GeneVAR: Causal MeanFlow for Autoregressive Gene-to-WSI Tile Synthesis

Jianwei Zhao[1], Fan Yang[2], Xin Li[1,#], Qiang Zhai[3], Ao Luo[4], Ziqi Ren[5], Zhicheng Jiao[6], and Hong Cheng[1]

UESTC[1], ALPACA AI LAB[2], SICAU[3], SWJTU[4], XDU[5], Brown[6]

1.Introduction

Why WSI:

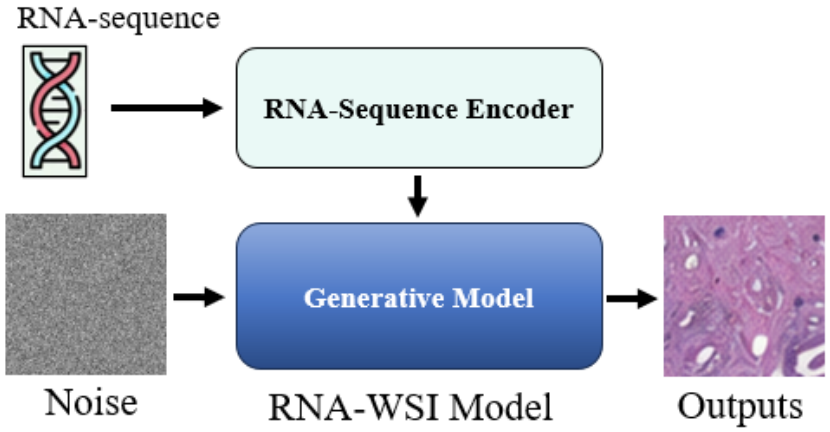
Histological Whole Slide Image (WSI) classification is essential in digital pathology, automating diagnosis and sub-typing while extracting insights from high-resolution scansto support prognosis and treatment planning

Why Gene-to-WSI:

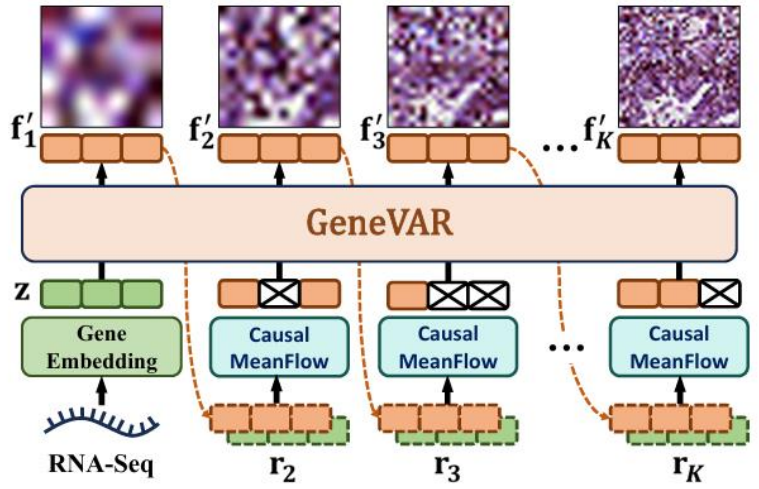
Understanding how transcriptomic programs shape tissue morphology remains a central challenge in computational pathology. Gene-to-WSI tile synthesis offers a principled generative framework to translate molecular profiles into histological images.

Strategy:

Most existing methods compress RNA-Seq into a single global embedding injected once at initialization, an oversimplified design that weakens transcriptomic signals. We present GeneVAR, an Autoregressive Gene-to-WSI model that reformulates synthesis as an iterative, coarse-to-fine generative process。



Traditional solutions



Ours GeneVAR

1.Introduction

CVPR JUNE 3-7, 2026 DENVER COLORADO

Motivation

Despite recent progress, Gene-to-WSI tile synthesis remains underdeveloped. Existing methods typically compress RNA-Seq profiles into low-dimensional embeddings to condition GANs or diffusion models. However, this design has three key limitations: (1) one-time conditioning leads to signal decay, causing generated images to drift from gene-driven morphology; (2) fixed-resolution synthesis imposes scale rigidity, reducing cross-scale consistency.

Embeddings are learned in a purely correlational manner, leaving models vulnerable to confounders. An iterative paradigm alone does not immunize the generative trajectory against confounders, such as tumor purity, and staining variability, whose correlational footprints can divert decoding from gene-driven semantics.

Solution

Coarse-to-Fine: To address these challenges, we reformulate Gene-to-WSI synthesis as an *iterative, coarse-to-fine* generative process in which transcriptomic signals are injected at multiple stages during generation rather than used once at initialization. Building on this principle, we propose Gene VAR-the first autoregressive Gene-to-WSI model that delivers step-wise molecular reinforcement while preserving structural coherence and causal fidelity.

Causal MeanFlow: r-CM is introduced as a gene-driven causal module intrinsically embedded in the auto regressive trajectory. Within this framework, a Structural Causal Model (SCM) governs generative updates from coarse to fine semantics, ensuring that causal factors, *i.e.*, semantic information regulated by gene expression, remain consistent, while non-causal factors are suppressed.

1.Introduction

Overall, the contributions of this work can be summarized as:

- **Paradigm Shift.** GeneVAR reformulates Gene-to-WSI synthesis as an iterative, coarse-to-fine generative process in which transcriptomic signals are injected at multiple stages across scales, overcoming the signal decay and rigidity of static global embeddings.

- **Causal Fidelity.** By integrating the novel Causal MeanFlow module into the autoregressive trajectory, GeneVAR disentangles transcriptomic effects from confounders, ensuring that morphological synthesis remains biologically grounded and causally aligned.

- **State-of-the-Art Performance.** Comprehensive evaluation on five TCGA cohorts demonstrates that GeneVAR attains the lowest FID and the highest accuracy in downstream classification, setting a new benchmark for both generative fidelity and functional utility.

2.Method

- **Multi-Scale Vector Quantization：**

We employ multi-scale vector quantization (MSVQ) to discretize a WSI tile $X \in R^{H\times W\times C}$ into K hierarchical token maps $S = \{s_k\}_{k=1}^{n}$, where each $s_k \in R^{h_k\times w_k}$ represents a discrete map at scale k.

- **Problem Reformulation：**

Gene-to-WSI synthesis should preserve transcriptomic guidance dynamically across scales. However, existing methods collapse RNA-Seq into static embeddings injected once, leading to signal decay, scale rigidity, and spurious correlations. To overcome these issues, we reformulate Gene-to-WSI synthesis as a coarse-to-fine autoregressive process, where an image is quantized into hierarchical token maps $S = \{s_k\}_{k=1}^{n}$ and generated sequentially under recurrent guidance from g:

$$s_k = P_\Theta(s_{<k}, \mathbf{g})$$

Here, $\mathbf{g}$ is injected at multiple stages rather than only once at initialization, ensuring that transcriptomic signals re-main persistently active throughout the generative trajectory while leveraging the distributional strength of autoregressive models.

2.Method

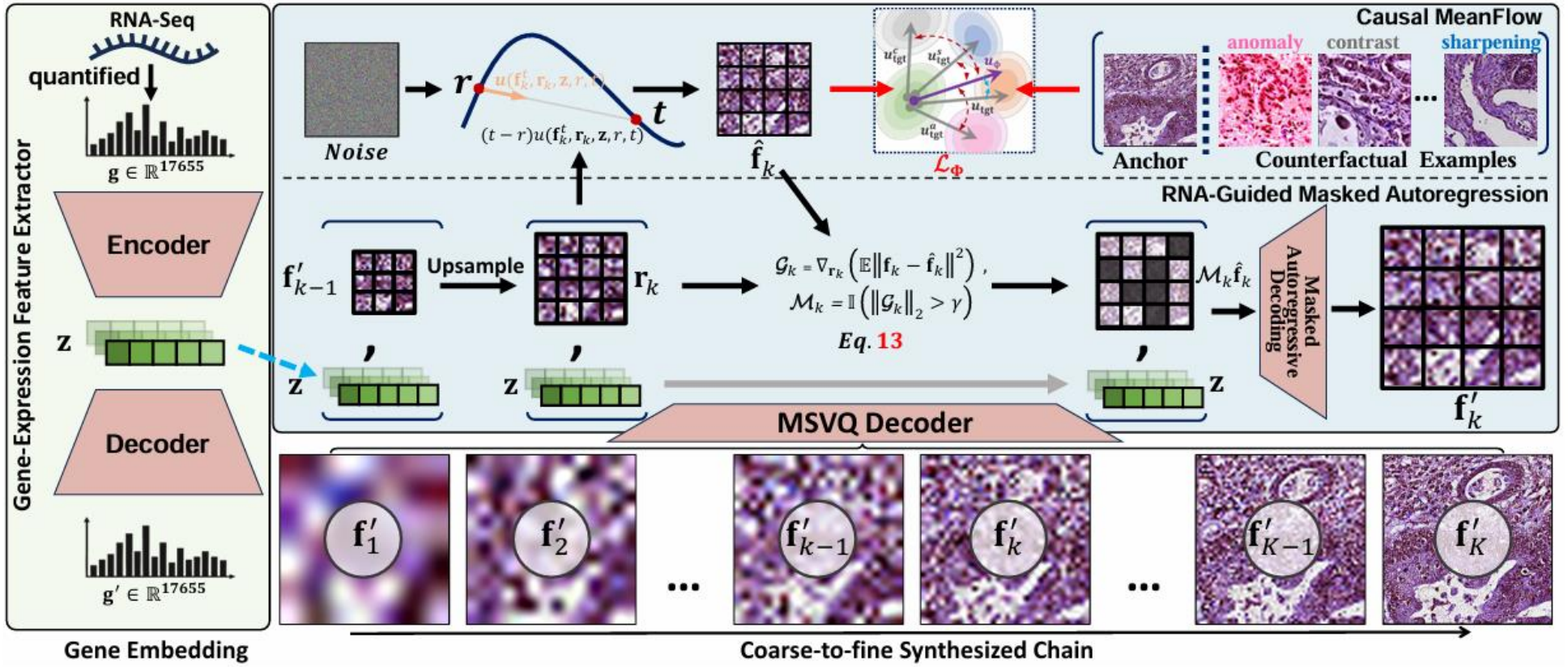


Figure 2. **Overview of GeneVAR**. RNA-Seq embeddings are injected into a multi-scale autoregressive pipeline, where morphology is reconstructed from $\mathbf{f}'_K$ via the MSVQ decoder and reinforced by Causal MeanFlow with counterfactual supervision (see Sec. 3 for details).

Fig. 2 presents an overview of our GeneVAR, an Autoregressive Gene-to-WSI model that reformulates synthesis as an iterative, coarse-to-fine generative process. At its core is a novel Causal MeanFlow module that reinforces transcriptome-informed guidance at multiple stages and mitigates non-causal factors through counterfactual-style interventions, thereby ensuring biological fidelity throughout the generative trajectory.

2.Method

Causal MeanFlow：

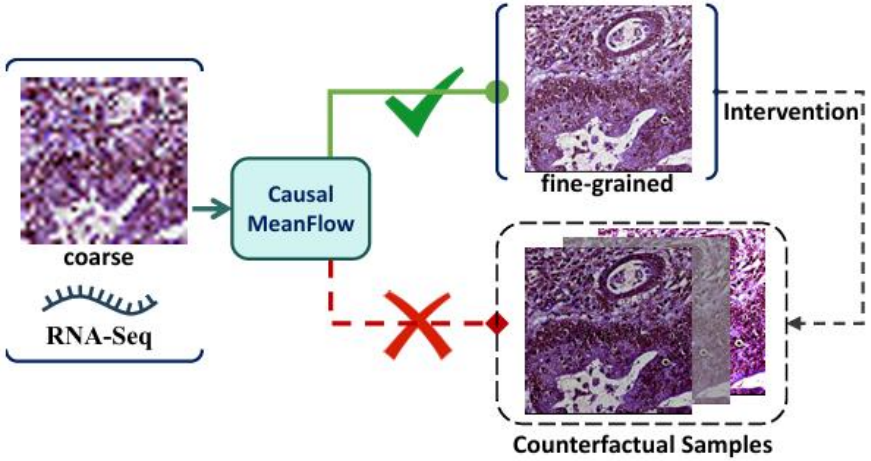
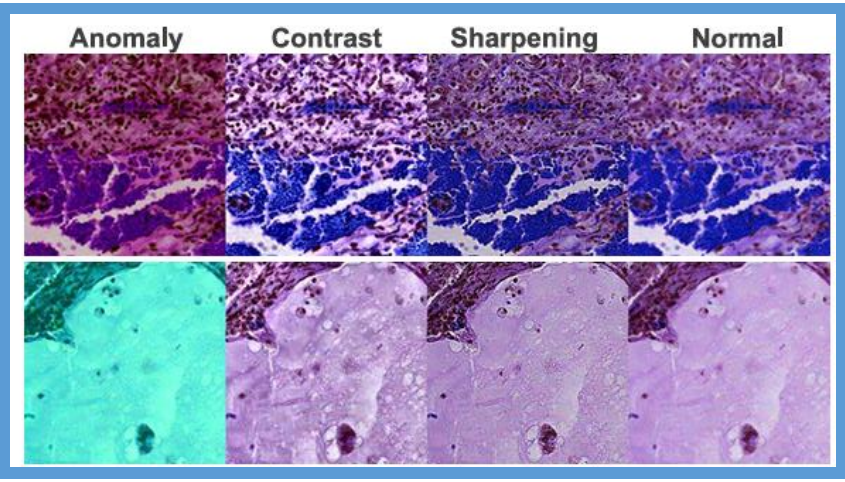


Figure 3. **Structural Causal Mechanism.** Our model transforms coarse gene-guided features into fine-grained morphology, while counterfactual interventions suppress non-causal variations.



Counterfactual examples construction and velocity formulation:

b.) Counterfactual Regularization. Autoregressive modeling can be viewed as an iterative reconstruction of quantized features, producing latent representations $\{\mathbf{f}_k^{'}\}_{k=1}^{K}$. Within our framework, $\mathbf{f}_k^{'}$ is reconstructed from $\mathbf{r}_k$ under the guidance of gene embedding $\mathbf{z}$, progressively enriching semantics across scales. To prevent spurious enhancements from non-causal factors such as tumor purity and staining [1, 23, 28, 38], we introduce a causal regularization strategy based on counterfactual interventions. Following the Structural Causal Model (SCM) [34] and the principle that interventions must destruct non-causal cues [45], we apply three perturbations, *i.e.,* color anomaly and contrast adjustment (to model stain variation) and sharpening (to simulate tumor purity differences), to generate counterfactual variants X^a, X^c, X^s for each WSI tile X, as illustrated in Fig. 3. After quantization, these yield feature maps $\{\mathbf{f}_k^a, \mathbf{f}_k^c, \mathbf{f}_k^s\}_{k=1}^{K}$, which enforce Φ to emphasize causal, scale-invariant morphology.

To endow r-$\mathcal{CM}$ with causal regularization, we enlarge the discrepancy between the predicted average velocity u_ϕ and the target velocities $\{u_{\text{tgt}}^a, u_{\text{tgt}}^c, u_{\text{tgt}}^s\}$ obtained from degraded counterfactuals, thereby stabilizing reconstruction. For example, u_{tgt}^a is not derived via costly partial derivatives in Eq. 6. Instead, we decompose average velocity into magnitude and direction, where the direction is computed by normalizing the flow between two sampled fields $\mathbf{f}_k^{a,t}$ and $\mathbf{f}_k^{a,r}$, and the magnitude is estimated by scaling $||u_{\text{tgt}}||$ with a stochastic factor λ. Targets $\{u_{\text{tgt}}^c, u_{\text{tgt}}^s\}$ are constructed analogously. The procedure is formally expressed as:

$$u_{\text{tgt}}^a = \lambda_a ||u_{\text{tgt}}|| \cdot \frac{\mathbf{f}_k^{a,t} - \mathbf{f}_k^{a,r}}{||\mathbf{f}_k^{a,t} - \mathbf{f}_k^{a,r}||}. \tag{9}$$

c.) Learning Objective. We further define a causality-driven learning objective to disentangle scale-invariant morphological semantics (causal factors) from degradation-related cues (non-causal factors). For the k-th iteration, u_{tgt} from $\mathbf{f}_k$ serves as the anchor, while $\mathbf{f}_{k-1}$ is treated as an extreme counterfactual, yielding u_{tgt}^l via Eq. 9. Additional degradation samples $\{u_{\text{tgt}}^a, u_{\text{tgt}}^c, u_{\text{tgt}}^s\}$ derived from independent WSIs, distinct from the source of u_{tgt}, forming the counterfactual set $\mathcal{C}_u$. This design blocks potential shortcuts that exploit non-causal signals and enforces reliance on scale-invariant morphology. The training objective is:

$$\mathcal{L}_\Phi = \mathbb{E}\big[||u_\phi - \text{sg}(u_{\text{tgt}})||^2\big] - \frac{\alpha}{N} \sum_{u_{\text{tgt}}^n \sim \mathcal{C}_u} \mathbb{E}\big[||u_\phi - \text{sg}(u_{\text{tgt}}^n)||^2\big], \tag{10}$$

where $\text{sg}(\cdot)$ denotes stop-gradient and α controls the strength of causal regularization.

2.Method

RNA-Guided Masked Autoregression:

We formulate GeneVAR as an autoregressive model over tokenized sequences, This factorization is realized by a ViT-based transformer with a causal mask, ensuring that generation proceeds strictly left-to-right under continuous transcriptomic conditioning, akin to GPT-2, as shown in **Eq.11**.

$$p(s_1, s_2, \cdots, s_K) = \prod_{k=1}^{K} p(s_k | s_{<k}, s_0 = \mathbf{z}), \tag{11}$$

a.) Gradient-Guided Masking. After r-$\mathcal{CM}$ converges, it produces stable reconstructions $\hat{\mathbf{f}}_k$ from $\mathbf{r}_k$ before passing them to the transformer. We define the gradient as the derivative of the MSE loss between $\mathbf{f}_k$ and $\hat{\mathbf{f}}_k$ with respect to $\mathbf{r}_k$, and compute its ℓ_2 norm along the channel dimension:

$$\begin{aligned} \mathcal{G}_k &= \nabla_{\mathbf{r}_k}(\mathbb{E}||\mathbf{f}_k - \hat{\mathbf{f}}_k||^2), \\ \mathcal{M}_k &= \mathbb{I}(\|\mathcal{G}_k\|_2 > \gamma), \gamma \sim \mathcal{N}(0,1)^{h_k \times w_k}. \end{aligned} \tag{12}$$

Here $\mathbb{I}(\cdot)$ denotes the indicator function. Tokens with large gradient responses are deemed RNA-conditioned salient regions, as perturbations there strongly affect reconstruction. These positions are therefore preferentially retained, dynamically injecting transcriptomic cues at the k-th iteration.

b.) Masked Autoregressive Decoding. Following class/text-conditioned VAR models, the latent gene embedding $\mathbf{z}$ is used as the initial condition token $\mathbf{r}_1$. For subsequent steps, tokens at positions with $\mathcal{M}_k = 0$ are replaced by a learnable embedding e. The autoregressive generation and cross-entropy training objective are formulated as:

$$\begin{aligned} \{s_k^{'}\}_{k=1}^{K} &= \mathcal{P}_\Theta(\{\mathbf{r}_1, \cdots, \mathcal{M}_k\hat{\mathbf{f}}_k, \mathcal{M}_K\hat{\mathbf{f}}_K\}), \\ \mathcal{L}_\Theta &= \sum_{k=1}^{K} \mathbb{CE}(s_k^{'}, s_k). \end{aligned} \tag{13}$$

For efficiency, masking is applied only when $k \geq K_m$, balancing accuracy and cost. Together, these advances endow GeneVAR with causality-aware, biologically grounded generation, which we validate in the next section through comprehensive experiments.

3.Quantitative Results

Table 1. **Quantitative comparison on TCGA benchmarks.** Fréchet Inception Distance (**FID, lower is better**) is reported along with model parameters (#Para) and generation steps (#Step). GeneVAR consistently achieves the best performance across all cohorts.

Method	Type	#Para	#Step	GBM	CESC	KIRP	COAD	LUAD	ALL
RNA-CDM [4]	Diffusion	1146M	2000	24.15	25.76	24.46	33.60	27.98	23.36
PathLDM [43]	Diffusion	400M	50	22.46	19.87	20.39	26.96	21.28	20.29
U-ViT [2]	Diffusion	297M	100	13.89	15.74	21.83	26.75	17.86	18.55
DiT [29]	Diffusion	305M	250	15.03	18.75	21.53	29.13	18.57	18.11
SiT [27]	Flow Matching	305M	25	14.63	19.48	22.10	29.47	19.52	18.84
LlamaGen [37]	Token-wise Autoregression	343M	256	15.48	18.34	19.89	27.91	17.52	17.43
VQGAN [11]	Token-wise Autoregression	227M	256	21.74	23.48	26.10	32.47	26.48	25.09
VAR [39]	Scale-wise Autoregressive	310M	10	12.96	17.21	17.32	25.84	15.40	16.83
ImageFolder [26]	Scale-wise Autoregressive	314M	10	23.75	23.97	25.81	33.25	25.46	24.81
GeneVAR (Ours)	Scale-wise Autoregressive	317M	10	**11.20**	**13.68**	**14.49**	**19.86**	**13.65**	**12.95**

1.GeneVAR achieves leading generative performance (FID) across all datasets, setting new benchmarks.

Table 2. **Cell distribution comparison.** Mean±std of cell proportions across all cohorts. Extended results are shown in Supplementary Material.

Dataset	Method	Neoplastic	Dead
GBM	VAR	9.96%±21.99	69.55%±34.59
	Ours	**5.04%±12.07**	**78.67%±26.91**
	Real	6.04%±16.29	77.49%±30.90
LUAD	VAR	25.06%±27.38	55.99%±30.00
	Ours	**20.20%±22.78**	**61.10%±26.56**
	Real	19.32%±25.21	63.12%±29.60

2 . While FID captures visual idelity, biological plausibility is critical for clinical relevance. We confirm that GeneVAR not only preserves overall cell composition but also generates morphologies consistent with cancer type, demonstrating biological realism beyond mere visual fidelity.

Table 3. **Tile classification performance.** ACC/F1 under varying substitution (p) and pretraining (q) ratios. Full results are provided in Supplementary Material.

Method	p=0.0		p=0.75	
	ACC	F1	ACC	F1
R-CDM [4]	0.573±0.020	0.556±0.022	0.492±0.016	0.472±0.046
VAR [39]	0.573±0.034	0.556±0.015	0.521±0.030	0.510±0.035
Ours	**0.579±0.032**	**0.570±0.039**	**0.592±0.029**	**0.588±0.031**
	q=0.5		q=1.0	
R-CDM [4]	0.618±0.026	0.612±0.030	0.650±0.021	0.641±0.031
VAR [39]	0.661±0.020	0.662±0.019	0.708±0.011	0.709±0.010
Ours	**0.722±0.023**	**0.722±0.020**	**0.767±0.015**	**0.766±0.017**

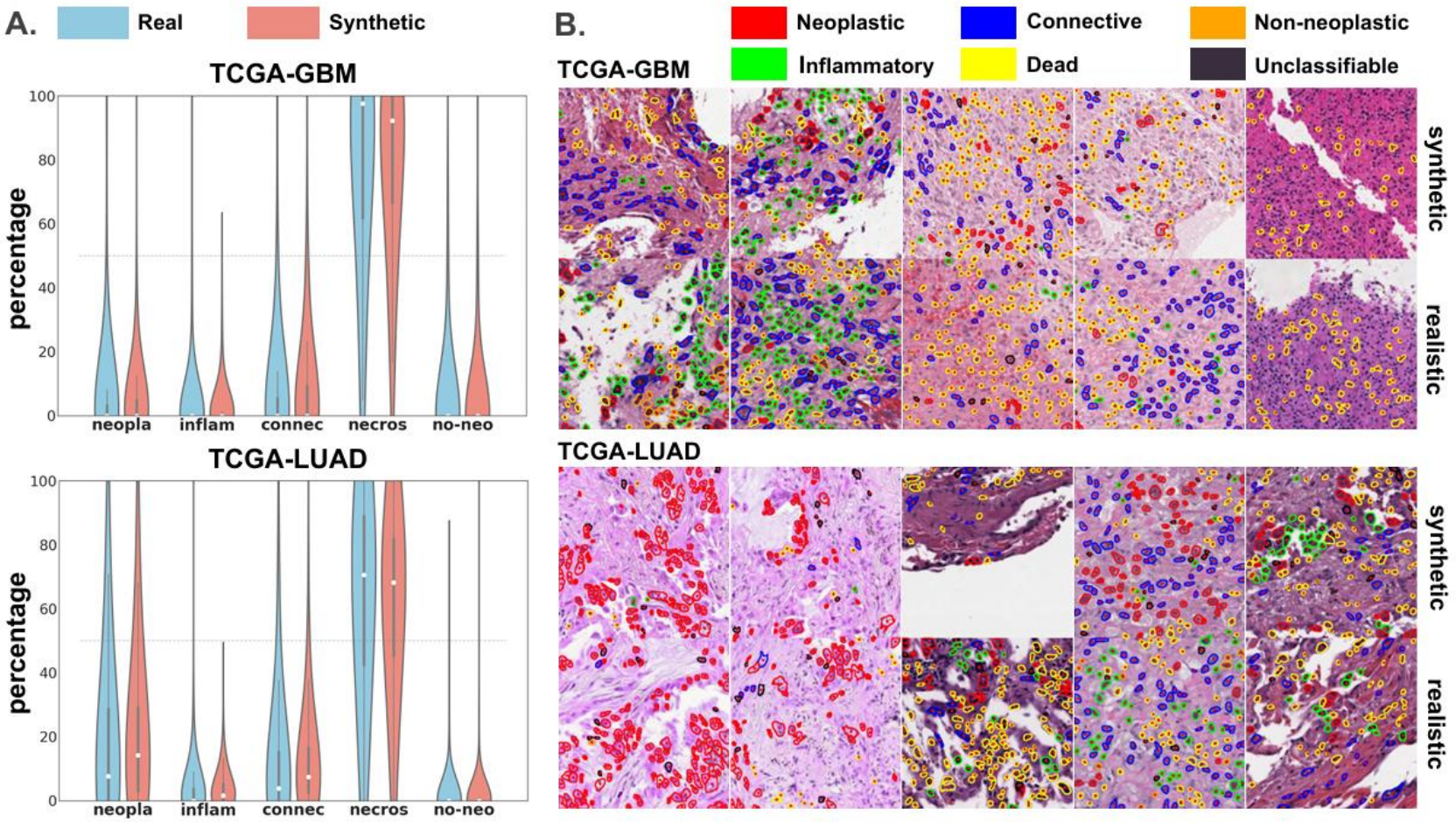
Table 4. **Performance Comparison** of MIL methods (w/ vs. w/o pretraining). Our GeneVAR generates 512 patch-wise tiles using each RNA-Seq.

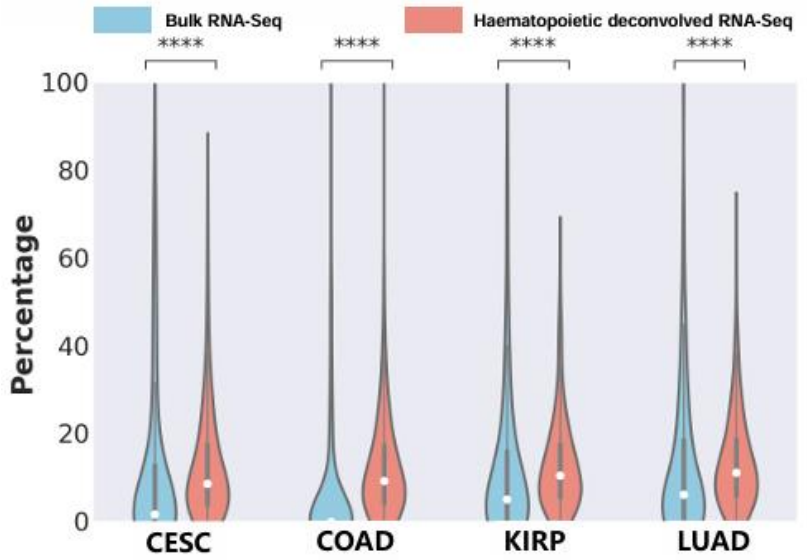
Method	ACC		F1-score		AUC	
	w/o	w/	w/o	w/	w/o	w/
TransMIL [36]	0.849	**0.877**	0.847	**0.875**	0.894	**0.934**
ACMIL [46]	0.767	**0.863**	0.749	**0.858**	0.817	**0.938**
WiKG [25]	0.767	**0.822**	0.753	**0.818**	0.820	**0.917**
MambaMIL [42]	0.836	**0.918**	0.833	**0.917**	0.925	**0.941**

3. For downstream evaluation (tile-level and slide-level classification), we train tile-level and WSI-level classifiers on synthetic tiles and report F1-score, accuracy (ACC), and AUC.

3.Qualitative Interpretability

CVPR JUNE 3-7, 2026 DENVER COLORADO





Synthetic Vs. Real tiles. Panel A. Cell-type distributions in real and synthetic tiles across TCGA cohorts. Panel B. HoverNet visualizations showing consistent cellular composition between synthetic and real samples.

Lymphocyte Distribution Comparison between WSIs generated from bulk RNA-Seq and those generated from hematopoietic-deconvolved expression. Lymphocyte proportions show consistently increased trends across TCGA-CESC, TCGA-COAD, TCGA-KIRP and TCGA-LUAD.

4.Conclusion

In this work, we introduced GeneVAR, an autoregressive Gene-to-WSI tile synthesis model that unites multi-stage transcriptomic conditioning with causality-aware modeling. By reformulating synthesis as an iterative, coarse-to-fine autoregressive process, GeneVAR continually reinforces transcriptomic signals, mitigating signal decay and ensuring cross-scale consistency. At its core lies the Causal Mean-Flow module, guided by RNA-Seq embeddings, which uses counterfactual interventions to suppress spurious variation and enforce causal fidelity in morphology. Beyond benchmarking, GeneVAR serves as a controllable tool for probing gene-morphology relationships under transcriptomic perturbations, opening new avenues for integrative studies in computational pathology.