GeneCaDiff: Hierarchical Tissue Image Synthesis from Gene Expression via Multi-Stage Cascaded Diffusion Model

Weitian Huang[1] , **Shuaibo Gao**[2] , **Bing Liu**[2] , **Xiaoqi Sheng**[1] , **Jiazhou Chen**[3] and **Hongmin Cai**[1,2,*]

[1]School of Future Technology, South China University of Technology, Guangzhou, China
[2]School of Computer Science and Engineering, South China University of Technology, Guangzhou, China
[3]School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China
hmcai@scut.edu.cn

Abstract

The scarcity of paired gene expression and pathology images datasets poses a major bottleneck for training large-scale pathology foundation models. Although gene-to-image generative models offer a promising solution, existing methods typically employ coarse-grained conditional control strategies, resulting in entanglement between background textures and cell layouts. To address this challenge, we propose GeneCaDiff, a three-stage cascaded diffusion model that explicitly aligns the generative process with the hierarchical organization of biological tissues. Two conditional DDPMs independently synthesize tissue backgrounds and cellular foreground components, and a subsequent ControlNet-based fusion generator utilizes niche and cellular community maps as dual conditions to synthesize realistic tissue images. Quantitative and qualitative evaluations demonstrate strong realism and diversity, while controllability experiments validate hierarchical, decoupled control over niche textures and community layouts.

1 Introduction

The rapid advancement of computational pathology has been driven by the emergence of large-scale foundation models, which are capable of learning rich feature representations from histopathology images [Chen *et al.*, 2024; Vorontsov *et al.*, 2024]. Trained on extensive collections of whole-slide images, these models have demonstrated strong performance in disease characterization and prognostic stratification. However, their development is fundamentally constrained by the limited availability of high-quality multimodal datasets that align molecular measurements with tissue morphology [Acosta *et al.*, 2022]. In particular, spatial transcriptomics datasets paired with histological images remain scarce and costly to acquire [Moses and Pachter, 2022], restricting the ability of foundation models to fully capture the relationship between gene expression and tissue structure [He *et al.*, 2020; Lewis *et al.*, 2021].

To mitigate this data shortage, cross-modal gene-to-image synthesis has emerged as a promising solution [Zhang *et al.*, 2025]. By establishing a mapping from molecular profiles to pixel space, generative models function as robust data augmentation engines, creating synthetic training samples that enrich the diversity of existing datasets. Early efforts based on GANs and diffusion models have demonstrated the feasibility of generating histology images conditioned on gene expression [Carrillo-Perez *et al.*, 2023; Carrillo-Perez *et al.*, 2025]. Despite these advances, most existing approaches adopt a coarse global conditioning paradigm, in which high-dimensional gene expression vectors are compressed into a single latent representation to generate entire image patches. This monolithic approach leads to severe attribute entanglement. Specifically, the microenvironmental background and spatial organization of the cellular community are coupled within a unified generation process, preventing fine-grained and interpretable control over distinct tissue components.

Recent advances in spatial representation learning offer a pivotal insight to address this entanglement [Singhal *et al.*, 2024]. Foundation models for spatial transcriptomics, such as Nicheformer [Tejada-Lapuerta *et al.*, 2025], have empirically demonstrated that gene expression profiles inherently encode distinct structural information, ranging from microenvironmental niche identities to cellular densities. This suggests that the signal required to distinguish background textures from cellular layouts is already latent within the transcriptomic data. Consequently, rather than a monolithic synthesis, we posit that these dimensions can be effectively decoupled and modeled explicitly to mirror the natural hierarchy of biological tissues.

Inspired by this principle, we propose GeneCaDiff, a three-stage cascaded diffusion model tailored for spatial transcriptomics. The generative process is decomposed into biologically meaningful stages. First, a conditional diffusion model synthesizes niche level background textures based on gene expression in specific microenvironments. Second, another independent conditional diffusion model generates cell appearances and assembles them into a structured cellular community layout. Finally, a tissue level fusion model based on ControlNet integrates the synthesized background and layout using dual structural conditions to generate the final

*Corresponding author.

histopathological image. This design enables decoupled hierarchical control over tissue appearance while maintaining visual coherence and realism.

The main contributions of this work are summarized as follows:

- A biologically aligned multi-stage cascaded diffusion model is proposed that addresses multi-condition gene-to-image generation by decomposing the synthesis process into independently controllable stages, enabling hierarchical and decoupled control over tissue structure.
- The proposed model synthesizes histopathological images from spatial transcriptomics by explicitly modeling niche backgrounds and cellular community structures.
- Extensive quantitative and qualitative experiments on the HEST-1k dataset demonstrate that our method achieves superior realism and diversity, while providing effective orthogonal controllability.

2 Related Work

2.1 Gene to Image Synthesis

The task of gene to image synthesis aims to establish an explicit mapping between transcriptomic data and cellular or tissue images, thereby translating molecular level states into observable visual phenotypes. A typical strategy is to encode high dimensional expression vectors derived from bulk RNA-seq, single cell RNA-seq, or spatial transcriptomics into latent conditional representations. Subsequently, these representations are fed into generative adversarial networks, variational autoencoders, or diffusion models. Through learning correspondences between expression patterns and visual phenotypes in a cross modal representation space, these models synthesize pathological images consistent with underlying molecular states, serving as computational surrogates or complements to real tissue sections.

Early studies predominantly adopted GAN-based approaches [Yi *et al.*, 2019]. For example, RNA-GAN embeds gene expression in the conditional space of a GAN to generate tissue images associated with specific expression backgrounds [Carrillo-Perez *et al.*, 2023]. These approaches empirically support the feasibility of gene expression-driven image generation. However, GAN training is often sensitive to hyperparameters and adversarial balance. This sensitivity makes it prone to instability and mode collapse, which in turn degrades sample diversity and coverage of the expression space.

With diffusion models demonstrating more stable training behavior and higher generation fidelity in natural image synthesis [Dhariwal and Nichol, 2021; Kazerouni *et al.*, 2023], generation driven by diffusion-based gene-expression has attracted increasing attention. A representative work, RNA-CDM [Carrillo-Perez *et al.*, 2025], introduces a cascaded diffusion architecture for the task of RNA-seq in tissue image generation. It synthesizes high quality tumor image patches from RNA data via multistage diffusion and demonstrates advantages over GANs in both image quality and training stability. Nevertheless, most existing methods still follow a top down global conditioning paradigm. They directly generate patch level images from a single global expression vector without explicitly decomposing or controllably modeling multiscale structural factors such as niche backgrounds, single cell morphology, and cellular community layouts. Building upon prior gene driven generation studies, this work further introduces a hierarchical multistage design to enhance cross scale structural expressiveness and interpretability.

2.2 Controllable Image Generation

To achieve fine grained control over generated content, controllable generation methods typically introduce structured conditional signals such as bounding boxes, keypoints, and segmentation maps to explicitly constrain image content and structure. Within the domain of diffusion model architectures, representative approaches, including ControlNet [Zhang *et al.*, 2023] and T2I-Adapter [Mou *et al.*, 2024], provide commonly adopted solutions for structured control. ControlNet replicates the backbone of a pretrained diffusion model and appends a trainable control branch. This mechanism encodes external conditions like Canny edges, depth maps, or segmentation masks into multi scale features that are fused with the main network at multiple layers during denoising. This design injects structural constraints throughout the denoising process while largely preserving the generative capacity of the base model. In contrast, T2I-Adapter adopts a more lightweight adapter branch. It maps conditional information into intermediate representations compatible with diffusion models and injects them at a few key layers, thereby achieving structural control with reduced parameter count and training cost.

2.3 Multi Stage Diffusion Architectures

Multi stage diffusion architectures decompose complex generation tasks into multiple phases. This allows different stages to focus on distinct resolutions or semantic levels, thereby enhancing controllability and expressive power. Typical cascaded diffusion models adopt a coarse to fine resolution hierarchy. For example, the Cascaded Diffusion Models first generate low resolution images and then progressively refine them through a sequence of super resolution diffusion models [Ho *et al.*, 2022]. Similarly, SR3 [Saharia *et al.*, 2022b] and its successor Imagen [Saharia *et al.*, 2022a] employ diffusion based super resolution as a core component. These methods achieve higher fidelity and stronger conditional controllability in natural image generation.

Beyond resolution cascades, some studies further construct multi stage diffusion architectures along temporal or semantic dimensions. For example, eDiff-I [Balaji *et al.*, 2022] observes that early diffusion steps are more sensitive to textual conditions, while later steps emphasize visual details. Consequently, it partitions the diffusion timeline into multiple intervals and trains specialized expert denoisers for different noise regimes. Similarly, another multistage approach clusters noise levels and assigns each cluster to a distinct stage [Zhang *et al.*, 2024]. This system utilizes customized multi-decoder U-Net architectures, featuring a shared encoder and stage-specific decoders, to improve training and sampling efficiency while maintaining generation quality.

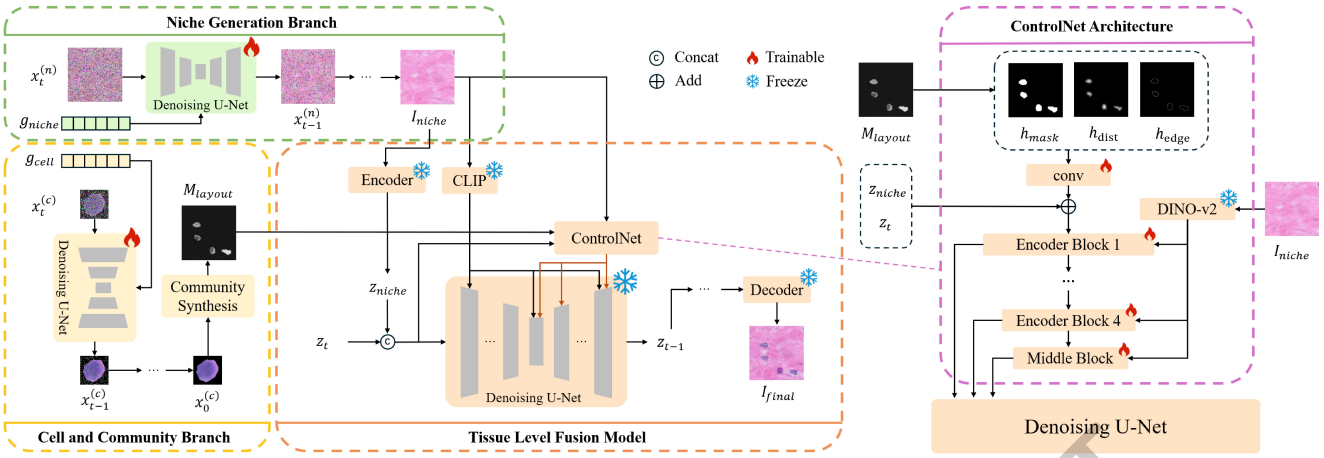

Figure 1: Schematic overview of GeneCaDiff. Three hierarchical components are designed to mimic tissue organization. First, the Niche Generation Branch synthesizes the microenvironmental background I_{niche} conditioned on the niche level gene expression g_{niche}. Second, the Cell and Community Branch generates individual cell appearances from g_{cell} and assembles them into a semantic layout map M_{layout}. Third, the Tissue Level Fusion Model integrates these outputs using a dual conditioning mechanism. The layout map is processed into structural hints containing masks and edges while the background provides texture guidance via DINO-v2 embeddings. The zoomed panel on the right details the ControlNet internals where these conditions guide the denoising process to produce the final image I_{final}.

The concept of multi-stage diffusion has also begun to emerge in computational pathology. For instance, RNA-CDM applies cascaded diffusion to map RNA sequencing data to whole slide tumor image tiles [Carrillo-Perez *et al.*, 2025]. This method synthesizes high quality WSI tiles from transcriptomic data and provides an important reference for pathological image synthesis driven by gene expression. However, stage partitioning in both natural image generation and pathological applications is primarily defined along the resolution or temporal axes. Furthermore, generation units typically remain at the level of entire patches or tiles. Such designs create difficulties in aligning diffusion stages explicitly with the hierarchical structure of tumor microenvironments including niche backgrounds, single cell morphology, and cellular community layouts. Building upon multistage diffusion architectures, this work aligns stage partitioning with biological hierarchy. We explicitly model niche backgrounds, cellular community structures, and tissue level rendering. Consequently, this enables fine grained and cross scale structural controllability in tissue image generation driven by gene expression.

3 Gene-guided Cascaded Diffusion Model

3.1 Methodology Overview

To address the challenge of attribute entanglement inherent in existing global conditioning methods, we propose a multi-stage cascaded diffusion architecture called GeneCaDiff. As illustrated in Figure 1, GeneCaDiff comprises three primary modules designed to mirror the hierarchical nature of tissue organization. First, the Niche Generation Branch is designed to reconstruct local microenvironmental backgrounds indicated by I_{niche} by interpreting niche level gene expression g_{niche}. Second, the Cell and Community Branch functions to synthesize individual cell appearances and assemble them into coherent spatial layouts, producing a structured map M_{layout}. Third, the Tissue Level Fusion Model acts as a unifying renderer. Using I_{niche} and M_{layout} as dual conditions, it produces the final high resolution histopathological image I_{final}. The synergistic operation of these components enables our approach to decouple the generation of stromal textures from cellular organizations, thereby ensuring both structural controllability and visual fidelity.

3.2 Niche Generation Branch

To reconstruct realistic microenvironmental backgrounds denoted as I_{niche} from niche level gene expression g_{niche}, we construct a conditional generative module based on the Denoising Diffusion Probabilistic Model [Ho *et al.*, 2020]. This branch serves to map the abstract molecular semantics of a niche into a visually coherent tissue texture, providing the fundamental stromal context for the subsequent composition.

The backbone of this module adopts a U-Net architecture enhanced with spatial transformer layers [Dhariwal and Nichol, 2021; Rombach *et al.*, 2022]. To effectively inject the molecular information, the input gene expression vector g_{niche} is first projected via a multilayer perceptron into a sequence of condition embeddings. These embeddings subsequently interact with the intermediate spatial features of the U-Net through cross attention mechanisms located at the downsampling, middle, and upsampling blocks. This design allows the gene expression condition to guide the feature synthesis across multiple resolution scales progressively.

To formally define the diffusion dynamics, we consider a forward process that gradually corrupts the ground truth image $x_0^{(n)}$ into a Gaussian noise distribution. The noisy image state $x_t^{(n)}$ at any arbitrary timestep t can be sampled directly using the closed form expression.

$$x_t^{(n)} = \sqrt{\bar{\alpha}_t} x_0^{(n)} + \sqrt{1 - \bar{\alpha}_t}\epsilon \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, I)$ and $\bar{\alpha}_t$ represents the cumulative noise schedule.

The network is optimized to reverse this corruption by estimating the noise component under the guidance of gene expression. The training objective minimizes the mean squared error between the actual noise and the predicted noise ϵ_θ, formulated as,

$$\mathcal{L}_{\text{niche}} = \mathbb{E}_{x_0^{(n)}, \epsilon, t, g_{niche}} \left[\|\epsilon - \epsilon_\theta(x_t^{(n)}, t, g_{niche})\|^2 \right] \quad (2)$$

During the inference phase, the model reconstructs the niche background I_{niche} by iteratively denoising a random Gaussian noise sample $x_T^{(n)}$. This reverse transition from step t to $t-1$ is guided by the conditional gene vector g_{niche} and is computed via the recurrence relation.

$$x_{t-1}^{(n)} = \frac{1}{\sqrt{\alpha_t}} \left(x_t^{(n)} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t^{(n)}, t, g_{niche}) \right) + \sigma_t \xi \quad (3)$$

where $\xi \sim \mathcal{N}(0, I)$ and σ_t is the variance term. Through T iterations of this sampling process, the system recovers the clean microenvironmental texture I_{niche}.

3.3 Cell and Community Branch

While the Niche Generation Branch provides the microenvironmental background, the Cell and Community Branch is responsible for synthesizing cellular appearance and organizing cells into biologically plausible spatial communities. This branch aims to model both the visual phenotype of individual cells and their collective spatial organization, thereby capturing mesoscopic tissue structure that is essential for realistic histopathological synthesis.

Given a high-dimensional single-cell gene expression vector g_{cell}, we first generate cell-level visual representations using a conditional denoising diffusion probabilistic model. The model architecture follows the same design principles as the niche generation branch, employing a U-Net backbone augmented with cross-attention layers. The gene expression vector is projected into a sequence of condition embeddings, which are injected into multiple layers of the denoising network via cross-attention. The network is trained to predict the noise component from the noisy cellular image state $x_t^{(c)}$ by minimizing the following objective function,

$$\mathcal{L}_{\text{cell}} = \mathbb{E}_{x_0^{(c)}, \epsilon, t, g_{cell}} \left[\|\epsilon - \epsilon_\phi(x_t^{(c)}, t, g_{cell})\|^2 \right] \quad (4)$$

Upon training, the model can synthesize specific cellular morphologies such as nuclear shape and chromatin texture that correspond to the underlying molecular state g_{cell}.

Once the individual cell library is synthesized, the challenge shifts to assembling these units into a biologically plausible community. We design a Hierarchical Placement Module that arranges the generated cells onto a canvas to produce the final semantic layout map M_{layout}. To accommodate different data availability scenarios, this module supports two distinct layout strategies.

In the absence of spatial coordinates, we simulate natural cellular aggregation patterns. This algorithm iteratively places cells based on geometric constraints. To ensure realistic density and avoid physical overlap, we utilize frequency domain cross correlation to identify valid candidate regions. Furthermore, a probability map based on distance fields is introduced to govern the balance between cellular clustering and dispersion. This mechanism ensures that the generated community exhibits structurally coherent boundaries rather than uniform randomness.

When spatial coordinate priors are available, the module directly utilizes these positions to place the synthesized cells. This mode enables precise and supervised reconstruction of tissue architecture. Consequently, the system can strictly adhere to specific spatial arrangements dictated by the user or ground truth data, resulting in a spatially accurate M_{layout} for the final fusion stage.

3.4 Tissue Level Fusion Model

To integrate the independent outputs from the preceding stages into a coherent whole, we introduce the Tissue Level Fusion Model. We select the Paint-by-Example architecture [Yang *et al.*, 2023] as the backbone, leveraging its pretrained generative priors. However, solely relying on its original CLIP-based conditioning is insufficient for the strict pixel-level alignment required in pathology. Consequently, we adopt a ControlNet-based strategy. We freeze the parameters of the main U-Net to preserve its learned priors. Consistent with the backbone design, the niche image I_{niche} undergoes a dual encoding process. Specifically, an AutoEncoderKL maps it into a background latent z_{niche} for subsequent concatenation with the noisy latent, while a CLIP encoder extracts its semantic embedding to preserve global consistency.

In parallel, we introduce a trainable ControlNet branch to enforce fine-grained structural and textural control. This branch is driven by a dual-conditioning mechanism. First, the layout M_{layout} is processed into a three-channel hint map h (masks, distance transforms, edges) and encoded via zero convolutions to lock geometric boundaries. Second, to capture the nuanced biological textures of the stroma, we employ a frozen DINO-v2 encoder to extract high-level stylistic embeddings from I_{niche}. Instead of simple concatenation, these embeddings act as global context cues and are injected into the ControlNet branch at multiple scales. To fuse these features, we implement a Cross Attention mechanism. In this operation, the intermediate spatial features of the ControlNet serve as the query Q. Meanwhile, the DINO-v2 embeddings serve as both the key K and value V via linear projections. The attention is calculated as,

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (5)$$

This operation maps the niche textures onto the layout structure.

Finally, the multi-scale features learned by the ControlNet are projected through zero convolution layers and added to the decoder features of the frozen main U-Net. This progressive injection ensures that the generated tissue I_{final} strictly adheres to the spatial constraints of M_{layout} while faithfully rendering the texture of I_{niche}.

The optimization objective is defined as,

$$\mathcal{L}_{\text{fusion}} = \mathbb{E}_{z_0,\epsilon,t,M_{layout},I_{niche}}\big[\|\epsilon - \epsilon_\varphi(z_t, t, M_{layout}, I_{niche})\|^2\big] \quad (6)$$

4 Experiments

This section evaluates the effectiveness of the proposed multi-stage diffusion architecture on the cross modal generation task mapping gene expression to histopathological images through both quantitative and qualitative analyses. We first describe the dataset construction and preprocessing procedures followed by the evaluation protocols and the baseline method. Subsequently, we report comparative results against this baseline approach and further analyze the controllability of the model with respect to microenvironmental backgrounds and cellular community layouts.

4.1 Datasets

We utilize the large scale multimodal HEST-1k dataset to validate our approach. This repository contains 1,108 spatial transcriptomics samples paired with H&E stained whole slide images covering 25 organs and 25 cancer subtypes across human and mouse species. It is distinguished by its rich alignment between gene expression profiles and morphological structures. To align with the high resolution requirements of our research, we specifically select a subset of 65 samples acquired via the Xenium platform for this study. To train the proposed three stage architecture, we construct specific ground truth pairs from these real tissue samples. We denote the original real H&E tissue patches as I^{gt}_{final}. Consistent with the notation defined in the method section, g_{niche} represents the niche level gene expression vector and g_{cell} denotes the single cell gene expression vector.

Niche Generation Data Preparation For the Niche Generation Branch, we derive ground truth background targets denoted as I^{gt}_{niche}. We apply the LaMa inpainting model [Suvorov *et al.*, 2022] to raw tissue patches I^{gt}_{final} to remove cellular structures effectively. The resulting supervision pairs consist of the niche gene expression g_{niche} and the processed background I^{gt}_{niche}. During inference, the model generates the predicted background I_{niche}.

Cell and Community Data Preparation For the Cell and Community Branch, single cell nuclear patches are cropped from the whole slide images to form the visual ground truth. These are paired with their corresponding single cell gene expression vectors g_{cell} to train the appearance synthesis module. Subsequently, the layout generation utilizes these synthesized cells to construct the community layout map M_{layout}.

Tissue Fusion Data Preparation The Tissue Level Fusion Model requires a robust alignment between generated conditions and real tissue targets. To achieve this, we employ a two phase training strategy. In the pretraining phase, the model learns to reconstruct the ground truth tissue image I^{gt}_{final} using oracle conditions derived from real data. These conditions include the inpainted background I^{gt}_{niche} and the ground truth cellular layout M^{gt}_{layout} extracted from real images. During the fine tuning phase, we bridge the domain gap by replacing these oracles with the model generated background I_{niche} from the Niche Generation Branch and the synthesized layout M_{layout} from the Cell and Community Branch, while maintaining the same supervision signal I^{gt}_{final}.

Method	Tissue Type	KID (10^2) ↓	LPIPS ↓
RNA-CDM	Breast Cancer	8.0996 ± 0.1658	0.6529 ± 0.1183
	Lung Diseased	7.0095 ± 0.1204	0.6594 ± 0.0984
	Bowel Healthy	8.0931 ± 0.4716	0.6752 ± 0.1846
	Bowel Cancer	7.3069 ± 0.1712	0.5848 ± 0.1130
	Bone Healthy	6.5750 ± 0.2841	0.7369 ± 0.1525
Ours	Breast Cancer	$\mathbf{5.9643 \pm 0.1377}$	$\mathbf{0.5896 \pm 0.1687}$
	Lung Diseased	$\mathbf{4.6245 \pm 0.1022}$	$\mathbf{0.6434 \pm 0.0903}$
	Bowel Healthy	$\mathbf{7.4193 \pm 0.2291}$	$\mathbf{0.6224 \pm 0.1198}$
	Bowel Cancer	$\mathbf{6.9507 \pm 0.1214}$	$\mathbf{0.5613 \pm 0.1045}$
	Bone Healthy	$\mathbf{3.6090 \pm 0.1028}$	$\mathbf{0.5792 \pm 0.1595}$

Table 1: Quantitative comparison with the baseline method across different data types.

4.2 Evaluation Metrics

To evaluate the generation quality comprehensively, we assess the model from two complementary perspectives focusing on distributional consistency and paired similarity. First, we employ the Kernel Inception Distance (KID) [Bińkowski *et al.*, 2018] to quantify the alignment between the generated and real data distributions in the deep feature space. Unlike FID, KID provides an unbiased estimator of the Maximum Mean Discrepancy (MMD) between Inception representations, making it more robust particularly given limited sample sizes. A lower KID score indicates superior overall realism and diversity. Second, to measure the fidelity of synthesized textures and local structures against the ground truth, we utilize the Learned Perceptual Image Patch Similarity (LPIPS) [Zhang *et al.*, 2018]. Calculated under a paired protocol, a lower LPIPS value signifies a higher degree of perceptual similarity between the generated images and their real counterparts.

4.3 Performance Evaluation

Quantitative Comparison

Given the nascent nature of the gene to image synthesis task, there is a scarcity of directly comparable open source architectures. Consequently, we focus our comparative analysis on RNA-CDM [Carrillo-Perez *et al.*, 2025] which represents the current state of the art in diffusion based histological generation. We confirm that the suggested multi-stage system works by testing it against this strong baseline on a diverse range of tissue types.

Table 1 presents the detailed comparative results stratified by tissue type. We evaluated the models on five distinct subsets: Breast Cancer, Lung Diseased, Bowel Healthy, Bowel Cancer, and Bone Healthy. The quantitative results show that our approach consistently outperforms the baseline across all tissue types, achieving lower KID and LPIPS scores in every subcategory. These consistent improvements demonstrate that the proposed hierarchical design generalizes robustly across diverse anatomical structures and both healthy

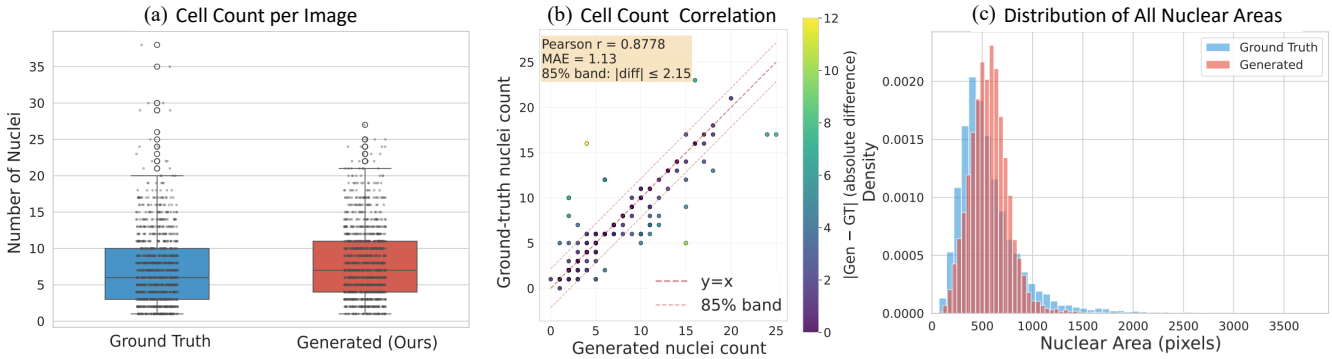

Figure 2: Evaluation of biological fidelity and consistency. (a) Box plots comparing the distribution of cell counts per image patch between Ground Truth and our Generated results. (b) Scatter plot showing the correlation between the number of nuclei in real versus synthesized images, with a Pearson correlation of 0.8778. (c) Kernel density estimation plots for nuclear area distributions, demonstrating that the synthesized cell morphologies closely match the physical properties of real tissues.

and pathological tissue conditions.

Beyond standard perceptual metrics, we extensively evaluated the biological fidelity of the generated tissues. This step is critical to ensure that the synthesized images are not just visually plausible but also clinically meaningful. We utilized a pretrained StarDist model to perform nuclei segmentation on both real and synthesized datasets. This allowed us to extract and compare key morphological features including cell counts and nuclear areas.

Figure 2 illustrates these biological consistency analyses. We first examined the cellular density preservation. The box plots in Figure 2a compare the cell count distributions per image patch. The results show that our model accurately reproduces the density statistics of the ground truth. The median values and interquartile ranges align closely between the real and generated sets. This indicates that the system avoids common generative issues such as overcrowding or sparse synthesis.

We further quantified this alignment through correlation analysis. Figure 2b plots the relationship between the ground truth and generated cell counts for paired samples. We observe a strong linear relationship characterized by a Pearson correlation coefficient of 0.8778. Additionally, the low Mean Absolute Error of 1.13 confirms that the model faithfully reconstructs the cellular organization from the input conditions. This implies that the system effectively preserves the spatial distribution of the community during the texture synthesis process, avoiding geometric distortions or semantic loss.

Finally, we assessed the geometric accuracy of individual cells. Figure 2c displays the kernel density estimation curves for nuclear areas. The distribution of the generated nuclei overlaps significantly with that of the real nuclei. This suggests that the model synthesizes cells with realistic sizes and shapes while effectively maintaining biological variance. These results collectively confirm that our approach better preserves key biological properties of tissue architecture.

Qualitative Comparison

Figure 3 illustrates the visual comparison involving three distinct categories defined as the ground truth I_{final}^{gt}, the baseline prediction $I_{baseline}$ generated by RNA-CDM, and the result I_{final} synthesized by our proposed architecture. While the baseline model generates visually plausible biological tissues, its predictions show considerable deviations from the specific ground truth. It produces inconsistent cell shapes and irregular spatial distributions. Additionally, the background stroma frequently appears fragmented. By comparison, although our method occasionally lacks sharp nuclear details and exhibits minor texture drift, it effectively maintains the overall structural integrity of the tissue. The synthesized images display clear cell boundaries and realistic community arrangements. Furthermore, the background stroma closely matches real H&E slides. These advantages directly result from our hierarchical strategy. By explicitly separating the cellular layout from the background, our design avoids the

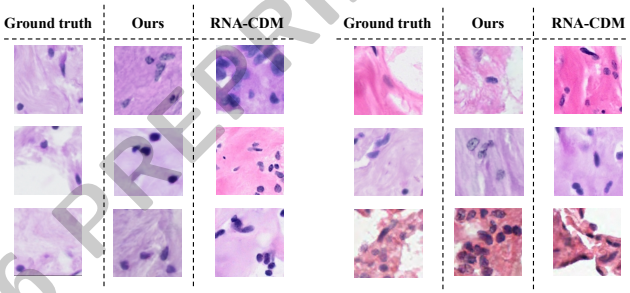

Figure 3: Qualitative visual comparison of synthesized histopathological images. The columns labeled as "Ground truth", "Ours", and "RNA-CDM" correspond to the real tissue targets (I_{final}^{gt}), our synthesized results (I_{final}), and the baseline predictions ($I_{baseline}$), respectively. Visual inspection reveals that images generated by the baseline method frequently exhibit considerable deviations in cellular morphology and spatial distribution, accompanied by visible structural fragmentation in the stromal background. In contrast, although our proposed architecture occasionally suffers from a lack of fine intra-nuclear details and minor texture drift, it faithfully restores the overall cellular phenotypes, spatial arrangements, and microenvironmental context, achieving a significantly higher visual alignment with the ground truth.

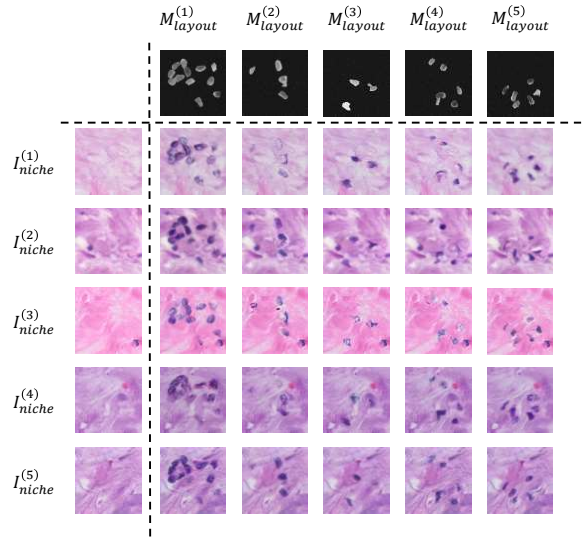

Figure 4: Visual demonstration of combinatorial controllability. The top row displays distinct cellular community layouts $M_{layout}^{(j)}$ and the left column presents representative niche backgrounds $I_{niche}^{(i)}$. The central grid contains the synthesized tissue images $I_{final}^{(i,j)}$ generated by the fusion model. This matrix illustrates the orthogonal decoupling capability of the architecture where rows exhibit consistent stromal textures with varying cellular arrangements and columns preserve spatial organizations under shifting microenvironmental contexts.

attribute entanglement common in standard global conditioning methods.

Controllability Analysis

A core objective of the proposed system is to achieve controllable generation over key structural factors. We focus on two observable structural dimensions defined as the niche background and the cellular community layout. The niche background represents the microenvironmental texture and style while the layout defines the spatial organization.

To evaluate this property, we perform a combinatorial grid analysis as shown in Figure 4. Specifically, for a given set of representative Niche backgrounds $I_{niche}^{(i)}$ and community layouts $M_{layout}^{(j)}$, we employ the trained Tissue Level Fusion Model to generate the corresponding matrix of synthesized tissues $I_{final}^{(i,j)}$. This grid allows the simultaneous observation of two distinct contrasts. First, we examine the effect of varying backgrounds while keeping the layout fixed. The generated images mainly change in global tone and stromal texture. However, the relative cell positions and community boundaries remain stable. Second, we vary the layout while fixing the background. In this scenario, the degree of cellular aggregation and spatial arrangement change significantly. Meanwhile, the background style and texture remain consistent. These observations indicate that the model achieves effective decoupling. The Niche background dominates the global style and the community layout governs the spatial organization. This demonstrates that these two factors exhibit orthogonal controllability during the generation process.

5 Conclusion

In this paper, we presented GeneCaDiff, a cascaded diffusion architecture designed to synthesize histopathological images from spatial transcriptomics data. Unlike global conditioning methods that often suffer from attribute entanglement, our system explicitly decomposes the generation process. We independently model the microenvironmental niche background and the cellular community layout. This decoupled design successfully aligns abstract molecular profiles with complex morphological phenotypes. Experimental evaluations on the HEST-1k dataset show that our approach demonstrates favorable performance over the state-of-the-art baseline. Our method achieves superior realism, evidenced by lower KID and LPIPS scores, and ensures high biological fidelity. Furthermore, GeneCaDiff enables precise, orthogonal control over tissue textures and spatial organizations. Consequently, this architecture provides a viable alternative for cross-modal data augmentation in computational pathology. Future work will leverage single-cell foundation models to extract interpretable embeddings. These semantic representations will enable more precise control over tissue image generation.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China (2024YFF1206600); in part by Guangdong S&T Programme (2025B0101130001); in part by the National Natural Science Foundation of China (62325204, 62572130, 62272326, 62502161); in part by the Guangdong Provincial Association for Science and Technology Youth Talent Cultivation Program (SKXRC2025375); in part by the China Postdoctoral Science Foundation (2025M782912).

Contribution Statement

Weitian Huang and Shuaibo Gao contributed equally to this work.

References

[Acosta *et al.*, 2022] Julián N Acosta, Guido J Falcone, Pranav Rajpurkar, and Eric J Topol. Multimodal biomedical ai. *Nature Medicine*, 28(9):1773–1784, 2022.

[Balaji *et al.*, 2022] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

[Bińkowski *et al.*, 2018] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

[Carrillo-Perez *et al.*, 2023] 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*, 3(8), 2023.

[Carrillo-Perez *et al.*, 2025] 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*, 9(3):320–332, 2025.

[Chen *et al.*, 2024] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024.

[Dhariwal and Nichol, 2021] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[He *et al.*, 2020] 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*, 4(8):827–834, 2020.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[Ho *et al.*, 2022] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022.

[Kazerouni *et al.*, 2023] Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, 88:102846, 2023.

[Lewis *et al.*, 2021] Sabrina M Lewis, Marie-Liesse Asselin-Labat, Quan Nguyen, Jean Berthelet, Xiao Tan, Verena C Wimmer, Delphine Merino, Kelly L Rogers, and Shalin H Naik. Spatial omics and multiplexed imaging to explore cancer biology. *Nature Methods*, 18(9):997–1012, 2021.

[Moses and Pachter, 2022] Lambda Moses and Lior Pachter. Museum of spatial transcriptomics. *Nature Methods*, 19(5):534–546, 2022.

[Mou *et al.*, 2024] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4296–4304, 2024.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[Saharia *et al.*, 2022a] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[Saharia *et al.*, 2022b] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022.

[Singhal *et al.*, 2024] Vipul Singhal, Nigel Chou, Joseph Lee, Yifei Yue, Jinyue Liu, Wan Kee Chock, Li Lin, Yun-Ching Chang, Erica Mei Ling Teo, Jonathan Aow, et al. Banksy unifies cell typing and tissue domain segmentation for scalable spatial omics data analysis. *Nature Genetics*, 56(3):431–441, 2024.

[Suvorov *et al.*, 2022] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2149–2159, 2022.

[Tejada-Lapuerta *et al.*, 2025] Alejandro Tejada-Lapuerta, Anna C Schaar, Robert Gutgesell, Giovanni Palla, Lennard Halle, Mariia Minaeva, Larsen Vornholz, Leander Dony, Francesca Drummer, Till Richter, et al. Nicheformer: a foundation model for single-cell and spatial omics. *Nature Methods*, pages 1–14, 2025.

[Vorontsov *et al.*, 2024] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, 30(10):2924–2935, 2024.

[Yang *et al.*, 2023] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023.

[Yi *et al.*, 2019] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, 58:101552, 2019.

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[Zhang *et al.*, 2023] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[Zhang *et al.*, 2024] Huijie Zhang, Yifu Lu, Ismail Alkhouri, Saiprasad Ravishankar, Dogyoon Song, and Qing Qu. Improving training efficiency of diffusion models via multi-stage framework and tailored multi-decoder architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7372–7381, 2024.

[Zhang *et al.*, 2025] Yuan Zhang, Xinfeng Zhang, Xiaoming Qi, Xinyu Wu, Feng Chen, Guanyu Yang, and Huazhu Fu. Content generation models in computational pathology: A comprehensive survey on methods, applications, and challenges. *IEEE Reviews in Biomedical Engineering*, 2025.