MedVTok: A General-Purpose Medical Visual Tokenizer

Chenglong Ma[1,2,*], Yuanfeng Ji[3,7,*], Junzhi Ning[4,*], Jiyao Liu[1,*], Yujie Wei[1], Shaohao Rui[2,6], Chaoyang Zhang[2], Wenjie Li[2,6], Wenhao Tang[4], Cheng Tang[4], Wei Li[4,6], Shujian Gao[1,4], Yujie Zhang[1,2], Ying Chen[4], Zilong Li[1], Zongyuan Ge[5], Diping Song[4], Jin Ye[4,5], Ming Hu[4,5], Junjun He[2,4,7,†], Hongming Shan[1,†]

[1] Fudan University [2] Shanghai Innovation Institute [3] Stanford University School of Medicine [4] Shanghai Artificial Intelligence Laboratory [5] Monash University [6] Shanghai Jiao Tong University [7] Ant Healthcare, Ant Group

* Equal contribution † Co-corresponding authors

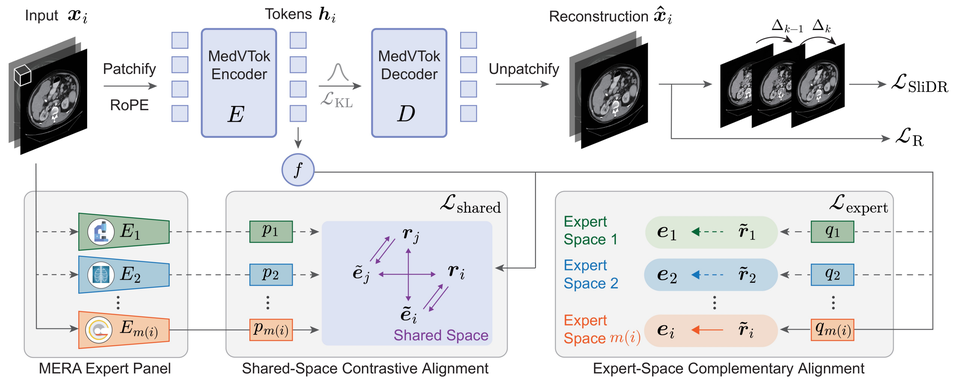
NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2026

A common visual interface for multimodal medical AI

- Medical AI needs to connect evidence of a person across examinations.
- However, separate tokenizers for different modalities and dimensions fragment visual representations, forcing downstream models to relearn how to adapt them.
- MedVTok maps multimodal 2D images and 3D volumes into **one reusable token space**. It preserves **anatomical detail and continuity** alongside **clinical semantics**, supporting both precise spatial tasks and clinical understanding.

Two complementary supervision signals



01 Slice Differential Regularization (SliDR): anatomical continuity

Voxel accuracy can miss inter-slice structure. Match first- and second-order differences to preserve transitions and curvature.

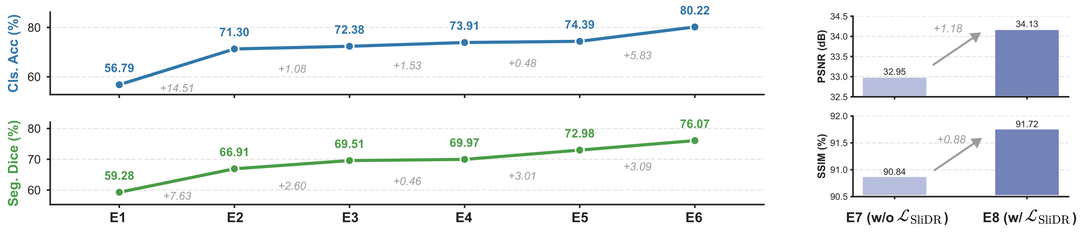
$$\mathcal{L}_{\text{SliDR}} = \frac{1}{T-1}\sum_{k=1}^{T-1}\|\Delta_k(\hat{\boldsymbol{x}}) - \Delta_k(\boldsymbol{x})\|_1 + \frac{1}{T-2}\sum_{k=2}^{T-1}\|\Delta_k^2(\hat{\boldsymbol{x}}) - \Delta_k^2(\boldsymbol{x})\|_1$$

02 Multi-Expert Representation Alignment (MERA): clinical semantics

Use the expert assigned to each modality. Align in shared space and back in expert space to retain complementary knowledge.

$$\mathcal{L}_{\text{shared}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp(\boldsymbol{r}_i\cdot\tilde{\boldsymbol{e}}_i/\tau)}{\sum_{j=1}^{B}\exp(\boldsymbol{r}_i\cdot\tilde{\boldsymbol{e}}_j/\tau)} + \log\frac{\exp(\boldsymbol{r}_i\cdot\tilde{\boldsymbol{e}}_i/\tau)}{\sum_{j=1}^{B}\exp(\boldsymbol{r}_j\cdot\tilde{\boldsymbol{e}}_i/\tau)}\right] \qquad \mathcal{L}_{\text{expert}} = \frac{1}{B}\sum_{i=1}^{B}\left(1 - \frac{\tilde{\boldsymbol{r}}_i\cdot\boldsymbol{e}_i}{\|\tilde{\boldsymbol{r}}_i\|\cdot\|\boldsymbol{e}_i\|}\right)$$

Ablations: what drives the gains?



E1 Baseline — Pool: Attn. | Experts: 0 | Shared loss ✗ | Expert loss ✗ | SliDR ✗ | 3D support ✗ — 3M images

E2 +MedImageInsight — Pool: Attn. | Experts: 1 | Shared loss ✗ | Expert loss ✗ | SliDR ✗ | 3D support ✗ — 3M images

E3 +Multiple experts — Pool: Attn. | Experts: Multi | Shared loss ✓ | Expert loss ✗ | SliDR ✗ | 3D support ✗ — 3M images

E4 +Expert loss — Pool: Attn. | Experts: Multi | Shared loss ✓ | Expert loss ✓ | SliDR ✗ | 3D support ✗ — 3M images

E5 Mean pooling — Pool: Mean | Experts: Multi | Shared loss ✓ | Expert loss ✓ | SliDR ✗ | 3D support ✗ — 3M images

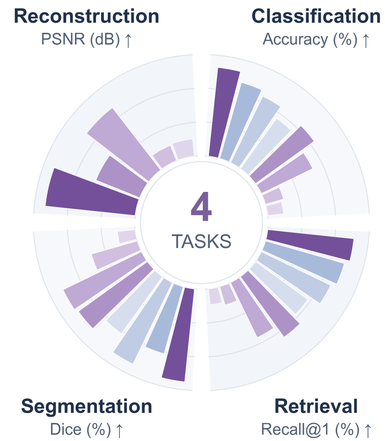
E6 +2D data — Pool: Mean | Experts: Multi | Shared loss ✓ | Expert loss ✓ | SliDR ✗ | 3D support ✗ — 28M images

E7 +3D data — Pool: Mean | Experts: Multi | Shared loss ✓ | Expert loss ✓ | SliDR ✗ | 3D support ✓ — 28M images + 70K volumes

E8 +SliDR — Pool: Mean | Experts: Multi | Shared loss ✓ | Expert loss ✓ | SliDR ✓ | 3D support ✓ — 28M images + 70K volumes

Broad utility with a frozen tokenizer

Low- and high-level representation utility

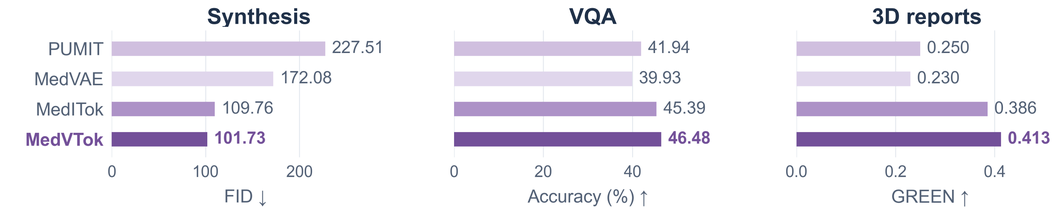


Model	2D recon. PSNR ↑ (dB)	Classif. Acc. ↑ (%)	Retrieval R@1 ↑ (%)	Segment. Dice ↑ (%)
MedVAE	27.95	51.32	51.30	38.86
PUMIT	27.78	52.41	52.53	53.63
AToken	38.22	67.50	69.12	73.03
MedITok	32.91	73.55	74.43	71.99
BiomedCLIP	—	69.45	70.52	64.63
MedSigLIP	—	74.12	75.38	74.62
MedImageInsight	—	76.21	77.34	64.38
MedVTok	**40.18**	**80.23**	**79.18**	**76.07**

— Not applicable to medical encoders.

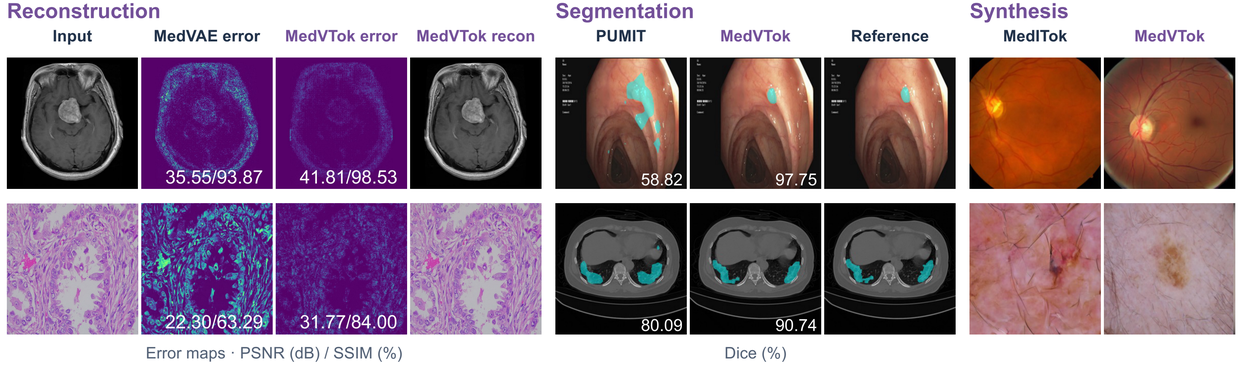
2D reconstruction: 9 modalities; classification / retrieval / segmentation: 9 / 9 / 8 benchmarks.

Generation, reasoning and reporting



Means over 7 modalities / 6 VQA benchmarks / 3 CT regions.

Qualitative results



- Contact: clma24@m.fudan.edu.cn ; yfj@stanford.edu
- Homepage: https://masaaki-75.github.io/medvtok