FIRM: MORPHOLOGY-GUIDED RESIDUAL EXPERTS FOR FINE-GRAINED PATHOLOGY IMAGE SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-grained pathology image synthesis requires modeling both diagnostic distinctions across categories and morphological diversity within each category. However, conventional category conditioning specifies the target diagnosis while leaving within-class variation largely implicit. We introduce Feature-Informed Residual Modulation (FIRM), a framework that structures pathology representations into inter-class diagnostic variation and intra-class morphology for conditional generator adaptation. FIRM first learns a shared pathology prior and constructs a diagnostic and morphological geometry from fixed training-image features, decomposing within-class morphology into discrete modes and continuous residuals. These structured conditions modulate class-routed low-rank residual experts, allowing each diagnostic category to specialize the shared generator while preserving fine-grained morphological variation. We further introduce paired pathology feature refinement, which aligns predicted clean images with their corresponding real training crops in feature space without additional sampling trajectories. Experiments on BRACS and SICAPv2 show consistent improvements in class-conditional distribution matching and feature-space coverage over existing generative baselines. Controlled interventions further demonstrate discrete and continuous morphology manipulation, while synthetic augmentation improves downstream classification.

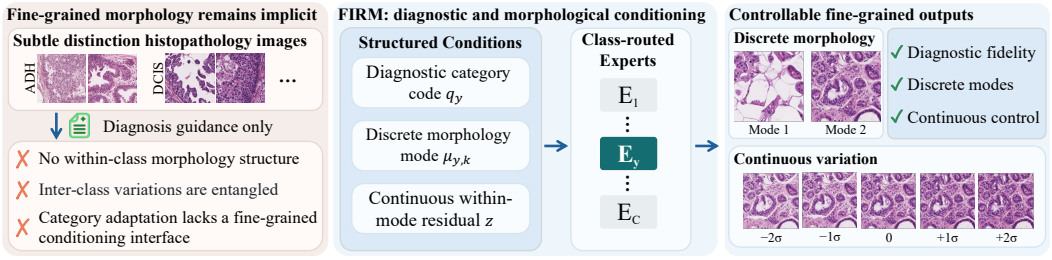


Figure 1: Overview of FIRM. Structured diagnostic codes, discrete morphology modes, and continuous residuals condition class-routed experts for fine-grained, controllable pathology synthesis.

1 INTRODUCTION

Pathology examines cellular morphology and tissue architecture to characterize diseases such as cancer, inflammatory lesions, and infectious diseases (Chiche & Adam, 2013; Neal et al., 2014; Perez et al., 2014). The digitization of tissue slides has enabled computational analysis of these visual patterns, and deep learning has supported tasks including tumor detection and grading (Tran et al., 2018; Wang et al., 2022b; Chen et al., 2024). These tasks involve recognizing diagnostic features across a wide range of tissue appearances, making representative training data essential for learning-based analysis.

Building such datasets is expensive because pathology supervision often depends on expert review, while exhaustive annotation at whole-slide scale is difficult to obtain (Campanella et al., 2019). Data

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

scarcity is particularly problematic for uncommon or diagnostically ambiguous lesions, and subtype recognition further requires sufficient coverage of heterogeneous appearances within each category (Vorontsov et al., 2024). Synthetic augmentation has therefore been explored to expand limited histopathology datasets, with previous studies showing improved classification when carefully selected synthetic samples are incorporated into training (Xue et al., 2021). This motivates generators that can provide additional category-specific examples while preserving meaningful within-class morphological diversity.

Meeting this need requires modeling both diagnostic differences between categories and morphological variation within each category. Breast histopathology is a representative example: proliferative lesions can occupy a subtle morphological continuum, and distinctions such as atypical ductal hyperplasia (ADH) versus ductal carcinoma in situ (DCIS) may depend on small architectural and cytological differences. Studies of atypical breast lesions report substantial case-dependent diagnostic disagreement, particularly for challenging ADH morphologies and borderline ADH/DCIS cases (Allison et al., 2016; Tozbikian et al., 2017). We therefore focus on fine-grained histopathology synthesis, where each diagnostic category should retain its identity while faithfully representing its internal morphological diversity.

Recent developments in generative models make this goal increasingly tractable. Modern diffusion and flow-based generators provide strong priors for high-fidelity image synthesis (Esser et al., 2024), while self-supervised pathology encoders learn feature spaces that capture diagnostically relevant tissue organization across spatial scales (Wang et al., 2022a; Chen et al., 2022). However, neither a category label nor an unstructured image embedding explicitly separates directions that distinguish diagnostic categories from those describing morphology within a category. This motivates using pathology representations not merely as conditioning signals, but as a structured geometry that decomposes inter- and intra-class variation and uses them to control category-specific generator adaptation.

As Figure 1 illustrates, we introduce Feature-Informed Residual Modulation (FIRM), which specializes the histopathology prior generator through class-routed residual experts. It supports both inter-class conditions and intra-class morphology control. Our main contributions are summarized as follows:

- We construct a *Diagnostic and Morphological Geometry* that derives category codes and within-class morphology coordinates from fixed training-image features, then organizes the morphology coordinates into discrete modes and continuous within-mode residuals.
- We develop *Morphology-Conditioned Residual Experts* that modulate low-rank updates to the frozen shared generator through category, centered mode, and residual gates.
- We introduce *Paired Pathology Feature Refinement* that continues flow-matching training with an additional loss between frozen-encoder features of decoded clean-image estimates and their source training crops, without additional sampling trajectories.
- Extensive experiments demonstrate improved generation quality and feature-space coverage over competing methods. Controlled interventions demonstrate discrete and continuous morphology control.

2 Related Work

Representation-guided pathology image synthesis. Early pathology generators model tissue appearance and morphological variation through structured latent spaces or diffusion models (Quiros et al., 2020; Moghadam et al., 2023). Subsequent approaches improve synthesis scale and controllability through coarse-to-fine whole-slide generation (Harb et al., 2024), report conditioning (Yellapragada et al., 2024), unpaired text–mask conditioning (Bhosale et al., 2025), and topology-guided cellular control (Li et al., 2025). More recently, learned pathology representations have become a direct conditioning interface: Graikos et al. (2024) condition diffusion models on self-supervised image embeddings, PixCell scales representation-guided generation to a pathology foundation model (Yellapragada et al., 2025). These works demonstrate the value of semantic and structural representations for pathology synthesis. FIRM instead explicitly organizes the feature geometry into diagnostic coordinates and class-relative morphology, represented by discrete modes and continuous residuals, to modulate category-specific residual experts.

Low-rank adaptation and conditional experts. LoRA (Hu et al., 2022) adapts pretrained models through trainable low-rank weight updates, while AdaLoRA (Zhang et al., 2023) further allocates low-rank capacity according to parameter importance. Sparse mixtures of experts use conditional routing to specialize subsets of model parameters (Shazeer et al., 2017), and recent approaches combine this principle with low-rank adaptation. MoLE (Wu et al., 2023), for example, learns gated compositions of multiple LoRA experts, while Diff-MoE (Cheng et al., 2025) introduces expert routing into diffusion transformers for timestep- and spatially adaptive specialization. FIRM uses a related parameter-efficient principle but separates expert selection from within-expert modulation: the diagnostic category selects the residual expert, whereas morphology modes and continuous residuals modulate its low-rank correction.

3 METHOD

As illustrated in Figure 2, given training images $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ with diagnostic categories $y_i \in \{1, \ldots, C\}$, our goal is to model each class-conditional image distribution while preserving its intra-class morphological diversity. FIRM first learns a shared pathology prior, then we introduce three components for fine-grained conditional generation. First, a *diagnostic and morphological geometry* decomposes pathology features into conditions. Second, these structured conditions modulate *morphology-conditioned residual experts*, where the requested category selects an expert that specializes the frozen shared prior. Finally, *paired pathology feature refinement* further aligns predicted clean images with their corresponding real training crops.

3.1 SHARED PATHOLOGY PRIOR

We learn a shared pathology prior by adapting a pretrained generator with LoRA (Hu et al., 2022) on training images from all diagnostic categories. Every image uses the same generic pathology prompt c_0. Let h_0 denote the normalized, packed latent encoding of a training image and $\epsilon \sim \mathcal{N}(0, I)$ denote Gaussian noise. Following flow matching (Lipman et al., 2023), we define the linear interpolation path and its target velocity as:

$$h_\sigma = (1 - \sigma)h_0 + \sigma\epsilon, \qquad v^\star = \epsilon - h_0, \qquad \sigma \in [0, 1]. \tag{1}$$

where σ specifies the position along the path: h_σ equals the clean latent at $\sigma = 0$ and the sampled noise at $\sigma = 1$. For each pair (h_0, ϵ), the target velocity $v^\star = \partial h_\sigma / \partial \sigma$ is constant along this path and provides the supervision for learning the generator's velocity field.

We optimize θ_0 using the flow-matching objective, written as:

$$\mathcal{L}_{\mathrm{FM}} = \frac{1}{M} \left\| v_{\theta_0}(h_\sigma, \sigma, c_0) - v^\star \right\|_2^2, \tag{2}$$

where θ_0 denotes the trainable shared adaptation parameters. The velocity predictor v_{θ_0} takes the interpolated latent h_σ, flow time σ, and prompt c_0 as inputs. M is the number of latent entries. Training minimizes its expectation over images, noise, and sampled flow times. Training across categories allows the shared adaptation to capture common patterns of tissue texture and histological organization, resulting in a generator G_0 that serves as the shared pathology prior. In subsequent stages, both the generative backbone and shared adaptation θ_0 remain frozen, while morphology-conditioned residual experts introduce category-local corrections.

3.2 DIAGNOSTIC AND MORPHOLOGICAL GEOMETRY

Let $u_i \in \mathbb{R}^d$ denote a fixed, normalized pathology feature of training image x_i. We construct the diagnostic and morphological geometry from these features to represent differences between diagnostic categories and morphological variation within each category.

Category and morphology coordinates. Let m_y denote the mean feature of category y, computed with equal weight for each slide, and let $\bar{m} = C^{-1} \sum_y m_y$ denote the mean across categories. We apply a regularized within-class whitening transform W to account for feature scales and correlations. In the whitened space, we form a weighted scatter matrix from differences between

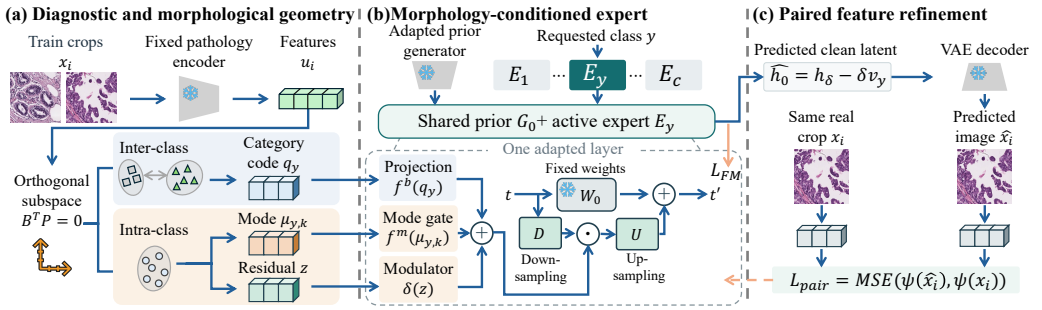


Figure 2: Overview of FIRM pipeline. (a) Diagnostic and morphological geometry constructed from orthogonal pathology subspace, (b) morphology-conditioned residual experts for each category, and (c) paired pathology feature refinement.

neighboring category means, with neighborhood relationships determined by feature similarity and classification confusion. Its leading eigenvectors define an orthonormal category basis B.

To describe within-class morphology, we center each image feature at its own category mean and project the whitened deviations onto the orthogonal complement of B. The leading directions of the remaining variation form a morphology basis P. The category code q_y and image-specific morphology coordinates s_i are then defined as:

$$q_y = B^\top W(m_y - \bar{m}), \qquad s_i = P^\top W(u_i - m_{y_i}), \qquad B^\top P = 0. \tag{3}$$

Here, q_y locates category y relative to the common reference $\bar{m}$ and is shared by all images in that category, while s_i describes the position of image x_i relative to its own category mean in the morphology subspace.

Discrete modes and continuous residuals. We organize these morphology coordinates into discrete modes by applying K-means separately within each category. This assigns image x_i to a mode $k_i \in \{1, \ldots, K_{y_i}\}$ with center μ_{y_i,k_i}. Let $n_{y,k}$ denote the number of training images assigned to mode k in category y, and let $n_y = \sum_k n_{y,k}$. We determine K_y by decreasing K from 10 until every cluster contains at least $\max(100, \lceil 0.05 n_y \rceil)$ images. The mode frequencies define a smoothed sampling prior $\pi_{y,k} = (n_{y,k} + 1)/(n_y + K_y)$. To retain variation among images assigned to the same mode, we use each image's displacement from its mode center. We define this residual r_i and its per-coordinate root-mean-square scale $\rho_{y,j}$ as:

$$r_i = s_i - \mu_{y_i,k_i}, \qquad \rho_{y,j} = \left(\frac{1}{n_y} \sum_{i:y_i=y} r_{i,j}^2 \right)^{1/2}. \tag{4}$$

Here, j indexes morphology coordinates. Element-wise division gives the scaled residual $\tilde{r}_i = r_i \oslash \rho_{y_i}$, placing its coordinates on comparable scales. We then obtain the continuous condition by limiting its norm to R:

$$z_i = \frac{\tilde{r}_i}{\max\{1, \|\tilde{r}_i\|_2 / R\}}. \tag{5}$$

The fixed R leaves scaled residuals with norm at most R unchanged and rescales larger ones to norm R while preserving their direction. The category code q_{y_i}, mode center μ_{y_i,k_i}, and residual condition z_i modulate the expert gates described next. The feature representation and all geometric quantities remain fixed throughout expert training.

3.3 Morphology-Conditioned Residual Experts

We use the diagnostic and morphological geometry from Section 3.2 to condition residual experts on the shared pathology prior G_0. Each category y has an expert E_y with trainable parameters Θ_y. For category-y examples, only E_y is active; training updates Θ_y while keeping G_0 frozen. For input

t at an adapted layer, the expert adds a gated rank-r correction to the frozen shared transformation W^0. Omitting layer and expert indices, we write:

$$t' = W^0 t + U\big[(Dt) \odot a(k, z)\big], \tag{6}$$

where D and U are trainable down- and up-projections shared across modes, $a(k, z) \in \mathbb{R}^r$ is the gate conditioned on mode k and residual z, and $\odot$ denotes elementwise multiplication. We initialize $U = 0$, so the expert initially preserves G_0.

To construct this gate, a trainable network first maps the fixed category code to $b = f^b(q_y)$, providing a component shared by all images in category y. A second network maps each mode center to a raw mode gate $g_k = f^m(\mu_{y,k})$. We express each mode's contribution relative to the common category component by subtracting the average mode gate under the sampling prior π_y:

$$\widetilde{g}_k = g_k - \sum_{s=1}^{K_y} \pi_{y,s} g_s. \tag{7}$$

The centered gates satisfy $\sum_k \pi_{y,k} \widetilde{g}_k = 0$, so $b + \widetilde{g}_k$ combines a category component with a mode-specific deviation. This gate is shared by images assigned to the same mode. To incorporate variation within that mode, the residual gate modulator maps the continuous condition z to an additional adjustment:

$$\delta(z) = \alpha \tanh(Az). \tag{8}$$

Here, A is a trainable linear projection with no bias initialized to 0, and $\alpha > 0$ bounds each entry of $\delta(z)$ in magnitude. Since z describes the scaled displacement from the mode center, this adjustment varies continuously within the mode and vanishes at its center, where $z = 0$. The category, mode, and residual components then form the complete gate $a(k, z) = b + \widetilde{g}_k + \delta(z)$ for equation 6.

3.4 Paired Pathology Feature Refinement

We first train the morphology-conditioned residual experts with flow matching to learn the conditional velocity field. We then continue training with paired pathology feature supervision, comparing each predicted clean image with its corresponding real pathology crop in feature space. Expert training replaces c_0 with the category-specific prompt c_y.

For training image x_i, let v_y denote the expert's velocity field, conditioned on the category prompt c_y, assigned mode k_i, and residual z_i. Rearranging the flow path in equation 1 gives a clean-latent estimate from the current velocity prediction:

$$\widehat{h}_0 = h_\sigma - \sigma v_y(h_\sigma, \sigma, c_y, k_i, z_i). \tag{9}$$

We reverse latent normalization and packing through $\mathcal{T}^{-1}$ and apply the frozen VAE decoder to obtain $\widehat{x}_i = \mathrm{Dec}(\mathcal{T}^{-1}(\widehat{h}_0))$. This reuses the velocity prediction already computed for flow matching and requires no additional sampling trajectory.

Let ψ denote a frozen image-feature encoder, including its input preprocessing, with output dimension d_ψ. We compare the features of $\widehat{x}_i$ with those of the same training crop x_i used to construct h_σ:

$$\mathcal{L}_{\mathrm{pair}}(\widehat{x}_i, x_i) = \frac{1}{d_\psi} \left\| \psi(\widehat{x}_i) - \psi(x_i) \right\|_2^2 . \tag{10}$$

The real-image features $\psi(x_i)$ are treated as fixed targets, with no gradient computed through the real-image branch. Both images undergo the same preprocessing, and the loss uses the feature vectors without output normalization. Although the decoder and feature encoder remain frozen, gradients pass through the predicted-image branch to update the expert projections, gate networks, and residual modulator.

To control decoding cost, each optimizer update selects one image, indexed by j, from its effective batch for feature supervision. When the selected image's flow time falls within the prescribed low-noise range, we add its paired loss from equation 10 to the flow-matching objective:

$$\mathcal{L} = \mathcal{L}_{\mathrm{FM}} + \lambda_y \mathcal{L}_{\mathrm{pair}}(\widehat{x}_j, x_j), \tag{11}$$

where $\mathcal{L}_{\mathrm{FM}}$ is the mean flow-matching loss over the effective batch. Thus, flow matching supervises every image in the batch, while the additional feature term supervises only the selected pair $(\widehat{x}_j, x_j)$.

Table 1: Cross-model comparison on histopathology image synthesis on two datasets. Results are macro-averages over classes. Bold indicates the best score within each dataset.

	BRACS					SICAPv2				
	FID ↓	KID ↓	Cov. ↑	Path-F ↓	Path-R ↑	FID ↓	KID ↓	Cov. ↑	Path-F ↓	Path-R ↑
FLUX.2	47.12	19.02	25.12	584.41	35.55	45.39	30.86	41.81	228.61	72.70
LDM	76.43	54.01	21.28	983.62	0.10	95.59	70.71	10.20	1035.72	0.09
PathLDM	47.54	22.88	33.58	598.35	5.82	42.38	27.61	41.68	421.32	12.50
PathUP	43.32	17.79	30.00	504.52	10.13	43.39	26.51	35.06	452.93	8.61
PixCell	64.52	37.84	13.47	1582.13	4.76	70.83	53.02	10.59	806.29	2.53
FIRM	**31.45**	**6.40**	**37.01**	**335.18**	**36.34**	**27.07**	**9.12**	**57.05**	**101.21**	**73.77**

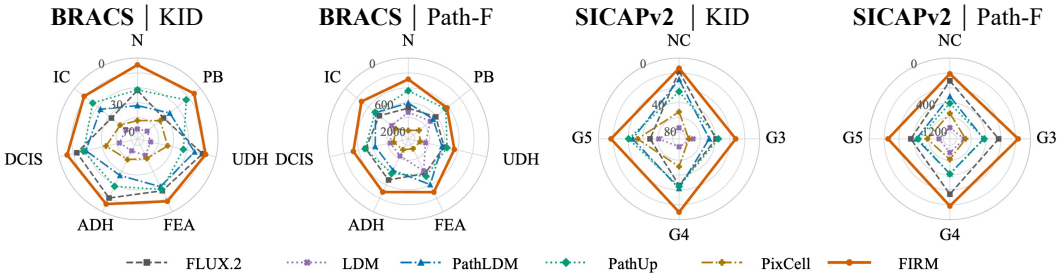


Figure 3: Per-class KID and Path-F on BRACS and SICAPv2. Metrics are normalized such that farther outward indicates better performance.

If the selected image's flow time is outside the prescribed range, the update uses $\mathcal{L} = \mathcal{L}_{\mathrm{FM}}$. We calibrate λ_y for each category using the relative gradient magnitudes of the two losses on training data, then hold it fixed during refinement.

Inference. For requested category y, we activate E_y and sample $k \sim \mathrm{Cat}(\pi_y)$. We then draw z uniformly from the stored residual conditions of training images assigned to mode k. The conditions (k, z) remain fixed throughout the generation trajectory. Inference requires neither the geometry encoder nor the feature-refinement branch.

4 Experiments

4.1 Datasets, Metrics and Experimental Details

We evaluate FIRM on two datasets covering breast and prostate histopathology. **BRACS** (Brancati et al., 2022) comprises seven diagnostic categories: normal (N), pathological benign (PB), usual ductal hyperplasia (UDH), flat epithelial atypia (FEA), atypical ductal hyperplasia (ADH), ductal carcinoma in situ (DCIS), and invasive carcinoma (IC). We use 30,413 training, 2,925 validation, and 5,300 test crops. **SICAPv2** (Silva-Rodríguez et al., 2020) contains prostate biopsy images labeled as non-cancerous tissue (NC) or Gleason patterns 3, 4, and 5 (G3, G4, G5), with 7,472 training, 2,487 validation, and 2,122 test images. All source patches are 512×512, acquired at $20\times$ magnification for BRACS and $10\times$ for SICAPv2. Training, validation, and test partitions are patient-disjoint.

We evaluate generation in both general-purpose and pathology-specific feature spaces. FID (Heusel et al., 2017), KID (Bińkowski et al., 2018), and Coverage (Cov.) (Naeem et al., 2020) are computed in Inception space, where Cov. measures the fraction of real samples covered by generated samples using the fifth-nearest real neighbor to define neighborhood radii. We further report pathology-feature Fr'echet distance (Path-F) and recall (Path-R) using a frozen Virchow encoder (Vorontsov et al., 2024). Path-F measures distributional discrepancy, while Path-R measures the fraction of real features covered by generated-feature neighborhoods defined by the fifth-nearest generated neighbor. Lower FID, KID, and Path-F, and higher Cov. and Path-R indicate better performance. Cov. and Path-R are reported in percentages, and KID is scaled by 10^3.

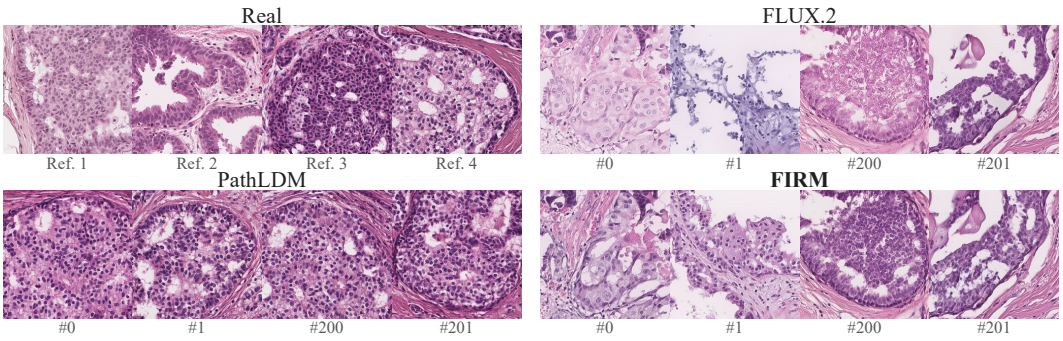


Figure 4: Qualitative comparison on BRACS ADH between real images and multiple pathology image synthesis methods, showing local structure and within-class morphological diversity.

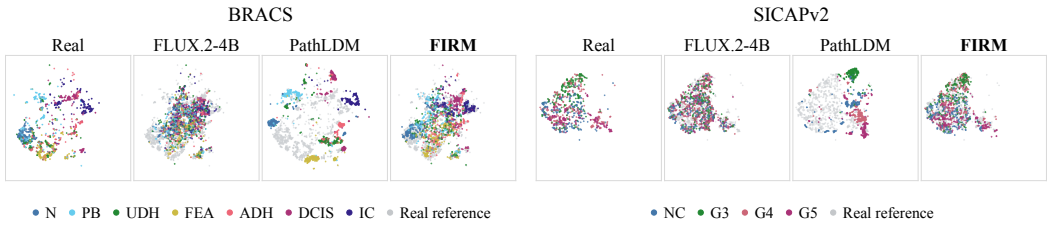


Figure 5: Class-conditional feature distributions on BRACS and SICAPv2. Virchow features are embedded in a shared t-SNE space. Gray points denote real-reference distribution for comparison.

FIRM is initialized from FLUX.2-klein-4B (Labs, 2025) and learns a shared pathology prior using training images across all categories. We then freeze this prior and train class-routed rank-32 experts with an effective batch size of eight. The standard expert schedule consists of 1,000 flow-matching updates followed by 1,000 paired feature-refinement updates. Frozen UNI2-h features (Chen et al., 2024) provide the representations for inter-class and intra-class geometry; implementation details are provided in Appendix A. We use AdamW with learning rates of 10^{-4} for the shared prior and 5×10^{-5} for expert parameters, weight decay 10^{-4}, and 200 warmup updates. FIRM generates 512×512 images with 28 sampling steps.

4.2 Inter-Model Comparison

We compare FIRM with general-purpose and pathology-specific generators: FLUX.2(4B) (Labs, 2025), LDM (Rombach et al., 2022), PathLDM (Yellapragada et al., 2024), PathUp (Li et al., 2024), and PixCell (Yellapragada et al., 2025). The FLUX.2 baseline adapts FLUX.2-klein-4B with a single class-conditional LoRA trained jointly across categories. We evaluate 1,000 generated images per class against entire real image pool and report macro-averaged scores. All image-derived conditions use training set only. Evaluation resizes 512-pixel outputs without cropping into 256×256.

FIRM improves distribution matching and coverage. FIRM achieves the best scores on all five metrics across both datasets (Table 1). On BRACS, FIRM achieves FID, KID, and Path-F scores of 31.45, 6.40, and 335.18, respectively. On SICAPv2, the reductions of FID, KID, and Path-F are 36.1%, 65.6%, and 55.7% relative to the strongest baseline for each metric. Cov. increases from 33.58% to 37.01% on BRACS and from 41.81% to 57.05% on SICAPv2, while Path-R improves from 35.55% to 36.34% and from 72.70% to 73.77%, respectively. These gains span both general-purpose and pathology-specific feature spaces.

FIRM consistently improves class-wise generation. Figure 3 reveals substantial variation across categories for competing methods. FIRM achieves the lowest KID and Path-F across all seven BRACS and four SICAPv2 categories. Competing methods exhibit uneven class-wise performance: FLUX.2, for example, obtains relatively low KID on NC but substantially higher values on G3 and

Table 2: Component ablations on BRACS, using 1,000 samples per class. Bold indicates the best macro-average.

Variant	FID ↓	KID ↓	Cov. ↑	Path-F ↓	Path-R ↑
Prior only	63.15	31.14	16.95	761.61	34.98
+ RE	36.74	12.03	31.93	383.02	35.41
+ DMG	33.16	8.19	35.01	340.06	36.27
+ PPFR	**31.45**	**6.40**	**37.01**	**335.18**	**36.34**

RE: Residual Expert; DMG:Diagnostic and Morphological Geometry; PPFR:Paired Pathology Feature Refinement

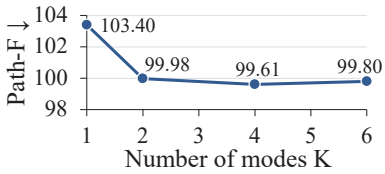


Figure 6: Sensitivity to the number of morphology modes K on SICAPv2.

Table 3: Ablation of diagnostic and morphological geometry on BRACS. We report per-class Path-F (↓) after removing conditioning methods, compared with the full FIRM model.

Variant	N	PB	UDH	FEA	ADH	DCIS	IC
w/o discrete mode gate	329.36	447.74	459.80	306.15	317.95	330.85	288.95
w/o continuous residual	316.98	436.14	457.49	307.91	311.62	336.28	285.35
FIRM	**284.88**	**418.91**	**452.75**	**292.63**	**293.08**	**326.05**	**277.90**

G5. FIRM's macro-level advantage therefore reflects improvements across categories rather than gains concentrated in a few favorable classes.

FIRM combines coherent morphology with broader feature coverage. In Figure 4, FIRM exhibits clear cellular detail across different tissue arrangements at sample indices, whereas FLUX.2 shows fragmented structures and PathLDM repeatedly produces similar rounded, densely packed layouts. In the shared Virchow-feature projection in Figure 5, we observe broader overlap between FIRM samples and real class-specific regions on both datasets. In contrast, PathLDM forms restricted clusters, while FLUX.2 concentrates in central regions, particularly on BRACS. These observations are consistent with FIRM's lower Path-F, higher coverage and recall in Table 1.

4.3 Ablation Studies

Component and geometry ablations. Table 2 progressively adds Residual Experts (RE), Diagnostic and Morphological Geometry (DMG), and Paired Pathology Feature Refinement (PPFR) to the shared prior. RE provides the largest gain, reducing FID from 63.15 to 36.74 and Path-F from 761.61 to 383.02. DMG further improves all metrics, and PPFR yields the best overall performance. Table 3 further isolates the two intra-class components of DMG. Removing either the discrete mode gate or continuous residual consistently increases Path-F across all seven BRACS categories, while the full model performs best for every class. These results support both category-specific adaptation and the complementary roles of discrete and continuous morphology conditioning.

Number of morphology modes. Figure 6 evaluates the number of modes K on SICAPv2. All multi-mode settings improve mean Path-F over $K = 1$ (103.40), with $K = 4$ achieving the lowest value of 99.61; $K = 2$ and $K = 6$ yield similar results of 99.98 and 99.80. This suggests that multiple morphology centers are beneficial, while performance is relatively insensitive to K once multiple modes are introduced.

Controlled morphological variation. Figure 7 probes the learned conditions under fixed generation noise. Switching the discrete mode with $z = 0$ changes tissue organization, while continuous control sweeps a mode-specific principal residual direction from -2 to $+2$ training SDs and measures the output shift along the same normalized morphology coordinate. Across all modes and 16 noise seeds, the mean signed Spearman correlations are 0.265, 0.575, and 0.804 for PB, UDH, and IC, respectively. Detailed definitions and evaluation protocols are provided in Appendix.

4.4 Downstream Task Evaluation

Table 4 evaluates synthetic augmentation on SICAPv2 using test F1 over three runs. Adding FIRM samples improves macro-F1 from 63.39 ± 2.18 with real-only training to 67.43 ± 1.21, exceeding

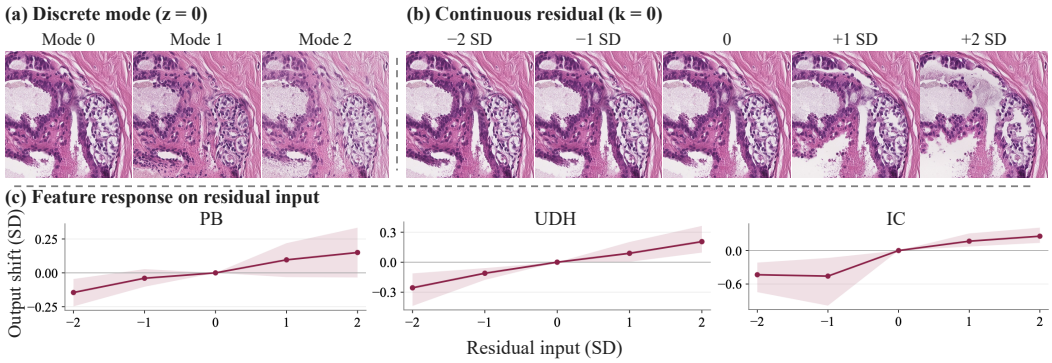


Figure 7: Controlled morphology interventions under fixed generation noise. (a) Switching the discrete mode at $z = 0$ changes local morphology. (b) Sweeping a mode-specific residual direction from -2 to $+2$ training SDs induces gradual within-mode variation. (c) Mean signed output shift along the corresponding normalized morphology coordinate over all modes and 16 noise seeds; shading denotes 95% bootstrap confidence intervals.

Table 4: Downstream ResNet50 classification on the SICAPv2 test set with synthetic augmentation. mean $\pm$ standard deviation F1 scores (%) are reported. Bold denotes the best mean in each column.

Training data	Average	NC	G3	G4	G5
Real only	63.39 ± 2.18	89.71 ± 0.86	56.79 ± 0.52	67.73 ± 1.16	39.31 ± 9.97
Real + repeated real	63.73 ± 0.98	90.27 ± 0.36	57.22 ± 1.08	69.44 ± 0.85	37.97 ± 4.60
Real + FLUX.2	65.38 ± 1.47	$\mathbf{91.07 \pm 0.38}$	57.29 ± 1.18	70.14 ± 0.80	43.01 ± 5.39
Real + FIRM	$\mathbf{67.43 \pm 1.21}$	90.95 ± 0.59	$\mathbf{57.58 \pm 0.08}$	$\mathbf{73.28 \pm 1.88}$	$\mathbf{47.91 \pm 4.74}$

repeated-real augmentation (63.73 ± 0.98) and FLUX.2 augmentation (65.38 ± 1.47). FIRM achieves the best F1 on G3, G4, and G5, with the largest gain over real-only training on the challenging G5 class ($39.31 \rightarrow 47.91$), while FLUX.2 performs slightly better on NC. These results indicate that FIRM-generated images provide useful complementary variation for downstream classification.

4.5 Limitations

FIRM currently focuses on patch-level histopathology synthesis under discrete diagnostic categories, which limits its direct use in settings that require broader spatial context. Future work may extend the framework to larger spatial scales, including region-level and whole-slide generation, while preserving fine-grained morphological controllability. In addition, incorporating complementary conditioning signals, such as textual descriptions, spatial annotations, or other clinical metadata, may further broaden its applicability. It would also be interesting to explore richer morphology representations and more flexible user-specified controls beyond the discrete modes and continuous residuals considered in this work, enabling more intuitive and diverse forms of pathology image synthesis.

4.6 Conclusion

We presented FIRM, a framework for fine-grained pathology image synthesis that structures diagnostic and morphological variation within a shared generative prior. FIRM derives diagnostic coordinates, discrete morphology modes, and continuous within-mode residuals from pathology representations, and uses them to modulate class-routed residual experts. Paired pathology feature refinement further improves feature-space alignment without additional sampling trajectories. Experiments on BRACS and SICAPv2 show consistent gains in distribution matching and coverage, while controlled interventions demonstrate both discrete and continuous morphology manipulation. Synthetic augmentation further improves downstream classification, highlighting the practical value of explicitly modeling inter- and intra-class variation for controllable pathology generation.

AI USE STATEMENT

We used generative AI tools to assist with language polishing, grammatical correction, and improving the clarity and conciseness of the manuscript. AI tools were also used to suggest alternative phrasings and improve the organization of selected passages. They were not used to generate experimental results or make scientific conclusions. All AI-assisted content was manually reviewed and revised by the authors for technical accuracy, consistency, and appropriate attribution. The authors take full responsibility for the final content of the manuscript.

REPRODUCIBILITY STATEMENT

We provide the model architecture, training objectives, optimization settings, dataset splits, and evaluation protocols in the main paper and appendices. Additional implementation details and hyperparameter settings are included in the supplementary material. To facilitate reproducibility and future research, we will publicly release the complete source code, training and evaluation scripts, and configuration files upon publication.

ACKNOWLEDGMENTS

The authors thank all colleagues and collaborators who provided valuable discussions, feedback, and support throughout this work.

REFERENCES

Kimberly H. Allison, Mara H. Rendi, Sue Peacock, Tom Morgan, Joann G. Elmore, and Donald L. Weaver. Histological features associated with diagnostic agreement in atypical ductal hyperplasia of the breast: illustrative cases from the b-path study. *Histopathology*, 69(6):1028–1046, 2016. doi: 10.1111/his.13035.

Mahesh Bhosale, Abdul Wasi, Yuanhao Zhai, Yunjie Tian, Samuel Border, Nan Xi, Pinaki Sarder, Junsong Yuan, David Doermann, and Xuan Gong. PathDiff: Histopathology image synthesis with unpaired text and mask conditions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22415–22424, 2025.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosuè Scognamiglio, Guillaume Jaume, Giuseppe De Pietro, Maurizio Di Bonito, Antonio Foncubierta, Gerardo Botti, Maria Gabrani, Florinda Feroce, and Maria Frucci. BRACS: A dataset for BReAst carcinoma subtyping in H&E histology images. *Database*, 2022:baac093, 2022. doi: 10.1093/database/baac093.

Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25: 1301–1309, 2019. doi: 10.1038/s41591-019-0508-1.

Richard J. Chen, Chengkuan Chen, Yicong Li, Tiffany Y. Chen, Andrew D. Trister, Rahul G. Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16144–16155, 2022.

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024. doi: 10.1038/s41591-024-02857-3.

Kun Cheng, Xiao He, Lei Yu, Zhijun Tu, Mingrui Zhu, Nannan Wang, Xinbo Gao, and Jie Hu. Diff-moe: Diffusion transformer with time-aware and space-adaptive experts. In *Forty-second International Conference on Machine Learning*, 2025.

Laurence Chiche and Jean-Philippe Adam. Diagnosis and management of benign liver tumors. In *Seminars in liver disease*, volume 33, pp. 236–247. Thieme Medical Publishers, 2013.

Patrick Esser, Sumith Kulal, Andreas Blattmann, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 12606–12633, 2024.

Alexandros Graikos, Srikar Yellapragada, Minh-Quan Le, Saarthak Kapse, Prateek Prasanna, Joel Saltz, and Dimitris Samaras. Learned representation-guided diffusion models for large-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8532–8542, 2024.

Robert Harb, Thomas Pock, and Heimo Müller. Diffusion-based generation of histopathological whole slide images at a gigapixel scale. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5131–5140, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2`, 2025.

Jingxiong Li, Sunyi Zheng, Chenglu Zhu, Yuxuan Sun, Pingyi Chen, Zhongyi Shui, Yunlong Zhang, Honglin Li, and Lin Yang. Pathup: Patch-wise timestep tracking for multi-class large pathology image synthesising diffusion model. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 3984–3993, 2024.

Jingxiong Li, Chenglu Zhu, Sunyi Zheng, Pingyi Chen, Yuxuan Sun, Honglin Li, and Lin Yang. Topofm: Topology-guided pathology foundation model for high-resolution pathology image synthesis with cellular-level control. *IEEE Transactions on Medical Imaging*, 44(10):4116–4127, 2025. doi: 10.1109/TMI.2025.3548872.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

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*, pp. 2000–2009, 2023.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7176–7185, 2020.

Lonzetta Neal, Nicole P Sandhu, Tina J Hieken, Katrina N Glazebrook, Maire Brid Mac Bride, Christina A Dilaveri, Dietlind L Wahner-Roedler, Karthik Ghosh, and Daniel W Visscher. Diagnosis and management of benign, atypical, and indeterminate breast lesions detected on core needle biopsy. In *Mayo Clinic Proceedings*, volume 89, pp. 536–547. Elsevier, 2014.

Amanda Arantes Perez, Débora Balabram, Marcio de Almeida Salles, and Helenice Gobbi. Ductal carcinoma in situ of the breast: correlation between histopathological features and age of patients. *Diagnostic pathology*, 9:1–6, 2014.

Adalberto Claudio Quiros, Roderick Murray-Smith, and Ke Yuan. PathologyGAN: Learning deep representations of cancer tissue. In *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, volume 121 of *Proceedings of Machine Learning Research*, pp. 669–695, 2020.

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*, 2022.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Julio Silva-Rodríguez, Adrián Colomer, María A. Sales, Rafael Molina, and Valery Naranjo. Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. *Computer Methods and Programs in Biomedicine*, 195: 105637, 2020. doi: 10.1016/j.cmpb.2020.105637.

Gary Tozbikian, Edi Brogi, Christina E. Vallejo, Dilip Giri, et al. Atypical ductal hyperplasia bordering on ductal carcinoma in situ: Interobserver variability and outcomes in 105 cases. *International Journal of Surgical Pathology*, 25(2):100–107, 2017. doi: 10.1177/1066896916662154.

Thanh Tran, Oh-Heum Kwon, Ki-Ryong Kwon, Suk-Hwan Lee, and Kyung-Won Kang. Blood cell images segmentation using deep learning semantic segmentation. In *2018 IEEE international conference on electronics and communication engineering (ICECE)*, pp. 13–16. IEEE, 2018.

Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, 30:2924–2935, 2024. doi: 10.1038/s41591-024-03141-0.

Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 81:102559, 2022a. doi: 10.1016/j.media.2022.102559.

Y Wang, B Acs, S Robertson, B Liu, Leslie Solorzano, Carolina Wählby, J Hartman, and M Rantalainen. Improved breast cancer histological grading using deep learning. *Annals of Oncology*, 33 (1):89–98, 2022b.

Xun Wu, Shaohan Huang, and Furu Wei. Mole: Mixture of lora experts. In *The Twelfth International Conference on Learning Representations*, 2023.

Yuan Xue, Jiarong Ye, Qianying Zhou, L. Rodney Long, Sameer Antani, Zhiyun Xue, Carl Cornwell, Richard Zaino, Keith C. Cheng, and Xiaolei Huang. Selective synthetic augmentation with histogan for improved histopathology image classification. *Medical Image Analysis*, 67:101816, 2021. doi: 10.1016/j.media.2020.101816.

Srikar Yellapragada, Alexandros Graikos, Zilinghan Li, Kostas Triaridis, Varun Belagali, Tarak Nath Nandi, Karen Bai, Beatrice S. Knudsen, Tahsin Kurc, Rajarsi R. Gupta, Prateek Prasanna, Ravi K. Madduri, Joel Saltz, and Dimitris Samaras. PixCell: A generative foundation model for digital histopathology images. *arXiv preprint arXiv:2506.05127*, 2025. doi: 10.48550/arXiv.2506.05127. Version 2.

Srikar Yellapragada et al. Pathldm: Text conditioned latent diffusion model for histopathology. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5182–5191, 2024.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations*, 2023.

A Additional Method Details

A.1 Diagnostic and Morphological Geometry

Pathology features. All geometry statistics are constructed exclusively from the training split and remain fixed during subsequent expert training. We use a frozen UNI2-h encoder to extract

pathology representations. Specifically, UNI2-h uses the released ViT-Giant/14 configuration with an input resolution of 224×224 and class-token pooling. Each RGB pathology crop is resized over its full field of view to 224×224 using bilinear interpolation, converted to $[0, 1]$, and normalized with ImageNet mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$. The resulting 1,536-dimensional pooled feature is ℓ_2-normalized and denoted by $u_i \in \mathbb{R}^{1536}$.

Slide-balanced means and whitening. Let $\mathcal{S}_y$ denote the set of training slides containing category y, and let $n_{y,s}$ be the number of category-y crops on slide s. For crop i from slide $s(i)$, we define

$$w_i = \frac{1}{C|\mathcal{S}_{y_i}|n_{y_i,s(i)}}, \qquad m_y = C \sum_{i:y_i=y} w_i u_i, \qquad \bar{m} = \frac{1}{C}\sum_{y=1}^{C} m_y. \tag{A1}$$

Thus, $\sum_i w_i = 1$, and each diagnostic category contributes total weight $1/C$ irrespective of its number of slides or crops.

We construct the within-class scatter matrix as

$$S_w = \sum_i w_i (u_i - m_{y_i})(u_i - m_{y_i})^\top, \tag{A2}$$

and apply shrinkage regularization,

$$S_\eta = (1-\eta)S_w + \eta \frac{\mathrm{tr}(S_w)}{d} I, \qquad \eta = 0.01. \tag{A3}$$

Let $S_\eta = V \,\mathrm{diag}(\ell_j) V^\top$,

The whitening transform is

$$W = V \,\mathrm{diag}\left(\max\left\{ \ell_j, \frac{\ell_{\max}}{1000} \right\}^{-1/2} \right) V^\top, \tag{A4}$$

where $\ell_{\max} = \max_j \ell_j$. The eigenvalue floor limits the condition number of the regularized covariance used for whitening to at most 1000. All geometry statistics are computed in float64.

Category-neighbor graph. The category graph combines feature similarity with empirical classification confusion. For graph construction only, we form an image-averaged category prototype

$$p_y = \frac{n_y^{-1} \sum_{i:y_i=y} u_i}{\left\| n_y^{-1} \sum_{i:y_i=y} u_i \right\|_2}. \tag{A5}$$

These prototypes are distinct from the slide-balanced means m_y used in the scatter matrices above.

We train a class-balanced logistic regressor on the training split and obtain out-of-fold predictions using five stratified, slide-grouped folds with random seed 37. No slide appears in both the training and held-out portions of a fold. Each regressor is optimized with `lbfgs` for at most 2,000 iterations. Let H_{ab} denote the row-normalized confusion frequency for true category a and predicted category b. For $a \neq b$, we define the symmetric pair score

$$T_{ab} = \frac{1}{2}\left(\frac{1 + p_a^\top p_b}{2} \right) + \frac{1}{2}\left(\frac{H_{ab} + H_{ba}}{2} \right). \tag{A6}$$

For each category a, we retain the two other categories with the largest T_{ab} values, with ties resolved by category index. Let $D_{ab} = T_{ab}$ when b is one of these directed neighbors of a, and $D_{ab} = 0$ otherwise. The final undirected graph weight is

$$\omega_{ab} = \max(D_{ab}, D_{ba}), \qquad \omega_{aa} = 0. \tag{A7}$$

Category bases. For a pair of diagnostic categories (a, b), define the whitened mean difference $\Delta_{ab} = W(m_a - m_b)$. The category scatter matrix is

$$S_c = \sum_{a<b} \omega_{ab} \Delta_{ab} \Delta_{ab}^\top. \tag{A8}$$

The columns of B are its leading nonzero eigenvectors, retaining at most $C - 1$ directions. We treat an eigenvalue as nonzero when it exceeds $d\,\varepsilon_{64}$ times the largest eigenvalue, where d is the feature dimension and ε_{64} is machine precision for float64. The resulting columns of B are orthonormal.

Morphology basis. We construct the morphology subspace from category-centered variation that is orthogonal to the diagnostic subspace. A complete QR decomposition of B yields an orthonormal complement Q satisfying $B^\top Q = 0$. For each training image, we compute

$$a_i = Q^\top W(u_i - m_{y_i}). \tag{A9}$$

We stack the weighted vectors $\sqrt{w_i}\, a_i^\top$ into a matrix and perform singular value decomposition. Let V_m contain the leading right singular vectors associated with the complementary variation. The morphology basis and the corresponding category and image coordinates are then

$$P = QV_m, \qquad q_y = B^\top W(m_y - \bar{m}), \qquad s_i = P^\top W(u_i - m_{y_i}). \tag{A10}$$

By construction, $B^\top P = 0$ up to numerical precision, preventing category directions from being reintroduced into the morphology subspace.

We choose the smallest morphology rank that explains 95% of the nonzero complementary variance, subject to a maximum dimension of 32. Both BRACS and SICAPv2 reach this upper bound, and therefore use $d_m = 32$. Because of the dimension cap, this does not imply that the retained components necessarily explain the full 95% variance.

Discrete modes and continuous residuals. Within each diagnostic category, we apply Euclidean K-means to the morphology coordinates s_i using ten initializations.Starting from $K = 10$, we decrease K until every cluster contains at least $\max(100, \lceil 0.05 n_y \rceil)$ training images; the first feasible value is retained as K_y. Cluster identifiers are canonicalized by lexicographically ordering their centers, after which each center is recomputed as the arithmetic mean of its assigned coordinates.

For category y, let $\mu_{y,k}$ denote the center of mode k, $n_{y,k}$ its number of assigned training samples, and K_y the retained number of modes. The smoothed sampling prior is

$$\pi_{y,k} = \frac{n_{y,k} + 1}{n_y + K_y}. \tag{A11}$$

For image i, the residual relative to its assigned mode is $r_i = s_i - \mu_{y_i,k_i}$. Following Eq. (4) in the main paper, we compute the category-level coordinate scale

$$\rho_{y,j} = \left(\frac{1}{n_y} \sum_{i:y_i=y} r_{i,j}^2 \right)^{1/2}, \tag{A12}$$

where the statistics are pooled over all modes within the same category. The normalized residual is $\widetilde{r}_i = r_i \oslash \rho_{y_i}$, and the continuous morphology condition is obtained by radial clipping,

$$z_i = \frac{\widetilde{r}_i}{\max(1, \|\widetilde{r}_i\|_2 / R)}, \qquad R = 12. \tag{A13}$$

This is the implementation of Eq. (5) in the main paper. When $K_y = 1$, the centered discrete mode contribution defined in Section 3.3 vanishes, while the continuous within-category residual condition remains active.

A.2 Morphology-Conditioned Residual Experts

Each diagnostic category y is associated with a separate residual expert E_y. The shared pathology prior remains frozen, and only the expert corresponding to the current category is optimized. Each expert uses rank $r = 32$ and adapts 60 linear transformations of the frozen generator.

For adapted layer ℓ, let $d_{\mathrm{in},\ell}$ and $d_{\mathrm{out},\ell}$ denote its input and output widths. The low-rank factors have shapes $D_\ell \in \mathbb{R}^{32 \times d_{\mathrm{in},\ell}}, U_\ell \in \mathbb{R}^{d_{\mathrm{out},\ell} \times 32}$. As in Eq. (6) of the main paper, the same low-rank factors are shared across morphology modes, while the rank-wise activation is modulated by the structured conditions.

Each category contains two conditioning MLPs. The category network f_b maps the fixed diagnostic code q_y to the layer-wise category gates, while the mode network f_m maps each mode center $\mu_{y,k}$ to

Table A1: Architecture and parameter counts of one morphology-conditioned residual expert. Each expert adapts $L = 60$ linear transformations with rank $r = 32$ and uses a $d_m = 32$ morphology representation. The category-code dimension is $d_c = 6$ for BRACS and $d_c = 3$ for SICAPv2. Parameter counts correspond to the implemented tensor shapes and include biases where applicable.

Component	Architecture	Parameters	Initialization
D, U	Rank-32 factors	33,423,360	D: normal; U: zero
f^b	$d_c \to 64 \to 1920$, SiLU	$124864 + 64d_c$	Linear defaults
f^m	$32 \to 64 \to 1920$, SiLU	126,912	Linear defaults
A	$32 \to 1920$, no bias	61,440	Zero

the corresponding raw mode gates. Both MLPs contain two linear layers with a hidden width of 64 and a SiLU nonlinearity. Their outputs contain $60 \times 32 = 1920$ entries, reshaped into one rank-32 gate for each adapted layer. Thus, the same conditioning network is shared across adapted layers within an expert, but produces a distinct gate for every layer.

The continuous residual modulator implements Eq. (8) of the main paper, $\delta(z) = \alpha \tanh(Az)$, with $\alpha = 0.25$. Here,$A \in \mathbb{R}^{1920\times32}$ is a bias-free linear projection whose output is likewise reshaped into 60 layer-specific rank-32 modulation vectors. The modulator therefore contains 61,440 parameters per category.

The down-projections D_ℓ use their standard initialization, whereas all up-projections U_ℓ are initialized to zero. Consequently, each newly initialized expert initially reproduces the frozen shared pathology prior regardless of the initial category or mode gates. The residual projection A is also initialized to zero, independently disabling continuous residual modulation at initialization. The two gate MLPs use PyTorch's default linear-layer initialization.

Expert training is class-specific and uses class-homogeneous batches. For every image, its discrete mode k_i and continuous residual z_i are retained individually. The resulting rank gate is broadcast over all spatial or token positions within the corresponding adapted layer. The geometry itself is fixed throughout expert training. Table A1 summarizes the trainable components of one category expert.

A.3 Paired Pathology Feature Refinement

Paired feature extraction. For the frozen encoder ψ in Section 3.4, we use the 2,048-dimensional pooled representation of an ImageNet-pretrained Inception-V3 network, giving $d_\psi = 2048$. The predicted clean image and its paired real training crop undergo identical differentiable preprocessing. Images are clipped to $[0, 1]$, resized to 342 pixels, center-cropped to 299×299, and normalized using the standard ImageNet statistics. The resulting feature vectors are used without output normalization. The paired feature objective is therefore exactly the mean-squared feature distance defined in Eq. (10) of the main paper.

The VAE decoder and Inception-V3 encoder remain frozen. No gradient is propagated through the real-image feature branch. Gradients from the predicted-image branch pass through the frozen decoder and feature encoder to the trainable expert projections, conditioning networks, and residual modulator.

Pair selection and accumulation. The effective batch size is $B_{\text{eff}} = 8$

Each optimizer update selects one example for paired feature supervision by cycling through the effective batch. For zero-indexed update t, the selected one-based position is $j_t = 1 + (t \bmod B_{\text{eff}})$ The paired objective is applied only when the selected image lies in the prescribed low-noise interval $0.1 \le \sigma_{j_t} \le 0.5$ Defining $I_t = \mathbb{I}[0.1 \le \sigma_{j_t} \le 0.5]$, the refinement objective for one optimizer update is

$$\mathcal{L}_t = \frac{1}{B_{\text{eff}}} \sum_{i=1}^{B_{\text{eff}}} \mathcal{L}_{\text{FM},i} + I_t \lambda_y \mathcal{L}_{\text{pair}} \left(\hat{x}_{j_t}, x_{j_t}\right). \tag{A14}$$

Thus, flow matching supervises every example in the effective batch, whereas the paired feature term is evaluated for at most one image per optimizer update. If the selected sample falls outside the eligible noise range, the update reduces to the flow-matching objective alone.

When gradient accumulation is used, the feature term on the selected microbatch is scaled by the number of accumulation steps before the standard accumulation divisor is applied. This preserves equation A14 at the effective-batch level. The image ordering, image-specific noise samples, and flow times are unchanged when the effective batch is partitioned into different microbatch sizes.

Training-only calibration. The feature-loss coefficient is calibrated independently for each diagnostic category before formal paired refinement. For up to three eligible calibration events, we measure the gradient norms of the flow-matching and unweighted paired-feature objectives with respect to the concatenated trainable expert parameters. We then set

$$\lambda_y = 0.1 \,\mathrm{median}_e \left(\frac{\|\nabla_{\Theta_y} \mathcal{L}_{\mathrm{FM}}^{(e)}\|_2}{\|\nabla_{\Theta_y} \mathcal{L}_{\mathrm{pair}}^{(e)}\|_2} \right). \tag{A15}$$

The calibration examines at most 16 optimizer updates and is used only to determine an initial relative gradient scale. The resulting λ_y is fixed throughout paired refinement rather than dynamically maintaining a constant gradient ratio.

Calibration is performed only on training data. Its temporary parameter updates are discarded, and formal refinement resumes from the source checkpoint with the corresponding model, optimizer, scheduler, and random states restored. Calibration steps are therefore excluded from the reported training budget. The shared pathology prior, VAE, and feature encoder remain frozen throughout refinement.

A.4 Training and Inference

FIRM is trained in three stages. First, we adapt the pretrained generator with a shared pathology LoRA using pooled training images and the dataset-specific generic pathology prompt. Second, the backbone and shared adaptation are frozen, and a separate morphology-conditioned residual expert is trained for each diagnostic category using flow matching. Third, each expert is continued with paired pathology feature refinement. The diagnostic and morphological geometry is computed before expert training and remains fixed throughout both expert stages.

At inference, neither UNI2-h nor the paired-feature encoder is required. For each generated image, we perform the following procedure:

1. Activate the residual expert E_y and category prompt for the requested diagnostic category y.
2. Sample a discrete morphology mode $k \sim \mathrm{Cat}(\pi_y)$,
   and then sample z uniformly from the stored training residual conditions assigned to mode k.
3. Draw the initial Gaussian noise and keep (y, k, z) fixed throughout all 28 sampling steps.
4. Decode the final latent representation using the frozen VAE decoder.

Therefore, the pathology feature encoders and geometry construction introduce only offline or training-time computation. Generation uses the shared pathology prior together with a single active category expert.

B Implementation and Evaluation Details

B.1 Dataset Processing

BRACS. We construct the BRACS crop dataset from the annotated regions at $20\times$ magnification. Each region is tiled into RGB crops of size 512×512 with a 128-pixel overlap between neighboring crops. Crops with a whitespace ratio greater than 0.8 are discarded. This preprocessing yields 38,638

Table A2: Class-wise numbers of pathology crops in the patient-disjoint training, validation, and test splits of BRACS and SICAPv2. All image-derived FIRM conditions are constructed exclusively from the corresponding training split.

Dataset	Class	Train	Validation	Test
BRACS	N	2,578	301	482
	PB	7,381	280	584
	UDH	1,750	186	407
	FEA	1,667	278	275
	ADH	1,611	166	392
	DCIS	6,364	539	565
	IC	9,062	1,175	2,595
SICAPv2	NC	3,088	685	644
	G3	1,204	625	393
	G4	2,662	979	853
	G5	518	198	232

crops in total, comprising 30,413 training, 2,925 validation, and 5,300 test images, consistent with the dataset statistics reported in the main paper.

The resulting training, validation, and test partitions contain 275, 37, and 69 slides from 103, 15, and 30 patients, respectively. Patient, slide, and crop identifiers are disjoint across the three partitions. The same preprocessing pipeline defines the image pools used throughout generation and downstream evaluation. All image-derived conditions used by FIRM, including the diagnostic and morphological geometry and stored residual conditions, are constructed exclusively from the training split.

SICAPv2. For SICAPv2, we retain the four official diagnostic categories and their original spreadsheet labels. The official class lists contain 12,081 images in total, divided into 7,472 training, 2,487 validation, and 2,122 test images. The source patches are 512×512 and correspond to $10\times$ magnification, following the setting used in the main paper. The training, validation, and test partitions contain 95, 29, and 31 slides, respectively, and are drawn from disjoint patient groups. As for BRACS, all geometry construction and image-derived conditioning use the training partition only.

B.2 Training Configuration

Shared pathology prior. We initialize FIRM from the released FLUX.2-klein-4B model.[1] The shared pathology prior is learned by adapting the pretrained generator with a rank-32 LoRA using images pooled across all diagnostic categories. The LoRA scaling parameter is also set to 32.

Within each dataset, all images use the same generic pathology prompt c_0. Specifically, we use ``breast histopathology H&E 20x tile'' for BRACS and ``prostate histopathology H&E 10x tile'' for SICAPv2. The shared adaptation is trained for 30,000 optimizer updates with AdamW, a learning rate of 10^{-4}, 200 warmup updates, and an effective batch size of eight. After this stage, both the pretrained backbone and the shared LoRA are frozen for all subsequent expert training.

Morphology-conditioned experts. A separate rank-32 residual expert is trained for each diagnostic category. Following the training procedure in Section 3, each expert first undergoes 1,000 flow-matching updates and is then continued for 1,000 updates with paired pathology feature refinement. The diagnostic and morphological geometry remains fixed throughout both stages.

Expert parameters are optimized with AdamW using $(\beta_1, \beta_2) = (0.9, 0.999), \epsilon = 10^{-8}$, a weight decay of 10^{-4}, and gradient-norm clipping at 1. The learning rate is 5×10^{-5} with 200 warmup updates followed by a constant schedule. The low-rank expert parameters, conditioning networks, and continuous residual modulator use the same learning rate. The effective batch size is eight throughout expert training.

[1] https://huggingface.co/black-forest-labs/FLUX.2-klein-4B

Unless otherwise stated, the reported FIRM models follow this configuration. Generation is performed at 512×512 resolution using 28 sampling steps and guidance scale 1. The discrete morphology mode and continuous residual condition sampled at the start of a trajectory remain fixed across all sampling steps.

B.3 Evaluation Protocol

Sampling and image preprocessing. For the main generative evaluation, we sample 1,000 images per diagnostic category and compute all metrics independently for each class before reporting the macro-average across classes. All image-derived generation conditions are obtained exclusively from the training split.

To ensure consistent evaluation across methods, generated images are evaluated over the same full field of view. Native 512×512 outputs are first resized to 256×256 without cropping, as described in the main paper. Encoder-specific preprocessing is then applied to these standardized evaluation images. Thus, the 256×256 evaluation resolution specifies the common image field presented to the evaluator, while the subsequent resizing or cropping described below is part of the fixed preprocessing associated with each feature encoder.

Evaluation feature spaces. We evaluate generation in both general-purpose and pathology-specific representation spaces.

For FID, KID, and Coverage, we use the raw 2,048-dimensional pooled representation of an ImageNet-pretrained Inception-V3 model from `torchvision` using standard ImageNet normalization. For pathology-specific evaluation, we use the released Virchow v1 ViT-H/14 model. RGB images are resized to 224×224 using bicubic interpolation and normalized using the ImageNet mean and standard deviation. We form a 2,560-dimensional representation by concatenating the class token with the mean patch token. The representation is used without output normalization. We refer to the Fréchet distance computed in this feature space as Path-F and to generative recall in this space as Path-R.

The three pathology-related encoders therefore serve distinct purposes in FIRM. UNI2-h is used only to construct the diagnostic and morphological geometry, Inception-V3 provides the paired feature objective and the general-purpose evaluation space, and Virchow provides the pathology-specific evaluation space.

Fréchet and kernel distances. Let μ_r, Σ_r and μ_g, Σ_g denote the empirical means and covariances of real and generated features, respectively. We compute the Fréchet distance as

$$D_{\mathrm{F}} = \|\mu_r - \mu_g\|_2^2 + \operatorname{tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r^{1/2}\Sigma_g\Sigma_r^{1/2}\right)^{1/2}\right). \tag{A16}$$

Applying equation A16 to Inception-V3 features gives FID, while applying the same form to Virchow features gives Path-F. All covariance and matrix operations use the full feature dimensionality.

KID is computed in Inception feature space using the unbiased squared maximum mean discrepancy estimator with the polynomial kernel

$$k(a, b) = \left(\frac{a^\top b}{d} + 1\right)^3, \tag{A17}$$

where d is the feature dimension. As in the main paper, reported KID values are multiplied by 10^3.

Coverage and pathology recall. Coverage and Path-R follow the neighborhood-based formulation with Euclidean feature distances and $k = 5$ nearest-neighbor radii. For real features $\{r_i\}_{i=1}^{n_r}$ and generated features $\{g_j\}_{j=1}^{n_g}$, let ρ_i denote the distance from r_i to its fifth-nearest other real feature. Coverage is defined in Inception feature space as

$$\text{Coverage} = \frac{1}{n_r}\sum_{i=1}^{n_r} \mathbb{I}\left[\min_j \|r_i - g_j\|_2 < \rho_i\right]. \tag{A18}$$

It measures the fraction of real-data neighborhoods reached by at least one generated sample.

Table A3: Training and sampling configuration used for FIRM on BRACS and SICAPv2. Expert training consists of 1,000 flow-matching updates followed by 1,000 paired feature-refinement updates; update counts are cumulative within each category expert.

Setting	
Backbone	FLUX.2-klein-4B
Training / sampling size	512×512
Shared-prior updates	30,000
Expert stages	1,000 FM + 1,000 paired
Rank / LoRA scaling	32 / 32
Effective batch	8
Optimizer	AdamW, $(0.9, 0.999)$
Prior / expert LR	10^{-4} / 5×10^{-5}
Residual-modulator LR	5×10^{-5}
Expert weight decay	10^{-4}
Warmup / schedule	200 updates / constant
Gradient norm cap	1
Precision	BF16; FP8 transformer path
Expert-training seed	71
Paired loss	$0.1 \leq \sigma \leq 0.5$, one pair/update
Inference	28 steps, guidance 1, BF16

For pathology recall, let γ_j denote the distance from generated Virchow feature g_j to its fifth-nearest other generated feature. We compute

$$\text{Path-R} = \frac{1}{n_r} \sum_{i=1}^{n_r} \mathbb{I}\left[\exists j : \|r_i - g_j\|_2 < \gamma_j\right]. \tag{A19}$$

Here, r_i and g_j are the raw Virchow representations. We use strict distance inequalities in both neighborhood metrics. Coverage and Path-R are reported as percentages.

Lower FID, KID, and Path-F indicate better distribution matching, whereas higher Coverage and Path-R indicate broader coverage of the corresponding real feature distributions. All reported aggregate values are macro-averages over diagnostic categories.

C Extended Quantitative Results

C.1 Per-Class Generation Results

Figures C.1 and C.1 expand the aggregate comparison in Table 1 of the main paper by reporting all five generation metrics for every diagnostic category and all six synthesis methods. The evaluation follows the same protocol described, 1,000 images are generated for each class, each class is evaluated independently against its corresponding real reference set, and the reported mean gives equal weight to all classes.

Consistent with the class-wise comparison summarized in Figure 3 of the main paper, FIRM achieves the lowest KID and Path-F for all seven BRACS categories and all four SICAPv2 categories. The remaining metrics are reported to expose class-specific behavior that is not visible from the macro-averages alone. Overall, the per-class results show that the aggregate gains of FIRM are not driven by a small subset of favorable categories, but are accompanied by consistent improvements in distribution matching across diagnostic classes.

C.2 Sensitivity to the Number of Morphology Modes

We further study the effect of the number of discrete morphology modes on SICAPv2. For this controlled experiment, we override the adaptive mode-selection procedure described in Appendix A and explicitly set $K \in \{1, 2, 4, 6\}$ for every diagnostic category. All configurations start from the same frozen shared pathology prior and use the same training budget of 2,000 expert updates,

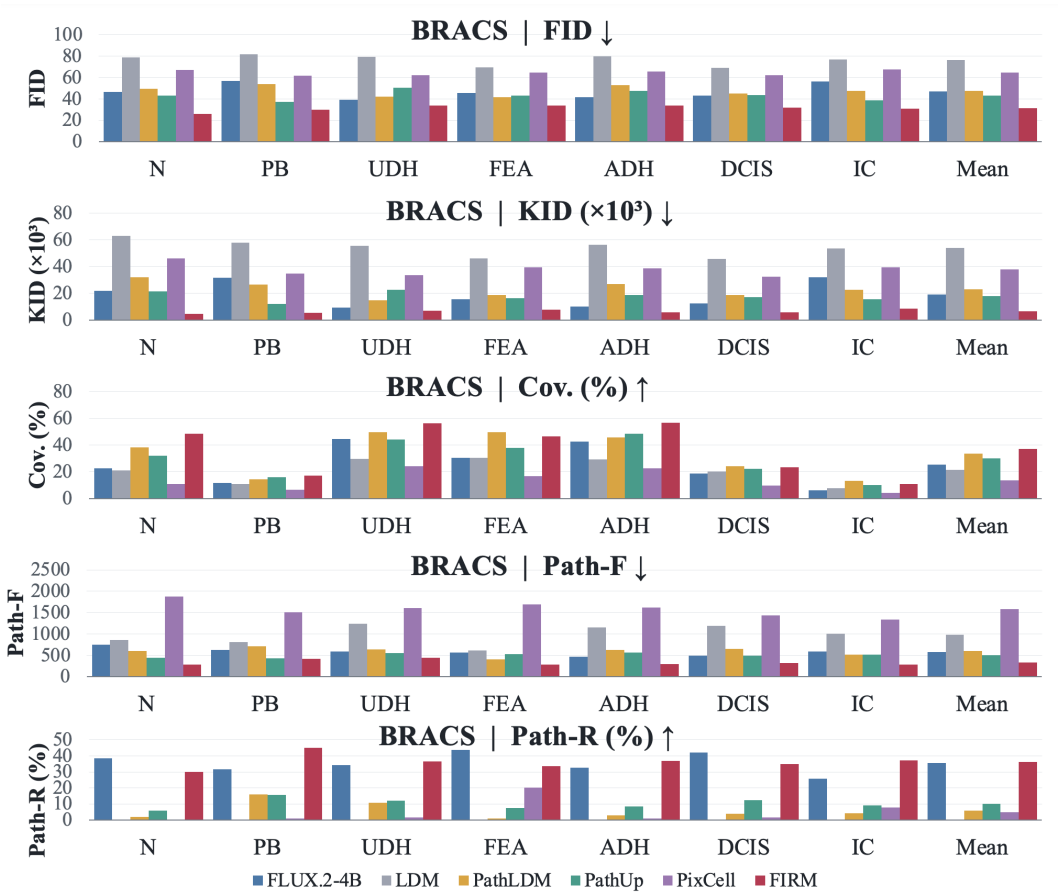


Figure A1: Per-class generative evaluation on BRACS. FID, KID, Coverage, Path-F, and Path-R are reported for all seven diagnostic categories and their macro-average. Lower FID, KID, and Path-F are better; higher Coverage and Path-R are better.

consisting of 1,000 flow-matching updates followed by 1,000 paired feature-refinement updates. When $K = 1$, the centered discrete mode gate vanishes by construction, while the continuous residual condition remains active.

For each setting, we generate 1,000 images per class at 512×512 resolution using 28 sampling steps and guidance scale 1. The initial image-noise seeds are shared across different values of K within each class to reduce variation unrelated to the mode configuration. Figure C.2 reports the resulting class-wise and macro-averaged Path-F.

All multi-mode configurations improve the four-class mean over $K = 1$: the mean Path-F decreases from 103.40 at $K = 1$ to 99.98, 99.61, and 99.80 at $K = 2$, 4, and 6, respectively. The best macro-average is obtained with $K = 4$, while the class-wise optimum varies: $K = 4$ gives the lowest Path-F for NC and G5, whereas $K = 6$ performs best for G3 and G4. These results support the use of multiple morphology centers, while also showing that increasing the number of modes does not yield a monotonic improvement.

C.3 Morphology Controllability Evaluation

We quantify continuous morphology control along a mode-specific direction estimated from the training residuals. For each category y and mode k, we compute the first principal-component unit direction $v_{y,k}$ of the normalized training residuals $\widetilde{r}_i = (s_i - \mu_{y,k}) \oslash \rho_y$. Let $\sigma_{y,k}$ denote the sample standard deviation of $v_{y,k}^\top \widetilde{r}_i$ within that mode. We intervene along this direction using

$$z_{y,k}(a) = \mathrm{cap}_{12}\left(a\sigma_{y,k}v_{y,k}\right), \qquad a \in \{-2, -1, 0, 1, 2\}, \tag{A20}$$

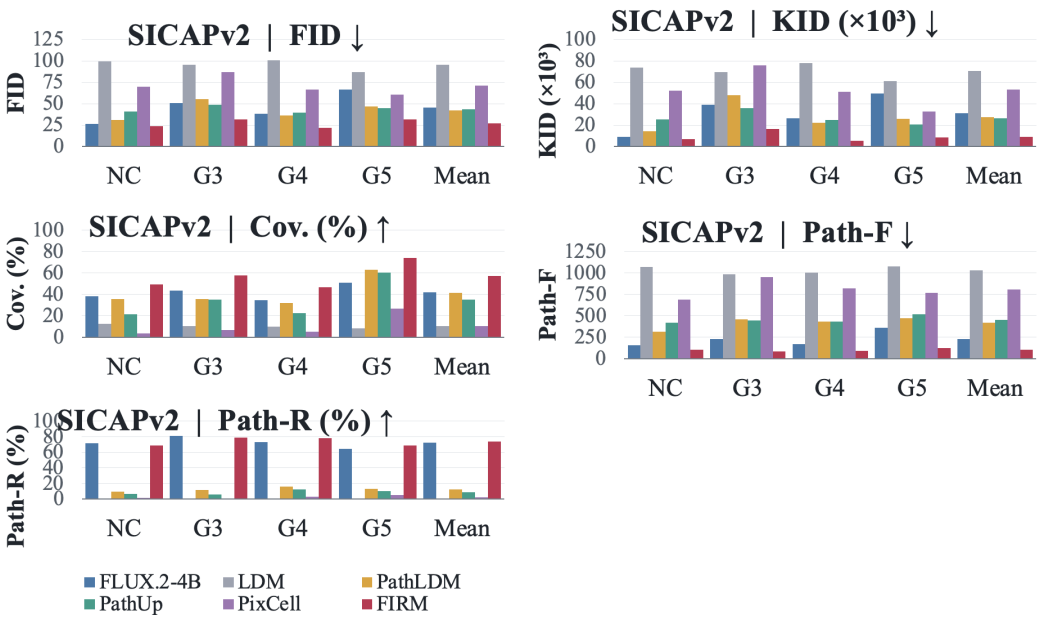


Figure A2: Per-class generative evaluation on SICAPv2. FID, KID, Coverage, Path-F, and Path-R are reported for all four diagnostic categories and their macro-average. Lower FID, KID, and Path-F are better; higher Coverage and Path-R are better.

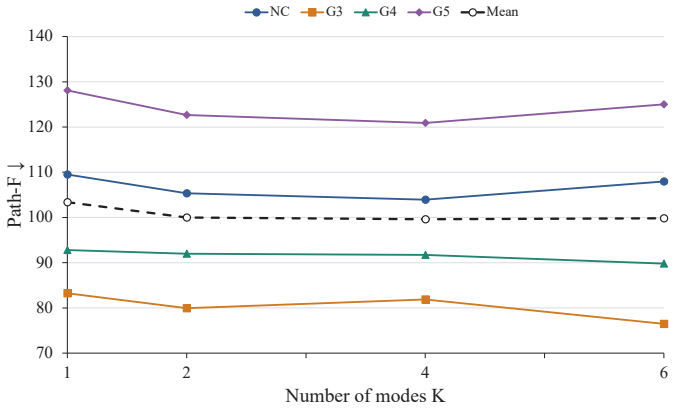


Figure A3: Per-class sensitivity of K on SICAPv2.

where cap_{12} is the same residual-norm cap used during training. Thus, the intervention level is expressed in standard-deviation (SD) units of the corresponding training-mode residual distribution.

For a generated image $x_{y,k,n}(a)$, we project its frozen pathology feature back to the same morphology space,

$$\widehat{s}_{y,k,n}(a) = P^\top W\left(\phi(x_{y,k,n}(a)) - m_y\right), \tag{A21}$$

where $\phi(\cdot)$ denotes the same fixed feature representation used to construct the geometry. The signed output response along the intervention direction is

$$t_{y,k,n}(a) = \frac{v_{y,k}^\top[(\widehat{s}_{y,k,n}(a) - \mu_{y,k}) \oslash \rho_y]}{\sigma_{y,k}}, \qquad \Delta t_{y,k,n}(a) = t_{y,k,n}(a) - t_{y,k,n}(0). \tag{A22}$$

so the zero-residual output serves as the reference and the response is reported in SD units.

For each fixed category, mode, and noise seed, we compute the signed Spearman correlation between the five intervention levels and their corresponding output coordinates. The class-level statistic

reported in Figure 7 is the equal-weight average over all modes and 16 noise seeds. For PB, UDH, and IC, the resulting mean signed correlations are 0.265, 0.575, and 0.804, respectively.

For discrete-mode stability, each generated output is assigned to its nearest frozen morphology-mode center within the requested category. A trajectory is considered mode-preserving when this nearest-center assignment is unchanged across all five intervention levels.

D Additional Qualitative Results

Figures A4 and A5 extend the qualitative comparison in the main paper to all diagnostic categories and all six synthesis methods. Each row corresponds to one diagnostic category and contains one real reference image together with one generated example from FLUX.2, LDM, PathLDM, PathUp, PixCell, and FIRM.

Generated examples use matched per-class sample indices across methods for reproducibility. Because the compared models have different architectures and sampling procedures, a shared sample index does not imply an identical latent or noise tensor across methods. Real-reference images are sampled independently from the corresponding real-image pool without replacement using NumPy seed 6061 in canonical class order. They serve only as visual references and are not paired targets for the generated images.

All displayed images use the full 256×256 evaluation field of view, without stain normalization or additional local cropping. For methods that natively generate 512×512 images, the full image is resized to the evaluation resolution without cropping, following the protocol used for the quantitative comparison in Section 4.2. These examples complement the focused ADH comparison in Figure 4 of the main paper by exposing category-specific appearance across the complete BRACS and SICAPv2 label sets.

E Downstream Classification Details

Training data and augmentation protocol. We evaluate the utility of synthetic images for downstream classification on SICAPv2. All four training settings use the complete real training split of 7,472 images. We compare: (i) real images only, (ii) real images augmented with repeated real samples, (iii) real images augmented with FLUX.2 samples, and (iv) real images augmented with FIRM samples.

For each augmented setting, we add 4,000 images, corresponding to 1,000 additional images per diagnostic category. The repeated-real setting samples 1,000 training images per class with replacement, while the two synthetic settings add 1,000 generated images per class. Therefore, all augmented settings use identical class counts and the same total number of training images. For both synthetic settings, all image-derived generation conditions are constructed exclusively from the SICAPv2 training split. No patient, slide, or crop from the validation or test partitions is used to construct the augmented training pools.

Classifier and optimization. We use a torchvision ResNet-50 initialized with ImageNet-1K V1 weights. The full network is fine-tuned, including the convolutional backbone and BatchNorm layers, together with a newly initialized classification head for the four SICAPv2 categories. Training minimizes unweighted cross-entropy over NC, G3, G4, and G5.

We optimize the classifier with AdamW using a backbone learning rate of 3×10^{-5} and a classification-head learning rate of 3×10^{-4}. The optimizer uses weight decay 10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.999)$, and gradient-norm clipping at 1. The learning-rate schedule consists of 50 linear warmup updates followed by cosine decay to zero. Model parameters and optimizer states are maintained in FP32, with BF16 autocast during training. The global batch size is 1,024 without gradient accumulation. Each run is trained for 1,000 optimizer updates, and the same optimization budget and training configuration are used for all four settings.

Classifier inputs use the full 256×256 image field of view. No additional local cropping is applied when constructing the downstream training examples.

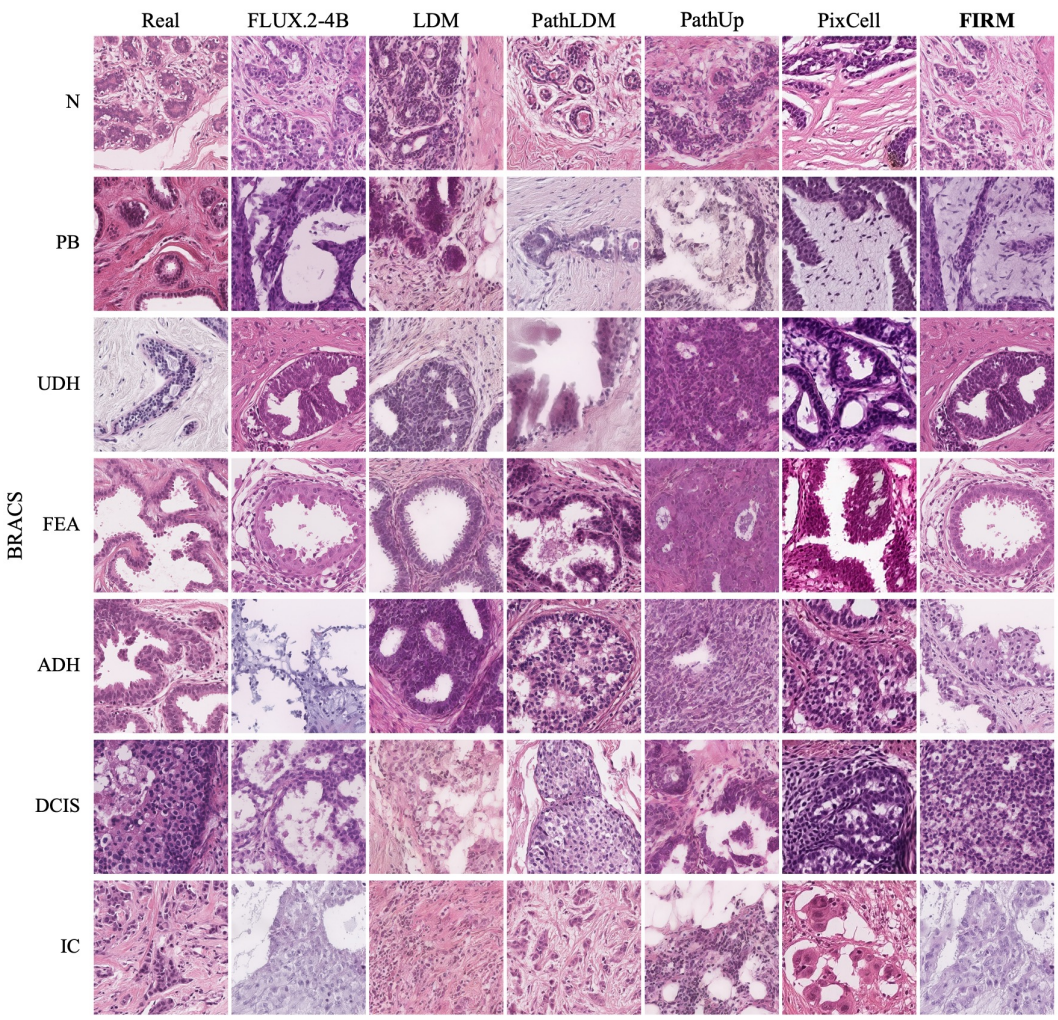


Figure A4: All-class qualitative comparison on BRACS. Each row corresponds to one diagnostic category and shows a real reference together with samples from the six evaluated synthesis methods. All images are displayed over the same full field of view.

Repeated runs and evaluation. We repeat each training setting with three fixed random seeds: 17, 1729, and 2609. All models are evaluated on the same patient-disjoint SICAPv2 test split used in the main paper. For each diagnostic category, we compute the class-wise F1 score. Macro-F1 is obtained by assigning equal weight to the four diagnostic categories. We report the mean and sample standard deviation across the three independent runs.

The resulting classification scores are reported in Table 4 of the main paper. Under the matched augmentation budget, FIRM augmentation improves macro-F1 from 63.39 ± 2.18 with real-only training to 67.43 ± 1.21, compared with 63.73 ± 0.98 for repeated-real augmentation and 65.38 ± 1.47 for FLUX.2 augmentation. FIRM also achieves the highest mean F1 on G3, G4, and G5, with the largest improvement over real-only training observed for G5.

F Parameter Efficiency

FIRM introduces category-specific residual experts on top of a single shared pathology prior. The shared rank-32 LoRA is stored once and remains frozen during expert training. For each diagnostic category, only its corresponding low-rank factors and conditioning modules are optimized, while the pretrained backbone, shared adaptation, VAE, and feature encoders remain fixed.

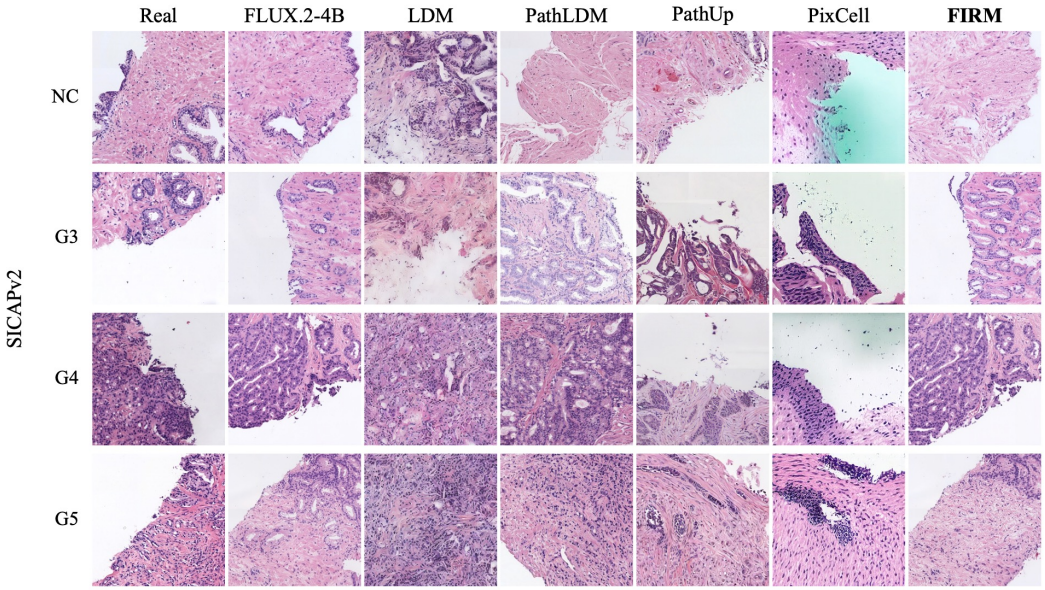


Figure A5: All-class qualitative comparison on SICAPv2. Each row corresponds to one diagnostic category and shows a real reference together with samples from the six evaluated synthesis methods. All images are displayed over the same full field of view.

Table A4: Adapter parameter counts and inference configuration for FIRM. The shared rank-32 LoRA is stored once, while each category has one residual expert containing its low-rank factors and conditioning modules. Only the requested category expert is active during generation.

Component	Value
Shared rank-32 LoRA	33,423,360
One complete class expert	33,736,960
Conditioning networks within one expert	313,600
Inference steps per image	28

Table A4 summarizes the adapter size and inference configuration. The shared pathology LoRA contains 33.42M parameters, while one complete morphology-conditioned expert contains approximately 33.74M parameters. Of these, only about 0.31M parameters belong to the conditioning networks, including the category gate, mode gate, and continuous residual modulator; the remaining parameters are the rank-32 low-rank factors. The values in the table correspond to the BRACS configuration with category-code dimension $d_c = 6$. For SICAPv2, $d_c = 3$, reducing the conditioning network and complete expert by only 192 parameters, while leaving the low-rank factors unchanged.

At inference, the shared pathology prior is combined with only the expert associated with the requested diagnostic category. Consequently, although the total stored adapter parameters grow linearly with the number of categories, the number of active expert parameters for a single generated image is independent of the total number of stored experts. The discrete morphology mode and continuous residual condition modulate the same low-rank path and therefore introduce no additional expert copies.

Generation uses the same 28-step sampling procedure as in the main experiments. Neither the UNI2-h encoder used to construct the diagnostic and morphological geometry nor the Inception-V3 encoder used for paired feature refinement is required during generation. These components therefore introduce only offline or training-time overhead and do not add additional encoder passes to the inference trajectory.