Layout Before Pixels: Topology-Anchored Transcriptome-to-Histology Generation

Jianwei Zhao[1] **Xin Li**[2,*] **Fan Yang**[2] **Qiang Zhai**[3] **Ao Luo**[4] **Hong Cheng**[1]
[1]UESTC [2]Alpaca AI Lab [3]SICAU [4]SWJTU

Abstract

Transcriptome-conditioned whole-slide image (WSI) synthesis aims to decode the complex mapping from molecular states to histologic phenotypes. Existing generative paradigms predominantly rely on a direct black-box transcriptome-to-image mapping, tasking a single generator with the simultaneous interpretation of transcriptome signals, inference of multicellular spatial organization, and rendering of fine-grained textures. This formulation, however, is structurally underconstrained; the underlying tissue architecture, the critical substrate linking gene expression to morphology, remains latent and only weakly supervised by pixel-level objectives. We present TOPOSCAPE, a topology-anchored framework for transcriptome-to-histology generation under a *Layout-Before-Pixels* principle. Instead of synthesizing pixels directly from transcriptomic embeddings, TOPOSCAPE first resolves molecular states into explicit, multi-class cellular topologies via an RNA-guided Topological Prior Generator, reinforced by persistence-based topological regularization. These resolved topologies are further reparameterized into a suite of topology-derived controls, including structure-aware initialization, continuous density fields, and distance representations, which guide a hierarchical, frequency-decoupled flow-matching trajectory. By introducing multicellular organization as an explicit structural mediator, TOPOSCAPE allows molecular states to deterministically shape both the macroscopic tissue architecture and the microscopic generative process. Across five TCGA benchmarks, TOPOSCAPE consistently outperforms state-of-the-art methods in generative fidelity, biological realism, and downstream predictive utility. All models and code are available at `https://github.com/JWZhao-uestc/TopoScape`.

1 Introduction

The rapid growth of digital pathology has spurred increasing interest in transcriptome-conditioned histology tile synthesis, which aims to generate WSI tiles from bulk RNA-sequencing profiles [3, 6]. This problem is scientifically compelling because it provides a direct route to study how gene-expression programs are expressed through tissue morphology, cellular composition, and spatial organization [4, 41, 16]. It is also practically valuable, with potential applications in molecularly informed pathology, synthetic data augmentation, and privacy-conscious data sharing when large paired molecular–histology cohorts are unavailable, costly, or difficult to release [20, 11]. Yet this task is fundamentally difficult. Transcriptomic inputs are highly abstract and high-dimensional, whereas their mapping to histologic appearance is intrinsically *one-to-many*, requiring the model to recover biologically consistent multicellular organization together with realistic visual detail.

Pioneering efforts have demonstrated the feasibility of generating histology tiles from bulk RNA-sequencing profiles, including gene-conditioned GAN-based frameworks [2], cascaded diffusion-based models [3], and recent autoregressive approaches such as GeneVAR [57]. Meanwhile, GeneFlow [45] employs rectified flow to synthesize histology images from single-cell gene expressions.

*Corresponding Author: xinli_uestc@hotmail.com

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

These methods typically inject compressed transcriptomic embeddings directly into a monolithic image generator, implicitly coupling molecular interpretation with visual rendering. As a result, pixel-centric training objectives bias optimization toward low-level appearance cues, such as texture and color, while leaving multicellular organization under-constrained. Consequently, without an explicit structural mediator, the model cannot fully exploit the relationship between transcriptomic signals and spatial cellular distributions, often producing images that are visually plausible yet structurally ungrounded.

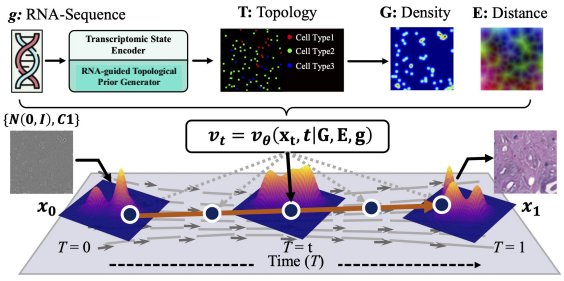


Figure 1: **Illustration of TOPOSCAPE.** Unlike direct transcriptome-to-image mapping, TOPOSCAPE first resolves RNA-seq profiles into multicellular topologies that guide flow-based histology synthesis. Our *Layout-Before-Pixels* design anchors generation to biologically grounded tissue structure.

Such structural under-specification motivates a *Layout-Before-Pixels* formulation, in which transcriptomic profiles are first resolved into an explicit organizational substrate before pixel-level synthesis begins. As shown in Fig. 1, we instantiate this principle with TOPOSCAPE, a topology-anchored transcriptome-to-histology generation framework that recasts synthesis as a structured progression from molecular state to cellular topology and ultimately to visual phenotype. Specifically, following established practice [2, 3], a transcriptomic state encoder first maps high-dimensional RNA-seq profiles into compact transcriptomic embeddings. The key departure from conventional pipelines lies in the utilization of these embeddings: rather than directly injecting them into a monolithic image generator, TOPOSCAPE decodes them into explicit multi-class cellular layouts via an RNA-guided Topological Prior Generator. This module leverages an RNA-conditioned MeanFlow [31] dynamics, which models cellular distributions as a continuous flow field optimized under persistence-based topological regularization. By enforcing these structural constraints, this module preserves cell-type-specific spatial organization, making tissue architecture an explicit generative target rather than a hidden variable left to emerge under image-level supervision.

Yet, explicitly resolving tissue architecture is only a partial solution. If the cellular layouts are treated as mere static conditioning maps, the topological guidance remains exogenous to the generative process, resulting in a weak coupling between structure and synthesis. To bridge this gap, we further reparameterize the resolved layouts into a suite of topology-derived controls, including structure-aware initial noise, Gaussian density fields, and distance representations. These controls are integrated into a Topology-Anchored Hierarchical Flow-Matching Renderer, a frequency-decoupled generator governed by a coarse-to-fine synthesis principle [35]. Specifically, the structure-aware noise *preconditions* a topology-aligned generative trajectory, while the Gaussian density fields and distance representations *regularize* low-frequency tissue organization and high-frequency morphology decoding, respectively. In this manner, topology is no longer appended as ancillary side information; instead, it intrinsically parameterizes the rendering process, steering the synthesis across initialization, semantic layout formation, and pixel-level refinement.

By coupling transcriptomic encoding, topology generation, and topology-anchored rendering, TOPOSCAPE builds a structured generative pathway from gene-expression profiles to tissue architecture and histologic appearance. Across five TCGA cancer cohorts, TOPOSCAPE consistently improves visual fidelity, biological realism, and downstream utility over direct transcriptome-to-image baselines. These results show that transcriptome-resolved cellular topology provides an effective structural mediator for molecularly grounded histology generation.

To conclude, our contributions are threefold: **i)** we formulate *Layout-Before-Pixels*, a topology-anchored generative paradigm that resolves multicellular organization as a structural prerequisite before transcriptome-conditioned image synthesis; **ii)** we propose TOPOSCAPE, a hierarchical framework that establishes a continuous generative bridge spanning the molecular-to-phenotypic continuum. By synergistically integrating RNA-guided Topological Prior Generator with a Topology-Anchored Hierarchical Flow Renderer, TOPOSCAPE transforms high-dimensional genomic profiles into biologically grounded histology via an explicit and controllable topological mediator; and **iii)** we achieve SOTA performance across five TCGA cancer benchmarks, validating the framework's

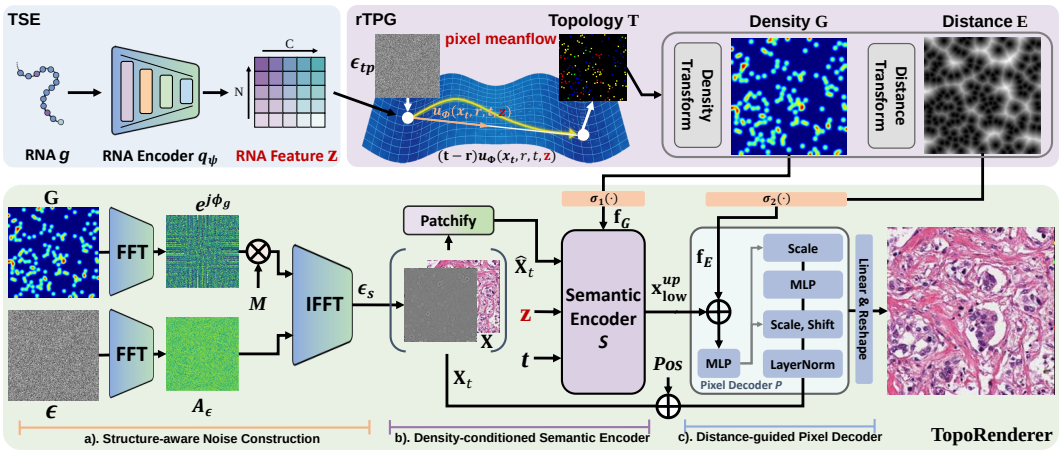


Figure 2: **Overall framework of TOPOSCAPE.** RNA-Seq embeddings are injected into a pixel-wise MeanFlow pipeline to generate topology $\mathbf{T}$ along with derived density $\mathbf{G}$ and distance $\mathbf{E}$. These are hierarchically integrated into TopoRenderer to synthesize tiles. Best viewed in color.

superior capability in generative fidelity, pathology-informed biological realism, and downstream predictive utility, thereby establishing a new standard for molecularly-aware digital pathology.

2 Related Work

Transcriptome-Conditioned Histology Generation. Transcriptome-conditioned histology generation aims to synthesize WSI tiles from molecular profiles, bridging gene expression and histologic phenotype. Compared with the widely studied image-to-transcriptomics direction [41, 19, 14, 51, 47, 5, 9, 59], this inverse problem remains underexplored. Existing methods, including gene-conditioned GANs [2], cascaded diffusion models [3], and autoregressive approaches such as GeneVAR [57], inject RNA-seq embeddings as global or iterative conditioning signals for image generation. However, multicellular organization remains implicit, although bulk transcriptomic profiles more naturally constrain cellular composition and tissue structure than pixel-level appearance. In contrast, our TOPOSCAPE resolves transcriptomic profiles into multicellular topology, introducing an explicit structural mediator for flow-based histology synthesis.

Structure-Aware Generative Modeling. Structure-aware generative models improve controllability by incorporating explicit spatial representations. In natural images, layout-to-image and controllable diffusion methods use semantic maps, masks, or adapters to guide composition [58, 53, 26, 36], while frequency-decoupled diffusion separates global structure from fine details [35]. Diffusion and flow-based models further provide continuous frameworks for conditional generation [18, 28, 29], with applications across diverse domains [55, 33, 32, 40]. In pathology, TopoCellGen [48] shows that modeling cellular topology improves biological realism. Unlike prior approaches relying on externally specified or image-derived structures, TOPOSCAPE infers multicellular topology directly from non-spatial transcriptomic profiles and uses it as a structural anchor for WSI generation.

3 Methodology

3.1 Preliminaries

Layout-Before-Pixels Formulation. We first formulate transcriptome-conditioned histology synthesis as a topology-mediated generative process. Given a bulk transcriptomic profile $\mathbf{g} \in \mathbb{R}^{17655}$, the objective is to generate a histology tile $\mathbf{X} \in \mathbb{R}^{H\times W\times 3}$ that is molecularly consistent with $\mathbf{g}$ and structurally plausible in tissue morphology. We encode $\mathbf{g}$ into a compact transcriptomic representation $\mathbf{z} \in \mathbb{R}^{200}$, which serves as the molecular state variable for subsequent generation. Rather than directly mapping $\mathbf{z}$ to pixels, we introduce cellular topology $\mathbf{T}$ as an explicit intermediate variable that spatializes the transcriptomic state before image rendering. Specifically, the generation process is organized as $\mathbf{g} \to \mathbf{z} \to \mathbf{T} \to \mathbf{X}$, where $\mathbf{T} = \mathcal{T}_\phi(\mathbf{z})$ represents transcriptome-induced cellular organization, and $\mathbf{X} = \mathcal{G}_\theta(\mathbf{T}, \mathbf{z})$ denotes topology-conditioned histology synthesis. This formulation departs from the direct transcriptome-to-image paradigm [2, 3] where tissue organization is treated as an emergent latent variable, and instead elevates multicellular organization into an explicit generative

target. Under this *Layout-Before-Pixels* principle, the molecular state $\mathbf{z}$ provides global biological conditioning, while the resolved topology $\mathbf{T}$ provides a deterministic spatial substrate. Their synergy transforms transcriptome-conditioned histology synthesis from a direct conditional image generation problem into a structured molecular-to-topological-to-phenotypic generation pathway.

Cellular Topology. We define cellular topology $\mathbf{T}$ as a point-based representation of multicellular organization within a WSI tile. Formally, $\mathbf{T} \in [0,1]^{H\times W\times K}$ denotes a K-channel cell-center topology map, where each channel corresponds to a specific cell category. For the k-th category, $\mathbf{T}^k$ encodes the spatial locations of cell centers belonging to that category. In practice, each cell center can be represented by a localized support region for stable learning, rather than by a dense pixel-wise cellular mask. This center-based formulation does not aim to recover exact cell boundaries or dense cellular morphology. Instead, it represents the spatial arrangement of category-specific cell centers as a compact topological prior. In our framework, $\mathbf{T}$ is further transformed into topology-aware structural controls, including Gaussian density maps, distance maps, and structure-aware initial noise. These derived controls anchor the subsequent flow-based WSI tile synthesis by providing spatial layout guidance, geometric conditioning, and topology-aligned initialization.

3.2 TOPOSCAPE

Based on the above formulation, we instantiate the topology-anchored generation process with TOPOSCAPE, a unified framework for transcriptome-conditioned histology synthesis. As illustrated in Fig. 2, TOPOSCAPE consists of three components: ① Transcriptomic State Encoder (**TSE**), which encodes the transcriptomic profile $\mathbf{g} \in \mathbb{R}^{17655}$ into a compact latent representation $\mathbf{z} \in \mathbb{R}^{200}$; ② RNA-guided Topological Prior Generator (**rTPG**), which predicts the cellular topology $\mathbf{T}$ conditioned on $\mathbf{z}$; and ③ Topology-Anchored Hierarchical Flow-Matching Renderer (**TopoRenderer**), which generates the histology tile $\mathbf{X}$ by jointly leveraging the molecular condition $\mathbf{z}$ and topology-derived structural controls from $\mathbf{T}$.

Transcriptomic State Encoder (TSE). The Transcriptomic State Encoder is the first component of TOPOSCAPE, responsible for converting high-dimensional molecular measurements into a compact latent state for downstream topology generation and tile synthesis. Given a transcriptomic profile $\mathbf{g} \in \mathbb{R}^{17655}$, this module produces a low-dimensional representation $\mathbf{z} \in \mathbb{R}^{200}$ that serves as the shared molecular condition throughout the generation pipeline. Following RNA-CDM [3], we implement the Transcriptomic State Encoder as a β-VAE with an inference encoder $q_\psi(\mathbf{z} \mid \mathbf{g})$ and a reconstruction decoder $h_\eta(\mathbf{g} \mid \mathbf{z})$. The encoder approximates the posterior distribution of the latent transcriptomic state, while the decoder reconstructs the input profile from the sampled latent code. The objective is defined as

$$\mathcal{L}_{\mathrm{TSE}} = -\mathbb{E}_{q_\psi(\mathbf{z}|\mathbf{g})}\left[\log h_\eta(\mathbf{g} \mid \mathbf{z})\right] + \beta\,\mathrm{KL}\left(q_\psi(\mathbf{z} \mid \mathbf{g}) \,\|\, p(\mathbf{z})\right), \tag{1}$$

where $p(\mathbf{z}) = \mathcal{N}(0, I)$ is the isotropic Gaussian prior and β balances profile reconstruction with latent regularization. After training, the latent code $\mathbf{z}$ serves as the shared transcriptomic state for **rTPG** and **TopoRenderer**.

RNA-guided Topological Prior Generator (rTPG). Given the transcriptomic state $\mathbf{z}$ from TSE, rTPG resolves molecular information into an explicit cell-center topology $\mathbf{T} \in [0,1]^{H\times W\times K}$. We implement this module through RNA-conditioned pixel MeanFlow dynamics, modeling topology formation as continuous pixel-space transport conditioned on $\mathbf{z}$.

a) RNA-guided Pixel Average-Velocity Modeling. To efficiently generate the cell-center topology $\mathbf{T}$, rTPG adopts a pixel-space MeanFlow formulation. Specifically, conditioned on the transcriptomic state $\mathbf{z}$, the model learns a transport trajectory from Gaussian noise ϵ_{tp} to the clean topology $\mathbf{T}$. Unlike conventional flow-based models that directly parameterize the instantaneous velocity field v, MeanFlow instead models an average velocity field u over a time interval, which enables one-step generation with reduced inference cost. Following [31], we formulate this process directly in pixel space, thereby avoiding latent-space construction and preserving the spatial fidelity required for cell-center topology generation. The relationship between the average velocity u and the instantaneous velocity v is given by the *MeanFlow Identity*:

$$\begin{aligned} v(x_t, t) &= u(x_t, r, t, \mathbf{z}) + (t-r)\frac{d}{dt}u(x_t, r, t, \mathbf{z}), \\ x_t &= (1-t)\mathbf{T} + t\epsilon_{tp}, \qquad \epsilon_{tp} \sim \mathcal{N}(0, \mathrm{I}), \end{aligned} \tag{2}$$

where x_t denotes the interpolated state between the clean topology $\mathbf{T}$ and Gaussian noise ϵ_{tp}.

Following JiT [24], we predict the clean topology endpoint rather than the velocity field, which is more favorable for 1-NFE generation. To this end, we define the topology estimator induced by the average velocity as

$$\widehat{\mathbf{T}}(x_t, r, t, \mathbf{z}) \triangleq x_t - t \cdot u(x_t, r, t, \mathbf{z}). \tag{3}$$

We parameterize this estimator with a network Φ, implemented as a standard Vision Transformer (ViT), where the conditions $(r, t, \mathbf{z})$ are injected as in-context tokens. The predicted topology is thus given by

$$\mathbf{T}_\Phi = \Phi(x_t, r, t, \mathbf{z}), \tag{4}$$

and the corresponding average velocity can be recovered from Eq. 3 as

$$u_\Phi(x_t, r, t, \mathbf{z}) = \frac{1}{t}\big(x_t - \mathbf{T}_\Phi\big). \tag{5}$$

b) Topology Learning Objective. The training objective of rTPG consists of a denoising term and a topological consistency term. Following [10], the denoising objective is defined in the instantaneous velocity space. Based on Eq. 2, we compute the temporal derivative $\frac{d}{dt}u_\Phi$ using the Jacobian-vector product (JVP), and construct the predicted instantaneous velocity as

$$V_\Phi = u_\Phi + (t - r)\Big[v_\Phi \partial_x u_\Phi + \partial_t u_\Phi\Big]_{\mathrm{sg}}, \qquad v_\Phi := \frac{1}{t}\big(x_t - \Phi(x_t, t, t, \mathbf{z})\big), \tag{6}$$

where $[\cdot]_{\mathrm{sg}}$ denotes the stop-gradient operator. The denoising loss is then defined as

$$\mathcal{L}_d^u = \|V_\Phi - v_{\mathrm{tgt}}\|^2, \qquad v_{\mathrm{tgt}} = \frac{dx_t}{dt} = \epsilon_{tp} - \mathbf{T}. \tag{7}$$

To further enforce structural consistency, we introduce a topology regularization term $\mathcal{L}_{tp}$ between the predicted topology $\mathbf{T}_\Phi$ and the ground-truth topology $\mathbf{T}$. Following [48], this term is defined over each of the K channels using the distance transform $D(\cdot)$ and the corresponding persistence diagram $Dgm(\cdot)$:

$$\mathcal{L}_{tp} = \frac{1}{K}\sum_{k=1}^{K} \mathcal{L}_{spc}\Big(Dgm(D(\mathbf{T}_\Phi^k)), Dgm(D(\mathbf{T}^k))\Big). \tag{8}$$

Finally, rTPG is optimized with $\mathcal{L}_{\mathrm{rTPG}} = \mathcal{L}_d^u + \lambda_{tp}\mathcal{L}_{tp}$, where λ_{tp} balances the denoising objective and the topology-preserving regularization. Please refer to Sec. A.1 for details of $\mathcal{L}_{spc}$.

c) Structural Control Derivation. After generating the cellular topology $\mathbf{T}$, we further derive two topology-based structural controls: a Gaussian density field $\mathbf{G}$ that captures cell-density organization and a distance representation $\mathbf{E}$ that encodes local geometric proximity to cell centers.

Topology-Anchored Hierarchical Flow-Matching Renderer (TopoRenderer). TopoRenderer synthesizes histology tiles from $\mathbf{T}$, using $\mathbf{G}$ for low-frequency structure and $\mathbf{E}$ for high-frequency morphology.

a) Structure-aware Noise Construction. To construct a topology-aligned source state for TopoRenderer, we derive a structure-aware noise ϵ_s from the Gaussian density field $\mathbf{G}$. Since $\mathbf{G}$ encodes the spatial distribution of cell centers, we use it as the structural reference for noise modulation. Specifically, we decompose both $\mathbf{G}$ and Gaussian noise ϵ in the Fourier domain as

$$\mathcal{F}(\mathbf{G}) = A_g \cdot e^{j\phi_g}, \qquad \mathcal{F}(\epsilon) = A_\epsilon \cdot e^{j\phi_\epsilon}. \tag{9}$$

To control the amount of structural information injected from $\mathbf{G}$, we introduce a mask $M \in \mathbb{R}^{H\times W}$, which is set to 1 inside a circle of radius r centered at $(\frac{H}{2}, \frac{W}{2})$, and 0 elsewhere. We then construct the structure-aware noise by combining the amplitude of ϵ with a phase field whose low-frequency components are inherited from $\mathbf{G}$:

$$\mathcal{F}(\epsilon_s) = A_\epsilon \cdot e^{j(\phi_\epsilon \odot (1-M) + \phi_g \odot M)}, \qquad \epsilon_s = \mathcal{F}^{-1}\big(\mathcal{F}(\epsilon_s)\big), \tag{10}$$

where $\odot$ denotes the Hadamard product and $\mathcal{F}^{-1}$ denotes the inverse Fourier transform.

b) Gaussian Density-conditioned Semantic Encoder $\mathcal{S}$. We adopt a DiT-based semantic encoder $\mathcal{S}$ to model topology-enhanced low-frequency semantics under the guidance of the Gaussian density field $\mathbf{G}$. Given a histology tile $\mathbf{X}$, we construct the interpolated state

$$\mathbf{X}_t = (1-t)\mathbf{X} + t\epsilon_s, \qquad v_s = \frac{d\mathbf{X}_t}{dt} = \epsilon_s - \mathbf{X}. \tag{11}$$

Table 1: **Quantitative comparison on TCGA benchmarks.** Fréchet Inception Distance (**FID**) and Fréchet Conch Distance (**FCD**) are reported, and ↓ means the lower is better.

Method	GBM		CESC		KIRP		COAD		LUAD		ALL	
	FID↓	FCD↓	FID↓	FCD↓	FID↓	FCD↓	FID↓	FCD↓	FID↓	FCD↓	FID↓	FCD↓
PathLDM [50]	22.46	72.87	19.87	43.13	20.39	80.27	26.96	77.20	21.28	91.14	20.29	68.29
U-ViT [1]	13.89	65.11	15.74	45.78	21.83	96.29	26.75	82.03	17.86	85.20	18.55	62.86
DiT [38]	15.03	63.25	18.75	54.34	21.53	76.12	29.13	64.93	18.57	59.54	18.11	56.36
SiT [34]	14.63	68.56	19.48	51.01	22.10	80.39	29.47	72.91	19.52	86.51	18.84	60.93
LlamaGen [43]	15.48	70.43	18.34	46.97	19.89	60.76	27.91	71.09	17.52	44.80	17.43	54.72
VQGAN [7]	21.74	76.13	23.48	58.21	26.10	85.89	32.47	89.61	26.48	54.72	25.09	56.43
VAR [44]	12.96	47.57	17.21	31.40	17.32	40.14	25.84	38.87	15.40	37.00	16.83	33.46
ImageFolder [25]	23.75	60.89	23.97	43.35	25.81	62.11	33.25	58.26	25.46	60.92	24.81	46.02
RNA-CDM [3]	24.15	68.60	25.76	41.36	24.46	70.24	33.60	59.34	27.98	47.10	23.36	54.40
GeneVAR [57]	11.20	49.95	13.68	29.06	14.49	40.44	19.86	36.48	13.65	37.49	12.95	28.84
TOPOSCAPE	**10.21**	**35.25**	**9.29**	**25.27**	**10.40**	**29.91**	**15.04**	**33.96**	**10.50**	**34.17**	**10.51**	**27.11**

Here, v_s denotes the ground-truth instantaneous velocity field for flow matching. To capture low-frequency semantic structure, $\mathcal{S}$ jointly takes the downsampled state $\hat{\mathbf{X}}_t$, a structural feature $\mathbf{f}_G$ extracted from $\mathbf{G}$ via $\sigma_1(\cdot)$, and the transcriptomic state $\mathbf{z}$ as inputs. These conditions are fused through an improved adaptive layer normalization (adaLN) mechanism, in which the modulation parameters are generated as

$$\alpha_1, \beta_1, \gamma_1, \alpha_2, \beta_2, \gamma_2 = \mathrm{MLP}(\mathbf{z} + \mathbf{f}_G + t). \tag{12}$$

The encoder output $\mathbf{x}_{\mathrm{low}} = \mathcal{S}(\hat{\mathbf{X}}_t, \mathbf{G}, \mathbf{z}, t)$ serves as the low-frequency semantic condition for the subsequent pixel-level decoding process.

c) Distance-guided Pixel Decoder $\mathcal{P}$. The distance representation $\mathbf{E}$ encodes continuous geometry by measuring each position's distance to the nearest cell center, providing local cues for high-frequency morphological refinement. The noised image $\mathbf{X}_t$ is embedded at full resolution into pixel features $\mathbf{h}$, while $\mathbf{E}$ is projected without downsampling into distance features $\mathbf{f}_E$ via $\sigma_2(\cdot)$. The upsampled semantic feature $\mathbf{x}_{\mathrm{low}}^{\mathrm{up}}$ and $\mathbf{f}_E$ are then fused through an adaLN-Zero block [38], injecting topology-aware geometric modulation into the pixel decoder:

$$\begin{aligned} \alpha, \beta, \gamma &= \mathrm{MLP}(\mathrm{SiLU}(\mathbf{x}_{\mathrm{low}}^{\mathrm{up}} + \mathbf{f}_E + t)), \\ \mathbf{h}_i &= \mathbf{h}_{i-1} + \mathrm{MLP}((1+\gamma) \odot \mathrm{LN}(\mathbf{h}_{i-1}) + \beta) \odot \alpha. \end{aligned} \tag{13}$$

Overall, $\mathbf{G}$ and $\mathbf{E}$ are hierarchically integrated into $\mathcal{S}$ and $\mathcal{P}$ to estimate the instantaneous velocity field. The prediction and optimization objective are defined as

$$v_s' = \mathcal{P}\big(\mathcal{S}(\hat{\mathbf{X}}_t, \mathbf{G}, \mathbf{z}, t), \mathbf{X}_t, \mathbf{E}, t\big), \quad \mathcal{L}_d^v = \|v_s' - v_s\|^2. \tag{14}$$

4 Experiments

We evaluate TOPOSCAPE on five TCGA cohorts used in RNA-CDM [3], spanning diverse cancer types and tissue morphologies: lung adenocarcinoma (LUAD, $n = 520$), kidney renal papillary cell carcinoma (KIRP, $n = 298$), colon adenocarcinoma (COAD, $n = 289$), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC, $n = 277$), and glioblastoma multiforme (GBM, $n = 212$). Each cohort contains diagnostic WSIs paired with bulk RNA-seq profiles. Following prior work [3, 57], gigapixel WSIs are partitioned into non-overlapping 256×256 tiles at $20\times$ magnification, so each RNA-seq profile is associated with a bag of histology tiles from its paired slide. To ensure direct and fair comparison with prior transcriptome-to-histology generation methods, we follow the established TCGA preprocessing and split protocol used in these studies, and train and evaluate all baselines under the same setting. We assess generated tiles with Fréchet Inception Distance (FID) [17] for visual fidelity and Fréchet Conch Distance (FCD), computed using the pathology foundation model Conch [30], for pathology-specific biological fidelity.

4.1 Main Results

Quantitative Comparison with SOTAs. We compare TOPOSCAPE with 10 representative generative baselines, including RNA-to-WSI methods (GeneVAR [57], RNA-CDM [3]) and general generative

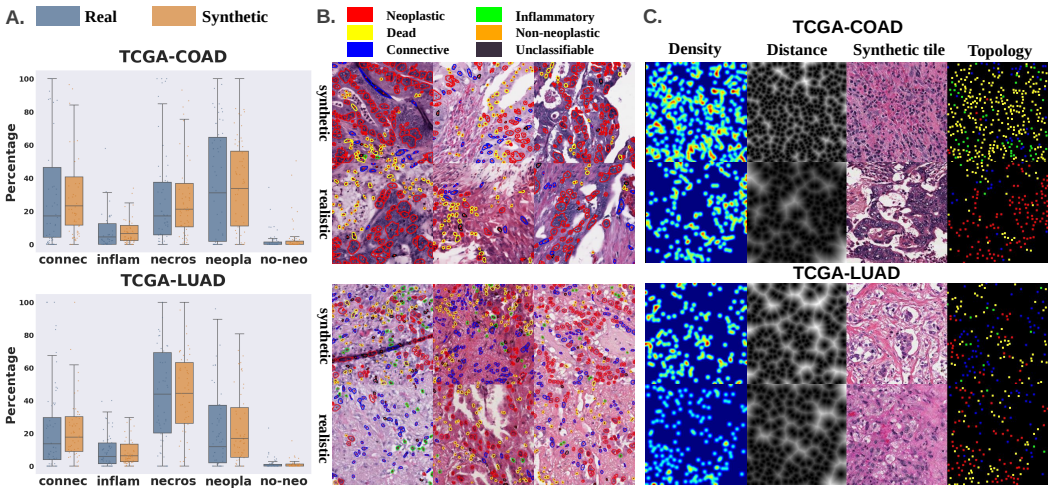


Figure 3: **Synthetic vs. real tiles**. *Panel A.* Cell-type distributions in real and synthetic tiles across TCGA cohorts. *Panel B.* HoverNet [12] visualizations showing consistent cellular composition between synthetic and real samples. *Panel C.* Density and distance maps, serving as the generative conditions, along with the corresponding synthetic tiles. Best viewed in color.

models adapted with RNA-seq conditioning for fair comparison. All models are trained on the same cohorts under consistent settings, and evaluated using FID and FCD on both individual cohorts and the pooled dataset (ALL), with 50K tiles uniformly sampled per RNA-seq profile. As shown in Tab. 1, TOPOSCAPE consistently achieves the best performance across all cohorts. It obtains the lowest FID on every dataset, with particularly large gains on challenging settings such as COAD and ALL, outperforming the second-best methods by substantial margins. More importantly, TOPOSCAPE also establishes new state-of-the-art results on FCD across all cohorts, indicating improved biological fidelity in terms of cellular composition and tissue organization. These results demonstrate that explicitly modeling cellular topology leads to consistent gains in both visual realism and pathology-aware representation.

Cell Distribution Consistency. Beyond distributional fidelity, we evaluate whether synthetic tiles preserve realistic cellular composition. For each cohort, we randomly sample 2,000 real tiles and generate RNA-conditioned synthetic counterparts. We then use HoverNet [12] to detect and classify five cell types: neoplastic, inflammatory, connective, dead, and non-neoplastic epithelial cells. As shown in Tab. 2, TOPOSCAPE produces cell-type proportions that most closely match real tissues, outperforming strong baselines such as GeneVAR [57] and VAR [44]. The agreement is particularly strong for representative cell populations, e.g., neoplastic cells in COAD and dead cells in LUAD. The visualizations in Fig. 3 further confirm that TOPOSCAPE generates cohort-consistent cellular composition and morphology, supporting the biological realism of topology-anchored synthesis.

Analysis for Cell Topology. Fig. 3(C) illustrates the rTPG-generated topology alongside the derived Gaussian density and distance map, demonstrating that the synthetic tiles faithfully preserve the underlying spatial structure in topology. Additionally, we compare cell distributions between rTPG-generated and real topologies; Fig. 4(A) shows that rTPG faithfully recapitulates the realistic cellular architecture. To further assess whether the generated tiles preserve both spatial organization and cellular composition, we apply HoverNet to the synthesized images; Fig. 4(B) confirms that the generated cell layout, which serves as the generative condition, is consistent with that recovered from the synthesized tiles.

Tile Classification with Synthetic Data. To assess the inter-class separability of synthetic data, we train a ResNet-50 [15] classifier on 5,000 real and 5,000 synthetic tiles from five cancer types, replacing real tiles with synthetic ones at different ratios p. As shown in Tab. 3, TOPOSCAPE causes only a minor performance drop (ACC: 0.654 $\rightarrow$ 0.642; F1: 0.653 $\rightarrow$ 0.638), while GeneVAR [57] degrades more substantially. Moreover, synthetic tile pretraining consistently improves classification, with TOPOSCAPE achieving the largest gains (+0.102 ACC, +0.104 F1), demonstrating its utility as effective training data.

WSI-Level Gains from Synthetic Tiles. To further demonstrate the effectiveness in computational pathology, we evaluated 8 SOTA multi-instance learning/regression (MIL/MIR) methods across

Table 2: **Cell distribution comparison.** Mean±std of cell proportions across cohorts. Extended results in Tab. 7 (Appendix).

Dataset	Method	Neoplastic	Dead
COAD	GeneVAR [57]	38.14%±28.25	38.25%±24.69
	Ours	**35.30%±26.28**	**25.33%±19.55**
	Real	35.83%±32.80	24.87%±24.84
LUAD	GeneVAR [57]	20.20%±22.78	61.10%±26.56
	Ours	**23.14%±21.72**	**44.63%±23.43**
	Real	23.09%±26.24	44.64%±29.01

Table 3: **Tile classification performance.** ACC/F1 under varying substitution (p) and pretraining (q) ratios. Full results in Tab. 8 (Appendix).

Method	p=0.0		p=0.75	
	ACC↑	F1↑	ACC↑	F1↑
GeneVAR [57]	0.657±0.022	0.645±0.026	0.602±0.032	0.597±0.034
Ours	**0.654±0.015**	**0.653±0.013**	**0.642±0.033**	**0.638±0.031**
	q=0.5		q=1.0	
GeneVAR [57]	0.680±0.012	0.679±0.013	0.734±0.031	0.734±0.032
Ours	**0.694±0.018**	**0.695±0.019**	**0.756±0.009**	**0.757±0.009**

Table 4: **Performance comparison** of MIL methods (w/ vs. w/o pretraining). For each RNA-seq profile, we synthesize 512 WSI tiles.

Method	ACC ↑		F1-score ↑		AUC ↑	
	w/o	w/	w/o	w/	w/o	w/
TransMIL	0.849	**0.877**	0.847	**0.876**	0.894	**0.936**
ACMIL	0.767	**0.890**	0.749	**0.890**	0.817	**0.943**
WiKG	0.767	**0.849**	0.753	**0.843**	0.820	**0.929**
MambaMIL	0.836	**0.932**	0.833	**0.931**	0.925	**0.948**

Table 5: **Performance comparison** of MIR methods (w/ vs. w/o augmentation). For each RNA-seq profile, we synthesize 512 WSI tiles.

Method	RMSE ↓		PCC ↑		SCC ↑	
	w/o	w/	w/o	w/	w/o	w/
AbReg	0.307	**0.293**	0.184	**0.209**	0.184	**0.205**
SEQUOIA	0.282	**0.273**	0.207	**0.239**	0.204	**0.230**
AbMIL	0.272	**0.260**	0.232	**0.262**	0.218	**0.253**
2dMamba	0.271	**0.262**	0.253	**0.274**	0.246	**0.266**

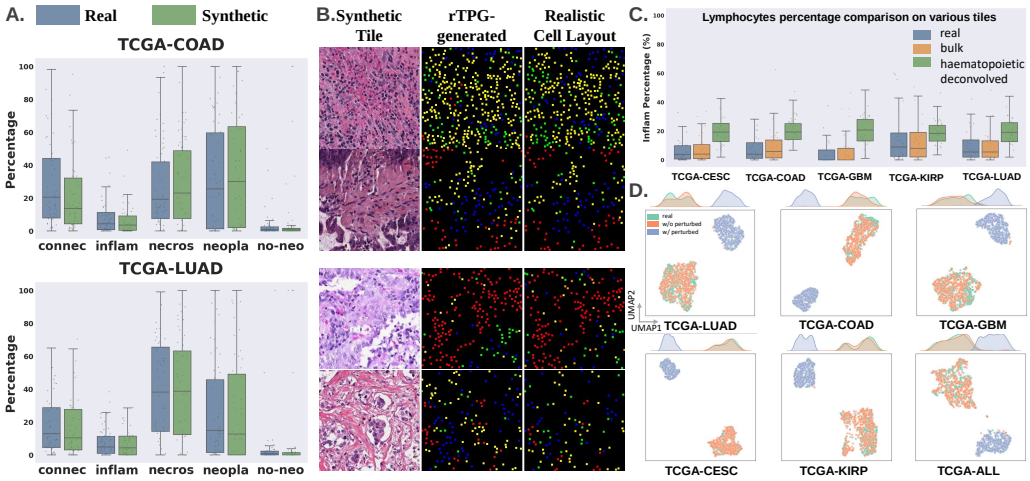


Figure 4: **Synthetic vs. real topology**. *Panel A.* Cell-type distributions in real and synthetic topologies across TCGA cohorts. *Panel B.* Comparison between rTPG-generated topology and real topology detected by HoverNet [12]. *Panel C & D.* Comparison of cell-type proportions and UMAP projections derived from bulk RNA-Seq versus deconvolved expression profiles. Best viewed in color.

two downstream tasks. For WSI classification [56], we evaluated TransMIL [42], ACMIL [54], WiKG [22] and MambaMIL [49], and assessed the performance gains following synthetic pretraining, while MambaMIL performs best, achieving 0.932 ACC, 0.931 F1, and 0.948 AUC. For slide-level transcriptomic estimation, we benchmarked AbReg [13], SEQUOIA [39], AbMIL [21], and 2dMamba [52], and consistently observed performance improvements under synthetic tile augmentation, with SEQUOIA achieving the largest gains (PCC+0.032, SCC +0.026). Please refer to Appendix Tab. 10 and Tab. 11 for full results.

RNA Perturbation Analysis. We further evaluate tile generation conditioned on haematopoietic deconvolved RNA-seq profiles. Given the enrichment of haematopoietic signals, an increased prevalence of lymphocytes is expected in the synthesized tiles. Consistent with this hypothesis, the median lymphocyte proportion is systematically higher in tiles generated from haematopoietic RNA-seq compared to those conditioned on bulk RNA-seq across cancer types (Fig. 4(C)). Moreover, UMAP projections of tiles generated from bulk versus deconvolved RNA-seq exhibit clear separation in the Conch [30] feature space, underscoring the effectiveness of RNA perturbation (Fig. 4(D)).

Table 6: **Ablation experiments.** We systematically analyze (a) the effectiveness of Topology & Structure Noise, (b) integration of Topology($\mathbf{T}$), Density ($\mathbf{G}$) and Distance map ($\mathbf{E}$), (c) compatibility of rTPG with different generative models.

	Settings	FID (↓)	FCD (↓)
I	base.(Δ)	14.39	39.54
	+ $\mathbf{T_1}$	36.78	79.60
	+ $\mathbf{T_2}$	24.34	43.47
	+ $\mathbf{T_3}$	**10.51**	**27.11**
II	noise ϵ	12.06	30.59
	noise ϵ_s	**10.51**	**27.11**

(a) Effect of Topo & Structure Noise.

Settings.	FID (↓)	FCD (↓)
base.(Δ)	14.39	39.54
only $\mathbf{T}$	13.63	35.71
only $\mathbf{G}$	12.15	**29.68**
only $\mathbf{E}$	26.94	54.76
$[\mathbf{T}, \mathbf{E}]$	**11.73**	30.04
$[\mathbf{G}, \mathbf{E}]$	**10.51**	**27.11**

(b) Analysis of $\mathbf{T}$, $\mathbf{G}$ and $\mathbf{E}$.

Settings	FID (↓)	FCD (↓)
JiT	16.74	52.37
+ $\mathbf{G}$	14.05	35.63
+ $\mathbf{T}$	14.97	37.45
SiT	18.84	60.93
+ $\mathbf{G}$	15.41	41.43
+ $\mathbf{T}$	16.19	43.21

(c) Compatibility of rTPG.

4.2 Ablation Study

Ablation for Effectiveness of Topology. We assess the necessity of RNA-guided topology generation, and set the RNA-guided DeCo [35] as control (Δ); results are reported in Tab. 6a(I). Randomly generated $\mathbf{T}_1$ and its derivatives ($\mathbf{G}_1$, $\mathbf{E}_1$) fed into TopoRenderer drastically increase FID (14.39→36.78) and FCD (39.54→79.60), indicating that biologically meaningful topology is critical. TopoCellGen [48] generates $\mathbf{T}_2$ with structurally plausible but randomly proportioned cells, mitigating this degradation; however, performance remains suboptimal (FID 24.34, FCD 43.47), suggesting the importance of accurate cell proportions. In contrast, rTPG-generated $\mathbf{T}_3$ yields topologies that are both structurally and proportionally valid.

Analysis for Density $\mathbf{G}$ and Distance $\mathbf{E}$. We evaluate hierarchical fusion of topology-derived maps (Tab. 6b). Direct conditioning on the point map $\mathbf{T}$ via TopoFM [23] serves as the control, yielding FID/FCD increases of 0.76 and 3.83, respectively, relative to the baseline DeCo. Adding the density map $\mathbf{G}$ declines FID (13.63→12.15) and FCD(35.71→29.68) compared to the control, while naively injecting the distance map $\mathbf{E}$ degrades performance. Ultimately, the fusion of T/G and E achieves the best results. This indicates nontrivial complementarity between $\mathbf{G}$ and $\mathbf{E}$, and demonstrates that hierarchical fusion outperforms direct encoding of $\mathbf{T}$.

Compatibility of rTPG-generated Topology. We evaluate the compatibility of rTPG-generated topology with generic image generators by augmenting SiT [34] and JiT [24] with topology guidance. As neither model employs frequency decoupling, we condition semantic encoding on $\mathbf{T}$ or $\mathbf{G}$. As shown in Tab. 6c, introducing rTPG-generated topology leads to significant FCD decreases across both variants (52.37→35.63 for JiT, and 60.93 → 41.43 for SiT). Moreover, conditioning on $\mathbf{G}$ yields larger gains than directly using $\mathbf{T}$. For JiT [24], FID(FCD) declines from 14.97(37.45) to 14.05(35.63); for SiT [34], from 16.19(43.21) to 15.41(41.43).

Analysis for Structure Noise ϵ_s. We emphasize that structured noise is beneficial (Tab. 6a(II)). By encoding spatial information from conditions into the noise and training it explicitly, the model better adheres to spatial layout constraints, reducing FID from 12.06 to 10.51. Correspondingly, accurate layouts yield a larger FCD gain of 3.48 (30.59→27.11).

5 Conclusion

We introduced TOPOSCAPE, a topology-anchored framework for transcriptome-conditioned histology image generation. By adopting a *Layout-Before-Pixels* paradigm, TOPOSCAPE explicitly separates cellular organization from pixel-level synthesis, enabling a structured mapping from non-spatial molecular profiles to spatially coherent tissue morphology. This design addresses the limitations of direct RNA-to-image generation by introducing topology as an intermediate representation that guides both global structure and local detail formation. Extensive experiments across multiple TCGA cohorts demonstrate that TOPOSCAPE achieves strong visual fidelity, improved biological realism, and consistent downstream utility. More broadly, our results highlight the importance of structural intermediates in generative modeling, suggesting topology-anchored generation as a promising direction for bridging molecular signals and spatially organized image synthesis.

Limitations. TOPOSCAPE follows established TCGA-based evaluation protocols [2, 3, 57], and broader external validation would further assess generalization. Its topology supervision relies on automated nuclei analysis, which may introduce detector-specific bias. Future work will explore expert-verified or multi-detector topology guidance and stronger conditional consistency analyses.

Acknowledgement. This research is supported by the National Natural Science Foundation of China (62203089, 62303092, 62103084), the China Postdoctoral Science Foundation under Grant Number GZC20252334 and 2025M770276, the Fundamental Research Funds for the Central Universities under grant No. 2682025CX116.

References

[1] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *CVPR*, 2023.

[2] Francisco Carrillo-Perez, Marija Pizurica, Michael G Ozawa, Hannes Vogel, Robert B West, Christina S Kong, Luis Javier Herrera, Jeanne Shen, and Olivier Gevaert. Synthetic whole-slide image tile generation with gene expression profile-infused deep generative models. *Cell Reports Methods*, 2023.

[3] 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.

[4] Eduard Chelebian, Christophe Avenel, and Carolina Wählby. Combining spatial transcriptomics with tissue morphology. *Nature Communications*, 2025.

[5] Youngmin Chung, Ji Hun Ha, Kyeong Chan Im, and Joo Sang Lee. Accurate spatial gene expression prediction by integrating multi-resolution features. In *CVPR*, 2024.

[6] Tong Ding, Sophia J Wagner, Andrew H Song, Richard J Chen, Ming Y Lu, Andrew Zhang, Anurag J Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, et al. A multimodal whole-slide foundation model for pathology. *Nature medicine*, 2025.

[7] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021.

[8] 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.

[9] Aniruddha Ganguly, Debolina Chatterjee, Wentao Huang, Jie Zhang, Alisa Yurovsky, Travis Steele Johnson, and Chao Chen. Merge: multi-faceted hierarchical graph-based gnn for gene expression prediction from whole slide histopathology images. In *CVPR*, 2025.

[10] Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J Zico Kolter, and Kaiming He. Improved mean flows: On the challenges of fastforward generative models. *arXiv preprint arXiv:2512.02012*, 2025.

[11] Mauro Giuffrè and Dennis L Shung. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *NPJ digital medicine*, 2023.

[12] 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.

[13] Mara Graziani, Niccolò Marini, Nicolas Deutschmann, Nikita Janakarajan, Henning Müller, and María Rodríguez Martínez. Attention-based interpretable regression of gene expression in histology. In *iMIMIC Workshop*, pages 44–60. Springer, 2022.

[14] Bryan He, Ludvig Bergenstråhle, Linnea Stenbeck, Abubakar Abid, Alma Andersson, Åke Borg, Jonas Maaskola, Joakim Lundeberg, and James Zou. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature biomedical engineering*, 2020.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[16] Siyu He, Yinuo Jin, Achille Nazaret, Lingting Shi, Xueer Chen, Sham Rampersaud, Bahawar S Dhillon, Izabella Valdez, Lauren E Friend, Joy Linyue Fan, et al. Starfysh integrates spatial transcriptomic and histologic data to reveal heterogeneous tumor–immune hubs. *Nature biotechnology*, 2025.

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.

[19] Danh-Tai Hoang, Gal Dinstag, Eldad D Shulman, Leandro C Hermida, Doreen S Ben-Zvi, Efrat Elis, Katherine Caley, Stephen-John Sammut, Sanju Sinha, Neelam Sinha, et al. A deep-learning framework to predict cancer treatment response from histopathology images through imputed transcriptomics. *Nature cancer*, 2024.

[20] Petr Holub, Heimo Müller, Tomáš Bíl, Luca Pireddu, Markus Plass, Fabian Prasser, Irene Schlünder, Kurt Zatloukal, Rudolf Nenutil, and Tomáš Brázdil. Privacy risks of whole-slide image sharing in digital pathology. *Nature Communications*, 2023.

[21] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

[22] Jiawen Li, Yuxuan Chen, Hongbo Chu, Qiehe Sun, Tian Guan, Anjia Han, and Yonghong He. Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. In *CVPR*, 2024.

[23] Jingxiong Li, Chenglu Zhu, Sunyi Zheng, Pingyi Chen, Yuxuan Sun, Honglin Li, and Lin Yang. Topofm: Topology-guided pathology foundation model for high-resolution pathology image synthesis with cellular-level control. *IEEE TMI*, 44(10):4116–4127, 2025.

[24] Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025.

[25] Xiang Li, Kai Qiu, Hao Chen, Jason Kuen, Jiuxiang Gu, Bhiksha Raj, and Zhe Lin. Imagefolder: Autoregressive image generation with folded tokens. In *ICLR*, 2025.

[26] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *CVPR*, 2023.

[27] Zhongxiao Li, Tianqi Su, Bin Zhang, Wenkai Han, Sibin Zhang, Guiyin Sun, Yuwei Cong, Xin Chen, Jiping Qi, Yujie Wang, et al. His-mmdm: Multi-domain and multi-omics translation of histopathological images with diffusion models. *Advanced Science*, 13(19):e18066, 2026.

[28] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[29] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[30] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature medicine*, 30(3):863–874, 2024.

[31] Yiyang Lu, Susie Lu, Qiao Sun, Hanhong Zhao, Zhicheng Jiang, Xianbang Wang, Tianhong Li, Zhengyang Geng, and Kaiming He. One-step latent-free image generation with pixel mean flows. *arXiv preprint arXiv:2601.22158*, 2026.

[32] Ao Luo, Xin Li, Fan Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Flowdiffuser: Advancing optical flow estimation with diffusion models. In *CVPR*, 2024.

[33] Ao Luo, Xin Li, Fan Yang, Yuezun Li, Zhaoquan Yuan, Shan Zhao, Bing Su, and Xiao Wu. Optical flow matching: Reframing optical flow as continuous transport dynamics. In *CVPR*, 2026.

[34] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *ECCV*, 2024.

[35] Zehong Ma, Longhui Wei, Shuai Wang, Shiliang Zhang, and Qi Tian. Deco: Frequency-decoupled pixel diffusion for end-to-end image generation. *arXiv preprint arXiv:2511.19365*, 2025.

[36] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, 2024.

[37] Kazuya Nishimura, Ryoma Bise, Shinnosuke Matsuo, Haruka Hirose, and Yasuhiro Kojima. Cell-type prototype-informed neural network for gene expression estimation from pathology images. *arXiv preprint arXiv:2603.18461*, 2026.

[38] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023.

[39] Marija Pizurica, Yuanning Zheng, Francisco Carrillo-Perez, Humaira Noor, Wei Yao, Christian Wohlfart, Antoaneta Vladimirova, Kathleen Marchal, and Olivier Gevaert. Digital profiling of gene expression from histology images with linearized attention. *Nature Communications*, 15 (1):9886, 2024.

[40] Jie Qiu, Xin Li, Fan Yang, Yan Wang, Dong Yu, Changying Wang, Linwei Tai, Yongxiang Chen, Youqin Chen, and Jianzhang Chen. Hyseg: Learning generative priors for structure-aware remote sensing segmentation. In *CVPR*, 2026.

[41] Benoît Schmauch, Alberto Romagnoni, Elodie Pronier, Charlie Saillard, Pascale Maillé, Julien Calderaro, Aurélie Kamoun, Meriem Sefta, Sylvain Toldo, Mikhail Zaslavskiy, et al. A deep learning model to predict rna-seq expression of tumours from whole slide images. *Nature communications*, 2020.

[42] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. In *NeurIPS*, 2021.

[43] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv*, 2024.

[44] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *NeurIPS*, 2024.

[45] Mengbo Wang, Shourya Verma, Aditya Malusare, Luopin Wang, Yiyang Lu, Vaneet Aggarwal, Mario Sola, Ananth Grama, and Nadia Lanman. Geneflow: Translation of single-cell gene expression to histopathological images via rectified flow. *NeurIPS*, 2026.

[46] 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.

[47] Ronald Xie, Kuan Pang, Sai Chung, Catia Perciani, Sonya MacParland, Bo Wang, and Gary Bader. Spatially resolved gene expression prediction from histology images via bi-modal contrastive learning. *NeurIPS*, 2023.

[48] Meilong Xu, Saumya Gupta, Xiaoling Hu, Chen Li, Shahira Abousamra, Dimitris Samaras, Prateek Prasanna, and Chao Chen. Topocellgen: Generating histopathology cell topology with a diffusion model. In *CVPR*, pages 20979–20989, 2025.

[49] Shu Yang, Yihui Wang, and Hao Chen. Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology. In *MICCAI*, 2024.

[50] Srikar Yellapragada, Alexandros Graikos, Prateek Prasanna, Tahsin Kurc, Joel Saltz, and Dimitris Samaras. Pathldm: Text conditioned latent diffusion model for histopathology. In *WACV*, 2024.

[51] Yuansong Zeng, Zhuoyi Wei, Weijiang Yu, Rui Yin, Yuchen Yuan, Bingling Li, Zhonghui Tang, Yutong Lu, and Yuedong Yang. Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *Briefings in Bioinformatics*, 2022.

[52] Jingwei Zhang, Anh Tien Nguyen, Xi Han, Vincent Quoc-Huy Trinh, Hong Qin, Dimitris Samaras, and Mahdi S Hosseini. 2dmamba: Efficient state space model for image representation with applications on giga-pixel whole slide image classification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3583–3592, 2025.

[53] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023.

[54] Yunlong Zhang, Honglin Li, Yunxuan Sun, Sunyi Zheng, Chenglu Zhu, and Lin Yang. Attention-challenging multiple instance learning for whole slide image classification. In *ECCV*, 2024.

[55] Jianwei Zhao, Xin Li, Fan Yang, Qiang Zhai, Ao Luo, Zicheng Jiao, and Hong Cheng. Focusdiffuser: Perceiving local disparities for camouflaged object detection. In *ECCV*, 2024.

[56] Jianwei Zhao, Xin Li, Fan Yang, Qiang Zhai, Ao Luo, Yang Zhao, Hong Cheng, and Huazhu Fu. Mexd: An expert-infused diffusion model for whole-slide image classification. In *CVPR*, 2025.

[57] Jianwei Zhao, Fan Yang, Xin Li, Qiang Zhai, Ao Luo, Ziqi Ren, Zhicheng Jiao, and Hong Cheng. Genevar: Causal meanflow for autoregressive gene-to-wsi tile synthesis. In *CVPR*, 2026.

[58] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *CVPR*, 2023.

[59] Junchao Zhu, Ruining Deng, Tianyuan Yao, Juming Xiong, Chongyu Qu, Junlin Guo, Siqi Lu, Mengmeng Yin, Yu Wang, Shilin Zhao, et al. Asign: an anatomy-aware spatial imputation graphic network for 3d spatial transcriptomics. In *CVPR*, 2025.

A Technical appendices and supplementary material

A.1 Calculation of Topology Loss

The calculation process of $\mathcal{L}_{tp}$ involves extracting spatial features via distance transforms and aligning topological features through 1-dimensional persistence diagrams.

First, for each channel of the target layout x_0 and the predicted noiseless layout $\hat{x}_0^t$, a distance transform is performed. The distance transform function $D(x)$ assigns each pixel in the channel the minimum Euclidean distance to the nearest cell (i.e., non-zero pixel), formally defined as:

$$D(x) = \min_{p \in \text{cells}} \|x - p\| \tag{15}$$

where p denotes the positions of cells in the layout. Through this operation, the binarized predicted layout is mapped to a distance map $\hat{x}_t^{edt} = D(b(\hat{x}_0^t))$, and the target layout is mapped to $x_0^{edt} = D(x_0)$.

Upon obtaining the distance transform maps, their 1-dimensional persistence diagrams are extracted, denoted as $Dgm(\hat{x}_t^{edt})$ and $Dgm(x_0^{edt})$, respectively. To quantify the discrepancy between these two topological distributions, the classic Wasserstein distance is employed to find an optimal matching γ^* between the sets of points in the diagrams. This matching process can be solved using methods such as the traditional Hungarian algorithm, which pairs each persistence dot in the predicted diagram $Dgm(\hat{x}_t^{edt})$ with either a target dot in $Dgm(x_0^{edt})$ or its projection on the diagonal. Once the optimal matching γ^* is determined, the spatial distribution consistency loss $\mathcal{L}_{\text{spc}}$ for a single class is defined as the sum of squared distances between each point in $Dgm(\hat{x}_t^{edt})$ and its match:

$$\mathcal{L}_{\text{spc}} = \sum_{q \in Dgm(\hat{x}_t^{edt})} \|q - \gamma^*(q)\|^2 \tag{16}$$

A.2 Implementation Details

Cell Topology Construction. We used HoverNet [12], trained on PanNuke [8], to classify cells into six categories: 0 (neoplastic), 1 (inflammatory), 2 (connective), 3 (dead), 4 (non-neoplastic), and 5 (non-classifiable). Each topology is stored as a '.npy' file using the same filename as its corresponding tile.

RNA-guided Topological Prior Generator (rTPG). rTPG adopts the network [31] for the network Φ, with model depth and feature dimension set to 12 and 768, respectively. The hyperparameter λ_{tp} is fixed at 5e-4 in line with TopoCellGen [48]. rTPG is trained using **Muon** with a learning rate of 1e-3 and a batch size of 32 for 100 epochs, using 100 topologies per WSI.

Topology-Anchored Hierarchical Flow-Matching Renderer (TopoRenderer). TopoRenderer follows the architecture of DeCo [35], with a 12-layer density-conditioned semantic encoder $\mathcal{S}$ and a 3-layer distance-guided pixel decoder $\mathcal{P}$. TopoRenderer is trained on two NVIDIA H800 GPUs for 400 epochs using 200 tile–topology pairs per WSI. Training is performed using **AdamW** with learning rate 1e-4, batch size 256, $\beta_1 = 0.9$, $\beta_2 = 0.95$, and weight decay 0.05. $\sigma_1(\cdot)$ is a two-layer convolution block with SiLU activation, while $\sigma_2(\cdot)$ denotes a two-layer linear projector. During training, $\mathbf{T}$ is sampled from the ground-truth topology. During inference, we replace the ground-truth topology with the rTPG-generated topology. The inference pipeline is summarized in Alg. 1. The radius r of mask M is set to 2.

Algorithm 1 TOPOSCAPE: Inference Pipeline

Input: RNA sequence $\mathbf{g}$, topology size $H \times W \times K$, image size $H \times W \times 3$, mask M, sample steps N
Phase 1 & 2: Encode RNA and topology generation
$\mathbf{z}$ = TSE($\mathbf{g}$)
$\epsilon_{tp} \sim \mathcal{N}(0, I)^{H \times W \times K}$
$\mathbf{T} = \Phi(\epsilon_{tp}, 0, 1, \mathbf{z})$

Phase 3: TopoRenderer
$\mathbf{G}, \mathbf{E}$ = derive_controls($\mathbf{T}$)
A_g, ϕ_g = fft($\mathbf{G}$)
$\epsilon \sim \mathcal{N}(0, I)^{H \times W \times 3}$
$A_\epsilon, \phi_\epsilon$ = fft(ϵ)
ϵ_s = ifft$\big(A_\epsilon * \exp(j * (\phi_\epsilon \odot (1 - M) + \phi_g \odot M))\big)$

$\mathbf{X}_t, \Delta t = \epsilon_s, 1/N$

for t in linspace(1, 0, N):
  $\mathbf{x}_{\text{low}} = \mathcal{S}$(downsample($\mathbf{X}_t$), $\mathbf{G}$, $\mathbf{z}$, t)
  $v'_s = \mathcal{P}$(upsample($\mathbf{x}_{\text{low}}$), $\mathbf{X}_t$, $\mathbf{E}$, t)
  $\mathbf{X}_t \leftarrow \mathbf{X}_t - \Delta t \cdot v'_s$

return $\mathbf{X}_0$

Dataset Split. We follow the established TCGA benchmark protocol adopted by prior RNA-to-histology generation studies [3, 57]. Specifically, stratified training/validation/test partitions are constructed at a 60%/20%/20% ratio over the preprocessed tile set, and the identical preprocessing and split setting is applied to all baselines for direct comparability.

Code Availability. The training scripts and pretrained models of TOPOSCAPE will be publicly.

A.3 Extended Experimental Results

Comprehensive Cell Distribution Comparison. Comprehensive cell distribution data across five datasets and five cell lineages are detailed in Tab. 7, serving as an extension of Tab. 2. The study utilized 2,000 RNA-Seq-derived synthetic tiles and 2,000 randomly sampled real-world counterparts for each dataset. For automated cell detection, we leveraged HoverNet [12] (trained via PanNuke [8])

Table 7: **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
	GeneVAR	5.04% ± 12.07	2.80% ± 7.60	8.97% ± 17.15	78.67% ± 26.91	4.49% ± 12.70
	Ours	8.77% ± 16.48	7.82% ± 16.68	15.80% ± 20.17	64.56% ± 33.06	3.04% ± 11.46
	Real	8.69% ± 18.27	7.81% ± 15.66	14.41% ± 22.64	65.29% ± 36.41	3.79% ± 12.81
CESC	VAR	30.92% ± 31.89	1.44% ± 5.20	19.04% ± 26.68	47.47% ± 32.18	1.10% ± 6.53
	GeneVAR	27.19% ± 24.82	2.05% ± 5.52	13.31% ± 19.57	56.67% ± 26.64	0.77% ± 3.21
	Ours	32.18% ± 27.60	7.36% ± 8.68	24.53% ± 23.19	33.25% ± 22.89	2.68% ± 5.33
	Real	31.57% ± 31.80	8.26% ± 12.65	22.70% ± 27.02	34.40% ± 28.22	3.05% ± 7.92
LUAD	VAR	25.06% ± 27.38	2.99% ± 6.84	15.23% ± 20.63	55.99% ± 30.00	0.71% ± 3.42
	GeneVAR	20.20% ± 22.78	3.95% ± 5.86	13.16% ± 15.14	61.10% ± 26.56	1.56% ± 6.64
	Ours	23.14% ± 21.72	9.24% ± 9.30	21.54% ± 16.64	44.63% ± 23.43	1.44% ± 3.98
	Real	23.09% ± 26.24	9.89% ± 11.91	20.41% ± 21.52	44.64% ± 29.01	1.95% ± 6.26
KIRP	VAR	21.59% ± 25.29	5.52% ± 8.01	9.22% ± 18.70	60.78% ± 26.97	2.72% ± 8.30
	GeneVAR	22.61% ± 23.44	2.91% ± 4.64	9.36% ± 13.89	63.32% ± 25.53	1.78% ± 4.76
	Ours	24.64% ± 22.46	12.19% ± 11.53	18.56% ± 17.63	40.16% ± 24.34	4.43% ± 9.09
	Real	26.20% ± 26.33	12.54% ± 14.59	16.12% ± 20.81	38.75% ± 28.53	6.38% ± 13.96
COAD	VAR	33.37% ± 33.08	2.01% ± 7.29	27.39% ± 28.00	36.08% ± 29.55	1.13% ± 6.71
	GeneVAR	38.14% ± 28.25	2.81% ± 5.00	20.09% ± 20.19	38.25% ± 24.69	0.69% ± 2.37
	Ours	35.30% ± 26.28	8.38% ± 8.28	28.85% ± 22.02	25.33% ± 19.55	2.12% ± 6.23
	Real	35.83% ± 32.80	8.79% ± 11.76	27.99% ± 28.48	24.87% ± 24.84	2.48% ± 6.41

Table 8: **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
GeneVAR	0.657 ± 0.022	0.645 ± 0.026	0.645 ± 0.026	0.636 ± 0.037	0.628 ± 0.005	0.623 ± 0.005	0.602 ± 0.032	0.597 ± 0.034
Ours	**0.654 ± 0.015**	**0.653 ± 0.013**	**0.662 ± 0.019**	**0.662 ± 0.022**	**0.634 ± 0.023**	**0.630 ± 0.029**	**0.642 ± 0.033**	**0.638 ± 0.031**
	q=25%		q=50%		q=75%		q=100%	
GeneVAR	0.665 ± 0.022	0.663 ± 0.024	0.680 ± 0.012	0.679 ± 0.013	0.719 ± 0.014	0.718 ± 0.014	0.734 ± 0.031	0.734 ± 0.032
Ours	**0.664 ± 0.025**	**0.663 ± 0.024**	**0.694 ± 0.018**	**0.695 ± 0.019**	**0.733 ± 0.011**	**0.733 ± 0.009**	**0.756 ± 0.009**	**0.757 ± 0.009**

to categorize cells into neoplastic, inflammatory, connective, dead, and non-neoplastic classes. Quantifications were initially performed at the tile level and subsequently averaged to derive global statistics. The comparative results reveal that TOPOSCAPE effectively replicates biological cell-type ratios, leading to superior morphological authenticity in the synthetic outputs. Visual evidence supporting these claims is provided in Fig. 5.

Comprehensive Cell Topology Analysis. We randomly sample 2,000 rTPG-generated topologies and an equal number of ground-truth topologies. We first evaluate cell-type distributions across five cohorts, and the complete results in Fig. 6 demonstrate consistent cell-count patterns. In addition, we employ TopoFD and TCE from TopoCellGen [48] to evaluate topological quality in Tab. 9. Our rTPG achieves comparable topological quality while maintaining efficient inference. rna-JiT denotes JiT [24] with the class embedding replaced by an RNA embedding.

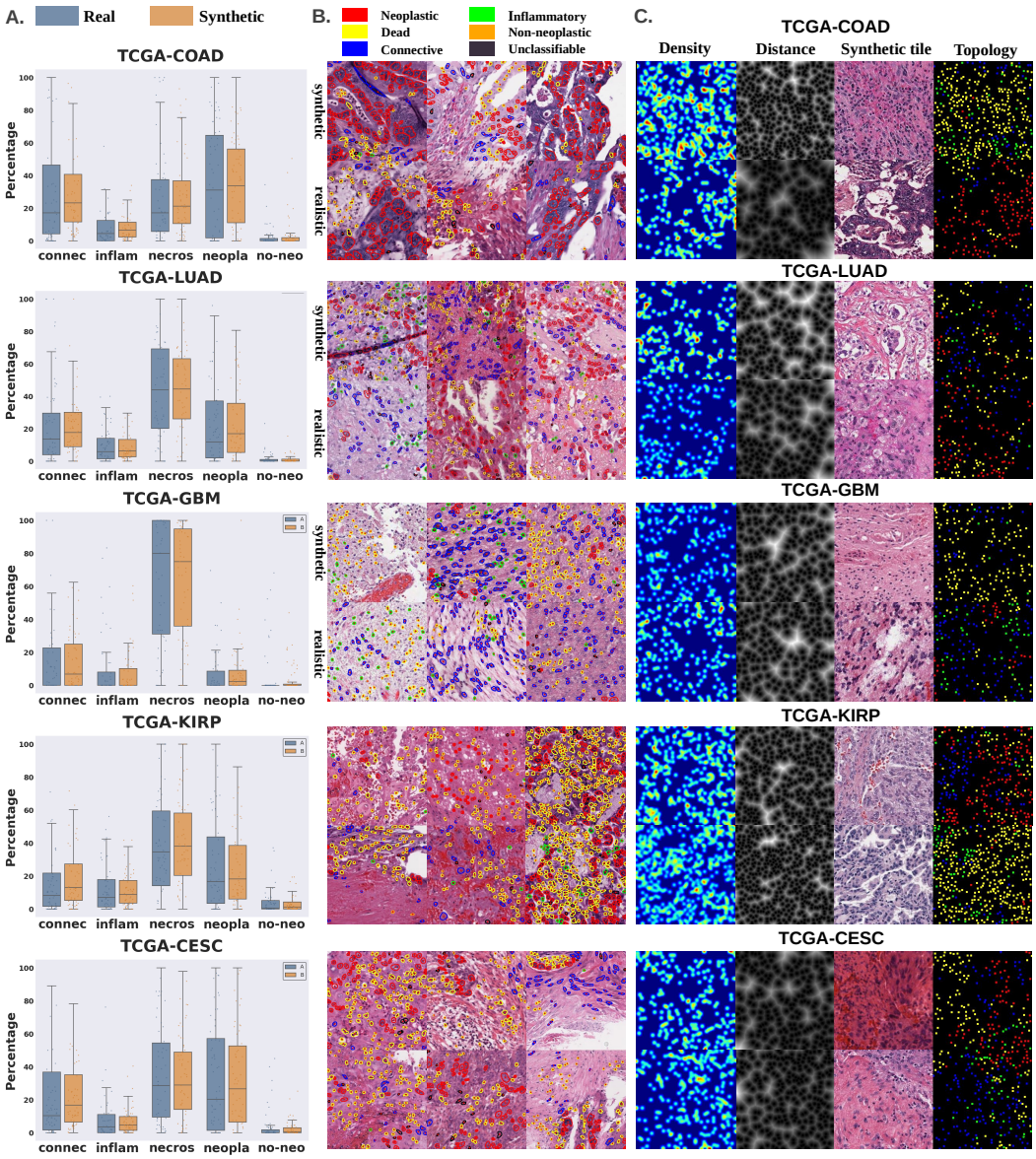


Figure 5: **Synthetic vs. real tiles**. *Panel A.* Cell-type distributions in real and synthetic tiles across TCGA cohorts. *Panel B.* HoverNet [12] visualizations showing consistent cellular composition between synthetic and real samples. *Panel C.* Density and distance maps used as generative conditions, together with the corresponding synthetic tiles. Best viewed in color.

Comprehensive Tile Classification Analysis. Performance metrics under various p and q distributions are thoroughly documented in Tab. 8. TOPOSCAPE is notably superior in sustaining classification accuracy as synthetic tile fractions grow within the 5,000 mixed samples, whereas traditional methods exhibit a sharp downward trajectory in performance. Additionally, we observe that the systematic addition of synthetic content during pretraining facilitates a steady improvement in classification outcomes, with TOPOSCAPE surpassing the benchmarks set by the two SOTA baselines. The methodology employed a 50-epoch training regimen for the classifiers, complemented by fixed 20-epoch pretraining intervals at each stage.

Comprehensive Slide-level Classification Analysis. Whole-slide image (WSI) classification is conducted within a multiple-instance learning (MIL) paradigm, where the ensemble of patches derived from a single WSI constitutes a "bag" linked to a slide-specific diagnosis: microsatellite stability (MSS, labeled 0) or microsatellite instability (MSI, labeled 1). The analytical pipeline involves extracting patch-level descriptors via a specialized feature extractor, which are subsequently processed

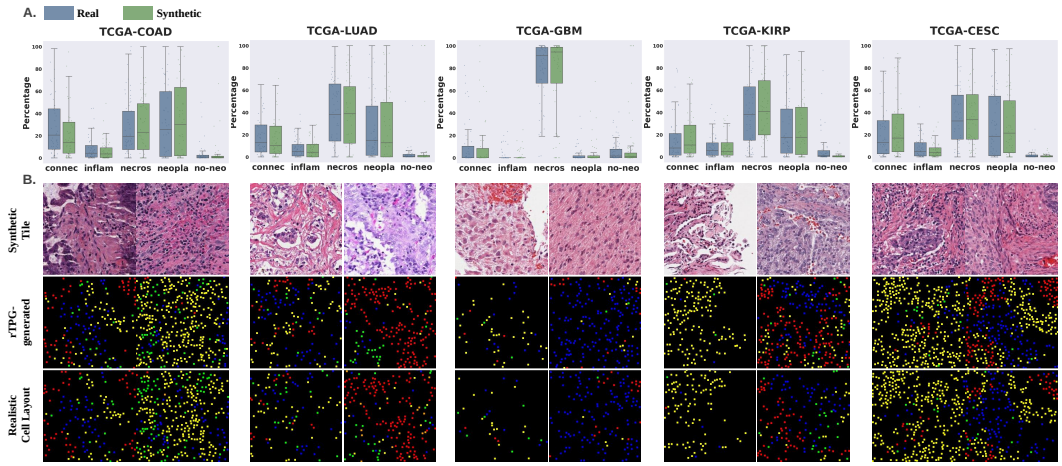


Figure 6: **Synthetic vs. real topology**. *Panel A.* Cell-type distributions in real and synthetic topologies across TCGA cohorts. *Panel B.* Comparison between rTPG-generated topology and real topology detected by HoverNet [12].

Table 9: Comparison of topology-generated methods on the TCGA-ALL cohort.

Method	TCE↓	TopoFD↓	NFE↓
rna-JiT	25.76	19.12	25
rTPG(ours)	25.44	21.60	1

Table 10: **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(MoCo V3)						
TransMIL	0.849	**0.877**	0.847	**0.876**	0.894	**0.936**
ACMIL	0.767	**0.890**	0.749	**0.890**	0.817	**0.943**
WiKG	0.767	**0.849**	0.753	**0.843**	0.820	**0.929**
MambaMIL	0.836	**0.932**	0.833	**0.931**	0.925	**0.948**
CTransPath						
TransMIL	0.822	**0.863**	0.821	**0.860**	0.862	**0.932**
ACMIL	0.808	**0.863**	0.800	**0.859**	0.836	**0.926**
WiKG	0.767	**0.849**	0.745	**0.846**	0.876	**0.931**
MambaMIL	0.781	**0.877**	0.766	**0.875**	0.867	**0.928**

by a classification architecture to determine the overall slide status. To rigorously evaluate the clinical efficacy of the generated patches, we utilized two distinct feature extraction backbones—ViT pre-trained with MoCoV3 and CTransPath—in conjunction with four leading WSI classification frameworks. Our experiments utilized the COAD MSI cohort, comprising 360 patients (298 MSS and 66 MSI cases), partitioned into a training set of 292 slides and a validation set of 72 slides. For every transcriptomic profile, TOPOSCAPE produces 512 synthetic patches to facilitate the pretraining of MIL classifiers. As detailed in Tab. 10, a universal performance improvement was observed across all classification models and feature extractors.

Comprehensive Slide-level Transcriptome Estimation Analysis. Slide-level transcriptome estimation from whole-slide images is formulated as a multiple-instance regression (MIR) problem, where gene expression profiles ($\in \mathbb{R}^N$, N: total number of genes) are predicted from a set of patches sampled from a single WSI. Following HisMMDM [27], we select 4,461 well-studied genes. The

Table 11: **Performance comparison of slide-to-transcriptome methods** with and without synthetic augmentation across four MIL/MIR models. Shaded cells indicate results with augmentation.

Method	RMSE		PCC		SCC	
	w/o	w/	w/o	w/	w/o	w/
Conch						
AbReg	0.284	**0.270**	0.260	**0.276**	0.249	**0.264**
SEQUOIA	0.300	**0.272**	0.258	**0.278**	0.244	**0.269**
AbMIL	0.280	**0.269**	0.261	**0.274**	0.253	**0.264**
2dMamba	0.293	**0.272**	0.245	**0.266**	0.230	**0.265**
CTransPath						
AbReg	0.307	**0.293**	0.184	**0.209**	0.184	**0.205**
SEQUOIA	0.282	**0.273**	0.207	**0.239**	0.204	**0.230**
AbMIL	0.272	**0.260**	0.232	**0.262**	0.218	**0.253**
2dMamba	0.271	**0.262**	0.253	**0.274**	0.246	**0.266**

COAD cohort (289 slides) is split into training and test sets with an 8:2 ratio. Following CPNN [37], we use Pearson and Spearman correlation coefficients (PCC and SCC) as evaluation metrics, which are standard in practical settings. We extended CPNN [37], utilizing two WSI foundation models to extract patch features: Conch [30] and CTransPath [46]. Each transcriptomic profile generates 512 synthetic tiles, which serve as a supplementary set to augment the original patch set. Extensive evaluations in Tab. 11 demonstrate that the synthetic tiles preserve gene expression information and further enhance clinical utility.

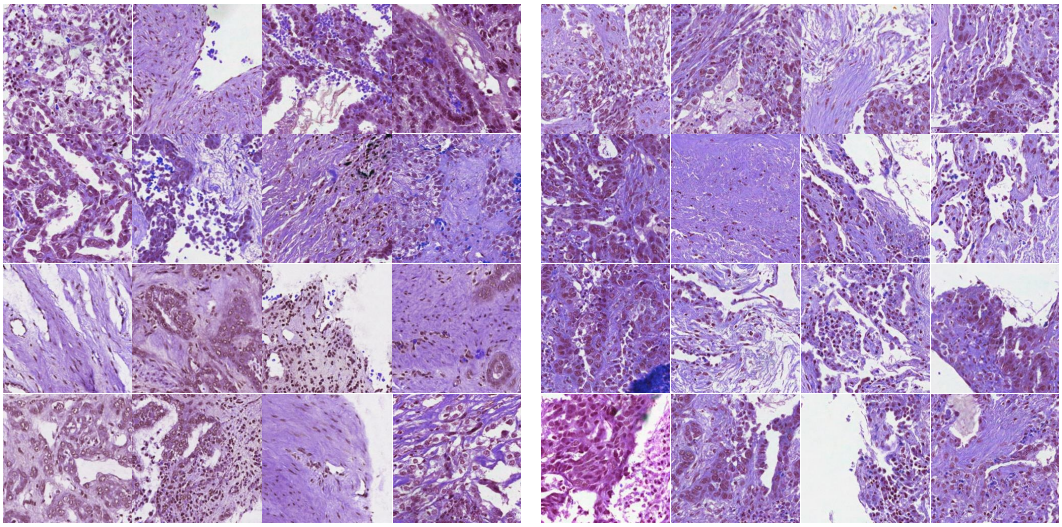
Comprehensive RNA Perturbation Analysis. Given the enrichment of haematopoietic transcriptional signatures, a higher representation of lymphocytic populations is anticipated in the generated tiles. Consistent with this expectation, we observe a systematic increase in median lymphocyte fraction in tiles conditioned on haematopoietic RNA-seq relative to those derived from bulk RNA-seq across multiple cancer types. As shown in Tab. 12, both TOPOSCAPE and GeneVAR consistently improve across all cohorts under perturbation. TOPOSCAPE achieves higher mean performance than GeneVAR in most settings, indicating a more responsive and biologically sensitive perturbation response.

Table 12: Comparison of TOPOSCAPE and GeneVAR under perturbation and non-perturbation settings across datasets. Values are reported as mean $\pm$ standard deviation.

Dataset	TOPOSCAPE		GeneVAR	
	w/o perturbation	w/ perturbation	w/o perturbation	w/ perturbation
COAD	$9.68\% \pm 11.55$	$19.94\% \pm 7.41$	$3.75\% \pm 12.17$	$10.42\% \pm 9.48$
LUAD	$8.70\% \pm 9.56$	$19.49\% \pm 8.91$	$9.45\% \pm 10.94$	$15.93\% \pm 16.65$
GBM	$8.45\% \pm 16.30$	$20.62\% \pm 9.44$	$4.63\% \pm 13.94$	$9.77\% \pm 8.94$
KIRP	$12.63\% \pm 14.39$	$18.49\% \pm 7.67$	$13.14\% \pm 13.96$	$17.51\% \pm 5.11$
CESC	$7.28\% \pm 9.79$	$19.03\% \pm 8.55$	$3.33\% \pm 10.11$	$10.58\% \pm 13.40$

A.4 Out-of-Distribution Generation

To evaluate the out-of-distribution generalization ability of TOPOSCAPE, we additionally introduce gene expression profiles from the GEO (Gene Expression Omnibus) repository. These external samples differ substantially from the training data in tissue processing procedures, sequencing platforms, and cohort composition, thereby constituting a distinct distributional setting. As illustrated in Fig. 7, TOPOSCAPE maintains stable morphological generation under these unseen conditions and produces structurally coherent, biologically consistent tissue patterns. These observations suggest that the learned correspondence between gene expression and histomorphology transfers beyond the training distribution and remains robust across heterogeneous datasets.



(a) Synthetic H&E-stained WSI tiles generated by TOPOSCAPE from GSE226069 RNA-seq profiles.

(b) Synthetic H&E-stained WSI tiles generated by TOPOSCAPE from GSM1228184 RNA-seq profiles.

Figure 7: **Synthetic H&E-stained WSI tiles** generated by TOPOSCAPE from out-of-distribution GEO RNA-seq profiles, including GSE226069 (lung cancer) and GSM1228184 (colorectal cancer).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately describe the proposed TOPOSCAPE, including rTPG and TopoRenderer, as well as the experimental scope across five TCGA benchmarks, consistent with the content presented in Sections 3–5 and Appendix A.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We clarify the limitations after Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors

should reflect on how these assumptions might be violated in practice and what the implications would be.

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not include theoretical results.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the experimental setup, datasets, and evaluation metrics. The appendix provides network architecture, training specifics, and hyperparameter details. Code will be provided upon acceptance.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.

- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code will be provided upon acceptance.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Experimental settings, including network architecture, optimizer, batch size, iterations, and hyperparameter search ranges, are detailed in the Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Tables 2, 3, and detailed appendix tables report results as mean and standard deviation over multiple runs.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The paper specifies the hardware used in the Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The study uses publicly available, de-identified biomedical datasets and focuses on algorithmic development for transcriptome-conditioned histology generation. The method is intended for research use, and generated images are not used for clinical diagnosis or decision making.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses potential benefits for molecularly informed pathology research, synthetic data augmentation, and privacy-conscious data sharing. It also notes that generated histology should not be used for clinical decision making without external validation and may require responsible release to mitigate misuse.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the existing datasets, models, and tools used in this work, including TCGA/GEO data resources, PanNuke, HoverNet, CONCH, and baseline methods. We follow their publicly available access policies and terms of use, and will include detailed license and access information in the released code and documentation.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The primary new asset is the source code for the proposed methods, which will be open source.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The research does not involve crowdsourcing experiments or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.