DiffVP: Differential Visual Semantic Prompting for LLM-Based CT Report Generation

Yuhe Tian[1,2,3], Kun Zhang[1,2,3*], Haoran Ma[1,2,3], Rui Yan[1,2,3], Yingtai Li[1,2,3], Rongsheng Wang[1,2,3], and Shaohua Kevin Zhou[1,2,3,4,5*]

[1] University of Science and Technology of China (USTC), Hefei, China
[2] Medical Imaging, Robotics, Analytic Computing & Learning (MIRACLE) Lab, Suzhou Institute for Advanced Research, USTC, Suzhou, China
[3] Jiangsu Provincial Key Laboratory of Multimodal Digital Twin Technology, Suzhou, China
[4] State Key Laboratory of Precision and Intelligent Chemistry, USTC, Hefei, China
[5] Biomedical Basic Research Center of Jiangsu Province, Suzhou, Jiangsu, China
[kkzhang@ustc.edu.cn], [skevinzhou@ustc.edu.cn] *Corresponding authors.

Abstract. While large language models (LLMs) have advanced chest CT report generation, existing methods typically encode 3D volumes holistically, failing to distinguish informative cues from redundant anatomical background. Inspired by radiological cognitive subtraction, we propose Differential Visual Prompting (DiffVP), which conditions report generation on explicit, high-level semantic scan-to-reference differences rather than solely on absolute visual features. DiffVP employs a hierarchical difference extractor to capture complementary global and local semantic discrepancies into a shared latent space, along with a difference-to-prompt generator that transforms these signals into learnable visual prefix tokens for LLM conditioning. These difference prompts serve as structured conditioning signals that implicitly suppress invariant anatomy while amplifying diagnostically relevant visual evidence, thereby facilitating accurate report generation without explicit lesion localization. On two large-scale benchmarks, DiffVP consistently outperforms prior methods, improving the average BLEU-1-4 by +10.98 and +4.36, respectively, and further boosts clinical efficacy on RadGenome-ChestCT (F1 score 0.421). Code will be available at `https://github.com/ArielTYH/DiffVP/`.

Keywords: 3D CT Report Generation · Vision-Language Models · Semantic Discrepancy Modeling

1 Introduction

Computed tomography (CT) is one of the most widely used imaging modalities in clinical practice, which plays a critical role in disease screening, diagnosis, and treatment planning across a broad range of conditions. Despite its clinical value, interpreting chest CT scans remains highly demanding for radiologists, requiring careful slice-by-slice review to produce reliable reports. Compared with natural

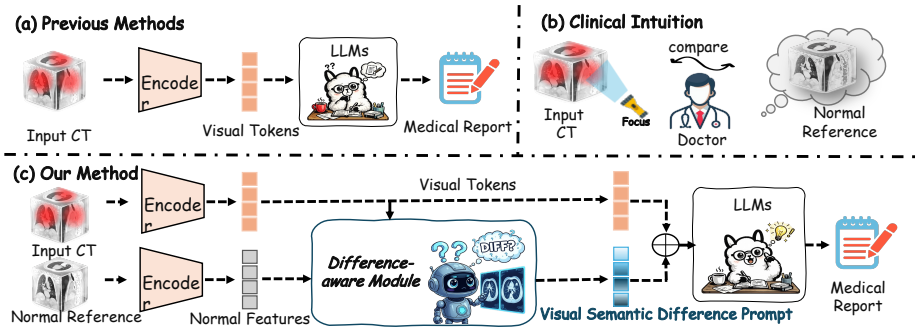


Fig. 1: Comparison of conventional methods and our proposed framework. (a) Previous methods feed whole-volume visual tokens uniformly into an LLM for report generation. (b) In clinical practice, radiologists often compare the current scan image with a normal reference image to identify differences. (c) Inspired by this, our method leverages a normal reference prior to derive deviation-aware visual prompts, guiding the LLM to generate more accurate and fine-grained reports.

scenarios [29,30], this challenge is particularly pronounced due to the inherently sparse nature of 3D CT volumes, where abnormal areas constitute only a small portion of the entire volume. Consequently, automated report generation, as an effective solution to assist clinicians, has attracted widespread research attention.

Early studies on CT report generation predominantly adopt an encoder-decoder paradigm [3,8,25,26], in which a visual encoder extracts representations from CT slices or volumes, followed by a language decoder that generates reports conditioned on these visual features. More recently, driven by the remarkable reasoning and language generation capabilities of large language models (LLMs), the field has shifted toward LLM-based medical report generation. R2GenGPT first introduced LLMs to this task [24], and Reg2RG further improves it by incorporating segmented local regions to provide finer-grained visual context [4]. As depicted in Fig. 1(a), existing LLM-based CT report generation typically encodes a CT volume into a sequence of visual tokens and feeds them into an LLM to generate the report. This paradigm typically passes all tokens to the LLM uniformly, implicitly assuming equal informativeness, and relies on the LLM to autonomously infer clinically salient cues from raw visual representations.

However, this uniform token-feeding strategy is fundamentally misaligned with the actual cognitive process of radiologists when composing CT reports: in clinical practice, radiologists rely on strong priors of normal anatomy and summarize imaging findings by contrasting the current scan with an expected internal reference [11]. In fact, this comparison is not performed at the pixel level; due to inherent inter-subject anatomical variations and inevitable spatial registration errors, raw scans lack direct comparability at the pixel scale, as

shown in Fig. 6(b). Furthermore, this comparative process does not require explicit contour delineation or precise localization of abnormal regions; instead, it serves as an abstract, high-level semantic cue designed to highlight deviations worthy of clinical attention.

In contrast, existing LLM-based paradigms provide no such visual reference signal, leading to two key issues. First, 3D CT volumes are dominated by redundant background anatomy (*e.g.* bones and muscles). Without reference-guided cues, the LLM must process these undifferentiated tokens equally, which can dilute clinically important information and hinder effective report generation. Second, abnormalities in 3D CT are often sparse. Radiologists identify them by leveraging normal anatomical priors, whereas LLMs conditioned on a single scan lack this contrastive reference, making it difficult to emphasize diagnostically relevant findings while avoiding redundancy- and sparsity-induced noise. As a result, the generated reports tend to be overly generic and often miss fine-grained descriptive details that are critical for clinical interpretability.

Drawing inspiration from this semantic comparative reasoning mechanism, we propose **Diff**erential **V**isual **P**rompting for chest CT Report Generation (**DiffVP**). Instead of explicitly localizing or segmenting lesions, DiffVP leverages normal CT scans as a reference prior and extracts semantic deviation-aware cues to guide attention toward clinically meaningful differences. By contrasting the input scan with a normal reference, DiffVP constructs a visual difference prompt that reweights the conditioning semantics, implicitly suppressing invariant anatomy while amplifying distinctive variations, promoting more fine-grained and informative report generation. Specifically, DiffVP comprises two modules. First, the Hierarchical Difference Extractor (HDE) projects both target and reference normal inputs into a unified, fixed-budget token space via a shared extractor. Rather than relying on alignment-sensitive pixel-level subtraction, it computes high-level semantic difference representations using hierarchically complementary global and local difference operators. This step effectively disentangles potentially diseased signals from the anatomical surroundings while resisting registration-induced noise. Second, the Difference-to-Prompt Generator (DPG) transforms these difference representations into learnable prefix [13] embeddings. These embeddings are then integrated into the LLM input sequence, serving as continuous cues that guide the model to generate diagnostic-sensitive reports.

Our main contributions are summarized as follows:

- We propose the Differential Visual Prompting (DiffVP) framework, a novel paradigm that introduces normal CT scans as visual reference priors for CT report generation. Unlike conventional methods that rely solely on absolute visual representations or are sensitive to low-level misalignment when forming comparisons, DiffVP models differences in a high-level semantic space. This enables LLMs to better focus on clinically important information without explicit abnormal region localization.
- We design a hierarchical difference extractor to capture complementary global and local semantic discrepancies, together with a difference-to-prompt generator that projects these signals into the LLM's input space. This design

effectively alleviates redundancy and sparsity issues inherent in 3D CT volumes while preserving diagnostic sensitivity and improving robustness to inter-subject anatomical variation.
– We conduct extensive experiments on two large-scale 3D chest CT datasets, Rad Genome-ChestCT and CTRG-Chest-548K. Results demonstrate that DiffVP consistently outperforms state-of-the-art methods across major metrics, producing reports with improved clinical accuracy, diagnostic relevance, and fine-grained clinical descriptions.

2 Related Work

2.1 3D CT Report Generation

Early studies on 3D CT report generation, such as CT2REP [8] and CT-ViT [17], largely follow encoder–decoder architectures that compress volumetric data into global representations. Such compression inevitably discards subtle pathological cues, limiting the clinical fidelity of the generated reports. There are also growing concerns about generational fairness [32–35] and model explainability [22, 23].

With the strong generation and reasoning abilities of large language models (LLMs), recent work increasingly adapts LLMs for medical report generation. Generalist foundation models, including RadFM [27], M3D [1], and Argus [14], use lightweight adapters to connect 3D visual encoders with pre-trained LLMs. To improve diagnostic precision, several methods further inject fine-grained guidance: Reg2RG [4] leverages anatomical region masks to extract region-aware features and align structures with text, while Dia-LLaMA [5] introduces diagnostic prompts to steer decoding. Other lines explore more elaborate pipelines. Recently, multi-agent clinical diagnosis has attracted increasing attention [12], with representative examples including the agent-based CT-Agent [16] and the retrieval-augmented MvKeTR [6]. Despite these advances, most methods still operate in a single-scan setting, extracting features only from the target volume. Lacking a normal reference for contrast, they struggle to suppress anatomical redundancy and background noise, which hinders isolating pathological changes from normal anatomy.

2.2 Difference-Aware Report Generation

Prior work has explored difference-aware strategies with distinct calculation and utilization paradigms. In longitudinal settings, Yun *et al.* [28] explicitly extracts isolated and localized disease-wise patch differences between current and prior scans to capture disease progression, while Song *et al.* [19] computes dynamic temporal residuals, treating them as implicit features to emphasize evolving regions. Several methods have also addressed abnormal-normal discrepancies in non-temporal settings. For instance, Park *et al.* [18] and Lyu *et al.* [15] perform simple, direct feature-vector subtraction between abnormal and normal 2D images. The subtracted difference vectors are then fused with tag information

and fed into traditional decoders for text generation. Alternatively, rather than employing paired reference images, Zhao *et al.* [36] constructs a latent decoupling memory matrix to learn and separate normal from pathological patterns across the entire dataset. More recently, Bian *et al.* [2] introduce DiffRGenNet, which retrieves similar reports and explicitly models image-report discrepancies through feature-difference estimation and flexible global-local aggregation for fine-grained difference-aware 2D X-ray report generation.

However, applying these mechanisms to LLM-based 3D report generation reveals numerous limitations. At the calculation level, extracting isolated patches or performing naive subtraction destroys cross-slice continuity, subsequently leading to the omission of minute lesions and introducing spatial noise. Furthermore, these differences are typically relegated to mere implicit auxiliary features. To effectively handle the inherent lesion sparsity and massive background redundancy in 3D CT volumes, our method hierarchically models discrepancies directly over the full 3D voxel-token sequence. This multi-scale approach preserves volumetric continuity and enables the fine-grained detection of structural deviations, effectively disentangling critical anatomy from the background. After extracting the hierarchical visual semantic differences, we innovatively and explicitly project them into learnable visual prefix tokens to serve as part of the LLM input prefix. This prefix-based injection provides a structured and controllable autoregressive signal, guiding the LLM to focus more attention on what differs between scans.

3 Methodology

3.1 Overview

Existing LLM-based CT report generation typically follows a unified paradigm: the input CT volume X is first transformed into a sequence of visual tokens $I \in \{v_i\}_{i=1}^{k}$ (k denotes the number of visual tokens), which are then directly fed into the LLM to generate the diagnostic report $\hat{T}$:

$$T \sim \mathrm{LLM}(\hat{T} \mid I),\ I = \Phi(X), \tag{1}$$

where $\Phi(\cdot)$ denotes the visual encoder. While effective to some extent, this paradigm fundamentally treats all visual tokens as equally informative and relies on the LLM to implicitly discover clinically relevant cues from the input CT.

In contrast, our Differential Visual Prompting CT report generation (**DiffVP**) explicitly models the discrepancies between the target input X and a normative reference X_{ref} in a shared latent space. The extracted difference features are then transformed into learnable prefix embeddings and prepended to the LLM input sequence as:

$$\begin{aligned} &T \sim \mathrm{LLM}\big(\hat{T} \mid [\Delta_{\mathrm{prompt}};\, I]\big), \\ &\quad \text{s.t.} \quad \Delta_{\mathrm{prompt}} = \mathrm{Diff}\big(\Phi(X),\, \Phi(X_{\mathrm{ref}})\big), \end{aligned} \tag{2}$$

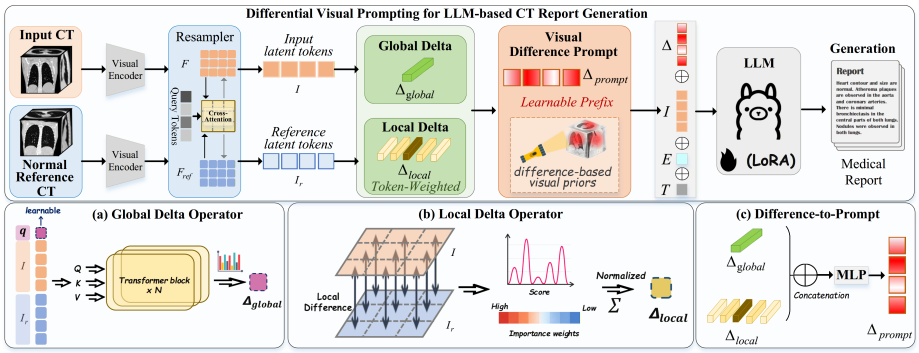


Fig. 2: Overall architecture. The framework takes a target CT volume and a normal reference CT as inputs. A shared visual encoder and resampler produce aligned latent tokens (I and I^r). A difference-aware module then derives (a) a global delta Δ_{global} by applying a Transformer with a learnable query over I and I^r, and (b) a local delta Δ_{local} by aggregating token-wise residuals with distance-based importance weights. These two signals are fused by (c) a difference-to-prompt generator to form a visual difference prompt, which is prepended as a soft prefix to an LLM (LoRA-tuned) to guide medical report generation.

where Δ_{prompt} denotes the learnable prompt tokens derived from the feature disparity between $\Phi(X)$ and $\Phi(X_{ref})$, functioning as discriminative semantic cues. $\text{Diff}(\cdot)$ is a difference module that captures visual changes between the target and reference 3D CT volumes. By injecting explicit differential cues, DiffVP effectively mitigates interference from the redundant anatomical background, steering the LLM to prioritize the description of potentially important areas to improve report generation quality. In the following, we introduce the visual encoder (Sec. 3.2), the hierarchical difference extractor (Sec. 3.3 and Sec. 3.3), and the difference-to-prompt generator (Sec. 3.4) with LLM input construction.

3.2 Visual Encoder

To extract dense anatomical representations from high-dimensional CT volumes, we employ a ResNet-based visual encoder as the backbone network. Given a batch of input volumes $X \in \mathbb{R}^{b\times s\times h\times w}$, the encoder processes the volumetric data to capture voxel-level visual patterns. The resulting feature maps are flattened along the spatial dimensions and denoted as $F \in \mathbb{R}^{b\times v\times c}$, where v corresponds to the number of voxel tokens and c is the encoder feature dimension. These dense volumetric features serve as the input for subsequent difference modeling.

3.3 Hierarchical Difference Extractor

We aim to capture visual differences between the input and reference scans to guide the LLM in report generation. In 3D CT volumes, organ- and structure-

level variations manifest at multiple spatial scales, ranging from global morphological shifts to subtle localized deviations within specific anatomical regions. Therefore, a single-scale comparison is often insufficient to represent such heterogeneous differences. To address this, we design the Hierarchical Difference Extractor (HDE) to model discrepancies at two complementary levels. Specifically, a *Global Delta Operator* captures holistic structural and semantic shifts, while a *Local Delta Operator* aggregates fine-grained deviations, jointly providing multi-scale difference cues for report generation.

Global Delta Operator Our objective is to capture global differential semantics between a target CT volume X and a normative reference X_{ref}, enabling the model to reason about *what has changed* at a holistic level rather than processing each input independently.

Both volumes are first encoded by a 3D visual encoder into dense volumetric feature maps, $F, F_{\text{ref}} \in \mathbb{R}^{b\times v\times c}$. However, due to the prohibitive magnitude of the voxel count v, directly operating on these features is computationally infeasible and introduces substantial redundancy, as most voxels correspond to common anatomical background with limited differential information.

Specifically, we adopt a Q-Former–style Visual Resampler to perform semantic downsampling. We initialize a shared set of n learnable latent queries $L \in \mathbb{R}^{n\times d}$, which serve as semantic anchors for both the target and reference volumes. Using cross-attention, these shared queries aggregate informative visual content from the dense features into compact latent representations:

$$\begin{aligned} I &= \text{Resampler}(F, L) \in \mathbb{R}^{b\times n\times d}, \\ I^r &= \text{Resampler}(F_{\text{ref}}, L) \in \mathbb{R}^{b\times n\times d}, \end{aligned} \tag{3}$$

By enforcing a shared query set, this resampling process projects both volumes into a common latent space, ensuring structural alignment and enabling direct comparison at the token level.

Given the resampled latent tokens I and I^r, we then explicitly model their global difference semantics. We introduce a learnable global difference query $\delta \in \mathbb{R}^d$, which is designed to summarize holistic discrepancies between the target and reference volumes. This query interacts jointly with both token sets through a Transformer layer:

$$\begin{aligned} \Delta_{\text{global}} &= \mathcal{D}_{diff}(I, I^r) \\ &= \text{Transformer}_\theta\left([\delta;\, I;\, I^r]\right)_\delta, \end{aligned} \tag{4}$$

Here, $[\cdot;\cdot]$ denotes concatenation along the token dimension, and $(\cdot)_\delta$ indicates extracting the output state corresponding to the query token δ. The resulting Δ_{global} encapsulates the overall structural and semantic shift between X and X_{ref}, providing a compact global difference representation to condition downstream report generation.

Local Delta Operator While Δ_{global} captures high-level semantic discrepancies, it may not explicitly encode the strength of localized deviations. To complement it, we introduce a *local-level difference aggregation* mechanism that operates directly on the aligned latent visual tokens.

Formally, given the input and reference tokens I, I^r, we define a local difference operator $\mathcal{A}_{\text{diff}}$ as

$$\Delta_{\text{local}} = \mathcal{A}_{\text{diff}}(I, I^r)\,, \tag{5}$$

where $\Delta_{\text{local}} \in \mathbb{R}^{b\times d}$ summarizes localized deviations via a geometry-aware weighting scheme in the latent space.

Specifically, we assign each aligned token pair (I_i, I_i^r) a normalized importance weight w_i based on their squared Euclidean distance:

$$w_i = \frac{\|I_i - I_i^r\|_2^2}{\sum_{j=1}^{N} \left\|I_j - I_j^r\right\|_2^2 + \epsilon}, \quad i = 1, \ldots, N. \tag{6}$$

This weighting emphasizes token pairs that deviate most from the reference while remaining robust to scale via normalization, where $\epsilon = 1 \times 10^{-5}$ is a small term added to prevent division by zero.

The local difference representation is then obtained by aggregating the residual vectors under these weights:

$$\mathcal{A}_{\text{diff}}(I, I^r) = \sum_{i=1}^{N} w_i \left(I_i - I_i^r\right). \tag{7}$$

This operator is parameter-free and grounded in the metric structure of the latent token space. It ensures that tokens exhibiting pronounced deviations contribute dominantly to Δ_{local}, providing a direct and interpretable difference evidence signal for the LLM during report generation.

3.4 Difference-to-Prompt Generator

The HDE module yields two complementary difference representations: a global descriptor Δ_{global} and a locality-sensitive descriptor Δ_{local}. The Difference-to-Prompt Generator (DPG) transforms these visual differences into a learnable prefix that can directly condition the LLM.

Since the extracted differences reside in the vision latent space, they must be projected into the LLM embedding space. To jointly encode holistic shifts and localized deviations, we first fuse Δ_{global} and Δ_{local} via concatenation and an MLP, and then expand the fused representation into a sequence of prefix tokens:

$$\Delta_{\text{prompt}} = \text{Projector}\big([\Delta_{\text{global}}\,;\,\Delta_{\text{local}}]\big). \tag{8}$$

The resulting $\Delta_{\text{prompt}} \in \mathbb{R}^{p\times d_{\text{LLM}}}$ consists of p learnable prefix tokens (p=16 in our study), with dimensionality matching the LLM embedding size. By injecting explicit scan-to-reference discrepancy cues initially, Δ_{prompt} acts as a dynamic soft prompt that reweights conditioning signals, suppressing background-dominated representations and enhancing report generation.

Diagnostic Semantic Guidance. To provide high-level semantic guidance, we incorporate predicted diagnostic classes E, a standard anchoring paradigm in medical report generation [10]. Specifically, E is generated by a pre-trained 3D ResNet-18 classifier optimized via Binary Cross-Entropy (BCE) loss over 18 report-independent disease labels (*e.g.* Cardiomegaly, Emphysema), preventing information leakage. During generation, this classifier remains frozen, and its predictions are converted into short textual prompts. Detailed class definitions, training specifics, and the rationale for this two-stage design are deferred to the Supplementary Material.

LLM Input Construction. To preserve complete visual context, we retain the original tokens I with the difference prefix Δ_{prompt}. We then integrate diagnostic guidance E and append structured textual prompts T to aid instruction following. The final input sequence for the LLM is formulated as:

$$\mathcal{X}_{\text{in}} = [\, \Delta_{\text{prompt}} \,;\, I \,;\, E \,;\, T \,]. \tag{9}$$

3.5 Training Objective

We train the model with a multi-task objective that combines report generation with an auxiliary disease classification task.

Report Generation Loss. We formulate report generation as conditional sequence modeling. Given the hybrid input embeddings $\mathcal{X}_{\text{in}}$ and the ground-truth report $Y = \{y_1, \ldots, y_L\}$, we minimize the negative log-likelihood (NLL):

$$\mathcal{L}_{\text{gen}} = -\sum_{t=1}^{L} \log P(y_t \mid y_{<t}, \mathcal{X}_{\text{in}}; \theta)\,. \tag{10}$$

Diagnostic Classification Loss. To encourage diagnostic consistency and regularize the visual representations, we add an auxiliary multi-label classification loss $\mathcal{L}_{\text{cls}}$ using binary cross-entropy (BCE) over K=18 disease categories.

Total Objective. The final training objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{gen}} + \mathcal{L}_{\text{cls}}. \tag{11}$$

4 Experiments

4.1 Datasets

RadGenome-ChestCT. The RadGenome-ChestCT dataset [31], derived from the CT-RATE dataset [7], comprises 25,692 CT-report pairs from 21,304 unique patients. We adhere to the official partition, designating 24,128 pairs for training and 1,564 pairs for evaluation. The data is pre-processed into ten anatomical regions with a standardized voxel spacing of 1×1×3 mm.

CTRG-Chest-548K. As a supplementary benchmark [20], it contains 1,804 CT report pairs, split into train and test sets with an 8:2 ratio. We further extract anatomical region masks and parse reports into region-level descriptions.

Table 1: Quantitative results on RadGenome-ChestCT and CTRG-Chest-548K. Models marked with † are foundation models. We report BLEU (BL), ROUGE-L (RG-L), and METEOR (MTR) scores. The best, second best, and third best results are highlighted in dark blue, medium blue, and light blue, respectively.

Dataset	Method	Year	BL-1	BL-2	BL-3	BL-4	RG-L	MTR
RadGenome-ChestCT	† RadFM	2023	44.20	34.49	28.06	23.65	31.53	39.94
	† M3D	2023	43.57	34.48	28.54	24.49	32.61	39.95
	R2GenGPT	2023	43.28	34.11	28.16	24.16	32.26	39.85
	MedVInT	2024	44.28	34.91	28.75	24.60	32.58	40.39
	CT2Rep	2024	44.42	34.43	27.94	23.56	30.99	40.16
	Reg2RG	2025	47.25	36.49	29.57	24.87	**36.65**	44.07
	Ours	-	**58.16**	**48.74**	**40.93**	**34.28**	32.32	**47.40**
CTRG-Chest-548K	† RadFM	2023	48.66	40.28	34.73	30.89	49.08	49.18
	† M3D	2023	46.27	39.02	34.23	30.86	50.24	49.26
	R2GenGPT	2023	41.82	36.37	32.70	30.10	**50.93**	47.05
	MedVInT	2024	47.38	39.60	34.28	30.68	49.53	49.32
	CT2Rep	2024	42.28	36.16	32.08	29.19	50.17	47.00
	Reg2RG	2025	49.63	41.43	35.91	32.04	47.76	49.71
	Ours	-	**50.37**	**46.57**	**41.94**	**37.57**	45.00	**50.44**

Normal Reference Set Construction. To facilitate explicit difference modeling, we construct a dedicated normal reference pool by selecting CT volumes where the associated reports indicate no abnormalities. Based on this criterion, we select 1,787 training and 124 evaluation samples from RadGenome-ChestCT, each used in its training and testing phase. Similarly, for the CTRG-Chest-548K dataset, we select 105 training and 43 testing samples as normal references. References are sampled strictly within the corresponding split, train-only for training, test-only for testing, preventing any train–test leakage. In the supplementary material, we further report a five-seed stability analysis with mean±std and two-sided t-tests under random reference pairing.

4.2 Implementation Details

We utilize a pretrained frozen 3D ResNet-18 as the visual encoder to extract voxel-level features ($d = 512$). For the language backbone, we adopt LLaMA-2-7B [21] equipped with Low-Rank Adaptation (LoRA) [9] for efficient fine-tuning. All models are optimized with a learning rate of 5×10^{-5}. We train the models for 10 epochs with a batch size of 1 per GPU. Experiments are conducted on a cluster of 4 NVIDIA H20 GPUs trained in parallel.

Table 2: Clinical Efficacy (CE) performance on RadGenome-ChestCT. The symbol † denotes generalist foundation models, while unmarked methods are specialized Report Generation Models.

Method	Clinical Efficacy (CE)		
	F1-Score	Recall	Precision
†RadFM	0.195	0.131	0.382
†M3D	0.148	0.090	0.407
R2GenGPT	0.110	0.066	0.340
MedVInT	0.212	0.148	0.377
CT2Rep	0.139	0.089	0.317
Reg2RG	0.253	0.181	**0.423**
Ours	**0.421**	**0.496**	0.366

Table 3: Impact of Visual Prefix Length. We vary the visual prefix length $p \in \{4, 8, 16, 24, 32\}$ on RadGenome-ChestCT. NLG performance increases with p and peaks at p=16 .

Metric	Prefix Length (p)				
	4	8	16	24	32
BLEU-1	53.97	57.98	**58.16**	56.89	56.24
BLEU-2	44.29	48.60	**48.74**	47.46	46.51
BLEU-3	36.37	40.78	**40.93**	39.79	38.43
BLEU-4	29.72	34.13	**34.28**	33.32	31.53
METEOR	43.37	**47.96**	47.40	47.33	46.19
ROUGE-1	59.63	63.48	**63.52**	62.33	61.82
ROUGE-2	38.41	42.60	**42.63**	41.49	40.38
ROUGE-L	31.46	**32.43**	32.32	32.27	31.23

4.3 Evaluation Metrics

We evaluate linguistic quality using standard generation metrics, specifically BLEU-n with n ranging from 1 to 4, METEOR, and ROUGE-1, ROUGE-2, and ROUGE-L, and medical correctness using Clinical Efficacy (CE) metrics. CE computes Precision, Recall, and F1 on 18 pathologies extracted by a RadBERT classifier. Due to label misalignment between RadBERT's training data and the primary abnormalities in CTRG-Chest-548K, CE metrics are reported only for RadGenome-ChestCT.

4.4 Comparison with SOTA Methods

Natural Language Generation Performance. Table 1 summarizes the NLG results on RadGenome-ChestCT and CTRG-Chest-548K. We compare DiffVP with large-scale medical foundation models (RadFM, M3D) and dedicated report generation methods (R2GenGPT, MedVInT, CT2Rep, Reg2RG). Overall, DiffVP achieves the best or competitive performance across most metrics on both datasets. Compared with the previous SOTA Reg2RG, DiffVP delivers large gains in n-gram precision, improving BLEU-1 by 23.08% and BLEU-4 by 37.83%. DiffVP also attains the highest METEOR, indicating better semantic alignment with reference reports. Although Reg2RG is slightly higher on ROUGE-L, DiffVP's stronger BLEU and METEOR suggest improved precision in describing clinically relevant findings rather than relying on long shared subsequences.

Clinical Efficacy Analysis. Table 2 reports CE results on RadGenome-ChestCT. DiffVP achieves the best F1-score (0.421) and the highest Recall among all methods, indicating improved sensitivity to clinically relevant findings and a reduced risk of missing positives. While Reg2RG is marginally higher in Precision, DiffVP provides a better overall balance between Precision and Recall.

Table 4: Ablation study on the effectiveness of different modules on RadGenome-ChestCT. **Global**, **Local**, and **E** denote the Global Delta, Local Delta, and Classification Prompt modules, respectively. CE Metrics denote clinical efficacy metrics including Precision, Recall, and F1.

ID	Module Settings			NLG Metrics								CE Metrics		
	Δ_{global}	Δ_{local}	E	B-1	B-2	B-3	B-4	M	R-1	R-2	R-L	Pre.	Rec.	F1
Baseline				55.29	46.55	39.36	33.17	44.11	62.12	41.95	31.94	0.295	0.109	0.159
#1			✓	55.62	46.89	39.65	33.44	44.80	62.31	42.19	32.35	0.311	0.217	0.256
#2	✓		✓	57.46	47.92	39.97	33.20	46.63	62.97	41.69	32.08	**0.373**	0.459	0.411
#3		✓	✓	57.68	48.24	40.47	33.89	**47.67**	63.19	42.25	**32.76**	0.360	0.464	0.406
Ours	✓	✓	✓	**58.16**	**48.74**	**40.93**	**34.28**	47.40	**63.52**	**42.63**	32.32	0.366	**0.496**	**0.421**

5 Ablation Studies and Analysis

5.1 Ablation Studies

Ablation Study on Model Components. Table 4 reports an ablation analysis of DiffVP. We start from a baseline with a pretrained ResNet18 encoder on RadGenome-ChestCT. Adding the inferred diagnostic anchor (#1) provides consistent improvements, notably in CE Recall, and offers a stable semantic context for subsequent difference prompting. Comparing #2 and #3 shows that modeling discrepancies at complementary levels is beneficial: the **Global Delta** captures holistic structural and semantic shifts and improves global consistency, while the **Local Delta** contributes fine-grained deviations that boost METEOR. Combining both yields the best BLEU-4, indicating that joint global consistency and local detail improves long-range coherence and overall report quality.

Impact of Visual Prefix Length. Table 3 studies the visual prefix length p. Performance increases as p grows from 4, with BLEU-1–4 peaking at p=16. While p=8 remains competitive for Recall, larger prefixes (24/32) consistently degrade performance, suggesting redundant conditioning. We therefore set p=16 by default.

5.2 Performance Analysis

Visual semantic difference map. Fig. 6(a) explicitly visualizes the critical discrepancy cue extracted by our difference-aware module. Given an input CT scan and its paired normal reference, we compute multi-scale discrepancy in the latent visual token space and project it back to the image plane for interpretability. This map is not obtained by simple pixel-wise subtraction; instead, it highlights semantically deviant regions relative to the normal prior, which is consistent with radiologists' high-level comparative reasoning (*e.g.* for the first pair in Fig. 6(a), the activation concentrates around the lung fields where the report describes subsegmental atelectasis in the right lower lobe and multiple pulmonary nodules, most prominent in the left upper-lobe lingular segment).

Impact of discrepancy level. Fig. 6(b) compares three variants:*w/o Diff*, which removes the discrepancy cue; *Pixel-level Diff*, which derives the cue from

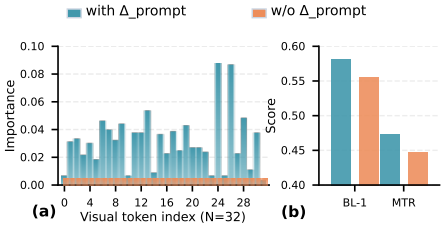


Fig. 3: Visual token importance. (a) Normalized importance distributions for N=32 latent tokens with (blue) and without (orange) the proposed Δ_{prompt}. (b) Performance gain by Δ_{prompt}.

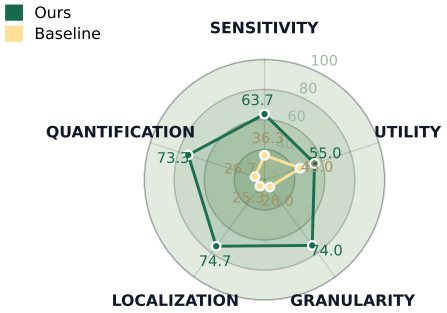


Fig. 4: LLM-as-a-Judge Evaluation.We employ GPT-5 for pairwise comparisons across five clinical axes: Granularity, Localization, Quantification, Sensitivity, and Utility

pixel-wise subtraction in the image space; and *Ours*, which computes discrepancy in the latent token space. Importantly, Pixel-level Diff and Ours share the same discrepancy-prompt injection form with the same prefix structure and length, and thus differ only in where the discrepancy is computed. Semantic discrepancy consistently yields better BLEU-1 and CE-F1, suggesting that clinically useful comparison is better captured at a semantic level than by raw pixel subtraction.

Analysis of the Δ_{prompt} Distribution. To clarify the visual prefix mechanism, Fig. 3(a) visualizes the token importance distribution, defined as the latent geometric discrepancy: $S_i = \|I_i - I_i^r\|_2^2$. This metric quantifies local clinical deviations from the reference in semantic space. As shown, the baseline (orange) exhibits a near-uniform distribution, failing to distinguish critical findings from massive anatomical redundancy. In sharp contrast, Δ_{prompt} (blue) induces a highly sparse distribution, where the top 8 tokens capture 45% of the total importance mass. This shift proves our prefix forces the model to focus on localized clinically-variant regions rather than treating all tokens equally. Such difference-guided focus aligns with the linguistic gains in Fig. 3(b), underscoring the necessity of prioritizing semantically variant regions for high-quality report generation.

LLM-as-a-Judge: Multi-dimensional Pairwise Evaluation. We use GPT-5 as a reference-guided judge for randomized pairwise comparisons in Figure 4 for multi-dimensional analysis. We evaluate descriptive quality using Granularity, Localization, and Quantification, which assess linguistic richness, anatomical specificity, and the use of measurable descriptors, respectively. We further assess clinical substance via Sensitivity and Utility, measuring coverage of reference findings and potential decision support. As shown in the radar chart, our method demonstrates a dominant advantage in descriptive quality, achieving superior win-rates across Granularity, Localization, and Quantification. This automated

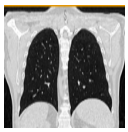
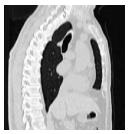
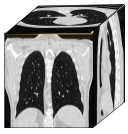
Visuals		Excerpt of the generated report
	GT	**Abdomen:** **A decrease in liver density, consistent with steatosis**, is observed. Bilateral adrenal glands were normal and no space-occupying lesion was detected. Thoracic aorta diameter is normal. **Lung:** Several **nonspecific nodules** are observed in both lungs, the largest of which is in the left lung lower lobe anteromedial segment (**4.6 mm**). There was no finding in favor of active infiltration in both lungs.
	Comparison Method	**Abdomen:** Bilateral adrenal glands were normal and **no space-occupying lesion was detected**. Upper abdominal organs included in the sections are **normal**. *[Missed Steatosis]* **Lung:** Trachea and both main bronchi are open. The lungs are **clear of active infiltration**. *[Missed Nodules]*
	Ours (w/o Δ prompt)	**Abdomen:** Bilateral adrenal glands were normal. Upper abdominal organs included in the sections are **normal**. *[Missed Steatosis]* No space-occupying lesion was detected in the liver. **Lung:** **No mass lesion** with distinguishable borders was detected. *[Missed Nodule]* However, when examined in the parenchyma window, **Linear subsegmental atelectasis was observed** in the middle lobe of the right lung. *[Fine-grained detail]*
	Ours	**Abdomen:** A **decrease in liver density** is noted, consistent with **steatosis**. Bilateral adrenal glands normal and no space-occupying lesion was detected. **Lung:** **Nonspecific nodules** are seen in the lung parenchyma. The largest nodule is measured at **4.6 mm** in the left lower lobe. There was no finding in favor of active infiltration.

Fig. 5: Qualitative case study comparing the Ground Truth, the Comparison Method, the ablation variant without the Δ_{prompt}, and the full model. Correctly identified pathologies are highlighted in **blue** and missed detections are marked with **crosses**.

assessment provides a more detailed way of analyzing data, rather than replacing report generation metrics.

Case Study Figure 5 shows qualitative comparisons. The comparison method, using a uniform input strategy, tends to generate template-like reports and miss subtle findings. In contrast, DiffVP uses reference-guided prompting to suppress redundant anatomy and emphasize informative evidence. The variant without Δ_{prompt} already captures linear subsegmental atelectasis missed by the baseline, while the full model with Δ_{prompt} further focuses on sparse deviations and recovers finer details such as the 4.6 mm measurement and steatosis.

6 Conclusion

In this paper, we propose DiffVP, a reference-guided framework that brings 3D CT report generation closer to the comparative reasoning used in clinical practice. DiffVP explicitly encodes scan-to-reference discrepancies and converts them into visual delta prompts via a hierarchical difference extractor and a difference-to-prompt generator. Injected as learnable prefixes to an LLM, these prompts provide structured visual guidance that suppresses redundant anatomical background and highlights informative evidence for report generation. Experiments

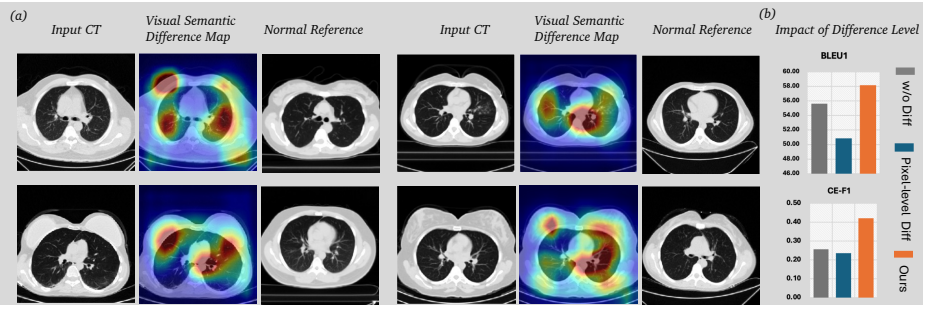


Fig. 6: Visual semantic discrepancy and discrepancy-level comparison. (a) Two examples showing the input CT, the *Visual Semantic Difference Map* produced by our method, and the paired normal reference. (b) Performance comparison among *w/o Diff*, *Pixel-level Diff* (pixel-wise subtraction), and *Ours* on BLEU-1 and CE-F1.

show that DiffVP consistently surpasses state-of-the-art baselines in both language metrics and clinical efficacy. This work supports reference-guided prompting as an effective direction for medical report generation.

Acknowledgment and Author Contributions

This work was supported in part by the National Natural Science Foundation of China under Grants 62502490 and 62271465; in part by the Natural Science Foundation of Jiangsu Province under Grants BK20250496 and BK20255001; in part by the China Postdoctoral Science Foundation under Grant 2024M763178; in part by the Jiangsu Funding Program for Excellent Postdoctoral Talent; in part by the National Key R&D Program of China under Grant 2025YFC3408300; and in part by the Suzhou Basic Research Program under Grant SYG202338.

Y.T. and K.Z. jointly led this study and contributed to the methodological design. Y.T. conducted the experimental validation and drafted the manuscript. K.Z. conceived the main idea, contributed to manuscript writing, and supervised the overall research process. H.M., R.Y., Y.L., and R.W. participated in discussions and manuscript revision. S. Kevin Z. provided overall supervision.

References

1. Bai, F., Du, Y., Huang, T., Meng, M.Q.H., Zhao, B.: M3d: Advancing 3d medical image analysis with multi-modal large language models. arXiv preprint arXiv:2404.00578 (2024)
2. Bian, M., Zhang, K., Zhao, D., Zhou, S.K.: Diffrgennet: Difference-aware medical report generation. In: Medical Imaging with Deep Learning. pp. 1–14 (2025)
3. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.16056 (2020)

4. Chen, Z., Bie, Y., Jin, H., Chen, H.: Large language model with region-guided referring and grounding for ct report generation. IEEE Transactions on Medical Imaging (2025)
5. Chen, Z., Luo, L., Bie, Y., Chen, H.: Dia-llama: Towards large language model-driven ct report generation. In: MICCAI. pp. 141–151. Springer (2025)
6. Deng, X., He, X., Bao, J., Zhou, Y., Cai, S., Cai, C., Chen, Z.: Mvketr: chest ct report generation with multi-view perception and knowledge enhancement. IEEE Journal of Biomedical and Health Informatics (2025)
7. Hamamci, I.E., Er, S., Almas, F., Simsek, A.G., Esirgun, S.N., Dogan, I., Dasdelen, M.F., Wittmann, B., Simsar, E., Simsar, M., et al.: A foundation model utilizing chest ct volumes and radiology reports for supervised-level zero-shot detection of abnormalities. CoRR (2024)
8. Hamamci, I.E., Er, S., Menze, B.: Ct2rep: Automated radiology report generation for 3d medical imaging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 476–486. Springer (2024)
9. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
10. Kale, K., Bhattacharyya, P., Jadhav, K.: Replace and report: Nlp assisted radiology report generation. In: Findings of the Association for Computational Linguistics: ACL 2023. pp. 10731–10742 (2023)
11. Kundel, H.L., Nodine, C.F.: Interpreting chest radiographs without visual search. Radiology **116**(3), 527–532 (1975)
12. Li, W., Yan, R., Zhang, X., Chen, L., Zhu, H., Zhao, J., Li, J., Li, M., Cao, W., Jiang, Z., et al.: Macd: Multi-agent clinical diagnosis with self-learned knowledge for llm. arXiv preprint arXiv:2509.20067 (2025)
13. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for 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. 4582–4597 (2021)
14. Liu, C., Wan, Z., Wang, Y., Shen, H., Wang, H., Zheng, K., Zhang, M., Arcucci, R.: Argus: benchmarking and enhancing vision-language models for 3d radiology report generation. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 16448–16460 (2025)
15. Lyu, C., Qiu, C., Han, K., Li, S., Sheng, V.S., Rong, H., Song, Y., Liu, Y., Liu, Z.: Automatic medical report generation combining contrastive learning and feature difference. Knowledge-Based Systems **305**, 112630 (2024)
16. Mao, Y., Xu, W., Qin, Y., Gao, Y.: Ct-agent: A multimodal-llm agent for 3d ct radiology question answering. arXiv preprint arXiv:2505.16229 (2025)
17. Marcos, L., Babyn, P., Alirezaie, J.: Pure vision transformer (ct-vit) with noise2neighbors interpolation for low-dose ct image denoising. Journal of Imaging Informatics in Medicine **37**(5), 2669–2687 (2024)
18. Park, H., Kim, K., Yoon, J., Park, S., Choi, J.: Feature difference makes sense: a medical image captioning model exploiting feature difference and tag information. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. pp. 95–102 (2020)
19. Song, S., Tang, H., Yang, H., Li, X.: Ddatr: Dynamic difference-aware temporal residual network for longitudinal radiology report generation. arXiv preprint arXiv:2505.03401 (2025)
20. 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)

21. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
22. Wang, C., Li, F., Hu, W., Yan, R., Zhang, K., Zhou, S.K.: Concept bottleneck models for explainable decision making: A survey of progress, taxonomy, and future directions. In: Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (2026)
23. Wang, C., Zhang, K., Liu, Y., He, Z., Tao, X., Zhou, S.K.: Mvp-cbm: multi-layer visual preference-enhanced concept bottleneck model for explainable medical image classification. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence. pp. 529–537 (2025)
24. Wang, Z., Liu, L., Wang, L., Zhou, L.: R2gengpt: Radiology report generation with frozen llms. Meta-Radiology **1**(3), 100033 (2023)
25. 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)
26. Wang, Z., Zhou, L., Wang, L., Li, X.: A self-boosting framework for automated radiographic report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2433–2442 (2021)
27. Wu, C., Zhang, X., Zhang, Y., Hui, H., Wang, Y., Xie, W.: Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. Nature Communications **16**(1), 7866 (2025)
28. Yun, H., Maeng, J., Kang, E., Suk, H.I.: Diff-rrg: Longitudinal disease-wise patch difference as guidance for llm-based radiology report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 152–161. Springer (2025)
29. Zhang, K., Li, J., Li, Z., Zhou, S.K.: Dh-set: Improving vision-language alignment with diverse and hybrid set-embeddings learning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24993–25003 (2025)
30. Zhang, K., Mao, Z., Liu, A.A., Zhang, Y.: Unified adaptive relevance distinguishable attention network for image-text matching. IEEE Transactions on Multimedia **25**, 1320–1332 (2022)
31. Zhang, X., Wu, C., Zhao, Z., Lei, J., Zhang, Y., Wang, Y., Xie, W.: Radgenome-chest ct: A grounded vision-language dataset for chest ct analysis. arXiv preprint arXiv:2404.16754 (2024)
32. Zhang, Z., Li, N., Liu, Q., Li, R., Gao, W., Mao, Q., Huang, Z., Yu, B., Tao, D.: The other side of the coin: Exploring fairness in retrieval-augmented generation. arXiv preprint arXiv:2504.12323 (2025)
33. Zhang, Z., Liu, Q., Hu, Z., Zhan, Y., Huang, Z., Gao, W., Mao, Q.: Enhancing fairness in meta-learned user modeling via adaptive sampling. In: Proceedings of the ACM Web Conference 2024. pp. 3241–3252 (2024)
34. Zhang, Z., Liu, Q., Hu, Z., Zhan, Y., Huang, Z., Gao, W., Mao, Q., Chen, E.: A hybrid adaptive sampling strategy for fair and accurate meta-learned user modeling. ACM Transactions on Information Systems **44**(1), 1–39 (2025)
35. Zhang, Z., Liu, Q., Jiang, H., Wang, F., Zhuang, Y., Wu, L., Gao, W., Chen, E.: Fairlisa: Fair user modeling with limited sensitive attributes information. In: Advances in Neural Information Processing Systems. vol. 36, pp. 41432–41450 (2023)
36. Zhao, G., Yan, Y., Zhao, Z.: Normal-abnormal decoupling memory for medical report generation. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 1962–1977 (2023)