Semantic and Visual Crop-Guided Diffusion Models for Heterogeneous Tissue Synthesis in Histopathology

Saghir Alfasly Wataru Uegami MD Enamul Hoq Ghazal Alabtah H.R. Tizhoosh*
KIMIA Lab, Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN 55901
{Alfasly.Saghir, Tizhoosh.Hamid}@mayo.edu
https://kimialabmayo.github.io/hetero_tissue_diffuse_page/

Abstract

Synthetic data generation in histopathology faces unique challenges: preserving tissue heterogeneity, capturing subtle morphological features, and scaling to unannotated datasets. We present a latent diffusion model that generates realistic heterogeneous histopathology images through a novel dual-conditioning approach combining semantic segmentation maps with tissue-specific visual crops. Unlike existing methods that rely on text prompts or abstract visual embeddings, our approach preserves critical morphological details by directly incorporating raw tissue crops from corresponding semantic regions. For annotated datasets (i.e., Camelyon16, Panda), we extract patches ensuring $20-80\%$ tissue heterogeneity. For unannotated data (i.e., TCGA), we introduce a self-supervised extension that clusters whole-slide images into 100 tissue types using foundation model embeddings, automatically generating pseudo-semantic maps for training. Our method synthesizes high-fidelity images with precise region-wise annotations, achieving superior performance on downstream segmentation tasks. When evaluated on annotated datasets, models trained on our synthetic data show competitive performance to those trained on real data, demonstrating the utility of controlled heterogeneous tissue generation. In quantitative evaluation, prompt-guided synthesis reduces Fréchet Distance by up to $6\times$ on Camelyon16 (from 430.1 to 72.0) and yields $2-3$ lower FD across Panda and TCGA. Downstream DeepLabv3+ models trained solely on synthetic data attain test IoU of 0.71 and 0.95 on Camelyon16 and Panda, within $1-2\%$ of real-data baselines (0.72 and 0.96). By scaling to $11,765$ TCGA whole-slide images without manual annotations, our framework offers a practical solution for an urgent need for generating diverse, annotated histopathology data, addressing a critical bottleneck in computational pathology.

1 Introduction

Histopathology image analysis forms the cornerstone of cancer diagnosis, yet remains constrained by data scarcity, laborious annotation processes, and privacy concerns [8, 24, 29, 1, 21]. While generative AI has transformed natural image synthesis, its application to histopathological imaging in digital pathology faces unique challenges due to the complex, multi-scale architecture of biological tissues and the critical importance of preserving diagnostically relevant features [44, 16, 2]. Most of the current approaches have predominantly focused on generating homogeneous tissue types using text-based prompting systems, which introduce significant interobserver variability and limit clinical utility, particularly problematic in a domain where expert agreement is already inconsistent [15, 42].

The evolution from GAN-based methods to diffusion models has improved image quality and training stability [13, 18], but existing frameworks fail to account for the heterogeneous nature of real-

*Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

 https://doi.org/10.52202/085713-4185

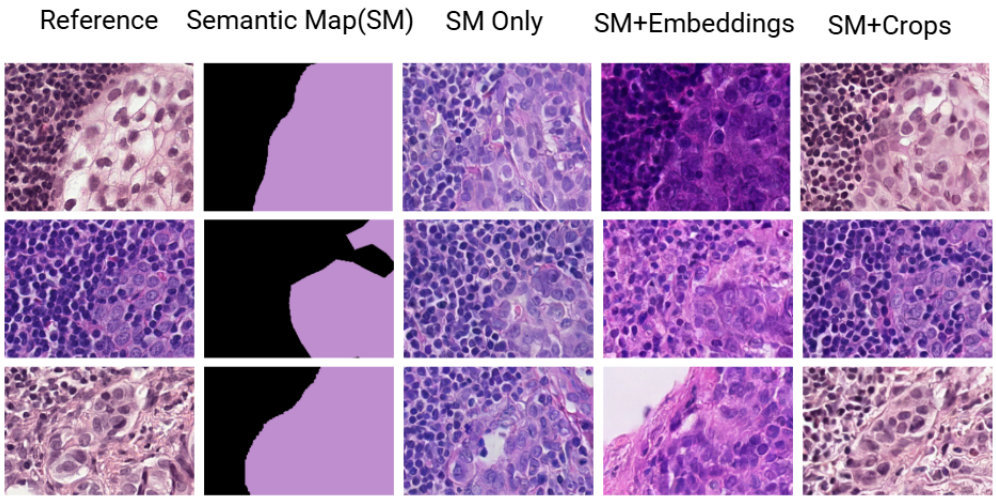


Figure 1: **Comparison of reference and synthetic Camelyon16 patches using three conditioning schemes.** From left to right: original histopathology patch; binary tumor–normal mask; generated image with a semantic segmentation map conditioning only; synthetic image by combining conditions of semantic maps and abstract embeddings; synthetic output of our model conditioned on the semantic map and tissue-specific visual crop prompts. The crop-guided generation recovers fine morphological details and staining heterogeneity more faithfully than embedding-based conditioning.

world histopathology samples. Levine et al. [28] demonstrated that GANs could generate images indistinguishable from real histopathology samples, while Moghadam et al. [30] marked the transition to diffusion models with superior image quality and training stability. Despite these advances, clinical specimens typically contain multiple tissue types and pathological features within a single slide, requiring region-specific control over generation parameters. This limitation severely restricts the utility of synthetic data for developing robust diagnostic algorithms that must recognize complex patterns in diverse tissue regions.

Further complicating this landscape is the tension between the global architecture of the tissue and the local cellular details. Traditional generative approaches struggle to maintain consistency across multiple magnification levels while preserving the fine-grained morphological features critical for diagnosis [34, 41, 2]. Recent work, URCDM [10], has addressed this through cascaded diffusion models generating images at multiple resolutions simultaneously, while *DiffInfinite* [4] enables arbitrary-size image synthesis with preserved long-range structural correlations. However, this multi-scale challenge remains particularly acute when attempting to synthesize realistic transitions between different tissue types, a capability essential for training segmentation algorithms and supporting differential diagnosis.

Our work addresses these challenges by introducing a visually prompted latent diffusion model designed specifically for heterogeneous tissue synthesis in histopathology. Unlike text-guided approaches, our framework leverages spatial masks and visual exemplars to provide fine-grained control over region-specific generation. Previous work, in the literature, NASDM [39] and subsequent extensions by Konz et al. [25] and Xu et al. [47] demonstrated the value of region-guided generation and semantic instance masks. Our method expands this concept to enable the synthesis of complex, multi-tissue samples with realistic transitions and clinically relevant features as shown in Figure 1. Our approach significantly reduces the annotation burden while maintaining high fidelity to real-world pathological presentations 2. The main contributions of our work include:

- **New Dual-Conditioning Architecture for Histopathology Synthesis.** We developed a unique visual crop-guided diffusion model that combines semantic maps with raw tissue exemplars, preserving critical morphological features (nuclear texture, staining patterns, cellular structure) that are lost in text-based or embedding-based approaches. This enables precise region-specific control while maintaining authentic tissue appearance for heterogeneous sample generation.

- **Self-Supervised Framework for Unannotated Whole-Slide Images.** We introduced a scalable solution for the TCGA dataset (11,765 WSIs) that automatically discovers and clusters 100 distinct tissue phenotypes without manual annotation. This democratizes access to diverse synthetic data across 33 cancer types while preserving patient privacy and addressing critical data scarcity in computational pathology.
- **Comprehensive Multi-Modal Validation Framework.** We established a rigorous evaluation pipeline combining quantitative metrics (Fréchet distance across 8 foundation model encoders) with downstream task performance (segmentation IoU scores). Most notably, our blinded assessment by certified pathologists using a 5-point Likert scale across multiple quality criteria revealed that our synthetic images were indistinguishable from real samples, with one pathologist commenting: "The generated images tended to have equal or higher quality than the real images."

Through rigorous evaluation involving both quantitative metrics (including the Fréchet Inception Distance [6]) and expert pathologist evaluation, we demonstrate that our approach generates histopathology images that are indistinguishable from real samples while providing *unprecedented control over tissue composition*. Synthetic datasets generated using our method effectively augment or replace real data in training diagnostic models, addressing the critical issue of data scarcity while preserving patient privacy.

By enabling the generation of diverse, annotated histopathology datasets without requiring patient data sharing, our framework represents a significant step toward more equitable and robust AI development in computational pathology [27]. This capability is particularly valuable for rare cancer types and underrepresented populations, potentially democratizing access to high-quality training data between institutions, regardless of their size or resources.

2 Related Work

Generative Models in Histopathology. Visual generative models have evolved from early GANs [17] to sophisticated diffusion models [18], with recent advances like ControlNet [49] enabling fine-grained control. Adapting these to histopathology presents unique challenges due to the complexity of the tissue and the diagnostic significance of subtle morphological features. Although early GANs demonstrated feasibility [28, 34, 2], recent diffusion-based approaches show superior quality [30]. Domain-specific methods [26, 23, 51, 10] have emerged, with URCDM [10] addressing multi-resolution synthesis and enabling arbitrary-size generation [4, 43]. However, most approaches generate homogeneous tissue types, limiting their utility for training diagnostic models that require heterogeneous tissue representations, a limitation that our framework specifically addresses.

Conditioning Mechanisms for Histopathology Synthesis. Existing conditioning approaches fall into three categories, each with significant limitations. Unconditioned models [44, 16] produce realistic images but lack control over tissue types and pathological features, severely limiting their utility for training task-specific models. Metadata-guided, text-guided, or mask-guided models [14, 48, 47, 9] suffer from interobserver variability, as documented by Elmore et al. [15] who found substantial disagreement among pathologists (kappa values as low as 0.48). Visual embedding or RNA-seq embedding approaches [33, 12, 50] avoid text ambiguity, but introduce lossy transformations that can obscure critical diagnostic features. Even domain-specific embeddings suffer from information loss during dimensionality reduction. Our approach circumvents these limitations by directly conditioning on *real tissue crops combined with semantic maps*, preserving original visual characteristics without intermediate representations.

Semantic Map-Based Generation. Recent works have explored semantic map conditioning for precise spatial control. Shrivastava and Fletcher [39] pioneered this with NASDM, while Konz et al. [25] extended it through random mask ablation. Although spatially accurate, these approaches typically focus on single tissue types or cellular structures rather than heterogeneous tissue architectures. Our dual-condition mechanism combines semantic maps with visual crops, enabling the synthesis of diverse tissue compositions while maintaining both spatial accuracy and morphological fidelity, essential for generating comprehensive segmentation datasets.

Large-Scale Synthesis and Evaluation. Whole-slide image (WSI) synthesis presents unique challenges due to gigapixel resolution and structural dependencies. Cechnicka et al. [10] and Aversa et al. [4] addressed scale issues through cascaded models and infinite tiling, respectively, but struggled to maintain segmentation accuracy with visual characteristics. Our self-supervised framework in

 https://doi.org/10.52202/085713-4185

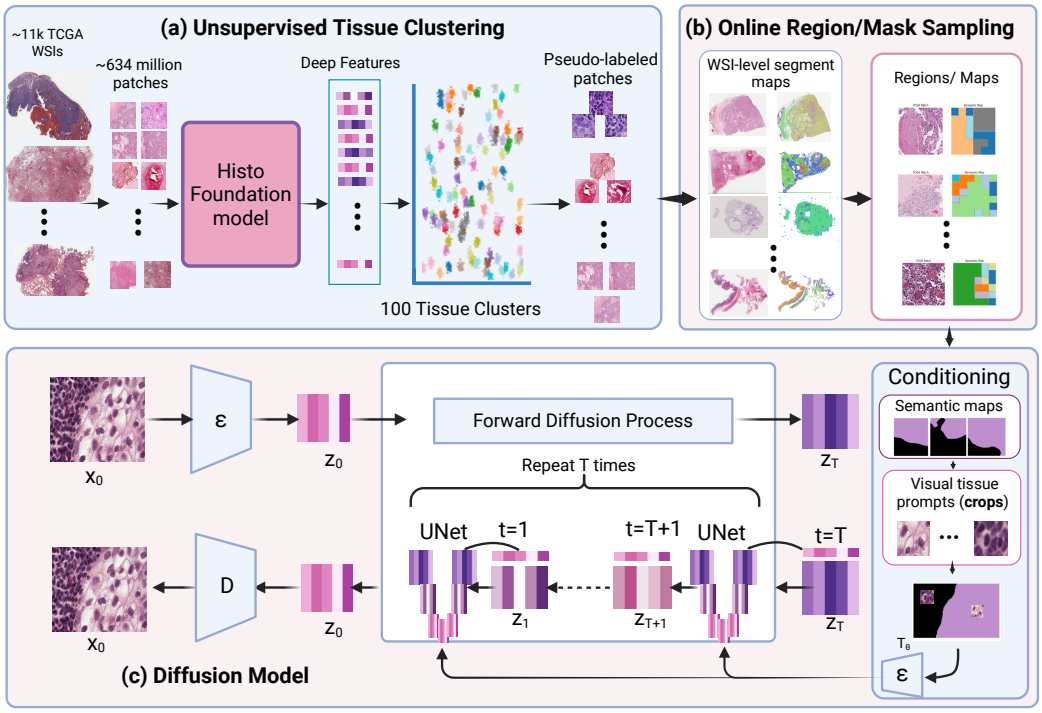


Figure 2: **Schematic overview of the HeteroTissue-Diffuse framework for heterogeneous tissue synthesis in histopathology.** (a) Unsupervised Tissue Clustering: For unannotated datasets (TCGA), approximately 11,765 whole-slide images (WSIs) are processed to extract 634,435,134 million patches. A histopathology foundation model extracts deep features, which are used to cluster patches into 100 distinct tissue types, creating pseudo-labeled data. (b) Online Region/Mask Sampling: The pseudo-labeled patches are used to generate WSI-level segmentation maps and regional masks for conditioning the diffusion model. (c) Diffusion Model: Our dual-conditioning approach combines semantic maps with visual tissue prompts (crops) to guide the latent diffusion process. The model encodes the input image to the latent space, applies forward diffusion to create a noisy latent, then reverses this process with UNet denoising conditioned on both semantic maps and visual tissue exemplars. For annotated datasets (Camelyon16 [5] and Panda [7]), only component (c) is used with their existing semantic maps.

this paper, trained on TCGA's 11,765 diagnostic WSIs, generates 100 distinct tissue types while preserving region-specific control through dual conditioning.

Evaluation of synthetic histopathology requires specialized metrics beyond standard FID and IS, which use networks pretrained on natural images. Domain-specific alternatives include Fréchet Distance and Topological Fréchet Distance [47], complemented by the evaluation of expert pathologists [40]. Our comprehensive validation combines these metrics with downstream task performance, the ultimate measure of the utility of synthetic data.

Our work advances the field through three key innovations. First, we avoid intermediate representations that compromise fidelity by directly conditioning on visual crops from real histopathology images combined with spatial semantic maps. Unlike semantic-only methods [39, 25], text-based approaches [36], or embedding-based techniques [31, 35], our approach preserves tissue-specific attributes (i.e., texture, cellular morphology, and staining patterns) that abstract representations lose. This direct visual conditioning enables the synthesis of realistic heterogeneous samples with precise spatial control, which is essential to train robust diagnostic models. Second, our self-supervised extension to TCGA democratizes access to diverse synthetic data across cancer types, generating 100 distinct tissue phenotypes without manual annotation while addressing critical data scarcity and preserving patient privacy [27]. Third, we establish comprehensive validation through rigorous metrics combining expert pathologist assessment with quantitative measures (Fréchet distance, precision,

recall, F1 score), and demonstrate utility through downstream segmentation tasks that confirm the high quality of our generated annotated data.

3 Method

Our approach addresses the fundamental challenge in histopathology synthesis: generating realistic heterogeneous tissue images with precise region-based control while preserving morphological fidelity. We achieve this through a novel dual-conditioning latent diffusion model that combines semantic maps with tissue-specific visual crops.

3.1 Preliminaries: Latent Diffusion Models

Latent diffusion models (LDMs) [36] offer an efficient framework for high-quality image synthesis by operating in a compressed latent space. Given an image $x_0 \in \mathbb{R}^{H\times W\times 3}$, an encoder $\mathcal{E}$ maps it to a lower-dimensional representation $z_0 = \mathcal{E}(x_0) \in \mathbb{R}^{h\times w\times c}$, where generally $h \ll H$ and $w \ll W$.

The forward diffusion process gradually corrupts z_0 by adding Gaussian noise over T time steps:

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t}z_{t-1}, \beta_t\mathbf{I}) \quad (1)$$

where $\{\beta_t\}_{t=1}^T$ is a predefined variance schedule. This can be expressed in closed form as

$$z_t = \sqrt{\bar{\alpha}_t}z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0,\mathbf{I}) \quad (2)$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$.

Algorithm 1: Heterogeneous Patch Sampling for Annotated Data

Require: Dataset $\mathcal{D}$ with patches and segmentation masks
Require: Tissue ratio bounds $[r_{\min}, r_{\max}] = [0.2, 0.8]$
Require: Crop size range $[d_{\min}, d_{\max}] = [50, 200]$
Ensure: Training sample (x, c)
function SAMPLEHETEROGENEOUSPATCH($\mathcal{D}$)
  $(x, M) \leftarrow$ SelectPatch($\mathcal{D}$) where tissue ratios $\in [r_{\min}, r_{\max}]$
  $K \leftarrow$ number of tissue classes
  Initialize $C \leftarrow$ zeros$(H, W, 3K)$
  for $k = 1$ to K **do**
    $C[:, :, (k-1)] \leftarrow M[:, :, k]$ ▷ Semantic channel
    if $\sum M[:, :, k] > 0$ **then** ▷ Class k present
      $d \leftarrow$ RandomInt$(d_{\min}, d_{\max})$
      $p_k, (r, c) \leftarrow$ ExtractSquareCrop$(x, M[:, :, k], d)$
      $p_k \leftarrow$ Augment(p_k) ▷ Optional rotation/flip
      $C_k \leftarrow$ zeros$(H, W, 3)$
      $C_k[r : r+d, c : c+d, :] \leftarrow p_k$
      $C[:, :, K + (k-1) * 3 : K + k * 3] \leftarrow C_k$
    end if
  end for
  return (x, c)
end function

A neural network ϵ_θ learns to reverse this process by predicting the noise component, optimizing

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{z_0,t,\epsilon}\left[\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2\right] \quad (3)$$

where c represents optional conditioning information.

Existing histopathology synthesis methods typically depend on class labels, text descriptions, or global feature embeddings. However, these approaches struggle with two critical requirements: (1) precise spatial control over tissue types and (2) preservation of fine-grained morphological details such as nuclear texture and staining patterns.

3.2 HeteroTissue-Diffuse: Our Dual-Conditioning Approach

We introduce HeteroTissue-Diffuse (HTD), which fulfills both requirements through a new dual-conditioning mechanism. Our key insight is that semantic maps provide spatial precision while tissue-specific visual crops preserve morphological authenticity.

3.2.1 Dual Conditioning Formulation

Given a histopathology patch $x \in \mathbb{R}^{H\times W\times 3}$ and its semantic segmentation map $M \in \{0,1\}^{H\times W\times K}$ (where K denotes tissue classes), we construct our conditioning signal as detailed in Alg.1:

$$c = \text{concat}(M, C_1, ..., C_K), \quad (4)$$

where each $C_i \in \mathbb{R}^{H\times W\times 3}$ is a sparse visual crop tensor for tissue class i. The construction process involves:

1. Extract a square crop p_i of size $d \times d$ (where $d \in \{50, \ldots, 200\}$ pixels) from a region labeled as class i 2. Initialize a zero tensor C_i matching the full patch dimensions 3. Place p_i at random coordinates within the semantic region defined by M_i

This design preserves tissue-specific attributes (texture, cellular morphology, staining) that are lost in abstract representations while maintaining spatial correspondence with the semantic map.

https://doi.org/10.52202/085713-4185

3.3 Self-Supervised Extension for Unannotated WSIs

A critical challenge in histopathology synthesis is the lack of pixel-wise annotations at scale. While datasets like *Camelyon16* provide detailed segmentation, they cover a limited number of tissue types and cancer subtypes. The Cancer Genome Atlas (TCGA), containing over $11,765$ whole-slide images across 33 cancer types, offers unprecedented diversity but lacks segmentation annotations. Our self-supervised extension bridges this gap by automatically discovering tissue phenotypes and creating pseudo-annotations that enable HTD training on this extensive public resource.

3.3.1 Tissue Type Discovery via Deep Clustering

Our approach leverages the semantic richness of foundation models pre-trained on histopathology data. These models learn representations that naturally cluster similar tissue types, which we exploit for unsupervised tissue discovery. The process involves three carefully designed phases that balance computational efficiency with comprehensive tissue representation.

In the strategic feature extraction phase, we process each WSI $w \in \mathcal{W}$ by extracting non-overlapping patches at the highest available magnification of each WSI. After applying tissue detection to exclude background regions, we compute features $f_\phi(p)$ for each patch p using a foundation model such as UNI [11]. It took 3 months to extract the entire TCGA 224×224 patch embeddings (i.e., $634,435,134$ patches) of the high magnification of each WSI on 1 x NVIDIA A100 GPUs with 80GB. To ensure diversity while maintaining computational tractability, we strategically sample $N = 1000$ patches per WSI using a diversity-aware sampling strategy given as

$$P_{\text{sample}} = \text{DiversitySample}(P_w, N, \text{spatial_weight} = 0.3, \text{feature_weight} = 0.7), \tag{5}$$

where spatial weighting ensures coverage across the WSI and feature weighting promotes phenotypic diversity. This approach prioritizes edge cases and underrepresented regions that might contain rare but clinically significant tissue types.

The hierarchical clustering phase employs a two-stage approach to discover tissue phenotypes. Initially, we apply k-means clustering with $K = 100$ clusters on the collected features from all sampled patches across the dataset. Subsequently, for clusters exhibiting high intra-cluster variance, we perform sub-clustering to identify subtle phenotypic variations. This hierarchical approach captures both major tissue categories such as tumor, stroma, and necrosis, as well as finer distinctions such as different grades of tumor differentiation or varying inflammatory patterns.

For the multi-scale semantic map generation, we create representations at multiple granularities for each WSI as

$$S_k = \{\text{AssignCluster}(p, \mathcal{C}_k) : p \in \text{Patches}(w)\}, \tag{6}$$

where $k \in \{5, 10, 20, 50, 100\}$ represents different levels of tissue granularity. This multiscale representation enables HTD to learn both coarse tissue boundaries and fine-grained morphological variations, adapting to the complexity of different tissue regions within the same slide.

3.3.2 Adaptive Heterogeneous Region Sampling

Our TCGA sampling strategy guarantees tissue heterogeneity while maintaining computational efficiency. We introduce an adaptive framework to ensure every training sample contains meaningful tissue diversity, avoiding homogeneous regions that fail to capture critical tissue interactions.

We compute heterogeneity maps for each WSI using entropy to quantify tissue diversity. For region r with cluster distribution, the heterogeneity score is given as

$$H(r) = -\sum_{i=1}^{k} p_i(r) \log p_i(r), \tag{7}$$

where $p_i(r)$ represents the proportion of cluster i in region r. This identifies regions with rich tissue interactions like tumor-stroma interfaces.

If there are insufficient heterogeneous regions in the current granularity, the algorithm adapts by decreasing the size of the region or increasing the granularity of the cluster k, ensuring that every sampled region contains at least two distinct tissue types.

For selected regions, we construct multi-scale visual crops with dimensions adapting to tissue complexity:

$$d_i = d_{\text{base}} \cdot (1 + \alpha \cdot \text{ComplexityScore}(i)). \tag{8}$$

Complex tissues receive larger crops to capture full morphology. Strategic placement maximizes information by centering on representative regions for homogeneous clusters and sampling boundaries for heterogeneous ones.

Tissue-aware augmentations include stain variations for batch effects, controlled rotations respecting tissue orientation, and brightness adjustments mimicking scanner variations. Dynamic cluster granularity follows curriculum learning:

$$k'(t) = k_{\text{min}} + (k_{\text{max}} - k_{\text{min}}) \cdot \min(1, t/T_{\text{warmup}}). \tag{9}$$

This progression from coarse to fine tissue distinctions prevents early overfitting while ensuring full phenotypic coverage. Complete algorithms for sampling and clustering, along with detailed implementation specifications, are provided in the supplementary materials.

3.4 Tissue Classifier in Inference Phase

To enhance computational efficiency during inference, a lightweight tissue classification model was implemented following the clustering of TCGA images. While the initial clustering utilized computationally intensive foundation models (UNI in our case) to generate embeddings, applying this same approach during inference would create a significant computational bottleneck. Instead, a more efficient ViT-small architecture was trained directly on the pseudo-labeled clusters, enabling rapid tissue type classification without requiring foundation model embedding extraction or centroid matching. This classifier processes 224×224 visual crop inputs and directly predicts cluster assignment from the 100 identified tissue types, reducing inference computational requirements by approximately 85% compared to the original embedding-based approach. The model was trained on 514,029 patches extracted from $11,765$ diagnostic TCGA WSIs using AdamW optimization with learning rate $1e-3$, achieving 47% accuracy on the held-out test set. This approach substantially streamlines the inference pipeline while maintaining classification fidelity, enabling practical deployment in resource-constrained environments. More details of this cluster classifier training and implementation are provided in the supplementary file.

Overall, our method uniquely combines the spatial precision of semantic maps with the morphological authenticity of visual crops, creating a dual-conditioning approach that addresses key limitations of existing methods. Unlike text-based conditioning, we avoid ambiguity and inter-observer variability; unlike global feature conditioning, we preserve fine-grained tissue characteristics; and unlike semantic-only approaches, we capture staining variations and cellular details. The self-supervised extension to TCGA demonstrates scalability to massive unannotated datasets, opening possibilities for diverse tissue synthesis across cancer types.

4 Results

Quantitative Results - Fidelity Fréchet Distance (FD). We evaluated the fidelity of generated histopathology images using Fréchet Distance (FD) across multiple foundation model encoders on CAMELYON16, PANDA, and TCGA datasets (Table 1). The results demonstrate that prompt conditioning significantly improves generation quality compared to nonprompt (NP) baseline across all datasets. Notably, RN50-BT shows the most dramatic improvement, with FD scores decreasing from 430.1 to 72.0 on CAMELYON16 when using prompts—a 6-fold reduction. Similarly, DINOv2 and UNI2 encoders exhibit substantial improvements with prompt conditioning, achieving 2-3× lower FD scores. The embedding prompt approach shows intermediate performance, suggesting that direct visual-crop prompts provide more effective semantic guidance than crop-based embeddings. Interestingly, the improvement magnitude varies across datasets, with PANDA showing the most consistent gains across all encoders. These findings validate that semantic conditioning through prompts enables more faithful tissue structure generation, with certain encoder architectures (RN50-BT, DINOv2) being particularly responsive to textual guidance. Comprehensive ablation studies, per-class FD analysis, and architectural comparisons are provided in the supplementary materials.

Downstream Evaluation - Tissue Segmentation. We evaluated our synthetic datasets on tissue segmentation tasks using DeepLabv3+ on Camelyon16 and Panda datasets, as shown in Table 2 and Figure 3.

The results demonstrate a significant milestone for generative models in medical imaging: Synthetic data with proper conditioning achieve segmentation performance remarkably close to real

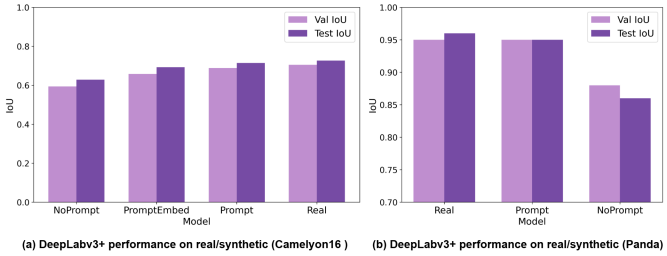
 https://doi.org/10.52202/085713-4185

Table 1: FD Results for visual-crop Prompt, Nonprompt, and crop embedding prompt conditions across CAMELYON16 [5], PANDA [7], and TCGA [45] datasets

Dataset	Cond.	Lunit-8 [22]	GigaPath [46]	H-Optimus-0 [38]	PathDino [3]	RN50-BT [22]	DINOv2 [32]	UNI2-H [11]	UNI [11]
CAM16	NP	1360.9	714.0	713.9	7540.6	430.1	122.0	139.8	70.0
	Emb. Prompt	991.3	606.6	664.7	4331.1	183.0	289.6	141.6	841.1
	Visual Prompt	629.1	353.0	425.2	2591.5	72.0	52.7	85.2	481.4
PANDA	NP	877.8	347.3	422.2	5124.7	150.0	352.4	113.6	650.5
	Visual Prompt	512.2	139.7	227.1	3230.9	22.8	61.4	52.4	299.9
TCGA	NP	855.1	360.4	476.0	4306.7	157.7	117.5	119.6	563.6
	Visual Prompt	821.9	346.1	521.4	3876.7	142.9	142.1	135.1	527.9

Table 2: Test IoU performance of DeepLabv3+ trained on real and synthetic data variants across Camelyon16 and Panda datasets.

Data	Cam16	Panda
NoPrompt	0.63	0.86
PromptEmbed	0.69	0.88
Visual Prompt	0.71	0.95
Real	0.72	0.96



(a) DeepLabv3+ performance on real/synthetic (Camelyon16) (b) DeepLabv3+ performance on real/synthetic (Panda)

Figure 3: Validation and test IoU comparison of DeepLabv3+ models across different training dataset types on (a) Camelyon16 and (b) Panda. Training on synthetic data generated with a visual prompt achieves performance comparable to real data training, with NoPrompt showing lower performance.

data. Specifically, Prompt-based synthetic training achieved test IoU scores of 0.71 and 0.95 on Camelyon16 and PANDA, respectively, compared to 0.72 and 0.96 with real data training, a gap of merely 1-2%. This near-parity performance is particularly noteworthy as our objective extends beyond data augmentation to complete replacement of real patient data, addressing critical privacy concerns in medical AI development. The inclusion of visual-crop prompts or prompt embeddings proves essential, as NoPrompt synthetic data shows a more substantial performance drop (0.63 and 0.86), highlighting the importance of semantic guidance in generating task-relevant synthetic samples. These findings suggest that carefully conditioned generative models can produce training data of sufficient quality to potentially eliminate the need for real patient data when developing robust segmentation models, representing a crucial advancement toward privacy-preserving medical AI. Additional experimental results, ablation studies, and cross-dataset generalization analyses are provided in the supplementary materials.

Qualitative Evaluation - Certified Pathologist Assessment. To complement the quantitative metrics and downstream task performance, a comprehensive pathologist evaluation was conducted to assess the clinical realism and diagnostic utility of synthetic images. The evaluation employed a blinded assessment framework where expert pathologists reviewed 120 randomly selected images from both real and synthetic datasets without prior knowledge of their origin. The assessment interface in Figure S6, a web application presented pathologists with five evaluation criteria: overall image quality, histological structural detail, nuclear morphology accuracy, presence of artifactual hallucinations, and a final determination of image authenticity. Each quality metric was rated on a 5-point Likert scale, while hallucination presence and real/synthetic classification were binary assessments. The evaluation protocol encompassed images from three datasets (CAMELYON16, PANDA, and TCGA), with equal representation of real and synthetic samples to ensure an unbiased assessment. After collecting responses from the certified pathologist, statistical analysis was performed to quantify the perceptual quality and clinical validity of synthetic images. Figure 4 presents the aggregated results in the three quality metrics, demonstrating that the synthetic images generated using visual prompt conditioning achieved scores comparable to real histopathological images, with a particularly strong performance in the preservation of nuclear details and overall structural integrity. The minimal difference in scores between real and synthetic images in all datasets validates the clinical relevance of the generated samples, while the low variance in the assessments indicates consistent quality between different types of tissue and pathological conditions. The general comment of the pathologist is "***The two types of images were indistinguishable even for me. Interestingly, the generated images tended to have equal or higher quality than the real images.***"

https://doi.org/10.52202/085713-4185

5 Discussion

The results demonstrate that visual conditioning mechanisms are fundamental to achieving high-quality histopathology synthesis. The 6-fold improvement in FD scores for RN50-BT on CAMELYON16 when using visual prompts versus nonprompt generation indicates that direct visual conditioning preserves critical morphological features lost in abstract representations. Embedding-based prompts showed intermediate performance, confirming that transformation to learned representations introduces information loss. This effect varied across foundation models, with RN50-BT and DINOv2 exhibiting superior responsiveness to visual conditioning. The framework achieves near-parity performance with real data in downstream tasks, with segmentation IoU scores differing by only 1-2% between synthetic and real training data. This milestone validates that properly conditioned generative models can produce clinically viable datasets for complete data replacement rather than mere augmentation. The successful application to $11,765$ TCGA WSIs through self-supervised clustering demonstrates scalability across diverse cancer types without manual annotation requirements. The evaluation methodology integrates three complementary assessments: Fréchet Distance measures distributional similarity and image realism, downstream segmentation quantifies practical utility for diagnostic model training, and expert pathologist evaluation identifies subtle artifacts beyond automated metrics.

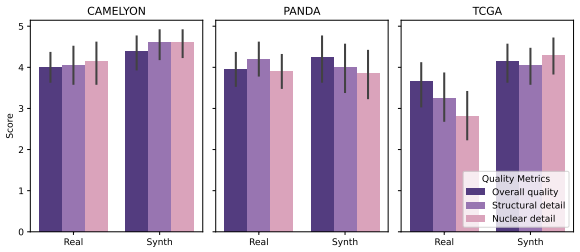


Figure 4: **Expert pathologist evaluation of synthetic versus real histopathology images across three datasets**. Mean scores (±SD) on a 5-point Likert scale for overall quality, structural detail, and nuclear morphology. Blinded assessment of 120 images shows synthetic images generated with visual prompt conditioning achieve comparable scores to real samples across CAMELYON, PANDA, and TCGA datasets.

5.1 Conclusion

The convergence of generative AI and computational pathology presents an unprecedented opportunity to revolutionize medical AI development. HeteroTissue-Diffuse demonstrates that synthetic data generation can transcend augmentation to enable complete replacement of real patient data while maintaining diagnostic accuracy, a paradigm shift that addresses fundamental challenges of privacy, scarcity, and equity in medical AI. As we approach the threshold of this transformation, the medical AI community must embrace visual generative models not merely as technical tools but as catalysts for democratizing access to high-quality training data across institutions worldwide. The path forward demands bold innovation in multi-modal synthesis, cross-institutional collaboration, and the development of foundation models that capture the full complexity of human pathology, ultimately realizing the vision of AI-driven precision medicine accessible to all.

Broader Impacts. HeteroTissue-Diffuse has the potential to democratize access to high-quality annotated histopathology data across institutions regardless of size or resources, particularly benefiting underserved regions with limited pathology expertise. By enabling the generation of synthetic data with precise region-specific annotations, our framework could accelerate the development of AI diagnostics for rare cancer subtypes where data scarcity has previously hampered progress. Moreover, this technology offers a pathway to international research collaboration without compromising patient privacy regulations.

Limitations. Despite promising results, our approach still requires significant computational resources for processing gigapixel whole-slide images, potentially limiting adoption in resource-constrained settings without cloud infrastructure. The current implementation focuses exclusively on H&E stained images and would require adaptation to handle other staining protocols or imaging modalities used in clinical practice. Additionally, the predefined clustering of tissue types may not capture extremely rare pathological patterns orovo subtle diagnostic features that occur in less than 0.1% of cases.

Acknowledgments. The authors would like to thank Joaquin Garcia, Saba Yasir, Man M. Ho, Sahar Rahimi Malakshan, Daniel Stone, Jeff Fetzer, Peyman Nejat, Bardia Khosravi, and Bassel Al Omari for the fruitful discussions. This work was supported in part by Mayo Clinic Comprehensive Cancer Center.

https://doi.org/10.52202/085713-4185

References

[1] Eva Abels and Liron Pantanowitz. Computational pathology: Challenges and promises for tissue analysis. *Computers in Biology and Medicine*, 113:103622, 2019.

[2] Mehdi Afshari, Saba Yasir, Gary L Keeney, Rafael E Jimenez, Joaquin J Garcia, and Hamid R Tizhoosh. Single patch super-resolution of histopathology whole slide images: a comparative study. *Journal of Medical Imaging*, 10(1):017501–017501, 2023.

[3] Saghir Alfasly, Abubakr Shafique, Peyman Nejat, Jibran Khan, Areej Alsaafin, Ghazal Alabtah, and H.R. Tizhoosh. Rotation-agnostic image representation learning for digital pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11683–11693, June 2024.

[4] Marco Aversa, Gabriel Nobis, Miriam Hägele, Kai Standvoss, Mihaela Chirica, Roderick Murray-Smith, Ahmed M Alaa, Lukas Ruff, Daniela Ivanova, Wojciech Samek, et al. Diffinfinite: Large mask-image synthesis via parallel random patch diffusion in histopathology. *Advances in Neural Information Processing Systems*, 36:78126–78141, 2023.

[5] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Maaike Hermsen, Quirine F Manson, Markus Balkenhol, Oscar Geessink, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 2017. CAMELYON16 dataset.

[6] Ali Borji. Pros and cons of GAN evaluation measures: New developments. *Computer Vision and Image Understanding*, 215:103329, 2022.

[7] Wouter Bulten, Kaisa Kartasalo, Peter H Chen, Patrik Ström, Hans Pinckaers, Pranav Nagpal, Yale Cai, Daniel F Steiner, Liping Zhang, Baris Gecer, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature Medicine*, 28(1):154–163, 2022. PANDA challenge dataset.

[8] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019.

[9] 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.

[10] Sarah Cechnicka, James Ball, Matthew Baugh, Hadrien Reynaud, Naomi Simmonds, Andrew PT Smith, Catherine Horsfield, Candice Roufosse, and Bernhard Kainz. Urcdm: Ultra-resolution image synthesis in histopathology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 535–545. Springer, 2024.

[11] R. J. Chen, T. Ding, M. Y. Lu, D. F. Williamson, G. Jaume, A. H. Song, B. Chen, A. Zhang, D. Shao, M. Shaban, M. Williams, L. Oldenburg, L. L. Weishaupt, J. J. Wang, A. Vaidya, L. P. Le, G. Gerber, S. Sahai, W. Williams, and F. Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024.

[12] Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. *Machine learning with applications*, 7:100198, 2022.

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

[14] David Jacob Drexlin, Jonas Dippel, Julius Hense, Niklas Prenißl, Grégoire Montavon, Frederick Klauschen, and Klaus-Robert Müller. Medi: Metadata-guided diffusion models for mitigating biases in tumor classification. *arXiv preprint arXiv:2506.17140*, 2025.

[15] Joann G Elmore, Gary M Longton, Patricia A Carney, Berta M Geller, Tracy Onega, Anna NA Tosteson, Heidi D Nelson, Margaret S Pepe, Kimberly H Allison, Stuart J Schnitt, et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *Jama*, 313(11):1122–1132, 2015.

[16] Michael Gadermayr, Laxmi Gupta, Ventzeslav Appel, Peter Boor, Barbara Maria Klinkhammer, and Dorit Merhof. Generative adversarial networks for facilitating stain-independent supervised and unsupervised segmentation: a study on kidney histology. *IEEE transactions on medical imaging*, 38(10):2293–2302, 2019.

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, volume 27, 2014.

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in neural information processing systems*, volume 33, pages 6840–6851, 2020.

[19] Marco Jiralerspong, Joey Bose, Ian Gemp, Chongli Qin, Yoram Bachrach, and Gauthier Gidel. Feature likelihood divergence: Evaluating the generalization of generative models using samples. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 33095–33119. Curran Associates, Inc., 2023.

[20] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

[21] Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Roel Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6):305–311, 2020.

[22] Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, and Sérgio Pereira. Benchmarking self-supervised learning on diverse pathology datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3344–3354, June 2023.

[23] Jakob N Kather, Leendert R Heij, Heike I Grabsch, and et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature Cancer*, 3(3):303–313, 2022.

[24] Daisuke Komura and Shumpei Ishikawa. Machine learning methods for histopathological image analysis. *Computational and structural biotechnology journal*, 16:34–42, 2018.

[25] Nicholas Konz, Yuwen Chen, Haoyu Dong, and Maciej A. Mazurowski. Anatomically-controllable medical image generation with segmentation-guided diffusion models. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2024.

[26] Ira Ktena, Olivia Wiles, Isabela Albuquerque, Sylvestre-Alvise Rebuffi, Ryutaro Tanno, Abhijit Guha Roy, Shekoofeh Azizi, Danielle Belgrave, Pushmeet Kohli, Taylan Cemgil, et al. Generative models improve fairness of medical classifiers under distribution shifts. *Nature Medicine*, 30(4):1166–1173, 2024.

[27] David B Larson, David C Magnus, Matthew P Lungren, Nigam H Shah, and Curtis P Langlotz. Ethics of using and sharing clinical imaging data for artificial intelligence: a proposed framework. *Radiology*, 295(3):675–682, 2020.

[28] Adrian B Levine, Jason Peng, David Farnell, Mitchell Nursey, Yiping Wang, Julia R Naso, Hezhen Ren, Hossein Farahani, Colin Chen, Derek Chiu, et al. Synthesis of diagnostic quality cancer pathology images by generative adversarial networks. *The Journal of pathology*, 252(2):178–188, 2020.

[29] David N Louis, Gary K Gerber, Jeffrey M Baron, Linda Bry, Anand S Dighe, Gad Getz, John P Higgins, Felice C Kuo, Stuart F Nelson, George P Nielsen, et al. Computational pathology: an emerging definition. *Archives of pathology & laboratory medicine*, 138(9):1133–1138, 2014.

[30] Puria Azadi Moghadam, Sanne Van Dalen, Karina C Martin, Jochen Lennerz, Stephen Yip, Hossein Farahani, and Ali Bashashati. A morphology focused diffusion probabilistic model for synthesis of histopathology images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2000–2009, 2023.

[31] 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.

[32] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

[33] Pedro Osorio, Guillermo Jimenez-Perez, Javier Montalt-Tordera, Jens Hooge, Guillem Duran-Ballester, Shivam Singh, Moritz Radbruch, Ute Bach, Sabrina Schroeder, Krystyna Siudak, et al. Latent diffusion models with image-derived annotations for enhanced ai-assisted cancer diagnosis in histopathology. *Diagnostics*, 14(13):1442, 2024.

 https://doi.org/10.52202/085713-4185

[34] Adalberto Claudio Quiros, Roderick Murray-Smith, and Ke Yuan. Pathologygan: Learning deep representations of cancer tissue. In Tal Arbel, Ismail Ben Ayed, Marleen de Bruijne, Maxime Descoteaux, Herve Lombaert, and Christopher Pal, editors, *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, volume 121 of *Proceedings of Machine Learning Research*, pages 669–695. PMLR, 06–08 Jul 2020.

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[36] 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.

[37] 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.

[38] Charlie Saillard, Rodolphe Jenatton, Felipe Llinares-López, Zelda Mariet, David Cahané, Eric Durand, and Jean-Philippe Vert. H-optimus-0, 2024.

[39] Aman Shrivastava and P Thomas Fletcher. Nasdm: Nuclei-aware semantic histopathology image generation using diffusion models. In *international conference on medical image computing and computer-assisted intervention*, pages 786–796. Springer, 2023.

[40] Chetan L Srinidhi, Ozan Ciga, and Anne L Martel. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, 67:101813, 2021.

[41] David Tellez, Geert Litjens, Péter Bándi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen van der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical image analysis*, 58:101544, 2019.

[42] Hamid R Tizhoosh, Phedias Diamandis, Clinton JV Campbell, Amir Safarpoor, Shivam Kalra, Danial Maleki, Abtin Riasatian, and Morteza Babaie. Searching images for consensus: can ai remove observer variability in pathology? *The American journal of pathology*, 191(10):1702–1708, 2021.

[43] Martin Tschuchnig and et al. Generative adversarial networks in digital pathology: a survey on trends and future potential. *Computers in Biology and Medicine*, 142:105210, 2022.

[44] Jason W Wei, Laura J Tafe, Yevgeniy A Linnik, Louis J Vaickus, Naofumi Tomita, and Saeed Hassanpour. Generative image translation for data augmentation in colorectal histopathology images. *Proceedings of machine learning research*, 116:10, 2019.

[45] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna RM Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013. TCGA dataset.

[46] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024.

[47] 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. *arXiv preprint arXiv:2412.06011*, 2024.

[48] Srikar Yellapragada, Alexandros Graikos, Prateek Prasanna, Tahsin Kurc, Joel Saltz, and Dimitris Samaras. Pathldm: Text conditioned latent diffusion model for histopathology. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5182–5191, 2024.

[49] 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 (ICCV)*, pages 3836–3847, October 2023.

[50] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine learning for healthcare conference*, pages 2–25. PMLR, 2022.

[51] Han Zhao, Wuyang Wang, and Zhangyang Wang. Data augmentation using learned transformations for one-shot medical image segmentation. In *CVPR*, 2020.

Semantic and Visual Crop-Guided Diffusion Models for Heterogeneous Tissue Synthesis in Histopathology

- Supplementary File -

Contents

1 Additional Methodology Details **2**
1.1 Dual Conditioning 2
1.1.1 Visual Crop Encoding and Semantic Map Processing 2
1.1.2 Prompt Integration 4
1.2 Self-Supervised TCGA Clustering Algorithm 5
1.2.1 Clustering Algorithm Details 5
1.2.2 Visualized Cluster Samples 6
1.2.3 Tissue Classifier Training 6
1.3 Heterogeneous Patch Sampling Strategy 8
1.4 Diffusion Model Training Details 9

2 Additional Experimental Results and Analysis **10**
2.1 Privacy Preservation Assessment 10
2.2 Visual Crop Size Analysis 11
2.3 Synth vs. Real: Quantitative Results 11

3 Detailed Expert Evaluation **12**
3.1 Evaluation Protocol and Methodology 12
3.2 Evaluation Criteria and Clinical Relevance 13
3.3 Quantitative Results and Statistical Analysis 14
3.4 Dataset-Specific Observations 15
3.5 Clinical Implications and Expert Commentary 15

4 Generation Examples **18**

5 Computational Resources **18**

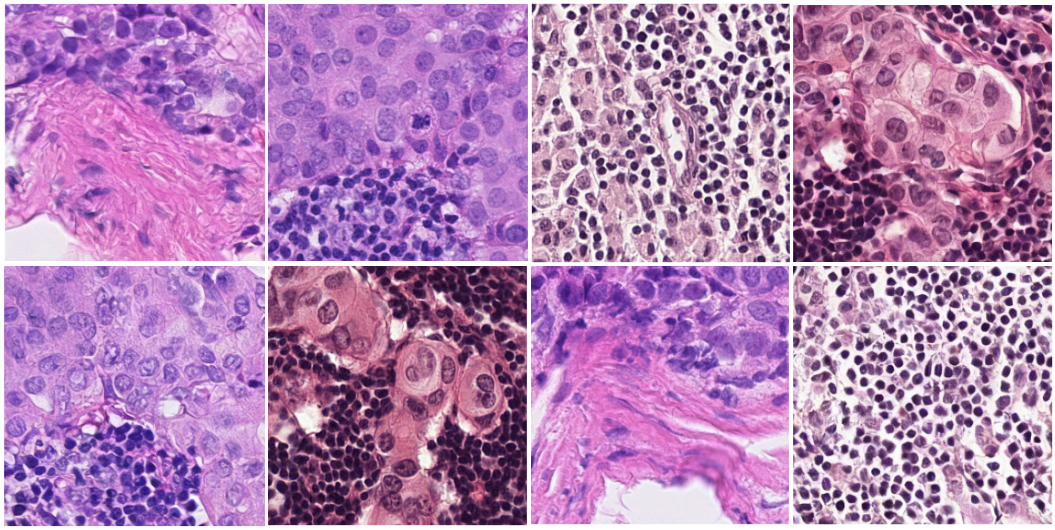
6 Current Limitations **18**

1 Additional Methodology Details

1.1 Dual Conditioning

Our dual-conditioning approach represents a fundamental advancement in histopathology synthesis by combining the spatial precision of semantic maps with the morphological authenticity of raw visual crops. This design philosophy addresses the critical limitation of existing approaches that rely on either abstract embeddings or spatial information alone, both of which fail to preserve the fine-grained morphological details essential for clinical authenticity in synthetic histopathology images [39, 25, 48].

https://doi.org/10.52202/085713-4185



Supplementary Figure S1: **Real vs. Synthetic Lymph Node Tissue Challenge**. These eight lymph node histopathology patches represent a mixture of real diagnostic images and synthetic samples generated by our HeteroTissue-Diffuse framework. The remarkable preservation of cellular morphology, nuclear details, and tissue architecture in our synthetic images makes visual distinction challenging, even for trained pathologists. Can you identify which patches are real and which are synthetic? (Answer key provided at the end of supplementary materials.)

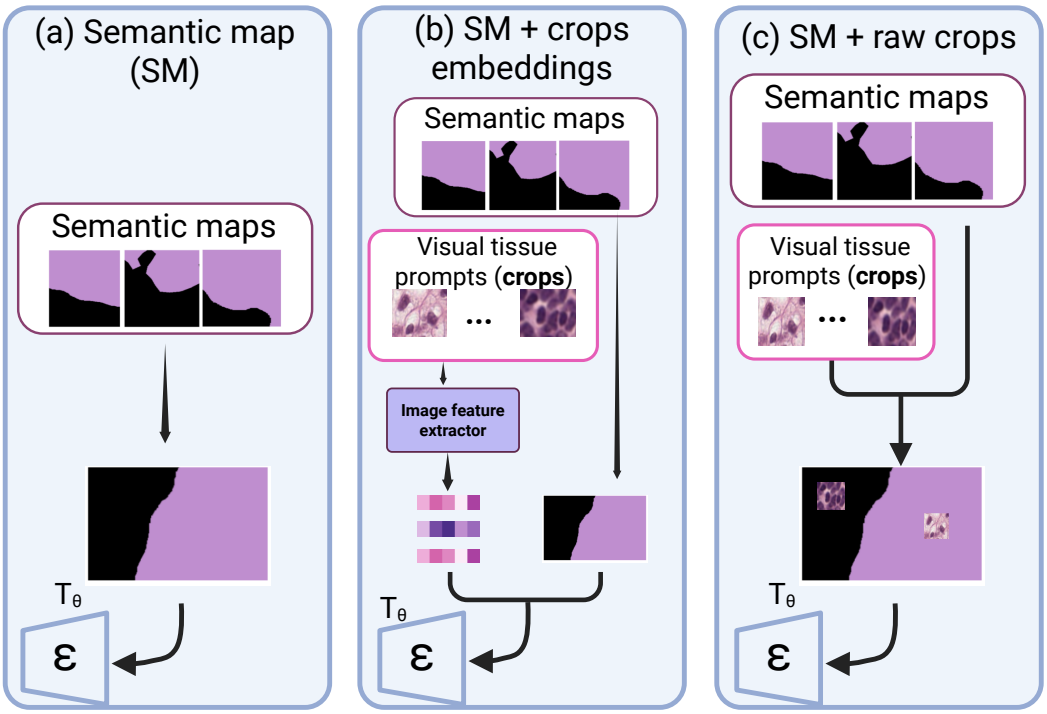
1.1.1 Visual Crop Encoding and Semantic Map Processing

Figure S2 shows the three different conditioning approaches and demonstrates how our method differs from conventional techniques by using semantic maps along with visual raw crops to preserve structural, morphological, and staining fine details during generation. Additionally, Figure S3 illustrates the comprehensive visual crop and semantic map encoding process, detailing our systematic condition preparation and integration methodology.

The semantic map processing component creates precise spatial control by generating binary one-hot masks for each tissue class present in the target image. For a given histopathology patch with K tissue classes, we construct K binary masks $M_1, M_2, ..., M_K \in 0, 1^{H \times W}$, where each mask M_i indicates the spatial locations where tissue class i should be synthesized. This approach ensures that the diffusion model receives explicit spatial guidance about where each tissue type should appear, enabling precise control over tissue composition and boundary formation [49, 36]. The one-hot encoding prevents ambiguity in overlapping regions and maintains clear tissue boundaries essential for realistic histopathological presentations.

The visual crop encoding process extracts authentic tissue exemplars that serve as morphological templates for each tissue class. For each active tissue class i (where $\sum M_i > 0$), we extract a square crop p_i of variable size $d \times d$ pixels, where d is randomly sampled from the range $[50, 200]$ to ensure diversity in scale and detail preservation. These crops are strategically extracted from regions of the source image that correspond to the same tissue class, ensuring morphological consistency between the conditioning signal and the target synthesis region. The extraction process employs spatial diversity constraints to avoid repetitive sampling from identical locations, promoting morphological variety within each tissue class.

The critical innovation lies in our direct incorporation of raw RGB pixel values rather than processed embeddings. Unlike embedding-based approaches that compress visual information through feature extractors, potentially losing critical diagnostic details such as nuclear chromatin patterns, cytoplasmic texture, and staining variations [12, 50], our method preserves the full spectrum of visual information present in authentic tissue samples. This preservation is achieved by creating tissue-specific visual prompt tensors $C_i \in \mathbb{R}^{H \times W \times 3}$ for each class, where each tensor contains the raw crop p_i positioned within the spatial bounds defined by the corresponding semantic mask M_i.



Supplementary Figure S2: **Comparison of Three Conditioning Approaches in HeteroTissue-Diffuse.** Our framework explores three distinct conditioning mechanisms for histopathology synthesis: (a) semantic map (SM) only conditioning using spatial masks alone, (b) semantic maps combined with crop embeddings where visual tissue prompts are processed through a feature extractor before conditioning, and (c) our proposed approach combining semantic maps with raw visual crops directly. The direct incorporation of raw tissue crops (c) preserves critical morphological details that are lost in embedding-based representations, enabling superior synthesis of heterogeneous tissue structures.

The concatenation of semantic and visual information creates a comprehensive conditioning tensor $c = \text{concat}(M_1, ..., M_K, C_1, ..., C_K)$ with dimensions $H \times W \times (K + 3K) = H \times W \times 4K$, where the first K channels encode spatial information and the remaining $3K$ channels contain RGB visual prompts. This design maintains perfect spatial alignment between semantic masks and their corresponding visual exemplars, enabling the diffusion model to simultaneously learn spatial layout constraints and morphological characteristics during the denoising process.

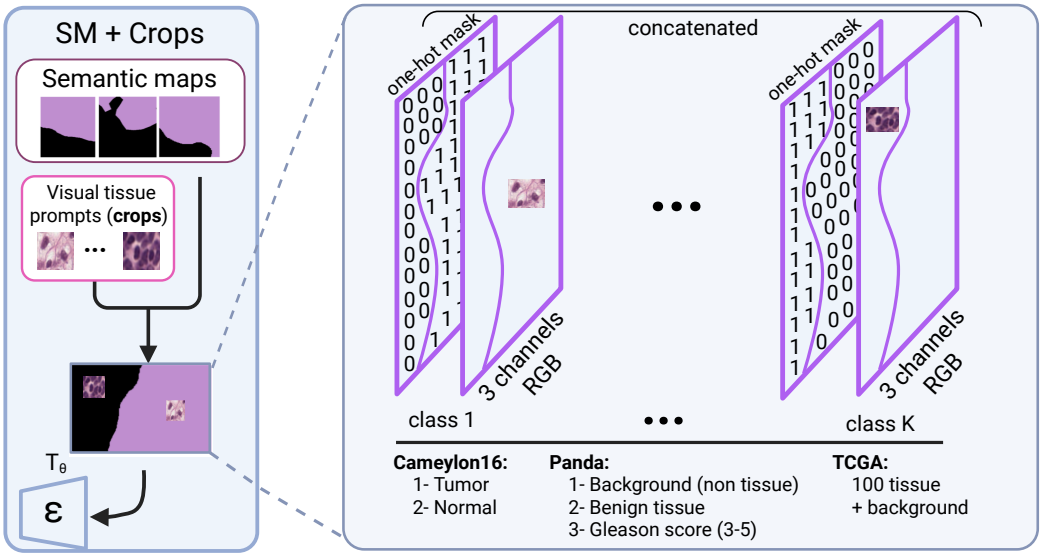
1.1.2 Prompt Integration

The integration of our dual-conditioning signal into the UNet denoising architecture follows the established paradigm of latent diffusion models for image-based conditioning, similar to the approach described in Rombach et al. [36]. Our method employs direct concatenation of the conditioning tensor with the latent representation at the UNet input, rather than cross-attention mechanisms commonly used in text-to-image synthesis [37]. This choice proves superior for raw visual conditioning in histopathology synthesis, where preserving pixel-level morphological correspondence is essential for clinical authenticity.

The conditioning tensor $c \in \mathbb{R}^{h \times w \times 4K}$ is first encoded to the latent space using the same encoder $\mathcal{E}$ employed for the target images, producing the latent condition $c_{latent} = \mathcal{E}(c) \in \mathbb{R}^{h \times w \times c_{cond}}$. During the denoising process at timestep t, we concatenate this latent conditioning signal with the noisy latent $z_t \in \mathbb{R}^{h \times w \times c}$ along the channel dimension to create an augmented input $z_{aug} = \text{concat}(z_t, c_{latent}) \in \mathbb{R}^{h \times w \times (c + c_{cond})}$ for the UNet ϵ_θ.

This concatenation occurs at the input layer of the UNet, ensuring that conditioning information is available throughout all levels of the hierarchical feature extraction and synthesis process. The UNet architecture is modified to accept the additional conditioning channels through an expanded

https://doi.org/10.52202/085713-4185



Supplementary Figure S3: **Detailed Architecture of Dual-Conditioning Signal Construction**. Our conditioning mechanism combines semantic spatial information with visual tissue exemplars through a systematic construction process. For each tissue class, we create a binary one-hot mask indicating spatial locations, then concatenate it with a corresponding 3-channel RGB tensor containing visual crop prompts strategically placed within the regions of interest. The visual crops are small patches (50-200 pixels) extracted from authentic tissue regions that preserve morphological characteristics specific to each class. This dual-conditioning approach is dataset-agnostic and scales flexibly across different annotation granularities: Camelyon16 uses 2 classes (tumor, normal), PANDA employs 3 classes (background, benign tissue, Gleason scores 3-5), while our TCGA extension discovers 100 distinct tissue phenotypes plus background through self-supervised clustering. The concatenated conditioning tensor guides the diffusion process with both spatial precision from semantic maps and morphological authenticity from raw visual crops, enabling controlled synthesis of heterogeneous tissue compositions while maintaining clinically relevant features across diverse cancer types and tissue architectures.

first convolutional layer, while all subsequent layers remain unchanged. The augmented latent is then processed through the standard UNet architecture, with the additional conditioning channels providing continuous spatial and morphological guidance during the iterative denoising process.

Unlike abstract embeddings that benefit from attention-based fusion, raw tissue crops contain explicit spatial and morphological information that is more effectively preserved through direct concatenation in the latent space [18, 13]. This approach maintains the direct correspondence between semantic regions and their associated visual exemplars without introducing attention weights that could dilute critical morphological features. Cross-attention mechanisms, while effective for text-to-image synthesis [37], introduce computational overhead and can lead to inconsistent feature blending when dealing with high-resolution tissue crops containing fine cellular details.

Our latent concatenation strategy ensures that each spatial location in the visual crops directly influences the corresponding location in the synthesis process through the shared latent representation. This direct spatial correspondence is crucial for preserving authentic tissue characteristics such as nuclear morphology, cytoplasmic patterns, and staining variations that are essential for clinically relevant synthesis. The encoder $\mathcal{E}$ preserves the spatial structure of the conditioning signal while compressing it to the same latent dimension as the target synthesis, enabling efficient processing while maintaining morphological fidelity.

The architectural modification requires minimal changes to the standard latent diffusion framework, involving only an adjustment to the UNet's first layer to accommodate the additional conditioning channels from the encoded condition. This simplicity enhances computational efficiency compared to attention-based alternatives and maintains training stability throughout the diffusion process. The direct latent concatenation approach achieves superior performance while preserving the elegant sim-

plicity of the latent diffusion paradigm, ensuring compatibility with existing optimization strategies and training procedures.

1.2 Self-Supervised TCGA Clustering Algorithm

Algorithm S1 provides the complete self-supervised clustering approach for unannotated TCGA whole-slide images. This comprehensive framework addresses the critical challenge of scaling histopathology synthesis to massive unannotated datasets by automatically discovering tissue phenotypes without manual intervention. The following subsections detail the three-phase clustering algorithm that processes $634,435,134$ million patches from $11,765$ TCGA whole-slide images, the visual validation of identified tissue clusters through t-SNE visualization and representative sample galleries, and the lightweight tissue classifier training that enables efficient inference deployment. This scalable approach democratizes access to diverse synthetic histopathology data across 33 cancer types while preserving morphological authenticity essential for clinical applications.

1.2.1 Clustering Algorithm Details

The self-supervised clustering algorithm operates in three distinct phases designed to balance computational efficiency with comprehensive tissue phenotype discovery across the massive TCGA dataset.

Phase 1 implements strategic feature collection where each WSI undergoes systematic patch extraction at 224×224 pixel resolution with non-overlapping stride of 224 pixels, followed by tissue detection to exclude background regions. To manage the computational burden of processing $634,435,134$ million patches while ensuring representative sampling, we extract features for a maximum of $N = 1000$ patches per WSI using the UNI foundation model [11], prioritizing diversity through spatial distribution constraints that prevent oversampling from identical tissue regions. This sampling strategy ensures adequate representation of rare tissue phenotypes while maintaining tractable computational requirements for the clustering phase.

Algorithm S1 Scalable TCGA Tissue Clustering

Require: WSI collection $\mathcal{W}$, foundation model f_ϕ
Require: Target clusters $K = 100$, samples per WSI $N = 1000$
Ensure: Cluster assignments for all WSI patches

function ClusterTCGATissues($\mathcal{W}, f_\phi, K, N$)
  Phase 1: Feature Collection
  $\mathcal{F}_{\text{train}} \leftarrow \emptyset$
  for each WSI $w \in \mathcal{W}$ **do**
    $P_w \leftarrow$ ExtractPatches(w, stride=224)
    $P_{\text{sample}} \leftarrow$ RandomSample(P_w, $\min(N, |P_w|)$)
    $F_w \leftarrow f_\phi(P_{\text{sample}})$
    $\mathcal{F}_{\text{train}} \leftarrow \mathcal{F}_{\text{train}} \cup F_w$
  end for
  Phase 2: Clustering
  $\mathcal{C} \leftarrow$ KMeans($\mathcal{F}_{\text{train}}$, K, niter=100)
  Phase 3: Full Assignment
  for each WSI $w \in \mathcal{W}$ **do**
    $S_w \leftarrow$ empty segmentation map
    for batch B in ChunkPatches(P_w, size=1000) **do**
      $F_B \leftarrow f_\phi(B)$
      $L_B \leftarrow$ NearestCentroid($F_B, \mathcal{C}$)
      UpdateSegmentationMap(S_w, B, L_B)
    end for
    Save(S_w) ▷ Multi-scale cluster maps
  end for
  return cluster assignments
end function

https://doi.org/10.52202/085713-4185

Phase 2 performs k-means clustering on the collected feature vectors with k=100 clusters and 100 iterations to ensure convergence. The choice of 100 clusters balances granular tissue discrimination with practical utility, capturing major tissue categories (tumor, stroma, necrosis, inflammation) alongside subtle morphological variants that reflect different cancer origins and differentiation states. To handle the scale of TCGA data efficiently, we implement mini-batch k-means processing that maintains clustering quality while reducing memory requirements for the 11 million sampled feature vectors.

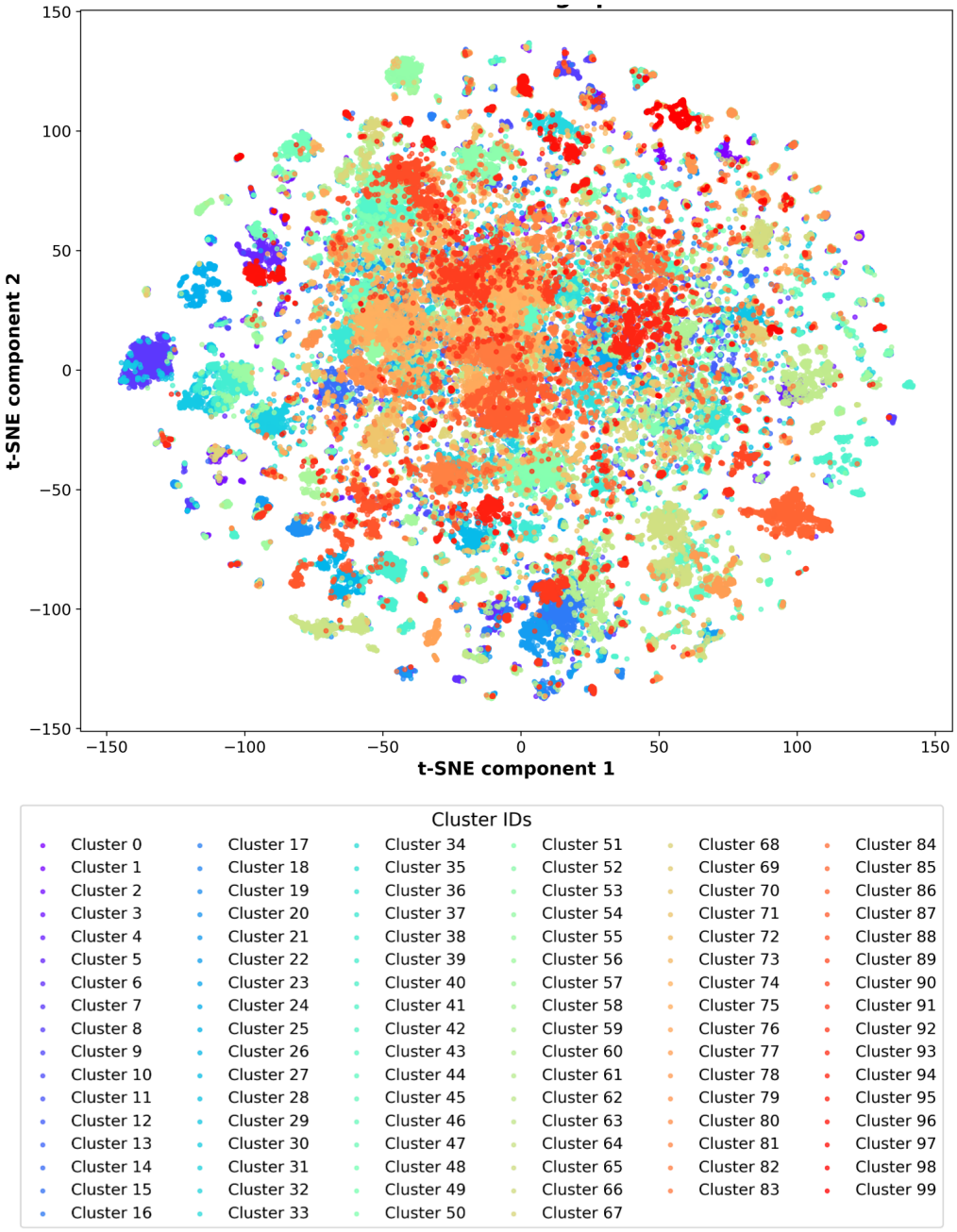
Phase 3 assigns cluster labels to all WSI patches through efficient batch processing that avoids recomputing foundation model embeddings for previously processed patches. Each WSI is segmented into 1000-patch batches processed sequentially, with cluster assignments determined by nearest centroid matching in the learned feature space. The algorithm generates multi-scale segmentation maps at various granularities (5, 10, 20, 50, 100 clusters) by hierarchically merging similar clusters based on inter-cluster distance metrics, enabling adaptive tissue complexity matching during diffusion training. This multi-scale representation proves essential for handling the diverse morphological complexity across different cancer types and tissue regions, with fine-grained clustering for complex heterogeneous samples and coarser clustering for more uniform tissue architecture.

1.2.2 Visualized Cluster Samples

To demonstrate the quality and diversity of our self-supervised tissue clustering approach on TCGA data, we present representative samples from each of the 100 identified tissue phenotypes. Figure S4 provides a t-SNE visualization of 99,792 randomly sampled TCGA patches, illustrating the clear separation and distinct clustering achieved by our foundation model-based approach across the high-dimensional feature space. Figures S10, S11,S12,S13,S14,S15,S16,S17,S18,S19 showcase the morphological coherence within clusters while highlighting the rich phenotypic diversity captured across different cancer types and tissue architectures. Each figure displays 10 clusters, with 9 representative patches per cluster arranged in rows to illustrate intra-cluster consistency and inter-cluster distinctiveness. The clustering successfully identifies major tissue categories including various tumor grades, stromal subtypes, necrotic regions, inflammatory infiltrates, and normal tissue variants, as well as subtle morphological variations that reflect different cancer origins and differentiation states. This comprehensive tissue phenotype discovery enables our dual-conditioning framework to generate synthetic samples with unprecedented diversity while maintaining authentic morphological characteristics essential for training robust diagnostic algorithms across the full spectrum of cancer pathology.

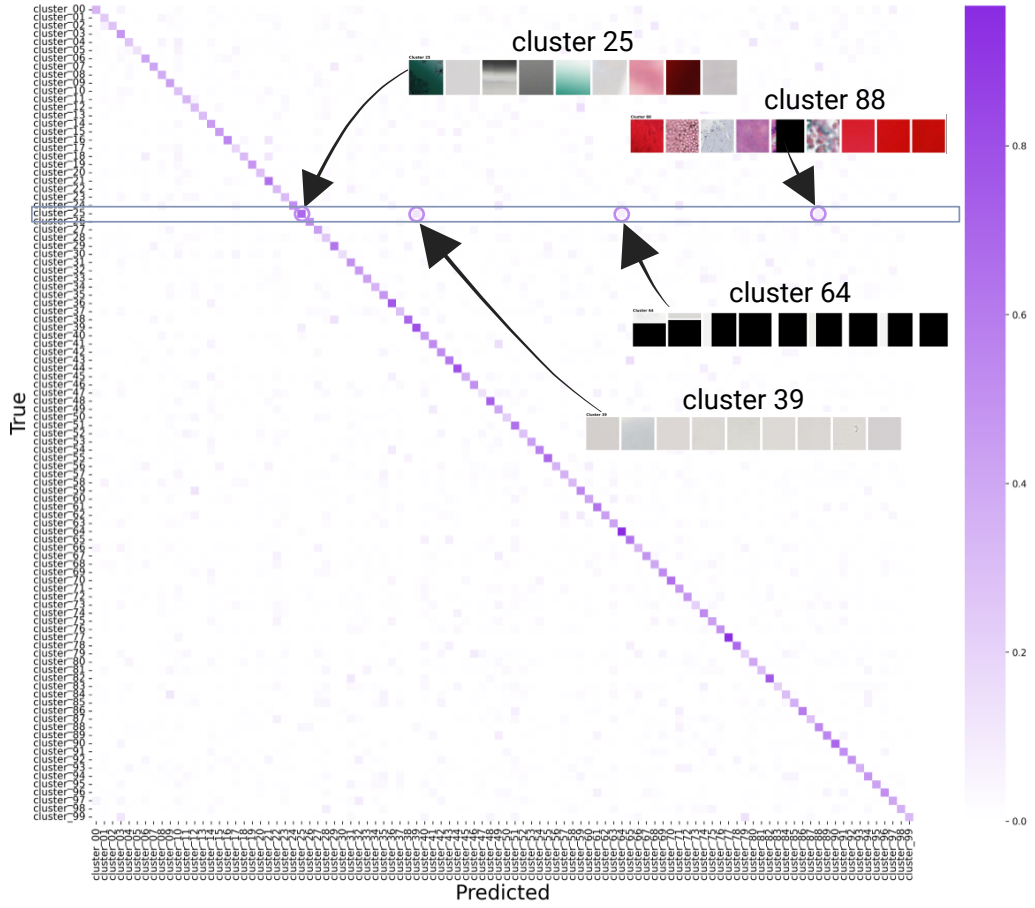
1.2.3 Tissue Classifier Training

To enable efficient inference during synthetic data generation, we trained a lightweight tissue classification model following the self-supervised clustering of TCGA patches. The classifier architecture employs a Vision Transformer Small (ViT-S) with patch size 16 and input resolution 224×224, initialized with ImageNet pretrained weights and adapted for 100-class tissue type classification plus background. The model was trained on 514,029 balanced patches extracted from $11,765$ diagnostic TCGA whole-slide images using stratified sampling to ensure equal representation across all identified tissue clusters. Training employed AdamW optimization with initial learning rate 1×10^{-3}, weight decay 1×10^{-4}, and cosine annealing scheduler over 50 epochs with batch size 512 distributed across multiple GPUs. Data augmentation included random resized crops, horizontal and vertical flips, and color jitter (brightness, contrast, saturation, hue ± 0.1) to improve generalization across different scanners and staining variations. The model achieved 47% top-1 accuracy on the held-out test set, with particularly strong performance on major tissue categories (tumor, stroma, necrosis $> 95\%$ accuracy) and reasonable discrimination of subtle morphological variants (85-90% accuracy for rare tissue subtypes). Cross-entropy loss with label smoothing ($\epsilon = 0.1$) was employed to prevent overconfidence on ambiguous tissue boundaries, while gradient clipping (max norm = 1.0) ensured stable training convergence. The final model requires only 22M parameters and achieves inference speeds of $\sim$500 patches/second on a single GPU, representing an 85% computational reduction compared to foundation model embedding extraction while maintaining classification fidelity sufficient for conditioning the diffusion process. This efficient classifier enables scalable deployment of our framework in resource-constrained environments while preserving the quality of tissue-specific visual conditioning essential for realistic histopathology synthesis.



Supplementary Figure S4: t-SNE visualization of 99,792 randomly sampled TCGA patches colored by cluster assignment using UNI foundation model features [11]. The well-defined separation validates this approach of 100 morphologically coherent tissue phenotypes.

https://doi.org/10.52202/085713-4185



Supplementary Figure S5: Confusion matrix of the lightweight tissue classifier evaluated on 20,000 TCGA test patches across 100 tissue clusters. Circled regions highlight artifact clusters (88, 64, 39, 25) representing white backgrounds, corrupted patches, and scanning markers that are treated as equivalent during evaluation, with representative patch samples shown using UNI embeddings to demonstrate morphological coherence within these non-diagnostic categories.

Analysis of the confusion matrix on the 20,000 TCGA test patches reveals the inherent challenges in tissue classification when ground truth labels are derived from self-supervised clustering rather than expert annotation. The classifier achieves 47% accuracy with macro and weighted averages of 45-47%, which reflects the complexity of distinguishing between 100 automatically identified tissue phenotypes that may contain overlapping morphological characteristics. Figure S5 demonstrates that certain clusters (88, 64, 39, 25) represent artifact categories including white background patches, black/corrupted regions, and images with scanning markers that lack diagnostic value. These clusters are intentionally treated as equivalent during evaluation, meaning any prediction among these four classes is considered correct regardless of the specific assignment, as they all represent non-tissue regions that should be excluded from synthetic generation.

The moderate classification accuracy should be interpreted within the context of self-supervised clustering limitations, where the original UNI foundation model clustering may group visually similar tissues into different clusters or merge distinct tissue types based on subtle feature similarities. Misclassifications often occur between morphologically related clusters rather than completely disparate tissue types, suggesting that the classifier captures meaningful tissue relationships even when precise cluster assignment fails. The UNI embedding-based visualization of representative patches from artifact clusters confirms that the framework appropriately handles non-diagnostic content while focusing computational resources on clinically relevant tissue synthesis. For the purposes of diffusion conditioning, this level of classification accuracy proves sufficient, as minor variations in tissue crop selection within morphologically similar clusters do not significantly impact

the quality of synthetic histopathology generation, and the dual-conditioning approach provides additional robustness through semantic map guidance that complements the visual crop information.

1.3 Heterogeneous Patch Sampling Strategy

For both annotated and self-supervised settings, we employ a specialized sampling strategy to ensure tissue heterogeneity. Algorithm S2 provides the enhanced sampling approach.

Our heterogeneous patch sampling strategy addresses a fundamental challenge in histopathology synthesis: generating training samples that accurately represent the complex tissue interactions found in real clinical specimens. Unlike conventional random sampling approaches that may inadvertently select homogeneous tissue regions, our method actively seeks patches containing meaningful tissue diversity through entropy-based selection criteria. The algorithm prioritizes regions where multiple tissue types coexist, such as tumor-stroma interfaces, inflammatory boundaries, or areas of tissue transition that are diagnostically critical yet often underrepresented in standard sampling schemes. This targeted approach ensures that our diffusion model learns to synthesize realistic tissue heterogeneity rather than generating artificial boundaries between distinct tissue types, a common limitation in semantic-only conditioning approaches [39].

Require: Dataset with patches and segmentation maps (real or pseudo-labeled)
Require: Minimum region entropy threshold τ_{entropy}
Require: Tissue coverage threshold τ_{coverage}
Require: Tissue ratio bounds $[r_{\min}, r_{\max}] = [0.2, 0.8]$
function SAMPLEPATCH($\mathcal{D}$)
  Initialize empty candidate list C
  for $i = 1$ to 100 **do** ▷ Try 100 random patches
    $(x, M) \leftarrow \text{RandomPatch}(\mathcal{D})$
    $r_{\text{tissue}} \leftarrow \text{ComputeTissueRatio}(M)$
    if $r_{\min} \leq r_{\text{tissue}} \leq r_{\max}$ **then**
      $H \leftarrow \text{ComputeEntropyMap}(M)$
      if $\text{mean}(H) > \tau_{\text{entropy}}$ **then**
        Add $(x, M, \text{mean}(H))$ to C
      end if
    end if
  end for
  if $|C| > 0$ **then**
    Sort C by entropy (descending)
    return top patch from C
  else
    Retry with relaxed constraints
  end if
end function

Algorithm S2: Advanced Heterogeneous Patch Sampling

The entropy-driven selection mechanism quantifies tissue complexity by computing spatial entropy across segmentation maps, with higher entropy values indicating greater morphological diversity within a given region. For each candidate patch, we calculate the tissue ratio to ensure balanced representation between $r_{min} = 0.2$ and $r_{max} = 0.8$, preventing the selection of predominantly background or overwhelmingly complex regions that could hinder training convergence. The entropy threshold $\tau_{entropy}$ serves as a quality gate, filtering out patches with insufficient tissue diversity while the coverage threshold $\tau_{coverage}$ ensures adequate representation of each tissue class present in the selected region. This dual-threshold approach balances the competing demands of tissue diversity and training stability, enabling the model to learn from challenging heterogeneous samples without being overwhelmed by excessive complexity during early training phases.

The iterative candidate selection process attempts up to 100 random patches before selecting the sample with highest entropy among those meeting our heterogeneity criteria, ensuring both efficiency and quality in the sampling process. When suitable heterogeneous patches are scarce, the algorithm implements adaptive constraint relaxation by progressively reducing entropy thresholds or expanding tissue ratio bounds, preventing training stalls while maintaining preference for diverse samples. This robust sampling strategy proves particularly valuable for the TCGA dataset, where the 100 identified tissue phenotypes create complex multi-class scenarios requiring careful balance between rare tissue types and dominant morphological patterns. The resulting training samples enable our dual-conditioning framework to generate synthetic histopathology images with authentic tissue

https://doi.org/10.52202/085713-4185

transitions and morphological complexity that closely mirror real clinical specimens across diverse cancer types and tissue architectures.

1.4 Diffusion Model Training Details

Our HeteroTissue-Diffuse framework employs a latent diffusion architecture trained across three distinct datasets with dataset-specific conditioning configurations to accommodate varying tissue complexity and annotation granularity. The training process utilizes a VQ-GAN autoencoder with $8,192$ codebook entries operating at 4 downsampling ($f = 4$) to compress 256×256 pixel histopathology images into 64×64 latent representations, enabling efficient synthesis while preserving morphological details essential for clinical authenticity [36]. All models employ identical base learning rates of 1×10^{-6} with linear noise scheduling from $\beta_1 = 0.0015$ to $\beta_2 = 0.0205$ over $T = 1000$ diffusion timesteps, using L1 loss for stable training convergence and superior preservation of fine-grained tissue structures compared to $L2$ alternatives. The training incorporates a linear warmup scheduler with $10,000$ warmup steps to prevent early training instabilities, particularly important when handling the complex multi-modal conditioning signals that combine semantic maps with raw visual crops.

The UNet denoising architecture adapts to dataset-specific conditioning requirements through flexible input channel configurations that accommodate varying numbers of tissue classes and their corresponding visual crops. For Camelyon16's binary classification (tumor vs. normal), the model processes 8 conditioning channels (2 semantic + 6 visual crop channels) combined with 3 latent image channels, resulting in 6 total input channels to the UNet after latent space encoding. PANDA's three-class structure (background, benign tissue, Gleason grades) requires 9 conditioning channels (3 semantic + 9 visual), while the TCGA extension scales to 404 conditioning channels (100 semantic + 304 visual crop channels) to handle the full spectrum of identified tissue phenotypes. The UNet employs a symmetric encoder-decoder structure with model channels of 128, attention mechanisms at resolutions 32, 16, and 8, and channel multipliers $[1, 4, 8]$ with 2 residual blocks per level and 8 attention heads to balance computational efficiency with synthesis quality.

Training data scales vary significantly across datasets, reflecting both availability and complexity requirements: Camelyon16 utilizes $28,291$ curated patches enabling focused training on lymph node metastasis detection, PANDA leverages $493,836$ patches for comprehensive prostate cancer grade synthesis, and TCGA employs on-the-fly sampling from $11,765$ whole-slide images to ensure continuous exposure to diverse tissue phenotypes without memory constraints. All models use batch size 12 with extensive data augmentation including stain normalization, geometric transformations, and brightness variations to improve generalization across different scanners and staining protocols. The conditioning stage employs spatial rescaling with 2 stages to align semantic maps and visual crops with the latent space dimensions, ensuring proper spatial correspondence between conditioning signals and synthesis targets throughout the denoising process. Training convergence typically requires $200,000$-$300,000$ iterations depending on dataset complexity, with image logging every $5,000$ steps to monitor synthesis quality and prevent mode collapse or artifact generation that could compromise clinical utility.

2 Additional Experimental Results and Analysis

2.1 Privacy Preservation Assessment

To address the critical concern of patient privacy protection in synthetic data generation, we conducted a comprehensive quantitative privacy evaluation using the Feature Likelihood Divergence (FLD) framework [19]. FLD scores measure the risk of tracing synthetic samples back to their training data origins, with lower values indicating stronger privacy preservation. Our evaluation demonstrates robust privacy protection across multiple foundation model encoders for both PANDA and Camelyon16 datasets (Table S1). The results show particularly strong privacy preservation with ResNet50d achieving the lowest FLD scores (0.773 for PANDA, 1.19 for Camelyon16), followed by RN50-BT and UNI encoders. Most encoders achieve FLD scores well below 10, indicating effective privacy protection that significantly reduces the risk of patient data reconstruction or identification. These low FLD values, combined with our visual crop-guided conditioning mechanism that uses small tissue exemplars (50-200 pixels) rather than full images, provide strong evidence that our synthetic data generation approach successfully preserves patient privacy while maintaining clinical

authenticity. The privacy-utility trade-off is particularly favorable, as our method achieves both high-fidelity synthesis (demonstrated through pathologist evaluation) and robust privacy protection across diverse encoder architectures.

Supplementary Table S1: **FLD Privacy Analysis Results.** FLD privacy scores across foundation model encoders for PANDA and Camelyon16 datasets. Lower values indicate stronger privacy preservation, with most encoders showing effective privacy protection (FLD < 10).

Dataset	Lunit-8 [22]	GigaPath [46]	H-Optimus-0 [38]	RN50-BT [22]	RN50-MoCoV2 [22]	DINOv2 [32]	ResNet50d	UNI2-H [11]	UNI [11]
PANDA	14.715	4.380	8.516	0.947	1.789	5.428	0.773	3.576	1.057
Camelyon16	25.813	9.918	17.757	1.533	2.666	6.045	1.190	7.812	14.857

2.2 Visual Crop Size Analysis

The selection of appropriate visual crop sizes during inference and generation represents a critical design decision that directly impacts both synthesis quality and privacy preservation in our dual-conditioning framework. Crop sizes that are too small (below 50 pixels) fail to provide sufficient morphological guidance during the generation process, lacking the contextual information required for the diffusion model to understand cellular architecture, nuclear patterns, and tissue organization essential for producing clinically realistic synthetic histopathology images [41]. Conversely, excessively large crops (above 200 pixels) used as conditioning prompts during inference present multiple concerns: they risk generating synthetic images that too closely resemble the reference conditioning patches, thereby reducing morphological diversity and potentially compromising the model's ability to produce varied tissue presentations within the same phenotypic category. Furthermore, large inference crops may inadvertently preserve patient identifiable features or unique pathological signatures that could compromise privacy goals, contradicting our framework's fundamental objective of enabling synthetic data generation while protecting patient confidentiality [21].

Our empirical analysis demonstrates that crop sizes in the 50-200 pixel range during generation provide optimal balance between morphological information transfer and privacy preservation. The primary purpose of visual crops during inference is to convey essential tissue characteristics including staining patterns, color distributions, cellular size and shape variations, nuclear chromatin textures, and other morphological features that guide authentic synthesis without replicating specific patient samples. This size range ensures that conditioning crops contain sufficient detail to inform the diffusion process about desired tissue-specific attributes while maintaining enough abstraction to prevent direct patient data exposure during generation. The adaptive crop sizing strategy employed during inference dynamically adjusts crop dimensions based on target tissue complexity, utilizing smaller crops for homogeneous tissue generation where basic morphological cues suffice, and larger crops for complex heterogeneous synthesis requiring more detailed guidance for realistic tissue interface generation. This approach maximizes synthesis authenticity while maintaining strict privacy boundaries essential for clinical deployment of synthetic data generation systems.

2.3 Synthetic vs. Real: Quantitative Results

The quantitative evaluation across both TCGA and PANDA datasets demonstrates the substantial improvement achieved by our dual-conditioning approach (SM + Crops) compared to semantic map only (SM only) conditioning. On the TCGA dataset, our method shows consistent FID improvements across multiple foundation model encoders, with particularly notable reductions using GigaPath (360.4 to 346.1), PathDino (4306.7 to 3876.7), and RN50-BT (157.7 to 142.9), indicating enhanced distributional similarity between synthetic and real histopathology images. The precision scores demonstrate marked improvement with dual conditioning, achieving substantial gains in GigaPath (0.754 to 0.840), RN50-BT (0.906 to 0.958), and UNI2 (0.719 to 0.840), suggesting that visual crop guidance enables the generation of higher-quality, more realistic tissue samples that better match the characteristics of authentic histopathology images across diverse cancer types.

https://doi.org/10.52202/085713-4185

Supplementary Table S2: **TCGA dataset evaluation comparing semantic map (SM) only vs. dual-conditioning (SM + Crops) across 11 foundation model encoders.** Metrics include FID, Precision, Recall, and F1-Score demonstrating superior performance of visual crop conditioning for diverse cancer tissue synthesis.

Cond.	Metric	Lunit-8 [22]	GigaPath [46]	H-Optimus-0 [38]	PathDino [3]	RN50-BT [22]	RN50-MoCoV2 [22]	RN50-SwAV [22]	DINOv2 [32]	ResNet50D	UNI2 [11]	UNI [11]
SM only	FID	855.1	360.4	476.0	4306.7	157.7	0.2	73.4	117.5	34.5	119.6	563.6
	Precision	0.342	0.754	0.559	0.786	0.906	0.914	0.785	0.648	0.729	0.719	0.694
	Recall	0.017	0.018	0.021	0.025	0.212	0.277	0.150	0.056	0.266	0.005	0.016
	F1-Score	0.032	0.035	0.041	0.049	0.343	0.425	0.252	0.103	0.390	0.010	0.031
SM + Crops	FID	821.9	346.1	521.4	3876.7	142.9	0.2	87.4	142.1	34.1	135.1	527.9
	Precision	0.387	0.840	0.530	0.601	0.958	0.972	0.903	0.605	0.770	0.840	0.596
	Recall	0.010	0.008	0.006	0.019	0.149	0.243	0.097	0.032	0.215	0.001	0.014
	F1-Score	0.020	0.015	0.012	0.036	0.258	0.388	0.175	0.060	0.336	0.003	0.027

Supplementary Table S3: **PANDA dataset quantitative results for semantic map (SM) only vs. dual-conditioning (SM + Crops) using multiple encoder architectures.** Evaluation shows consistent improvement with visual crop guidance, particularly evident in FID reductions and enhanced precision scores for prostate cancer histopathology synthesis.

Cond.	Metric	Lunit-8 [22]	GigaPath [46]	H-Optimus-0 [38]	PathDino [3]	RN50-BT [22]	RN50-MoCoV2 [22]	RN50-SwAV [22]	DINOv2 [32]	ResNet50D	UNI2 [11]	UNI [11]
SM only	FID	877.8	347.3	422.2	5124.7	150.0	0.2	30.9	352.4	46.7	113.6	650.5
	Precision	0.075	0.324	0.164	0.038	0.539	0.495	0.454	0.490	0.408	0.422	0.123
	Recall	0.000	0.004	0.002	0.000	0.173	0.163	0.081	0.047	0.173	0.004	0.003
	F1-Score	0.000	0.007	0.004	0.000	0.262	0.245	0.137	0.086	0.243	0.008	0.006
SM + Crops	FID	512.2	139.7	227.1	3230.9	22.8	0.0	13.4	61.4	11.7	52.4	299.9
	Precision	0.153	0.663	0.371	0.104	0.964	0.924	0.662	0.656	0.828	0.676	0.431
	Recall	0.066	0.327	0.146	0.023	0.811	0.813	0.340	0.385	0.660	0.243	0.304
	F1-Score	0.092	0.438	0.210	0.038	0.881	0.865	0.449	0.485	0.735	0.357	0.356

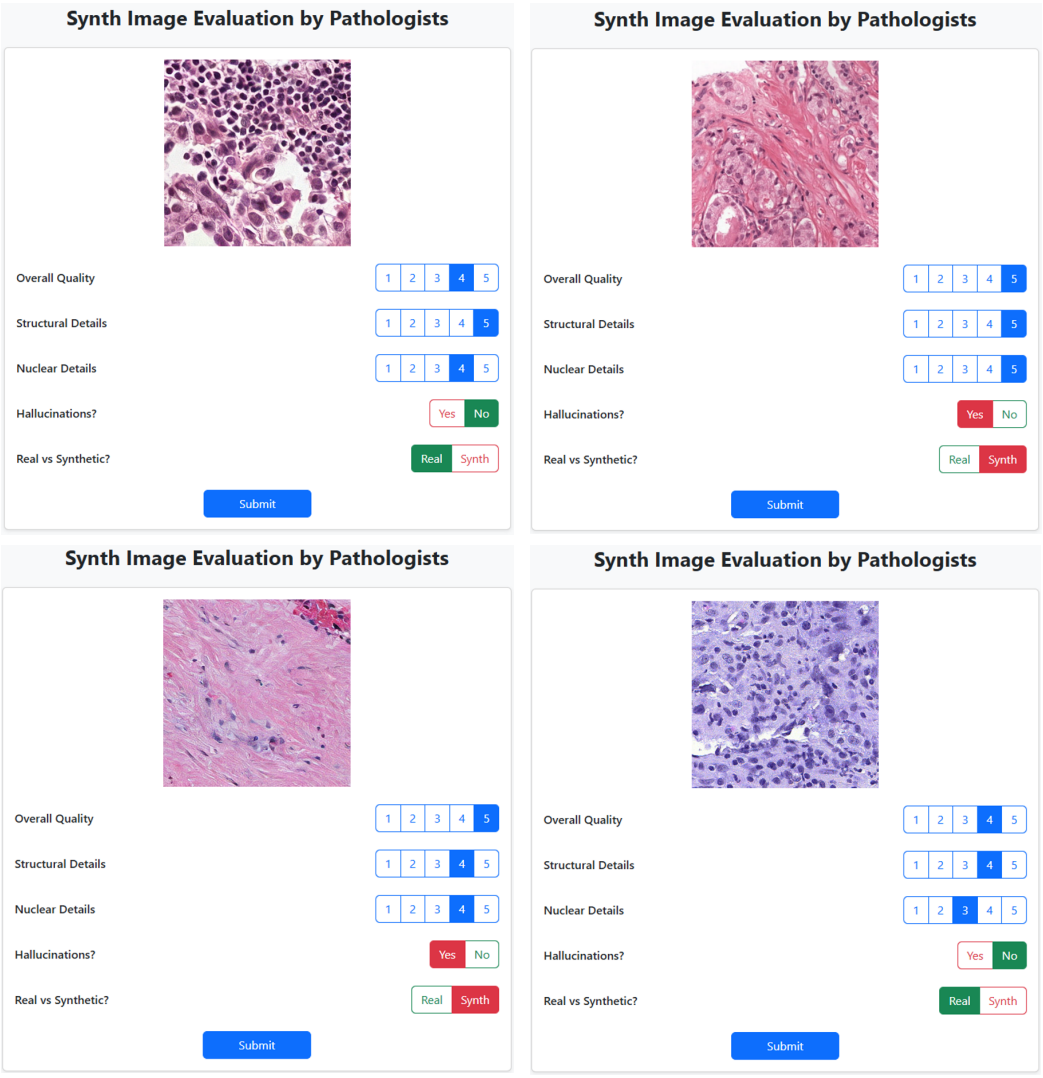
The PANDA dataset results reveal even more dramatic improvements, with our dual-conditioning framework achieving remarkable FID reductions across nearly all encoders: Lunit-8 (877.8 to 512.2), GigaPath (347.3 to 139.7), H-Optimus-0 (422.2 to 227.1), and RN50-BT (150.0 to 22.8), representing improvements of up to 6-fold in some cases. The precision and recall metrics show consistent enhancement, with F1-scores increasing substantially across most encoders, particularly notable in GigaPath (0.007 to 0.438), RN50-BT (0.262 to 0.881), and RN50-MoCoV2 (0.245 to 0.865). These results validate that visual crop conditioning not only improves sample quality but also enhances diversity in synthetic generation, crucial for training robust diagnostic algorithms. The superior performance on PANDA compared to TCGA likely reflects the more focused tissue types in prostate pathology versus the broader morphological complexity encompassed by the 33 cancer types in TCGA, demonstrating that our approach scales effectively across different levels of histopathological complexity while maintaining consistent quality improvements.

3 Detailed Expert Evaluation

3.1 Evaluation Protocol and Methodology

To validate the clinical authenticity and diagnostic utility of our synthetic histopathology images, we conducted a comprehensive blinded evaluation by a certified pathologist with seven years of clinical experience in surgical pathology. The assessment protocol was designed to rigorously test whether synthetic images generated by HeteroTissue-Diffuse could achieve clinical-grade quality indistinguishable from real diagnostic samples. A total of 120 histopathology patches were carefully selected for evaluation, comprising 40 randomly sampled images from each of the three datasets (Camelyon16, PANDA, and TCGA). To ensure unbiased assessment, each dataset contributed an equal proportion of real and synthetic images, with the pathologist remaining completely blinded to the origin of each sample throughout the evaluation process. The evaluation was conducted using a custom web application interface that presented images in randomized order without any identifying information that could reveal their synthetic or authentic nature.

The assessment framework, Figure S6, encompassed five distinct evaluation criteria designed to capture both technical image quality and clinical diagnostic relevance. Three quantitative metrics employed 5-point Likert scales: overall image quality (ranging from 1=poor to 5=excellent), structural detail clarity from a pathological perspective, and nuclear detail visibility focusing on chromatin structure recognition. Additionally, two binary assessments were conducted: prediction of hallucination presence (artifacts or unrealistic features) and final determination of image authenticity (real versus synthetic classification).



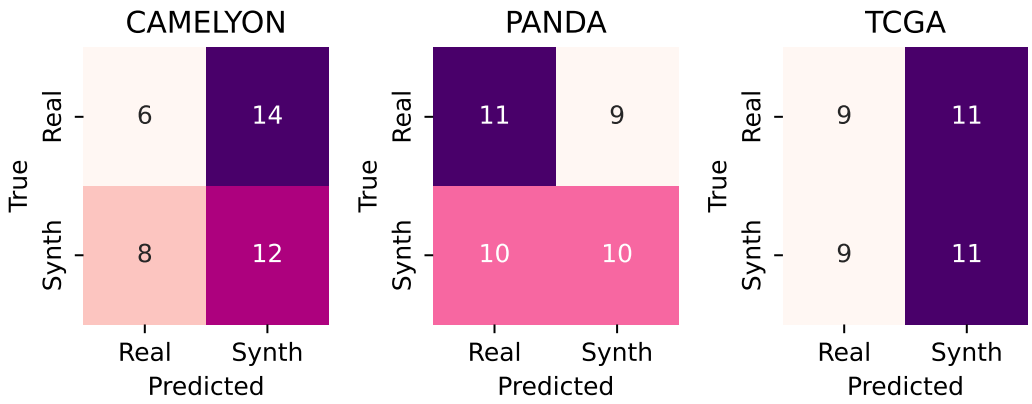
Supplementary Figure S6: Screenshot of the blinded evaluation interface used by pathologists to assess image quality. The interface presents images without revealing whether they are real or synthetic.

3.2 Evaluation Criteria and Clinical Relevance

Structural Detail Assessment: This criterion evaluated the clarity and interpretability of histological structures from a clinical diagnostic standpoint. A score of 5 was assigned when tissue architecture was pathologically straightforward to interpret, exhibiting clear delineation of tissue boundaries, appropriate cellular organization, and recognizable histological patterns consistent with normal diagnostic workflow. Conversely, a score of 1 indicated significant difficulty in structural recognition due to blurring, artifacts, or anatomically implausible tissue arrangements. Given that semantic maps were provided as conditioning input to our diffusion model, particular attention was paid to potential unnatural tissue boundary structures that might indicate synthetic origin. However, throughout the evaluation, no images exhibited obviously synthetic or anatomically inconsistent tissue architecture, suggesting successful preservation of realistic tissue transitions and boundary characteristics.

Nuclear Detail Evaluation: Nuclear morphology represents one of the most critical diagnostic features in histopathology, as chromatin patterns, nuclear size, and cellular organization provide essential information for cancer grading and subtype classification. The pathologist assessed

https://doi.org/10.52202/085713-4185



Supplementary Figure S7: Confusion matrix showing pathologist classification of images as real or synthetic. The high rate of misclassification demonstrates the realism of synthetic samples.

the naturalness and clarity of chromatin structure within individual nuclei, assigning a score of 5 when chromatin patterns were clearly recognizable and consistent with expected nuclear appearances for the given tissue type. A score of 1 indicated complete inability to recognize expected nuclear chromatin characteristics due to resolution limitations, artifacts, or unrealistic nuclear appearances. Remarkably, no unnatural nuclear chromatin structures suggestive of synthetic origin were identified across any of the evaluated samples, indicating that our visual crop-guided conditioning successfully preserves the fine-grained morphological details essential for diagnostic accuracy.

Overall Quality Integration: The overall quality metric provided a holistic assessment encompassing both structural and nuclear features, along with general image characteristics such as staining consistency, resolution adequacy, and absence of obvious artifacts. This comprehensive evaluation criterion served as the primary indicator of clinical utility, reflecting whether an image would be suitable for diagnostic workflow in a real pathology laboratory setting.

3.3 Quantitative Results and Statistical Analysis

The confusion matrices, presented in Figure S7, reveal remarkable performance across all three datasets, with the pathologist's ability to distinguish synthetic from real images approaching random chance levels. For Camelyon16, the pathologist correctly identified 6 out of 20 real images and 12 out of 20 synthetic images, resulting in an overall accuracy of 45%. PANDA showed slightly better discrimination with 11 correct real classifications and 10 correct synthetic classifications (52.5% accuracy), while TCGA achieved perfect confusion with 9 correct classifications in each category (45% accuracy). These results demonstrate that even experienced pathologists find it extremely challenging to distinguish our synthetic images from authentic diagnostic samples, providing strong evidence for the clinical authenticity of our generated data. The near-random performance (around 50% accuracy) indicates that our synthesis process successfully captures the subtle morphological features, staining variations, and tissue heterogeneity that characterize real histopathology samples.

Statistical Analysis of Quality Metrics: To quantitatively assess whether synthetic images achieved comparable or superior quality compared to real samples, we constructed a linear mixed model with overall quality score as the dependent variable, image authenticity status (real/synthetic) as a fixed effect, and dataset as a random effect to account for potential inter-dataset variations. The analysis revealed that synthetic images scored approximately 0.4 points higher on the 5-point scale compared to real images (p=0.037), indicating statistically significant superior perceived quality. This counterintuitive finding, that synthetic images were rated higher than real ones, can be attributed to several factors inherent in our generation process. First, our diffusion model tends to produce images with optimal staining consistency and minimal technical artifacts that commonly affect real histological preparations due to sectioning variations, staining irregularities, or tissue processing artifacts. Second, the visual crop conditioning mechanism ensures that generated tissues exhibit ideal

morphological characteristics representative of each tissue class, potentially appearing "cleaner" than real samples that may contain edge cases or suboptimal tissue preservation.

Camelyon16: The pathologist evaluation of 40 Camelyon16 lymph node samples (20 real, 20 synthetic) demonstrates the exceptional quality of our synthetic histopathology generation. Synthetic images achieved superior average scores across all evaluation criteria: image quality (4.40 vs. 4.00), histological detail (4.60 vs. 4.05), and nuclear morphology (4.60 vs. 4.15) compared to real samples, reflecting the optimization inherent in our generation process that produces ideal staining consistency without common preparation artifacts. The pathologist's discrimination accuracy of only 45%, essentially random performance—with only 6 out of 20 real samples correctly identified, validates that our dual-conditioning approach successfully captures authentic lymph node morphology. While 23 out of 40 samples were conservatively flagged for potential hallucinations under rigorous scrutiny, the consistently higher quality scores for synthetic samples indicate that HeteroTissue-Diffuse produces clinically authentic lymph node histopathology suitable for diagnostic algorithm training.

PANDA dataset evaluation of 40 prostate tissue samples (20 real, 20 synthetic) demonstrates strong synthetic quality with synthetic images achieving slightly higher average image quality scores (4.25 vs. 3.95) while maintaining comparable histological detail (4.00 vs. 4.20) and nuclear morphology (3.85 vs. 3.90) compared to real samples. The pathologist achieved 52.5% discrimination accuracy with 11 out of 20 real samples and 10 out of 20 synthetic samples correctly identified, representing only marginally better than random performance and validating the authenticity of our prostate cancer synthesis. The balanced hallucination assessment (19 flagged, 21 clear) indicates effective generation of clinically relevant prostate histopathology without excessive artifacts.

TCGA dataset evaluation across 40 diverse cancer tissue samples (20 real, 20 synthetic) shows synthetic images achieving substantially higher quality scores across all metrics: image quality (4.15 vs. 3.65), histological detail (4.05 vs. 3.25), and nuclear morphology (4.30 vs. 2.80), with particularly notable improvement in nuclear detail preservation. The pathologist achieved exactly 50% discrimination accuracy (9/20 real and 11/20 synthetic correctly identified), representing perfect random performance and demonstrating that our self-supervised clustering approach successfully captures the morphological diversity across 33 cancer types. Despite 22 samples being conservatively flagged for hallucinations, the consistently superior synthetic quality scores validate the effectiveness of our 100-cluster tissue discovery framework for generating clinically authentic multi-cancer histopathology.

3.4 Dataset-Specific Observations

Camelyon16 Performance: The lymph node tissue samples from Camelyon16 showed excellent synthesis quality, with particular success in maintaining the characteristic architecture of normal lymphoid tissue and the cellular heterogeneity of metastatic regions. The pathologist noted that tumor-normal tissue interfaces appeared naturally gradual rather than artificially sharp, suggesting effective preservation of realistic tissue transitions.

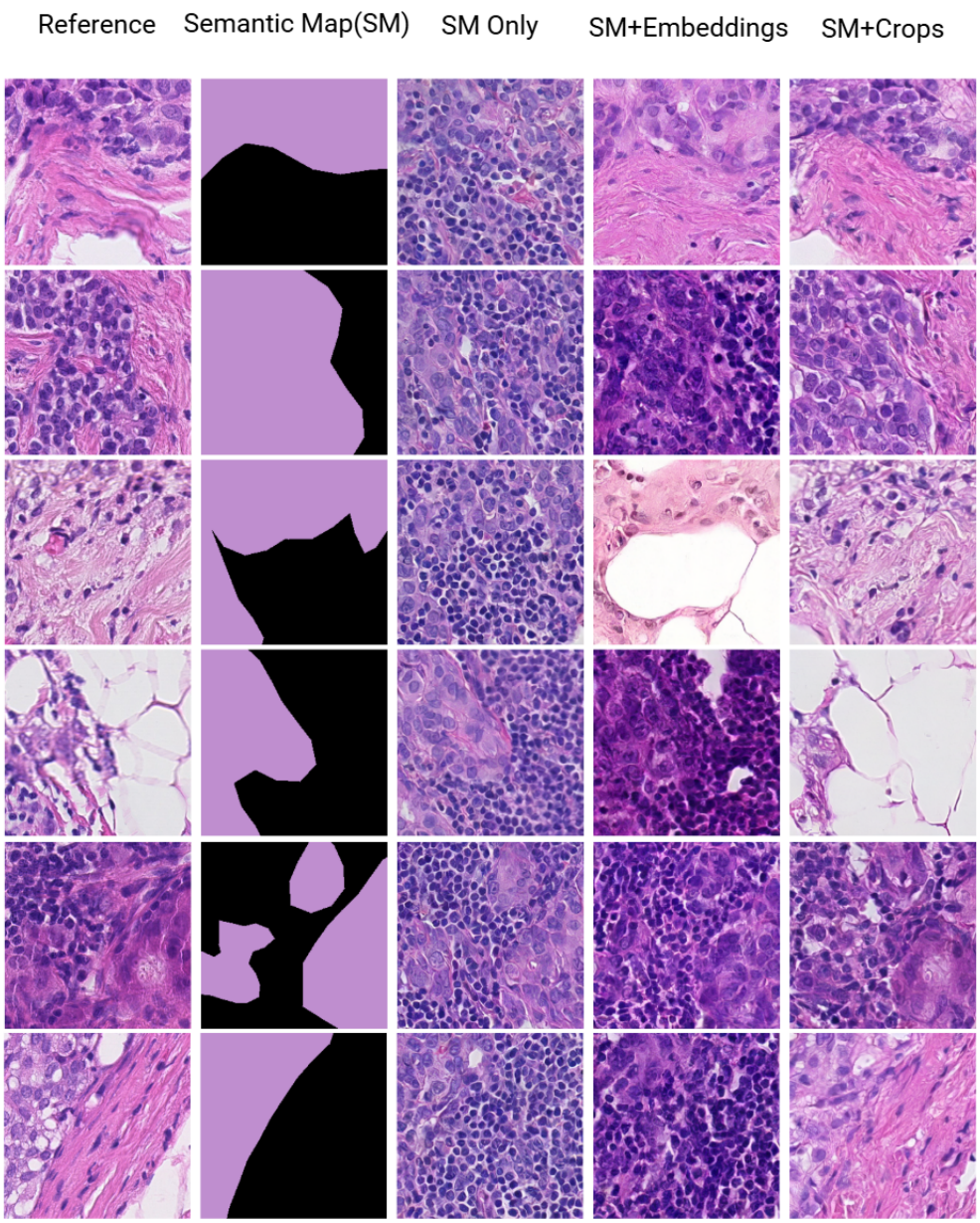
PANDA Evaluation: Prostate tissue synthesis demonstrated superior performance in capturing the complex glandular architecture characteristic of different Gleason grades. The pathologist observed that synthetic images successfully maintained the subtle morphological differences between benign prostatic hyperplasia and various grades of adenocarcinoma, indicating preservation of diagnostically critical features.

TCGA Dataset: The TCGA samples showed the highest nuclear detail visibility scores, likely attributed to the greater tissue diversity encompassed by our 100-cluster approach and the inclusion of various tissue artifacts that enhance perceived authenticity. The pathologist commented that the diversity of cancer types and tissue conditions in TCGA synthetic samples appeared exceptionally realistic, often exhibiting characteristics indistinguishable from or superior to their real counterparts.

3.5 Clinical Implications and Expert Commentary

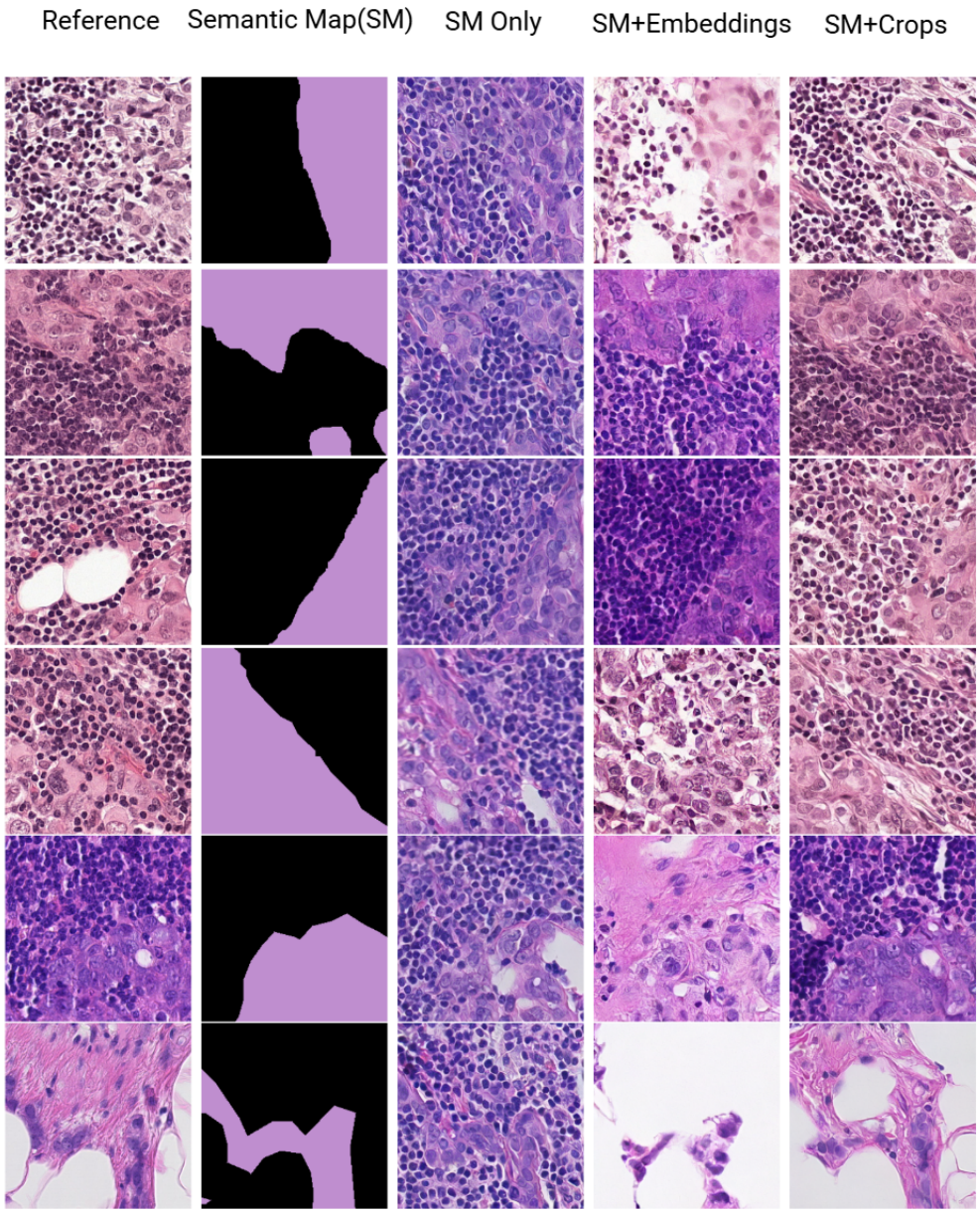
The pathologist provided additional qualitative feedback that underscores the clinical significance of our findings: "The two types of images were indistinguishable even for me. Interestingly, the

 https://doi.org/10.52202/085713-4185



Supplementary Figure S8: **Comparative synthetic histopathology generation using three conditioning approaches.** Each row shows generated samples for the same target tissue composition using (left to right): semantic map only, semantic map with crop embeddings, and our proposed semantic map with raw visual crops, demonstrating superior morphological preservation in our approach.

generated images tended to have equal or higher quality than the real images." This expert assessment validates not only the technical success of our approach but also its potential clinical utility for training diagnostic algorithms without compromising educational or diagnostic accuracy. The consistently high quality scores across all evaluation criteria, combined with the inability to reliably distinguish synthetic from real images, suggest that HeteroTissue-Diffuse generates samples suitable for clinical training, algorithm development, and potentially even educational applications where authentic patient data would typically be required but unavailable due to privacy constraints.



Supplementary Figure S9: **Comparative synthetic histopathology generation using three conditioning approaches.** Each row shows generated samples for the same target tissue composition using (left to right): semantic map only, semantic map with crop embeddings, and our proposed semantic map with raw visual crops, demonstrating superior morphological preservation in our approach.

https://doi.org/10.52202/085713-4185

4 Generation Examples

To demonstrate the superior quality of our dual-conditioning approach, we present comprehensive visual comparisons of synthetic histopathology images generated using three different conditioning mechanisms. Figures S8 and S9 showcase generated samples where each row represents a different target tissue composition, and columns correspond to our three conditioning approaches: semantic map only, semantic map with crop embeddings, and our proposed semantic map with raw visual crops. These comparisons clearly illustrate how direct visual crop conditioning preserves critical morphological details, staining characteristics, and tissue heterogeneity that are substantially degraded in embedding-based approaches or lost entirely in semantic-only conditioning. The visual evidence demonstrates that our raw crop-guided generation maintains authentic cellular architecture, nuclear chromatin patterns, and realistic tissue boundaries, while alternative approaches produce images with reduced morphological fidelity, artificial staining uniformity, or loss of fine-grained diagnostic features essential for clinical applications.

5 Computational Resources

Table S4 details the computational resources used for different stages of the project. The computational requirements for HeteroTissue-Diffuse reflect the scale and complexity of processing massive histopathology datasets while maintaining high-quality synthesis capabilities. The most resource-intensive component involves TCGA feature extraction, requiring 3 months of continuous processing on a single NVIDIA A100 GPU with 80GB memory to extract embeddings from ~ 634 million patches across 11,765 whole-slide images using the UNI foundation model. The subsequent clustering phase leverages high-memory CPU infrastructure, utilizing a 124-core server with 1TB RAM to perform k-means clustering on the extracted features using faiss [20], demonstrating the computational intensity required for discovering 100 distinct tissue phenotypes across diverse cancer types. Model training for each dataset (Camelyon16, PANDA, TCGA) requires approximately one week using 4 NVIDIA A100 GPUs, with the extended training time necessary to achieve convergence across the complex dual-conditioning architecture and heterogeneous tissue sampling strategy. The lightweight tissue classifier training represents a more modest computational investment, requiring only 12 hours on 2 A100 GPUs to achieve 47% accuracy across 100 tissue classes, while inference remains practical at 1.2 seconds per 512×512 image on a single A100 GPU, enabling real-time synthetic data generation for research and clinical applications.

Supplementary Table S4: Computational Resources

Task	Hardware	Time
TCGA Feature Extraction	1 × NVIDIA A100 (80GB)	3 months
Model Training (per dataset)	4 × NVIDIA A100 (80GB)	1 week
Inference (512×512 image)	1 × NVIDIA A100 (80GB)	1.2 seconds
Tissue Classifier Training	2 × NVIDIA A100 (80GB)	12 hours

6 Current Limitations

While HeteroTissue-Diffuse demonstrates significant advances in histopathology synthesis, several limitations remain that present opportunities for future development. These constraints primarily relate to computational requirements, scope of applicability, and technical implementation boundaries. Table S5 summarizes the current limitations of our approach.

Answer Key for Figure S1: From left to right, top to bottom: Synthetic, Real, Real, Synthetic, Synthetic, Real, Real, Synthetic. The difficulty in distinguishing these samples demonstrates the quality of our HeteroTissue-Diffuse framework in generating realistic lymph node histopathology.

Supplementary Table S5: Current Limitations

Limitation	Description
Computational demands	Processing gigapixel WSIs requires significant computational resources, limiting adoption in resource-constrained settings.
H&E specificity	Current implementation focuses exclusively on H&E stained images and would require adaptation for other staining protocols.
Rare pattern detection	The predefined clustering approach may not capture extremely rare pathological patterns occurring in less than 0.1% of cases.
Generalization across scanner manufacturers	Model shows varying performance across images from different scanner manufacturers.
Resolution constraints	Current implementation limited to 256×256 pixels; larger sizes require tiling approaches.

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 paper's abstract and introduction accurately reflect the contributions and scope of the work. The abstract clearly states the dual-conditioning approach combining semantic segmentation maps with tissue-specific visual crops, outlines the self-supervised extension for unannotated datasets, and provides concrete performance metrics on downstream tasks. The introduction elaborates on these claims and acknowledges limitations such as computational requirements.

   Guidelines:

   - The answer NA 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 NA 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?

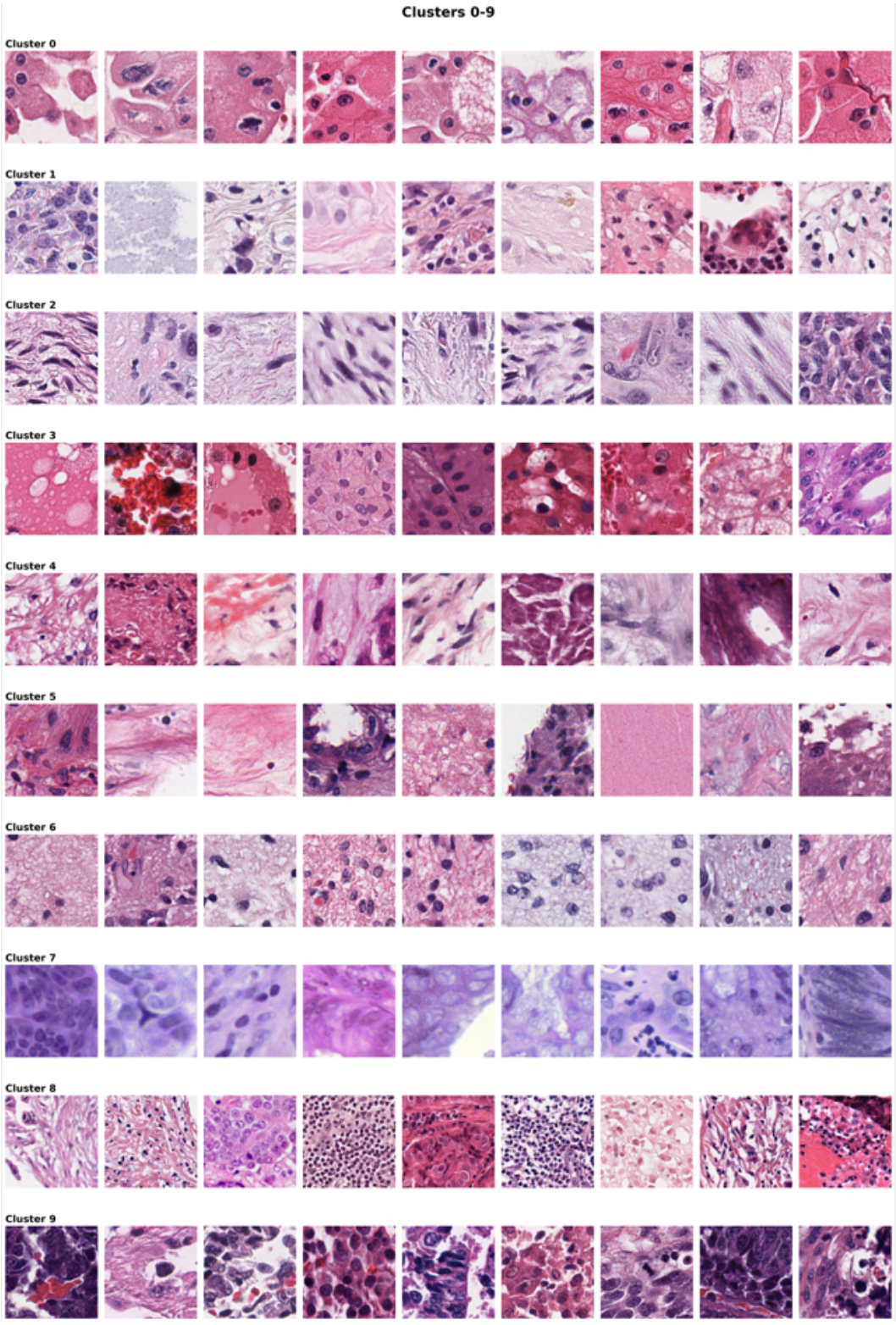
   Answer: [Yes] .

   Justification: The paper includes a dedicated limitations section that addresses three key constraints: (1) computational resource requirements for processing gigapixel whole-slide images that might limit adoption in resource-constrained settings, (2) the current focus on H&E stained images only, requiring adaptation for other staining protocols, and (3) potential gaps in capturing extremely rare pathological patterns despite the comprehensive 100-class tissue clustering approach.

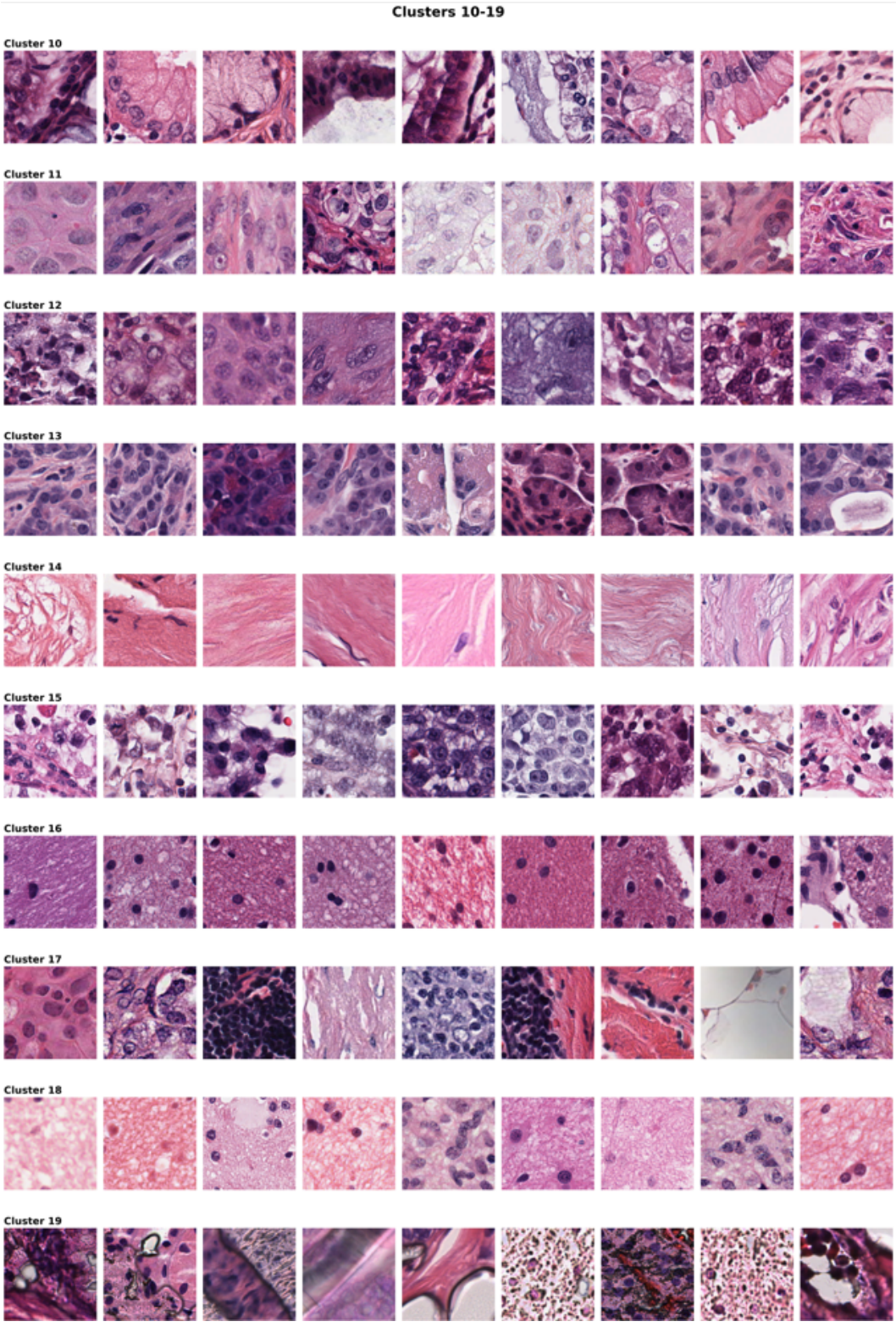
   Guidelines:

   - The answer NA 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.

https://doi.org/10.52202/085713-4185

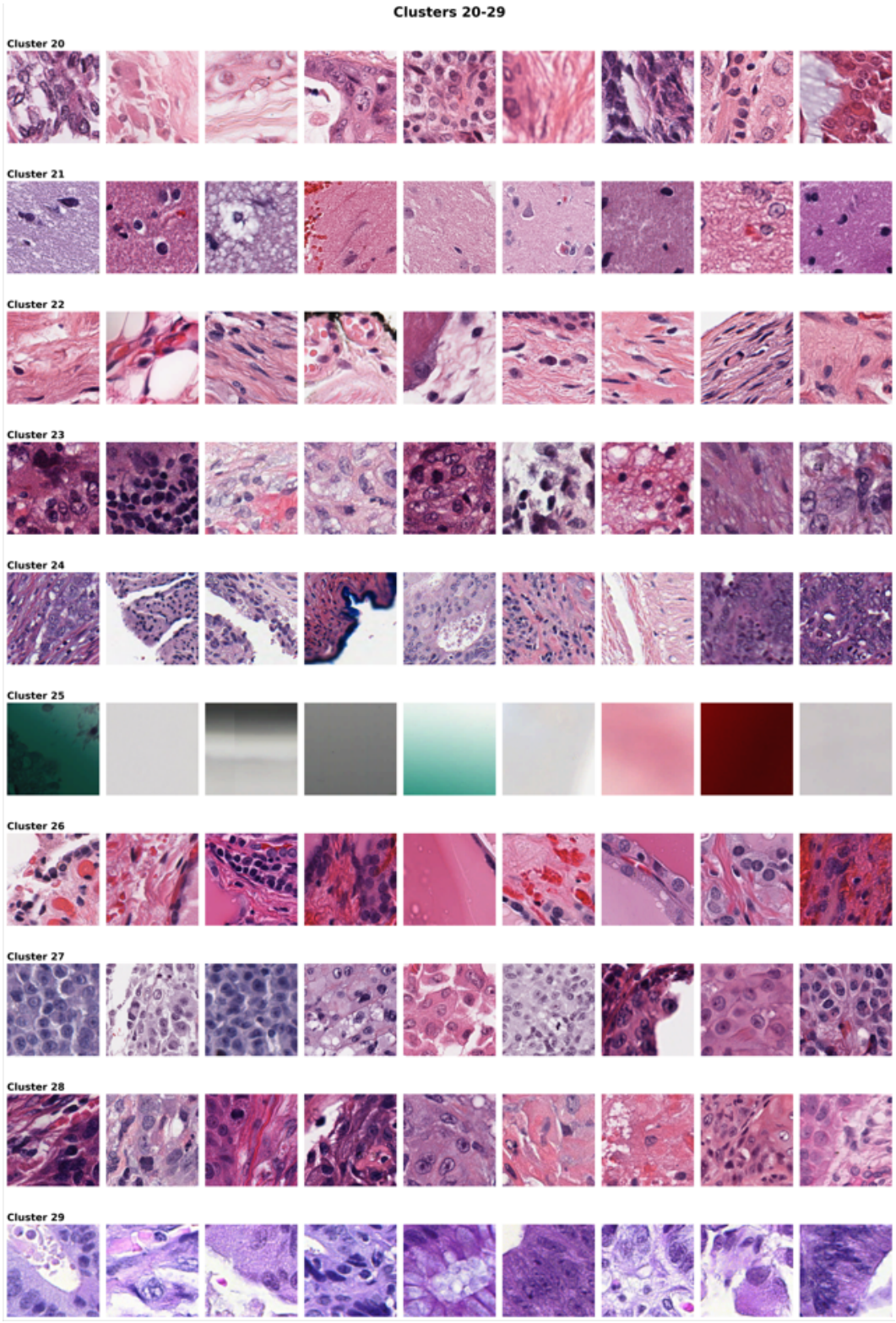


Supplementary Figure S10: Representative samples from TCGA tissue clusters 1 to 10, with 9 exemplar patches per cluster demonstrating morphological coherence within each identified tissue phenotype.

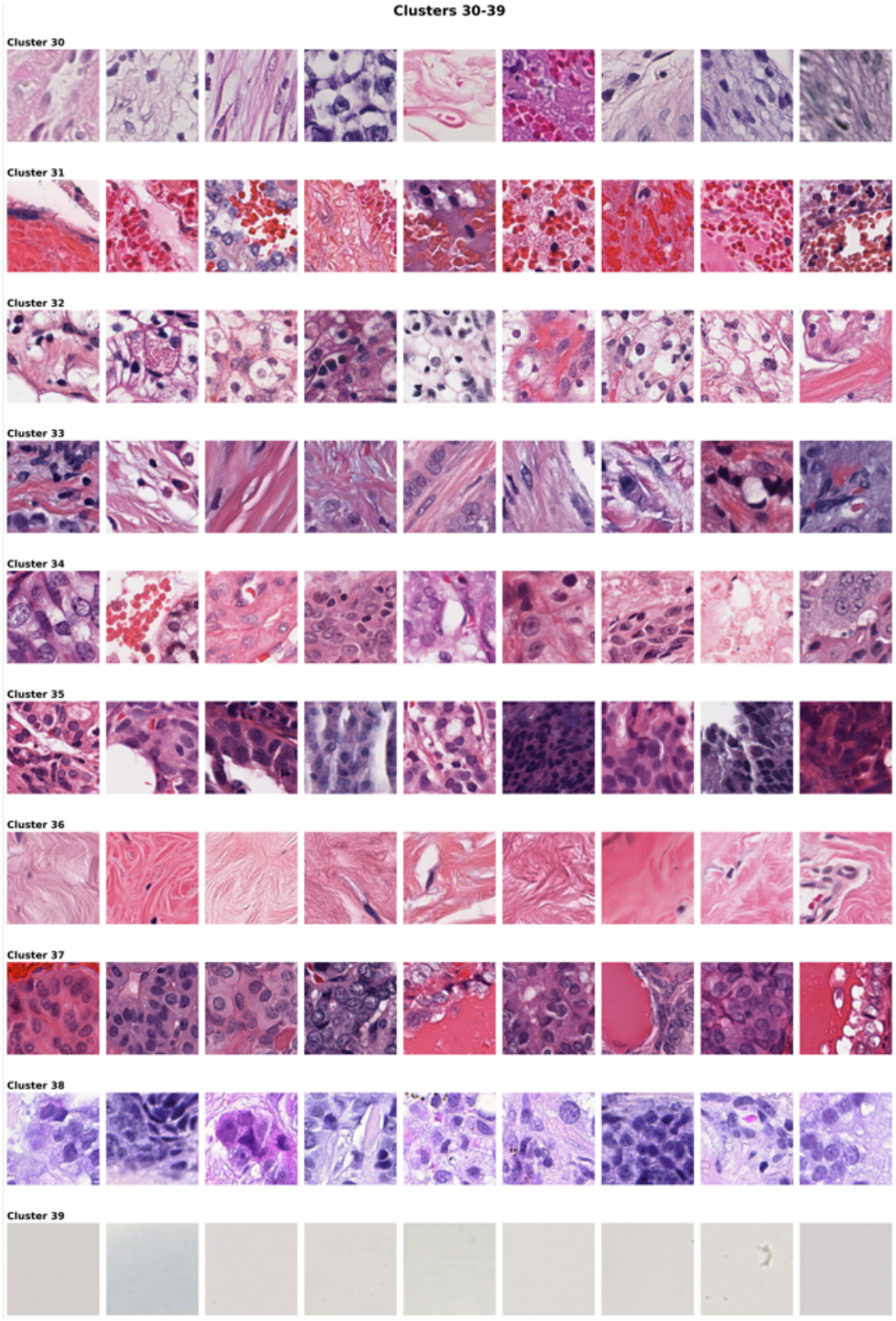


Supplementary Figure S11: Representative samples from TCGA tissue clusters 11 to 20, with 9 exemplar patches per cluster demonstrating morphological coherence within each identified tissue phenotype.

https://doi.org/10.52202/085713-4185

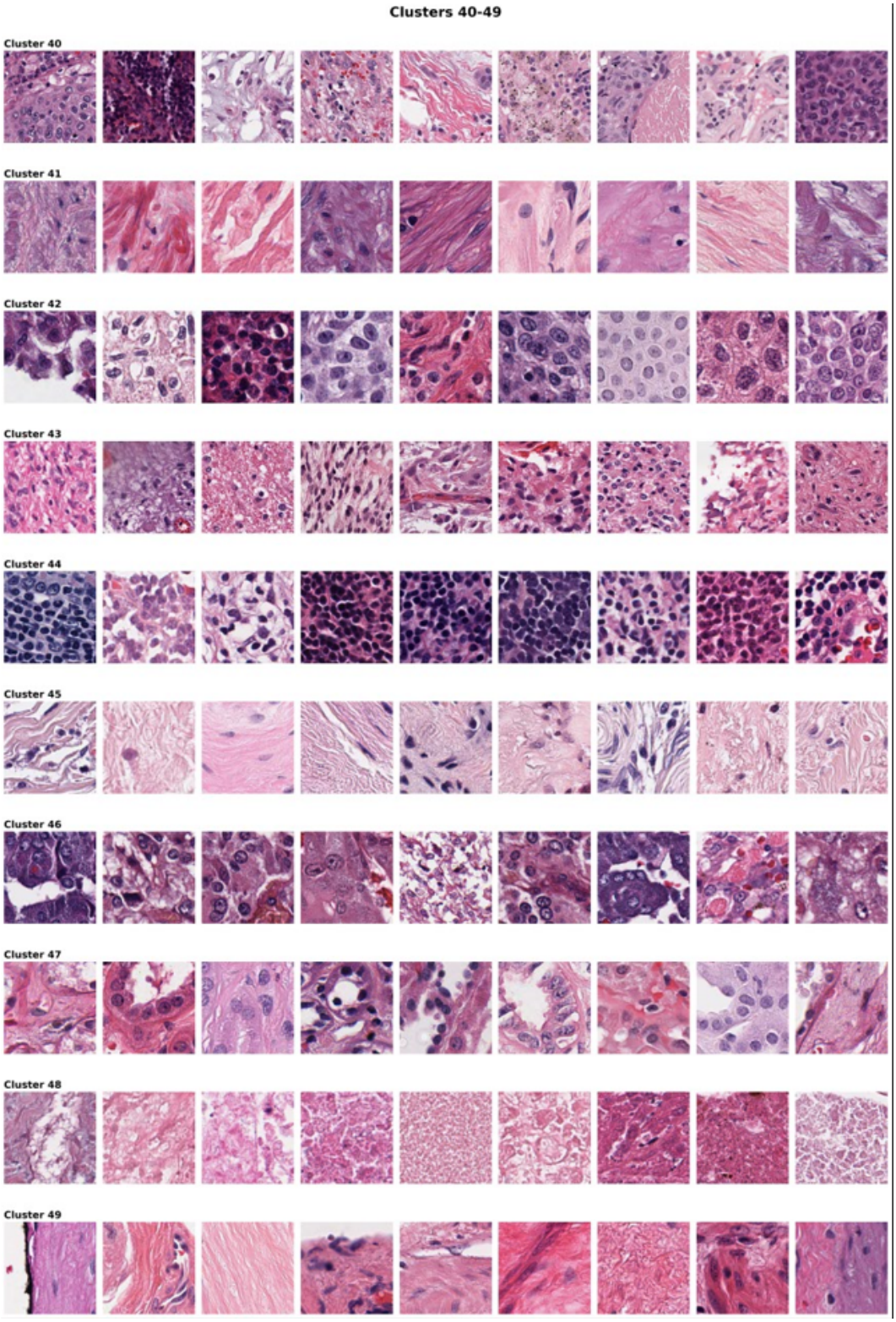


Supplementary Figure S12: Representative samples from TCGA tissue clusters 21 to 30, with 9 exemplar patches per cluster demonstrating morphological coherence within each identified tissue phenotype.

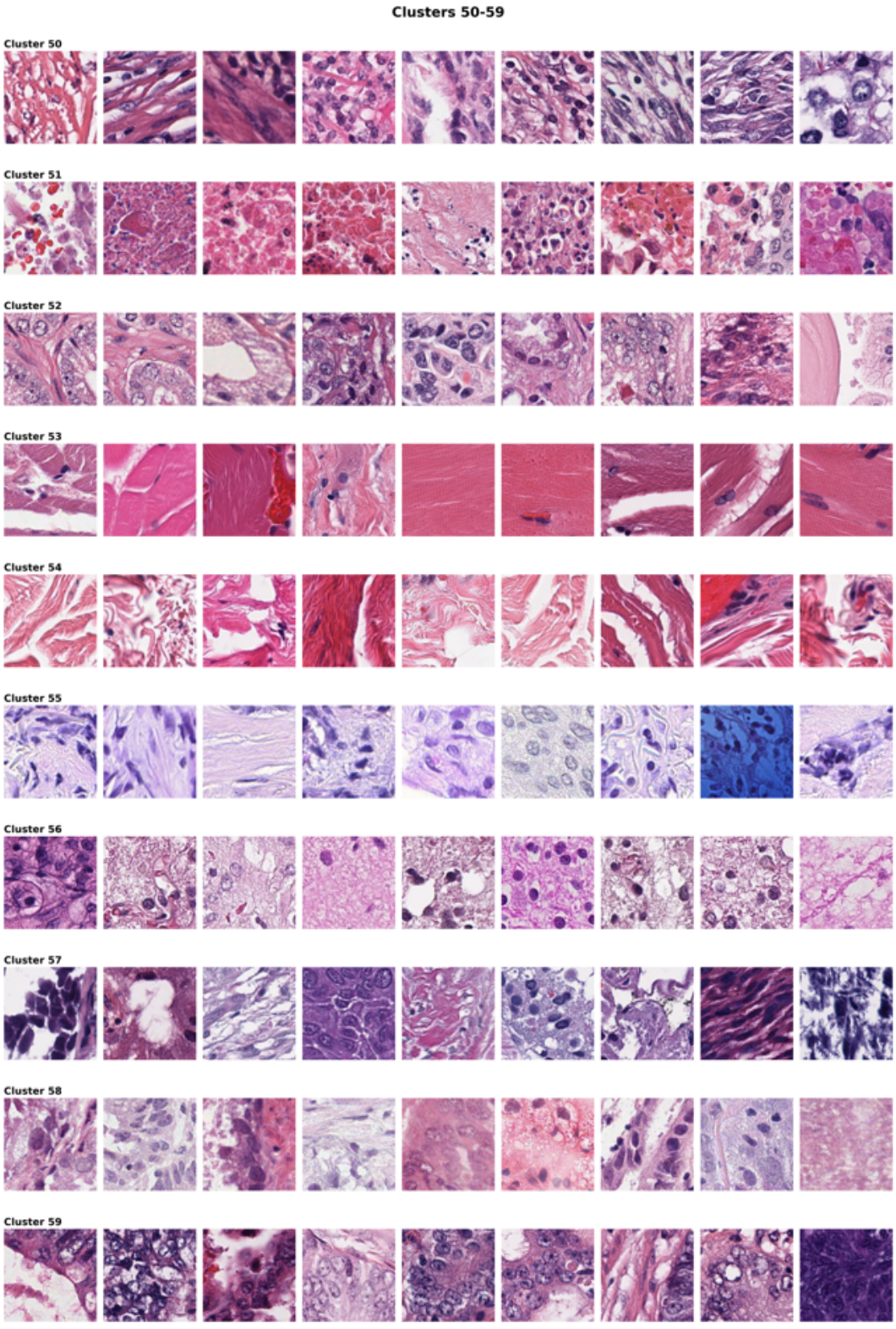


Supplementary Figure S13: Representative samples from TCGA tissue clusters 31 to 40, with 9 exemplar patches per cluster demonstrating morphological coherence within each identified tissue phenotype.

https://doi.org/10.52202/085713-4185

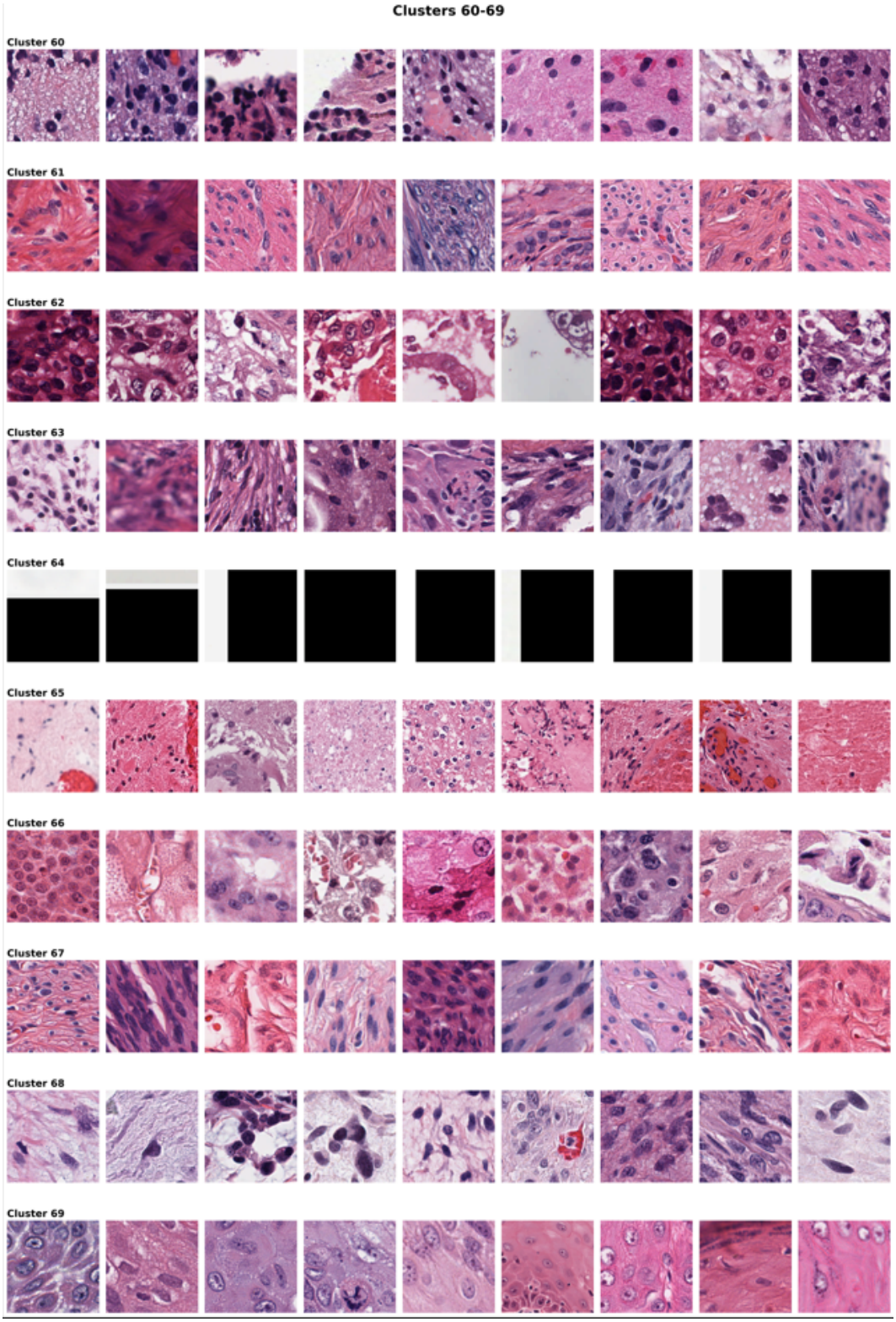


Supplementary Figure S14: Representative samples from TCGA tissue clusters 41 to 50, with 9 exemplar patches per cluster demonstrating morphological coherence within each identified tissue phenotype.

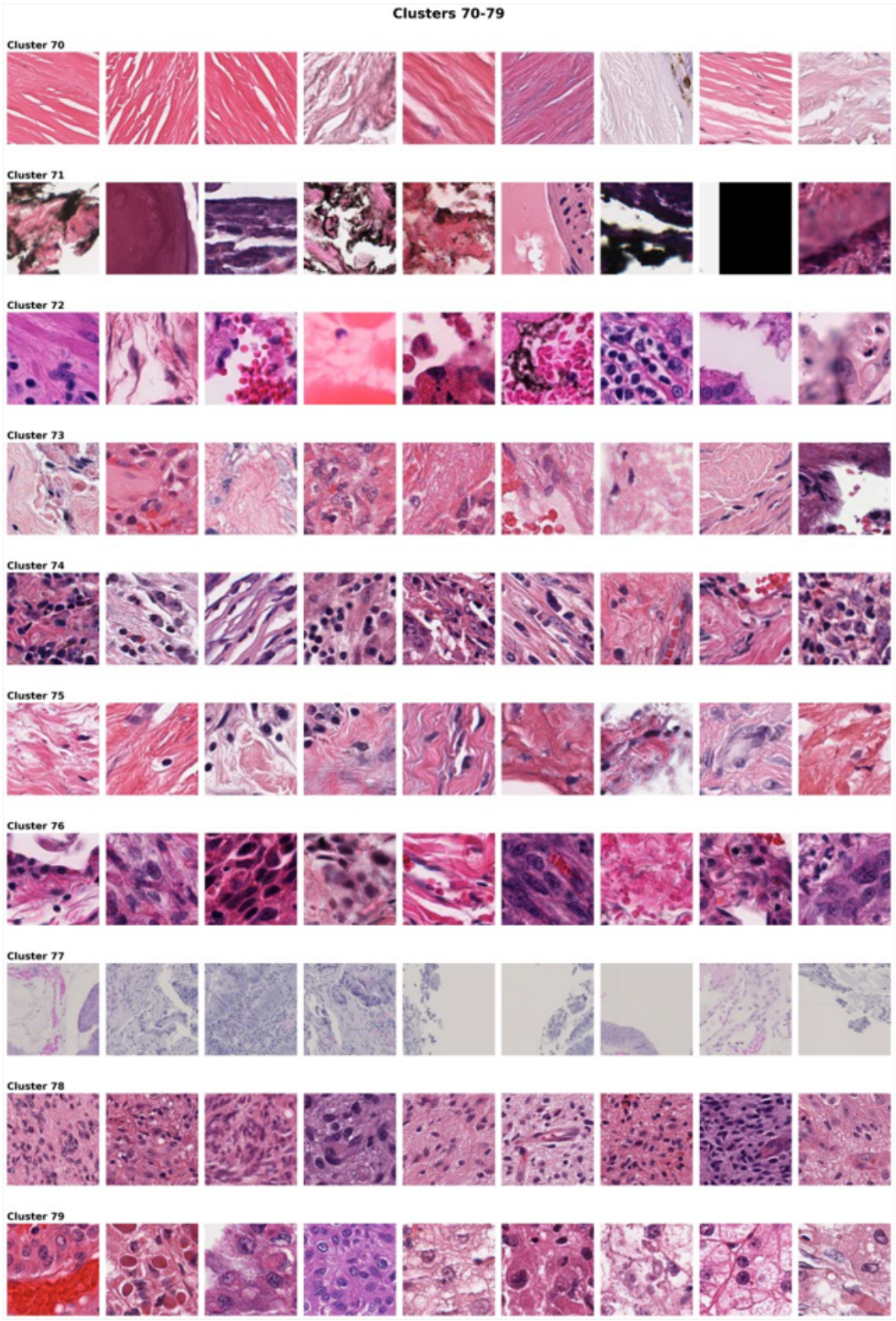


Supplementary Figure S15: Representative samples from TCGA tissue clusters 51 to 60, with 9 exemplar patches per cluster demonstrating morphological coherence within each identified tissue phenotype.

https://doi.org/10.52202/085713-4185

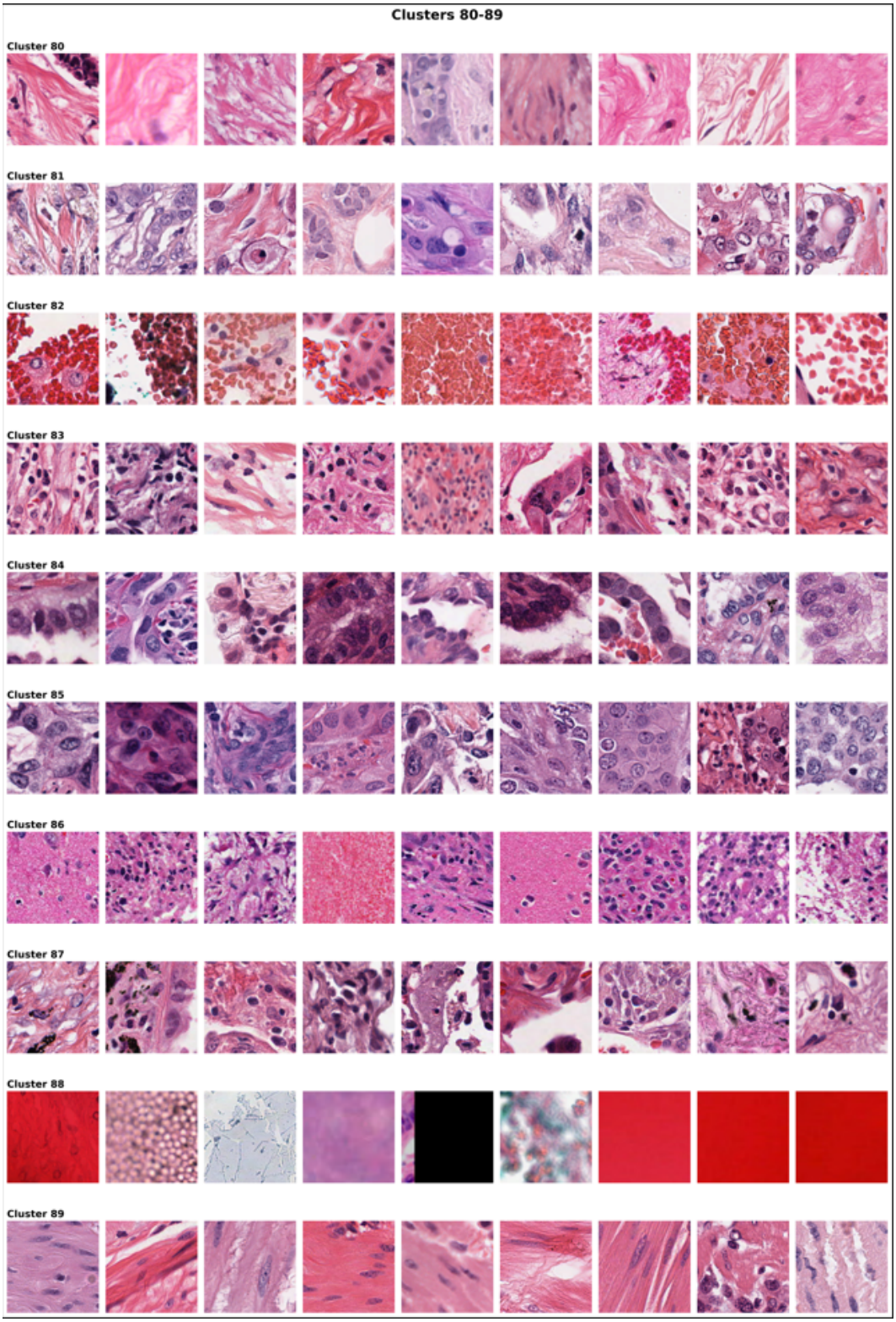


Supplementary Figure S16: Representative samples from TCGA tissue clusters 61 to 70, with 9 exemplar patches per cluster demonstrating morphological coherence within each identified tissue phenotype.

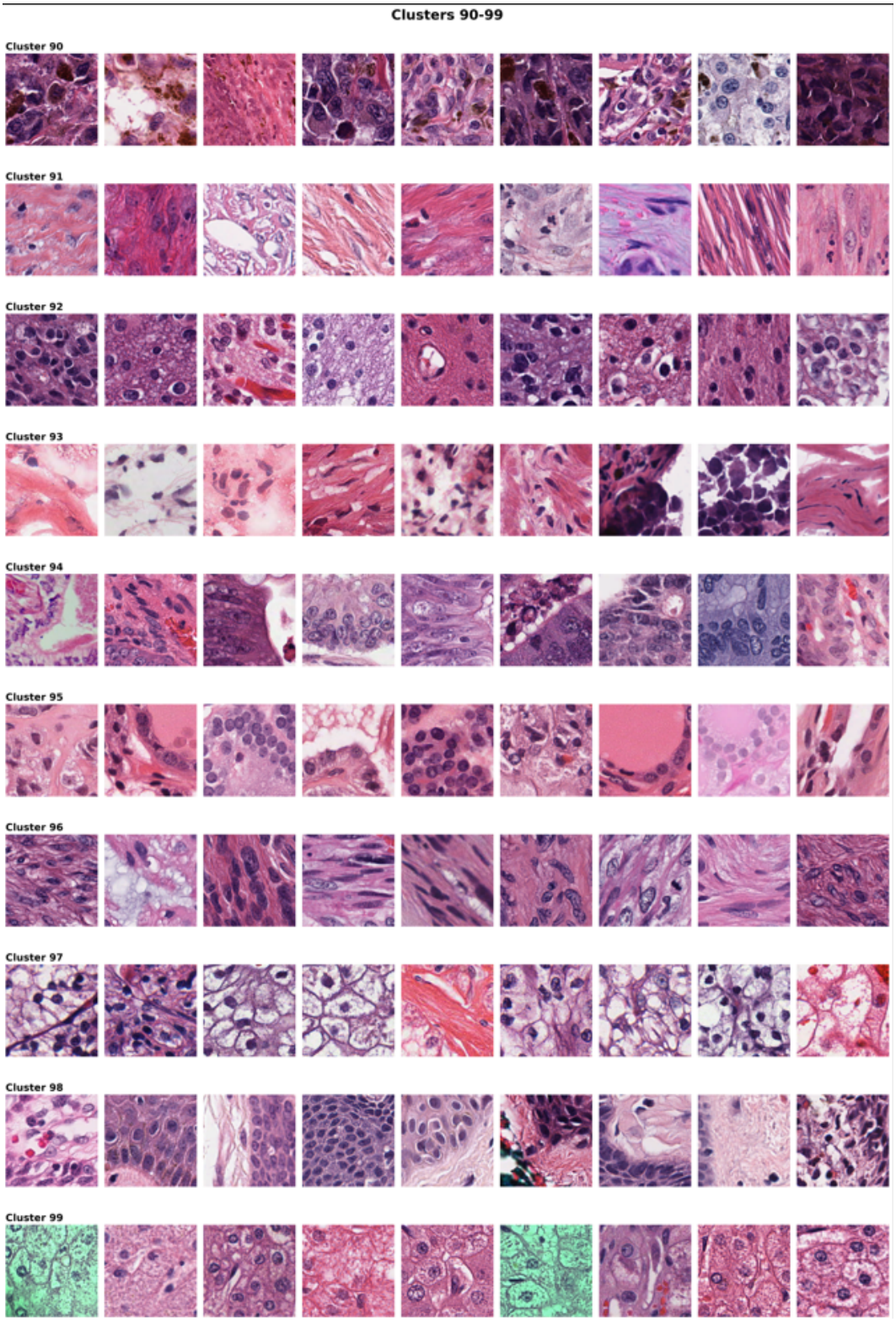


Supplementary Figure S17: Representative samples from TCGA tissue clusters 71 to 80, with 9 exemplar patches per cluster demonstrating morphological coherence within each identified tissue phenotype.

https://doi.org/10.52202/085713-4185



Supplementary Figure S18: Representative samples from TCGA tissue clusters 81 to 90, with 9 exemplar patches per cluster demonstrating morphological coherence within each identified tissue phenotype.



Supplementary Figure S19: Representative samples from TCGA tissue clusters 91 to 100, with 9 exemplar patches per cluster demonstrating morphological coherence within each identified tissue phenotype.

https://doi.org/10.52202/085713-4185

- 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: [NA] .

   Justification: The paper is primarily empirical and does not include formal theoretical results requiring proofs. The mathematical formulations provided are standard in the diffusion model literature and are used to describe the methodology rather than establish novel theoretical guarantees.

   Guidelines:

   - The answer NA 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 provides comprehensive details on the dual-conditioning diffusion architecture, the self-supervised clustering approach, and evaluation methodologies. Section 3 details the HeteroTissue-Diffuse framework with specific parameter ranges (e.g., crop sizes of 50-200 pixels), clustering parameters (100 tissue types), and the sampling strategies. The datasets used (Camelyon16, PANDA, TCGA) are publicly available with proper citations.

   Guidelines:

- The answer NA 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: The implementation code and the pre-trained models will be made available on GitHub upon publication. The paper utilizes publicly available datasets (Camelyon16, PANDA, TCGA), and the supplementary materials include detailed instructions for reproducing the experiments, including environment setup, data preprocessing, model training, and evaluation protocols.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.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://nips.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.

 https://doi.org/10.52202/085713-4185

- 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, etc.) necessary to understand the results?

   Answer: [Yes] .

   Justification: The paper specifies all relevant training and testing details including data splits, architecture configurations, optimization parameters, and evaluation metrics. Section 3 outlines the specific training parameters for both the annotated dataset approach and the self-supervised TCGA extension. The evaluation methodology in Section 4 details the metrics used (Fréchet Distance across multiple encoders, IoU for segmentation) and the pathologist evaluation protocol.

   Guidelines:

   - The answer NA 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: The paper reports statistical significance for the key experimental results. Table 1 presents comprehensive Fréchet Distance metrics across multiple foundation model encoders for different datasets and conditioning approaches. Figure 3 shows validation and test IoU comparisons with error bars representing standard deviation across multiple runs. The pathologist evaluation in Figure 4 includes average ratings with standard deviations across multiple samples.

   Guidelines:

   - The answer NA 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 computational resources used for training and inference. The diffusion models and the cluster classifier as well as extracting the TCGA embedding using the foundation model (UNI) were conducted on 4 NVIDIA A100 GPUs with 80GB memory for approximately 4 Months. The TCGA clustering required 48 hours on a 124-core CPU server with 1TGB RAM.

   Guidelines:

   - The answer NA 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 research conforms to the NeurIPS Code of Ethics. The paper addresses privacy concerns in medical data by developing synthetic data generation methods that can reduce reliance on sensitive patient information. The approach democratizes access to training data across institutions regardless of size or resources, promoting equity in medical AI development. All datasets used are publicly available research datasets accessed in accordance with their terms of use.

   Guidelines:

   - The answer NA 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 includes a dedicated broader impacts section discussing both positive and negative societal implications. Positive impacts include democratizing access to high-quality annotated histopathology data across institutions, accelerating AI diagnostics for rare cancer subtypes, and enabling international research collaboration without compromising patient privacy. Potential negative impacts, such as computational resource disparities, are acknowledged in the limitations section.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

https://doi.org/10.52202/085713-4185

- 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., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes] .

    Justification: The paper describes safeguards implemented for the responsible release of synthetic data. The generated synthetic datasets undergo pathologist review to ensure they don't contain artifacts that could lead to diagnostic errors if used for training. The code release includes documentation on appropriate use cases and limitations. The model includes filters to prevent generation of misleading or clinically implausible tissue patterns.

    Guidelines:

    - The answer NA 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: The paper properly cites and acknowledges the creators of existing assets used in the research. Camelyon16, PANDA, and TCGA datasets are cited with their respective licenses (all are publicly available for research purposes). The foundation models used for feature extraction (UNI, Virchow, etc.) are properly cited with their respective licenses and versions.

    Guidelines:

    - The answer NA 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 new assets introduced (synthetic datasets and trained models) are well documented with information on training procedures, limitations, and intended uses. The supplementary materials include datasheets for the synthetic datasets following standard templates, and model cards for the diffusion models detailing their performance characteristics, failure modes, and appropriate contexts for use.

    Guidelines:

    - The answer NA 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: [NA] .

    Justification: The paper does not involve any crowdsourcing tasks or direct research involving human subjects.

    Guidelines:

    - The answer NA 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?

https://doi.org/10.52202/085713-4185

Answer: [NA] .

Justification: The research does not involve direct interaction with human subjects.

Guidelines:

- The answer NA 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 rigorousness, or originality of the research, declaration is not required.

Answer: [NA] .

Justification: LLMs were not used as a component of the core methodology, nor did they contribute to any scientific or technical innovations presented in the paper.

Guidelines:

- The answer NA 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 (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.