MaLViL: Multi-axis Low-rank Vision-LSTM for Medical Image Segmentation

Afshin Bozorgpour[1], Sina Ghorbani Kolahi[2], Moein Heidari[3], Ilker Hacihaliloglu[3], and Dorit Merhof[1,4]

[1] Faculty of Informatics and Data Science, University of Regensburg, Germany
[2] Independent Computer Science Researcher, Tehran, Iran
[3] University of British Columbia, Vancouver, BC, Canada
[4] Fraunhofer Institute for Digital Medicine MEVIS, Bremen, Germany
dorit.merhof@ur.de

Abstract. Vision-LSTM (ViL) enables efficient global modeling, but its cost still scales with the number of spatial tokens, so existing segmenters confine ViL to a coarse bottleneck and lose fine anatomical detail. Rasterizing 2D features into a 1D sequence further breaks adjacency across the orthogonal scan axis. We propose MaLViL, a Multi-axis Low-rank Vision-LSTM network that extends ViL across decoder resolutions. Bidirectional low-rank ViL (Bi-LRViL) reasons on a compact orthonormal subspace and preserves detail through an orthogonal residual; scale-aware SaLViL restores cross-axis neighbors before serialization; and a Cross-Directional Mixer (CDM) fuses orthogonal horizontal and vertical traversal paths. Statistics-Guided Skip Modulation (SGSM) further retains boundary cues in encoder skips. On skin-lesion, ultrasound, and multi-organ CT benchmarks, MaLViL achieves competitive or state-of-the-art segmentation accuracy, while reducing ViL operator memory by up to 83× at fine decoder resolutions. Code is available at: github.com/xmindflow/malvil.

Keywords: Medical image segmentation · Vision-LSTM · xLSTM · Low-rank approximation · Multi-axis sequence modeling

1 Introduction

Medical image segmentation (MIS) demands both precise delineation of fine anatomical boundaries and an understanding of long-range context [19]. Convolutional neural networks (CNNs) provide strong local inductive biases [2], but their effective receptive fields can limit global structural reasoning [15]. Vision Transformers (ViTs) address this limitation through self-attention, although its quadratic dependence on the number of image tokens is costly for high-resolution dense prediction [9]. More recently, state-space models such as Mamba have offered linear sequence modeling [12,21]; nevertheless, preserving two-dimensional locality while efficiently modeling global context remains an open challenge.

Vision-LSTM (ViL) [1], which adapts the matrix-memory mechanism of xLSTM [4] to visual sequences, provides an appealing alternative through bidirectional global modeling. Its application to dense segmentation, however, is constrained by the cost of processing long spatial sequences [10]. For a feature map with $N{=}HW$ tokens and channel width d, a ViL block incurs a sequence-dependent cost of $O(Nd^2)$, making shallow, high-resolution decoder stages particularly demanding in computation and activation memory. Existing designs therefore tend to concentrate ViL processing at low-resolution stages, limiting its contribution to multi-scale reconstruction. Moreover, rasterizing a two-dimensional feature map into a one-dimensional sequence separates spatial neighbors orthogonal to the scan direction, weakening the locality required for boundary-sensitive segmentation.

To address these limitations, we propose MaLViL, a Multi-axis Low-rank Vision-LSTM architecture that extends ViL reasoning across decoder resolutions. Its central component, scale-aware low-rank ViL (SaLViL), projects dense spatial tokens onto a compact orthonormal subspace of rank $p \ll N$, applies bidirectional ViL to the resulting sequence, and lifts the response back to the image grid. Although projection and reconstruction retain an $O(Np)$ cost, the expensive ViL operation is performed on p rather than N tokens. A learnable channel-wise gate balances global subspace reasoning with information in the orthogonal residual. SaLViL further splits channels into multi-scale groups and, before rasterization, applies axis-aligned convolutions across the direction orthogonal to the ViL scan so that neighbors broken by flattening remain accessible. It is embedded in the Cross-Directional Mixer (CDM), which runs SaLViL on two rotated views (horizontal and vertical traversal paths) and fuses their shared response with cross-directional disagreement to recover oriented boundaries. In parallel, Statistics-Guided Skip Modulation (SGSM) separates smooth and high-frequency encoder information, suppressing semantic clutter without discarding useful boundary cues. Together, these components couple compact global modeling with explicit spatial and frequency-aware refinement.

MaLViL is evaluated across skin-lesion, breast-ultrasound, and multi-organ CT segmentation benchmarks against representative CNN-, Transformer-, state-space-, and ViL/xLSTM-based methods. Direct comparison with bottleneck- and encoder-based ViL variants assesses whether low-rank modeling throughout the decoder offers benefits beyond inserting the same block family at low resolution. The main contributions are: ❶ *Multi-resolution low-rank ViL:* an orthogonal subspace formulation that performs the costly bidirectional ViL operation on p compressed tokens and uses a learnable high-frequency residual gate to preserve local detail; ❷ *Spatial- and frequency-aware refinement:* SaLViL and CDM restore cross-directional context, while SGSM suppresses irrelevant skip responses without discarding boundary information; and ❸ *Comprehensive validation:* multi-modal experiments demonstrate consistent improvements over diverse segmentation architectures, supported by component ablations and operator-level memory analysis.

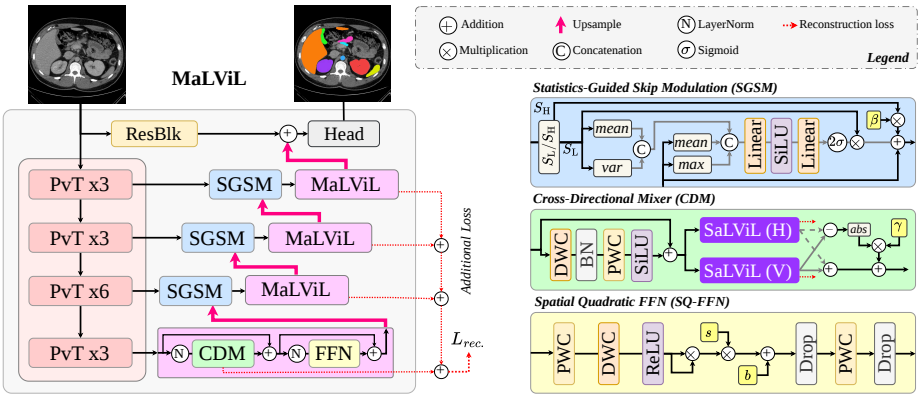

Fig. 1. MaLViL architecture. A hierarchical encoder feeds four MaLViL Blocks; the three encoder skips are fused by SGSM. Each block contains a Cross-Directional Mixer (CDM) followed by a Spatial Quadratic FFN (SQ-FFN). The dashed path denotes the auxiliary low-rank reconstruction loss.

2 Methodology

Overview. MaLViL is a hierarchical encoder–decoder for medical image segmentation (Figure 1). A hierarchical encoder produces features $\{E_i\}_{i=1}^{4}$ at resolutions from 1/4 to 1/32 of the input. Four MaLViL Blocks then reconstruct the prediction from coarse to fine, and Statistics-Guided Skip Modulation (SGSM) fuses the three available encoder skips:

$$D_4 = \mathcal{M}_4(E_4), \qquad D_i = \mathcal{M}_i(\mathrm{SGSM}\,(\mathcal{U}(D_{i+1}), E_i))\,, \; i = 3, 2, 1, \tag{1}$$

where $\mathcal{M}_i$ and $\mathcal{U}$ denote a MaLViL Block and upsampling, respectively. A shallow residual feature of the input is added after the final upsampling before the prediction head.

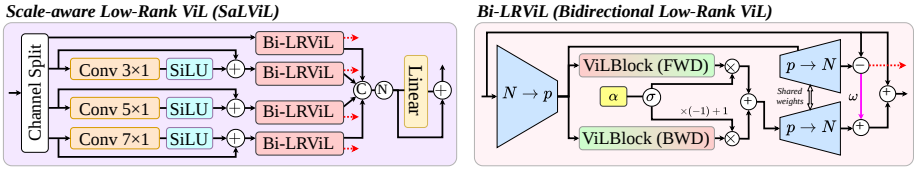

Fig. 2. SaLViL and Bi-LRViL. Left: scale-specific branches recover orthogonal neighbors before Bi-LRViL sequence modeling. Right: rank-p projection with opposite forward/backward ViL scans and gated fusion (α, ω).

MaLViL Block. For tokens $X \in \mathbb{R}^{B\times N\times C}$, $N = HW$, the block alternates directional context modeling and local spatial refinement within a pre-normalized

residual formulation:

$$\tilde{X} = X + \gamma_c \odot \mathrm{CDM}(\mathrm{LN}(X)), \qquad Y = \tilde{X} + \gamma_f \odot \mathrm{SQFFN}(\mathrm{LN}(\tilde{X})), \quad (2)$$

where γ_c and γ_f control the contributions of the two residual paths. This organization first establishes long-range, multi-directional context and then restores local spatial detail.

Bidirectional Low-Rank ViL (Bi-LRViL). Applying ViL to all N spatial tokens is costly at shallow decoder stages. Bi-LRViL uses a learnable basis $V \in \mathbb{R}^{N\times p}$ ($p \ll N$, fixed per stage), kept orthonormal by QR each forward ($V^\top V = I$). Each Bi-LRViL has its own V; shared-weight CDM reuses the same V on H/V paths. Projection and lifting are

$$Z = V^\top X \in \mathbb{R}^{p\times C}, \qquad \hat{X} = VZ = VV^\top X, \qquad X_\perp = X - \hat{X}. \quad (3)$$

The expensive sequence operation therefore acts on p coefficients rather than all N tokens, while $X_\perp$ preserves spatial information outside the compact subspace. ViL is then applied only to the p compressed coefficients. Along this serialized path, opposite forward and backward traversals capture context from both ends of the sequence; their outputs are mixed by a channel-wise gate $\alpha = \sigma(a)$:

$$\bar{Z} = \alpha \odot \mathrm{ViL}_f(Z) + (1-\alpha) \odot \mathrm{ViL}_b(Z), \quad Y = X + V(\bar{Z} - Z) - \omega \odot X_\perp, \quad (4)$$

where $\omega = \sigma(w)$ down-weights the orthogonal residual $X_\perp$ when injecting the global update, preserving fine local structure that lies outside the low-rank subspace. The reconstruction regularizer $\ell_{\mathrm{rec}} = \mathrm{MSE}(X, \hat{X})$ encourages the learned basis to retain informative spatial variation; orthonormality is enforced by QR, not by ℓ_{rec}.

SaLViL. Rasterizing a $H\times W$ map into $N{=}HW$ tokens preserves adjacency along the scan direction but separates neighbors across the orthogonal axis. SaLViL recovers this lost two-dimensional structure before low-rank sequence modeling (Figure 2). Channels are partitioned into scale-specific groups with complementary receptive fields (e.g., 1×1, 3×1, 5×1, 7×1 in the current orientation). Within each group, lightweight axis-aligned convolutions are applied before serialization to aggregate context across the axis opposite to the ViL traversal, thereby bridging neighbors that would otherwise be distant in the flattened sequence. The enriched features are then rasterized and passed to Bi-LRViL, which performs global reasoning on only $p \ll N$ subspace coefficients per branch. Branch outputs are concatenated and adaptively remixed with the input.

Cross-Directional Mixer (CDM). A single rasterization path is inherently anisotropic: row-major scanning privileges one axis over the other. CDM therefore evaluates SaLViL on two orthogonal traversal paths (the native map and a 90°-rotated view), so horizontal and vertical neighborhoods are serialized and reasoned over symmetrically. Denoting the two responses by F_H and F_V, CDM retains their shared structure while restoring oriented detail suppressed by plain averaging:

$$\mu = \frac{1}{2}(F_H + F_V), \qquad A = \frac{1}{2}\big(|F_H - \mu| + |F_V - \mu|\big), \qquad F_{\mathrm{CDM}} = \mu + \gamma_h \odot A. \quad (5)$$

Because $A = \frac{1}{2}|F_H - F_V|$ for two views, the learned scale γ_h re-injects cross-directional disagreement as a direction-sensitive correction. A lightweight 3×3 residual pathway further supplies isotropic neighborhood context.

Statistics-Guided Skip Modulation (SGSM). Given decoder feature D and encoder skip S, SGSM decomposes the skip as $S_l = \mathrm{AvgPool}_{3\times3}(S)$ and $S_h = S - S_l$. A compact MLP uses first- and second-order statistics of the smooth skip and decoder context to predict a channel gate:

$$q = [\mu(S_l), \mathrm{var}(S_l), \mu(D), \max(D)], \qquad g = 2\sigma(\mathrm{MLP}(q)). \tag{6}$$

Fusion is

$$\mathrm{SGSM}(D, S) = D + g \odot S_l + \beta \odot S_h, \tag{7}$$

where $g \in [0, 2]$ adaptively suppresses or enhances smooth semantic content, and the learned channel scale β controls the propagation of high-frequency boundary detail. This decomposition prevents strong encoder activations from overwhelming the decoder while retaining useful fine structure.

Spatial Quadratic FFN (SQ-FFN) and objective. SQ-FFN complements global reasoning with a lightweight local transformation. After channel expansion and depthwise spatial mixing, a learnable quadratic activation $s \odot \mathrm{ReLU}(x)^2 + b$ strengthens salient local responses before projection to the original width. The network is optimized using

$$\mathcal{L} = \mathcal{L}_{\mathrm{seg}} + \lambda \mathcal{L}_{\mathrm{rec}}, \qquad \mathcal{L}_{\mathrm{rec}} = \frac{1}{4}\sum_{i=1}^{4} \overline{\ell}_{\mathrm{rec}}^{\,i}, \quad \lambda = 0.01, \tag{8}$$

where $\mathcal{L}_{\mathrm{seg}}$ is the benchmark-specific segmentation loss and $\overline{\ell}_{\mathrm{rec}}^{\,i}$ averages reconstruction errors over the SaLViL branches and directional responses at decoder stage i.

3 Experiments and Results

Datasets and Experimental Setups.
All experiments are implemented in PyTorch and conducted on a single NVIDIA A5000 GPU with 24 GB memory. MaLViL uses an ImageNet-pretrained PVTv2-B2 encoder [24] with decoder ranks p=32/64/128/256 (coarse→fine). Reported methods follow their established benchmark splits, and all retrained baselines use the same data partitions as MaLViL.

Synapse. The dataset contains 30 abdominal CT volumes. Following TransUNet [9], 18 volumes are used for training and 12 for testing. MaLViL is trained for 350 epochs with AdamW, batch size 8, and an initial learning rate of 10^{-4}. The retrained nnU-Net baseline [14] operates at its native 512×512 resolution. Its complexity is measured for one 2D slice, with one multiply–accumulate reported as one FLOP to match the convention used in the comparison tables.

Skin-lesion and ultrasound segmentation. We evaluate PH^2, HAM10000, ISIC 2017/2018, and BUSI using their established splits. The ISIC 2017 split

Table 1. Evaluation results on the Synapse dataset (best results in bold; second-best results underlined).

Methods	Params	FLOPs	Spl.	RKid.	LKid.	Gal.	Liv.	Sto.	Aor.	Pan.	Average	
											DSC↑	HD95↓
R50 U-Net [9]	30.42 M	-	85.87	78.19	80.60	63.66	93.74	74.16	87.74	56.90	74.68	36.87
TransUNet [9]	96.07 M	88.91 G	85.08	77.02	81.87	63.16	94.08	75.62	87.23	55.86	77.49	31.69
Swin-UNet [7]	27.17 M	6.16 G	90.66	79.61	83.28	66.53	94.29	76.60	85.47	56.58	79.13	21.55
VM-UNet [21]	44.27 M	6.52 G	89.51	82.76	86.16	69.41	94.17	81.40	86.40	58.80	81.08	19.21
PVT-EMCAD-B2 [19]	26.76 M	5.60 G	92.17	84.10	88.08	68.87	95.26	83.92	88.14	68.51	83.63	15.68
MSA^{2}Net [15]	112.77 M	15.56 G	<u>92.69</u>	84.24	88.30	<u>74.35</u>	95.59	84.03	<u>89.47</u>	69.30	84.75	<u>13.29</u>
2D D-LKA Net [3]	101.64 M	19.92 G	91.22	<u>84.92</u>	<u>88.38</u>	73.79	94.88	84.94	88.34	67.71	84.27	20.04
GLM-SFNet [8]	14.35 M	4.64 G	91.98	84.35	87.49	**74.78**	95.14	85.31	88.32	<u>71.24</u>	<u>84.82</u>	**11.87**
UxLSTM-Bot [10]	15.02 M	16.56 G	91.65	73.56	80.21	64.58	94.77	76.78	87.14	58.88	78.45	19.71
nnU-Net [14]	46.35 M	60.02 G	92.04	81.48	86.68	65.20	**96.31**	<u>85.54</u>	**90.65**	**74.80**	84.09	19.82
MaLViL	37.49 M	6.56 G	**92.74**	**85.71**	**89.71**	72.98	<u>96.28</u>	**87.64**	88.83	69.91	**85.48**	15.23

contains 1,250 training, 150 validation, and 600 test images. Skin-lesion models are trained for 40 epochs at 224×224, while BUSI training uses 50 epochs at 256×256. Both settings use AdamW, batch size 8, an initial learning rate of 10^{-4}, and the preprocessing and augmentation protocol of [6].

3.1 Comparative Results

Synapse. Table 1 reports MaLViL's performance on the Synapse dataset, where it achieves a 0.66% DSC improvement over the previous best GLM-SFNet [8], and surpasses MSA^{2}Net [15] and 2D D-LKA Net [3] by 0.73% and 1.21%, respectively, demonstrating its effectiveness across diverse abdominal organs. For additional reference against relevant model families, we retrain nnU-Net [14] at 512×512 and UxLSTM-Bot [10] at 224×224, using the same split. At their respective operating resolutions, MaLViL exceeds nnU-Net by 1.39 percentage points in mean DSC (85.48 vs. 84.09); their computational profiles are 6.56 and 60.02 G, respectively. MaLViL also improves over the bottleneck-only UxLSTM baseline by 7.03 percentage points (85.48 vs. 78.45), supporting the benefit of low-rank ViL throughout the decoder. Notably, nnU-Net remains highly competitive on well-defined structures such as the aorta and pancreas, whereas MaLViL yields more consistent gains across the smaller, harder organs, resulting in the best overall average. Fig. 3 further confirms MaLViL's superior multi-scale accuracy for challenging structures such as the gallbladder, pancreas, and stomach.

Skin Benchmarks. Table 2 shows that MaLViL consistently outperforms CNN-, Transformer-, hybrid, and Mamba-based architectures across the evaluated skin-lesion benchmarks. It improves upon CENet [5] by 0.64 and 0.46 percentage points in DSC on PH^2 and HAM10000, respectively, and attains a state-of-the-art DSC of 91.23% on ISIC 2018. On ISIC 2017 it reaches 92.08% DSC, second only to GLM-SFNet [8] (92.18%) while attaining the highest accuracy (97.15%). Fig. 4 qualitatively illustrates the resulting lesion boundaries on PH^2, HAM10000, and BUSI.

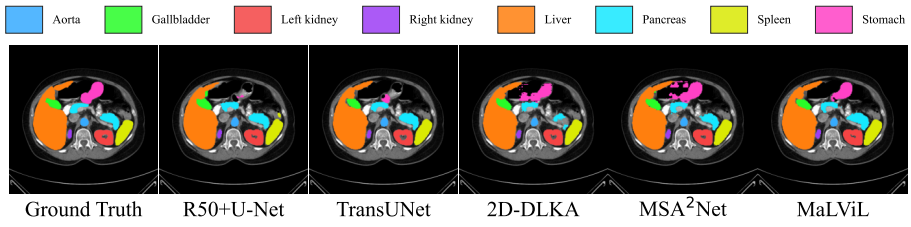

Fig. 3. Qualitative comparison on the Synapse multi-organ dataset.

Table 2. Comparison on PH^2, HAM10000, ISIC'17, ISIC'18, and BUSI datasets.

Methods	PH^2		HAM10000	
	Dice	Acc.	Dice	Acc.
U-Net [20]	89.36	92.33	91.67	95.67
TransUNet [9]	88.40	92.00	93.53	96.49
Swin-Unet [7]	94.49	96.78	92.63	96.16
Att-UNet [18]	90.03	92.76	92.68	96.10
UCTransNet [23]	90.93	94.08	93.46	96.84
MissFormer [13]	85.50	90.50	92.11	96.21
CENet [5]	95.04	97.19	94.71	97.10
UxLSTM-Enc [10]	91.51	94.42	93.71	96.89
MaLViL	**95.68**	**97.71**	**95.17**	**97.33**

Methods	ISIC2017		ISIC2018	
	DSC↑	ACC↑	DSC↑	ACC↑
U-Net [20]	89.89	96.13	88.51	95.39
VM-UNet [21]	90.70	96.45	89.91	95.54
VM-UNet v2 [26]	90.45	96.37	89.02	95.51
DermoSegDiff [6]	91.43	96.72	89.66	95.75
EGE-UNet [22]	90.73	96.42	88.19	95.10
GLM-SFNet [8]	**92.18**	97.08	90.64	96.04
UltraLight VM-UNet [25]	90.91	96.46	89.40	95.58
UxLSTM-Enc [10]	90.19	96.38	89.78	96.04
MaLViL	92.08	**97.15**	**91.23**	**96.75**

Methods	BUSI
	Dice↑
U-Net [20]	74.04
PraNet [11]	75.41
CaraNet [17]	77.34
Swin-UNet [7]	77.38
TranFuse [28]	79.36
U-Kan [16]	76.40
UWT-net [27]	79.65
UxLSTM-Enc [10]	72.80
MaLViL	**81.76**

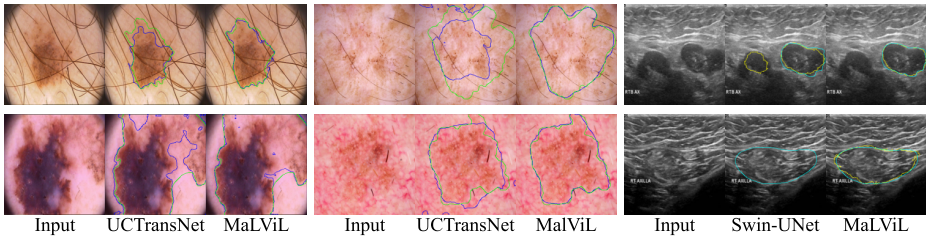

Fig. 4. Qualitative segmentation on PH^2, HAM10000, and BUSI.

Ultrasound. As reported in Table 2, MaLViL achieves the highest listed BUSI Dice score of 81.76%, exceeding UWT-Net [27] by 2.11 percentage points. The examples in Fig. 4 show accurate lesion localization and boundary delineation.

Ablation Studies. Figure 5 visualizes a representative ISIC 2017 example at the 28×28 decoder resolution. Bi-LRViL produces complementary forward and backward responses, which combine into a more coherent lesion representation. The horizontal and vertical SaLViL responses emphasize different spatial patterns, while CDM consolidates them into a coherent target region. SGSM further decomposes the encoder skip S into a smooth semantic component S_{low} and a detail component S_{high} before producing the modulated skip output. Each activation map is normalized independently for visualization.

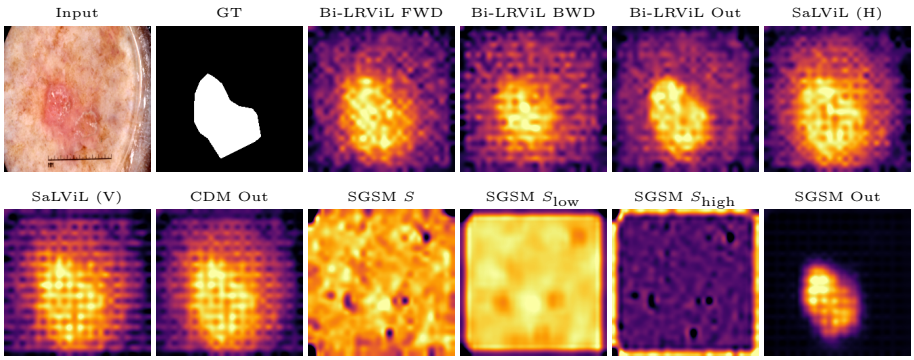

Fig. 5. Internal features on ISIC 2017 (28×28): Bi-LRViL, SaLViL/CDM, and SGSM activations (normalized per map).

Table 3. Component and efficiency analysis of MaLViL. (a) Progressive PH2 ablation with whole-network forward-pass resources at 224×224 (batch size 1); excluded components use identity mappings. (b) Operator-level profiling of full- versus low-rank bidirectional ViL in isolation at each decoder resolution (batch size 8).

(a) Component contribution

Configuration	Dice	FLOPs(G)	Mem.(MB)	Params(M)
Full ViL	93.72	8.83	1543	**32.89**
+ Low-rank	94.30	**5.91**	**180**	33.81
+ Bidirectional	94.57	6.05	190	36.21
+ SaLViL	94.93	6.02	193	36.94
+ CDM	95.32	6.60	195	37.34
+ SGSM	95.41	6.60	196	37.49
+ L_{rec} **(MaLViL)**	**95.68**	6.60	196	37.49

(b) Resolution efficiency

Stage	N	Mem.(MB)		FLOPs(G)		Reduction
		Full	LR	Full	LR	
1	56^2	10796.4	**129.5**	14.17	**0.55**	**83**×
2	28^2	740.2	**62.7**	2.01	**0.29**	12×
3	14^2	102.8	**48.0**	0.67	**0.34**	2.1×
4	7^2	56.3	**53.8**	0.34	**0.27**	1.0×

Component and efficiency analysis. Table 3(a) summarizes a progressive PH^2 ablation at 224×224 (batch size 1). Low-rank projection yields the largest efficiency gain (8.83→5.91 GFLOPs, 1543→180 MB) and improves Dice from 93.72% to 94.30%. Table 3(b) reports operator-level memory and computational costs for full and low-rank bidirectional ViL in isolation at each decoder resolution (batch size 8), rather than for the end-to-end network. At 56×56, low-rank ViL reduces operator memory by 83×, illustrating why MaLViL can apply ViL throughout the decoder where full-sequence modeling becomes impractical.

4 Conclusions

We presented MaLViL, a multi-resolution decoder that makes Vision-LSTM reasoning practical beyond the bottleneck through orthogonal low-rank projection. Bi-LRViL, SaLViL, and CDM capture global context across orthogonal traversal paths, while SGSM retains informative boundary cues during skip fusion. Experiments on skin-lesion, ultrasound, and abdominal CT segmentation demonstrate competitive or state-of-the-art accuracy, with substantial ViL memory savings at fine decoder resolutions.

Acknowledgments. The authors gratefully acknowledge the computational and data resources provided by the Leibniz Supercomputing Centre (www.lrz.de).

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

References

1. Alkin, B., Beck, M., Pöppel, K., Hochreiter, S., Brandstetter, J.: Vision-lstm: xlstm as generic vision backbone. arXiv preprint arXiv:2406.04303 (2024)
2. Azad, R., Aghdam, E.K., Rauland, A., Jia, Y., Avval, A.H., Bozorgpour, A., Karimijafarbigloo, S., Cohen, J.P., Adeli, E., Merhof, D.: Medical image segmentation review: The success of u-net. arxiv. arXiv preprint arXiv:2211.14830 (2022)
3. Azad, R., Niggemeier, L., Hüttemann, M., Kazerouni, A., Aghdam, E.K., Velichko, Y., Bagci, U., Merhof, D.: Beyond self-attention: Deformable large kernel attention for medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1287–1297 (January 2024)
4. Beck, M., Pöppel, K., Spanring, M., Auer, A., Prudnikova, O., Kopp, M., Klambauer, G., Brandstetter, J., Hochreiter, S.: xlstm: Extended long short-term memory. Advances in Neural Information Processing Systems **37**, 107547–107603 (2024)
5. Bozorgpour, A., Kolahi, S.G., Azad, R., Hacihaliloglu, I., Merhof, D.: Cenet: Context enhancement network for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 120–129. Springer (2025)
6. Bozorgpour, A., Sadegheih, Y., Kazerouni, A., Azad, R., Merhof, D.: Dermosegdiff: A boundary-aware segmentation diffusion model for skin lesion delineation. In: International workshop on predictive intelligence in medicine. pp. 146–158. Springer (2023)
7. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. pp. 205–218. Springer (2022)
8. Chen, J., Qi, F., Chang, C., Hu, Q., Fu, K., Wang, X., Liu, K.: Glm-sfnet: Global-local vision-mamba with semantic fusion for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 224–234. Springer (2025)
9. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
10. Chen, T., Ding, C., Zhu, L., Xu, T., Ji, D., Wang, Y., Zang, Y., Li, Z.: xlstm-unet can be an effective 2d & 3d medical image segmentation backbone with vision-lstm (vil) better than its mamba counterpart. arXiv preprint arXiv:2407.01530 (2024)
11. Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: Pranet: Parallel reverse attention network for polyp segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 263–273. Springer (2020)
12. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. In: First conference on language modeling (2024)

13. Huang, X., Deng, Z., Li, D., Yuan, X., Fu, Y.: Missformer: An effective transformer for 2d medical image segmentation. IEEE transactions on medical imaging **42**(5), 1484–1494 (2022)
14. Isensee, F., Jaeger, P.F., Kohl, S.A.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)
15. Kolahi, S.G., Chaharsooghi, S.K., Khatibi, T., Bozorgpour, A., Azad, R., Heidari, M., Hacihaliloglu, I., Merhof, D.: Msa $\hat{2}$ net: Multi-scale adaptive attention-guided network for medical image segmentation. arXiv preprint arXiv:2407.21640 (2024)
16. Li, C., Liu, X., Li, W., Wang, C., Liu, H., Liu, Y., Chen, Z., Yuan, Y.: U-kan makes strong backbone for medical image segmentation and generation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 39, pp. 4652–4660 (2025)
17. Lou, A., Guan, S., Ko, H., Loew, M.H.: Caranet: context axial reverse attention network for segmentation of small medical objects. In: Medical Imaging 2022: Image Processing. vol. 12032, pp. 81–92. SPIE (2022)
18. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
19. Rahman, M.M., Munir, M., Marculescu, R.: Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11769–11779 (2024)
20. 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)
21. Ruan, J., Li, J., Xiang, S.: Vm-unet: Vision mamba unet for medical image segmentation. ACM Transactions on Multimedia Computing, Communications and Applications (2024)
22. Ruan, J., Xie, M., Gao, J., Liu, T., Fu, Y.: Ege-unet: an efficient group enhanced unet for skin lesion segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 481–490. Springer (2023)
23. Wang, H., Cao, P., Wang, J., Zaiane, O.R.: Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 2441–2449 (2022)
24. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pvt v2: Improved baselines with pyramid vision transformer. Computational visual media **8**(3), 415–424 (2022)
25. Wu, R., Liu, Y., Liang, P., Chang, Q.: Ultralight vm-unet: Parallel vision mamba significantly reduces parameters for skin lesion segmentation. arxiv 2024. arXiv preprint arXiv:2403.20035 (2024)
26. Zhang, M., Yu, Y., Jin, S., Gu, L., Ling, T., Tao, X.: Vm-unet-v2: rethinking vision mamba unet for medical image segmentation. In: International symposium on bioinformatics research and applications. pp. 335–346. Springer (2024)
27. Zhang, P., Ouyang, X., Peng, R.: Uwt-net: Mining low-frequency feature information for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 615–624. Springer (2025)
28. Zhang, Y., Liu, H., Hu, Q.: Transfuse: Fusing transformers and cnns for medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 14–24. Springer (2021)