ProFuseGPT: Progressive Fusion with Contrastive Refinement for Long-Sequence Medical Report Generation

Shaowei Shen[1], Jie Yang[2], Shanhao Zhan[1], Lianfen Huang[1,3,4(✉)], Zexin Huang[5], Zhibin Gao[6], and Xiaohong Yang[1]

[1] School of Informatics, Xiamen University, Xiamen, Fujian, China
[2] National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, Fujian, China
[3] School of Information Science and Technology, Xiamen University Tan Kah Kee College, Zhangzhou, Fujian, China
[4] Key Laboratory of Intelligent Manufacturing Equipment and Industrial Internet Technology, Fujian Provincial Universities, Xiamen, Fujian, China
[5] Department of Electronic Engineering, Tsinghua University, Beijing, China
[6] Navigation Institute, Jimei University, Xiamen, Fujian, China
lfhuang@xmu.edu.cn

Abstract. Automatic medical report generation (MRG) holds promise for alleviating radiologists' workload, which has spurred growing interest in MRG for stroke diagnosis. However, existing approaches often fail to effectively model its long-range spatial dependencies and suppress phase-level noise in multi-slice sequences, leading to diluted pathological signals and unstable cross-modal alignment. To address this, we propose a novel framework integrating a progressive fusion mechanism (PFM) and intermediate state refinement via contrastive alignment (ISRCA), inspired by radiologists' clinical workflow. PFM progressively refines pathological representations through anatomically deviation-aware adaptive weighting, suppressing noise from normal slices while enhancing salient abnormalities from local deviations to global context integration. ISRCA adopts a teacher-student distillation approach using contrastive learning to mitigate noise propagation and stabilize intermediate report features. Experimental results on two stroke imaging datasets demonstrate that our method outperforms existing approaches in natural language generation (NLG) metrics, highlighting the effectiveness of PFM and ISRCA in handling long-sequence medical images and advancing stroke imaging report generation.

Keywords: Report Generation, Contrastive Refinement, Cross-modal Alignment, Large Language Models.

1 Introduction

Automatic medical report generation (MRG) aims to leverage the growing volume of medical data to alleviate radiologists' workload and support clinical decision-making.

The majority of existing work has focused on generating radiology reports from chest X-rays (CXR)[1]-[5], typically using only one or a few images, with an emphasis on identifying localized lesions and achieving basic cross-modal alignment [6]-[8]. In recent years, however, brain report generation has garnered increasing clinical attention [3], [9], [10], spurring growing interest in MRG for stroke diagnosis. This task presents a fundamentally different challenge: unlike CXR, stroke assessment requires radiologists to examine numerous magnetic resonance imaging (MRI) slices across multiple anatomical planes and scanning phases to accurately evaluate lesion location, extent, and severity. Modeling such long-sequence, multi-slice imaging data imposes significantly higher demands on both feature extraction and cross-modal semantic alignment. Specially, the model must effectively aggregate critical pathological features across phases without overlooking subtle yet clinically significant abnormalities [11]; Furthermore, the abundance of image data risks introducing redundant or even distracting information, complicating the diagnostic reasoning process [8]. Consequently, achieving precise feature representation and robust cross-modal alignment in this complex, sequential imaging context remains a key challenge in the field [8], [11].

Current MRG approaches predominantly adopt an encoder–decoder paradigm, in which medical images are first encoded into visual features, which are then fed into a text decoder to automatically produce diagnostic reports. Early methods, such as HRNN and Up-Down, employed LSTM-based architectures to sequentially decode these visual representations. Subsequent advances, including WCL and WGAM, introduced weakly supervised learning to better localize pathological regions. More recently, Transformer-based frameworks, such as R2GenCMN and XProNet, have significantly improved cross-modal semantic alignment through memory-augmented mechanisms and hierarchical attention structures. To further enhance alignment between visual and textual modalities, researchers have proposed various strategies: integrating diagnostic labels or external knowledge bases on the image side to distinguish normal from abnormal regions, leveraging textual hierarchies to guide report generation, and designing granularity-aware modules or hybrid retrieval techniques (e.g., WGAM-HI [7]) to improve semantic consistency. While these methods achieve strong performance in single-phase or small-scale imaging scenarios, e.g., those involving one or a few CXR images, they struggle when applied to long-sequence medical imaging. In such settings, the sheer volume of input slices leads to phase-level information dilution, while redundant or noisy data accumulates across stages, degrading model performance.

Building upon advances in large language models, multimodal large language models (MLLMs) have revolutionized the field. LLaVA-Med [12] and HILT [13] have demonstrated the potential of LLMs (e.g., LLaMA3-8B) for medical vision–language tasks, while instruction-tuned frameworks such as PromptMR [14] have achieved state-of-the-art results by aligning visual features with LLM embeddings. Concurrently, models like mPLUG-Owl [11], Qwen2-VL [15], and InternVL2.5 [16] have extended these capabilities to multi-image understanding, and studies including R2Gengpt [3] have explored LLM-based chest X-ray report generation. Despite these advances, modeling and progressively aligning multi-phase image sequences within LLMs remains

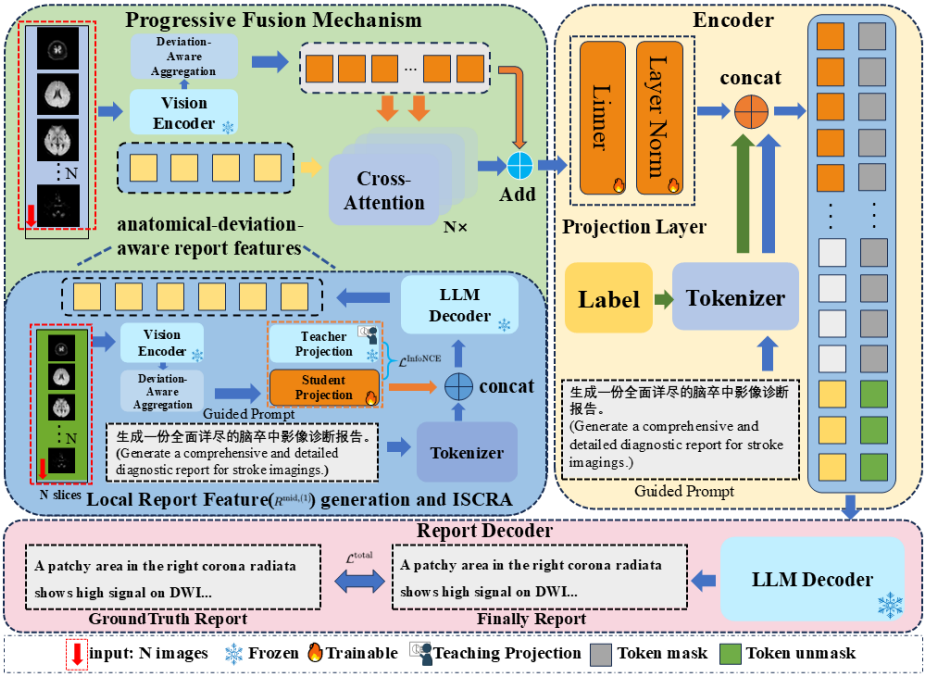


Fig. 1. ProFuseGPT employs ISRCA (blue) to generate $R_{\text{mid}}^{(1)}$ via anatomically deviation-aware adaptive weighting, integrates PFM (green) for adaptive weighting and cross-attention refinement of intermediate features, and generates diagnostic reports via encoder (yellow) and decoder (pink); iterative contrastive learning uses the optimal projection layer as teacher.

highly challenging [8], particularly in medical domains where sequential image coherence is critical for accurate diagnosis.

Motivated by the above challenges in stroke imaging report generation, we propose a novel framework integrating the progressive fusion mechanism (PFM) for multislice feature modeling and intermediate state refinement via contrastive alignment (ISRCA) to enhance cross-modal alignment. This framework progressively accumulates diagnostic information and mitigates noise propagation across phases, inspired by the radiological workflow. The architecture is illustrated in Fig. 1. Our key contributions are:

- We propose a two-stage framework for long-sequence stroke imaging report generation implementing radiologists' diagnostic workflow: global pathological context establishment followed by cross-modal refinement. This design models slice dependencies while mitigating noise dilution in lengthy image sequences.
- We introduce PFM with adaptive weighting based on anatomical deviation. By emphasizing slices with atypical patterns relative to learnable reference features, PFM enhances pathological saliency and suppresses redundant normal tissue, addressing the core challenge of noise propagation in multi-slice analysis.
- We propose ISRCA, which stabilizes intermediate representation learning through teacher-student contrastive learning at the first stage. This mechanism

prevents error accumulation during progressive refinement and ensures diagnostic relevance of intermediate features.

- Extensive experiments on StrokeQD and CTRG-Brain datasets demonstrate that our method outperforms state-of-the-art approaches, including Transformer-based models and vision-language models, with consistent gains in natural language generation (NLG) metrics and clinical accuracy, and we will open-source our code upon acceptance to support reproducibility and community research.

2 Methodology

2.1 Progressive Fusion Mechanism

Inspired by radiologists' diagnostic workflow, where initial global assessment of the full image sequence is followed by focused refinement of salient pathological details, we propose *progressive fusion mechanism (PFM)*, which implements this two-phase clinical reasoning through two functionally progressive stages ($T = 2$). These stages embody radiologists' progressive refinement principle: Stage 1 establishes a global pathological context from the complete sequence, while Stage 2 performs cross-modal refinement between visual features and intermediate textual representations to enhance diagnostic specificity. This design captures progressive information refinement through representation-level evolution, making it particularly suitable for long-sequence medical imaging where global context is diagnostically indispensable.

Given that non-lesion content dominates clinical imaging sequences, we introduce an adaptive weighting mechanism to suppress redundant normal tissue representations while enhancing diagnostically relevant patterns. A pre-trained Swin Transformer [17] encoder extracts features $\mathbf{Z}_i \in \mathbb{R}^d$ from each image I_i, where d denotes the feature dimensionality. We compute a reference feature $\mathbf{Z}^{\text{ref}} \in \mathbb{R}^d$ representing typical anatomical patterns, initialized as the centroid of visual features from 500 expert-labeled slices annotated as lesion-free. During training, $\mathbf{Z}^{\text{ref}}$ is maintained as a learnable parameter with L2 regularization ($\lambda^{\text{ref}} = 0.01$) to preserve anatomical consistency while adapting to dataset-specific characteristics. The adaptive feature aggregation at stage t is:

$$\overline{Z_t} = \sum_{i=1}^{N} \underbrace{\frac{\| \mathbf{Z}_i - \mathbf{Z}^{\text{ref}} \|_2}{\sum_{j=1}^{N} \| \mathbf{Z}_j - \mathbf{Z}^{\text{ref}} \|_2}}_{\text{adaptive weight } w_i^{(t)}} \cdot \mathbf{Z}_i, \tag{1}$$

where N is the total number of input images, and $w_i^{(t)}$ denotes the adaptive weight assigned to slice i at stage t, computed based on its deviation from the anatomical reference. This weighting mechanism effectively enhances feature diversity by amplifying contributions from slices exhibiting atypical patterns while suppressing redundant normal tissue representations, providing an efficient strategy for pathological saliency

enhancement in long-sequence medical image analysis. A trainable projection matrix $W^{\text{proj}} \in \mathbb{R}^{d_h \times d}$ maps the aggregated visual features into the language model's embedding space $\mathcal{H}(t)$, where d_h denotes the hidden dimension of the language model:

$$H(t) = \begin{cases} W^{\text{proj}} \cdot \overline{Z_1}, & t = 1, \\ W^{\text{proj}} \cdot \left(\underbrace{W^{\text{align}} Z_2^{\text{mult}}}_{\in \mathbb{R}^d} + \overline{Z_2} \right), & t = 2, \end{cases} \tag{2}$$

where $W^{\text{align}} \in \mathbb{R}^{d \times d}$ is a learnable projection matrix that aligns the cross-modal feature Z_2^{mult} to the same dimensional space as visual features for residual connection, and Z_2^{mult} represents the cross-modal feature obtained through multi-head cross-attention between current visual features and the intermediate report representation from the previous stage.

Stage 1: Global Pathological Assessment. At the initial stage ($t = 1$), the Swin Transformer encoder extracts features from all N slices. These features are adaptively aggregated using the weights defined in Eq. (1) to obtain $\overline{Z_1}$, which emphasizes regions with significant deviation from typical anatomy. The projected feature $H(1) = W^{\text{proj}} \cdot \overline{Z_1}$ is then concatenated with the instruction prompt X_p ("Generate a comprehensive stroke imaging diagnosis report") to form the initial input sequence: $W_p^{(1)} = \{Human : \langle \text{Img} \rangle H(1) \langle /\text{Img} \rangle, X_p\}$. After tokenization and embedding, this input is processed by the language model to produce the intermediate report representation $R^{\text{mid},(1)} = \text{LLM}^{\text{hidden}}(W_p^{(1)}; \theta_c)$, which captures the global pathological context.

Stage 2: Cross-Modal Feature Refinement. In the refinement stage ($t = 2$), the current aggregated visual feature $\overline{Z_2}$ (computed using identical weighting as Eq. (1)) is fused with the intermediate report representation $R^{\text{mid},(1)}$ generated from stage 1. This fusion is performed through a multi-head cross-attention mechanism (MHCA) that enables deep integration between visual and textual modalities. For each attention head i ($1 \le i \le h$, where h denotes the number of attention heads), the query, key, and value matrices are computed as:

$$Q_i = W_i^Q \overline{Z_2}, K_i = W_i^K R^{\text{mid},(1)}, V_i = W_i^V R^{\text{mid},(1)}, \tag{3}$$

where $W_i^Q \in \mathbb{R}^{d_k \times d}$ projects visual features to query space, and $W_i^K, W_i^V \in \mathbb{R}^{d_k \times d_h}$ project textual features to key and value spaces, with d_k denoting the dimensionality

of each attention head and output restored to $\mathbb{R}^d$. The cross-modal feature for each head is then computed as:

$$\begin{cases} Z_i = \text{Softmax}\left(\dfrac{Q_i K_i^\top}{\sqrt{d_k}}\right) V_i, \\ Z_2^{\text{mult}} = \text{Concat}(Z_1, Z_2, \ldots, Z_h) W^O, \end{cases} \tag{4}$$

where Concat(·) denotes the concatenation operation across all attention heads, and $W^O \in \mathbb{R}^{hd_k \times d}$ is a learnable output projection matrix that combines the multi-head outputs into a unified representation. The refined feature $Z_2^{\text{mult}} + \overline{Z_2}$ is then projected into the language model's embedding space via W^{proj} to obtain $H(2)$, which serves as the enhanced visual input for final report generation.

Report Generation. At the final stage ($t = 2$), the complete input sequence is constructed as $W_p^{(2)} = \{\text{Human: } \langle \text{Img} \rangle H(2) \langle /\text{Img} \rangle, X_p, X_r\}$, where X_r denotes the ground-truth diagnostic report. During training, the language modeling loss is computed using masked language modeling, where only tokens corresponding to the report portion are considered for optimization. The negative log-likelihood loss is formulated as:

$$\mathcal{L}^{\text{nll}}(\theta; W_p^{(2)}) = -\sum_{i=1}^{n} \log p(x_i \mid W_p^{(2)}, X_{<i}), \tag{5}$$

where θ represents all trainable parameters, x_i denotes the i-th token in the ground-truth report, and $X_{<i}$ represents all preceding tokens in the report sequence.

2.2 Contrastive Optimization for Intermediate State Refinement

To address the challenge that intermediate features in PFM receive supervision only through the final report loss, potentially leading to error propagation and suboptimal intermediate representations, we introduce *intermediate state refinement via contrastive alignment (ISRCA)*. Motivated by our observation that progressive fusion exhibits low decoding confidence and suboptimal NLG metrics during early training epochs, we employ the projection layer from an initially trained single-stage variant of our framework that demonstrates robust pathological encoding as a frozen teacher to guide the student model through divergence minimization. Specifically, we select the projection layer output at stage $t = 1$ (i.e., $H^{\text{teac}}(1)$) as the reference for intermediate feature alignment.

During training, the student model aligns its intermediate projection $H^{\text{stud}}(1)$ with the teacher's counterpart $H^{\text{teac}}(1)$ using contrastive learning. Formally, the student's

feature $q = H^{\text{stud}}(1)$ serves as the query, the teacher's feature $k^+ = H^{\text{teac}}(1)$ serves as the positive sample, and other student features within the same mini-batch serve as negative samples $\{k^-\}$. The InfoNCE loss [22] measures the normalized similarity between query and samples:

$$\mathcal{L}^{\text{InfoNCE}}(q,k^+) = -\log \frac{\exp(sim(q,k^+)/\tau)}{\exp(sim(q,k^+)/\tau) + \sum_{k^-} \exp(sim(q,k^-)/\tau)}, \tag{6}$$

where $sim(\cdot)$ denotes cosine similarity and $\tau = 0.07$ is a temperature hyperparameter. The total training objective integrates three components, the final report generation loss $\mathcal{L}^{\text{nll}}$, the contrastive refinement loss $\mathcal{L}^{\text{InfoNCE}}$, and the reference feature regularization loss $\mathcal{L}^{\text{ref}} = \lambda^{\text{reg}} \| \mathbf{Z}^{\text{ref}} - \mathbf{Z}^{\text{ref},(0)} \|_2^2$, into a unified objective:

$$\mathcal{L}^{\text{total}} = \mathcal{L}^{\text{nll}} + \lambda \mathcal{L}^{\text{InfoNCE}} + \beta \mathcal{L}^{\text{ref}}, \quad \lambda = 0.5, \beta = 0.01. \tag{7}$$

This multi-objective optimization strategy ensures that intermediate representations maintain diagnostic relevance throughout training, thereby enhancing the model's ability to generate accurate and clinically relevant reports.

3 EXPERIMENTS

3.1 Experimental Setup

Dataset. In this study, we evaluate our model on two medical imaging datasets: the publicly available StrokeQD dataset [23] for ischemic stroke imaging report generation, and the newly introduced CTRG-Brain dataset [24] for brain CT report generation. The StrokeQD dataset contains 22,626 MRI-DWI images and corresponding diagnostic reports from 1,181 anonymized ischemic stroke patients, curated by a medical consortium. To ensure consistent report length, we analyzed the text distribution and followed a standard 7:2:1 split for training, validation, and testing [25]. The CTRG-Brain dataset comprises 6,000 Chinese medical reports and a total of 160,336 brain CT images. We adopt the mainstream data partitioning strategy [7], splitting the dataset into training, validation, and test sets at a ratio of 7:1:2. Following prior practice [8], we select 24 representative slices from each scan for structured input.

Implementation Details. For feature extraction, we use a pretrained Swin Transformer[1] with a hidden dimension of 1,024, and for text generation, we employ the Llama3-Chinese-8B model[2] [26], optimized for Chinese medical text. A learnable

[1] https://huggingface.co/microsoft/swin-base-patch4-window7-224
[2] https://huggingface.co/FlagAlpha/Llama3-Chinese-8B-Instruct

projection layer aligns visual features with the LLM embedding space. Training follows two independent rounds: *(i) PFM pretraining*, 75 epochs optimizing projection/fusion layers with frozen vision encoder and LLM; the checkpoint exhibiting *optimal diagnostic token confidence and balanced NLG metrics* on validation set is frozen as teacher; *(ii) ISRCA refinement*, a randomly initialized student trains for 75 epochs under frozen teacher guidance via InfoNCE alignment of intermediate projections $H(\mathrm{l})$, alongside cross-entropy loss. Both stages employ AdamW ($lr = 1\times10^{-4}$) with cosine annealing. $\mathbf{Z}^{\mathrm{ref}}$ is updated at 5×10^{-5} with L2 regularization ($\lambda^{\mathrm{reg}} = 0.01$). During inference, beam search (width=3) generates reports with two progressive fusion stages.

Table 1. Performance comparison of proposed ProFuseGPT with other SOTA models.

Dataset	Methods	B1	B2	B3	B4	RG	M	C	F1
StrokeQD	R2GEN†[1]	48.2	36.6	29.0	23.1	44.3	28.4	42.7	–
	R2GENCMN†[2]	46.2	33.6	25.8	20.0	41.6	27.4	46.7	–
	R2GENGPT†[3]	<u>51.1</u>	<u>39.6</u>	<u>32.0</u>	<u>25.9</u>	<u>47.1</u>	<u>29.6</u>	44.7	–
	TSGET†[4]	47.3	35.4	27.8	22.3	43.9	28.4	52.5	–
	Qwen2-VL-7b†[15]	45.6	35.8	29.4	24.1	47.0	28.3	50.0	–
	InternVL2.5-8b†[16]	49.8	38.7	30.9	24.7	46.7	29.5	<u>53.6</u>	–
	OURS(ProFuseGPT)	**52.0**	**40.7**	**33.1**	**26.9**	**47.8**	**30.2**	**59.4**	**–**
CTRG-Brain	R2GEN†[1]	48.9	40.8	35.6	31.6	53.3	30.4	77.7	74.7
	R2GENCMN[2]	49.1	40.0	34.4	30.1	48.6	29.9	84.2	69.8
	R2GENGPT†[3]	51.2	<u>43.1</u>	<u>37.8</u>	<u>33.5</u>	<u>53.8</u>	31.7	<u>85.2</u>	73.8
	TSGET†[4]	51.4	42.6	36.7	32.0	52.4	31.3	79.9	<u>75.9</u>
	Qwen2-VL-7b†[15]	50.1	41.5	35.7	31.2	50.7	30.3	66.5	69.8
	InternVL2.5-8b†[16]	50.5	40.9	34.9	30.2	48.4	30.2	46.6	74.7
	HRNN[18]	42.3	28.3	21.2	17.1	39.3	27.2	20.9	70.2
	Up-Down[19]	45.8	34.8	28.5	24.4	42.5	31.6	27.3	70.2
	WCL[20]	49.5	36.5	29.4	25.1	42.8	31.3	33.3	64.5
	XProNet[21]	50.6	41.3	34.4	29.1	51.7	31.5	83.3	70.1
	WGAM[6]	49.4	36.7	29.6	25.4	42.4	**32.0**	31.9	68.9
	WGAM-HI[7]	50.4	39.3	30.5	26.1	43.8	31.4	33.2	67.4
	LLaVA-med[12]	50.0	39.3	32.0	26.3	46.7	31.4	38.6	64.3
	HILT[13]	50.7	39.9	33.1	27.9	46.1	30.8	43.7	68.4
	PromptMRG[14]	48.1	38.3	31.6	26.5	47.4	31.0	50.3	69.2
	SDSC[8]	<u>51.5</u>	42.0	35.7	30.9	49.0	31.4	80.0	70.6
	OURS(ProFuseGPT)	**52.3**	**44.0**	**38.5**	**34.1**	**53.9**	**32.0**	**87.2**	**76.5**

† denotes re-implementation results.

Evaluation Metrics. We evaluate the generated reports using various NLG metrics, computed by the *pycocoevalcap* tool [3]. These metrics include BLEU(B)[27], METEOR(M) [28], ROUGE(RG) [29], and CIDEr(C) [30]. To measure pathological accuracy, we compute a Clinical Evaluation (CE) F1 score based on 24 expert-

[3] https://github.com/tylin/coco-caption

identified keywords [7], which is only applied to the CTRG-Brain dataset due to the absence of structured pathology terms in StrokeQD.

3.2 Comparison with State-of-the-Art Methods

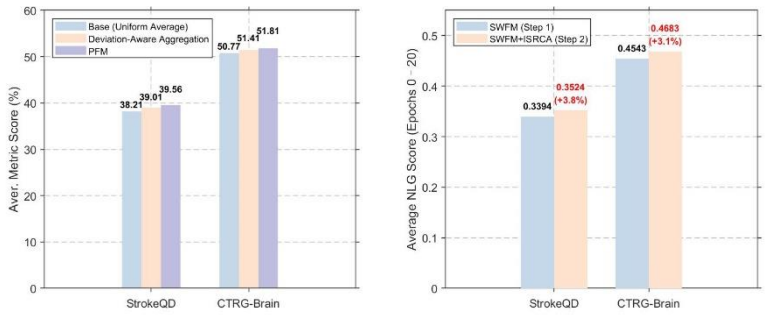
We compare ProFuseGPT with methods on StrokeQD and CTRG-Brain, including CNN-LSTM-based (HRNN [18], Up-Down [19]), Transformer-based (R2GEN [1], R2GENCMN [2], XProNet [21]), weakly supervised (WCL [19], WGAM [5], and vision-language models (LLaVA-med [12], HILT [13], InternVL2.5 [16], SDSC [8]). As shown in Table 1, most existing methods struggle to model inter-slice dependencies. ProFuseGPT achieves the best performance on both datasets. On StrokeQD, it outperforms all baselines with BLEU-4 of 0.269, ROUGE of 0.478, and CIDEr of 0.594. On CTRG-Brain, ProFuseGPT achieves BLEU-4 of 0.341, ROUGE of 0.539, and F1 of 0.765, demonstrating its superiority in medical report generation.

Table 2. Ablation studies: **Base** (Uniform Average), **+PFM** (progressive fusion), **+ISRCA** (full model).

Model	**StrokeQD**						
	B1	**B2**	**B3**	**B4**	**RG**	**M**	**C**
Base	49.5	38.4	31.3	25.6	46.8	29.1	46.8
+PFM	51.4	40.0	32.2	26.1	46.8	29.9	50.5
+ISRCA	**52.0**	**40.7**	**33.1**	**26.9**	**47.8**	**30.2**	**59.4**
Model	**CTRG-Brain**						
	B1	**B2**	**B3**	**B4**	**RG**	**M**	**C/F1**
Base	50.4	42.4	37.2	33.1	53.4	31.2	84.1/74.4
+PFM	51.7	43.5	38.1	33.8	53.7	31.2	87.2/75.3
+ISRCA	**52.3**	**44.0**	**38.5**	**34.1**	**53.9**	**32.0**	**87.2/76.5**

3.3 Ablation Study

Ablation Study I: Component Contribution Analysis. We conduct comprehensive ablation studies on StrokeQD and CTRG-Brain datasets to evaluate the contribution of each proposed component. As shown in Table 2, we define three configurations: *Base (Uniform Average of all slice features)*, *+PFM*, and *+ISRCA.* Results demonstrate consistent performance improvement from adding each component. On StrokeQD, PFM improves average NLG metrics by +3.5%, while the full ISRCA model achieves a substantial gain of +8.5%. On CTRG-Brain, PFM elevates average NLG metrics by +2.2% and improves clinical F1 by +1.2%, whereas ISRCA further boosts NLG metrics by +3.1% and clinical F1 by +2.8%. To further analyze the PFM mechanism, we compare



(a) (b)

Fig. 2. (a) Component ablation: Base (Uniform Average), *Deviation-Aware*, and PFM. PFM achieves optimal performance, confirming complementary benefits of anatomical deviation weighting and cross-modal refinement. (b) Early-training convergence (epochs 0–20): ISRCA consistently improves Average scores on both StrokeQD and CTRG-Brain.

three feature aggregation variants in Fig. 2 (a): Uniform Average, Deviation-Aware (adaptive weighting only), and full PFM. The results show that PFM achieves optimal performance, confirming the complementary benefits of anatomical deviation weighting and cross-modal refinement. These findings collectively demonstrate that PFM and ISRCA are complementary mechanisms essential for long-sequence medical image understanding.

Ablation Study II: Component-wise Mechanism Analysis. We conduct a component-wise analysis comparing two configurations: *(i) PFM alone (without curriculum guidance)*, and *(ii) the full model integrating PFM with the ISRCA instance-sequencing mechanism*. The analysis focuses on the critical early-training phase (epochs 0-20), where optimization trajectories are established. As shown in Fig. 2(b), ISRCA yields consistent gains over the PFM baseline within this period. Specifically, the average NLG score increases by +3.8% (from 0.339 to 0.352) on StrokeQD, while average score improves by +3.1% (from 0.454 to 0.468) on CTRG-Brain. This demonstrates that ISRCA's anatomy-guided instance sequencing stabilizes representation learning during early optimization. Performance advantages established in this early phase propagate to final convergence, directly contributing to the improved report generation quality observed in Table 2.

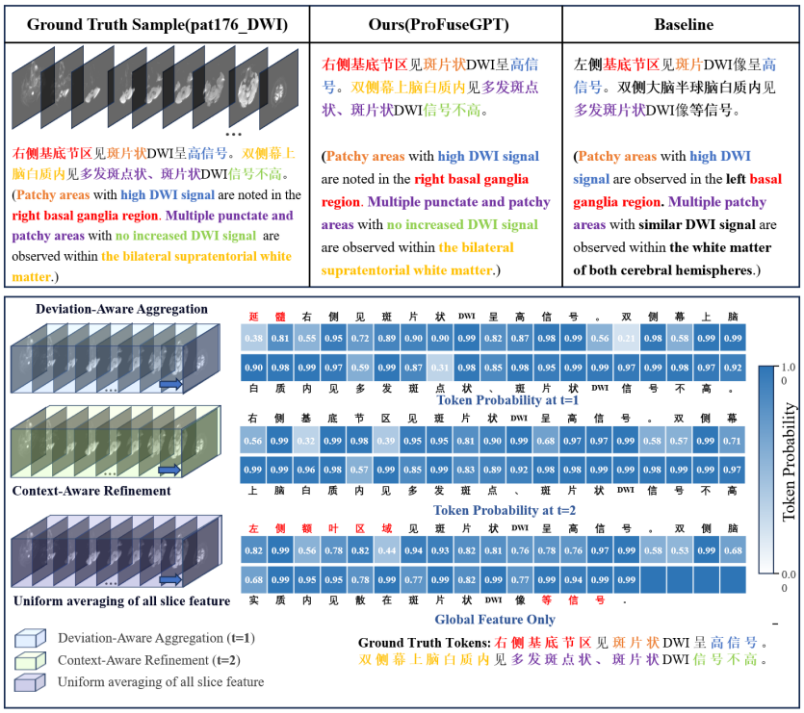


Fig. 3. Token probability evolution and report generation for anatomical regions using different features.

Ablation Study III: Token-level Confidence Analysis. Token-level analysis demonstrates PFM's stage-wise anatomical mislocalization correction. Fig. 3 shows the Uniform Average of all slice feature baseline mislocalizes lesions to the left frontal region (left: 0.82, frontal: 0.56, lobe: 0.78) and generic parenchyma (cerebrum: 0.68, parenchyma: 0.68). Stage 1 (Deviation-Aware Aggregation) detects right-sided abnormalities but erroneously attributes pathology to the medulla (tokens: 0.38/0.81), while identifying supratentorial white matter involvement (bilateral: 0.21, tentorial: 0.58, supratentorial: 0.99, white matter: 0.90/0.98). Stage 2 (Context-Aware Refinement) corrects localization to the clinically accurate right basal ganglia region, with its four-token sequence scoring 0.32, 0.99, 0.98, and 0.39; the middle tokens representing core anatomy achieve high confidence (0.99/0.98). Critically, medulla tokens vanish at Stage 2. Pathological descriptors progressively refine: "bilateral" confidence rises from 0.21 to 0.57, and the unstable pattern token in "patchy-dot" lesions (0.31 at Stage 1) stabilizes to 0.92/0.99 at Stage 2. DWI hyperintensity maintains high confidence across stages (DWI:0.99), contrasting with the baseline's inconsistent detection (0.76 for primary lesion). This evolution validates PFM's two-stage clinical reasoning, initial deviation detection followed by cross-modal refinement that eliminates anatomical hallucinations while preserving pathological specificity.

4 Conclusion

We propose ProFuseGPT, a framework integrating PFM with ISRCA for long-sequence stroke report generation. PFM implements radiologist-inspired two-stage reasoning through anatomically deviation-aware weighting, while ISRCA stabilizes intermediate representations via teacher-student contrastive learning. Experiments on StrokeQD and CTRG-Brain demonstrate consistent superiority over state-of-the-art methods, achieving BLEU-4 of 26.9 and CIDEr of 59.4 on StrokeQD, and BLEU-4 of 34.1 with clinical F1 of 76.5 on CTRG-Brain (all metrics $\times$ 100 scaled). Ablation studies confirm complementary component benefits and ISRCA's early-training convergence acceleration (+3.8%/+3.1% gains in epochs 0-20). The framework effectively balances global pathological context modeling with cross-modal refinement. Future work will extend to multi-modal integration and clinician evaluation for clinical deployment.

References

1. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1439–1449 (2020)
2. Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 5904–5914 (2021)
3. Wang, Z., Liu, L., Wang, L., Zhou, L.: R2gengpt: Radiology report generation with frozen llms. Meta-Radiology 1(3), 100033 (2023)
4. Yi, X., Fu, Y., Liu, R., Zhang, H., Hua, R.: Tsget: two-stage global enhanced transformer for automatic radiology report generation. IEEE Journal of Biomedical and Health Informatics 28(4), 2152–2162 (2024)
5. Wang, Z., Tang, M., Wang, L., Li, X., Zhou, L.: A medical semantic-assisted transformer for radiographic report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 655–664. Springer (2022)
6. Yang, S., Ji, J., Zhang, X., Liu, Y., Wang, Z.: Weakly guided hierarchical encoder-decoder network for brain ct report generation. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 568–573. IEEE (2021)
7. Zhang, X., Yang, S., Shi, Y., Ji, J., Liu, Y., Wang, Z., Xu, H.: Weakly guided attention model with hierarchical interaction for brain ct report generation. Computers in Biology and Medicine 167, 107650 (2023)
8. Zheng, C., Ji, J., Shi, Y., Zhang, X., Qu, L.: 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, pp. 16542–16552 (2024)
9. You, D., Liu, F., Ge, S., Xie, X., Zhang, J., Wu, X.: Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In: Interna-tional Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 72–82. Springer (2021)

10. Yang, J., Shen, S., Chen, N., Huang, L., Gao, Z., Chao, H.C.: Towards factual con-sistency in clinical summarization: A self-correction strategy. Human-Centric Compu-ting and Information Sciences 15 (2025)
11. Ye, J., Xu, H., Liu, H., Hu, A., Yan, M., Qian, Q., Zhang, J., Huang, F., Zhou, J.: mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. arXiv preprint arXiv:2408.04840 (2024)
12. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems 36, 28541–28564 (2023)
13. Liu, C., Wan, Z., Wang, Y., Shen, H., Wang, H., Zheng, K., Zhang, M., Arcucci, R.: Benchmarking and boosting radiology report generation for 3d high-resolution medical images. arXiv preprint arXiv:2406.07146 1(2) (2024)
14. Jin, H., Che, H., Lin, Y., Chen, H.: Promptmrg: Diagnosis-driven prompts for medical report generation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 2607–2615 (2024)
15. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
16. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
17. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
18. Krause, J., Johnson, J., Krishna, R., Fei-Fei, L.: A hierarchical approach for generating descriptive image paragraphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 317–325 (2017)
19. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6077–6086 (2018)
20. Yan, A., He, Z., Lu, X., Du, J., Chang, E., Gentili, A., McAuley, J., Hsu, C.N.: Weakly supervised contrastive learning for chest x-ray report generation. In: Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 4009–4015 (2021)
21. Wang, J., Bhalerao, A., He, Y.: Cross-modal prototype driven network for radiology report generation. In: European Conference on Computer Vision, pp. 563–579. Springer (2022)
22. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
23. Zhang, S., Xu, S., Tan, L., Wang, H., Meng, J.: Stroke lesion detection and analysis in mri images based on deep learning. Journal of Healthcare Engineering 2021(1), 5524769 (2021)
24. Tang, Y., Yang, H., Zhang, L., Yuan, Y.: Work like a doctor: Unifying scan localizer and dynamic generator for automated computed tomography report generation. Expert Systems with Applications 237, 121442 (2024)
25. Zhang, S., Tan, L., Han, Q., Wang, H., Meng, J.: Automatic report generation on a large-scale stroke mri dataset. In: 2023 IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT), pp. 123–128. IEEE (2023)
26. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)

27. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002)
28. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72 (2005)
29. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81 (2004)
30. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4566–4575 (2015)