U-VLM: Hierarchical Vision Language Modeling for Report Generation

Pengcheng Shi[1], Minghui Zhang[3,2], Kehan Song[1], Jiaqi Liu[1], Yun Gu(✉)[2], and Xinglin Zhang(✉)[1]

[1] Medical Image Insights Co. Ltd., Shanghai, China
[2] Shanghai Jiao Tong University, Shanghai, China
[3] Shanghai Artificial Intelligence Laboratory, Shanghai, China
geron762@sjtu.edu.cn, xinglin.zhang@imagecore.com.cn

Abstract. Automated radiology report generation is key for reducing radiologist workload and improving diagnostic consistency, yet generating accurate reports for 3D medical imaging remains challenging. Existing vision-language models face two limitations: they do not leverage segmentation-pretrained encoders, and they inject visual features only at the input layer of language models, losing multi-scale information. We propose U-VLM, which enables hierarchical vision-language modeling in both training and architecture: (1) progressive training from segmentation to classification to report generation, and (2) multi-layer visual injection that routes U-Net encoder features to corresponding language model layers. Each training stage can leverage different datasets without unified annotations. U-VLM achieves state-of-the-art performance on CT-RATE (F1: 0.414 vs 0.258, BLEU-mean: 0.349 vs 0.305) and AbdomenAtlas 3.0 (F1: 0.624 vs 0.518 for segmentation-based detection) using only a 0.1B decoder trained from scratch, demonstrating that well-designed vision encoder pretraining outweighs the benefits of 7B+ pre-trained language models. Ablation studies show that progressive pretraining significantly improves F1, while multi-layer injection improves BLEU-mean. Code is available at https://github.com/yinghemedical/U-VLM.

Keywords: Vision-Language Model · U-Net · Progressive Training · Report Generation

1 Introduction

Automated radiology report generation for 3D medical imaging is key for reducing radiologist workload and improving diagnostic consistency. However, generating accurate reports requires multi-scale visual understanding: global context for anatomical regions, and fine-grained details for lesion detection. Existing 3D medical VLMs [?,?,?,?,?,?,?,?] inject visual features only at the input layer of language models, losing multi-scale information during generation. Furthermore, no prior end-to-end VLM leverages dense per-voxel supervision from segmentation (Table ??)—despite SuPreM [?] showing segmentation pretraining transfers more effectively than self-supervised approaches, and RadGPT [?] demonstrating

Table 1: Comparison with related work on 3D radiology report generation. Existing methods inject visual features only at the language model input layer, and none leverage segmentation-pretrained U-Net for end-to-end report generation. VE: Vision Encoder. LD: Language Decoder. E2E: End-to-End. VE Init: CL (Contrastive Learning, CLIP [?]-style), VQ Recon (vector-quantized reconstruction), Cls (classification), Seg (segmentation). Perc.: Perceiver. AP: Attention Pooling.

Method	**VE**	**LD**	**VE Init**	**LD Init**	**E2E**
CT2Rep [?]	Causal Trans.	Trans. Dec.	Scratch	Scratch	✓
RadFM [?]	ViT+Perc.	MedLLaMA-13B	Scratch	Fine-tune	✓
M3D-LaMed [?]	ViT+Perc.	LLaMA-2-7B	CL	LoRA	✓
Merlin [?]	3D ResNet	RadLlama-7B	CL+Cls	LoRA	✓
CT-CHAT [?]	ViT+AP	LLaMA-3.1-70B	CL	LoRA	✓
MedM-VL [?]	ViT+AP	Qwen2.5-3B	CL	Fine-tune	✓
BTB3D [?]	Causal Conv.	LLaMA-3.1-8B	VQ Recon	LoRA	✓
RadGPT [?]	U-Net Enc.	–	Seg	–	✗
U-VLM (Ours)	U-Net Enc.	0.1B	Seg+Cls	Scratch	✓

segmentation-based methods outperform end-to-end VLMs in lesion detection (though it bypasses end-to-end learning via deterministic rules).

U-Net [?] dominates segmentation because its hierarchical encoder and skip connections preserve multi-scale information, with nnU-Net [?,?] providing a robust learning foundation. Existing 3D medical VLMs (Table ??) lack this multi-scale hierarchy and inject visual features only at the language model input layer. ConvLLaVA [?] shows that hierarchical backbones produce over 4× fewer visual tokens than ViT at the same resolution while preserving rich visual information, motivating our hierarchical U-Net encoder; DeepStack [?] shows that distributing visual tokens across multiple language model layers—rather than concentrating them at the input—enables models to handle 4× more visual tokens, motivating our multi-layer injection. We extend these insights to 3D medical imaging.

We propose U-VLM, a vision-language framework that enables hierarchical modeling in both training and architecture (Fig. ??): progressive training from segmentation to classification to report generation first captures fine-grained spatial structures, then disease patterns, and finally supports report generation; multi-layer visual injection routes U-Net encoder features to corresponding language model layers, extending skip connections [?] to vision-language modeling and preserving multi-scale information throughout generation.

To validate U-VLM, we train on CT-RATE [?] (25,692 chest CTs) and AbdomenAtlas 3.0 [?] (9,262 abdominal CTs). U-VLM achieves F1 of 0.414 and BLEU-mean of 0.349 on CT-RATE, surpassing BTB3D (F1: 0.258, BLEU-mean: 0.305), and outperforms both end-to-end methods and segmentation-based detection on AbdomenAtlas (F1: 0.624 vs 0.518). U-VLM uses only a 0.1B decoder

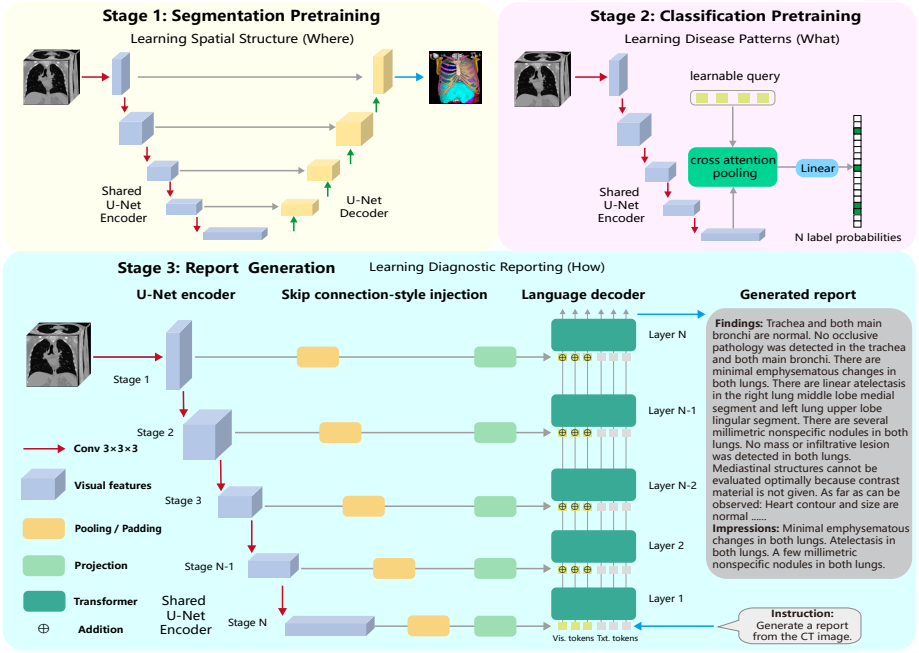

Fig. 1: U-VLM framework. **Stage 1:** Segmentation pretraining for learning fine-grained spatial structures. **Stage 2:** Classification pretraining for disease pattern recognition. **Stage 3:** Report generation via multi-layer injection (deep encoder → early language layers, shallow encoder → later layers).

trained from scratch, while compared methods use 7B+ pre-trained models. Each training stage leverages different datasets without unified annotations. Our contributions (backed by Tables **??**–**??**) are: (1) a progressive pretraining pipeline from segmentation to classification to report generation; (2) multi-layer visual injection routing U-Net features to corresponding language model layers; (3) demonstrating that vision encoder pretraining outweighs the benefits of 7B+ pre-trained language models.

2 U-VLM

U-VLM trains a shared U-Net encoder through three progressive stages, then connects it to a language decoder via multi-layer visual injection. Fig. **??** shows the framework overview.

2.1 Progressive Training

The shared U-Net encoder is sequentially optimized through three stages following curriculum learning [**?**]: Stage 1 learns spatial localization ("where") from

segmentation annotations, Stage 2 recognizes disease patterns ("what") from classification labels, and Stage 3 generates reports ("how") from image-report pairs. Each stage can leverage different datasets without unified annotations.

Stage 1: Segmentation Pretraining. The full U-Net learns fine-grained spatial structures through dense per-voxel supervision. We explore different segmentation granularities: (i) coarse anatomy only, (ii) coarse anatomy + lesions, and (iii) fine-grained anatomy + lesions, to analyze the impact on downstream report generation (Sec. **??**). Given input volume $\mathbf{X} \in \mathbb{R}^{D\times H\times W}$ and ground-truth masks $\mathbf{Y}_{seg}$, the U-Net produces multi-scale encoder features $\{\mathbf{f}_i\}_{i=1}^{N}$ and decoder output $\hat{\mathbf{Y}}_{seg}$:

$$\mathcal{L}_{seg} = \mathcal{L}_{dice}(\mathbf{Y}_{seg}, \hat{\mathbf{Y}}_{seg}) + \mathcal{L}_{ce}(\mathbf{Y}_{seg}, \hat{\mathbf{Y}}_{seg}) \tag{1}$$

Stage 2: Classification Pretraining. The decoder is replaced with a classification head that uses learnable query vectors $\mathbf{Q} \in \mathbb{R}^{M\times D}$ to aggregate encoder features via cross-attention. Given multi-label disease annotations $\mathbf{y} \in \{0,1\}^C$:

$$\hat{\mathbf{y}} = \sigma(\text{Linear}(\text{Flatten}(\text{CrossAttn}(\mathbf{Q}, \mathbf{f}_N)))), \quad \mathcal{L}_{cls} = \text{BCE}(\mathbf{y}, \hat{\mathbf{y}}) \tag{2}$$

Stage 3: Report Generation. The pretrained encoder connects to a language decoder via multi-layer visual injection (Sec. **??**). The language modeling loss is:

$$\mathcal{L}_{gen} = -\sum_{t=1}^{T} \log P(w_t | w_{<t}, \{\mathbf{h}_j\}_{j=1}^{L}) \tag{3}$$

where $\mathbf{h}_j$ denotes the hidden state at language layer j, enriched by injected visual features at vision token positions.

2.2 Multi-Layer Visual Injection

Standard VLMs typically inject visual features only at the input layer, causing multi-scale spatial details to vanish in deeper language layers. Following U-Net skip connections [**?**] and DeepStack [**?**], we inject features from each encoder stage S_i into specific language model layers L_j.

Feature Alignment. To enable injection across layers, features from different stages must share a consistent token sequence length K. We select a reference stage S_r (where $1 \le r \le N$, typically $r = N-1$) to define K. For any stage S_i with feature $\mathbf{f}_i$:

$$\tilde{\mathbf{f}}_i = \text{Align}_r(\mathbf{f}_i) = \begin{cases} \text{Pool}(\mathbf{f}_i) & \text{if } i < r \\ \mathbf{f}_i & \text{if } i = r \\ \text{Pad}(\mathbf{f}_i) & \text{if } i > r \end{cases} \tag{4}$$

Shallower stages ($i < r$) are downsampled via adaptive pooling, while deeper stages ($i > r$) are padded with zeros to match length K. A projection layer $\text{Proj}(\cdot)$ maps $\tilde{\mathbf{f}}_i$ to the language hidden dimension D.

Skip Connection-Style Injection. Analogous to U-Net's skip connections, we inject multi-scale features across language layers: deep encoder stages (global

semantics) feed early language layers, while shallow stages (fine-grained details) feed later layers. Following DeepSeek-OCR 2 [?], we adopt a hybrid attention mask where vision tokens attend bidirectionally while text tokens use causal attention. The hidden states at layer j are updated as:

$$\mathbf{h}_j^{(v)} = \mathrm{LM}_j\left(\mathbf{h}_{j-1}^{(v)} + \mathrm{Proj}_j\left(\mathrm{Align}_r(\mathbf{f}_{N-j+1})\right)\right), \quad \mathbf{h}_j^{(t)} = \mathrm{LM}_j\left(\mathbf{h}_{j-1}^{(t)}\right) \tag{5}$$

where visual features are injected only at vision token positions; text tokens pass through without injection. The mapping pairs encoder stage $N-j+1$ with language layer j (e.g., $\mathbf{f}_N \to L_1$). The decoder has N layers—one per encoder stage—yielding a one-to-one mapping ($\mathbf{f}_{N-j+1} \to L_j$ for $j = 1, \ldots, N$), so every language layer receives visual features from a distinct stage.

3 Experiments

3.1 Datasets and Setup

We evaluate U-VLM on two 3D CT datasets with no data leakage between training stages. (I) **CT-RATE** [?]: 25,692 chest CT volumes with reports and 18-class multi-label abnormality classification. For Stage 1, we use 2,628 cases from ReXGroundingCT [?] (a subset of CT-RATE training set) covering 14 lesion categories, augmented with pseudo-labels for anatomical structures from TotalSegmentator [?] (33 classes), ATM22 [?] (airway), and HiPaS [?] (pulmonary vessels). Stages 2–3 use CT-RATE's native labels. (II) **AbdomenAtlas 3.0** [?]: 9,262 abdominal CT volumes with per-voxel lesion annotations, 38 fine-grained anatomy classes, structured reports, and 3-class lesion classification (liver, kidney, pancreas). All three stages use only AbdomenAtlas data (same as RadGPT), with no external datasets. Unlike RadGPT which reports tumor detection, we evaluate lesion detection since the publicly available AbdomenAtlas 3.0 does not include malignancy information.

Implementation. We implement U-VLM based on nnU-Net [?,?]. The U-Net encoder (ResEncoder) has 6 stages with features [32, 64, 128, 256, 320, 320]. Input volumes are resized to $256 \times 256 \times 192$; segmentation uses patches of $128 \times 128 \times 96$. The lightweight language decoder (0.1B parameters, 6 layers, 512 hidden dim, 8 heads) is trained from scratch, as task-specific lightweight models match adapted large models [?]. For comparison, we also evaluate Qwen3-4B [?] with both LoRA [?] (rank 64, α=128) and full parameter fine-tuning. All experiments use batch size 2 on a single NVIDIA A100 80GB GPU. Training uses SGD with learning rate 0.01 for segmentation (Stage 1), and AdamW with base learning rate 2e-5 for classification and report generation (Stages 2–3), where the encoder uses 0.1× multiplier when unfrozen or 0 when frozen. Metrics include BLEU-mean and macro-averaged F1, Precision, and Recall following CT-CHAT [?] and BTB3D [?]. Report-level clinical labels are extracted by task-specific evaluators: for CT-RATE, the official RadBERT text classifier (threshold 0.5) extracts the 18 abnormality labels from generated reports, applied identically to predictions and ground-truth reports; for AbdomenAtlas, a DeepSeek LLM fills a structured

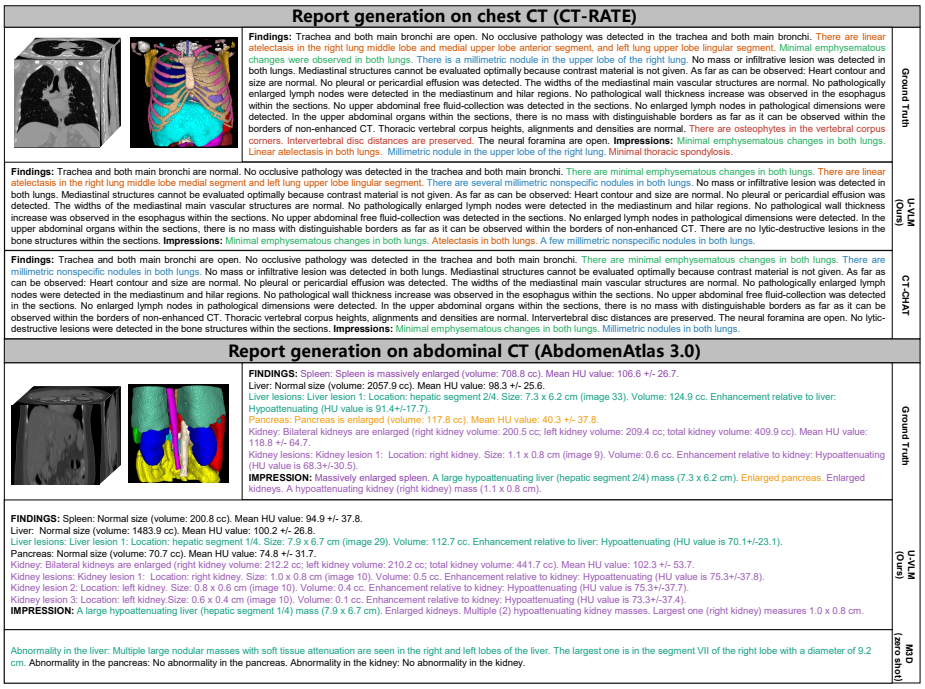

Fig. 2: Qualitative results of segmentation and report generation on chest CT (CT-RATE) and abdominal CT (AbdomenAtlas 3.0). We visualize the input 3D CT volumes alongside segmentation predictions: Seg(F+L) for chest CT and Seg(C+L) for abdominal CT. For all reports, text is color-coded to highlight abnormalities, maintaining consistent colors for the same pathology, while normal descriptions are shown in black.

template (yes/no/uncertain for each organ) to extract liver, kidney and pancreas lesion labels, following the RadGPT evaluation protocol.

3.2 Results

Main Results. Table **??** compares U-VLM against existing methods on CT-RATE. U-VLM achieves F1 of 0.414, surpassing BTB3D-16 (0.258) by 60% relative improvement, and BLEU-mean improves from 0.305 to 0.349—using only a 0.1B decoder trained from scratch, while compared methods use 7B+ pretrained models. Table **??** evaluates lesion detection on AbdomenAtlas 3.0. U-VLM with Seg(C+L) achieves best F1 across all organs (59.5% pancreas, 64.8% kidney, 62.9% liver), outperforming both end-to-end methods (M3D, RadFM) and segmentation-based detection following RadGPT protocol [**?**] (nnU-Net). Fig. **??** shows qualitative results on both datasets.

Ablation Studies. Tables **??**–**??** analyze key design choices. *Notations:* None: random initialization (no pretraining); Seg(C): coarse anatomy; Seg(C+L):

Table 2: Report generation on CT-RATE.

Method	F1	Prec.	Rec.	B-1	B-2	B-3	B-4	B-mean
CT2Rep [?]	0.160	0.435	0.128	0.372	0.292	0.243	0.213	0.280
Merlin [?]	0.160	0.295	0.112	0.231	0.163	0.124	0.099	0.154
CT-CHAT [?]	0.184	0.450	0.158	0.373	0.284	0.231	0.198	0.272
MedM-VL [?]	0.167	0.323	0.160	0.390	0.295	0.240	0.205	0.283
BTB3D-8 [?]	0.187	0.260	0.150	0.411	0.307	0.245	0.215	0.295
BTB3D-16 [?]	0.258	0.260	0.260	0.420	0.318	0.256	0.225	0.305
U-VLM (Ours)	**0.414**	0.491	0.429	**0.474**	**0.367**	**0.300**	**0.256**	**0.349**

Table 3: Lesion detection on AbdomenAtlas 3.0.

Model	Pancreatic Lesion			Kidney Lesion			Liver Lesion		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
End-to-End Report Generation									
M3D (zero-shot) [?]	11.8	3.6	5.6	23.5	9.2	13.2	16.5	10.7	13.0
RadFM (zero-shot) [?]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
U-VLM Seg(C+L)	61.0	58.2	**59.5**	54.8	79.4	**64.8**	58.1	68.7	**62.9**
U-VLM Seg(F+L)	51.2	39.1	44.3	59.8	70.2	64.6	45.8	75.6	57.1
Segmentation-Based Detection (RadGPT-style)									
nnU-Net Seg(C+L) [?]	69.1	50.9	58.6	34.2	90.8	49.7	32.3	87.8	47.2
nnU-Net Seg(F+L) [?]	86.7	35.5	50.3	49.1	74.8	59.3	49.2	48.1	48.6

coarse anatomy + lesions; Seg(F+L): fine-grained anatomy + lesions. B-M: BLEU-mean. SC: Skip Connections; Frz: Frozen Encoder; VT: Visual Tokens; Q4B-L/F: Qwen3-4B with LoRA/full fine-tuning. Classification metrics are extracted from generated reports by the evaluators described in Sec. **??**. *(1) Progressive training validates dense supervision.* The full Seg→Cls→Rep pipeline achieves F1=0.415 on CT-RATE, while skipping segmentation (Cls→Rep: 0.292) or classification (Seg→Rep: 0.238) degrades performance significantly. Table **??** shows segmentation pretraining improves classification by +23% (CT-RATE) and +73% (AbdomenAtlas), consistent with SuPreM [**?**]. This confirms that dense per-voxel supervision from segmentation enables learning finer-grained spatial structures than training from scratch. *(2) Multi-layer injection preserves multi-scale information.* Skip connection-style injection improves B-M from 0.303 to 0.349 on CT-RATE and from 0.413 to 0.437 on AbdomenAtlas, while maintaining F1 (0.415→0.414 and 0.620→0.624), enhancing report fluency without sacrificing diagnostic accuracy. This supports our hypothesis that routing hierarchical features to corresponding language layers prevents information loss during generation. *(3) Segmentation granularity is task-dependent.* Optimal granularity varies by dataset: Seg(F+L) achieves best F1 on CT-RATE (0.415) where fine-grained localization aids 18-class classification, while Seg(C+L) yields better lesion detection on AbdomenAtlas (F1=0.624 vs 0.553) where fine-grained

Table 4: Classification ablation on CT-RATE and AbdomenAtlas 3.0.

Training	CT-RATE			AbdomenAtlas 3.0		
	F1	**Precision**	**Recall**	**F1**	**Precision**	**Recall**
None → Cls	0.451	0.541	0.404	0.373	0.304	0.557
Seg(C) → Cls	0.546	0.587	0.535	<u>0.617</u>	0.568	0.679
Seg(C+L) → Cls	<u>0.553</u>	0.595	0.541	**0.646**	0.595	0.716
Seg(F+L) → Cls	**0.555**	0.598	0.539	0.578	0.567	0.589

anatomy segmentation (e.g., liver segments, pancreas subdivisions) may distract from 3-class lesion classification. *(4) Encoder freezing preserves learned representations.* Frozen encoder (F1=0.415) outperforms fine-tuning (0.362), likely because freezing prevents catastrophic forgetting of discriminative features learned during classification pretraining. *(5) Visual token count shows diminishing returns.* Increasing tokens (384→3072) yields no improvement, suggesting the bottleneck lies in other components rather than visual token quantity. *(6) Vision encoder pretraining outweighs decoder scale.* Our 0.1B decoder outperforms Qwen3-4B with both LoRA and full fine-tuning, consistent with UniRec [**?**]. On CT-RATE, 0.1B achieves F1=0.415 vs Qwen3-4B's 0.167 (LoRA) and 0.156 (full); on AbdomenAtlas, the gap narrows (0.624 vs 0.596). Pre-trained LLMs may have difficulty adapting to small medical datasets with domain-specific output distributions.

3.3 Limitations

Our method has several limitations. First, precise numerical statements in generated reports (e.g., lesion dimensions, volumes, and attenuation values) are produced by the language decoder and are not explicitly grounded in image measurements, so they may be hallucinated; as shown in Fig. **??**, the generated report also misses or alters clinically important findings (e.g., splenic and pancreatic enlargement, and the number/laterality of renal lesions), which are not fully captured by F1 and BLEU. Second, the segmentation annotations for Stage 1 on CT-RATE are generated by automated models (e.g., the lesion pseudo-labels from ReXGroundingCT), so the quality of encoder pretraining is bounded by the accuracy of these annotations. We leave clinically grounded numerical estimation to future work.

4 Conclusion

U-VLM enables hierarchical vision-language modeling in both training and architecture for 3D radiology report generation. By leveraging different datasets at each stage without requiring unified annotations, U-VLM achieves state-of-the-art performance on CT-RATE (F1: 0.414 vs 0.258, BLEU-mean: 0.349 vs 0.305,

Table 5: Report generation ablation on CT-RATE and AbdomenAtlas 3.0.

					CT-RATE		Abdomen	
Training	**SC**	**Frz**	**VT**	**Dec.**	**F1**	**B-M**	**F1**	**B-M**
None → Rep		✓	384	0.1B	0.062	0.251	0.264	0.336
Cls → Rep		✓	384	0.1B	0.292	0.294	0.384	0.391
Seg(C+L) → Rep		✓	384	0.1B	0.232	0.297	0.348	0.341
Seg(F+L) → Rep		✓	384	0.1B	0.238	0.278	0.398	0.392
Seg(C) → Cls → Rep		✓	384	0.1B	0.399	0.298	0.592	0.423
Seg(C+L) → Cls → Rep		✓	384	0.1B	0.409	0.279	<u>0.620</u>	0.413
Seg(F+L) → Cls → Rep		✓	384	0.1B	**0.415**	0.303	0.560	0.407
Seg(C+L) → Cls → Rep	✓	✓	384	0.1B	0.407	0.298	**0.624**	**0.437**
Seg(F+L) → Cls → Rep	✓	✓	384	0.1B	<u>0.414</u>	**0.349**	0.553	0.417
Seg(C+L) → Cls → Rep		✓	3072	0.1B	0.404	0.278	0.618	0.406
Seg(F+L) → Cls → Rep		✓	3072	0.1B	0.413	0.286	0.563	0.398
Seg(C+L) → Cls → Rep		✓	384	Q4B-L	0.268	0.275	0.596	0.382
Seg(C+L) → Cls → Rep		✓	384	Q4B-F	0.275	0.288	0.588	0.391
Seg(F+L) → Cls → Rep		✓	384	Q4B-L	0.167	0.281	0.525	0.404
Seg(F+L) → Cls → Rep		✓	384	Q4B-F	0.156	<u>0.304</u>	0.537	0.412
Seg(C+L) → Cls → Rep			384	0.1B	0.374	0.289	0.592	<u>0.431</u>
Seg(F+L) → Cls → Rep			384	0.1B	0.362	0.285	0.536	0.429

Table **??**) and AbdomenAtlas 3.0 (F1: 0.624 vs 0.518 for segmentation-based detection, Table **??**) with only a 0.1B decoder trained from scratch. Our ablation studies (Tables **??**–**??**) show that progressive pretraining significantly improves F1, while multi-layer injection improves BLEU-mean—demonstrating that both dense supervision from progressive pretraining and skip connection-style visual injection yield complementary benefits for 3D medical vision-language tasks. This flexibility in leveraging diverse annotation types enables data aggregation across institutions, suggesting potential for scalable unified medical AI without costly unified labeling.

Acknowledgments. This work is supported by supported by National Key R&D Program of China (2022ZD0212400), Science and Technology Commission of Shanghai Municipality (No.25511102602), Natural Science Foundation of China (62373243), and Shanghai Municipal Commission of Economy and Informatization (AI for Science Program: 2025-GZL-RGZN-BTBX-02023).

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

References

1. Baharoon, M., Luo, L., Moritz, M., Kumar, A., Kim, S.E., Zhang, X., Zhu, M., Alabbad, M.H., Alhazmi, M.S., Mistry, N.P., et al.: Rexgroundingct: A 3d chest ct dataset for segmentation of findings from free-text reports. NEJM AI **3**(7), AIdbp2501220 (2026) 5
2. 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) 1, 2, 7
3. Bassi, P.R., Yavuz, M.C., Hamamci, I.E., Er, S., Chen, X., Li, W., Menze, B., Decherchi, S., Cavalli, A., Wang, K., et al.: Radgpt: Constructing 3d image-text tumor datasets. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 23720–23730 (2025) 1, 2, 5, 6
4. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th annual international conference on machine learning. pp. 41–48 (2009) 3
5. Blankemeier, L., Kumar, A., Cohen, J.P., Liu, J., Liu, L., Van Veen, D., Gardezi, S.J.S., Yu, H., Paschali, M., Chen, Z., et al.: Merlin: a computed tomography vision–language foundation model and dataset. Nature pp. 1–11 (2026) 1, 2, 7
6. Chu, Y., Luo, G., Zhou, L., Cao, S., Ma, G., Meng, X., Zhou, J., Yang, C., Xie, D., Mu, D., et al.: Deep learning-driven pulmonary artery and vein segmentation reveals demography-associated vasculature anatomical differences. Nature Communications **16**(1), 2262 (2025) 5
7. Du, Y., Chen, Z., Xie, Y., Feng, W.B.H., Shi, W., Su, Y., Huang, C., Jiang, Y.G.: Unirec-0.1 b: Unified text and formula recognition with 0.1 b parameters. arXiv preprint arXiv:2512.21095 (2025) 5, 8
8. Ge, C., Cheng, S., Wang, Z., Yuan, J., Gao, Y., Song, J., Song, S., Huang, G., Zheng, B.: Convllava: Hierarchical backbones as visual encoder for large multimodal models. arXiv preprint arXiv:2405.15738 (2024) 2
9. 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) 1, 2, 7
10. Hamamci, I.E., Er, S., Shit, S., Reynaud, H., Yang, D., Guo, P., Edgar, M., Xu, D., Kainz, B., Menze, B.: Better tokens for better 3d: Advancing vision-language modeling in 3d medical imaging. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems 1, 2, 5, 7
11. 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.: Generalist foundation models from a multimodal dataset for 3d computed tomography. Nature Biomedical Engineering pp. 1–19 (2026) 1, 2, 5, 7
12. 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) 5
13. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021) 2, 5
14. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jaeger, P.F.: nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 488–498. Springer (2024) 2, 5, 7

15. Li, W., Yuille, A., Zhou, Z.: How well do supervised models transfer to 3d image segmentation. In: The Twelfth International Conference on Learning Representations. vol. 1 (2024) 1, 7
16. Meng, L., Yang, J., Tian, R., Dai, X., Wu, Z., Gao, J., Jiang, Y.G.: Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for lmms. Advances in Neural Information Processing Systems **37**, 23464–23487 (2024) 2, 4
17. 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) 2
18. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015) 2, 4
19. Shi, Y., Yang, S., Zhu, X., Wang, H., Fu, X., Li, M., Wu, J.: Medm-vl: What makes a good medical lvlm? In: International Workshop on Agentic AI for Medicine. pp. 290–299. Springer (2025) 1, 2, 7
20. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5), e230024 (2023) 5
21. Wei, H., Sun, Y., Li, Y.: Deepseek-ocr 2: Visual causal flow. arXiv preprint arXiv:2601.20552 (2026) 5
22. 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) 1, 2, 7
23. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025) 5
24. Zhang, M., Wu, Y., Zhang, H., Qin, Y., Zheng, H., Tang, W., Arnold, C., Pei, C., Yu, P., Nan, Y., et al.: Multi-site, multi-domain airway tree modeling. Medical image analysis **90**, 102957 (2023) 5