Just 32 Tokens: Medical Referring Image Segmentation via Next-Token Mask Prediction

Xinyu Chen, Gaoyang Pang, Zixuan Xu, Geng Wang, Haiyao Yu, Luping Zhou✉, and Yonghui Li✉

School of Electrical and Computer Engineering, University of Sydney, Sydney, Australia
luping.zhou@sydney.edu.au; yonghui.li@sydney.edu.au

Abstract. Medical Referring Image Segmentation (MRIS) aims to segment target regions in medical images based on natural language descriptions. While multimodal large language models (MLLMs) with next-token prediction offer strong cross-modal understanding, adapting them to MRIS remains challenging: fine-grained boundaries are difficult to preserve under compact representations, diverse clinical descriptions require robust linguistic reasoning, and token-based mask generation lacks inherent spatial coherence. We propose 32TokenMRISeg, an autoregressive (AR) framework that reformulates MRIS as next-token prediction and enables a multimodal LLM to generate segmentation masks directly using a highly compact representation—just 32 discrete tokens from a 1024-entry codebook. To maintain segmentation fidelity under this strict token budget, the framework integrates mask-level supervision to refine boundary details, chain-of-thought prompting to better align clinical descriptions with anatomical cues, and a token-order prediction signal to stabilize autoregressive decoding. Together, these lightweight mechanisms allow 32TokenMRISeg to produce accurate and spatially coherent masks with practical inference efficiency. Experiments on QaTa-COV19 and MosMedData+ demonstrate that our method achieves state-of-the-art performance despite its compact 32-token mask representation.

Keywords: Medical Referring Image Segmentation · Multimodal Large Language Model · Autoregressive Generation.

1 Introduction

Medical Referring Image Segmentation (MRIS) has emerged as a more flexible paradigm that segments target regions based on free-form natural language descriptions [1]. Early MRIS approaches employed traditional single-modal backbones with textual prompts incorporated at the encoder or decoder stage, or developed self-guided frameworks iterating between vision and language processing [2]. The emergence of large-scale vision-language models, particularly CLIP [3], spurred further interest due to impressive cross-modal generalization capabilities, with custom decoders designed to exploit CLIP's rich semantic

space [4]. Recent hybrid architectures have integrated cross-attention mechanisms to fuse image and text features: LViT [1] employed pixel-level attention for multimodal alignment, TGCAM [5] combined cross-attention with iterative text enhancement, GuideDecoder [2] applied language guidance in the decoder, and SGSeg [6] developed self-guided segmentation frameworks. However, these methods typically rely on complex multimodal fusion modules, multi-stage decoders, or external segmentation models like Segment Anything Model (SAM) [7].

With the rapid development of multimodal large language models (MLLMs) based on next-token prediction [8,9], their powerful image and text understanding capabilities offer new opportunities for MRIS. Recent HiMTok [10] explores hierarchical mask tokenization for image segmentation with MLLMs. Concurrent NTP-MRISeg [11] also formulates MRIS as next-token mask prediction, but relies on a much longer 1024-token representation. Unlike conventional fusion-based architectures, MLLMs can naturally process both visual and textual inputs through unified autoregressive (AR) generation. However, their strong understanding capabilities typically rely on long token sequences for image generation, which leads to slow inference and low efficiency. Furthermore, widely used large AR models such as Emu3 [8] and QwenVL [12] do not possess the capabilities required for MRIS: Emu3 produces high-quality images but offers limited Chain-of-Thought(CoT) reasoning and relies on long token sequences, while QwenVL provides strong reasoning but cannot output images. Neither is designed for pixel-level segmentation, underscoring the need for a dedicated AR formulation.

In this work, we introduce **32TokenMRISeg**, a unified AR framework that reformulates MRIS as next-token prediction and enables a multimodal LLM to generate segmentation masks directly. Modern large AR models typically rely on **long token sequences** to achieve high performance, but this leads to substantial inference latency, limiting their practical use for MRIS. To address this, each mask in our framework is represented using only **32 discrete tokens** from a 1024-entry codebook—an aggressive token budget that significantly shortens decoding but also makes boundary preservation challenging. To mitigate the spatial-detail loss introduced by this compact representation, the framework incorporates three lightweight design choices: **Mask-Level Supervision (MLS)** to refine the decoded mask and recover fine boundaries, **Chain-of-Thought (CoT)** prompting to help the model organize spatial cues from the referring description, and **Token Order Prediction (TOP)** to stabilize autoregressive decoding through an auxiliary ordering signal. These components together enable 32TokenMRISeg to maintain strong spatial fidelity under an extremely compact token representation, achieving state-of-the-art accuracy on two MRIS benchmarks while remaining highly efficient.

Our contributions are as follows: **First**, we propose an AR framework for MRIS that enables an MLLM to directly generate segmentation masks through a compact and efficient mask tokenization scheme, representing each mask with only 32 discrete tokens to achieve fast inference. **Second**, we incorporate three lightweight strategies-MLS, CoT prompting, and TOP-to enhance segmentation

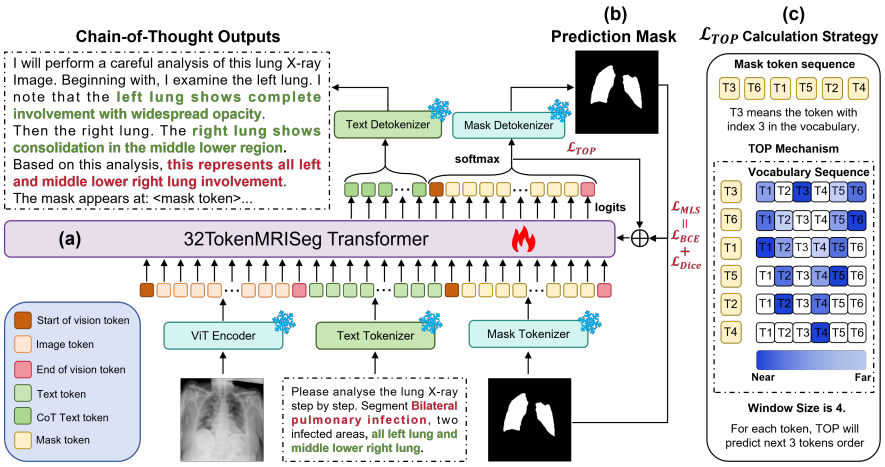

Fig. 1. Overall framework of 32TokenMRISeg. **(a) 32TokenMRISeg:** the model predicts each token in the sequence based on preceding tokens. **(b) Mechanism of MLS:** decoding the predicted mask token sequence and computing the BCE loss and Dice loss using the predicted mask with the ground-truth mask. **(c) Mechanism of TOP:** predicting the relative ordering of upcoming tokens rather than exact token identities.

fidelity under this strict token budget. **Third**, we demonstrate new state-of-the-art performance on two MRIS benchmarks, showing that reducing the mask sequence to 32 tokens retains high performance while achieving over 33× faster inference than the common 1024-token long-sequence AR generation.

2 Method

2.1 32TokenMRISeg Framework

Figure 1 illustrates the overall architecture of our proposed 32TokenMRISeg framework. Given an input medical image $I \in \mathbb{R}^{H\times W\times 3}$ and its corresponding clinical text description T, our objective is to generate a segmentation mask $M \in \{0,1\}^{H\times W}$ that accurately delineates the referred pathological regions.

The framework consists of: (1) a vision encoder that extracts visual features from the input image, (2) a text tokenizer that encodes clinical descriptions into discrete tokens, (3) a multimodal transformer that generates mask tokens autoregressively, and (4) a mask tokenizer which is trained to reconstruct binary masks from compact 1D quantized tokens for mask tokenization and the reverse process. We employ a pretrained Vision Transformer (ViT) as the image encoder to extract visual features $\mathbf{F}_v \in \mathbb{R}^{N_v\times D}$, where N_v represents the number of visual tokens and D denotes the feature dimension. We utilize a pretrained text tokenizer to convert the clinical description into a sequence of text tokens $\mathbf{F}_t = \{t_1, t_2, ..., t_{N_t}\}$, where N_t is the length of the text token sequence.

The visual features and text tokens are concatenated to form a unified multimodal sequence $\mathbf{S} = [\mathbf{F}_v; \mathbf{F}_t]$, which serves as input to the 32TokenMRISeg transformer. Following the AR paradigm, the model predicts mask tokens sequentially: $\mathbf{M}_{tok} = \{m_1, m_2, ..., m_{32}\}$, where each mask token m_i is tokenized from the mask tokenizer with a codebook of size 1024. The probability of generating the mask token sequence is formulated as:

$$p(\mathbf{M}_{tok}|\mathbf{S}) = \prod_{i=1}^{32} p(m_i|\mathbf{S}, m_{<i}) \tag{1}$$

where $m_{<i}$ denotes all previously generated mask tokens. Finally, the mask token detokenizer assembles the segmentation mask M from the predicted token sequence $\mathbf{M}_{tok}$. By unifying image, text, and mask within a single AR framework, 32TokenMRISeg eliminates the need for specialized fusion architectures while enabling end-to-end training with standard MLLM objectives.

2.2 Mask-Level Supervision

Mask learning through the compact 32-token representation and token-level cross-entropy loss alone is insufficient for capturing fine-grained spatial details, especially lesion boundaries. To address this, we introduce explicit MLS that directly constrains the predicted segmentation masks.

Specifically, during training, we decode the predicted mask token sequence $\mathbf{M}_{tok}$ into a continuous mask $\hat{M}$ using the mask detokenizer, and compute the MLS loss by comparing $\hat{M}$ with the ground-truth mask M. Following standard practices in medical image segmentation, we employ a combination of Binary Cross-Entropy (BCE) loss and Dice loss:

$$\mathcal{L}_{MLS} = \mathcal{L}_{BCE}(\hat{M}, M) + \mathcal{L}_{Dice}(\hat{M}, M). \tag{2}$$

The BCE loss ensures pixel-wise accuracy, while the Dice loss optimizes region overlap and handles class imbalance common in medical images. By combining token-level and mask-level supervision, the model generates token sequences that decode into accurate segmentation masks with well-defined boundaries.

2.3 Chain-of-Thought Reasoning

Medical referring expressions often involve complex anatomical descriptions and spatial relationships that require careful interpretation. To leverage the reasoning capabilities of pretrained MLLMs, we introduce CoT that encourages the model to explicitly articulate its diagnostic reasoning before generating mask tokens.

Specifically, we guide the model to analyze the clinical description, identify relevant anatomical landmarks, and reason about spatial relationships. During inference, the model generates intermediate reasoning steps enclosed within special tokens <think> and </think> before producing the mask token sequence.

This explicit reasoning process serves two purposes: (1) it decomposes complex referring expressions into structured anatomical understanding, and (2) it establishes clearer correspondences between linguistic descriptions and visual structures before mask generation. The CoT mechanism requires no architectural modifications—it exploits the pretrained MLLM's natural language generation capabilities. By encouraging step-by-step analysis, CoT enhances accurate localization of pathological regions described in nuanced clinical terminology, improving segmentation accuracy for challenging cases.

2.4 Token Order Prediction

While AR models trained with teacher forcing have achieved remarkable success in language modeling, they suffer from exposure bias during inference: the model is trained on ground-truth contexts but must rely on its own predictions when testing. This discrepancy can lead to error accumulation, particularly problematic for fine-grained MRIS where early prediction errors in the token sequence may propagate and degrade overall mask quality.

To mitigate this issue, we introduce a TOP mechanism that learns to predict the relative ordering of upcoming tokens rather than exact token identities. Specifically, given the current context tokens $m_{<i}$, we construct a target ranking sequence $\mathbf{k}_i \in \mathbb{R}^V$ for position i, where $V = 1024$ is the codebook size. The ranking scores are assigned based on token proximity: tokens that appear sooner in the ground-truth sequence receive higher scores, while tokens appearing later or not at all receive lower scores. The TOP objective is formulated as:

$$\mathcal{L}_{TOP} = -\sum_{i=1}^{I} \text{softmax}(\mathbf{k}_i) \cdot \log(\text{softmax}(\mathbf{u}_{TOP}(\mathbf{h}_i))), \tag{3}$$

where $\mathbf{h}_i$ is the hidden representation at position i, $\mathbf{u}_{TOP} : \mathbb{R}^D \to \mathbb{R}^V$ is a linear projection layer (TOP head), I is the window size and $\mathbf{k}_i$ contains the proximity-based ranking scores. Unlike standard cross-entropy loss that treats all incorrect tokens equally, TOP explicitly encourages the model to distinguish between tokens based on their temporal proximity in the sequence.

3 Experiments

3.1 Experimental Settings

Datasets We conduct experiments on two publicly available benchmark datasets for medical referring image segmentation. Specifically, the **QaTa-COV19** dataset comprises 9,258 chest X-ray images of COVID-19 patients with corresponding infection region annotations. Each image is accompanied by textual descriptions that specify the infection characteristics, including bilateral or unilateral involvement, the number of infected areas, and their approximate anatomical locations (e.g., "upper," "middle," or "lower" regions of the left and right lungs).

Table 1. Comparisons with SOTA methods on QaTa-COV19 and MosMedData+. † represents that the results are reported by the original paper. ∗ represents that the results are implemented by official open-source code.

Method	Backbone	Pub. Year	QaTa-COV19		MosMedData+	
			Dice↑	mIoU↑	Dice↑	mIoU↑
TransUNet* [14]	Hybrid	arXiv 2021	78.63	69.13	71.24	58.44
U-Net++* [15]	CNN	TMI 2019	79.62	70.25	71.75	58.39
nnU-Net* [16]	CNN	Nat. Methods 2020	80.42	70.81	72.59	60.36
TGANet* [17]	CNN	MICCAI 2022	79.87	70.75	71.81	59.28
LViT† [1]	Hybrid	TMI 2023	83.66	75.11	74.57	61.33
RefSegformer* [18]	Transformer	TMI 2024	84.09	75.48	74.98	61.70
LGA† [19]	SAM	MICCAI 2024	84.65	76.23	75.63	62.52
RecLMIS† [20]	CNN	TMI 2024	85.22	77.00	77.48	65.07
ARSeg† [21]	Hybrid	MICCAI 2025	83.88	72.63	72.36	58.38
SGSeg† [6]	Hybrid	MICCAI 2024	87.40	77.80	-	-
GuideDecoder†* [2]	Hybrid	MICCAI 2023	89.78	81.45	77.75	63.60
TGCAM† [5]	Hybrid	MICCAI 2024	90.60	82.81	77.82	63.69
32TokenMRISeg	**Transformer**	**-**	**91.53**	**84.39**	**79.50**	**65.98**

The dataset is divided into training, validation, and testing sets containing 5,716, 1,429, and 2,113 samples, respectively. The **MosMedData+** dataset contains 2,729 CT scan slices capturing lung infections. Similar to QaTa-COV19, each CT slice is paired with structured text annotations describing the infection patterns and spatial distributions across different lung regions. The dataset is split with 2,183 images for training, 273 for validation, and 273 for testing.

Evaluation Metrics We employ two widely-used metrics to quantitatively assess segmentation performance: the Dice Similarity Coefficient (Dice or DSC) and mean Intersection over Union (mIoU). Both metrics range from 0 to 1, with higher values indicating better segmentation quality.

Implementation Details We implement 32TokenMRISeg based on InternVL2.5-8B as the backbone MLLM and the mask tokenizer is built upon TiTok [13]. All experiments are conducted on two NVIDIA RTX A6000 Ada GPUs. Input images are resized to 448×448 pixels with dynamic patch support. We employ LoRA fine-tuning (rank 64) and train for 5 epochs using AdamW optimizer with learning rate 1×10^{-5} and cosine scheduling.

3.2 Comparison with SOTA

We compare our 32TokenMRISeg with SOTA medical image segmentation methods on both QaTa-COV19 and MosMedData+ datasets shown in Table 1 and Figure 2. The comparison includes unimodal segmentation methods and multimodal language-guided segmentation approaches.

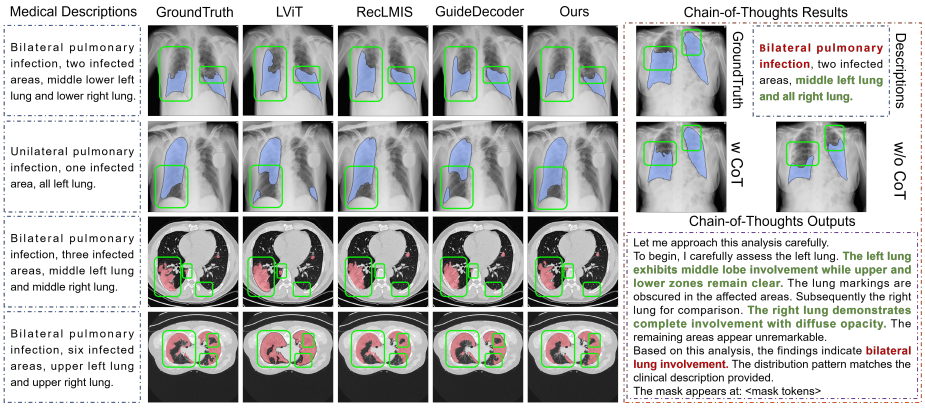

Fig. 2. The visualization of the main comparison with SOTA Methods on QaTa-COV19 and MosMedData+. The column titled "Medical Descriptions" denotes the input textual referring prompt, while the column titled "Image" signifies the input image. The column titled "GroundTruth" represents the ground truth segmentation target. The column titled "Ours" is the visualization result of our 32TokenMRISeg. The blue and red regions indicate the segmented infection areas for QaTa-COV19 and MosMedData+.

On the **QaTa-COV19** dataset, our method achieves 91.53% Dice and 84.39% mIoU, surpassing the previous SOTA TGCAM by 0.93% and 1.58%. Compared to GuideDecoder, which employs a text-guided architecture, 32TokenMRISeg demonstrates improvements of 1.75% in Dice and 2.94% in mIoU. The performance gains are even more substantial when compared to earlier methods: outperforming LViT by 7.87% in Dice and 9.28% in mIoU, and exceeds the best unimodal method nnU-Net by 11.11% in Dice and 13.58% in mIoU. On the **MosMedData+** dataset, 32TokenMRISeg achieves 79.50% Dice and 65.98% mIoU, establishing new SOTA performance. Our method surpasses TGCAM by 1.68% in Dice and 2.29% in mIoU, and outperforms GuideDecoder by 1.75% in Dice and 2.38% in mIoU. Notably, the improvement over RecLMIS, the previous best-performing method, is 2.02% in Dice and 0.91% in mIoU.

These results demonstrate that 32TokenMRISeg achieves better segmentation accuracy while maintaining a streamlined architecture with only 32 tokens per mask. This compact representation reduces inference overhead while preserving the rich cross-modal understanding capabilities of MLLMs, allowing the model to interpret complex clinical descriptions and generate accurate segmentation masks. The consistent improvements across both datasets validate the effectiveness of our proposed TOP, MLS, and CoT reasoning mechanisms.

3.3 Ablation Study

To validate the effectiveness of each proposed component, we conduct comprehensive ablation studies, as shown in Table 2. We systematically evaluate the contribution of three key components: Explicit MLS, CoT reasoning, and TOP.

Table 2. Ablation study of proposed components on QaTa-COV19 and MosMedData+

	MLS	TOP	CoT	QaTa-COV19		MosMedData+	
				Dice(%)↑	mIoU(%)↑	Dice(%)↑	mIoU(%)↑
(1)	×	×	×	88.51	81.03	76.25	60.75
(2)	✓	×	×	90.37	83.14	78.00	62.30
(3)	×	✓	×	89.00	81.45	77.79	62.19
(4)	×	×	✓	89.16	81.66	76.70	61.14
(5)	✓	✓	×	91.05	83.58	79.31	65.29
(6)	✓	✓	✓	**91.53**	**84.39**	**79.50**	**65.98**

Configuration (1) represents our baseline model without any proposed components, achieving 88.51% Dice and 81.03% mIoU on QaTa-COV19, and 76.25% Dice and 60.75% mIoU on MosMedData+. This baseline demonstrates competitive performance thanks to its compact 32-token representation and pretrained MLLM backbone. **Configurations (2)-(6)** demonstrate the incremental contributions of each component. Introducing MLS alone **(2)** yields the most significant individual gain, validating that direct mask supervision effectively compensates for fine-grained detail loss from compact tokenization. Adding TOP **(3)** shows that enforcing sequential dependencies improves spatially coherent mask generation. CoT reasoning **(4)** enhances the model's understanding of complex clinical descriptions. Combining MLS and TOP **(5)** produces synergistic effects that exceed individual contributions. The full model **(6)**, incorporating all three components, achieves a better performance of 91.53% Dice and 84.39% mIoU on QaTa-COV19, and 79.50% Dice and 65.98% mIoU on MosMedData+, confirming that all components contribute meaningfully to achieve SOTA accuracy.

Mask Token Length Analysis The number of mask tokens determines the granularity of mask representation. We investigate how different token lengths affect segmentation performance on QaTa-COV19. With 8 tokens, the representation is too coarse to capture mask boundaries and fine-grained details, achieving 84.13% Dice and 72.61% mIoU. Using 16 tokens improves performance to 89.69% Dice and 81.31% mIoU, which is sufficient for simple infection patterns with clear boundaries such as all lung involvement. Our 32-token configuration achieves 91.53% Dice and 84.39% mIoU, effectively representing more complex infection masks with irregular boundaries. However, expanding more tokens would reduce computational efficiency, as a longer token sequence raises the decoding overhead in MLS during training. These results demonstrate that 32 tokens provide a better balance between representation capacity and computational efficiency.

Inference Speed Analysis Autoregressive MLLMs typically rely on long token sequences to ensure generation quality, introducing substantial inference latency. Our 32TokenMRISeg requires only 1.22 seconds per mask image, with each mask token taking approximately 38.12 ms. In comparison, standard AR models like Emu3 that generate 1024 tokens per mask would require nearly 40 seconds per

image—a 33× slowdown. This demonstrates that our 32TokenMRISeg achieves both high segmentation accuracy and practical inference speed.

4 Conclusion

We introduced 32TokenMRISeg, an AR framework that generates MRIS masks directly using a compact 32-token representation. Supported by mask-level supervision, chain-of-thought prompting, and token-order prediction, the framework preserves boundary detail and decoding stability despite its small token budget. Experiments on QaTa-COV19 and MosMedData+ show that 32TokenMRISeg achieves state-of-the-art accuracy with practical inference efficiency, demonstrating the effectiveness of compact AR formulations for MRIS.

Acknowledgments. This work was supported by Australian Research Council (ARC) under Grant DP250100462, Grant LP240100451, and Grant IL250100082.

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

References

1. Zihan Li, Yunxiang Li, Qingde Li, Puyang Wang, Dazhou Guo, Le Lu, Dakai Jin, You Zhang, and Qingqi Hong. LViT: language meets vision transformer in medical image segmentation. *IEEE Trans. Med. Imag.*, 43(1):96–107, 2023.
2. Yi Zhong, Mengqiu Xu, Kongming Liang, Kaixin Chen, and Ming Wu. Ariadne's thread: Using text prompts to improve segmentation of infected areas from chest x-ray images. In *Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, pages 724–733. Springer, 2023.
3. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 8748–8763. PMLR, 2021.
4. Yaxiong Chen, Minghong Wei, Zixuan Zheng, Jingliang Hu, Yilei Shi, Shengwu Xiong, Xiao Xiang Zhu, and Lichao Mou. CausalCLIPSeg: Unlocking CLIP's potential in referring medical image segmentation with causal intervention. In *Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, pages 77–87. Springer, 2024.
5. Yunpeng Guo, Xinyi Zeng, Pinxian Zeng, Yuchen Fei, Lu Wen, Jiliu Zhou, and Yan Wang. Common vision-language attention for text-guided medical image segmentation of pneumonia. In *Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, pages 192–201. Springer, 2024.
6. Shuchang Ye, Mingyuan Meng, Mingjian Li, Dagan Feng, and Jinman Kim. Enabling Text-free Inference in Language-guided Segmentation of Chest X-rays via Self-guidance. In *Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, pages 242–252. Springer, 2024.

7. Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. MedCLIP-SAM: Bridging text and image towards universal medical image segmentation. In *Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, pages 643–653. Springer, 2024.
8. Xinlong Wang, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Zhen Li, Yuqi Wang, et al. Multimodal learning with next-token prediction for large multimodal models. *Nature*, pages 1–7, 2026.
9. Zayd MK Zuhri, Erland Hilman Fuadi, and Alham Fikri Aji. Predicting the order of upcoming tokens improves language modeling. *arXiv preprint arXiv:2508.19228*, 2025.
10. Tao Wang, Changxu Cheng, Lingfeng Wang, Senda Chen, and Wuyue Zhao. Himtok: Learning hierarchical mask tokens for image segmentation with large multimodal model. *arXiv:2503.13026*, 2025.
11. Xinyu Chen, Yiran Wang, Gaoyang Pang, Jiafu Hao, Chentao Yue, Luping Zhou, and Yonghui Li. Medical referring image segmentation via next-token mask prediction. *arXiv preprint arXiv:2511.05044*, 2025.
12. Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.
13. Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. In *Adv. Neural Inf. Process. Syst.*, volume 37, pages 128940–128966, 2024.
14. Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.
15. Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imag.*, 39(6):1856–1867, 2019.
16. Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods*, 18(2):203–211, 2021.
17. Nikhil Kumar Tomar, Debesh Jha, Ulas Bagci, and Sharib Ali. TGANet: Text-guided attention for improved polyp segmentation. In *Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, pages 151–160. Springer, 2022.
18. Jianzong Wu, Xiangtai Li, Xia Li, Henghui Ding, Yunhai Tong, and Dacheng Tao. Towards robust referring image segmentation. *IEEE Trans. Med. Imag.*, 2024.
19. Jihong Hu, Yinhao Li, Hao Sun, Yu Song, Chujie Zhang, Lanfen Lin, and Yen-Wei Chen. LGA: A language guide adapter for advancing the SAM model's capabilities in medical image segmentation. In *Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, pages 610–620. Springer, 2024.
20. Xiaoshuang Huang, Hongxiang Li, Meng Cao, Long Chen, Chenyu You, and Dong An. Cross-modal conditioned reconstruction for language-guided medical image segmentation. *IEEE Trans. Med. Imag.*, 2024.
21. Qijie Wang, Xian Lin, and Zengqiang Yan. Towards robust medical image referring segmentation with incomplete textual prompts. In *Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, pages 636–646. Springer, 2025.