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

Jianwei Zhao[1], Fan Yang[2], Xin Li[2,*], Qiang Zhai[3], Ao Luo[4], Ziqi Ren[5], Zhicheng Jiao[6], and Hong Cheng[1]
[1]UESTC, [2]Alpaca AI Lab, [3]SICAU, [4]SWJTU, [5]XDU, [6]Brown

Abstract

This appendix provides a comprehensive set of supplementary materials that extend and deepen the contributions presented in the main paper. We begin by revisiting the theoretical underpinnings of MeanFlow, offering a detailed derivation that clarifies its role in modeling average velocity and its integration into GeneVAR. Next, we describe the construction of counterfactual samples and the implementation details of training, which together provide greater transparency and reproducibility of our method. Beyond methodology, we report expanded experimental results, including extensive ablation studies, complete quantitative evaluations, and large-scale comparisons with state-of-the-art baselines. Finally, we present supplementary qualitative visualizations that highlight the diversity and realism of the synthesized tiles across multiple cancer types. Taken together, these materials serve to reinforce the rigor of our approach, demonstrate its robustness across varied conditions, and provide deeper insights into its practical implications.

1. Additional Analysis: Generative Updates through Causal Inference

We model the generative update process with a causal graph (Fig. 1) grounded in Pearl's structural causal model [10]. The graph consists of four nodes: R (RNA sequence profile), X (coarse semantics), Y (fine-grained semantics), and C (confounders).

$R \rightarrow X$, $R \rightarrow Y$. These pathways show that transcriptomic programs inferred from RNA-Seq profiles exert a systematic regulatory influence on WSI morphology. They shape the emergence of coarse structural patterns (X), refine fine-grained cellular semantics (Y).

$C \rightarrow X$, $C \rightarrow Y$. These pathways indicate that confounders such as tumor purity, staining variability, and acquisition-related factors substantially influence WSI morphology, thereby introducing systematic biases in downstream analyses [7].

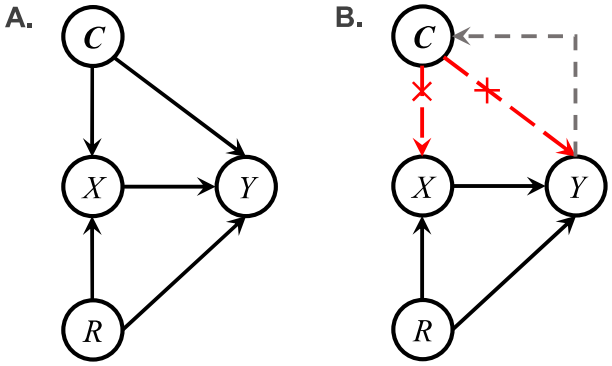


Figure 1. **Visualizations of the Causal Graph** for generative updates from coarse to fine semantics. **A**. The original causal graph. **B**. The causal graph after suppressing the spurious correlation between X and Y through counterfactual training. The red dashed line indicates the non-causal pathway that is suppressed during learning, whereas the gray dashed line marks the counterfactual interventions applied in training.

$X \rightarrow Y$. This pathway indicates that r-$\mathcal{CM}$ incrementally transforms coarse semantic structures (X) into fine-grained morphological patterns (Y) under the guidance of transcriptomic signals (R).

In this causal graph, the confounder C creates a backdoor path $X \leftarrow C \rightarrow Y$ that injects spurious associations between the two variables and obstructs reliable semantic refinement. Building on this formulation, our update procedure targets a central goal: preserving the causal factors represented by X and R, while suppressing the non-causal variations introduced by C. Following the insight of [12], we generate samples that contain such non-causal factors through controlled interventions and incorporate a causally guided metric learning scheme within the average velocity fields to attenuate the backdoor influence. This causal treatment allows the model to separate biologically meaningful morphological semantics from acquisition- or purity-induced degradations.

*Corresponding author: Xin Li (*xinli_uestc@hotmail.com*)

2. Further Derivation of MeanFlow

MeanFlow [4] introduces the concept of an average velocity u, defined as the displacement between two time steps t and r, normalized by their interval. This reformulation reduces the multi-step integration of conventional flow matching to a single-step computation, thereby simplifying the generative process. Given the correlation between the average velocity and the instantaneous velocity as defined in Eq. 1:

$$(t-r)u(\mathbf{f}_k^t, \mathbf{r}_k, \mathbf{z}, t, r) = \int_r^t v(\mathbf{f}_k^\tau, \tau)\, d\tau, \tag{1}$$

differentiating both sides with respect to t, we can rearrange this formulation to obtain u,

$$u(\mathbf{f}_k^t, \mathbf{r}_k, \mathbf{z}, t, r) = v(\mathbf{f}_k^t, t) - (t-r)\frac{d}{dt}u(\mathbf{f}_k^t, \mathbf{r}_k, \mathbf{z}, t, r). \tag{2}$$

where $\frac{d}{dt}u(\mathbf{f}_k^t, \mathbf{r}_k, \mathbf{z}, t, r)$ denotes the time derivative.

By further expanding the time derivative $\frac{d}{dt}u$ via the chain rule, u can be explicitly expressed as:

$$\begin{aligned}\frac{d}{dt}u(\mathbf{f}_k^t, \mathbf{r}_k, \mathbf{z}, t, r) = &\underbrace{\frac{d\mathbf{f}_k^t}{dt}}_{v(\mathbf{f}_k^t, t)}\partial_{\mathbf{f}_k}u + \underbrace{\frac{d\mathbf{r}_k}{dt}}_{0}\partial_{\mathbf{r}_k}u + \underbrace{\frac{d\mathbf{z}}{dt}}_{0}\partial_{\mathbf{z}}u \\ &+ \underbrace{\frac{dt}{dt}}_{1}\partial_t u + \underbrace{\frac{dr}{dt}}_{0}\partial_r u, \\ &= v(\mathbf{f}_k^t, t)\partial_{\mathbf{f}_k}u + \partial_t u.\end{aligned} \tag{3}$$

In conclusion, the total derivative can be written as a Jacobian–vector product (JVP), where $[\partial_{\mathbf{f}_k}u, \partial_{\mathbf{r}_k}u, \partial_{\mathbf{z}}u, \partial_t u, \partial_r u]$ constitutes the Jacobian matrix of $u(\mathbf{f}_k^t, \mathbf{r}_k, \mathbf{z}, t, r)$ along the tangent vector $[v, 0, 0, 1, 0]$. Since the ground-truth function u is inaccessible, we employ a network Φ to approximate it. Specifically, in Eq. 3, u is replaced by u_ϕ learned via Φ, and the target average velocity u_{tgt} is ultimately given by:

$$\begin{aligned}u_{\text{tgt}} &= v(\mathbf{f}_k^t, t) - (t-r)(v(\mathbf{f}_k^t, t)\partial_{\mathbf{f}_k}u_\phi + \partial_t u_\phi), \\ \text{where} \quad u_\phi &= \Phi(\mathbf{f}_k^t, \mathbf{r}_k, \mathbf{z}, t, r).\end{aligned} \tag{4}$$

Here, $\partial_{\mathbf{f}_k}u_\phi$ and $\partial_t u_\phi$ can be efficiently computed using the `jvp` interface in PyTorch. The procedure for training Φ with the target u_{tgt} in Eq. 4 is summarized in Alg. 1.

3. Counterfactual Sample Construction

Effective interventions should perturb non-causal factors while preserving causal content, thereby enabling meaningful causal inference. To this end, we empirically apply three interventions to the ground-truth tiles to construct counterfactual samples, as illustrated in Fig. 2.

Color Anomaly. Following CWNet [12], we apply a color degradation procedure to the original tile X in order to suppress color disturbances. The degradation is defined as:

$$X^a = \Delta H(X) + \Delta S(X) + \sum_{K\in\{R,G,B\}} \Delta K(X) + \epsilon, \tag{5}$$

where H, S, and K denote the hue, saturation, and RGB channel offsets, respectively.

Contrast Adjustment. To prevent spurious semantic enhancement caused by contrast variation, we introduce a contrast adjustment intervention:

$$X^c = \sum_{K\in\{R,G,B\}} \Big[\alpha_K \cdot (X_K - \mu_K) + \mu_K\Big] + \epsilon, \tag{6}$$

where α_K and μ_K are the contrast coefficient and channel-wise mean of the K-th channel.

Sharpening Operation. Similarly, to mitigate spurious semantic enhancement introduced by local spatial frequency distributions, we design a sharpening intervention:

$$X^s = X + \alpha \cdot (X - \text{Blur}(X)) \cdot k, \tag{7}$$

where α and k control the sharpening intensity and lightness coefficient, respectively, and $\text{Blur}(\cdot)$ denotes Gaussian blurring.

4. Implementation Details

β-VAE. The value of β follows the setting of [1] and is fixed at 0.005. This configuration consistently stabilizes optimization and yields a disentangled latent structure that supports the subsequent generative updates.

MSVQ $\mathcal{Q}$. All scales share a unified codebook $\mathcal{V} \in \mathbb{R}^{4096\times 32}$, comprising 4096 entries of 32-dimensional vectors, with the number of discrete token maps per image fixed at $K = 10$. A DINOv2-based encoder [8] is used to obtain continuous latent representations, followed by a decoder that reconstructs images from the quantized *token maps*. Subsequently, we randomly sample approximately 31 tiles per WSI associated with each RNA-Seq profile, constructing a total of 50,000 tiles, yielding an rFID of 2.11.

Causal MeanFlow. r-$\mathcal{CM}$ adopts the SiT architecture [9] for the network Φ, with model depth and feature dimension set to 6 and 768, respectively. The hyperparameter α is fixed at 0.1, and λ is uniformly sampled from $[0.8, 1.2]$. r-$\mathcal{CM}_\Phi$ is trained for 100 epochs using 50 tiles per WSI. Additionally, we randomly sample approximately 31 tiles per WSI associated with each RNA-Seq profile, yielding a total of 50,000 tiles for rFID evaluation.

RNA-Guided Masked Autoregression. The VAR transformer follows the standard architecture with depth $d = 16$, head count $h = d$, and width $w = 64h$. Training is performed using an AdamW optimizer with initial learning rate

Algorithm 1 Causal MeanFlow: Training

Note: `jvp` interface in PyTorch returns the function output and $\frac{du}{dt}$.

```
# fn(f_k, r_k, z, t, r): function to predict u
# f_k: gt quantinized feature at k scale
# r_k: token embeddings map at k scale
# z: compact molecular prior from RNA_Seq
# t,r: two sampled timestamps
# Counterfactual batch:
# f_k^l: r_k, f_k^a: color anomaly, f_k^c: contrast, f_k^s: sharpening

N = len([f_k^l, f_k^a, f_k^c, f_k^s])
t, r = sample_t_r()
ϵ = randn_like(f_k)

f_k^t = (1 - t) * f_k + t * ϵ
v_k^t = ϵ - f_k^t

u, du/dt = jvp(fn, (f_k, r_k, z, t, r), (v_k^t, 0, 0, 1, 0))
u_tgt = v_k^t - (t - r) * du/dt
u_tgt_m = ||u_tgt||

for f_k^j in {f_k^l, f_k^a, f_k^c, f_k^s}:
    ϵ_j = randn_like(f_k^j)
    f_k^{j,t} = (1-t) * f_k^j + t * ϵ_j
    f_k^{j,r} = (1-r) * f_k^r + r * ϵ_j
    d_j = f_k^{j,t} - f_k^{j,r}
    norm(d_j) = d_j / ||d_j||
    u_tgt^j = λ_j * u_tgt_m * norm(d_j)

C_u = {u_tgt^l, u_tgt^a, u_tgt^c, u_tgt^s}
error = u - sg(u_tgt)
for u_tgt^n in C_u:
    error_inven += (u - sg(u_tgt^n))

loss = metric(error) - α/N * metric(error_inven)
```

10^{-4}, batch size 256, $\beta_1 = 0.9$, $\beta_2 = 0.95$, and weight decay 0.05. After pretraining r-$\mathcal{CM}$, we freeze its parameters and integrate it into the training of $\mathcal{P}_\Theta$ for 200 epochs using 200 tiles per WSI. The MSVQ module $\mathcal{Q}$ is trained jointly on the same tile sampling protocol.

Code Availability. The implementation of GeneVAR, including training scripts and pretrained models, will be made publicly available upon publication.

5. Extended Experimental Results

Comprehensive Cell Distribution Comparison. Tab. 1 reports the complete cell distributions of five cell types across five datasets. For each dataset, we synthesize 2,000 tiles using all RNA-Seq profiles and randomly sample 2,000 real tiles. HoverNet [5], trained on the PanNuke dataset [3], is employed to detect neoplastic, inflammatory, connective, dead, and non-neoplastic cell types. For each dataset, cell distributions are first quantified at the tile level, and then aggregated to obtain overall statistics across the 2,000 tiles. The results demonstrate that synthetic tiles generated by GeneVAR better capture realistic cell-type distributions and further improve the morphological fidelity of synthesized tissue. Representative visualizations are provided in Fig. 3.

Comprehensive Tile Classification Analysis. Comprehensive results under different proportions p and q are reported in Tab. 2. For GeneVAR, classification accuracy does not decline even as the fraction of synthetic tiles steadily increases within the 5000 mixed tiles, whereas all competing methods exhibit significant performance degradation. Moreover, classification accuracy continues to improve with the progressive incorporation of synthetic tiles

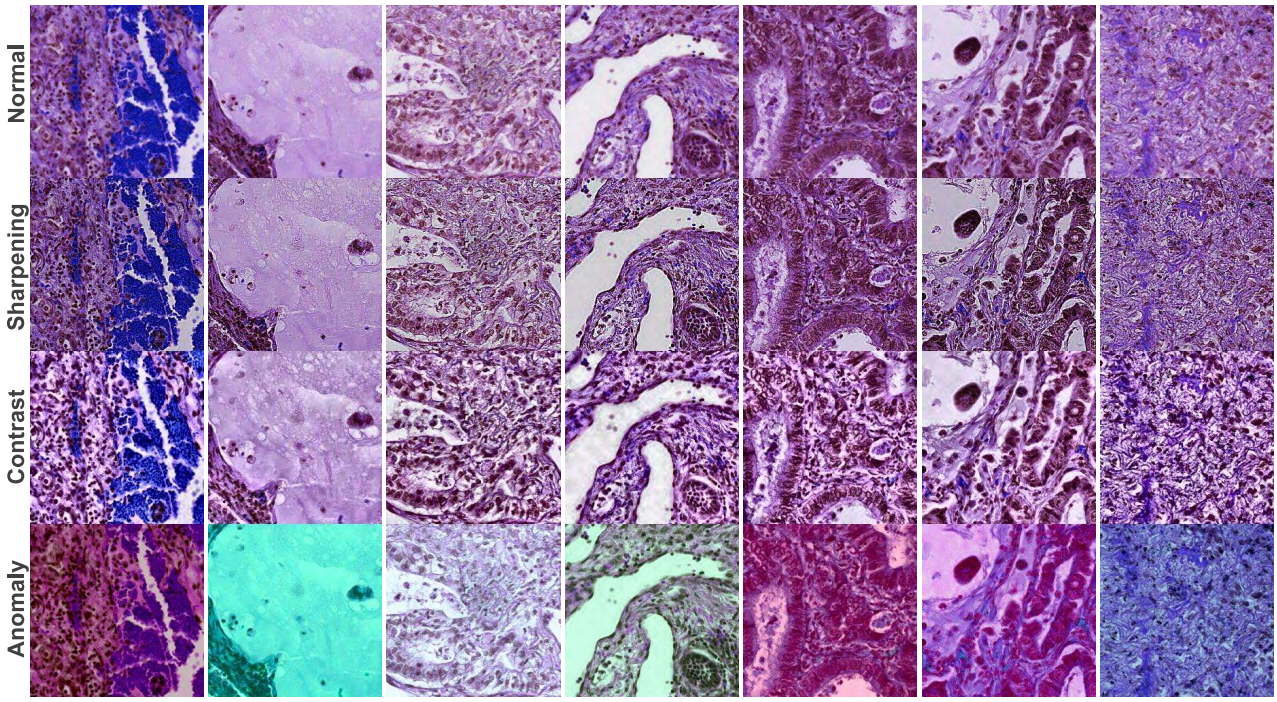


Figure 2. **Visualizations of counterfactual samples**. Our degradation procedures perturb non-causal factors while maintaining the morphology information.

Table 1. **Comparison of various cell distributions.** We systematically analyze the distributions of cell populations at the tile level across five cancer types, reporting both the mean and standard deviation (mean±std) of the corresponding cell proportion.

Dataset	Method	Neoplastic	Inflammatory	Connective	Dead	Non-Neoplastic
GBM	VAR	9.96% ± 21.99	1.64% ± 5.46	15.87% ± 26.84	69.55% ± 34.59	2.95% ± 12.01
	Ours	5.04% ± 12.07	2.80% ± 7.60	8.97% ± 17.15	78.67% ± 26.91	4.49% ± 12.70
	Real	6.04% ± 16.29	2.82% ± 8.86	8.39% ± 19.19	77.49% ± 30.90	5.24% ± 16.06
CESC	VAR	30.92% ± 31.89	1.44% ± 5.20	19.04% ± 26.68	47.47% ± 32.18	1.10% ± 6.53
	Ours	27.19% ± 24.82	2.05% ± 5.52	13.31% ± 19.57	56.67% ± 26.64	0.77% ± 3.21
	Real	25.67% ± 28.43	2.73% ± 9.43	12.21% ± 21.66	58.77% ± 30.56	0.60% ± 2.95
LUAD	VAR	25.06% ± 27.38	2.99% ± 6.84	15.23% ± 20.63	55.99% ± 30.00	0.71% ± 3.42
	Ours	20.20% ± 22.78	3.95% ± 5.86	13.16% ± 15.14	61.10% ± 26.56	1.56% ± 6.64
	Real	19.32% ± 25.21	3.57% ± 6.86	12.46% ± 19.40	63.12% ± 29.60	1.52% ± 7.04
KIRP	VAR	20.59% ± 25.29	3.52% ± 8.01	9.22% ± 18.70	60.78% ± 26.97	2.72% ± 8.30
	Ours	22.61% ± 23.44	2.91% ± 4.64	9.36% ± 13.89	63.32% ± 25.53	1.78% ± 4.76
	Real	23.07% ± 26.71	2.44% ± 5.78	8.04% ± 16.01	64.43% ± 29.85	2.00% ± 7.23
COAD	VAR	33.37% ± 33.08	2.01% ± 7.29	27.39% ± 28.00	36.08% ± 29.55	1.13% ± 6.71
	Ours	38.14% ± 28.25	2.81% ± 5.00	20.09% ± 20.19	38.25% ± 24.69	0.69% ± 2.37
	Real	35.54% ± 32.95	3.32% ± 7.77	19.87% ± 24.85	40.49% ± 29.66	0.76% ± 3.45

during pretraining. Notably, GeneVAR achieves larger gains than the two SOTA baselines. Classification models are trained for 50 epochs, with each pretraining stage fixed at 20 epochs.

Table 2. **Tile classification performance** under different substitution ratios p of real tiles with synthetic ones, and varying proportions q of synthetic tiles used for pretraining. Four different proportions for the pretraining 25% (1,250 samples), 50% (2,500 samples), 75% (3,750 samples) and 100% (5,000 samples). All experiments are conducted under a 5-fold cross-validation (CV) protocol to guarantee the stability and reliability of the results.

Method	p=0.0		p=0.25		p=0.50		p=0.75	
	ACC	F1-score	ACC	F1-score	ACC	F1-score	ACC	F1-score
RNA-CDM	0.573 ± 0.020	0.556 ± 0.022	0.563 ± 0.024	0.520 ± 0.028	0.533 ± 0.014	0.516 ± 0.036	0.492 ± 0.016	0.472 ± 0.046
VAR	0.573 ± 0.034	0.556 ± 0.015	0.576 ± 0.020	0.565 ± 0.026	0.553 ± 0.022	0.561 ± 0.017	0.521 ± 0.030	0.510 ± 0.035
Ours	**0.579 ± 0.032**	**0.570 ± 0.039**	**0.580 ± 0.020**	**0.569 ± 0.020**	**0.590 ± 0.037**	**0.585 ± 0.042**	**0.592 ± 0.029**	**0.588 ± 0.031**
	q=25%		q=50%		q=75%		q=100%	
RNA-CDM	0.612 ± 0.021	0.606 ± 0.020	0.618 ± 0.026	0.612 ± 0.030	0.635 ± 0.017	0.637 ± 0.018	0.650 ± 0.021	0.641 ± 0.031
VAR	0.628 ± 0.025	0.625 ± 0.024	0.661 ± 0.020	0.662 ± 0.019	0.695 ± 0.013	0.698 ± 0.010	0.708 ± 0.011	0.709 ± 0.010
Ours	**0.656 ± 0.023**	**0.654 ± 0.022**	**0.722 ± 0.023**	**0.722 ± 0.020**	**0.747 ± 0.016**	**0.747 ± 0.015**	**0.767 ± 0.015**	**0.766 ± 0.017**

Table 3. **Performance comparison of MIL methods** with and without pretraining across four SOTA models. Shaded cells indicate results with pretraining.

Method	ACC		F1-score		AUC	
	w/o	w/	w/o	w/	w/o	w/
ViT pre-trained with MoCo V3						
TransMIL	0.849	0.877	0.847	0.875	0.894	0.934
ACMIL	0.767	0.863	0.749	0.858	0.817	0.938
WiKG	0.767	0.822	0.753	0.818	0.820	0.917
MambaMIL	0.836	0.918	0.833	0.917	0.925	0.941
CTransPath pre-trained with SRCL						
TransMIL	0.822	0.849	0.821	0.843	0.862	0.900
ACMIL	0.808	0.849	0.800	0.847	0.836	0.936
WiKG	0.767	0.836	0.745	0.829	0.876	0.912
MambaMIL	0.781	0.849	0.766	0.844	0.867	0.912

Comprehensive WSI Classification Analysis. WSI classification follows the standard multi-instance learning framework, where all tiles extracted from an individual WSI form a tile-wise bag associated with a slide-level label: 0 for microsatellite stability (MSS) and 1 for microsatellite instability (MSI). Tile features are first extracted by a designated feature extractor and then fed into a classification model to predict the slide-level label. To assess the clinical utility of synthetic tiles, we adopt two tile feature extractors—ViT [2] pre-trained with MoCoV3 and CTransPath [11] pre-trained with SRCL—together with four state-of-the-art WSI classification models. We evaluate on the COAD MSI dataset [6], which includes 298 MSS and 66 MSI slides from 360 patients. Specifically, 292 slides for training and 72 slides for testing. For each RNA-Seq profile, GeneVAR generates 512 synthetic tiles to construct tile-wise bags for classifier pretraining. As shown in Tab. 3, all classifiers consistently achieve significant gains across both feature extractors.

Table 4. **Ablation on the injection timing of r-$\mathcal{CM}$.** "Time" denotes the relative training time per epoch, while "Inject." indicates the number of r-$\mathcal{CM}$ injection operations. Shaded cell indicates the best FID.

K_m	FID ↓	Time	Inject.
7	14.13	1.51×	4
8	12.78	1.30×	3
9	12.95	1.00×	2
10	15.60	0.86×	1

5.1. Additional Ablation Results

Ablation for K_m. Injecting r-$\mathcal{CM}$ at every autoregressive step incurs considerable computational redundancy. To address this, we analyze token maps across scales $\{1, 2, 3, 4, 5, 6, 8, 10, 13, 16\}$ and find that the final two token maps account for 62.5% of the total sequence length (680). Based on this observation, the initial value of K_m is empirically set to 9. We then evaluate GeneVAR under a range of K_m values. As reported in Tab. 4, setting $K_m = 8$ yields only a marginal improvement in FID (0.14) but comes at the cost of a 1.30× increase in per-epoch training time compared to $K_m = 9$.

Ablation for α. We carefully investigate the effect of the causal regularization strength controlled by α, where $\alpha = 0$ corresponds to disabling the causal constraint. As depicted in Tab. 5a, increasing α encourages the model to push the predicted average velocity away from those derived from counterfactual samples, thereby improving the semantic and biological consistency of the *coarse-to-fine* refinement. However, an excessively large α causes the model to overemphasize the repulsion from counterfactual veloci-

Table 5. **Ablation on the strength of causal regularization (a) and stochastic factor λ (b).** "rFID" denotes the reconstruction quality, while "FID" indicates the overall performance within the autoregressive framework. Shaded cell indicates the best results.

α	rFID ↓	FID ↓
0.00	4.41	15.23
0.01	4.08	15.17
0.05	3.76	14.30
0.10	2.03	12.95
0.15	2.73	13.54
0.20	3.38	14.97

(a) Ablation on α.

λ	rFID ↓	FID ↓
[1.0, 1.0]	3.85	14.03
[0.9, 1.0]	3.64	13.78
[0.9, 1.1]	3.29	13.90
[0.8, 1.1]	2.81	13.43
[0.8, 1.2]	2.03	12.95
[0.9, 1.2]	2.46	13.61

(b) Ablation on λ.

ties, making it difficult for the model to learn the true underlying average velocity. Subsequently, $\alpha = 0.10$ achieves the best performance, consistently across both rFID and FID.

Ablation for λ. We meticulously evaluate the effect of broadening the sampling range of λ. As the interval expands from the fixed value of 1.0 to [0.8, 1.2], the model exhibits consistent performance improvements, 1.83 (rFID) and 1.08 (FID). However, further enlarging the range leads to performance degradation. These observations suggest that an appropriately bounded stochastic factor λ, which scales the magnitude, enhances the robustness of causal learning.

5.2. Supplementary Visualization Results

To further illustrate the generative capacity of GeneVAR, we provide additional qualitative results in Fig. 4. These synthetic H&E-stained tiles are generated across five representative cancer types (GBM, CESC, LUAD, KIRP, and COAD). The examples demonstrate that GeneVAR is capable of producing diverse histological patterns that remain visually realistic while preserving disease-specific morphology. Such results highlight the robustness of GeneVAR in synthesizing tissue structures that are consistent with biological priors, thereby complementing the quantitative evaluations reported in the main paper.

5.3. Out-of-Distribution Generation

To further examine the generalization capacity of GeneVAR beyond the distributions encountered during training, we incorporate gene expression profiles obtained from the GEO (Gene Expression Omnibus) repository. These profiles originate from studies that differ in tissue preparation, sequencing protocols, and sample characteristics, placing them entirely outside the model's training domain. The qualitative results in Fig. 6 provide a visual assessment of how GeneVAR responds to such distributional shift. We observe that the model continues to produce coherent and biologically plausible morphological patterns, suggesting that the learned gene–morphology mapping is not limited to the specific conditions of the training set but extends to previously unseen datasets.

5.4. Impact of Deconvolved Gene Expression

Additionally, we further examined whether generating tiles from deconvolved gene expression would produce measurable differences in the representation of specific cell types within the synthetic tiles. To carry out this comparison, we used the haematopoietic deconvolved RNA-seq profiles provided in [1] and generated a corresponding set of synthetic tiles based on these deconvolved gene signals. When inspecting the tiles produced from haematopoietic gene expression, we observed the expected increase in the percentage of lymphocytes, which reflects the biological information contained in the deconvolved profiles. This trend is clearly visible in Fig. 5.

References

[1] Francisco Carrillo-Perez, Marija Pizurica, Yuanning Zheng, Tarak Nath Nandi, Ravi Madduri, Jeanne Shen, and Olivier Gevaert. Generation of synthetic whole-slide image tiles of tumours from rna-sequencing data via cascaded diffusion models. *Nature Biomedical Engineering*, 2025. 2, 6

[2] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021. 5

[3] Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benes, Simon Graham, Mostafa Jahanifar, Syed Ali Khurram, Ayesha Azam, Katherine Hewitt, and Nasir Rajpoot. Pannuke dataset extension, insights and baselines. *arXiv preprint arXiv:2003.10778*, 2020. 3

[4] Zhengyang Geng, Mingyang Deng, Xingjian Bai, J Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv*, 2025. 2

[5] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical image analysis*, 2019. 3, 8

[6] Jakob Nikolas Kather, Alexander T Pearson, Niels Halama, Dirk Jäger, Jeremias Krause, Sven H Loosen, Alexander Marx, Peter Boor, Frank Tacke, Ulf Peter Neumann, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nature Medicine*, 2019. 5

[7] Tiancheng Lin, Zhimiao Yu, Hongyu Hu, Yi Xu, and Chang-Wen Chen. Interventional bag multi-instance learning on whole-slide pathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19830–19839, 2023. 1

[8] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv*, 2023. 2

[9] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 2

[10] Bernhard Schölkopf. Causality for machine learning. In *Probabilistic and causal inference: The works of Judea Pearl*, pages 765–804. 2022. 1

[11] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 2022. 5

[12] Tongshun Zhang, Pingping Liu, Yubing Lu, Mengen Cai, Zijian Zhang, Zhe Zhang, and Qiuzhan Zhou. Cwnet: Causal wavelet network for low-light image enhancement. *arXiv*, 2025. 1, 2

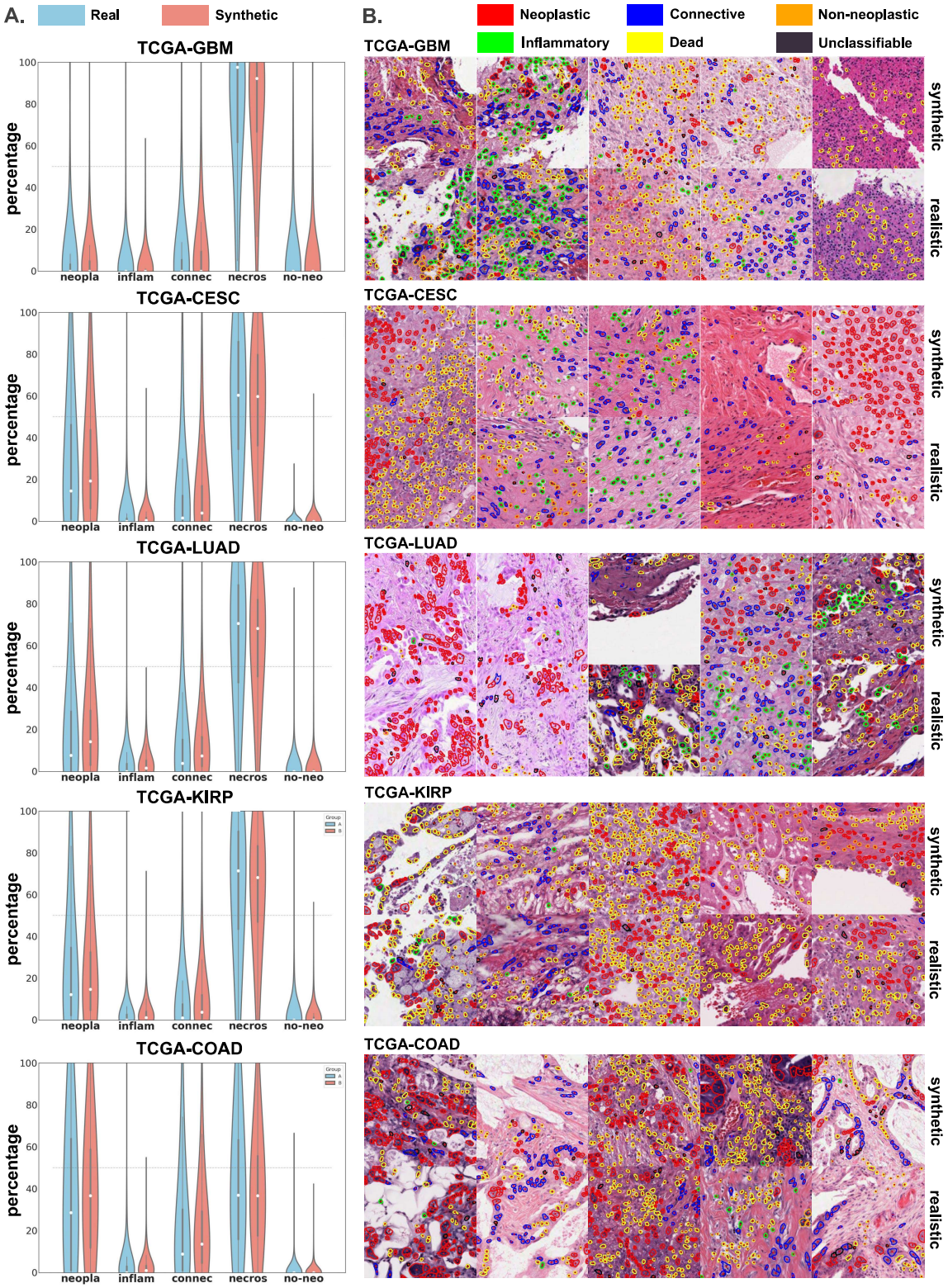
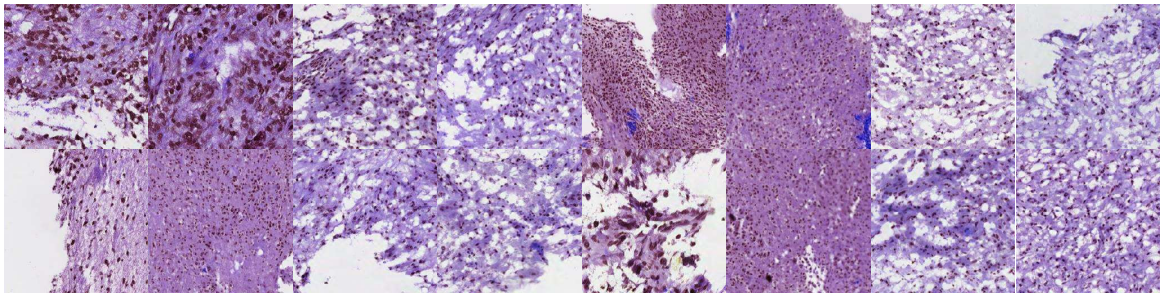
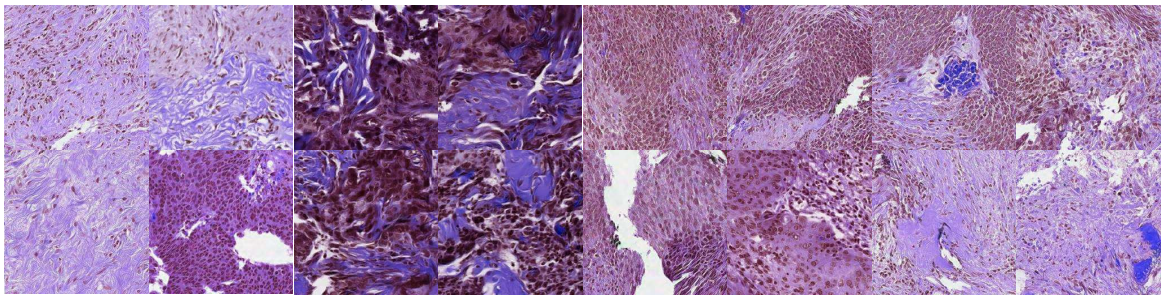


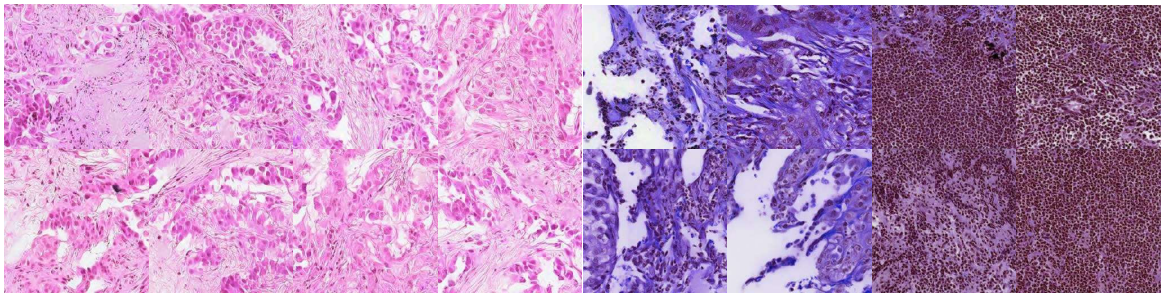
Figure 3. **Panel A.** Comparison of cell-type distributions between synthetic and real tiles across five cancer types. **Panel B.** Visualization of cells detected by HoverNet [5], showing neoplastic, inflammatory, connective, dead, and non-neoplastic populations in both synthetic and real samples.



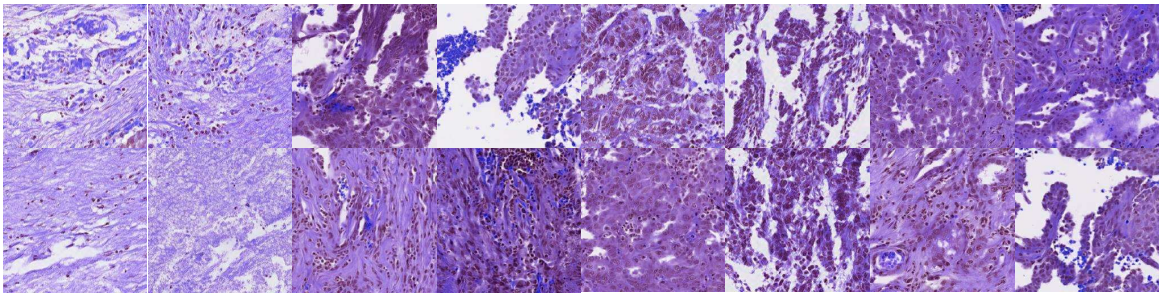
(a) Synthetic H&E-stained WSI tiles from GeneVAR for GBM.



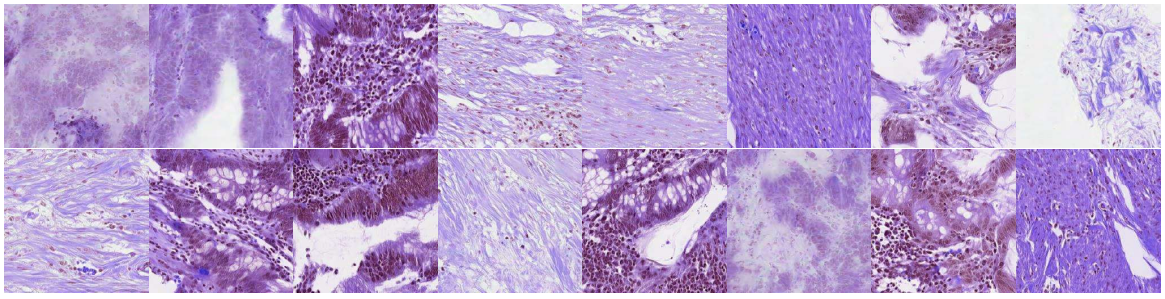
(b) Synthetic H&E-stained WSI tiles from GeneVAR for CESC.



(c) Synthetic H&E-stained WSI tiles from GeneVAR for LUAD.



(d) Synthetic H&E-stained WSI tiles from GeneVAR for KIRP.



(e) Synthetic H&E-stained WSI tiles from GeneVAR for COAD.

Figure 4. **Synthetic H&E-stained WSI tiles** generated by GeneVAR across five cancer types: (a) GBM, (b) CESC, (c) LUAD, (d) KIRP, and (e) COAD. The results demonstrate that GeneVAR can synthesize diverse and realistic histological patterns.

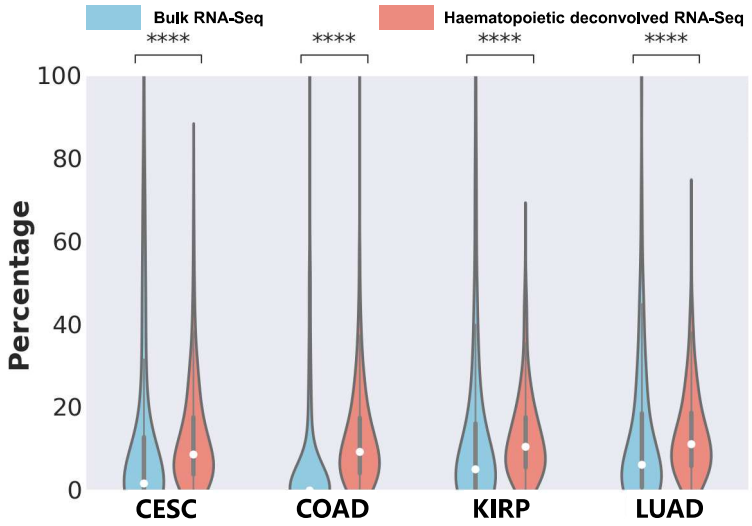
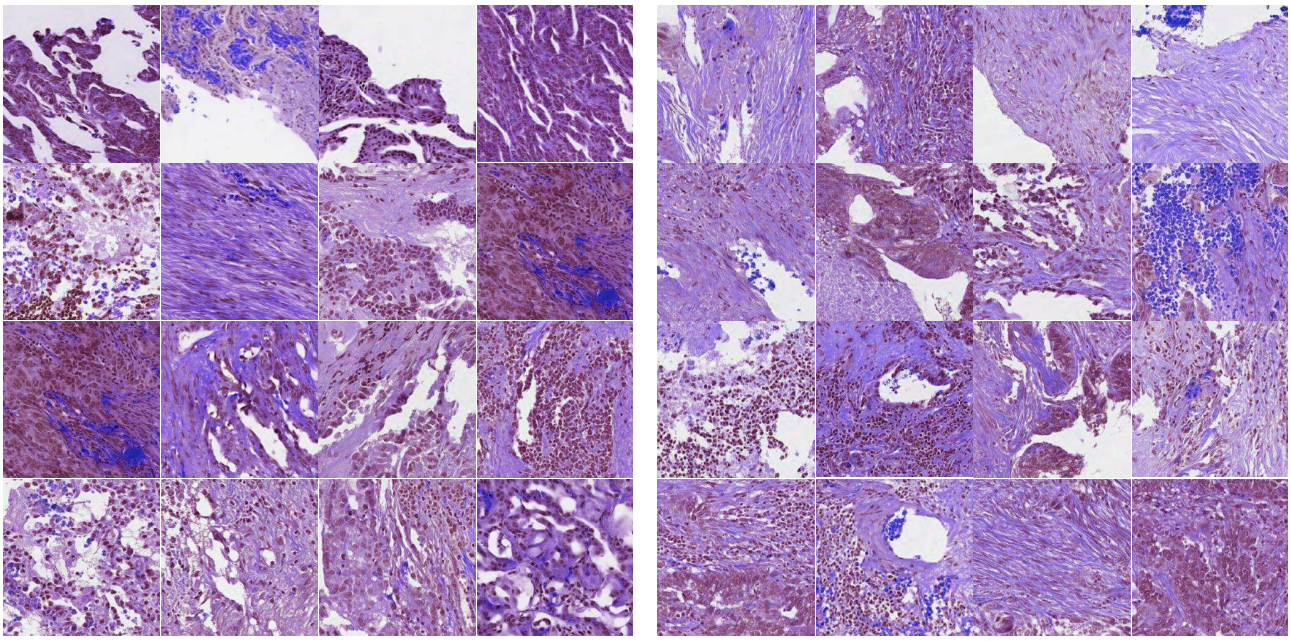


Figure 5. **Lymphocyte Distribution Comparison** between WSIs generated from bulk RNA-Seq and those generated from hematopoietic-deconvolved expression. Lymphocyte proportions show consistently increased trends across TCGA-CESC, TCGA-COAD, TCGA-KIRP, and TCGA-LUAD, with differences between groups reaching extremely high statistical significance (p-value < 0.0001, $\star\star\star\star$).



(a) Synthetic H&E-stained WSI tiles from GeneVAR using GSE226069.

(b) Synthetic H&E-stained WSI tiles from GeneVAR using GSM1228184.

Figure 6. **Synthetic H&E-stained WSI tiles** generated by GeneVAR using gene expression profiles out of the training distribution, specifically lung cancer RNA-Seq GEO: GSE226069 and colorectal cancer RNA-Seq GEO: GSM1228184.