Hybrid Attention-State Space Model Backbone for Cancer Pathology Vision Foundation Model with Superior Downstream Task Performance

Won June Cho[1], Daeky Jeong[1], Hyeongyeol Lim[1], Hongjun Yoon[1*]

[1]DEEPNOID Inc, Korea

Abstract

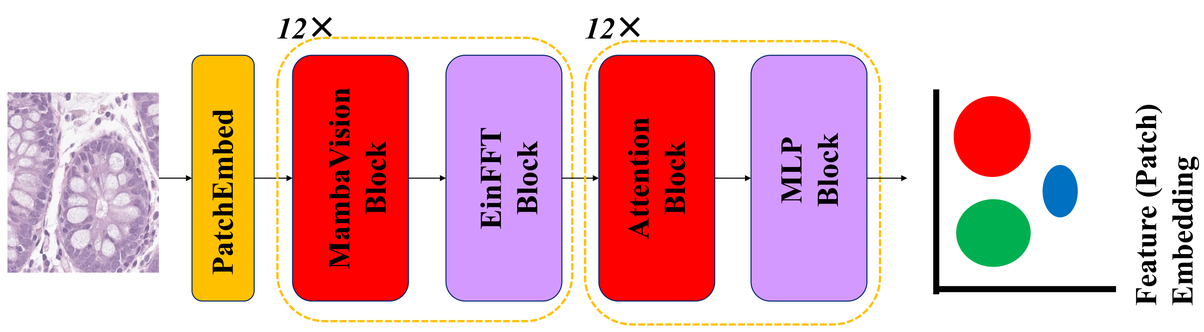
Colorectal cancer (CRC) management increasingly depends on the precise histologic and molecular characterization of pathology slides—including subtyping, polyp and adenoma grading, microsatellite instability (MSI) prediction, biomarker assessment, and survival prediction. Digital histopathology vision foundation models (VFMs) trained on whole slide images (WSIs) have shown promise across these tasks, but current Vision Transformer (ViT)-based VFMs still underperform on clinically demanding endpoints, particularly those tied to molecular phenotypes where the relevant tissue features are subtle and lie beyond human perception. To address this, we propose **MV_{Hybrid}**, a hybrid vision backbone combining state space models (SSMs) with ViT layers, motivated by the observation that SSMs initialized with negative real eigenvalues exhibit a strong bias toward the low-frequency morphological signals correlated with molecular states. We evaluate MV_{Hybrid} across the breadth of downstream tasks central to CRC pathology AI. MV_{Hybrid} matched or outperformed ViT baselines across nearly every downstream metric. The largest and most clinically meaningful gains were in gene expression (spatial transcriptomics) prediction—MV_{Hybrid} achieved roughly 57% higher correlation than that of the best-performing ViT. We also observed consistent gains of MV_{Hybrid}'s downstream performance in clinically significant Microsatellite Instability (MSI), polyp/adenoma, and tissue classification. Similar trends could also be found in downstream survival prediction and zero-shot retrieval tasks.

Introduction

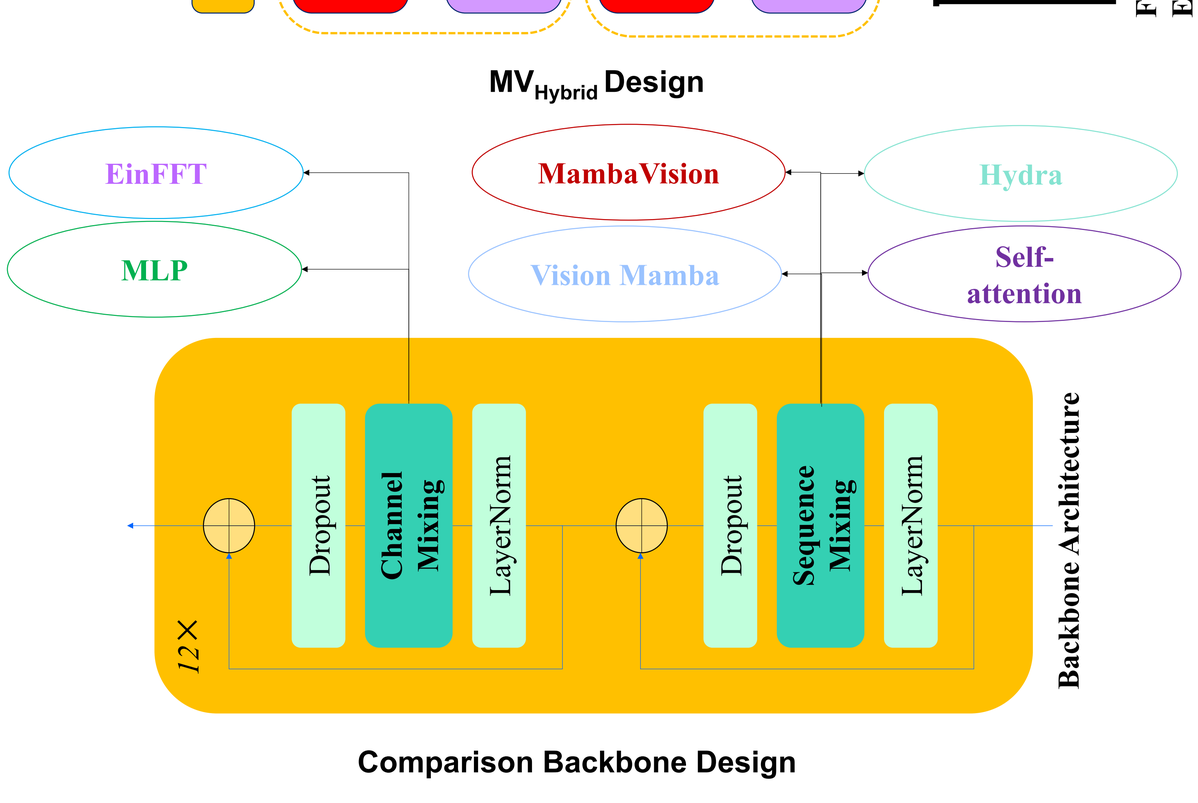
Despite being pretrained on millions of WSIs, current pathology VFMs have recently been shown to learn site- and hospital-specific features over true biological ones, indicating that scaling data alone is unlikely to close the performance gap on molecular tasks. WSIs also differ quantitatively from natural images by containing far larger proportions of high-frequency features, while molecular phenotype signals are assumed to be inherently low-frequency—a mismatch that motivates moving beyond ViT-only backbones. We therefore propose the backbone architecture itself as a meaningful lever for pathology AI, and present what is, to our knowledge, the first study to pretrain and evaluate multiple backbone architectures—spanning ViT, pure state space, and hybrid designs—on the same colorectal cancer pretraining dataset.

Materials & Methods

MV_{Hybrid} and five comparison backbones (vanilla ViTs and pure/hybrid Mamba-family architectures) were pretrained identically on the same public CRC WSI dataset (about 750k 256 x 256 patches from publicly available HunCRC and IMP-CRS-2024) using DINOv2 self-supervised learning method. The pretrained encoders were then evaluated on CRC downstream tasks spanning histologic classification (MHIST, UniToPatho), MSI classification (TCGA-CRC-MSI), spatial transcriptomics/biomarker prediction (HEST), zero-shot patch retrieval (NCT-CRC-100K, 9 tissue classes), and overall survival prediction (TCGA-CRC).



MV_{Hybrid} Design



Comparison Backbone Design

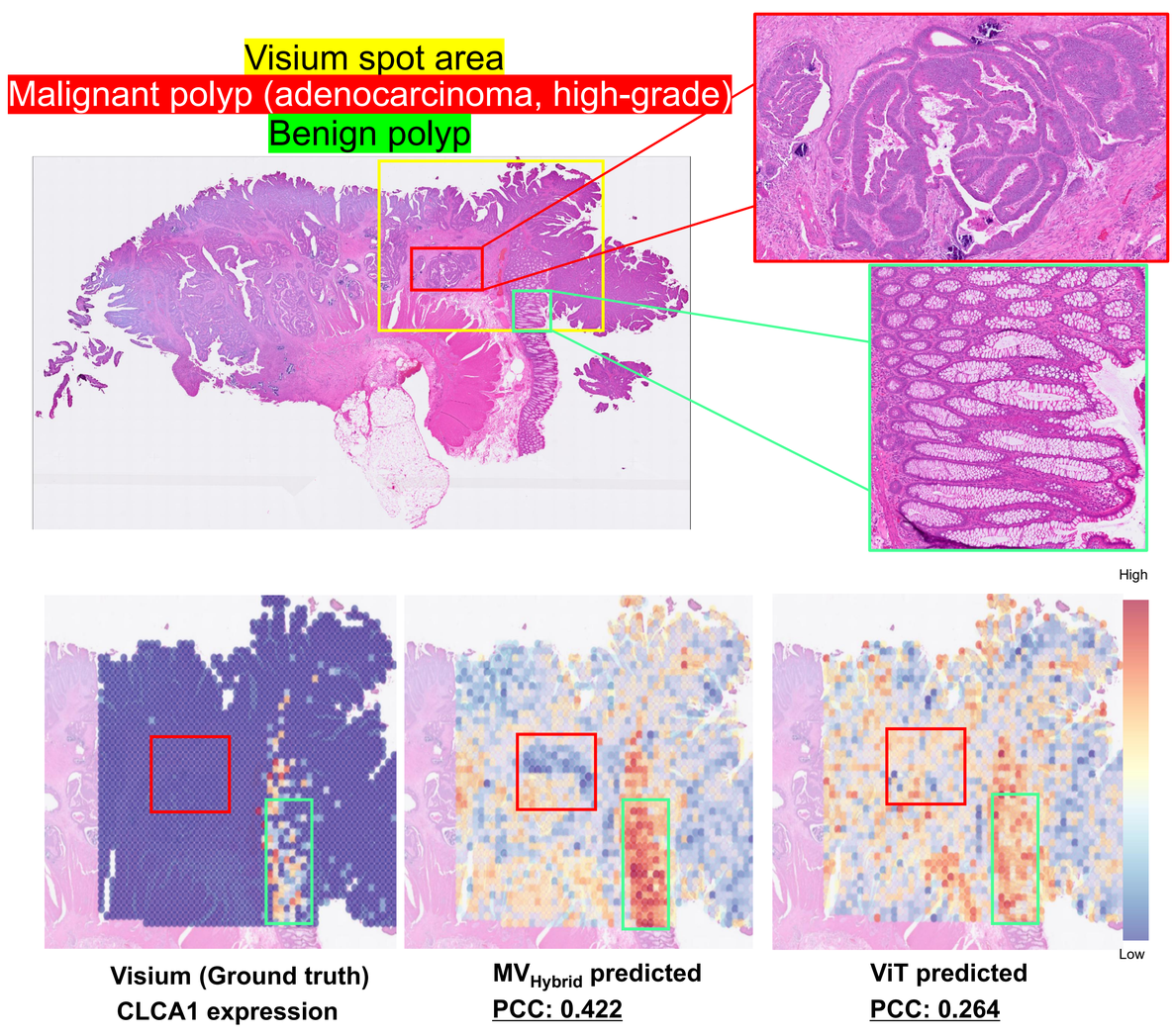
Results

Dataset & Evaluation	Metric	ViT†	ViT	ViM_{EinFFT}	$Hydra_{EinFFT}$	$Hydra_{Hybrid}$	MV_{Hybrid} (Ours)
TCGA-CRC-MSI (WSI, Binary) CLAM (k=10)	AUROC	0.706 ±0.116	0.746 ±0.134	0.730 ±0.089	**0.772 ±0.103**	0.763 ±0.103	0.765 ±0.090
	Mean Acc	0.657 ±0.139	0.696 ±0.086	0.668 ±0.072	0.707 ±0.098	0.718 ±0.102	**0.75 ±0.073**
MHIST (Patch, Binary) Linear Probing	AUROC	0.831	0.804	0.793	0.837	0.855	**0.863**
	Balanced Acc	0.713	0.705	0.689	0.72	0.758	**0.768**
UniToPatho (Patch, 6 class) Linear Probing	Mean AUROC	0.801	0.789	0.791	0.806	0.802	**0.82**
	Balanced Acc	0.416	0.403	0.396	0.416	0.405	**0.463**

Downstream CRC Classification Results

Dataset & Evaluation	Metric	ViT†	ViT	ViM_{EinFFT}	$Hydra_{EinFFT}$	$Hydra_{Hybrid}$	MV_{Hybrid} (Ours)
NCT-CRC-100K (Patch, 9 class) Zero-shot retrieval	Acc@1	0.674	0.648	0.762	0.763	0.774	**0.789**
	MVAcc@5	0.714	0.693	0.81	0.78	0.808	**0.825**
HEST (WSI) Ridge	Mean Pearson Correlation	0.415 ±0.055	0.365 ±0.042	0.397 ±0.065	0.404 ±0.064	0.415 ±0.068	**0.460 ±0.082**
HEST (WSI) XGBoost	Mean Pearson Correlation	0.332 ±0.055	0.340 ±0.081	0.347 ±0.108	0.374 ±0.077	0.368 ±0.073	**0.425 ±0.073**
TCGA-CRC (WSI) GMM	Mean c-index	0.623 ±0.081	0.620 ±0.132	0.601 ±0.056	**0.677 ±0.074**	0.651 ±0.071	0.658 ±0.076

Downstream Retrieval, Gene expression/Survival Prediction Results



Visium (Ground truth) CLCA1 expression | **MV_{Hybrid} predicted PCC: 0.422** | **ViT predicted PCC: 0.264**

Example Gene Expression Prediction on CLCA1 of a Visium CRC sample

Conclusion

- MV_{Hybrid} consistently outperforms current ViT-based VFM architectures across the molecular, morphological, and retrieval tasks that define clinical tasks in CRC.
- Backbone architecture—not only dataset scale—is a meaningful lever for advancing clinically useful VFMs, particularly on the molecular endpoints where ViT-only models underperform.
- These results position MV_{Hybrid} as a strong candidate for next-generation pathology foundation model training, possibly replacing current ViT-based models.