

A Unified 3D Vision-Language Framework for Structured Multi-Study Volumetric Report Generation

Rahul Shukla^{1*}, Ruiming Wu^{2*}, Hongzhuo Chen³, Shu Yang², Duy Duong-Tran², Duy Minh Ho Nguyen⁴, Mathias Niepert⁴, Cameron Beeche², James Gee², Jeffrey Duda², Rakesh Sharma², Christos Davatzikos², Walter Witschey², Bojian Hou^{2†}, and Li Shen^{2†}

¹ School of Engineering and Applied Science, University of Pennsylvania

² Perelman School of Medicine, University of Pennsylvania

³ School of Arts and Sciences, University of Pennsylvania

⁴ International Max Planck Research School for Intelligent Systems (IMPRS-IS)

Abstract. While medical vision-language models (VLMs) show promise for automated radiology report generation, existing approaches typically rely on 2D encoders that discard essential volumetric data. Furthermore, they universally assume a single scan per patient, overlooking the multi-study imaging common in longitudinal clinical care. In this paper, we propose a unified native 3D VLM framework designed for structured multi-study CT and MRI report generation. Our dual-projector architecture bridges specialized 3D vision encoders with a clinically adapted language model through a contrastively grounded intermediate semantic space. To prevent the degradation of vision-text representations during fine-tuning, we introduce a joint generative-contrastive objective that maintains contrastive alignment across all training stages. Additionally, we implement and evaluate distinct aggregation strategies to effectively process patients with multiple historical imaging studies. Evaluated on the Penn Medicine BioBank (9,384 patients; 29,523 scans), our framework outperforms existing 2D and 3D medical VLMs across standard lexical metrics and multi-dimensional clinical assessments spanning anatomical coverage, critical findings, diagnostic plausibility, and report quality.

Keywords: 3D Medical Vision-Language Model · Radiology Report Generation · Contrastive Learning · Multi-study.

1 Introduction

Radiology report generation directly guides patient care, yet drafting reports is cognitively complex and high-stakes: clinicians must synthesize 3D anatomy, detect subtle abnormalities across slices, and translate visual evidence into standardized text. As imaging volumes outpace the available radiology workforce [22],

* Equal contribution. †Corresponding authors.

automated medical vision-language models (VLMs) offer a scalable way to reduce clinician burden. More broadly, clinical NLP studies confirm the value of such task-specific adaptation: MentalGPT [29] instruction-tunes LLMs for empathetic mental-health support, MentalChat16K [30] supplies domain-specific dialogues for training and evaluation, and DALK [11] grounds biomedical question answering in a literature-derived knowledge graph. None of these, however, address structured, image-grounded radiology reporting. Healthcare VLMs built for imaging, such as LLaVA-Med [10] and Med-Flamingo [16], rely mainly on 2D encoders such as CLIP [21], while treating volumes as independent slices, as in prior slice-based adaptations [4], discards spatial continuity essential for CT and MRI interpretation. Native 3D VLMs, including CT2Rep [6], RadFM [26], E3D-GPT [9], and Med3DVLM [28], narrow this gap, but key limitations remain. First, they pair spatial encoders directly with general-purpose language models (e.g., Vicuna [5], LLaMA [23]) that lack clinical grounding. Second, they confine contrastive alignment [32, 13] to pretraining and discard it during generative fine-tuning, allowing vision-text representations to degrade as the model fits language patterns—a phenomenon we term *clinical drift*. Finally, existing systems assume a single study and cannot aggregate the longitudinal or multi-phase scans common in clinical practice.

In this paper, we present a unified, native 3D vision-language framework for structured CT and MRI report generation that departs from prior approaches which retrofit 2D architectures to volumetric data. Instead, we (i) introduce a 3D dual-projector framework that couples state-of-the-art volumetric encoders (CT-FM [18] and Merlin [3]) with a domain-adapted clinical language model (MMed-Llama-3 [20]) within a clinically grounded ClinicalBERT [1] space. Building on this aligned representation, we (ii) formulate a joint Generative-InfoNCE objective [17] that sustains cross-modal consistency during instruction tuning. To move beyond the single-scan assumption, we (iii) extend our framework to the multi-study setting through three aggregation strategies: semantic retrieval, token stacking, and embedding averaging. Finally, to rigorously assess clinical reliability, we (iv) implement a multi-dimensional clinical LLM-as-a-Judge evaluation pipeline measuring anatomical coverage, critical findings, diagnostic plausibility, and overall report quality, thereby complementing traditional lexical metrics with clinically grounded fidelity assessment.

2 Method

Our framework consists of a frozen 3D vision encoder $\mathcal{E}$, a dual-projector alignment module, and a medical language model $\mathcal{L}$, trained in two stages following the visual instruction tuning paradigm [14]. An overview is shown in Fig. 1.

3D Vision Encoding and Dual-Projector Alignment. Given a CT volume $V \in \mathbb{R}^{D \times H \times W}$, we extract volumetric features $\mathbf{z}_v = \mathcal{E}(V)$ using a pretrained frozen encoder. We evaluate CT-FM [18], a SegResNet pretrained on 148k CTs via contrastive learning, and Merlin [3], an I3D ResNet pretrained on 15k abdominal CTs with joint contrastive and EHR supervision.

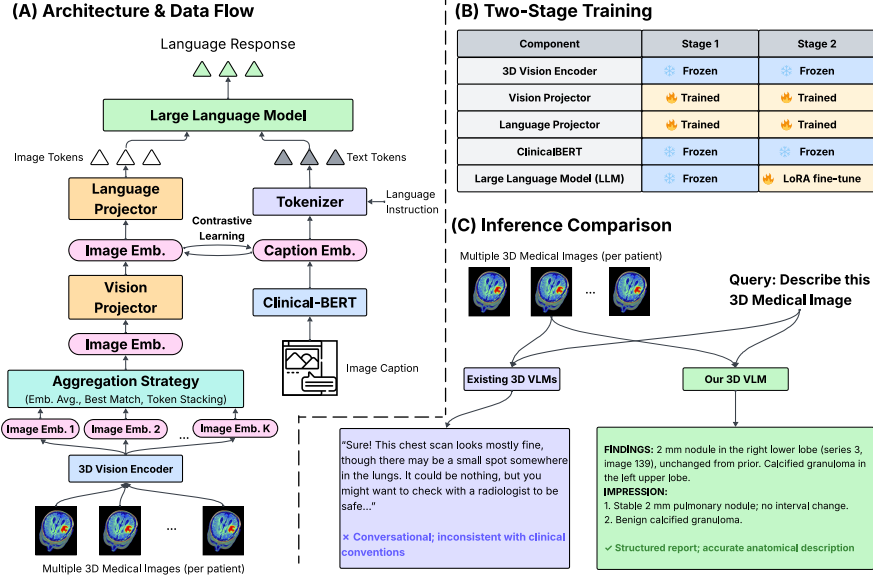


Fig. 1. Overview of our native 3D VLM framework. **(A)** Architecture: multiple 3D scans per patient are processed through one of three aggregation strategies, then routed through a dual-projector pipeline: Vision Projector $\mathcal{P}_v$ maps to ClinicalBERT space for InfoNCE alignment, Language Projector $\mathcal{P}_l$ maps to language model token space for generation. **(B)** Two-stage training: Stage 1 (frozen LM) jointly optimizes both projectors with generative-contrastive loss; Stage 2 applies LoRA fine-tuning with sustained contrastive regularization. **(C)** Inference comparison: existing 3D VLMs produce conversational outputs inconsistent with clinical conventions, while our model generates structured reports with accurate anatomical descriptions.

Projecting directly from a heterogeneous 3D feature space into language token space as done in frozen-LLM approaches like R2GenGPT [25] risks producing visually-grounded but clinically uninterpretable representations. We therefore introduce a dual-projector architecture that first anchors representations within an intermediate ClinicalBERT semantic space before generative decoding. Specifically, a vision projector $\mathcal{P}_v$ first maps $\mathbf{z}_v$ into the ClinicalBERT embedding space as:

$$\mathbf{z}_c = \mathcal{P}_v(\mathbf{z}_v) \in \mathbb{R}^{768}.$$

Aligning visual features to ClinicalBERT forces $\mathcal{P}_v$ to learn representations that are semantically meaningful in the clinical domain rather than arbitrary geometric mappings. A language projector $\mathcal{P}_l$ then bridges this clinical semantic space into the language model’s token space, $\mathbf{z}_l = \mathcal{P}_l(\mathbf{z}_c) \in \mathbb{R}^{d_c}$. We evaluate MMed-Llama-3-8B, pretrained on biomedical corpora, against Vicuna-7B [5] as a general-domain baseline, with both backbones fine-tuned under identical LoRA

configurations to isolate the effect of domain adaptation. Detailed architectural specifications for $\mathcal{P}_v$ and $\mathcal{P}_l$ are provided in Section 3.

Two-Stage Training with Sustained Contrastive Alignment. Both training stages share a joint Generative-InfoNCE objective. Given a batch of N image-text pairs, the InfoNCE loss [17] treats matched scan-report pairs as positives against all other pairings in the batch, enforcing visual grounding in the ClinicalBERT space:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\text{sim}(\mathbf{z}_{c,i}, \mathbf{t}_i)/\tau)}{\sum_{j=1}^N \exp(\text{sim}(\mathbf{z}_{c,i}, \mathbf{t}_j)/\tau)}$$

where $\mathbf{t}_i = \text{ClinicalBERT}(R_i)$ and τ is a temperature hyperparameter. The total training objective is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \lambda \cdot \mathcal{L}_{\text{InfoNCE}}$$

In stage 1, we set the contrastive weight to λ_1 and train both projectors from scratch with the encoder and language model frozen. This establishes the initial visual grounding in the clinical semantic space. In stage 2, we apply LoRA fine-tuning to the language model, continue training the projectors and set the contrastive weight to λ_2 to avoid clinical drift. Setting $\lambda_1 = \lambda_2 = 0$ recovers a purely generative baseline; we systematically vary these weights to characterize when sustained alignment benefits generation quality. Unlike prior contrastive VLMs [15] that apply alignment only during pretraining, both our stages retain a nonzero contrastive term, thereby actively maintaining visual grounding during generative adaptation. We stop after supervised instruction tuning and do not apply preference optimization, which can impair LLM calibration [27]; reliable uncertainty estimates are particularly important in clinical deployment. Calibration-aware bilevel optimization [8] and semantic-sampling evaluation for open-ended outputs [24] are complementary directions, but report-level calibration is outside our present scope.

Multi-Study Aggregation Strategies. For a patient with K studies yielding independent embeddings $\{\mathbf{z}_{v,1}, \dots, \mathbf{z}_{v,K}\}$, we evaluate three aggregation methods. (1) To mirror a radiologist synthesizing all available scans, we employ **Embedding Averaging**, which computes the mean $\bar{\mathbf{z}}_v = 1/K \sum_{k=1}^K \mathbf{z}_{v,k}$ of the features before projection. (2) To quantify the upper bound of single-study selection, we evaluate a **Semantic Retrieval** baseline, leveraging our Stage 1 alignment to select the most relevant study via cosine similarity in the clinical semantic space: $\mathbf{z}_v^* = \arg \max_k \cos(\mathcal{P}_v(\mathbf{z}_{v,k}), \text{ClinicalBERT}(R))$.¹ (3) To explicitly preserve temporal distinctions as a discrete baseline, we apply **Token Stacking**, concatenating projected embeddings $\mathbf{Z}_{\text{stack}} = [\mathcal{P}_l(\mathcal{P}_v(\mathbf{z}_{v,1})); \dots; \mathcal{P}_l(\mathcal{P}_v(\mathbf{z}_{v,K}))] \in \mathbb{R}^{K \times d_{\mathcal{L}}}$ into a single sequence, allowing the LLM to cross-attend across time points.

¹ Because R is unavailable at inference, this strictly serves as an analytical oracle benchmark. Our deployable inference strategies (averaging and stacking) operate entirely without text-guided retrieval, relying solely on zero-leakage visual features.

3 Experimental Setup

Dataset. We utilize the Penn Medicine BioBank (PMBB) of 9,384 patients, partitioned at the patient level to prevent leakage: 6,568 training, 1,407 validation, and 1,409 locked test samples. Each sample pairs one or more 3D volumes (20,873 CT, 8,650 MRI; median 3 scans/patient) with uncensored ‘Findings’ and ‘Impressions’ sections drawn directly from radiology reports, preserving real-world noise. The cohort spans chest (41.8%), abdomen (33.1%), and pelvis (15.8%) as the most prevalent anatomical regions.

Implementation Details. All image volumes are resampled to $1.5 \times 1.5 \times 3.0$ mm. To linearly map intensities to $[0, 1]$, CT volumes are clipped to encoder-specific Hounsfield ranges ($[-1000, 1000]$ for Merlin; $[-1024, 2048]$ for CT-FM), while MRI volumes use per-volume 1st–99th percentile clipping. After encoder-specific spatial center-cropping or zero-padding, both modalities are processed by the same frozen 3D encoders. Offline-extracted embeddings pass through vision ($\mathcal{P}_v$, 3-layer) and language ($\mathcal{P}_l$, 4-layer) MLP projectors with LayerNorm, GELU, and 0.1 dropout. **Stage 1** trains projectors for 1 epoch ($\text{LR}=2 \times 10^{-3}$, batch size 32). **Stage 2** applies rank-16 LoRA ($\alpha = 32$) to the LLM for 3 epochs (LoRA $\text{LR}=1 \times 10^{-6}$, projector $\text{LR}=5 \times 10^{-5}$). Training uses DeepSpeed ZeRO-2 in BF16 on two NVIDIA A40s. Inference uses nucleus sampling [7] ($T = 0.7$, $p = 0.9$, max 512 tokens).

Experimental Design & Baselines. Our experiments investigate: (1) Contrastive Regularization (titrating λ_1, λ_2); (2) Backbone Ablations (CT-FM vs. Merlin; MMed-Llama vs. Vicuna); and (3) Aggregation Strategies (semantic retrieval, token stacking, embedding averaging). We benchmark against fine-tuned 2D and 3D SOTA VLMs: LLaVA-Med (CLIP + Vicuna-7B) [10], RadFM (3D ViT + LLaMA-1 3B) [26], and Med3DVLM (DCFormer + Qwen-2.5-7B) [28]. LLaVA-Med is adapted for 3D input by extracting uniformly spaced 2D slices and concatenating their CLIP-encoded patch tokens as a flattened visual input.

Evaluation Metrics. We compute standard Lexical Metrics (BLEU-4 [19], METEOR [2], ROUGE-L [12]). Recognizing their poor correlation with true clinical correctness [31], we complement these with a custom LLM-as-a-Judge Protocol [33]. As detailed in the right panel of Fig. 2, GPT-4o scores generated reports (out of 10) across four dimensions calibrated for vision-only assessment: Anatomical Coverage (0–3), Critical Findings (0–3), Clinical Plausibility (0–2), and Report Quality (0–2).

4 Results and Discussion

Mitigating Clinical Drift (λ Tuning). We first establish the necessity of our joint Generative-InfoNCE objective. Table 1 tracks the impact of contrastive weights (λ_1, λ_2) on our default configuration (CT-FM + MMed-Llama with embedding averaging). The (0,0) configuration represents a purely generative baseline achieving a total clinical score of 4.78. Introducing symmetric regularization

Examples of a Generated Report with Caption Reference	Clinical LLM as a Judge: Evaluation Axes
Caption: The uterus is normal in size and appearance. No adenomyosis-related cysts are seen, and the endometrium appears unremarkable. There is a 3 × 2 mm cystic lesion in the right ovary, most consistent with a physiologic follicle. No suspicious adnexal mass or free fluid is identified. No suspicious osseous lesions are seen. Impression: Normal appearing ovaries. Report Generation: User: Describe this 3D medical image in detail. Assistant: Generated Report: The uterus appears normal in size and morphology, and the endometrium is unremarkable. A tiny 3 × 2 mm cystic focus is present in the right ovary, consistent with a normal follicle. No suspicious adnexal mass or free fluid is seen. No suspicious osseous lesions are identified. Impression: The ovaries appear normal.	1. ANATOMICAL COVERAGE (0–3 points): Focus on coverage of VISIBLE anatomical structures in the imaging study. 3 → Systematically covers all visible regions in exam. 2 → Covers most visible regions with minor gaps. 1 → Incomplete coverage of major visible regions. 0 → Wrong body region entirely or severely incomplete. 2. CRITICAL FINDINGS (0–3 points): Focus on VISUAL pathology observable in the image, including nodules, masses, calcifications, fluid, or fractures. 3 → Identifies all major VISIBLE pathology from reference. 2 → Identifies most visible pathology with minor omissions. 1 → Misses significant visible findings or contradicts reference. 0 → Misses critical acute visible findings or contains major contradictions. 3. CLINICAL PLAUSIBILITY (0–2 points): Assesses whether VISUAL findings described are medically coherent and anatomically possible. 2 → All visual findings are medically and anatomically possible. 1 → Mostly plausible with some questionable visual claims. 0 → Contains obvious hallucinations or impossible findings. 4. REPORT QUALITY (0–2 points): Assesses the structure and clarity of VISUAL findings description. 2 → Clear, well-organized visual findings description. 1 → Acceptable but suboptimal organization or detail. 0 → Poor structure, unclear descriptions, or unprofessional wording.

Fig. 2. Qualitative generation and evaluation protocol. **(Left)** Example generated report (CT-FM, $\lambda_1 = 0.7$, $\lambda_2 = 0.9$) capturing precise 3D-grounded anatomical details (e.g., a 3 × 2 mm cystic lesion) without hallucination. **(Right)** Multi-dimensional LLM-as-a-Judge scoring criteria across all four clinical axes.

(0.5, 0.5) improves the score to 5.44, confirming that sustained alignment prevents the language model from overfitting to statistical text priors. However, the left table highlights a highly non-linear optimization landscape. In lower-weight regimes, the model is sensitive to stage imbalance; delaying heavier regularization to Stage 2 (0.1, 0.5) yields a peak of 5.55, whereas its inverse (0.5, 0.1) scores 5.47. Crucially, the broader grid search reveals a ceiling effect: pushing constraints to their maximum (0.9, 0.9) degrades performance to 5.16, indicating that over-regularization overwhelms the cross-entropy objective and stifles generative fluency. The highest clinical fidelity is thus achieved not by maximizing constraints, but by finding a “sweet spot” balancing spatial grounding with linguistic decoding. Our empirical optimum ($\lambda_1 = 0.7$, $\lambda_2 = 0.9$) achieves the peak clinical score of 5.65; we fix these weights for all subsequent experiments.

Architecture and Aggregation. Table 2 isolates the architectural and multi-study choices. Replacing MMed-Llama with Vicuna (Group A) plummets the clinical score to 2.48, proving domain-adapted priors are essential. CT-FM outperforms Merlin (5.65 vs. 4.93), leveraging its 10× larger pretraining corpus and alignment-optimized embeddings. Group B highlights an aggregation-alignment interaction. Without contrastive regularization, token stacking (5.08) beats averaging (4.78) by preserving discrete, unaligned representations. Conversely, active contrastive alignment reverses this: averaging (5.65) surpasses stacking (5.39)

Table 1. Effect of contrastive regularization (λ_1, λ_2) on the validation set. **Left:** Clinical and lexical metrics for key configurations. **Right:** Grid search of Total Clinical Score across all evaluated weights, showing (0.7, 0.9) as optimal. (—: not evaluated)

Contr. Wt.		Clinical Quality (GPT-4o)					Lexical
λ_1	λ_2	Cov.	Crit.	Plau.	Qual.	Tot.	BLEU-4
0.0	0.0	1.45	0.73	1.65	0.93	4.78	0.100
0.1	0.5	1.90	0.98	1.59	1.08	5.55	0.183
0.5	0.1	1.92	0.97	1.55	1.02	5.47	0.180
0.5	0.5	1.90	0.96	1.55	1.04	5.44	0.177
0.7	0.9	1.93	1.09	1.56	1.07	5.65	0.181

(λ_1, λ_2)	0	0.1	0.3	0.5	0.7	0.9
0	4.78	—	—	—	—	—
0.1	5.35	5.28	5.50	5.55	5.40	5.34
0.3	5.30	5.23	5.13	5.26	5.25	5.32
0.5	5.32	5.47	5.40	5.44	5.39	5.39
0.7	5.17	5.55	5.52	5.29	5.45	5.65
0.9	—	5.05	—	5.27	—	5.16

Table 2. Ablation of architectural choices and multi-study aggregation strategies. Unless noted, models use CT-FM, MMed-Llama, embedding averaging, and optimal contrastive setting ($\lambda_1 = 0.7, \lambda_2 = 0.9$) on the **validation set**. Best results in **bold**.

		Clinical Quality (GPT-4o)					Lexical Metrics		
Ablation Axis	Configuration	Cov.	Crit.	Plau.	Qual.	Tot.	B-4	R-L	MTR
Group A: Vision & Language Backbones									
Language Backbone (Encoder = CT-FM fixed)	Vicuna-7B-Instruct	1.11	0.36	0.59	0.42	2.48	0.06	0.17	0.21
	MMed-Llama-3-8B	1.93	1.09	1.56	1.07	5.65	0.18	0.26	0.29
Vision Encoder (Decoder = MMed-Llama-3-8B fixed)	Merlin (1,692-dim)	1.79	0.89	1.34	0.92	4.93	0.16	0.24	0.27
	CT-FM (512-dim)	1.93	1.09	1.56	1.07	5.65	0.18	0.26	0.29
Group B: Multi-Study Aggregation Strategies									
Without Contrastive ($\lambda = 0$)	Embedding Averaging	1.45	0.73	1.65	0.93	4.78	0.10	0.19	0.22
	Token Stacking	1.63	0.81	1.63	1.01	5.08	0.13	0.21	0.25
With Contrastive ($\lambda > 0$)	Semantic Retrieval	1.80	0.86	1.33	0.95	4.94	0.15	0.20	0.28
	Token Stacking	1.85	0.93	1.56	1.04	5.39	0.14	0.23	0.27
	Embedding Averaging	1.93	1.09	1.56	1.07	5.65	0.18	0.26	0.29

and retrieval (4.94). This proves our objective structures the latent space such that averaging seamlessly preserves longitudinal context, whereas retrieval discards relevant studies and stacking dilutes attention with redundant tokens.

Comparison to State-of-the-Art. Table 3 benchmarks our framework on the held-out test set. We achieve a Total Clinical Score of 5.49 (reflecting expected variance from our 5.65 validation). Compared to the strongest 2D baseline, LLaVA-Med (4.75), our +0.74 outperformance is driven by superior 3D grounding, yielding massive gains in Anatomical Coverage (+0.37) and Critical Findings (+0.16). Notably, LLaVA-Med achieves marginally higher Plausibility (1.67 vs. 1.62) despite trailing substantially on lexical overlap (BLEU-4 0.08 vs. our 0.18); we revisit this divergence below. Furthermore, we drastically outperform native 3D baselines Med3DVLM (3.09) and RadFM (4.19). Their pure generative fine-tuning induces clinical drift and semantic collapse, crippling diagnostic and lexical accuracy. Conversely, our joint Generative-InfoNCE objective sustains contrastive alignment to fully leverage 3D features, averting this degradation and achieving state-of-the-art clinical and NLP performance.

Table 3. Comparison with state-of-the-art medical VLMs on the **test set**. Clinical metrics (GPT-4o): Cov. = Coverage (0–3), Crit. = Critical Findings (0–3), Plau. = Plausibility (0–2), Qual. = Quality (0–2), Tot. = Total (0–10). Lexical: B-4 = BLEU-4, R-L = ROUGE-L, MTR = METEOR. Best results in **bold**.

Model	Clinical Quality (GPT-4o)					Lexical Metrics		
	Cov.	Crit.	Plau.	Qual.	Tot.	B-4	R-L	MTR
LLaVA-Med (fine-tuned)	1.49	0.79	1.67	0.80	4.75	0.08	0.18	0.19
Med3DVLM (fine-tuned)	1.30	0.47	0.85	0.47	3.09	0.04	0.13	0.18
RadFM (fine-tuned)	1.44	0.68	1.28	0.80	4.19	0.02	0.15	0.13
Ours (CT-FM + MMed-Llama)	1.86	0.95	1.62	1.06	5.49	0.18	0.28	0.26

Hallucination and Dataset Bias. The Plausibility/BLEU-4 divergence noted above is consistent with LLaVA-Med relying on fluent language priors rather than grounded visual evidence (Fig. 2), since high plausibility without lexical fidelity indicates agreement with general clinical phrasing rather than the specific scan. This is an indirect signal, not a finding-level audit; we treat manual hallucination annotation as necessary future work rather than claiming it here. Separately, our uncured, multi-modality cohort avoids disease-specific filtering but has not been formally tested for diagnostic or demographic skew, which we flag as an open validation question.

5 Conclusion

In this paper, we presented a unified native 3D vision-language framework for structured multi-study CT and MRI report generation. By introducing a ClinicalBERT grounded dual-projector architecture and a joint Generative-InfoNCE objective, our framework sustains 3D spatial grounding during generative fine-tuning to mitigate clinical drift. To process patients with multiple imaging studies, we evaluated three multi-study aggregation strategies, finding parameter-free embedding averaging to be optimal. Our approach achieves the highest overall clinical and lexical scores among evaluated 2D and 3D baselines. Future work will pursue radiologist-led hallucination auditing, extend to additional modalities (e.g., PET), and integrate broader EHR context with explicit temporal embeddings to explicitly model longitudinal workflows.

Acknowledgments. This work was supported in part by seed funding from the PSOM AI2D Center at Penn and by the National Institutes of Health (NIH) under grants 2T32EB020087, 5R01HL169378, 1R21EB036734, and 1OT2OD038048.

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

References

1. Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jindi, D., Naumann, T., McDermott, M.B.A.: Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. pp. 72–78 (2019)
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. Blankemeier, L., Cohen, J.P., Kumar, A., Van Veen, D., Gardezi, S.J.S., Parekh, M., Sber, M., Langlotz, C., Chaudhari, A.: Merlin: A vision language foundation model for 3D computed tomography. arXiv preprint arXiv:2406.06512 (2024)
4. Chen, H., Shukla, R., Wu, R., Yang, S., Duong-Tran, D., Nguyen, D.M.H., Niepert, M., Beeche, C., Gee, J., Duda, J., Sharma, R., Davatzikos, C., Witschey, W., Hou, B., Shen, L.: Adapting vision-language models for 3D CT/MRI understanding on PMBB via slice selection and explanation analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops (VADH). pp. 2273–2282 (2025)
5. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. LMSYS Blog (2023)
6. 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)
7. Holtzman, A., Buys, J., Du, L., Forbes, M., Choi, Y.: The curious case of neural text degeneration. In: International Conference on Learning Representations (2020)
8. Jin, R., Wang, Z., Dai, Z., Xiao, J., Hou, B.: Beyond post-hoc temperature scaling: Bilevel optimization for LLM calibration. In: Conference on Language Modeling (2026)
9. 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)
10. 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. In: Advances in Neural Information Processing Systems. vol. 36 (2023)
11. Li, D., Yang, S., Tan, Z., Baik, J.Y., Yun, S., Lee, J., Chacko, A., Hou, B., et al.: DALK: Dynamic co-augmentation of LLMs and KG to answer alzheimer’s disease questions with scientific literature. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 2187–2205 (2024)
12. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out (2004)
13. Lin, W., Zhao, Z., Zhang, X., Wu, C., Zhang, Y., Wang, Y., Xie, W.: PMC-CLIP: Contrastive language-image pre-training using biomedical documents. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 525–536. Springer (2023)
14. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems. vol. 36 (2023)
15. Liu, L., Sun, X., Xiang, T., Zhuang, Z., Yin, L., Tan, M.: Contrastive vision-language alignment makes efficient instruction learner. arXiv preprint arXiv:2311.17945 (2023)

16. Moor, M., Huang, Q., Wu, S., Yasunaga, M., Zakka, C., Dalmia, Y., Reis, E.P., Rajpurkar, P., Leskovec, J.: Med-flamingo: a multimodal medical few-shot learner. arXiv preprint arXiv:2307.15189 (2023)
17. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
18. Pai, S., Hadzic, I., Bontempi, D., Bressem, K., Kann, B.H., Fedorov, A., Mak, R.H., Aerts, H.J.W.L.: Vision foundation models for computed tomography. arXiv preprint arXiv:2501.09001 (2025)
19. 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)
20. Qiu, P., Wu, C., Zhang, X., Lin, W., Wang, H., Zhang, Y., Wang, Y., Xie, W.: Towards building multilingual language model for medicine. *Nature Communications* **15**, 8384 (2024)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
22. Rimmer, A.: Radiologist shortage leaves patient care at risk, warns royal college. *BMJ* **359**, j4683 (2017)
23. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
24. Wang, Z., Xiao, J., Jin, R., Yang, S., Hou, B., et al.: A semantic-sampling framework for evaluating calibration in open-ended question answering. arXiv preprint arXiv:2605.08432 (2026)
25. Wang, Z., Liu, L., Wang, L., Zhou, L.: R2GenGPT: Radiology report generation with frozen LLMs. *Meta-Radiology* **1**(3), 100033 (2023)
26. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. *Nature Communications* **16**, 7866 (2025)
27. Xiao, J., Hou, B., et al.: Restoring calibration for aligned large language models: A calibration-aware fine-tuning approach. arXiv preprint arXiv:2505.01997 (2025)
28. Xin, Y., Ates, G.C., Gong, K., Shao, W.: Med3DVLM: An efficient vision-language model for 3D medical image analysis. *IEEE Journal of Biomedical and Health Informatics* (2025)
29. Xu, J., Wei, T., Hou, B., Orzechowski, P., Yang, S., Demiris, G., Shen, L.: Mental-GPT: Harnessing AI for compassionate mental health support. In: AMIA Annual Symposium Proceedings (2024)
30. Xu, J., Wei, T., Hou, B., et al.: MentalChat16K: A benchmark dataset for conversational mental health assistance. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 5367–5378 (2025)
31. Yu, F., Endo, M., Krishnan, R., et al.: Evaluating progress in automatic chest X-ray radiology report generation. *Patterns* **4**(9), 100802 (2023)
32. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine Learning for Healthcare Conference. pp. 2–25. PMLR (2022)
33. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging LLM-as-a-judge with MT-Bench and chatbot arena. *Advances in Neural Information Processing Systems* **36** (2023)