:343–348.

Yuhao Tang, Haichen Yang, Liyan Zhang, and Ye Yuan. 2024. Work like a doctor: Unifying scan localizer and dynamic generator for automated computed tomography report generation. *Expert Systems with Applications*, 237:121442.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Futian Wang, Yuhan Qiao, Xiao Wang, Fuling Wang, Yuxiang Zhang, and Dengdi Sun. 2025. R2genkg: Hierarchical multi-modal knowledge graph for llm-based radiology report generation. *ArXiv*, abs/2508.03426.

Jun Wang, Abhir Bhalerao, and Yulan He. 2022. Cross-modal prototype driven network for radiology report generation. In *Computer Vision – ECCV 2022*, pages 563–579, Cham. Springer Nature Switzerland.

Jun Wang, Abhir Bhalerao, Terry Yin, Simon See, and Yulan He. 2024. Camanet: Class activation map guided attention network for radiology report generation. *IEEE Journal of Biomedical and Health Informatics*, 28(4):2199–2210.

An Yan, Zexue He, Xing Lu, Jiang Du, Eric Chang, Amilcare Gentili, Julian McAuley, and Chun-Nan Hsu. 2021. Weakly supervised contrastive learning for chest X-ray report generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4009–4015, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Benjamin Yan, Ruochen Liu, David E. Kuo, Subathra Adithan, Eduardo Pontes Reis, Stephen Kwak, Vasantha Kumar Venugopal, Chloe O'Connell, Agustina Saenz, Pranav Rajpurkar, and Michael Moor. 2023. Style-aware radiology report generation with radgraph and few-shot prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14676–14688.

Xiulong Yi, You Fu, Jianzhi Yu, Ruiqing Liu, Hao Zhang, and Rong Hua. 2025. Lhr-rfl: Linear hybrid-reward-based reinforced focal learning for automatic radiology report generation. *IEEE Transactions on Medical Imaging*, 44(3):1494–1504.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. 2023. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9):100802.

Ke Zhang, Hanliang Jiang, Jian Zhang, Qingming Huang, Jianping Fan, Jun Yu, and Weidong Han. 2023a. Semi-supervised medical report generation via graph-guided hybrid feature consistency. *IEEE Transactions on Multimedia*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

Xiaodan Zhang, Yanzhao Shi, Junzhong Ji, Chengxin Zheng, and Liangqiong Qu. 2025. Mepnet: Medical entity-balanced prompting network for brain ct report generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25940–25948.

Xiaodan Zhang, Sisi Yang, Yanzhao Shi, Junzhong Ji, Ying Liu, Zheng Wang, and Huimin Xu. 2023b. Weakly guided attention model with hierarchical interaction for brain ct report generation. *Computers in Biology and Medicine*, 167:107650.

Haifeng Zhao, Yufei Zhang, Leilei Ma, Shuo Xu, and Dengdi Sun. 2025. Bridging vision and language: Optimal transport-driven radiology report generation via llms. *Preprint*, arXiv:2507.03908.

Chengxin Zheng, Junzhong Ji, Yanzhao Shi, Xiaodan Zhang, and Liangqiong Qu. 2024a. See detail say clear: Towards brain CT report generation via pathological clue-driven representation learning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16542–16552.

Chengxin Zheng, Junzhong Ji, Yanzhao Shi, Xiaodan Zhang, and Liangqiong Qu. 2024b. See detail say clear: Towards brain CT report generation via pathological clue-driven representation learning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16542–16552, Miami, Florida, USA. Association for Computational Linguistics.

Zijian Zhou, Miaojing Shi, Meng Wei, Oluwatosin Alabi, Zijie Yue, and Tom Vercauteren. 2024. Large model driven radiology report generation with clinical quality reinforcement learning. *CoRR*, abs/2403.06728.