MULTI-FREQUENCY CONSISTENCY AND SPECTRAL-BASIS CACHING FOR SPECTRAL-AWARE STAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Virtual staining enables the generation of biologically informative microscopy images without physical staining procedures. However, conventional evaluation and training objectives may overlook fine-grained biological structures. Downstream evaluation is more biologically relevant but requires task-specific annotations, predictive models, or expert assessment, limiting the scalability of downstream evaluation. Meanwhile, diffusion-based virtual staining relies on iterative denoising and can incur substantial inference cost. We propose Spectral-Aware Staining, a spectral-guided framework for virtual staining that improves staining quality assessment and diffusion acceleration. First, we introduce Multi-Frequency Spectral-Aware Consistency (MSAC), a spectral-aware consistency metric that evaluates virtual staining quality by comparing retained multi-frequency spectral subspaces. By enforcing multi-frequency consistency, the MSAC better preserves biologically relevant morphology. Second, we propose Spectral-Basis Cache (SBCache), a structure-aware caching strategy for diffusion inference. SBCache estimates denoising redundancy in a learned graph-spectral space and enables reuse of intermediate computations while achieving a favorable quality–efficiency trade-off. Both MSAC and SBCache share the same trained neural spectral decomposer, which predicts ordered graph-spectral bases from feature graphs and replaces computationally expensive eigendecomposition with efficient network inference. Experiments on public microscopy benchmarks and virtual-staining frameworks demonstrate that MSAC more effectively evaluates the quality of generated stains and that Spectral-Aware Staining generates high-quality stained images more efficiently.

1 INTRODUCTION

Deep learning-based virtual staining provides an efficient alternative to physical staining by predicting fluorescence or immunohistochemistry channels from label-free microscopy (Christiansen et al., 2018; Han et al., 2026) or routinely stained histology (Bian et al., 2025; Oh et al., 2026). The synthesized stains can support cell segmentation, organelle quantification, and phenotypic profiling (Cross-Zamirski et al., 2022; 2023). Their quality should therefore depend not only on pixel-level similarity, but also on the preservation of biologically meaningful structures.

Virtual-staining methods are still mainly evaluated using peak signal-to-noise ratio (PSNR), Pearson correlation coefficient (PCC), and structural similarity index (SSIM) (Wang et al., 2004; Bai et al., 2023). These commonly used metrics may overlook small and sparse structures because of averaged intensity or local image statistics. As shown in Figure 1, a structurally more faithful prediction can receive worse scores on conventional metrics despite producing substantially better segmentation results. Similar mismatches between image-quality metrics and task-relevant structural preservation have also been observed in other medical imaging applications (Dohmen et al., 2025b). Downstream tasks provide a more biologically relevant evaluation (Sengupta et al., 2025), but usually require expert annotations or assessment. Their outcomes may also depend on the selected analysis pipeline. Therefore, an automatic full-reference metric that better reflects structural consistency is needed.

Efficiency presents another challenge for diffusion-based virtual staining. Although iterative denoising can recover fine microscopic structures, the denoising process requires repeated network evaluations. TeaCache reuses computations according to timestep-modulated raw-feature differences

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.

(Liu et al., 2025). SeaCache observes that such distances can entangle content evolution with noise and uses a Fourier transform to decompose intermediate features into frequency components before distance estimation (Chung et al., 2026). However, the Fourier basis remains fixed across inputs and does not adapt to image-specific feature affinities, motivating a graph-adaptive representation for estimating denoising redundancy.

Learned perceptual objectives enrich raw-pixel comparisons (Johnson et al., 2016; Zhang et al., 2018), and recent foundation models provide dense, high-dimensional visual representations (Siméoni et al., 2026). However, direct feature matching mixes structural variations associated with different frequencies. Graph-spectral decomposition offers a natural way to decompose dense, high-dimensional features into frequency components that characterize variation over the feature graph. Eigenvectors of a graph Laplacian matrix organize image variations into ordered graph-frequency components: lower-frequency components vary smoothly over strongly connected regions, while components with larger eigenvalues exhibit greater variation across graph connections (Shuman et al., 2013; Ortega et al., 2018). Selecting the retained components defines a multi-frequency spectral subspace for structural comparison. However, repeatedly computing image-specific spectral bases through numerical eigendecomposition is costly, motivating an efficient learned approximation.

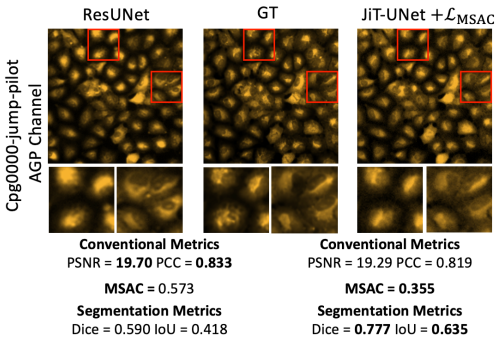


Figure 1: An example of conventional metrics' pitfall. Ground-truth stains (center) are compared with UNet (left) and JiT (right) predictions on the cpg0000-jump-pilot AGP channel. JiT-UNet $+\mathcal{L}_{\mathrm{MSAC}}$ better preserves the target structures and yields higher downstream Dice and IoU, yet PSNR and PCC favor the blurrier UNet predictions. In contrast, MSAC metric agrees with visual inspection and downstream segmentation.

In this work, we propose Spectral-Aware Staining, a spectral-guided framework for virtual staining that improves staining quality assessment and diffusion acceleration. The first component is Multi-Frequency Spectral-Aware Consistency (MSAC), which defines a new spectral-aware metric to evaluate the consistency between generated and reference stains over a retained multi-frequency spectral subspace. By enforcing multi-frequency consistency, MSAC better preserves biologically relevant morphology. The second component is Spectral-Basis Cache (SBCache) for diffusion inference. SBCache measures denoising redundancy in a learned graph-spectral space constructed by the proposed spectral decomposer. By using the spectral decomposer to decompose intermediate features into multi-frequency spectral bases, SBCache captures structure-relevant variations and reduces the influence of noise on structural-consistency estimation, leading to more reliable cache decisions for fine-grained microscopic structures. Both MSAC and SBCache share the same neural spectral decomposer for efficient extraction of graph spectral bases. The decomposer predicts ordered spectral bases from feature graphs and replaces computationally expensive eigendecomposition with a lightweight network evaluation.

Experiments on public microscopy benchmarks and virtual-staining frameworks demonstrate that MSAC more effectively evaluates the quality of generated stains and that Spectral-Aware Staining generates high-quality stained images more efficiently.

Our main contributions are:

- We propose Spectral-Aware Staining, a unified framework for more accurate and efficient virtual staining by leveraging multi-frequency spectral representations to improve structural fidelity.
- We introduce MSAC, a spectral-aware metric that provides a more reliable assessment of virtual-staining quality. By enforcing multi-frequency consistency, MSAC better preserves biologically relevant morphology.
- We propose SBCache, which exploits graph-spectral representations to identify redundancy in diffusion denoising trajectories and accelerates inference with a favorable quality–efficiency trade-off.

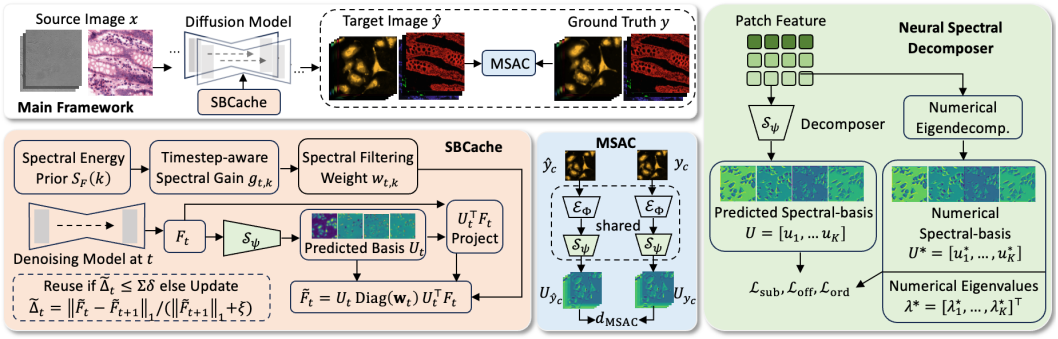


Figure 2: Overview of Spectral-Aware Staining, a unified spectral-guided framework for virtual staining evaluation and diffusion acceleration. Given a source-domain image, the goal of virtual staining is to generate a corresponding target-domain image. In the target domain, Multi-Frequency Spectral-Aware Consistency (MSAC) provides a spectral-aware consistency measure for reliable evaluation. By enforcing multi-frequency consistency, MSAC better preserves biologically relevant morphology. During inference, Spectral-Basis Cache (SBCache) exploits spectral representations to identify redundancy between adjacent denoising steps and accelerate diffusion sampling. Both MSAC and SBCache share the trained neural spectral decomposer to obtain ordered graph-spectral bases for multi-frequency consistency analysis and cache-efficient inference.

- Extensive experiments on public microscopy benchmarks and virtual-staining frameworks demonstrate that MSAC more effectively evaluates the quality of generated stains and that Spectral-Aware Staining generates high-quality stained images more efficiently.

2 Method

Overview. This section presents the proposed Spectral-Aware Staining framework, as shown in Figure 2. The framework consists of two main components: MSAC, which evaluates staining consistency between retained multi-frequency spectral subspaces, and SBCache, which improves diffusion inference by measuring redundancy in graph-spectral feature representations. By enforcing multi-frequency consistency, MSAC better preserves biologically relevant morphology. Both MSAC and SBCache share the same trained neural spectral decomposer to efficiently extract ordered spectral bases without costly numerical eigendecomposition.

2.1 Neural Spectral Decomposer

Numerical eigendecomposition is prohibitively expensive when repeatedly applied during generator training or diffusion inference. Existing neural spectral methods, including SpectralNet, SpecNet2, and SBL (Shaham et al., 2018; Chen et al., 2022; Sun et al., 2026), avoid per-sample eigendecomposition but learn spectral embeddings through unsupervised clustering or eigensolver objectives rather than regressing ordered, per-image numerical bases. Figure 3 qualitatively compares selected predicted basis maps with numerical spectral targets. Further quantitative comparisons are provided in Appendix C.5. We instead amortize eigendecomposition by learning directly from numerical spectral targets generated offline.

We introduce a neural spectral decomposer $\mathcal{S}_\phi : \mathbb{R}^{N\times d} \to \mathbb{R}^{N\times K}$, which maps the patch-feature to an orthonormal spectral basis: $U = \mathcal{S}_\phi(F) \quad \text{s.t.} \quad U^\top U = I_K$. The network operates on the regular $H_p \times W_p$ patch layout. The predicted basis maps are orthogonalized using a differentiable reduced QR factorization. $\mathcal{S}_\phi$ denotes the complete mapping, including this orthogonalization step.

We use numerical eigendecomposition as offline supervision for the decomposer. Let $F \in \mathbb{R}^{N\times d}$ denote a patch-feature matrix, where N is the number of patches. From F, we construct a positive cosine-affinity graph and denote the corresponding normalized graph Laplacian by L (Von Luxburg, 2007). The first $K+1$ numerical eigenpairs, ordered by ascending eigenvalues, are computed as $(U^\star, \Lambda^\star) = \mathrm{Eig}_{K+1}(L)$. Since the smallest eigenpair corresponds to the trivial zero-frequency mode of the normalized graph Laplacian, we discard the trivial eigenpair and retain the subsequent K nontrivial eigenpairs. After re-indexing, we denote them by $U^\star = [u_1^\star, \dots, u_K^\star] \in \mathbb{R}^{N\times K}$

and $\lambda^\star = [\lambda_1^\star, \ldots, \lambda_K^\star]^\top$. The corresponding diagonal eigenvalue matrix is $\Lambda^\star = \mathrm{Diag}(\lambda^\star)$. The construction of L and the graph-frequency interpretation of the eigenpairs are detailed in the supplementary material. The numerical basis $U^\star$ provides the training target, and K controls the retained graph-frequency bandwidth.

The direct basis-wise loss $\|U - U^\star\|_F^2$ is unsuitable because numerical and predicted eigenvectors may use opposite signs. Repeated eigenvalues admit arbitrary orthogonal rotations within the associated eigenspaces, and closely spaced eigenvectors can be unstable under perturbations (Lim et al., 2023; Yang et al., 2026). We therefore align the retained spectral subspaces using the normalized projection distance on the Grassmann manifold (Bendokat et al., 2024). The subspace-alignment loss $\mathcal{L}_{\mathrm{sub}}$ is defined as:

$$\mathcal{L}_{\mathrm{sub}}(U, U^\star) = 1 - \frac{1}{K}\left\|U^\top U^\star\right\|_F^2 = \frac{1}{K}\sum_{k=1}^{K}\sin^2\theta_k, \tag{1}$$

where $\theta_1, \ldots, \theta_K$ are the principal angles between $\mathrm{span}(U)$ and $\mathrm{span}(U^\star)$. This objective is invariant to sign flips, permutations, and orthogonal rotations of the basis vectors.

The subspace-alignment loss constrains the predicted bases to span the numerical target subspace. Nevertheless, invariance to rotations and column permutations within the aligned subspace leaves individual basis correspondence and low-to-high-frequency ordering unresolved. We therefore construct the numerical-target Rayleigh matrix $C = U^\top U^\star \Lambda^\star U^{\star\top} U \in \mathbb{R}^{K\times K}$. The diagonal entries of C describe the Rayleigh energies of the predicted basis vectors, and the off-diagonal entries of C measure coupling between different spectral directions. The off-diagonal loss $\mathcal{L}_{\mathrm{off}}$ is defined as:

$$\mathcal{L}_{\mathrm{off}}(C) = \frac{1}{K^2}\left\|\mathrm{off}(C)\right\|_F^2, \tag{2}$$

where $\mathrm{off}(C)$ denotes the matrix obtained by setting the diagonal entries of C to zero. We further align the Rayleigh energies with the ordered numerical eigenvalues. The spectral-ordering loss $\mathcal{L}_{\mathrm{ord}}$ is defined as:

$$\mathcal{L}_{\mathrm{ord}}(C, \lambda^\star) = \frac{1}{K}\left\|\mathrm{diag}(C) - \lambda^\star\right\|_1. \tag{3}$$

The off-diagonal term separates mixed spectral directions, and the ordering term associates them with ascending graph frequencies. Both terms remain invariant to global sign flips.

The spectral-decomposer loss $\mathcal{L}_S$ is defined as:

$$\mathcal{L}_S = \mathcal{L}_{\mathrm{sub}}(U, U^\star) + \alpha_{\mathrm{off}}\mathcal{L}_{\mathrm{off}}(C) + \alpha_{\mathrm{ord}}\mathcal{L}_{\mathrm{ord}}(C, \lambda^\star). \tag{4}$$

Numerical eigendecomposition is used only to generate offline spectral targets from the training feature graphs and requires no manual annotations. In all experiments, we set $K = 16$. We use $\alpha_{\mathrm{off}} = 0.2$ and $\alpha_{\mathrm{ord}} = 1$. We find that retaining 16 nontrivial spectral bases provides a favorable balance across the evaluated staining and caching metrics. A detailed analysis of the effect of K is provided in the supplementary material.

2.2 Multi-Frequency Spectral-Aware Consistency for Virtual Staining

Let $x \in \mathbb{R}^{H\times W\times C_x}$ denote a source-domain microscopy image and $y \in \mathbb{R}^{H\times W\times C_y}$ denote the corresponding ground truth staining. A virtual-staining model G_θ predicts target-domain image $\hat{y} = G_\theta(x)$. Because different staining channels represent distinct subcellular structures, we compute spectral consistency independently for each channel. For the c-th target channel, the staining-image decomposer produces $U_{y_c} = \mathcal{S}_\phi(\mathcal{E}_\Phi(y_c)) \in \mathbb{R}^{N\times K}$, and the generated channel is represented as $U_{\hat{y}_c} = \mathcal{S}_\phi(\mathcal{E}_\Phi(\hat{y}_c))$. The same frozen encoder $\mathcal{E}_\Phi$ and decomposer $\mathcal{S}_\phi$ are shared across all channels.

For each channel, we define the MSAC metric as the normalized projection distance between the generated and target spectral subspaces:

$$d_{\mathrm{MSAC}}(\hat{y}_c, y_c) = 1 - \frac{1}{K}\left\|U_{\hat{y}_c}^\top U_{y_c}\right\|_F^2. \tag{5}$$

The channelwise measure in Eq. (5) evaluates multi-frequency structural consistency for the c-th stain by comparing the subspaces jointly spanned by the retained graph-spectral modes. A lower

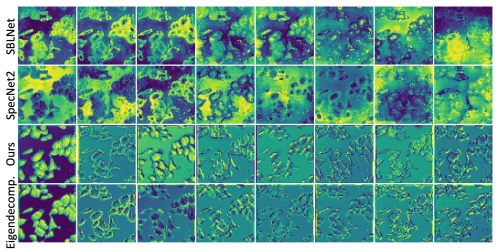


Figure 3: Qualitative comparison of different spectral decomposition methods. For SBLNet and SpecNet2, the predicted spectral bases are first sorted in ascending order of their estimated Dirichlet energies. We then visualize eight selected spectral bases and compare them with numerical eigendecomposition and our predicted bases.

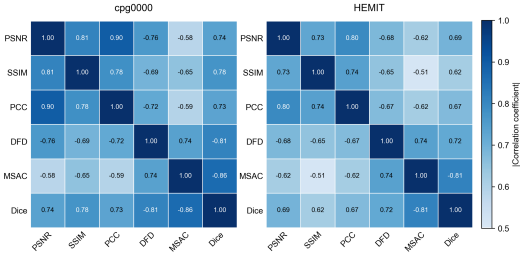


Figure 4: Correlation heatmaps between image quality metrics and downstream biological quality on the cpg0000 and HEMIT datasets. Each observation corresponds to one test image and one target channel. Predictions are generated by JiT-UNet without MSAC supervision

value indicates stronger agreement between the generated and target multi-frequency spectral subspaces.

Beyond evaluation, MSAC can also provide differentiable supervision for virtual-staining models. The MSAC loss $\mathcal{L}_{\mathrm{MSAC}}$ is defined as the average channelwise MSAC metric:

$$\mathcal{L}_{\mathrm{MSAC}} = \frac{1}{C_y} \sum_{c=1}^{C_y} d_{\mathrm{MSAC}}(\hat{y}_c, y_c). \tag{6}$$

During generator training, $\mathcal{E}_\Phi$ and $\mathcal{S}_\phi$ remain frozen. Gradients through the generated-image branch are retained to update G_θ. The target branch can be detached or precomputed. The overall staining loss $\mathcal{L}_{\mathrm{stain}}$ is defined as:

$$\mathcal{L}_{\mathrm{stain}} = \mathcal{L}_{\mathrm{origin}} + \alpha_{\mathrm{MSAC}} \mathcal{L}_{\mathrm{MSAC}}, \tag{7}$$

where $\mathcal{L}_{\mathrm{origin}}$ is the original model-specific objective. This plug-and-play formulation can be integrated into image translation and generation models with diverse architectures. We set $\alpha_{\mathrm{MSAC}} = 0.01$ in all experiments. Detailed implementation and experimental settings are provided in the supplementary material.

2.3 Spectral-Basis Cache for Diffusion Models

We further apply the neural spectral decomposer across diffusion timesteps and introduce SBCache. SBCache evaluates adjacent-step redundancy after graph-spectral filtering. SBCache requires no modification to the diffusion backbone or training procedure and serves as a plug-and-play module during inference.

Let y_t denote the latent variable at denoising step t. We extract the timestep-modulated input feature $F_t = \chi(y_t, t) \in \mathbb{R}^{N \times C}$, where χ is the fixed feature mapping used by the original cache policy, such as the pre-attention input of the first transformer block. Computing the cache criterion at this early location avoids most denoiser computation when the cached output is reused.

Offline Spectral-Prior Estimation. We run the original diffusion model without caching and collect $\mathcal{D}_{\mathrm{feat}} = \left\{ F_t^{(m)} \mid m = 1, \ldots, M \text{ and } t \in \mathcal{T} \right\}$, where m indexes training samples and $\mathcal{T}$ denotes the set of diffusion timesteps from which F_t is extracted. The same timestep set, feature location, and normalization are used during collection and inference.

The neural spectral decomposer $\mathcal{S}_\phi$ predicts ordered orthonormal bases for the collected features: $U_t = \mathcal{S}_\phi(F_t) = [u_{t,1}, \ldots, u_{t,K}] \in \mathbb{R}^{N \times K}$ s.t. $U_t^\top U_t = I_K$. SBCache therefore requires neither an additional decomposer-training stage nor online eigendecomposition or basis sorting.

For each $k = 1, \ldots, K$, we estimate a dataset-level spectral energy prior from the collected features:

$$S_F(k) = \frac{1}{M|\mathcal{T}|C} \sum_{m=1}^{M} \sum_{t \in \mathcal{T}} \left\| u_{t,k}^{(m)\top} F_t^{(m)} \right\|_2^2. \tag{8}$$

Table 1: MSAC-enhanced staining evaluation. All values are averaged over target channels. Bold denotes the best value in each metric column.

Method	cpg0000-jump-pilot							HEMIT					
	Stain				Seg.		Cnt.	Stain				Seg.	
	PCC ↑	PSNR ↑	SSIM ↑	MSAC ↓	Dice ↑	IoU ↑	MAE ↓	PCC ↑	PSNR ↑	SSIM ↑	MSAC ↓	Dice ↑	IoU ↑
ResUNet	.810	18.749	.657	.564	.554	.398	5.74	.675	18.538	.778	.689	.227	.141
$+\mathcal{L}_{\text{MSAC}}$	.847	19.186	.713	.460	.602	.425	4.18	.658	18.376	.759	.570	.540	.417
PyramidPix2Pix	.824	19.189	.682	.529	.567	.401	5.01	.712	17.472	.688	.668	.384	.320
$+\mathcal{L}_{\text{MSAC}}$	.845	19.305	.707	.436	.586	.413	4.33	.728	17.210	.689	.582	.433	.346
DiffStain	.866	19.774	.732	.444	.593	.426	1.95	.744	21.277	.687	.554	.542	.421
$+\mathcal{L}_{\text{MSAC}}$	**.874**	20.324	.744	.419	.632	.470	1.20	.800	23.437	.837	.438	.599	.439
StainDiffuser	.860	20.415	.711	.473	.604	.430	1.26	.731	21.141	.748	.537	.529	.411
$+\mathcal{L}_{\text{MSAC}}$	.860	20.358	.708	.412	.629	.465	.95	.798	22.106	.809	**.431**	.597	**.465**
JiT-UNet	.856	19.477	.692	.497	.585	.414	1.85	.787	22.924	.605	.625	.570	.403
$+\mathcal{L}_{\text{MSAC}}$	.868	**20.992**	**.749**	**.396**	**.638**	**.478**	**.75**	**.806**	**23.708**	**.842**	.436	**.609**	.449

The term $\|u_{t,k}^{(m)\top} F_t^{(m)}\|_2^2$ measures the projected energy of $F_t^{(m)}$ at the k-th spectral-basis frequency. Squared magnitudes make $S_F(k)$ invariant to eigenvector sign flips. The prior is computed once and fixed during inference.

Graph-Spectral Filtering. At timestep t, the shared decomposer $\mathcal{S}_\phi$ produces U_t for F_t. Given the spectral energy prior $S_F(k)$, we introduce a timestep-aware spectral gain to characterize how each graph-frequency band is modulated under the current diffusion noise level. Specifically, with $x_t = a_t x_0 + b_t \epsilon$, the gain of the k-th spectral band is defined as $g_{t,k} = a_t S_F(k)/(a_t^2 S_F(k) + b_t^2)$. We use this Wiener-inspired gain as a schedule-dependent weighting heuristic for nonlinear denoiser features, rather than assuming an exact linear noise model in feature space. $S_F(k)$ describes the intrinsic energy distribution of graph-frequency bands, and the timestep coefficients (a_t, b_t) determine their noise-dependent modulation. Since the overall gain magnitude varies across timesteps, we normalize the gains to obtain the final spectral weights: $w_{t,k} = g_{t,k} / \left(\frac{1}{K}\sum_{\ell=1}^{K} g_{t,\ell}\right)$ s.t. $\frac{1}{K}\sum_{k=1}^{K} w_{t,k} = 1$. This normalization prevents timestep-dependent gain scales from dominating the cache distance.

Let $\mathbf{w}_t = [w_{t,1}, \ldots, w_{t,K}]^\top$. The filtered feature is

$$\widetilde{F}_t = U_t \,\mathrm{Diag}(\mathbf{w}_t) U_t^\top F_t. \tag{9}$$

The projection $U_t^\top F_t$ decomposes the feature into the retained spectral-frequency components. Each frequency component is weighted according to the noise level and the corresponding empirical spectral energy before reconstruction in the token space.

Cache Distance And Refresh Rule. We retain the relative ℓ_1 distance and accumulated-threshold rule of SeaCache. The SBCache distance and its accumulator are

$$\widetilde{\Delta}_t = \frac{\|\widetilde{F}_t - \widetilde{F}_{t+1}\|_1}{\|\widetilde{F}_{t+1}\|_1 + \xi}, \qquad A \leftarrow A + \widetilde{\Delta}_t, \tag{10}$$

where ξ is a small constant. If $A \le \delta$, the cached denoiser output is reused. Otherwise, a full denoiser evaluation is performed, the cache is updated, and A is reset to zero.

The threshold δ controls the speed–fidelity trade-off. SBCache changes only the representation used to estimate redundancy and can therefore be viewed as a drop-in replacement for SeaCache's fixed Fourier representation with a learned graph-spectral representation.

3 Experiments

3.1 Datasets and Metrics

Datasets. We evaluate on two paired virtual-staining benchmarks. The cpg0000-jump-pilot subset of the Cell Painting Gallery (Weisbart et al., 2024) contains brightfield images of A549 and U2OS cells under chemical and genetic perturbations. Each sample has three brightfield input planes and five Cell Painting fluorescence targets corresponding to DNA, endoplasmic reticulum (ER), RNA,

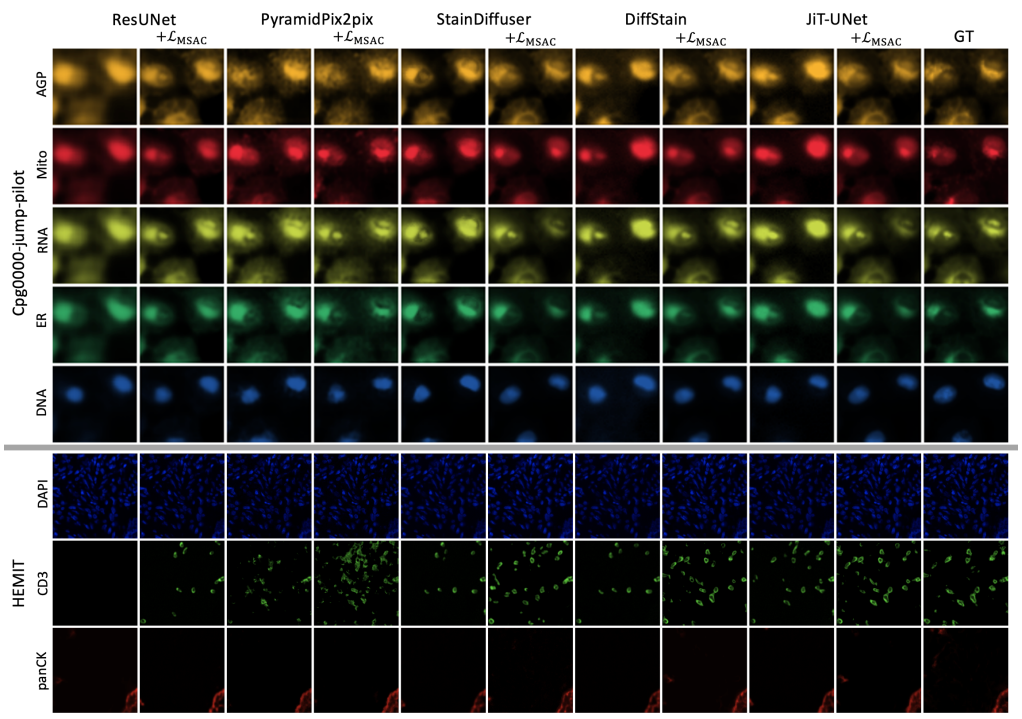


Figure 5: Qualitative comparison on both datasets. For each backbone, the unlabeled result denotes the original prediction and is followed by the corresponding $+\mathcal{L}_{\text{MSAC}}$ result. The final column is the ground truth. The upper five rows show the AGP, Mito, RNA, ER, and DNA channels of cpg0000. The lower three rows show DAPI, CD3, and panCK of HEMIT. On cpg0000, the MSAC training objective restores fine-grained structures while baseline predictions merge into blobs. On HEMIT, the CD3 row is particularly discriminative: the $+\mathcal{L}_{\text{MSAC}}$ variants recover many T cells missed by the baseline models and better match the ground-truth distribution.

Table 2: Controlled degradation analysis under matched PSNR. Structural perturbations are compared with Gaussian-noise controls whose PSNR is matched to the corresponding structural degradation. For structural removal, r denotes the ratio of removed connected components; σ_b and σ_n denote the boundary blur strength and Gaussian noise magnitude, respectively.

PSNR 35.0					PSNR 27.0				
Degradation	PCC↑	SSIM↑	DFD↓	MSAC↓	Degradation	PCC↑	SSIM↑	DFD↓	MSAC↓
Gaussian noise ($\sigma_n = 0.0217$)	.985	.779	.035	.025	Gaussian noise ($\sigma_n = 0.0496$)	.954	.546	.076	.141
Small structure removal ($r = 29.8\%$)	.982	.912	.059	.121	Small structure removal ($r = 47.3\%$)	.933	.961	.136	.549
Boundary blur ($\sigma_b = 1.61$)	.987	.983	.061	.075	Boundary blur ($\sigma_b = 3.82$)	.931	.931	.142	.425

actin/Golgi/plasma membrane (AGP), and mitochondria (Mito). We use a fixed image-level split containing 4,000 training, 200 validation, and 500 test images. The split is generated at the well level to avoid leakage among highly correlated microscopy acquisitions and the same split is shared by all methods and ablation studies. HEMIT (Bian et al., 2025) contains cellularly aligned H&E and multiplex-immunohistochemistry patches. Its three targets are DAPI, pan-cytokeratin (panCK), and CD3. We follow the standard split of 3,717/630/945 training/validation/test images.

Metrics. We evaluate synthesized stains using PCC, PSNR, and SSIM, which measure intensity fidelity and image-level structural similarity, together with the proposed MSAC metric for multi-frequency structural consistency in the graph-spectral space. Higher PCC, PSNR, and SSIM and lower MSAC indicate better staining quality.

To determine whether image fidelity transfers to biological analysis, we apply the same standard CellProfiler pipeline (McQuin et al., 2018) independently to every real and synthesized stain to obtain channel-wise binary masks. We compare the resulting masks using Dice and IoU. On cpg0000, we additionally identify nuclei from the DNA channel and report the mean absolute error (MAE) between synthesized-image and real-image cell counts.

3.2 Analysis of Structure-Aware Metric Reliability

Conventional fidelity metrics may fail to reflect biologically meaningful structural changes when the affected regions occupy only a small fraction of the image. To evaluate whether MSAC captures structural information beyond pixel-level similarity, we perform controlled degradation and correlation analysis.

For controlled degradation, we introduce structural perturbations including small-structure removal and boundary blurring, and compare them with PSNR-matched Gaussian-noise controls. As shown in Table 2, at matched PSNR, both DFD and MSAC assign larger discrepancies to small-structure removal and boundary blurring than to Gaussian noise. PCC changes relatively little, whereas SSIM penalizes Gaussian noise more strongly than the tested structural perturbations. These results characterize the metrics' different responses to controlled distortions.

We further compare MSAC with a DINOv3-based feature distance (DFD). Given two images $\hat{y}$ and y, DFD is defined as

$$\mathrm{DFD}(\hat{y}, y) = \frac{1}{CN} \sum_{c=1}^{C} \left\| \overline{\mathcal{E}}_\phi(\hat{y}_c) - \overline{\mathcal{E}}_\phi(y_c) \right\|_F^2, \tag{11}$$

where $\overline{\mathcal{E}}_\phi$ denotes the token-wise normalized features extracted from a frozen DINOv3 encoder. Lower DFD indicates higher feature-level similarity.

To avoid potential circularity caused by MSAC-based optimization, we perform correlation analysis using JiT-UNet predictions without MSAC supervision. Each image–channel pair constitutes one observation. As shown in Figure 4, MSAC exhibits the strongest absolute correlation with CellProfiler-derived Dice among the compared metrics on both datasets, suggesting closer alignment with this downstream structural proxy.

Together, these results characterize MSAC's responses to the tested distortions and its association with a downstream structural proxy.

3.3 MSAC-enhanced Staining Evaluation

We evaluate MSAC-enhanced training on five representative virtual-staining backbones, including ResUNet (Diakogiannis et al., 2020), PyramidPix2Pix (Liu et al., 2022), DiffStain (Han et al., 2026), StainDiffuser (Kataria et al., 2024), and a JiT model (Li & He, 2026). For each backbone, we add MSAC as an additional spectral consistency objective and keep the original training framework unchanged. Detailed training settings are provided in the supplementary material.

Table 1 shows that, with the MSAC training objective, downstream structural measures improve consistently across the evaluated architectures and modalities, while conventional fidelity metrics exhibit mixed changes. At the reported precision, 23 of the 30 conventional metrics comparisons improve and one is unchanged after adding MSAC loss. More importantly, all 25 downstream comparisons improve: Dice and IoU increase for every backbone on both datasets, and DNA-count MAE decreases for every backbone on cpg0000.

The qualitative results in Figure 5 further demonstrate that MSAC recovers fine-grained biological structures that are often blurred or merged in baseline predictions. On cpg0000, MSAC-enhanced models restore fine spatial patterns in Mito, AGP, RNA and ER channels, producing structures closer to the ground truth. On HEMIT, the CD3 channel provides a particularly clear example, where MSAC recovers more T-cell regions missed by baseline models and better matches the target spatial distribution.

The results also reveal that improvements in biological structure are not necessarily accompanied by uniform gains in intensity fidelity. For example, StainDiffuser exhibits a slight decrease in cpg0000 PSNR, and ResUNet shows lower HEMIT intensity scores, yet both obtain substantial improvements in downstream mask agreement. These results indicate that MSAC provides structural consistency complementary to conventional metrics.

Table 3: SBCache comparison on cpg0000-jump-pilot. Bold denotes the best value within each refresh-ratio group.

Method	δ	Ref. ↓	Lat. ↓	PCC ↑	PSNR ↑	SSIM ↑	MSAC ↓
Uncached	–	100	1.00	.868	20.992	.749	.396
TeaCache	.21	50	.56	.834	20.123	.714	.433
SeaCache	.19	50	.53	.838	20.307	.719	.423
SBCache (ours)	.17	50	**.52**	**.843**	**20.476**	**.724**	**.414**
TeaCache	.42	30	.39	.811	19.209	.677	.472
SeaCache	.41	30	.37	.822	19.554	.689	.455
SBCache (ours)	.39	30	**.36**	**.833**	**19.907**	**.701**	**.433**

Table 4: Ablation of pretrained models and spectral decomposition on JiT-UNet using cpg0000-jump-pilot.

Model		Stain				Seg.		Cnt.
		PCC ↑	PSNR ↑	SSIM ↑	MSAC ↓	Dice ↑	IoU ↑	MAE ↓
CLIP	$\mathcal{L}_{\text{perc}}$	.859	19.842	.708	.476	.596	.426	1.55
	$\mathcal{L}_{\text{MSAC}}$	.864	20.532	.734	.423	.621	.457	.98
DINOv2	$\mathcal{L}_{\text{perc}}$	.861	20.025	.716	.460	.603	.434	1.40
	$\mathcal{L}_{\text{MSAC}}$	.866	20.741	.741	.410	.629	.468	.86
DINOv3	$\mathcal{L}_{\text{perc}}$	.862	20.146	.721	.451	.607	.439	1.31
	$\mathcal{L}_{\text{MSAC}}$	**.868**	**20.992**	**.749**	**.396**	**.638**	**.478**	**.75**

3.4 Comparative Results of SBCache

We compare SBCache with TeaCache and SeaCache using 50-step sampling with JiT-UNet trained with MSAC. Since different caching distances have different numerical scales, we adjust each method's threshold δ to achieve the same refresh ratios of 50% and 30%, rather than comparing identical thresholds. The refresh ratio denotes the percentage of sampling steps that execute the full denoiser, and relative latency is normalized by uncached inference.

Table 3 reports the results on cpg0000-jump-pilot. At both matched refresh ratios, SBCache achieves better staining quality across all reported metrics and lower relative latency than the caching baselines. Compared with uncached inference, however, SBCache has lower PCC, PSNR, and SSIM and higher MSAC at both refresh ratios. These results support the effectiveness of the learned graph-spectral distance for cache decisions, yielding a favorable quality–efficiency trade-off under the same full-denoiser evaluation budget.

3.5 Effectiveness of Spectral Decomposition

To disentangle the contribution of pretrained visual representations from that of spectral decomposition, we compare features generated by CLIP (Radford et al., 2021), DINOv2 (Oquab et al., 2024), and DINOv3 (Siméoni et al., 2026) using JiT-UNet. All variants use the same feature-map resolution, spectral-decomposer architecture, number of bases, and training schedule.

For each encoder variant, $\mathcal{L}_{\text{perc}}$ follows the normalized feature-matching formulation in Eq. 11, using the corresponding frozen CLIP-, DINOv2-, or DINOv3-based representation. DFD denotes the DINOv3-specific instance used for metric evaluation. For a fair comparison, all MSAC scores in Table 4 are computed using the same default DINOv3-based evaluator.

The results in Table 4 separate the benefits of pretrained representations and spectral decomposition. Direct feature matching improves over JiT-UNet for all three pretrained encoders, while the corresponding MSAC-enhanced variants achieve larger gains in Dice, IoU, and counting accuracy. DINOv3 provides the strongest overall performance, and the consistent advantage of MSAC over direct feature matching with CLIP and DINOv2 demonstrates that the improvement originates from the proposed MSAC rather than a specific pretrained encoder.

4 Discussion and Conclusion

We presented Spectral-Aware Staining, a framework that leverages multi-frequency spectral representations to improve virtual-staining quality assessment and diffusion acceleration. MSAC provides a spectral consistency metric for assessing generated stains. By enforcing multi-frequency consistency, MSAC better preserves biologically relevant morphology. SBCache extends the learned spectral representation to diffusion inference, enabling more reliable redundancy estimation during denoising. Both MSAC and SBCache share the same trained neural spectral decomposer to efficiently extract the required spectral bases without costly numerical eigendecomposition. Experiments on public microscopy benchmarks and virtual-staining frameworks demonstrate that MSAC more effectively evaluates the quality of generated stains and that Spectral-Aware Staining generates high-quality stained images more efficiently.

AI USE STATEMENT

Generative AI tools were used during the preparation of this manuscript to assist with language editing and polishing, the retrieval and discovery of potentially relevant literature, and the drafting of selected passages and sections. All AI-generated or AI-assisted content was critically reviewed, verified, and revised by the authors. References suggested by AI tools were checked against the original publications before being cited. The authors made all final decisions regarding the technical content, methodology, experiments, analysis, and conclusions, and take full responsibility for the accuracy, originality, and integrity of the manuscript.

ETHICS STATEMENT

This work uses existing research datasets and does not involve new recruitment, intervention, or data collection from human participants. We follow the licenses and terms of use of the datasets, pretrained models, and external services used in this study. No personally identifiable information is used or released. The proposed method is intended for research purposes and has not been validated for direct clinical diagnosis or treatment.

REPRODUCIBILITY STATEMENT

Appendices B and C provide the graph construction, objective derivations, preprocessing, architectures, optimization settings, and cache implementation. Appendix D provides additional experimental results and parameter analyses. All datasets we used are publicly available. The training scripts and evaluation code will be released upon publication to facilitate replication.

REFERENCES

Bijie Bai, Xilin Yang, Yuzhu Li, Yijie Zhang, Nir Pillar, and Aydogan Ozcan. Deep learning-enabled virtual histological staining of biological samples. *Light: Science & Applications*, 12(1):57, 2023.

Thomas Bendokat, Ralf Zimmermann, and P.-A. Absil. A Grassmann manifold handbook: Basic geometry and computational aspects. *Advances in Computational Mathematics*, 50(1):6, 2024.

Chang Bian, Beth Phillips, Tim Cootes, and Martin Fergie. HEMIT: H&E to multiplex-immunohistochemistry image translation with dual-branch Pix2pix generator. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024 Workshops*, volume 15401, pp. 184–197, Cham, 2025. Springer Nature Switzerland.

Åke Björck and Gene H. Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, 27(123):579–594, 1973.

Ziyu Chen, Yingzhou Li, and Xiuyuan Cheng. SpecNet2: Orthogonalization-free spectral embedding by neural networks. In *Proceedings of Mathematical and Scientific Machine Learning*, volume 190 of *Proceedings of Machine Learning Research*, pp. 33–48. PMLR, 2022.

Eric M. Christiansen, Samuel J. Yang, D. Michael Ando, Ashkan Javaherian, Gaia Skibinski, Scott Lipnick, Elliot Mount, Alison O'Neil, Kevan Shah, Alicia K. Lee, Piyush Goyal, William Fedus, Ryan Poplin, Andre Esteva, Marc Berndl, Lee L. Rubin, Philip Nelson, and Steven Finkbeiner. In silico labeling: Predicting fluorescent labels in unlabeled images. *Cell*, 173(3):792–803.e19, 2018.

Jiwoo Chung, Sangeek Hyun, MinKyu Lee, Byeongju Han, Geonho Cha, Dongyoon Wee, Youngjun Hong, and Jae-Pil Heo. SeaCache: Spectral-evolution-aware cache for accelerating diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14283–14294, 2026.

Jan Oscar Cross-Zamirski, Elizabeth Mouchet, Guy Williams, Carola-Bibiane Schönlieb, Riku Turkki, and Yinhai Wang. Label-free prediction of Cell Painting from brightfield images. *Scientific Reports*, 12(1):10001, 2022.

Jan Oscar Cross-Zamirski, Praveen Anand, Guy Williams, Elizabeth Mouchet, Yinhai Wang, and Carola-Bibiane Schönlieb. Class-guided image-to-image diffusion: Cell Painting from bright-field images with class labels. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3802–3811. IEEE, 2023.

Foivos I. Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020.

Melanie Dohmen, Mark A. Klemens, Ivo M. Baltruschat, Tuan Truong, and Matthias Lenga. Similarity and quality metrics for MR image-to-image translation. *Scientific Reports*, 15(1):3853, 2025a.

Melanie Dohmen, Tuan Truong, Ivo M. Baltruschat, and Matthias Lenga. Five pitfalls when assessing synthetic medical images with reference metrics. In *Deep Generative Models*, volume 15224, pp. 150–159, Cham, 2025b. Springer Nature Switzerland.

Yikai Han, Jimao Jiang, and Yuru Pei. DiffStain: Conditioned diffusion-based semantic virtual staining with mask guidance. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15972, pp. 139–148, Cham, 2026. Springer Nature Switzerland.

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, pp. 6626–6637. Curran Associates, Inc., 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Justin Johnson, Alexandre Alahi, and Fei-Fei Li. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision – ECCV 2016*, volume 9906, pp. 694–711. Springer, 2016.

Tushar Kataria, Beatrice Knudsen, and Shireen Y. Elhabian. StainDiffuser: Multitask dual diffusion model for virtual staining. *arXiv preprint arXiv:2403.11340*, 2024.

Nicholas Konz, Yuwen Chen, Hanxue Gu, Haoyu Dong, and Maciej A. Mazurowski. Rethinking perceptual metrics for medical image translation. In *Medical Imaging with Deep Learning (MIDL) 2024 Short Papers*, pp. 1–5, 2024. URL https://openreview.net/forum?id=acNu3glOTh.

Tianbo Li, Zekun Shi, Jiaxi Zhao, and Min Lin. Amortized eigendecomposition for neural networks. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36115–36125, 2026.

Derek Lim, Joshua Robinson, Lingxiao Zhao, Tess Smidt, Suvrit Sra, Haggai Maron, and Stefanie Jegelka. Sign and basis invariant networks for spectral graph representation learning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Q-UHqMorzil.

Feng Liu, Shiwei Zhang, Xiaofeng Wang, Yujie Wei, Haonan Qiu, Yuzhong Zhao, Yingya Zhang, Qixiang Ye, and Fang Wan. Timestep embedding tells: It's time to cache for video diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7353–7363, 2025.

Shengjie Liu, Chuang Zhu, Feng Xu, Xinyu Jia, Zhongyue Shi, and Mulan Jin. BCI: Breast cancer immunohistochemical image generation through pyramid Pix2pix. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1815–1824, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Claire McQuin, Allen Goodman, Vasiliy Chernyshev, Lee Kamentsky, Beth A. Cimini, Kyle W. Karhohs, Minh Doan, Liya Ding, Susanne M. Rafelski, Derek Thirstrup, Winfried Wiegraebe, Shantanu Singh, Tim Becker, Juan C. Caicedo, and Anne E. Carpenter. CellProfiler 3.0: Next-generation image processing for biology. *PLoS Biology*, 16(7):e2005970, 2018.

Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8364–8375, 2022.

Hyun-Jic Oh, Junsik Kim, Zhiyi Shi, Yichen Wu, Yu-An Chen, Peter K. Sorger, Hanspeter Pfister, and Won-Ki Jeong. Virtual multiplex staining for histological images using a marker-wise conditioned diffusion model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(10): 8168–8176, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Antonio Ortega, Pascal Frossard, Jelena Kovačević, José M. F. Moura, and Pierre Vandergheynst. Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE*, 106 (5):808–828, 2018.

Fanhao Qiu, Yangyang Zhang, and Zhengxia Wang. Graph-semantic guided learning for virtual immunohistochemistry staining on consecutive histology sections. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8556–8564, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241. Springer, 2015.

Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pp. 15:1–15:10. ACM, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TIdIXIpzhoI.

Sourya Sengupta, Jianquan Xu, Phuong Nguyen, Frank J. Brooks, Yang Liu, and Mark A. Anastasio. On the utility of virtual staining for downstream applications as it relates to task network capacity. *Biomedical Optics Express*, 16(10):4224–4242, 2025.

Uri Shaham, Kelly Stanton, Henry Li, Boaz Nadler, Ronen Basri, and Yuval Kluger. SpectralNet: Spectral clustering using deep neural networks. In *International Conference on Learning Representations*, 2018.

David I. Shuman, Sunil K. Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 30(3):83–98, 2013.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michael Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=2NlGyqNjns. Featured Certification.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32211–32252. PMLR, 2023.

Diya Sun, Yuru Pei, and Tianbing Wang. Unsupervised deep spectral basis learning for generalized eigendecomposition and spectral embedding. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2026. Early Access.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Erin Weisbart, Ankur Kumar, John Arevalo, Anne E. Carpenter, Beth A. Cimini, and Shantanu Singh. Cell Painting Gallery: An open resource for image-based profiling. *Nature Methods*, 21 (10):1775–1777, 2024.

Zherui Yang, Tao Du, and Ligang Liu. Learning Laplacian eigenspace with mass-aware neural operators on point clouds. In *ACM SIGGRAPH 2026 Conference Papers*, pp. 6:1–6:11. ACM, 2026.

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

A Related Work

Microscopy Virtual Staining. Virtual staining aims to synthesize fluorescence or immunostaining images from label-free microscopy modalities, providing an alternative to physical staining. Early studies demonstrated that deep networks can recover biologically meaningful labels from transmitted-light images (Christiansen et al., 2018). Subsequent works extended virtual staining to Cell Painting and pathology applications, showing that generated images can support morphological profiling and multiplex analysis (Cross-Zamirski et al., 2022; Bian et al., 2025). Recently, diffusion-based models have further improved the fidelity and diversity of synthesized stains (Han et al., 2026; Kataria et al., 2024; Oh et al., 2026). GSGStain (Qiu et al., 2026) introduces graph-semantic guided learning for virtual immunohistochemistry staining on consecutive histology sections. Different from GSGStain, which imposes graph-semantic consistency directly for staining supervision, our MSAC measures similarity in learned multi-frequency graph-spectral subspaces and additionally serves as an evaluation measure.

Evaluation Metrics for Medical Image Synthesis. Image synthesis quality is commonly assessed using PSNR, SSIM, LPIPS, and FID (Wang et al., 2004; Zhang et al., 2018; Heusel et al., 2017). Although effective for general image comparison, these metrics may not faithfully reflect biologically meaningful structures, especially sparse or fine-grained patterns in medical images. Recent studies have reported similar mismatches between conventional image-quality metrics and task-specific structural fidelity in medical imaging applications (Dohmen et al., 2025a;b; Konz et al., 2024). Unlike these metrics, which mainly measure intensity, perceptual distance, or distribution similarity, our method evaluates structural consistency between retained multi-frequency spectral subspaces.

Amortized Eigendecomposition. Li et al. (2024) introduce amortized eigendecomposition for neural networks, replacing repeated exact eigendecomposition during training with QR orthogonalization and an eigen loss. Different from amortized eigendecomposition, our goal is not general eigensolver acceleration but an ordered spectral representation tailored to structural comparison and diffusion reuse decisions.

Diffusion Acceleration And Caching. Diffusion acceleration has been explored through fast samplers, distillation, quantization, and architectural optimization (Salimans & Ho, 2022; Song et al., 2023). Recently, training-free caching methods reduce inference cost by reusing intermediate computations between adjacent denoising steps. TeaCache estimates redundancy using timestep-aware feature differences and selectively reuses cached outputs (Liu et al., 2025), while SeaCache introduces spectral filtering to obtain more informative cache decisions from diffusion trajectories (Chung et al., 2026). Nevertheless, existing approaches measure redundancy using predefined feature spaces or fixed spectral filters. Our SBCache instead constructs a learned graph-spectral representation and evaluates redundancy along ordered structural frequency bands, providing a structure-aware cache criterion without modifying the denoising model.

B More Method Details

B.1 Numerical Spectral Decomposition

Using the feature matrix F defined in the main paper, we normalize each row as $\bar{f}_i = f_i/(\|f_i\|_2+\epsilon)$ and the normalized feature matrix $\bar{F} = [\bar{f}_1^\top, \dots, \bar{f}_N^\top]^\top$. The positive cosine-affinity graph is $[A]_{ij} = \max(\bar{f}_i^\top \bar{f}_j, 0)$. Positive-thresholded feature affinities are commonly used to build Laplacian matrix from patch features (Melas-Kyriazi et al., 2022). With $d_i = \sum_j [A]_{ij}$ and the degree matrix $D = \mathrm{Diag}(d_1, \dots, d_N)$, the symmetric normalized graph Laplacian is $L = I_N - D^{-1/2} A D^{-1/2}$. L is real, symmetric, and positive semidefinite (Von Luxburg, 2007).

Let the ordered eigendecomposition be $L = \widetilde{U}\widetilde{\Lambda}\widetilde{U}^\top$, with $0 \le \widetilde{\lambda}_1 \le \cdots \le \widetilde{\lambda}_N$. We compute the first $K+1$ eigenpairs, discard the trivial first pair, and re-index pairs $2{:}K+1$ using the notation of the main paper $(U^\star, \Lambda^\star)$.

Graph-Frequency Ordering. For any graph signal $z \in \mathbb{R}^N$, the normalized graph Dirichlet energy is

$$z^\top L z = \frac{1}{2} \sum_{i,j=1}^{N} [A]_{ij} \left(\frac{z_i}{\sqrt{d_i}} - \frac{z_j}{\sqrt{d_j}} \right)^2 . \tag{12}$$

Low energy therefore corresponds to slow variation across strongly connected patches. Each retained unit-norm eigenvector satisfies ${u_k^\star}^\top L u_k^\star = \lambda_k^\star$, where the associated eigenvalue quantifies its graph smoothness. Consequently, the retained eigenvectors are ordered by increasing eigenvalue, progressing from smoother to higher-index modes within the retained part of the graph spectrum, following the standard graph-frequency interpretation of Laplacian eigenvectors (Shuman et al., 2013). The numerical eigen-decompositions of the training graphs are conducted once and serve as supervision in neural spectral decomposer training process. Virtual-staining training never differentiates through numerical eigen-decomposition.

B.2 Subspace Loss

Reduced QR and numerical eigendecomposition give $U^\top U = I_K$ and ${U^\star}^\top U^\star = I_K$. Define the projectors $P = UU^\top$ and $P^\star = U^\star {U^\star}^\top$. The normalized projection-distance-based subspace loss is defined as:

$$\mathcal{L}_{\text{sub}}(U, U^\star) = \frac{1}{2K} \|P - P^\star\|_F^2 . \tag{13}$$

Expanding the squared norm gives

$$\|P - P^\star\|_F^2 = \operatorname{tr}(P^2) + \operatorname{tr}((P^\star)^2) - 2\operatorname{tr}(PP^\star). \tag{14}$$

Both projectors have rank K, so the traces of their squares are K. Cyclic invariance of the trace yields

$$\operatorname{tr}(PP^\star) = \operatorname{tr}\left(U^\top U^\star {U^\star}^\top U\right) = \|U^\top U^\star\|_F^2 . \tag{15}$$

Substitution into Eq. (13) yields the subspace loss in the main paper. Thus, implementation requires only the $K \times K$ overlap $U^\top U^\star$, reducing the additional overlap-storage cost from $O(N^2)$ to $O(K^2)$; storing the input bases still requires $O(NK)$ memory.

Consider the singular value decomposition of $U^\top U^\star$,

$$U^\top U^\star = Q \operatorname{Diag}(\sigma_1, \ldots, \sigma_K) R^\top . \tag{16}$$

The singular values satisfy $\sigma_k = \cos\theta_k$, where $\theta_k \in [0, \pi/2]$ are the principal angles between $\operatorname{span}(U)$ and $\operatorname{span}(U^\star)$ (Björck & Golub, 1973; Bendokat et al., 2024). Hence

$$\|U^\top U^\star\|_F^2 = \sum_{k=1}^{K} \cos^2\theta_k . \tag{17}$$

Substitution recovers the principal-angle form reported in the main paper. The loss is zero exactly when the retained subspaces coincide with those from the numerical eigen-decomposition.

B.3 Rayleigh Objectives

Let $M = U^\top U^\star \in \mathbb{R}^{K \times K}$. The Rayleigh matrix in the main paper can be evaluated entirely in the spectral embedding space:

$$C = M \Lambda^\star M^\top . \tag{18}$$

Given the estimated spectral basis $U = [u_1, \ldots, u_K]$ and the ground truth $U^\star = [u_1^\star, \ldots, u_K^\star]$, the (i, j)-th entry of the Rayleigh matrix (C) is defined as:

$$(C)_{ij} = \sum_{m=1}^{K} \lambda_m^\star (u_i^\top u_m^\star)(u_j^\top u_m^\star). \tag{19}$$

In particular, the i-th entry on the diagonal is defined as: $(C)_{ii} = \sum_m \lambda_m^\star (u_i^\top u_m^\star)^2$, indicating the Rayleigh energy of predicted spectral basis u_i under the retained numerical operator $U^\star \Lambda^\star {U^\star}^\top$. An off-diagonal entry is the corresponding coupling between two predicted spectral bases.

Off-Diagonal Separation. The squared off-diagonal Frobenius norm of the Rayleigh matrix (C) expands as

$$\| \operatorname{off}(C)\|_F^2 = \sum_{\substack{i,j=1 \\ i \neq j}}^{K} C_{ij}^2. \tag{20}$$

Suppose the subspace loss has aligned the retained subspaces. Then $U = U^\star R$ for an orthogonal $R \in \mathbb{R}^{K \times K}$, and Eq. (18) becomes $C = R^\top \Lambda^\star R$. If the retained eigenvalues are distinct, $\operatorname{off}(C) = 0$ only when R is a signed permutation. Thus, $\mathcal{L}_{\text{off}}$ removes rotations that mix graph frequencies but does not determine their column order. For a repeated eigenvalue, rotations within the corresponding eigenspace remain unpenalized, as required by the intrinsic non-uniqueness of its eigenvectors.

Low-to-High-Frequency Ordering. Let $r = \operatorname{diag}(C)$. The ordering term in the main paper compares this vector with $\lambda^\star$ using an ℓ_1 distance. After off-diagonal separation, a signed permutation gives $r = (\lambda^\star_{\pi(1)}, \ldots, \lambda^\star_{\pi(K)})^\top$. Since the target eigenvalues are ascending, $\mathcal{L}_{\text{ord}}$ selects $\pi(i) = i$. Permutations or rotations inside a repeated-eigenvalue block remain equivalent because they do not change the ordered eigenvalue vector.

The objectives act on complementary parts of C:

$$C - \Lambda^\star = \operatorname{off}(C) + \operatorname{Diag}(\operatorname{diag}(C) - \lambda^\star)\,. \tag{21}$$

The two right-hand matrices are orthogonal under the Frobenius inner product, so

$$\|C - \Lambda^\star\|_F^2 = \| \operatorname{off}(C)\|_F^2 + \| \operatorname{diag}(C) - \lambda^\star\|_2^2. \tag{22}$$

The implemented objective uses a squared Frobenius penalty for the off-diagonal component and a robust ℓ_1 penalty for the diagonal energies. Finally, replacing U by UD_s, with $D_s = \operatorname{Diag}(s_1, \ldots, s_K)$ and $s_i \in \{-1, 1\}$, transforms C into $D_s C D_s$ but leaves $\mathcal{L}_{\text{sub}}$, $\mathcal{L}_{\text{off}}$, and $\mathcal{L}_{\text{ord}}$ unchanged. The full objective therefore respects the unavoidable sign ambiguity of eigenvectors.

C Additional Implementation Details

C.1 Data Preprocessing

Microscopy images in cpg0000-jump-pilot have a spatial resolution of 1080×1080 pixels. Its three brightfield planes form the source domain, and its five fluorescence channels form the target domain. Each field is divided into a non-overlapping 2×2 grid, and each tile is resized to 256×256 by bilinear interpolation before it is passed to a staining model. HEMIT (Bian et al., 2025) provides paired 1024×1024 patches; H&E is the source domain and the DAPI, CD3, and panCK mIHC channels are the target domain. For both datasets, source and target images are scaled directly to $[0, 1]$. Synchronized horizontal and vertical flips, each with probability 0.5, are applied to source–target pairs during training only. The same resizing and intensity scaling, without random flips, are used for validation and testing.

MSAC operates on a higher-resolution feature grid than the staining backbones. Before feature distillation, spectral-target generation, or MSAC evaluation, each tile is resized to 512×512. The compact encoder then returns a 64×64 feature map for the neural spectral decomposer. This resize strategy defines the graph resolution; upsampling does not recover structures absent from the original image.

C.2 Distilled Feature Encoder

We use a frozen DINOv3 ViT-B/16 teacher (Siméoni et al., 2026) to train the compact feature encoder $\mathcal{E}_\Phi$ used by MSAC. Each input single-channel target-domain image is replicated to three channels before teacher inference. The student is a four-stage CNN with widths $(32, 64, 128, 192)$ at spatial scales $1/2$, $1/4$, $1/8$, and $1/16$. Each stage contains expanded depthwise-separable residual blocks with GroupNorm and GELU. A 3×3 convolution followed by a 1×1 projection maps the $1/16$ feature map to the teacher token dimension of 768; global average pooling followed by a linear layer predicts the teacher class token. The student's 192-channel internal representation is therefore projected to 768 channels before token-level distillation and is never compared directly

with the teacher representation. The 128-channel $1/8$ feature map, with spatial resolution 64×64, is retained as F for spectral decomposition and is not passed through the distillation projection head.

Let $\widetilde{T}_{\mathrm{p}}, \widetilde{S}_{\mathrm{p}} \in \mathbb{R}^{N_p \times 768}$ denote the teacher and student patch tokens before normalization. Their row-normalized versions are denoted as T_{p} and S_{p}, respectively. The class tokens are similarly normalized and denoted by t_{cls} and s_{cls}.

For a uniformly sampled set $\mathcal{Q}$ of at most 64 patch locations, we construct the pairwise patch-similarity matrices

$$\Gamma_T = T_{\mathrm{p},\mathcal{Q}} T_{\mathrm{p},\mathcal{Q}}^{\top}, \qquad \Gamma_S = S_{\mathrm{p},\mathcal{Q}} S_{\mathrm{p},\mathcal{Q}}^{\top}. \tag{23}$$

The patch-level feature matching loss and class-token alignment loss are defined as

$$\mathcal{L}_{\text{patch-mse}} = \frac{1}{N_p \times 768} \left\| \widetilde{S}_{\mathrm{p}} - \widetilde{T}_{\mathrm{p}} \right\|_F^2, \tag{24}$$

and

$$\mathcal{L}_{\mathrm{cls}} = 1 - s_{\mathrm{cls}}^{\top} t_{\mathrm{cls}}. \tag{25}$$

The patch-relation consistency term aligns the pairwise similarity structures between teacher and student representations. The complete distillation loss $\mathcal{L}_{\mathrm{dist}}$ is defined as:

$$\mathcal{L}_{\mathrm{dist}} = \mathcal{L}_{\text{patch-mse}} + 0.1\mathcal{L}_{\mathrm{cls}} + 0.2 \left\| \Gamma_S - \Gamma_T \right\|_F^2. \tag{26}$$

All five target channels are mixed when training the shared encoder. The same intensity scaling and random flips are used for all channels. We train for 300 epochs with AdamW (Loshchilov & Hutter, 2019), batch size 32, learning rate 10^{-3}, weight decay 5×10^{-2}, cosine learning-rate decay, and unit-norm gradient clipping. After distillation, $\mathcal{E}_{\Phi}$ remains frozen and in evaluation mode.

C.3 Neural Spectral Decomposer

Offline Numerical Eigen-Decomposition. The preprocessed 512×512 staining images are passed through the frozen encoder. Its retained feature map has shape $128 \times 64 \times 64$, giving $N = 4096$ graph nodes. We use the dense positive cosine-affinity graph of the training data and conduct numerical eigen-decomposition. We retain $K = 16$ nontrivial eigenpairs per graph, which serve as a supervision in neural spectral decomposer training.

Architecture and Optimization. The decomposer $\mathcal{S}_{\phi}$ receives the frozen 128-channel feature map. Two coordinate maps scaled to $[-1, 1]$ are concatenated to the input, followed by three 3×3 convolution–GELU layers with 384 hidden channels and a 1×1 convolution producing K raw basis maps. The resulting maps are passed through the reduced-QR layer described in the main paper.

We minimize the objective $\mathcal{L}_S$ in the main paper using the reported coefficients and the AdamW optimizer for 300 epochs, with the learning rate 10^{-4}, weight decay 10^{-2}, cosine learning-rate decay, and unit-norm gradient clipping. The batch size is set to 32. Both $\mathcal{E}_{\Phi}$ and $\mathcal{S}_{\phi}$ are frozen when evaluating MSAC, optimizing a staining model, or running SBCache.

C.4 Controlled Degradation Protocol

To evaluate metric sensitivity under controlled structural perturbations, we generate degraded images directly from ground-truth staining images. No model-generated predictions or MSAC-trained models are involved.

For small-structure removal, connected components are first extracted from foreground regions and sorted according to their areas. The smallest components are removed with ratios of 29.8% and 47.3% for the two degradation levels. The removed regions are filled with local background intensities to avoid introducing artificial boundaries. These perturbations modify only 4.62% and 15.38% of image pixels, respectively. For boundary degradation, Gaussian filtering is applied only to boundary regions with strengths $\sigma_b = 1.61$ and $\sigma_b = 3.82$. For Gaussian-noise controls, zero-mean Gaussian noise is added with $\sigma_n = 0.0217$ and $\sigma_n = 0.0496$. The noise magnitudes are adjusted such that the resulting PSNR matches the corresponding structural perturbations. All metrics are computed between degraded images and their original ground-truth counterparts.

Table 5: Quantitative comparison of spectral decomposition methods. MAE_λ measures the discrepancy between Rayleigh–Ritz eigenvalue estimates and numerical reference eigenvalues; d_{sub} denotes the normalized subspace error. Lower values are better. Bold denotes the best result in each column.

	cpg0000		HEMIT	
Method	$\mathrm{MAE}_\lambda \downarrow$	$d_{\mathrm{sub}} \downarrow$	$\mathrm{MAE}_\lambda \downarrow$	$d_{\mathrm{sub}} \downarrow$
SBLNet	0.208	0.796	0.135	0.513
SpecNet2	0.195	0.7640	0.152	0.587
Ours	**0.0289**	**0.223**	**0.0237**	**0.206**

C.5 Quantitative Evaluation of Spectral Decomposition

We quantitatively compare SBLNet, SpecNet2, and our neural spectral decomposer on cpg0000 and HEMIT. All methods are evaluated using the same normalized graph Laplacian L and its first $K = 16$ nontrivial numerical eigenpairs, denoted by $(U^\star, \boldsymbol{\lambda}^\star)$.

Given a predicted basis matrix Z, we obtain an orthonormal basis Q of its column space and estimate eigenvalues through Rayleigh–Ritz projection:

$$\widehat{\boldsymbol{\lambda}} = \mathrm{eigvalsh}(Q^\top L Q), \qquad \mathrm{MAE}_\lambda = \frac{1}{K}\sum_{k=1}^{K}\left|\widehat{\lambda}_k - \lambda_k^\star\right|, \tag{27}$$

where eigvalsh returns eigenvalues in ascending order. This procedure is applied to all methods, including ours, using the full graph Laplacian rather than the truncated teacher operator. We additionally measure the normalized subspace error:

$$d_{\mathrm{sub}} = 1 - \frac{1}{K}\|Q^\top U^\star\|_F^2. \tag{28}$$

Both measures are invariant to orthogonal changes of basis within the predicted subspace.

Table 5 shows that our method achieves lower Ritz-eigenvalue MAE and subspace error on both datasets. These results complement the qualitative comparison in Figure 3, supporting more accurate recovery of the retained eigenvalues and spectral subspaces.

C.6 MSAC Training for Staining Backbones

Each backbone is first optimized with its original objective $\mathcal{L}_{\mathrm{origin}}$ until validation convergence. Starting from the same converged model and optimizer checkpoint, we conduct an additional training stage under two settings. The baseline setting continues to optimize $\mathcal{L}_{\mathrm{origin}}$, whereas the MSAC-enhanced setting uses the combined staining objective defined in the main paper. Both settings use the same number of additional epochs, mini-batch order, augmentations, optimizer state, learning-rate schedule, and sampling configuration. Thus, the baseline and MSAC-enhanced models receive the same amount of additional optimization.

ResUNet and PyramidPix2Pix. For ResUNet (Ronneberger et al., 2015; Diakogiannis et al., 2020), $\mathcal{L}_{\mathrm{MSAC}}$ is evaluated on the full-resolution prediction $\hat{y} = G_\theta(x)$ and added to the original reconstruction loss; the architecture is unchanged. PyramidPix2Pix (Liu et al., 2022) retains its adversarial, full-resolution ℓ_1, and Gaussian-pyramid multi-scale objectives. MSAC is applied only to the full-resolution generator output and never to the discriminator.

Diffusion and Flow Models. DiffStain (Han et al., 2026) retains semantic-mask guidance, and StainDiffuser (Kataria et al., 2024) retains its staining and H&E-segmentation denoising branches. MSAC is attached only to the clean staining estimate; auxiliary branches and original losses are unchanged. For an ϵ-prediction model,

$$y_t = \sqrt{\bar{\alpha}_t}\, y + \sqrt{1-\bar{\alpha}_t}\, \epsilon,$$
$$\hat{y}_0 = \frac{y_t - \sqrt{1-\bar{\alpha}_t}\, \epsilon_\theta(y_t, t, x)}{\sqrt{\bar{\alpha}_t}},$$

Table 6: Repeated random split evaluation of MSAC-enhanced staining on cpg0000-jump-pilot.

Method	Stain				Seg.		Cnt.
	PCC ↑	PSNR ↑	SSIM ↑	MSAC ↓	Dice ↑	IoU ↑	MAE ↓
ResUNet	.810±.004	18.749±.112	.657±.006	.564±.009	.554±.010	.398±.009	5.74±.31
+$\mathcal{L}_{\text{MSAC}}$	.847±.003	19.186±.095	.713±.005	.460±.008	.602±.008	.425±.007	4.18±.24
PyramidPix2Pix	.824±.005	19.189±.126	.682±.007	.529±.010	.567±.009	.401±.008	5.01±.28
+$\mathcal{L}_{\text{MSAC}}$	.845±.004	19.305±.118	.707±.006	.436±.008	.586±.008	.413±.007	4.33±.25
DiffStain	.866±.003	19.774±.104	.732±.005	.444±.007	.593±.008	.426±.007	1.95±.17
+$\mathcal{L}_{\text{MSAC}}$	**.874±.002**	20.324±.089	.744±.004	.419±.006	.632±.006	.470±.006	1.20±.12
StainDiffuser	.860±.004	20.415±.098	.711±.006	.473±.008	.604±.007	.430±.006	1.26±.13
+$\mathcal{L}_{\text{MSAC}}$	.860±.003	20.358±.094	.708±.005	.412±.006	.629±.006	.465±.005	.95±.10
JiT-UNet	.856±.004	19.477±.119	.692±.007	.497±.009	.585±.009	.414±.008	1.85±.16
+$\mathcal{L}_{\text{MSAC}}$	.868±.003	**20.992±.081**	**.749±.004**	**.396±.005**	**.638±.006**	**.478±.005**	**.75±.08**

with $\epsilon \sim \mathcal{N}(0, I)$. The added term is $\mathcal{L}_{\text{MSAC}}(\hat{y}_0, y)$; the original denoising objective continues to supervise the predicted noise (Ho et al., 2020; Saharia et al., 2022).

For flow models, we use JiT (Li & He, 2026) as the representative because its pixel-space *x-pred* parameterization directly produces a clean staining estimate to which MSAC can be applied. Under our limited-data and limited-compute setting, we replace JiT's transformer denoiser with a convolutional residual U-Net. The U-Net provides local and translation-equivariant inductive biases that are useful for cellular structures in microscopy images, while its multi-scale skip connections preserve fine spatial details. Time embeddings condition the residual blocks, and self-attention is retained at the bottleneck. We otherwise follow JiT's x-pred formulation and train the clean-image prediction with the standard mean-squared-error objective, $\mathcal{L}_{\text{origin}} = \|\hat{y}_0 - y\|_2^2$. MSAC is evaluated on the same $\hat{y}_0$.

For the matched additional training stage, both settings restore the converged model and optimizer, set the learning rate to 10^{-5}, and train for 50 epochs with AdamW, batch size 4, zero weight decay, cosine decay, and unit-norm gradient clipping. The baseline setting uses the original objective, whereas the MSAC-enhanced setting uses the combined objective defined in the main paper. Inputs and targets use the $[0, 1]$ scaling.

C.7 SBCache Implementation

SBCache uses the same 50-step sampler and timestep grid as the uncached JiT-UNet. Both settings remain unchanged across cache policies. The output of the JiT-UNet encoder bears a spatial resolution of 64×64 with 128 feature channels. We flatten its spatial dimensions and use $F_t \in \mathbb{R}^{4096 \times 128}$ as the cache feature and pass it directly to the shared neural spectral decomposer.

Uncached sampling collects one F_t record at every denoiser evaluation. The collected records are passed through the shared frozen $\mathcal{S}_\phi$, and the dataset-level prior $S_F(k)$ is aggregated as defined in the main paper. Offline preparation therefore consists only of uncached feature collection, decomposer inference, and spectral-energy aggregation.

The prior and shared $\mathcal{S}_\phi$ are frozen during sampling. Refresh distances and accumulators are maintained independently for every sample, and the first and final denoiser evaluations are always refreshed. When a step is reused, the cached quantity is JiT-UNet's clean-image estimate. Velocity is recomputed from the current state and time to avoid stale timestep scaling. Refresh ratio counts full denoiser evaluations, and relative latency uses the same batch size, sampler, and model checkpoint for all policies.

Online latency includes encoder feature extraction, neural spectral decomposition and QR orthogonalization, spectral filtering, cache decisions, and denoiser computation. One-time offline prior estimation is excluded from online latency.

Table 7: Repeated random split evaluation of SBCache on cpg0000-jump-pilot.

Method	δ	Efficiency		Stain			
		Ref. ↓	Lat. ↓	PCC ↑	PSNR ↑	SSIM ↑	MSAC ↓
Uncached	–	100	1.00	.868±.003	20.992±.081	.749±.004	.396±.005
TeaCache	.21	50.1±.3	.561±.007	.834±.004	20.123±.102	.714±.003	.433±.007
SeaCache	.19	50.2±.4	.531±.008	.838±.004	20.307±.096	.719±.006	.423±.008
SBCache (ours)	.17	49.8±.7	**.519±.010**	**.843±.003**	**20.476±.109**	**.724±.005**	**.414±.006**
TeaCache	.42	30.3±.5	.391±.007	.811±.005	19.209±.118	.677±.004	.472±.009
SeaCache	.41	29.9±.4	.371±.009	.822±.004	19.554±.107	.689±.007	.455±.008
SBCache (ours)	.39	30.1±.9	**.359±.009**	**.833±.004**	**19.907±.118**	**.701±.006**	**.433±.007**

Table 8: Ablation of pretrained models and spectral decomposition on JiT-UNet using HEMIT.

Model		Stain				Seg.	
		PCC ↑	PSNR ↑	SSIM ↑	MSAC ↓	Dice ↑	IoU ↑
CLIP	$\mathcal{L}_{\text{perc}}$	.786	22.936	.616	.572	.578	.411
	$\mathcal{L}_{\text{MSAC}}$	.791	23.021	.703	.469	.597	.435
DINOv2	$\mathcal{L}_{\text{perc}}$	.793	23.128	.757	.550	.582	.416
	$\mathcal{L}_{\text{MSAC}}$	.803	23.564	.829	.452	.603	.442
DINOv3	$\mathcal{L}_{\text{perc}}$	.795	23.186	.794	.528	.586	.420
	$\mathcal{L}_{\text{MSAC}}$	**.806**	**23.708**	**.842**	**.436**	**.609**	**.449**

Table 9: SBCache comparison on HEMIT.

Method	δ	Efficiency		Stain			
		Ref. ↓	Lat. ↓	PCC ↑	PSNR ↑	SSIM ↑	MSAC ↓
Uncached	–	100	1.00	.806	23.708	.842	.436
TeaCache	.21	60	.65	.751	22.364	.794	.484
SeaCache	.16	60	.60	**.759**	**22.583**	.809	.465
SBCache (ours)	.15	**60**	**.58**	.758	22.422	**.813**	**.457**
TeaCache	.41	40	.45	.701	20.540	.739	.531
SeaCache	.36	40	.40	.715	**20.920**	.753	.508
SBCache (ours)	.33	**40**	**.39**	**.719**	20.284	**.757**	**.485**

D Additional Experimental Results

D.1 Repeated Random Split Results on cpg0000-jump-pilot

For robustness evaluation on cpg0000-jump-pilot, we perform repeated random split experiments at the well level, ensuring that tiles and fields from the same well never occur in different partitions. Each split contains 4,000 training, 200 validation, and 500 test images. We evaluate five independent split configurations and report the mean and standard deviation across splits. Table 6 reports the virtual staining performances, as well as the downstream segmentation and nuclei counting accuracy, on the backbone generative models with and without the proposed MSAC loss. We note that when given MSAC loss for model training, all reported Dice, IoU, and count-MAE comparisons improve. Intensity metrics do not improve uniformly. Table 7 reports caching results across the repeated splits. At approximately matched refresh ratios, SBCache maintains better staining scores and lower mean relative latency than the compared caching methods on cpg0000.

D.2 Additional Experiments on HEMIT

Table 8 provides additional ablation study of the pretrained-models on the HEMIT dataset. For CLIP, DINOv2, and DINOv3 teacher, we distill a separate compact feature encoder and train a separate neural spectral decomposer on the resulting feature maps. All variants use 64×64 feature maps, 16 retained bases, and the same decomposer architecture and optimization schedule. To ensure a common evaluation scale, every MSAC value in the cpg0000 table of the main paper and the HEMIT table below is computed with the default encoder and decomposer trained from DINOv3. In the following table, $\mathcal{L}_{\text{perc}}$ denotes direct perceptual feature matching, whereas $\mathcal{L}_{\text{MSAC}}$ uses spectral decomposition of the same pretrained representation.

Table 9 provides comparison of the proposed SBCache with existing caching methods, including TeaCache and SeaCache on the HEMIT dataset. Our SBCache shows competitive performances regarding both efficiency and virtual staining.

D.3 Parameter Analysis

Neural Spectral Decomposer Objective and Coefficients. For evaluation, $\mathcal{L}_{\text{sub}}$ denotes the subspace discrepancy defined in the main paper, with lower values indicating better alignment. We define the subspace alignment metric $d_{\text{sub}} = 1 - 1/K \left\| U^\top U^\star \right\|_F^2$ to evaluate the neural spectral decomposer. When given K=16 spectral bases, we measure the alignment regarding the low-index

Table 10: Neural spectral decomposition performances under different values of α_{ord} and α_{off}.

		$d_{\text{sub}}\downarrow$				$d_{\text{off}}\downarrow$	$d_{\text{ord}}(\%)\downarrow$
		Overall	Low index	Mid index	Higher index		
w/o $\mathcal{L}_{\text{off}}, \mathcal{L}_{\text{ord}}$		.249	.226	.306	.329	.124	31.3
α_{off}	.05	.242	.237	.282	.347	.051	18.8
	.10	.230	.223	.271	.327	.035	18.8
	.20	**.223**	**.220**	**.260**	.316	.027	**0**
	.50	.231	.232	.269	.314	**.019**	**0**
α_{ord}	0	.232	.225	.295	**.305**	.039	25.0
	.25	.227	.226	.268	.328	.031	**0**
	.50	.230	.223	.263	.327	.029	**0**
	1	**.223**	**.220**	**.260**	.316	.027	**0**
	2	.229	.225	.264	.314	.030	**0**

Table 11: Parameter analysis of α_{MSAC} for JiT-UNet on cpg0000-jump-pilot.

α_{MSAC}	PCC ↑	PSNR ↑	SSIM ↑	MSAC ↓	Dice ↑	IoU ↑	MAE ↓
0	.856	19.477	.692	.497	.585	.414	1.85
.001	.860	19.864	.711	.468	.601	.433	1.42
.005	.865	20.631	.739	.412	.627	.463	.91
.01	**.868**	**20.992**	**.749**	**.396**	**.638**	**.478**	**.75**
.02	.867	20.861	.746	.401	.635	.473	.82
.05	.861	20.214	.727	.429	.615	.448	1.10

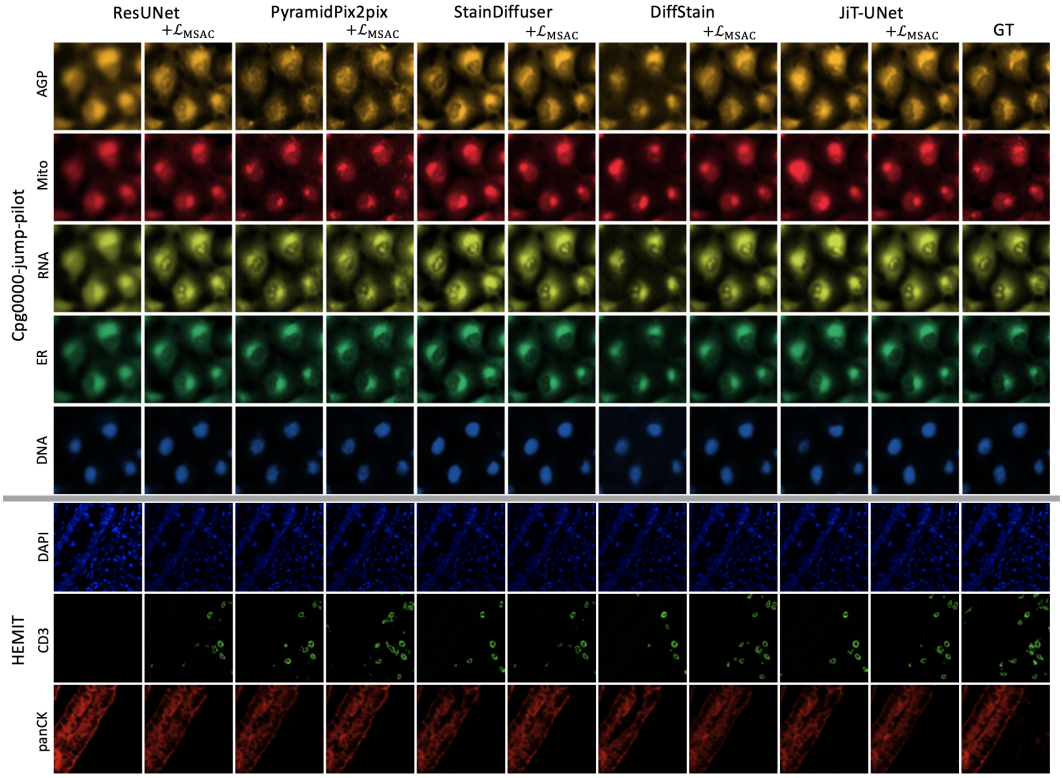


Figure 6: Additional qualitative comparisons on cpg0000-jump-pilot and HEMIT.

(1st–5th), middle-index (6th–10th), higher-index (11th–16th), and full retained (1st–16th) groups. For each group, the overlap uses only its columns and the normalization uses its group size. These relative index groups do not identify the high-frequency end of the full spectrum. The off-diagonal metric $d_{\text{off}} = \|\operatorname{off}(C)\|_F^2/(\|C\|_F^2 + \epsilon)$. The ordering metric d_{ord} is the percentage of adjacent diagonal Rayleigh energies that violate ascending order.

Table 10 report neural spectral decomposition performances with different values of α_{off} and α_{ord}. Note that $\mathcal{L}_{\text{sub}}$ aligns the retained eigenspace, $\mathcal{L}_{\text{off}}$ suppresses mixing between spectral directions, and $\mathcal{L}_{\text{ord}}$ establishes their column order. The setting $(\alpha_{\text{off}}, \alpha_{\text{ord}}) = (0.2, 1)$ provides the best overall balance. Larger coefficients do not consistently improve the diagnostics and can slightly degrade overall alignment.

MSAC Coefficient. Table 11 varies α_{MSAC} for JiT-UNet under the matched additional training protocol. The zero-weight baseline and selected 0.01 setting correspond to the main comparison.

The sweep shows a broad improvement for $\alpha_{\text{MSAC}} \in [0.005, 0.02]$, with 0.01 providing the best overall fidelity–utility trade-off. A larger coefficient begins to overemphasize spectral alignment relative to the original staining objective.

Table 12: Parameter analysis of the retained basis count K on cpg0000. The two panels reproduce the original MSAC-training and SBCache results.

(a) MSAC training

K	PCC ↑	PSNR ↑	SSIM ↑	MSAC ↓	Dice ↑	IoU ↑	MAE ↓
4	.858	19.598	.738	.421	.622	.458	.81
8	.865	20.317	.745	.404	.634	.467	**.73**
16	.868	20.992	**.749**	**.396**	**.638**	**.478**	.75
32	**.870**	**21.013**	.706	.397	.625	.454	1.19
48	.865	20.986	.689	.415	.609	.427	1.52

(b) SBCache

K	δ	Ref. ↓	Lat. ↓	PCC ↑	PSNR ↑	SSIM ↑	MSAC ↓
4	.15	50	**.52**	.834	20.293	**.726**	.422
8	.15	50	.53	.840	20.390	.721	.415
16	.17	50	.52	**.843**	**20.476**	.724	**.414**
32	.18	50	.53	.822	20.357	.703	.455
48	.20	50	.54	.806	20.029	.695	.468

Number of Retained Bases. For each value of K, we retrain the spectral decomposer and use the resulting K-basis model in MSAC training or SBCache, while keeping all other JiT-UNet, training, and cache settings fixed. To ensure that changing K does not also change the evaluation criterion, every MSAC score in Table 12 is recomputed using a separate fixed evaluator that retains the default $K = 16$ bases.

The results show a trade-off in the number of retained modes. Relative to $K = 16$, the $K = 4$ and $K = 8$ settings have lower SSIM, Dice, and IoU, although $K = 8$ gives the lowest count MAE. The $K = 16$ setting provides a favorable balance across MSAC training and SBCache. These comparisons do not isolate which individual graph modes account for the changes. When K is increased to 32 or 48, additional higher-index modes are included within the low-order graph spectrum. The observed degradation is consistent with less useful retained variation, but these experiments do not isolate noise as its cause. Although PCC and PSNR peak slightly at $K = 32$, SSIM, Dice, IoU, and count MAE deteriorate and the cache quality–efficiency trade-off also worsens; the degradation becomes more pronounced at $K = 48$. Considering the results on both components, we use $K = 16$ throughout.

D.4 Additional Qualitative Comparisons

Figure 6 provides additional examples for all five staining backbones. Across both datasets, MSAC-enhanced predictions retain sharper boundaries and finer spatial variations that are attenuated or merged in the original predictions. These selected visual results are consistent with the improved mask agreement under the evaluated CellProfiler pipeline; they do not establish agreement with independent biological annotations.