HiFi-Rep: Leveraging High-Resolution Vision Foundation Models with Tri-Planar Slice Context for CT Report Generation

Keetawan Limaroon*, Punnawat Namwongsa*, and Sansiri Tarnpradab✉

Department of Computer Engineering, Faculty of Engineering,
King Mongkut's University of Technology Thonburi, Bangkok, Thailand
sansiri.tarn@mail.kmutt.ac.th

Abstract. Automated radiology report generation from 3D CT volumes holds immense potential to alleviate radiologist workloads. Current methods, however, face a critical trade-off between efficiency and resolution: many often downsample volumes to fit memory constraints, discarding fine-grained anatomical details essential for diagnosis, while 3D-native models remain computationally prohibitive. To address this, we propose **HiFi-Rep**, a streamlined framework that leverages high-fidelity visual features without the burden of 3D encoders. Our method processes native 512×512 resolution slices across tri-planar views using the pre-trained DINOv3 vision foundation model. Furthermore, to resolve sequence length bottlenecks, we introduce **HiFi-S2D**, an efficient aggregation module that employs Space-to-Depth transformation. This allows for spatial-preserving aggregation, condensing volumetric depth while strictly retaining high-resolution in-plane diagnostic cues for MedGemma, a specialized medical LLM. We validate HiFi-Rep on the large-scale CT-RATE report generation benchmark, where it establishes a new state-of-the-art. Compared with the best-performing baseline, our proposed approach yields a 9.2% improvement in BLEU-1 and achieves the highest reported METEOR score among existing methods. These results demonstrate that combining robust 2D foundation models with structure-preserving aggregation offers a superior strategy for 3D medical understanding, paving the way for scalable, high-precision automated reporting. Our code is publicly available at `https://github.com/rqobistp/HiFi-Rep`.

Keywords: Medical Report Generation · Vision-Language Models · Feature Aggregation · Foundation Models · CT-RATE

1 Introduction

Automated radiology report generation is a powerful tool in medical AI that aims to improve efficiency and consistency in diagnostics while addressing critical challenges such as radiologist shortages and diagnostic discrepancies [23,26].

* Equal contribution.

Although Large Language Models (LLMs) have demonstrated remarkable proficiency in interpreting 2D medical images like X-rays [24], applying them to 3D volumetric scans like Computed Tomography (CT) presents significantly greater challenges. A single CT scan contains hundreds of slices. To read it correctly, a system must understand the entire organ's structure while spotting tiny details, such as small lung nodules or hairline fractures. The main challenge is finding a way to process this massive amount of data accurately without incurring prohibitive computational costs.

Early attempts to solve this used native 3D models (like 3D-ViTs). However, these models require immense memory and computing power. To make them work, pioneering methods like CT2Rep [6] and E3D-GPT [10] had to aggressively shrink the input images to very small sizes (e.g., $224 \times 224 \times 112$). Unfortunately, shrinking the image often deletes the diagnostic details needed for an accurate report [26]. This limitation led researchers to shift toward slice-based approaches, which look at 2D slices individually before combining them.

However, current slice-based methods [8,12] still face a resolution-efficiency paradox. They typically rely on complex, resource-intensive mechanisms to fuse slice information. Due to the prohibitive computational costs of such aggregation, these models are forced to aggressively shrink the input images (e.g., to 256×256 in SAMF [8]), inadvertently discarding critical diagnostic details. Furthermore, these methods often utilize rigid vision backbones that lack flexibility regarding input resolution, failing to capture the high-density features essential for detecting subtle lesions [22,3]. Consequently, they neglect useful multi-view data (Coronal and Sagittal) to save memory, and utilize generic language models lacking specialized medical knowledge, leading to inaccurate reports.

To address these challenges, we propose **HiFi-Rep**, an architecture that builds upon the premise established by recent efficient frameworks like Raptor [1], which demonstrated that 2D foundation models can handle 3D volumes via linear aggregation strategies. While successful in classification tasks, we argue that their aggressive dimensionality reduction is overly compressive for report generation, where preserving fine-grained anatomical details is crucial. To overcome this limitation, we introduce the HiFi-S2D aggregation module, a Space-to-Depth-based transformation applied after inter-slice mean pooling [21]. Rather than relying on destructive projection, this operation reduces token sequence length by rearranging spatial information into the channel dimension [17,11], thereby preserving structural fidelity while satisfying LLM constraints.

This efficiency allows us to process native 512×512 slices using the training-free DINOv3 [22], a vision foundation model that provides high-quality dense features. As a result, optimization is concentrated solely on the specialized medical LLM, MedGemma [20]. By coupling high-fidelity visual cues with a domain-specific language model, we mitigate the medical knowledge gap inherent in generic models, thus ensuring that fine-grained findings are translated into accurate clinical terminology. Our main contributions are as follows:

– We propose **HiFi-Rep**, a novel high-resolution framework that establishes a new state-of-the-art on the CT-RATE report generation task. By effi-

ciently bridging 2D general-purpose vision features with volumetric context for a specialized medical LLM, our method significantly outperforms existing baselines in both lexical fidelity and clinical accuracy.
– We demonstrate the efficacy of harnessing the high-quality dense features of the DINOv3, vision foundation model, for volumetric medical imaging. We show that these training-free, robust, domain-agnostic features can effectively generalize to medical tasks.
– We introduce a parameter-free aggregation module, HiFi-S2D, designed to compress visual tokens via a Space-to-Depth operation. This mechanism rearranges spatial pixels into the channel dimension, acting as a spatial-preserving aggregation that balances depth condensation with in-plane fidelity to resolve the long-sequence bottleneck.
– We devise an interleaved tri-planar instruction strategy that enables precise 3D spatial alignment through natural language cues. This allows the LLM to comprehend multi-view anatomical contexts without requiring explicit 3D positional encodings.

2 Methodology

2.1 Overview

We propose **HiFi-Rep**, a streamlined framework for high-resolution 3D report generation. As illustrated in Fig. 1, our pipeline consists of three stages: (1) High-Fidelity Visual Encoding, extracting dense features from native resolution slices using a pre-trained DINOv3 backbone; (2) HiFi-S2D Aggregation, a parameter-free module that compresses volumetric sequences via inter-slice pooling and Space-to-Depth transformation; and (3) Tri-Planar LLM Decoding, where a specialized medical LLM (MedGemma) synthesizes reports from multi-view inputs guided by interleaved textual instructions.

2.2 High-Fidelity Visual Encoding

Given a CT volume, we extract slices along three anatomical planes: Axial, Coronal, and Sagittal. Unlike prior methods that downsample inputs, we process slices at their native resolution of 512×512 to preserve fine-grained diagnostic cues. We utilize the vision foundation model DINOv3 [22] as our feature extractor.

Formally, we represent an input volume view as a sequence of 2D slices $\mathcal{V} = \{x_i\}_{i=1}^{D}$, where $x_i \in \mathbb{R}^{H \times W}$ denotes the i-th slice. We apply the pre-trained DINOv3 backbone $\Phi(\cdot)$ to each slice independently to generate a stack of patch embeddings $\mathcal{E}$:

$$\mathcal{E} = \{\Phi(x_i)\}_{i=1}^{D} \in \mathbb{R}^{D \times N \times C}, \tag{1}$$

where D is the number of slices, N is the number of patches per slice, and C is the feature dimension. This training-free approach ensures robust, domain-agnostic feature extraction without the computational burden of fine-tuning a massive 3D encoder.

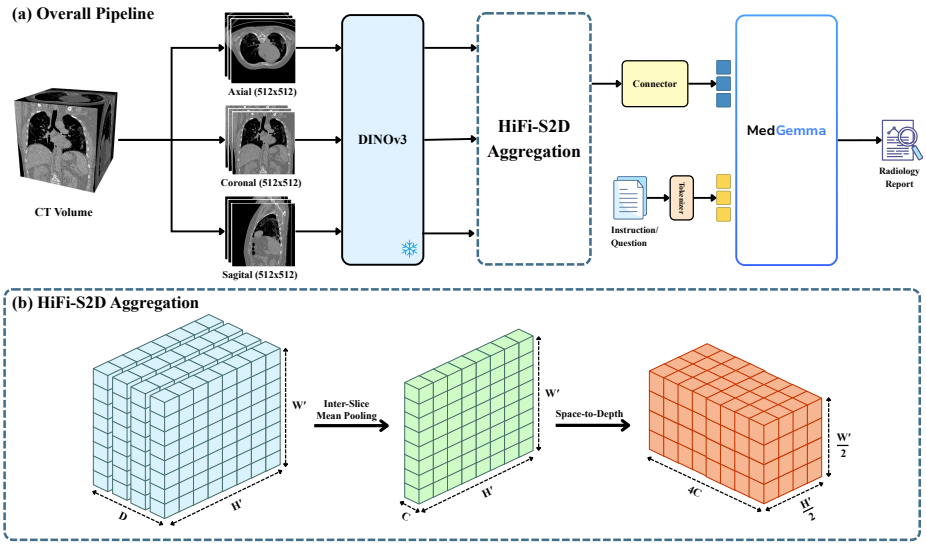

Fig. 1. Overview of the HiFi-Rep framework. (a) Frozen DINOv3 encodes triplanar slices, projecting features into MedGemma with textual instructions. (b) HiFi-S2D condenses volumetric features via inter-slice mean pooling and Space-to-Depth ($r = 2$), reducing sequence length while preserving high-resolution spatial details.

2.3 HiFi-S2D Aggregation

Standard flattening of volumetric features results in prohibitive sequence lengths for LLMs. To resolve this, we introduce **HiFi-S2D**, a parameter-free mechanism applied independently to each anatomical view (Axial, Coronal, and Sagittal).

For a given view v, let the encoded slice features be $\mathcal{E}^{(v)} \in \mathbb{R}^{D\times N\times C}$. We first reshape the patch sequence N into a 2D spatial grid $H' \times W'$ (where $N = H'W'$). The compression process then involves two steps:

1. **Inter-Slice Mean Pooling:** We aggregate anatomical context across the depth dimension of the specific view to form a unified 2D representation. This reduces the volume stack $\mathcal{E}^{(v)}$ to a condensed feature map $F_{avg}^{(v)}$:

$$F_{avg}^{(v)} = \frac{1}{D}\sum_{i=1}^{D} \mathcal{E}_i^{(v)} \in \mathbb{R}^{1\times H'\times W'\times C}, \tag{2}$$

   where $\mathcal{E}_i^{(v)}$ represents the reshaped feature map of the i-th slice.
2. **Space-to-Depth Transformation:** To efficiently encode high-resolution features without the information loss associated with downsampling or the massive token overhead of patch splitting [11], we apply a Space-to-Depth operation (Pixel Shuffle). This operation shifts spatial complexity into the channel dimension, preserving high-frequency diagnostic details:

$$F_{avg}^{(v)} \xrightarrow{\text{S2D}} Z'^{(v)} \in \mathbb{R}^{\frac{H'}{r}\times\frac{W'}{r}\times(C\cdot r^2)}. \tag{3}$$

Finally, a learnable linear projector maps the aggregated features $Z'^{(v)}$ from each view to the LLM's embedding dimension D_{LLM}. These projected tokens are then concatenated to form the complete visual input sequence.

2.4 Tri-Planar Instruction & Decoding

We leverage the language modeling capabilities of MedGemma-4B [20], a specialized medical LLM, to generate reports. Notably, we discard its native 2D vision encoder (MedSigLIP) entirely. Instead, we project our high-fidelity, multi-view features derived from DINOv3 into the LLM's embedding space. To enable precise 3D spatial alignment without complex positional embeddings, we devise an **Interleaved Tri-Planar Instruction** strategy. We concatenate the visual tokens from all three views (Z_{ax}, Z_{cor}, Z_{sag}) and inject natural language markers to guide the LLM's attention. The instruction prompt is structured as follows:

> *"You are an expert radiologist. Analyze the following multi-view CT features:*
> ***Axial View:*** $< Z_{ax} >$
> ***Coronal View:*** $< Z_{cor} >$
> ***Sagittal View:*** $< Z_{sag} >$
> *Based on these views, describe the findings ..."*

This text-guided structuring allows the LLM to comprehend multi-view anatomical contexts directly, ensuring that findings are localized and clinically accurate.

3 Experiments and Results

3.1 Experimental Setup

Dataset and Preprocessing. We validate our method on the CT-RATE report generation benchmark [7], a large-scale public dataset comprising chest CT volumes paired with radiology reports. The dataset is officially partitioned into a training set of 47,149 volume-report pairs and a validation set of 3,039 pairs.

Our preprocessing employs dynamic intensity scaling: background is filtered via a global median threshold, followed by 1-99% clipping and Min-Max normalization to $[0, 1]$. Tri-planar slices are then extracted, resized to 512×512, and ImageNet-standardized to align with the DINOv3 backbone.

Implementation Details. Our framework is implemented in PyTorch [19] using the Hugging Face ecosystem [5,16,25]. While 16 NVIDIA A100 GPUs were utilized to accelerate data-parallel training, the model maintains a minimal memory footprint suitable for single-GPU execution. To preserve visual robustness, the DINOv3 backbone [22] remains frozen, while we optimize the linear projector and LoRA modules [9] ($r = 64, \alpha = 128$) injected into MedGemma. Visual tokens are compressed via a Pixel Shuffle factor of $r = 2$. Training is performed using

AdamW [15] with a batch size of 16 and 512×512 resolution across two stages: (1) *Feature Alignment*, training only the projector for 5 epochs (LR 5×10^{-4}, constant schedule); and (2) *Instruction Tuning*, fine-tuning the projector and LoRA for 30 epochs (LR 1×10^{-4}, cosine annealing with 5% warmup).

Evaluation Metrics. We evaluate report generation performance using BLEU-1/4 [18], ROUGE-L [13], and METEOR [2] to capture lexical fidelity. Additionally, we employ BERTScore [27] to quantify high-level semantic alignment with ground-truth reports.

3.2 Results and Discussion

The quantitative results are presented in Table 1. Notably, HiFi-Rep establishes a new state-of-the-art on the CT-RATE report generation task, outperforming both volumetric and slice-based paradigms. By harnessing high-quality dense features from the pre-trained DINOv3 foundation model, we show that robust 2D representations encode volumetric context more effectively than resource-constrained native 3D architectures like E3D-GPT. Furthermore, when paired with a specialized medical LLM, our approach significantly improves upon the strongest slice-based baseline, SAMF + Ao2D. Consequently, we achieve a substantial 9.2% gain in BLEU-1 and secure the highest METEOR score of 0.424, validating that preserving native input resolution yields richer diagnostic cues than complex aggregation strategies applied to downsampled features.

Table 1. Comparative analysis of our approach against state-of-the-art methods for medical report generation on chest CT-scans. Bold values indicate best performance, while underlined values represent second-best results among baselines.

Model	Bleu1	Bleu4	RougeL	Meteor	Bert F1
3D Volumetric Methods					
E3D-GPT [10]	0.412	-	-	<u>0.418</u>	<u>0.880</u>
CT-AGRG [4]	-	0.172	0.280	0.196	0.867
CT2Rep [6]	0.309	0.172	0.243	0.173	0.865
CT-Chat [7]	0.395	-	0.321	0.219	-
2D Slice-Based Methods					
MS-VLM [12]	-	0.232	**0.438**	0.396	-
SAMF [8]	0.423	0.203	0.338	0.356	0.879
SAMF + *Ao2D* [8]	<u>0.440</u>	**0.261**	<u>0.417</u>	0.417	**0.889**
HiFi-Rep (Ours)	**0.532**	<u>0.255</u>	0.363	**0.424**	0.877

Complementing these quantitative gains, our qualitative analysis (Fig. 2) highlights three clinical validity patterns. First, Semantic Fidelity demonstrates robust visual understanding through clinically equivalent descriptions rather

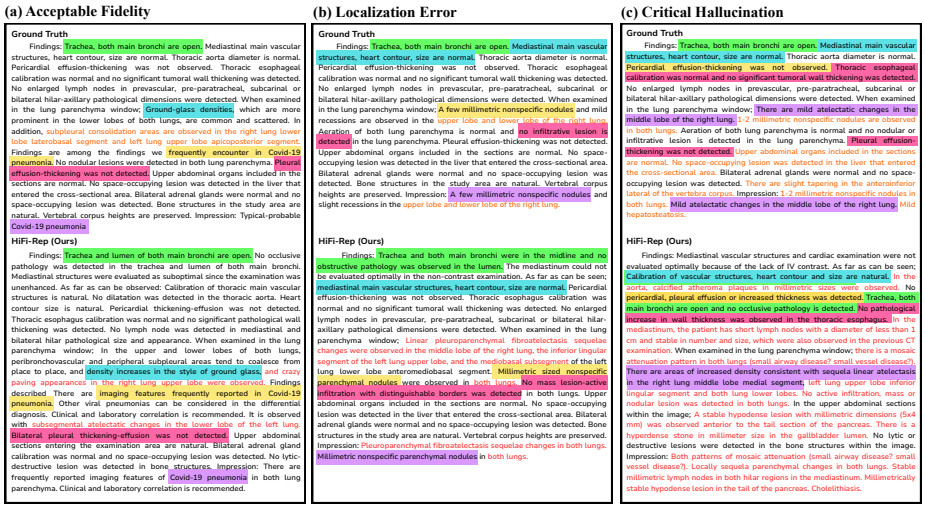

Fig. 2. Qualitative comparison of HiFi-Rep against ground truth. Matching colors denote semantic alignment. Orange in the ground truth represents omissions, while red in HiFi-Rep identifies hallucinations.

than rigid template matching. Second, Localization Sensitivity exhibits strong detection of fine-grained pathologies despite occasional struggles with precise spatial grounding. Lastly, Hallucinations and Omissions persist in non-target regions, reflecting the inherent trade-offs between generative fluency and factual strictness common in volumetric reporting.

3.3 Ablation Studies

We conduct ablation studies to validate three core components: (1) HiFi-S2D aggregation, (2) tri-planar anatomical context, and (3) interleaved instructions. All experiments follow the default configuration unless otherwise stated.

Effectiveness of HiFi-S2D Aggregation. To validate our premise that preserving spatial fidelity is critical for report generation, we compare **HiFi-S2D** against two destructive aggregation strategies: *Baseline Linear* [14] and *Random Projection* (Raptor) [1]. As shown in Table 2, HiFi-S2D consistently outperforms these baselines across all metrics. Unlike Random Projection, which aggressively compresses features and discards high-frequency diagnostic cues, HiFi-S2D spatially preserves and rearranges in-plane details into the channel dimension, compensating for pooled depth via tri-planar context. This structural preservation ensures the LLM retains access to fine-grained anatomical information, directly translating to superior lexical and semantic accuracy compared to destructive alternatives.

Table 2. Comprehensive ablation study. We analyze the impact of (a) aggregation strategies, (b) multi-view context, and (c) instruction prompting. Rows highlighted in gray denote our proposed **HiFi-Rep** configuration, which consistently yields the best balance of performance.

Model Configuration	Bleu-1	Bleu-4	Rouge-L	Meteor	Bert-F1
(a) Aggregation Strategy					
Baseline Linear [14]	0.498	0.219	0.329	0.393	0.872
Random Projection [1]	0.499	0.233	0.344	0.403	0.875
HiFi-S2D (Ours)	**0.532**	**0.255**	**0.363**	**0.424**	**0.877**
(b) Anatomical Views					
Axial Only	0.520	0.246	0.354	0.414	0.876
Axial + Coronal	0.524	0.248	0.357	0.418	0.876
Tri-Planar (A+C+S)	**0.532**	**0.255**	**0.363**	**0.424**	**0.877**
(c) Instruction Strategy					
Standard Concatenation	0.510	0.249	**0.365**	0.416	**0.878**
Interleaved Instructions	**0.532**	**0.255**	0.363	**0.424**	0.877

Impact of Tri-Planar Context. Medical diagnosis relies on correlating findings across anatomical planes. To validate our multi-view approach, we investigate the impact of incrementally adding anatomical views in Table 2. Relying solely on the Axial view yields suboptimal performance due to limited vertical continuity. However, the full tri-planar configuration achieves the highest accuracy across all metrics, showing significant step-wise improvements (paired t-test, $p < 0.01$). Qualitatively, while the Axial-only model missed vertically-spanning pathologies like a hiatal hernia, the Tri-Planar model identified them (e.g., "Sliding type hiatal hernia was observed at the lower end of the esophagus"), matching the ground truth. This monotonic improvement confirms that aggregating multi-view 2D slices is a prerequisite for approximating volumetric understanding, capturing complex structures that single-view inputs miss.

Impact of Instruction Strategy. Finally, we analyze the role of our Interleaved Tri-Planar Instruction strategy compared to a naive *Standard Concatenation* baseline. As shown in Table 2, while the baseline remains competitive on general similarity metrics with negligible variance, our Interleaved strategy delivers marked improvements where it matters most. Specifically, we observe a substantial 2.2% gain in BLEU-1 alongside improved METEOR scores, confirming that explicitly labeling views provides the necessary spatial cues to boost diagnostic precision without compromising semantic coherence.

4 Conclusion and Future Work

We introduced **HiFi-Rep**, a high-resolution framework that establishes a new state-of-the-art on CT-RATE by effectively bridging 2D vision features with vol-

umetric context. Our experiments validate that frozen DINOv3 representations generalize robustly to medical tasks, eliminating the need for computationally prohibitive 3D encoders. Crucially, we demonstrated that the sequence length bottleneck is resolved via our parameter-free HiFi-S2D aggregation, which acts as a spatial-preserving module by balancing depth condensation with multi-view context. Furthermore, we confirmed that precise 3D alignment is achievable through interleaved natural language cues, avoiding explicit 3D positional encoding. While current evaluations rely on standard NLP benchmarks for direct baseline comparability, future validation will incorporate clinical efficacy metrics and expert reader studies to assess true diagnostic validity. Subsequent extensions of this framework will explore multi-choice VQA, interactive medical conversations, and the integration of emerging high-resolution medical-specific encoders. Finally, incorporating reinforcement learning promises to foster reasoning-driven models, ultimately enhancing diagnostic transparency and reliability.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. An, U., Jeong, M., Lee, S.A., Gorla, A., Yang, Y., Sankararaman, S.: Raptor: Scalable train-free embeddings for 3d medical volumes leveraging pretrained 2d foundation models. In: Forty-second International Conference on Machine Learning (2025)
2. 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)
3. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. In: The Twelfth International Conference on Learning Representations (2024)
4. Di Piazza, T., Lazarus, C., Nempont, O., Boussel, L.: Ct-agrg: Automated abnormality-guided report generation from 3d chest ct volumes. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI). pp. 01–05 (2025)
5. Gugger, S., Debut, L., Wolf, T., Schmid, P., Mueller, Z., Mangrulkar, S., Sun, M., Bossan, B.: Accelerate: Training and inference at scale made simple, efficient and adaptable. `https://github.com/huggingface/accelerate` (2022)
6. Hamamci, I.E., Er, S., Menze, B.: CT2Rep: Automated Radiology Report Generation for 3D Medical Imaging . In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. vol. LNCS 15012. Springer Nature Switzerland (2024)
7. Hamamci, I.E., Er, S., Wang, C., Almas, F., Simsek, A.G., Esirgun, S.N., Dogan, I., Durugol, O.F., Hou, B., Shit, S., et al.: Developing generalist foundation models from a multimodal dataset for 3d computed tomography. arXiv preprint arXiv:2403.17834 (2024)
8. Hosseini, A., Ibrahim, A., Serag, A.: From Slices to Volumes: Multi-Scale Fusion of 2D and 3D Features for CT Scan Report Generation . In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. vol. LNCS 15965. Springer Nature Switzerland (2025)

9. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
10. Lai, H., Jiang, Z., Yao, Q., Wang, R., He, Z., Tao, X., Wei, W., Lv, W., Zhou, S.K.: E3d-gpt: enhanced 3d visual foundation for medical vision-language model. arXiv preprint arXiv:2410.14200 (2024)
11. Laurençon, H., Marafioti, A., Sanh, V., Tronchon, L.: Building and better understanding vision-language models: insights and future directions. In: Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models (2024)
12. Lee, C., Park, S., Shin, C.I., Choi, W.H., Park, H.J., Lee, J.E., Ye, J.C.: Read like a radiologist: Efficient vision-language model for 3d medical imaging interpretation. Medical Image Analysis **111**, 104077 (2026)
13. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
14. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
15. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
16. Mangrulkar, S., Gugger, S., Debut, L., Belkada, Y., Paul, S., Bossan, B., Tietz, M.: PEFT: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft` (2022)
17. Marafioti, A., Zohar, O., Farré, M., noyan, M., Bakouch, E., Jiménez, P.M.C., Zakka, C., allal, L.B., Lozhkov, A., Tazi, N., Srivastav, V., Lochner, J., Larcher, H., Morlon, M., Tunstall, L., Werra, L.V., Wolf, T.: SmolVLM: Redefining small and efficient multimodal models. In: Second Conference on Language Modeling (2025)
18. 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)
19. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
20. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
21. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1874–1883 (2016)
22. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
23. Sloan, P., Clatworthy, P., Simpson, E., Mirmehdi, M.: Automated radiology report generation: A review of recent advances. IEEE Reviews in Biomedical Engineering **18**, 368–387 (2025)
24. Wang, P., Lu, W., Lu, C., Zhou, R., Li, M., Qin, L.: Large language model for medical images: A survey of taxonomy, systematic review, and future trends. Big Data Mining and Analytics **8**(2), 496–517 (2025)

25. von Werra, L., Belkada, Y., Tunstall, L., Beeching, E., Thrush, T., Lambert, N., Huang, S., Rasul, K., Gallouédec, Q.: TRL: Transformers reinforcement learning. `https://github.com/huggingface/trl` (2020)
26. Wu, J., Wang, Y., Zhong, Z., Liao, W., Trayanova, N., Jiao, Z., Bai, H.X.: Vision-language foundation model for 3d medical imaging. npj Artificial Intelligence **1**(1), 17 (2025)
27. Zhang*, T., Kishore*, V., Wu*, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. In: International Conference on Learning Representations (2020)