PV-WSI: RETHINKING PSEUDO-VIDEO LEARNING FOR WHOLE SLIDE IMAGE SEGMENTATION VIA SPATIAL DISCONTINUITY MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Whole-slide image (WSI) segmentation for computational pathology is challenged by gigapixel scale, extreme morphological heterogeneity, and fragmented cross-patch context. Recent pseudo-video formulations attempt to treat WSI patches as sequential frames, but inherit a critical flaw: they impose temporal smoothness priors from natural video onto inherently discontinuous pathological landscapes. This mismatch leads to erroneous context propagation and degraded segmentation, especially across structurally dissimilar regions. We introduce PV-WSI, a pseudo-video paradigm that explicitly models both spatial continuity and discontinuity. Unlike prior work that passively adapts video architectures, PV-WSI reconstructs pseudo-temporal structure. Specifically, we propose Adaptive WSI Pseudo-Video Construction, which identifies pathology-informative anchor patches and dynamically assembles morphology-consistent sequences, eliminating traversal bias and enforcing structural alignment. Furthermore, we develop Distance-Constrained Gate State Propagation, a principled distance-aware propagation mechanism that selectively transmits reliable contextual states while suppressing spurious long-range dependencies across weakly correlated regions. Extensive experiments show that PV-WSI consistently outperforms state-of-the-art methods across magnification levels, improving both tumor segmentation and multi-class tissue parsing. These results demonstrate that explicitly modeling spatial discontinuity is essential for pseudo-video learning and support PV-WSI as a strong framework for large-scale WSI understanding. Code will be available at Github.

1 INTRODUCTION

Whole-slide image (WSI) segmentation is a fundamental task in computational pathology, underpinning diagnosis, prognosis, and downstream quantitative analysis in medical image analysis [33]. However, WSI segmentation remains inherently challenging due to gigapixel resolution and extreme morphological heterogeneity. A single slide may contain diverse tissue structures—such as tumor nests, stroma, necrosis, and inflammatory infiltrates—resulting in complex spatial organization across multiple scales [15]. Since end-to-end processing of WSIs is computationally prohibitive, most methods adopt patch-based strategies, dividing slides into tiles and performing local segmentation [2].

While effective, this **patch-centric paradigm** fundamentally disrupts global tissue continuity. Existing approaches—ranging from CNN-based models [30], multi-scale architectures [35], and transformer-based methods [12], to recent foundation models such as SAM [17], SAM-Med2D [9], and WSI-SAM [19]—primarily operate on isolated patches or limited local neighborhoods [4; 38; 5]. As a result, cross-patch dependencies remain weakly modeled, often leading to fragmented structures, inconsistent predictions, and degraded boundary quality. This limitation is particularly pronounced in regions with complex morphology, e.g., tumor boundaries, where coherent spatial context is critical.

To address this issue, recent work has explored **pseudo-video formulations**, which reinterpret WSI patches as sequential frames and apply video-style modeling [41; 45]. The key intuition is that neighboring patches share local morphological continuity, analogous to temporal coherence in natural videos. By converting spatial adjacency into sequential structure, such methods enable the use of temporal propagation, memory mechanisms, and recurrent or state-space updates, e.g., Mamba-based

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.

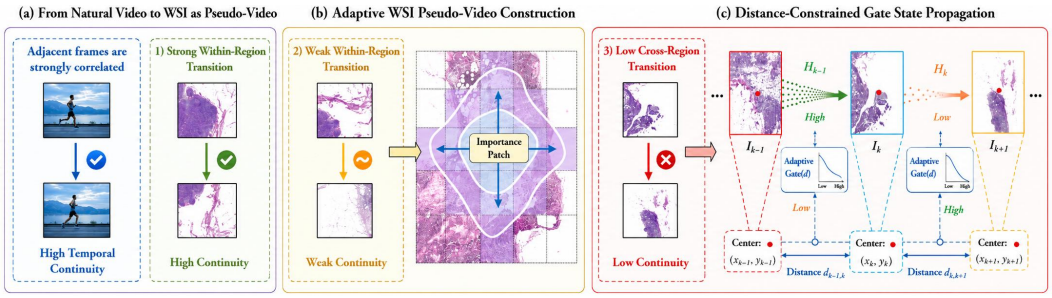

Figure 1: **Motivation and overview of WSI pseudo-video modeling**. Unlike natural videos with strong temporal continuity, WSI pseudo-videos exhibit spatially heterogeneous inter-patch transitions, including (1) *strong within-region similarity*, (2) *weak within-region similarity* caused by local morphological variations, and (3) *low-similarity cross-region transitions* with abrupt structural discontinuities. Our method addresses this challenge by constructing pathology-aware pseudo-video sequences and selectively propagating historical states through distance-constrained gating, explicitly modeling both spatial continuity and discontinuity for more reliable WSI segmentation.

TV3S [13], to capture long-range dependencies across gigapixel tissue regions. Early efforts have demonstrated the promise of this paradigm, ranging from general video segmentation frameworks such as SAM2 [29] to medical adaptations including MedSAM2 [24] and pathology-specific extensions such as Path-SAM2 [41] and PathSeqSAM [45].

However, this line of work suffers from a critical and largely overlooked limitation: it directly transfers temporal smoothness priors from natural video to WSIs [25; 8; 7]. As shown in Figure 1(a), unlike videos, where adjacent frames are typically strongly correlated, WSI patches exhibit spatially heterogeneous transitions. Adjacent patches may belong to the same tissue region with strong continuity, exhibit intra-region variation, or cross abrupt boundaries between distinct tissue types. This inherent spatial discontinuity *violates* the assumptions underlying conventional video architectures. As a result, naive sequential modeling can propagate misleading context across weakly correlated regions, blurring boundaries and degrading segmentation consistency.

In this work, we introduce **PV-WSI**, a pathology-aware pseudo-video framework that explicitly models both spatial continuity and discontinuity. Its primary contribution is not importance scoring, SSIM traversal, or distance gating in isolation, but a coupled reformulation of pseudo-time: AWPVC determines which patch transitions should be presented to a sequential model, and DCGSP determines whether contextual state should cross the residual transitions. Rather than passively adapting video architectures, PV-WSI actively reconstructs pseudo-temporal structure to better align with the underlying organization of tissue. First, we propose *Adaptive WSI Pseudo-Video Construction (AWPVC)*, which replaces fixed raster-scan ordering with a pathology-aware traversal strategy (Figure 1(b)). AWPVC identifies informative anchor patches and dynamically assembles morphology-consistent sequences, prioritizing structurally coherent regions and reducing traversal bias. This results in a more faithful sequential representation that strengthens within-region continuity. Second, we develop *Distance-Constrained Gate State Propagation (DCGSP)*, a distance-aware mechanism for robust context propagation (Figure 1(c)). Instead of uniformly propagating historical states, DCGSP selectively transmits reliable information between spatially close and semantically related patches, while suppressing or resetting spurious dependencies across distant or weakly correlated regions. By explicitly handling discontinuities, the model can re-establish context after structural transitions, preventing error accumulation across tissue boundaries.

Our main contributions are summarized as follows:

- We identify and formalize the mismatch between temporal smoothness in video models and spatial discontinuity in WSIs, revealing a key limitation of existing pseudo-video approaches.
- We propose PV-WSI, a new pseudo-video framework that jointly models spatial continuity and discontinuity for WSI segmentation.
- We introduce AWPVC for morphology-consistent pseudo-video construction and DCGSP for distance-aware, reliable context propagation.

- We demonstrate consistent improvements over prior methods on multiple benchmarks, validating the effectiveness of discontinuity-aware pseudo-video modeling.

2 Related Work

Whole-slide image segmentation. Whole slide image (WSI) segmentation is challenging due to the gigapixel scale of WSIs, heterogeneous tissue morphology, staining variations, sparse annotations, and limited cross-region context. Existing methods commonly process WSIs in a patch-wise manner using biomedical segmentation architectures such as U-Net [30], U-Net++ [43], and nnU-Net [14]; multi-scale models such as HookNet [35]; and transformer or state-space designs such as Swin-UNet [3], UNETR [12], and VM-UNet [31]. Foundation models including SAM [17], MedSAM [23], and WSI-SAM [19] provide transferable representations or promptable masks. Slide-level MIL methods and pathology encoders such as CLAM [21], TransMIL [32], HIPT [5], CTransPath [37], and UNI [6] address related WSI aggregation problems but are not interchangeable dense segmentation baselines.Nevertheless, most methods still rely on local or patch-wise inference, which can weaken tissue continuity across heterogeneous regions. In contrast, PV-WSI reorganizes WSI patches into a pathology-aware sequence to better preserve spatial coherence.

Video-inspired context propagation. Video segmentation models exploit temporal continuity to propagate information across frames, using mechanisms such as optical flow [34], space-time memory [25], long-term memory [7], or selective state spaces [11]. Medical SAM 2 [44] and MedSAM2 adapt video-style propagation to medical data. PathSeqSAM already uses distance-aware attention for variable inter-slice spacing [45]; distance-aware weighting is therefore not itself unique. However, WSIs differ from natural videos: neighboring patches are not governed by smooth temporal motion, but by tissue organization with both continuous and abrupt transitions. Directly imposing video-style smoothness may therefore cause misleading propagation. PV-WSI addresses this limitation through adaptive sequence construction and distance-aware state renewal, enabling useful context propagation while reducing error accumulation across heterogeneous regions.

3 Method

To alleviate pseudo-video discontinuity, PV-WSI introduces two modules: Adaptive WSI Pseudo-Video Construction (AWPVC) for improving *within-region continuity* through adaptive traversal of structurally similar pathology-related patches, and Distance-Constrained Gate State Propagation (DCGSP) for handling *cross-region discontinuity* via distance-guided state propagation and renewal. The overall framework of PV-WSI is presented in Figure 2.

3.1 Adaptive WSI Pseudo-Video Construction

Given a WSI, we divide it into a set of candidate patches $\mathcal{P} = \{p_i\}_{i=1}^{N}$, where each patch p_i is associated with its spatial coordinate $c_i = (x_i, y_i)$. AWPVC aims to construct an ordered pseudo-video sequence $\mathcal{V} = \{I_k\}_{k=1}^{T}$ that follows tissue continuity rather than a fixed raster-scan order, as illustrated in Figure 1(b).

Importance-Aware Initialization. We first estimate the importance of each patch to identify informative tissue regions. The importance score is defined as

$$S_{\text{imp}}(p_i) = E_{\text{den}}(p_i) + C_{\text{con}}(p_i) \tag{1}$$

where $E_{\text{den}}(p_i)$ measures edge density and $C_{\text{con}}(p_i)$ measures intensity contrast. Patches with rich structural boundaries or high tissue variation tend to obtain larger scores. The initial pseudo-video frame is selected as the most informative patch:

$$I_1 = \arg\max_{p_i \in \mathcal{P}} S_{\text{imp}}(p_i), \tag{2}$$

providing a pathology-aware starting point and avoiding initialization from low-content regions.

Similarity-Guided Sequential Traversal. Starting from I_1, AWPVC iteratively retrieves the next frame from the remaining unvisited patch set $\mathcal{U}_{k-1}$. In the implementation, the continuation set

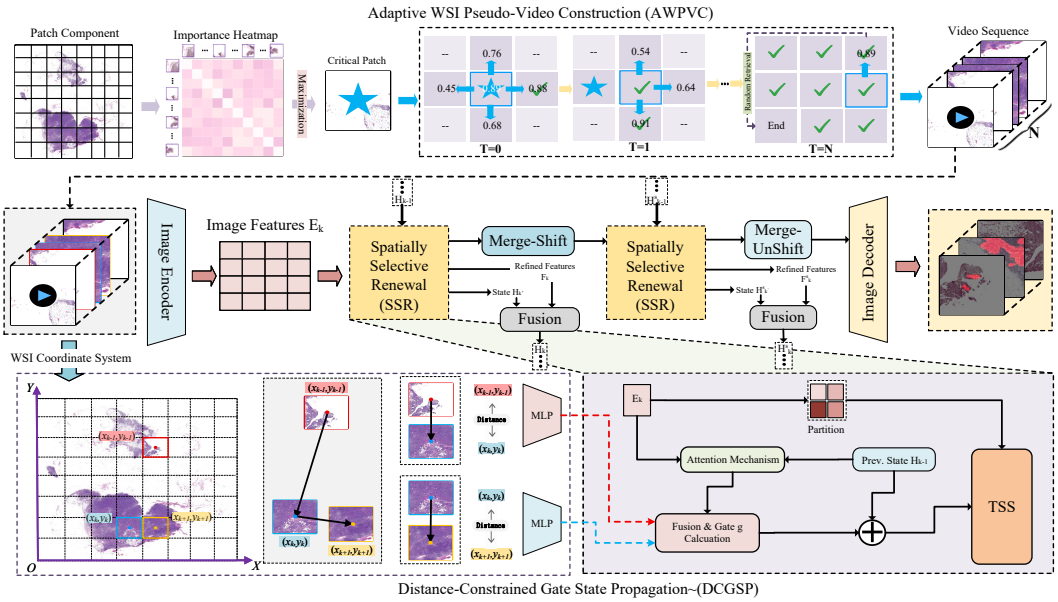

Figure 2: **Overall framework of PV-WSI**. Given a WSI, AWPVC first selects a pathology-informative critical patch and adaptively constructs a pseudo-video sequence based on structural similarity. The encoded pseudo-frames are then processed by DCGSP, which uses WSI coordinates and distance-aware gating to selectively propagate historical states. The refined features are finally decoded and projected back to the WSI coordinate system for segmentation.

$\mathcal{N}_{k-1} \subseteq \mathcal{U}_{k-1}$ contains at most the four spatial neighbors of the current patch. For each local candidate, we compute its structural similarity to the current frame and select

$$I_k = \arg \max_{p_i \in \mathcal{N}_{k-1}} S_{\text{sim}}(p_i, I_{k-1}), k > 1. \tag{3}$$

This strategy encourages adjacent pseudo-video frames to exhibit high morphological and appearance consistency, thereby improving within-region continuity. In practice, we find SSIM to be an effective choice for S_{sim}, as validated in our ablation studies.

Region-Aware Adaptive Restart. Due to the strong spatial heterogeneity of WSI tissues, highly similar continuation candidates may not always exist. Instead of random fallback, AWPVC performs a region-aware adaptive restart when the maximum similarity falls below a threshold τ:

$$\begin{aligned} I_k = \arg \max_{p_i \in \mathcal{U}_{k-1}} & \left[\lambda S_{\text{imp}}(p_i) + (1-\lambda) D_{\text{div}}(p_i, \mathcal{V}_{1:k-1})\right], \\ & \text{if } \mathcal{N}_{k-1} = \emptyset \text{ or } \max_{p_i \in \mathcal{N}_{k-1}} S_{\text{sim}}(p_i, I_{k-1}) < \tau. \end{aligned} \tag{4}$$

Here, $D_{\text{div}}(p_i, \mathcal{V}_{1:k-1})$ measures the diversity between candidate patch p_i and previously visited frames using color histogram differences. The coefficient $\lambda \in [0, 1]$ balances patch importance and exploration of under-sampled regions. This strategy preserves local continuity when high-similarity candidates are available, while adaptively restarting from informative yet diverse regions when continuity breaks occur. Each selected patch is removed from $\mathcal{U}$, and restarting continues within the remainder until $\mathcal{U}$ is empty; hence every valid tissue patch appears exactly once, while preprocessing-excluded background is omitted. We use color histogram disparity to capture global appearance diversity across the WSI. A comparison with alternative restart strategies is provided in Appendix B.

3.2 Distance-Constrained Gate State Propagation

Although AWPVC improves within-region continuity by constructing spatially coherent pseudo-video sequences, adjacent frames may still originate from spatially distant or heterogeneous tissue regions, as shown in Figure 1(c). In such cases, directly propagating the previous hidden states may introduce irrelevant or misleading context. To address this, we propose Distance-Constrained Gate State Propagation, which leverages absolute WSI coordinates to model inter-frame spatial relationships and adaptively regulates hidden-state propagation through distance-aware gating.

Global WSI Coordinate Anchoring For each selected pseudo-video frame I_k, we use its absolute WSI coordinate $c_k = (x_k, y_k)$ as a global spatial anchor. Given two consecutive pseudo-video frames I_{k-1} and I_k, their displacement in the WSI coordinate system is defined as

$$\Delta c_k = c_k - c_{k-1} = (\Delta x_k, \Delta y_k) \tag{5}$$

where $\Delta x_k = x_k - x_{k-1}$ and $\Delta y_k = y_k - y_{k-1}$. The displacement vector preserves the directional relationship between adjacent frames, which is useful for modeling spatially aware state propagation. Since WSIs are usually gigapixel-level images, directly using absolute coordinate distances may cause numerical instability and scale inconsistency. Therefore, we compute the spatial distance on the patch grid by normalizing the coordinate distance with the patch stride s_p:

$$d_k = \frac{\|\Delta c_k\|_2}{s_p} = \frac{\|c_k - c_{k-1}\|_2}{s_p}. \tag{6}$$

Here, d_k measures the normalized spatial separation between adjacent pseudo-video frames. A small d_k indicates that two frames are spatially close and likely to share local tissue context, while a large d_k suggests a possible cross-region transition, where historical memory should be propagated cautiously.

Spatially Selective Renewal For each pseudo-video frame I_k, an image encoder $\mathcal{E}$ first extracts its feature representation: $E_k = \mathcal{E}(I_k)$. We denote the historical contextual information accumulated from preceding frames as the previous hidden state H_{k-1}, which is propagated to the current frame. Given the current image feature E_k and the previous hidden state H_{k-1}, SSR first retrieves task-relevant historical information via cross-attention, similar to [40]:

$$\tilde{H}_{k-1} = \mathrm{Attn}(Q = E_k, K = H_{k-1}, V = H_{k-1}). \tag{7}$$

Instead of directly inheriting the entire historical memory, SSR adaptively regulates state propagation using a coordinate- and content-aware gating mechanism. We first compute the semantic agreement between the current feature and the aligned historical state:

$$s_k = \frac{E_k \cdot \tilde{H}_{k-1}}{\|E_k\| \, \|\tilde{H}_{k-1}\|}. \tag{8}$$

The propagation gate is then estimated from both spatial displacement and semantic agreement:

$$g_k = \sigma\left(\mathrm{MLP}_g\left([\Delta x_k, \Delta y_k, d_k, s_k]\right)\right), \tag{9}$$

where Δx_k and Δy_k denote directional offsets between adjacent frames, and d_k represents the normalized spatial distance defined in Eq. (6). To further suppress unreliable long-range propagation, we impose a Gaussian distance prior [28]:

$$\rho_k = \exp\left(-\frac{d_k^2}{2\sigma_d^2}\right), \qquad \hat{g}_k = \rho_k \cdot g_k. \tag{10}$$

The renewed hidden state is obtained by adaptively interpolating between the aligned historical context and the current observation:

$$\bar{H}_k = \hat{g}_k \odot \tilde{H}_{k-1} + (1 - \hat{g}_k) \odot \phi(E_k), \tag{11}$$

where $\phi(\cdot)$ projects E_k into the hidden-state space. Consequently, SSR favors historical propagation when adjacent frames are spatially close and semantically consistent, while relying more heavily on the current feature when the historical context is distant or mismatched.

Finally, the renewed state is synchronized with the current feature by a Temporal State Space module, similar to [11]:

$$F_k, H'_k = \mathrm{TSS}(E_k, \bar{H}_k), \tag{12}$$

producing the refined representation F_k and the updated hidden state H'_k.

3.3 Feature-Conditioned Attention Fusion with Shifted-Window State Refinement

The refined feature F_k from SSR and the incoming historical state H'_k described in Section 3.2 are processed via two separate paths. To integrate the information, we first perform residual attention

fusion between H'_k and F_k. The refined state H'_k serves as the query, while the current frame feature F_k provides the keys and values:

$$Q = H'_k W_q, \quad K = F_k W_k, \quad V = F_k W_v \tag{13}$$

$$H_k = H'_k + \mathrm{Attn}(Q, K, V) \tag{14}$$

Here, H_k is the true hidden state that will be propagated to the next frame. The residual connection preserves historical information, and the attention mechanism enables the state to adaptively extract relevant visual content from F_k.

Next, F_k is refined through a spatial reorganization and context aggregation pipeline. A Merge-Shift operation first rearranges the features to capture local dependencies. The shifted features are then processed by an SSR module, which leverages spatially compressed context to model long-range cross-frame dependencies. The two SSR applications are complementary rather than duplicate: the first operates on unshifted windows to preserve intra-region morphology, whereas the second operates after shifting so information can cross the window boundaries isolated in the first stage. Finally, a Merge-UnShift operation restores the original spatial resolution, yielding the refined feature F_k^+. After projection, F_k^+ serves as the decoder input with well-integrated spatial-temporal information.

3.4 Segmentation Prediction and Training Objective

Given the fused feature F_k^+, a decoder $\mathcal{D}$ predicts the per-frame segmentation map $\hat{Y}_k = \mathcal{D}(F_k^+)$. After processing all T pseudo-video frames, patch-level predictions are projected back to the original WSI coordinates $\{c_k\}_{k=1}^T$ to form the final whole-slide segmentation mask without post-processing.

We train PV-WSI using a hybrid Dice–BCE loss averaged over all pseudo-video frames:

$$\mathcal{L}_{\mathrm{seg}} = \frac{1}{T} \sum_{k=1}^{T} \left[\beta \mathcal{L}_{\mathrm{Dice}}^{(k)} + (1-\beta) \mathcal{L}_{\mathrm{BCE}}^{(k)} \right] \tag{15}$$

where $\beta \in [0, 1]$ balances region-level overlap and pixel-wise classification, and $\mathcal{L}_{\mathrm{Dice}}^{(k)}$ and $\mathcal{L}_{\mathrm{BCE}}^{(k)}$ denote the standard Dice and binary cross-entropy losses for the k-th frame.

4 Experiments

4.1 Datasets, protocol, and metrics

We evaluate on two breast H&E segmentation benchmarks. CAMELYON16 [10] uses 110 annotated tumor-positive training WSIs and 42 annotated tumor-positive test WSIs. Non-tissue background is removed by Otsu thresholding [27], and non-overlapping 1024×1024 patches are extracted at $1.25\times$, $2.5\times$, $5\times$, $10\times$, and $20\times$. BCSS [1] follows the official data preparation and benchmark partition, using the 1,484-patch H&E setting of Li et al. [18] for Tumor, Stroma, Inflammatory, Necrosis, and Other. BCSS sequences are restricted to one annotated source ROI, so this experiment measures ROI-level tissue parsing rather than full-slide long-range inference.

All methods are evaluated under a unified benchmark protocol with consistent data splits, inputs, preprocessing, training, and evaluation procedures, while retaining architecture-specific configurations where necessary. Test data are strictly excluded from training and model selection, and all reported results are descriptive point estimates; full training settings, metric definitions, evaluation protocols, and implementation details are provided in Appendix

4.2 Main results

Quantitative Results. Quantitative results in Table 1 and Table 2 demonstrate the effectiveness and generalizability of PV-WSI. On CAMELYON16, PV-WSI consistently achieves the best performance across all magnification levels from 1.25 $\times$ to 20 $\times$. Compared with the strongest baseline H&E, it improves the average Dice and mIoU by 7% and 8.11%, respectively, with particularly large gains at low magnifications, indicating its ability to capture both global context and fine-grained boundaries. On BCSS, PV-WSI achieves the best average performance across five tissue categories,

with 84.13% Dice and 72.90% IoU. It obtains the highest scores on Stroma, Inflammatory, Necrosis, and Other, while remaining competitive on Tumor. Compared with the strongest prior baseline MFTrans, PV-WSI improves the average Dice and IoU by 2.34% and 3.36%, verifying its capability for both binary tumor and multi-class tissue segmentation. We also evaluate the Boundary F1 (BF1) metric; detailed results and analysis are presented in Appendix

Table 1: CAMELYON16 comparison across five magnifications (%).

	1.25×		2.5×		5×		10×		20×		Avg.	
Method	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU
U-Net [30]	50.47	49.12	53.59	51.25	56.68	53.69	68.54	65.42	71.69	69.16	60.19	57.73
Swin-UNet [3]	52.67	50.31	58.34	54.26	63.11	61.28	74.58	70.66	78.61	72.18	65.46	61.74
UNETR [12]	48.91	46.10	61.38	58.85	61.57	58.64	71.72	65.58	75.79	71.11	63.87	60.06
VM-UNet [31]	66.22	64.35	78.91	76.27	87.63	84.35	88.12	85.09	90.72	88.61	82.32	79.73
nnWNet [42]	65.84	63.10	76.87	73.17	85.83	83.05	87.67	84.32	88.62	85.31	80.97	77.79
SAM [17]	23.32	20.36	25.35	21.69	29.63	26.32	56.25	54.12	63.58	61.32	39.63	36.76
PathSeqSAM [45]	61.48	59.12	75.43	72.43	84.65	82.17	85.41	83.10	88.27	85.39	79.05	76.44
Path-SAM2 [41]	63.67	60.98	77.12	73.68	85.63	83.66	86.19	84.32	89.42	86.96	80.41	77.92
SAM-Med2D [9]	22.35	20.36	23.36	21.59	25.65	22.69	48.50	45.32	65.21	62.12	37.01	34.42
SegAnyPath [36]	56.96	52.31	74.68	70.41	83.11	76.15	86.06	82.36	89.89	84.60	78.14	73.17
MedSAM [23]	59.04	56.10	72.44	70.69	82.69	79.63	88.36	86.12	90.65	88.01	78.64	76.11
MedSAM2 [24]	64.02	62.39	75.69	73.12	86.58	84.12	89.42	86.96	91.32	88.36	81.41	78.99
MFTrans [39]	60.95	56.18	74.89	71.11	80.75	77.41	83.19	81.24	85.30	77.74	77.02	72.74
H&E branch [18]	72.16	69.65	76.89	73.39	85.45	83.68	88.71	86.48	91.52	88.96	82.95	80.43
PV-WSI	**83.42**	**81.69**	**88.67**	**86.74**	**91.69**	**90.41**	**92.47**	**91.09**	**93.49**	**92.68**	**89.95**	**88.52**

Table 2: Comparison results on the BCSS dataset across different tissue categories (%).

	Tumor		Inflam.		Necrosis		Stroma		Other		Avg.	
Method	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
FCN-8s [20]	85.51	74.69	80.00	66.67	71.20	55.28	72.30	56.62	66.60	49.93	75.12	60.64
HookNet [35]	89.56	81.09	81.67	69.02	55.54	38.45	50.34	33.64	46.12	29.97	64.65	50.43
UNet++ [43]	88.89	80.08	81.72	69.09	75.61	60.78	66.64	49.97	63.80	46.84	75.33	61.35
Attention U-Net [26]	90.59	82.81	84.12	72.59	78.72	64.91	73.34	57.90	70.34	54.25	79.42	66.49
Swin-UNet [3]	88.16	78.83	80.98	68.04	75.37	60.48	63.54	46.56	68.82	52.46	75.37	61.27
VM-UNet [31]	88.29	79.04	80.54	67.42	72.91	57.37	67.13	50.52	61.76	44.68	74.13	59.81
CONCH [22]	63.21	46.21	73.42	58.13	35.48	21.57	80.89	67.91	58.65	41.49	62.33	47.06
nnWNet [42]	90.27	82.27	83.16	71.17	81.33	68.53	66.91	50.27	71.79	55.99	78.69	65.65
MedSAM2 [24]	80.99	68.05	71.64	55.81	51.56	34.73	51.68	34.84	42.28	26.81	59.63	44.05
Path-SAM2 [41]	81.32	68.52	72.40	56.74	53.23	36.27	53.34	36.37	44.37	28.51	60.93	45.28
MFTrans [39]	89.23	80.55	83.75	72.04	80.79	67.77	82.79	70.63	72.41	56.75	81.79	69.55
H&E branch [18]	**92.02**	**85.22**	84.23	72.76	77.74	63.59	81.02	68.12	71.43	55.56	81.29	69.05
PV-WSI	90.83	83.24	**86.34**	**75.96**	**82.41**	**70.08**	**84.83**	**73.65**	**76.23**	**61.58**	**84.13**	**72.90**

Qualitative Visualizations. To further validate the effectiveness of our method, we first present qualitative comparisons on WSI tumor segmentation. As shown in Figure 3, PV-WSI achieves more complete tumor localization and more accurate boundary delineation than recent foundation model based methods. In particular, the regions highlighted by the *red boxes* and *green arrows* show that PV-WSI better captures **irregular tumor boundaries** and **small lesion areas**, whereas competing methods often produce fragmented masks, missed regions, or inaccurate contours. We further provide multi-class tissue segmentation results on the BCSS dataset in Figure 4. PV-WSI generates segmentation maps that are more consistent with the ground truth across different tissue categories, while baseline methods suffer from category confusion, missing regions, or over-smoothed boundaries. These results demonstrate that our method is not only effective for multi-resolution WSI tumor segmentation, but also exhibits strong multi-class segmentation capability.

4.3 Ablation Study

Ablation on Key Components. As shown in Table 3, we ablate each proposed module from PV-WSI. Removing DCGSP decreases the average Dice, mIoU, and BF1 by 2.40%, 2.60%, and 1.51%, respectively, while removing AWPVC leads to average drops of 2.47%, 2.91%, and 1.38%. These results confirm that both distance-aware state renewal and pathology-aware sequence construction bring meaningful gains. Removing both modules, i.e., retaining only the pseudo-video backbone,

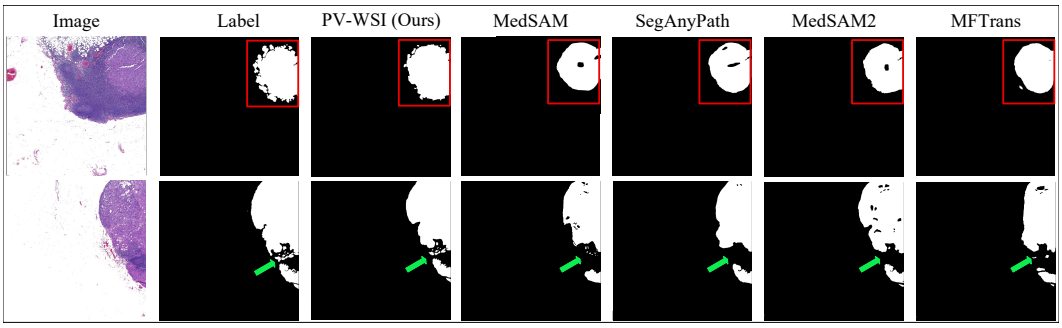

Figure 3: Qualitative comparison of WSI tumor segmentation (CAMELYON16, 5× magnification).

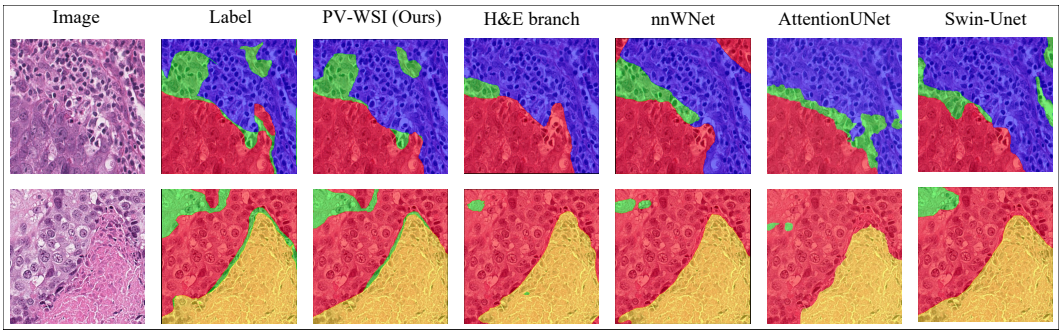

Figure 4: Qualitative comparison of multi-class tissue segmentation on the BCSS dataset.

Table 3: Same-backbone component ablation on CAMELYON16 (%). Configurations are expressed cumulatively from the shared backbone.

	1.25×			2.5×			5×		
Configuration	Dice	mIoU	BF1	Dice	mIoU	BF1	Dice	mIoU	BF1
Backbone	78.51	77.69	73.18	83.96	82.38	77.48	89.37	87.26	81.98
+ DCGSP	81.39	79.96	76.12	84.71	82.97	78.95	90.26	88.19	82.64
+ AWPVC	80.97	79.11	75.58	85.68	83.97	79.54	89.94	87.97	82.18
+ AWPVC + DCGSP	**83.42**	**81.69**	**77.18**	**88.67**	**86.74**	**81.19**	**91.69**	**90.41**	**83.47**

causes the largest degradation, with average reductions of 3.98% in Dice, 3.84% in mIoU, and 3.07% in BF1. Full PV-WSI consistently achieves the best performance at all magnifications, reaching 83.42%, 88.67%, and 91.69% Dice at 1.25×, 2.5×, and 5×, respectively, which demonstrates that AWPVC and DCGSP are complementary and essential for robust multi-resolution WSI segmentation.

Ablation on Pseudo-Video Construction. We ablate different strategies for constructing WSI pseudo-videos in Figure 5. Selecting and traversing frames solely by importance scores achieves competitive performance, but may disrupt morphological continuity between adjacent pseudo-frames. In contrast, pure SSIM-based construction preserves tissue-level continuity, yet can start from less informative regions. Our strategy combines both advantages: it selects the initial frame using pathology-aware importance and performs subsequent traversal based on SSIM structural similarity. This yields the best Dice at both 2.5× and 5×, outperforming the pure SSIM strategy by 0.93% and 1.23%, respectively. These results suggest that effective WSI pseudo-video construction requires both an informative starting patch and morphology-consistent frame transitions.

Ablation on Network Configurations. Figure 6 further evaluates different network configurations at 2.5× magnification. Removing the fusion module consistently decreases Dice, IoU, and BF1, indicating that effective integration between current image features and renewed historical states is important for WSI segmentation. The degradation becomes more evident when the attention mechanism is removed, suggesting that attention-guided interaction is critical for aligning historical context with the current pseudo-frame. Compared with the variant without attention, the full PV-WSI

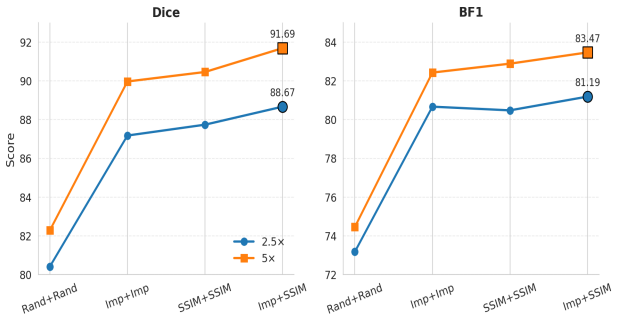

Figure 5: Ablation results of different start selection and video construction strategies. *Rand*: random initialization; *Imp*: Importance-aware initialization.

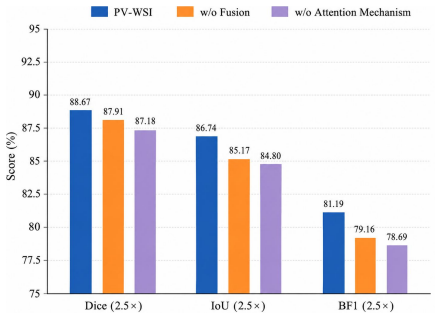

Figure 6: Ablation results of network configurations, including fusion between H'_k and F_k; E_k and H_{k-1}.

improves Dice, IoU, and BF1 by 1.49%, 1.94%, and 2.50%, respectively. Overall, these results demonstrate that both continuity-aware pseudo-video construction and attention-based feature-state interaction contribute to robust WSI segmentation.

Table 4: End-to-end comparison on CAMELYON16 at $1.25\times$. Times are per WSI except training time per epoch.

Method	Params (M)	Preproc.	Train	Inference	Dice
MFTrans	50.56	5.11 s	420.50 s	45.15 s	60.95
MedSAM	4.06	5.11 s	132.92 s	42.86 s	59.04
MedSAM2	11.72	5.11 s	126.67 s	30.58 s	64.02
H&E branch	78.50	5.11 s	374.83 s	48.40 s	72.16
PV-WSI	18.10	8.86 s	218.11 s	**28.57 s**	**83.42**

4.4 Analysis of Efficiency.

Table 4 reports the efficiency comparison on CAMELYON16 at $1.25\times$ magnification. Compared with prior methods, PV-WSI introduces a slightly higher WSI preprocessing cost, increasing from 5.11s to 8.86s. This overhead comes from the proposed AWPVC module, which adaptively constructs pseudo-video sequences by estimating patch importance, computing structural similarity, and performing region-aware restart. Despite this extra preprocessing step, PV-WSI achieves the fastest inference time among all compared methods, requiring only 28.57s per WSI. More importantly, PV-WSI substantially improves segmentation performance, achieving a Dice score of 83.42%, which is significantly higher than competing baselines. These results indicate that the proposed pseudo-video construction introduces only moderate preprocessing overhead while providing a favorable trade-off between computational efficiency and segmentation accuracy.

5 Conclusion

In this paper, we present PV-WSI, a pseudo-video segmentation framework that adapts video-style modeling to the spatially heterogeneous continuity of WSI patches. By combining Adaptive WSI Pseudo-Video Construction with Spatially Selective Renewal, PV-WSI builds morphology-consistent patch sequences and selectively regulates historical state propagation based on spatial distance and contextual relevance. Experiments on CAMELYON16 and BCSS demonstrate consistent improvements in segmentation accuracy, boundary quality, and spatial consistency across multiple magnifications and tissue categories. Despite these promising results, current limitations include the additional preprocessing overhead introduced by pseudo-video construction. Future work will focus on more efficient sequence generation and broader multi-organ and multi-center evaluations.

REFERENCES

[1] Mohamed Amgad, Habiba Elfandy, Hagar Hussein, Lamees A. Atteya, Mai A. T. Elsebaie, Lamia S. Abo Elnasr, Rokia A. Sakr, Hazem S. E. Salem, Ahmed F. Ismail, Anas M. Saad, Joumana Ahmed, Maha A. T. Elsebaie, Mustafijur Rahman, Inas A. Ruhban, Nada M. Elgazar, Yahya Alagha, Mohamed H. Osman, Ahmed M. Alhusseiny, Mariam M. Khalaf, Abo-Alela F. Younes, Ali Abdulkarim, Duaa M. Younes, Ahmed M. Gadallah, Ahmad M. Elkashash, Salma Y. Fala, Basma M. Zaki, Jonathan Beezley, Deepak R. Chittajallu, David Manthey, David A. Gutman, and Lee A. D. Cooper. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics*, 35(18):3461–3467, 2019. doi: 10.1093/bioinformatics/btz083.

[2] Raja Muhammad Saad Bashir, Muhammad Shaban, Shan E. Ahmed Raza, Syed Ali Khurram, and Nasir M. Rajpoot. A novel framework for coarse-grained semantic segmentation of whole-slide images. In *Medical Image Understanding and Analysis*, volume 13413 of *Lecture Notes in Computer Science*, pp. 425–439. Springer, 2022. doi: 10.1007/978-3-031-12053-4_32.

[3] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-Unet: Unet-like pure transformer for medical image segmentation. In *Computer Vision – ECCV 2022 Workshops*, volume 13803 of *Lecture Notes in Computer Science*, pp. 205–218. Springer, 2023. doi: 10.1007/978-3-031-25066-8_9.

[4] L. Chan, M. S. Hosseini, C. Rowsell, K. N. Plataniotis, and S. Damaskinos. Histosegnet: Semantic segmentation of histological tissue type in whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10662–10671, 2019.

[5] R. J. Chen, C. Chen, Y. Li, T. Y. Chen, A. D. Trister, R. G. Krishnan, and F. 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.

[6] 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.

[7] Ho Kei Cheng and Alexander G. Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *Proceedings of the European Conference on Computer Vision*, pp. 640–658. Springer, 2022.

[8] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 11781–11794, 2021.

[9] J. Cheng, J. Ye, Z. Deng, J. Chen, T. Li, H. Wang, Y. Su, Z. Huang, J. Chen, L. Jiang, H. Sun, J. He, S. Zhang, M. Zhu, and Y. Qiao. Sam-med2d. *arXiv preprint arXiv:2308.16184*, 2023.

[10] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 2017. doi: 10.1001/jama.2017.14585.

[11] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024. URL `https://openreview.net/forum?id=tEYskw1VY2`.

[12] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, and Daguang Xu. UNETR: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 574–584, 2022.

[13] Syed Ariff Syed Hesham, Yun Liu, Guolei Sun, Henghui Ding, Jing Yang, Ender Konukoglu, Xue Geng, and Xudong Jiang. Exploiting temporal state space sharing for video semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24211–24221, 2025.

[14] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnu-net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. doi: 10.1038/s41592-020-01008-z.

[15] Jakob Nikolas Kather, Cleo-Aron Weis, Francesco Bianconi, Susanne M Melchers, Lothar R Schad, Timo Gaiser, Alexander Marx, and Frank Gerrit Zöllner. Multi-class texture analysis in colorectal cancer histology. *Scientific Reports*, 6:27988, 2016. doi: 10.1038/srep27988.

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

[17] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W. Y. Lo, P. Dollár, and R. Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

[18] Tengyue Li, Shuangli Song, Qiong Wang, Simon Fong, Wei Song, Juntao Gao, Yi Pan, Nilanjan Dey, and Robert Simon Sherratt. Edge-guided cross-modal fusion network for multi-resolution breast cancer segmentation in smart digital pathology. *IEEE Transactions on Consumer Electronics*, 72(1):342–352, 2026. doi: 10.1109/TCE.2025.3646038.

[19] Hong Liu, Haosen Yang, Paul J. van Diest, Josien P. W. Pluim, and Mitko Veta. WSI-SAM: Multi-resolution segment anything model (SAM) for histopathology whole-slide images. In *Proceedings of the MICCAI Workshop on Computational Pathology*, volume 254 of *Proceedings of Machine Learning Research*, pp. 25–37. PMLR, 2024. URL https://proceedings.mlr.press/v254/liu24a.html.

[20] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015. URL https://openaccess.thecvf.com/content_cvpr_2015/html/Long_Fully_Convolutional_Networks_2015_CVPR_paper.html.

[21] Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5:555–570, 2021. doi: 10.1038/s41551-020-00682-w.

[22] Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024. doi: 10.1038/s41591-024-02856-4.

[23] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang. Segment anything in medical images. *Nature Communications*, 15:654, 2024.

[24] Jun Ma, Zongxin Yang, Sumin Kim, Bihui Chen, Mohammed Baharoon, Adibvafa Fallahpour, Reza Asakereh, Hongwei Lyu, and Bo Wang. Medsam2: Segment anything in 3d medical images and videos. *arXiv preprint arXiv:2504.03600*, 2025.

[25] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9226–9235, 2019.

[26] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention U-Net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018. doi: 10.48550/arXiv.1804.03999.

[27] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. doi: 10.1109/TSMC.1979.4310076.

[28] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[29] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Ha6RTeWMd0.

[30] 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. doi: 10.1007/978-3-319-24574-4_28.

[31] Jiacheng Ruan, Jincheng Li, and Suncheng Xiang. VM-UNet: Vision mamba UNet for medical image segmentation. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2025. doi: 10.1145/3767748.

[32] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, and Y. Zhang. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. In *Advances in Neural Information Processing Systems 34*, pp. 2136–2147, 2021.

[33] Andrew H Song, Guillaume Jaume, Drew F K Williamson, Ming Y Lu, Anurag Vaidya, Tiffany R Miller, and Faisal Mahmood. Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering*, 1(12):930–949, 2023. doi: 10.1038/s44222-023-00096-8.

[34] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, pp. 402–419. Springer, 2020. doi: 10.1007/978-3-030-58536-5_24.

[35] Mart van Rijthoven, Maschenka Balkenhol, Karina Siliņa, Jeroen van der Laak, and Francesco Ciompi. HookNet: Multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images. *Medical Image Analysis*, 68:101890, 2021. doi: 10.1016/j.media.2020.101890.

[36] Chong Wang, Yajie Wan, Shuxin Li, Kaili Qu, Xuezhi Zhou, Junjun He, Jing Ke, Yi Yu, Tianyun Wang, and Yiqing Shen. SegAnyPath: A foundation model for multi-resolution stain-variant and multi-task pathology image segmentation. *IEEE Transactions on Medical Imaging*, 44(10):3924–3937, 2025. doi: 10.1109/TMI.2024.3501352.

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

[38] R. Wetteland, K. Engan, T. Eftestøl, V. Kvikstad, and E. A. M. Janssen. A multiscale approach for whole-slide image segmentation of five tissue classes in urothelial carcinoma slides. *Technology in Cancer Research & Treatment*, 19:1533033820946787, 2020. doi: 10.1177/1533033820946787.

[39] Sungkyu Yang, Woohyun Park, Kwangil Yim, and Mansu Kim. MFTrans: A multi-resolution fusion transformer for robust tumor segmentation in whole slide images. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 4595–4605, 2025. doi: 10.1109/WACV61041.2025.00451.

[40] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 2491–2502, 2021.

[41] M. Zhang, L. Wang, Z. Chen, Y. Ge, and X. Tao. Path-sam2: Transfer sam2 for digital pathology semantic segmentation. *arXiv preprint arXiv:2408.03651*, 2024.

[42] Yanfeng Zhou, Lingrui Li, Le Lu, and Minfeng Xu. nnwnet: Rethinking the use of transformers in biomedical image segmentation and calling for a unified evaluation benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20852–20862, 2025.

[43] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 3–11. Springer, 2018. doi: 10.1007/978-3-030-00889-5_1.

[44] Jiayuan Zhu, Abdullah Hamdi, Yunli Qi, Yueming Jin, and Junde Wu. Medical sam 2: Segment medical images as video via segment anything model 2. *arXiv preprint arXiv:2408.00874*, 2024.

[45] M. Zhu, Y. Liu, M. Li, and J. Wang. Pathseqsam: Sequential modeling for pathology image segmentation with sam2. *arXiv preprint arXiv:2504.10526*, 2025.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, manuscript reorganization, conversion of an existing manuscript into the ICLR LaTeX format, and bibliographic metadata correction. They were not used to generate experimental data, annotations, experimental settings, or numerical results. All AI-assisted revisions were checked against the original manuscript, experimental records, and publication metadata. The authors remain fully responsible for the final content, analyses, and claims.

ETHICS STATEMENT

This work uses the publicly available CAMELYON16 and BCSS benchmark datasets and involves no new patient recruitment, specimen collection, or clinical intervention. All experiments are retrospective algorithmic evaluations conducted on previously released data. The proposed method has not been prospectively evaluated or validated for clinical deployment. In particular, the present experiments do not establish generalization to other organs, medical centers, scanners, staining protocols, or normal-slide screening scenarios.

REPRODUCIBILITY STATEMENT

Dataset membership, preprocessing, coordinate handling, optimization, checkpoint selection, metric definitions, and baseline adaptation are described in Section 4 and Appendix A. Section 3 specifies pseudo-video construction, state propagation, feature fusion, and segmentation prediction. Appendix B reports component ablations, alternative context operators, spatial-continuity statistics, backbone portability, transition perturbation, efficiency, restart-strategy, and parameter-sensitivity experiments. Code and trained models will be released upon acceptance.

A DATASET AND EVALUATION DETAILS

CAMELYON16. CAMELYON16 contains H&E-stained lymph-node WSIs for breast-cancer metastasis assessment [10]. We use 110 annotated tumor-positive WSIs from the official training partition for model training and 42 annotated tumor-positive WSIs from the official test partition for evaluation. The split is performed at the WSI level rather than the patch level, ensuring that patches extracted from the same slide never appear in both training and testing sets. All five magnification settings inherit the same slide membership, preventing scale-dependent changes in the evaluated cohort. Because normal WSIs are not included, this experiment evaluates the delineation of metastatic regions on annotated tumor-positive slides and does not measure normal-slide specificity, slide-level screening sensitivity, or false-positive rates across the complete CAMELYON16 cohort.

Non-tissue background is removed using Otsu thresholding [27]. Each valid tissue region is divided into non-overlapping 1024×1024 patches at $1.25\times$, $2.5\times$, $5\times$, $10\times$, and $20\times$ magnification. All compared methods use the same tissue masks, patch coordinates, and ground-truth projections at each magnification. Patch-level predictions are mapped back to their original WSI coordinates before evaluation, ensuring that all methods are compared on the same spatial support.

BCSS. BCSS contains pathologist-annotated breast-tissue regions derived from TCGA WSIs [1]. We follow the official data preparation, benchmark partition, and class grouping and use the 1,484-patch H&E setting at $0.25\ \mu$m/px described by Li et al. [18]. The five evaluated tissue classes are Tumor, Stroma, Inflammatory, Necrosis, and Other. Every patch retains its source WSI/ROI identifier and original top-left coordinate.

To prevent artificial context from being created across unrelated tissue regions, AWPVC constructs each sequence exclusively from patches belonging to the same annotated ROI. Patches from different ROIs or WSIs are never included in the same pseudo-video. After inference, predictions are restored to their original coordinates to reconstruct the five-class ROI mask. Pixels outside the annotated ROI are ignored rather than assigned to the Other class. Consequently, the BCSS experiment evaluates ROI-level multi-class tissue parsing and does not constitute a full-slide long-range segmentation experiment.

Training and automatic segmentation protocol. All experiments are implemented in PyTorch on one NVIDIA A800 GPU. Unless otherwise stated, PV-WSI and its controlled variants are trained for 50 epochs with a batch size of 2 using Adam [16], an initial learning rate of 10^{-4}, weight decay of 5×10^{-4}, and polynomial learning-rate decay with power 0.9. RandomHSV augmentation is applied during training to improve robustness to stain variation. Training-loss diagnostics indicate convergence by epoch 50 for PV-WSI and the evaluated baselines.

No validation set is used. Test data are excluded from training, hyperparameter tuning, early stopping, and checkpoint selection. Checkpoints are selected using the training objective within the converged training stage. This protocol avoids selecting checkpoints according to test performance, although it does not eliminate overfitting or model-selection uncertainty. Because repeated-seed experiments, confidence intervals, and significance tests are unavailable, all reported differences should be interpreted as descriptive point estimates.

All methods use the same dataset-specific split, patch inputs, valid-tissue support, coordinate reconstruction, and evaluation implementation. Controlled PV-WSI ablations additionally use the same encoder, decoder, loss, training schedule, and optimization settings unless the corresponding component is explicitly changed. Independently designed baselines retain their official architectures and recommended method-specific configurations when available. The comparison therefore controls the data and evaluation pipeline but is not claimed to be parameter-matched across every architecture.

Ablation implementation details. The conventional traversal used when AWPVC is absent is a raster scan over the same valid-tissue patches, ordered from the top left, left to right and then top to bottom, while skipping excluded background. Patch membership and sequence length are unchanged. In the fusion ablation, the feature/state fusion blocks are replaced by identity mappings, so the renewed state and current feature continue through their original paths with unchanged dimensions. State propagation is unidirectional: each forward state summarizes all previously visited patches, while DCGSP suppresses unreliable transitions. A reverse pass would add computation, memory, and a second fusion path with largely overlapping context, so the evaluated model uses one pass.

All methods use the same dataset-specific split, patch inputs, valid tissue support, augmentation, coordinate reconstruction, and evaluation implementation. Training uses one NVIDIA A800, 50 epochs, batch size 2, Adam [16], learning rate 10^{-4}, weight decay 5×10^{-4}, and polynomial decay of power 0.9. Training-loss diagnostics confirmed convergence by epoch 50 for PV-WSI and every baseline. No validation set is used: tests are held out from training, tuning, early stopping, and checkpoint selection; checkpoints are selected by the training objective within the converged stage. This avoids test-selected checkpoints but does not eliminate overfitting or model-selection uncertainty. Repeated-seed dispersion, confidence intervals, and significance tests are unavailable, so all differences are descriptive point estimates.

Metrics. Dice measures foreground overlap on CAMELYON16; its mIoU is the unweighted mean of foreground and background IoU. BCSS mDice, mIoU, and mBF1 average the five tissue classes. The five-scale CAMELYON16 averages do not pool pixels across magnifications. SAM-family models are evaluated as automatic dense predictors without ground-truth boxes, masks, or manually selected points. Architecture-specific configurations are retained under the shared benchmark pipeline.

For foreground prediction P and ground truth G, $\mathrm{Dice} = 2|P \cap G|/(|P| + |G|)$ and $\mathrm{IoU} = |P \cap G|/|P \cup G|$. CAMELYON16 reports foreground Dice and $\mathrm{mIoU} = \frac{1}{2}(\mathrm{IoU}_{\mathrm{fg}} + \mathrm{IoU}_{\mathrm{bg}})$ over the valid evaluation support. BCSS averages class-wise values over five classes. Boundary F1 matches predicted and reference boundary pixels within the shared tolerance radius. All metrics are percentages and gains are percentage-point differences.

B Additional Ablations and Diagnostics

Complementarity of AWPVC and DCGSP. To determine whether the improvements arise from pseudo-video construction, state propagation, or their combination, we compare the shared backbone with DCGSP alone, AWPVC alone where available, and the complete AWPVC+DCGSP configuration. All configurations retain the same encoder, decoder, data split, loss, and optimization protocol. On CAMELYON16, the complete model improves the shared backbone by an average of 3.98 Dice, 3.84 mIoU, and 3.07 BF1 points across the reported $1.25\times$, $2.5\times$, and $5\times$ settings. AWPVC and DCGSP also improve the backbone independently at each reported scale, showing that neither module merely duplicates the contribution of the other.

Table 5 reports the corresponding cumulative comparison on BCSS. Adding DCGSP to the backbone increases mDice from 81.67 to 82.36, mIoU from 66.97 to 68.18, and mBF1 from 75.41 to 76.09. Adding AWPVC on top of DCGSP further increases these values to 84.13, 72.90, and 78.15, respectively. Relative to the backbone, the complete model therefore gains 2.46 mDice, 5.93 mIoU, and 2.74 mBF1 points. These results indicate that sequence construction and selective state renewal provide complementary benefits for both region overlap and boundary delineation. Because an AWPVC-only BCSS result is unavailable, the BCSS experiment does not independently quantify the contribution of AWPVC without DCGSP.

Table 5: Same-backbone component ablation on BCSS (%). Configurations are cumulative from the shared backbone.

Configuration	mDice	mIoU	mBF1
Backbone	81.67	66.97	75.41
+ DCGSP	82.36	68.18	76.09
+ AWPVC + DCGSP	**84.13**	**72.90**	**78.15**

Additional high-magnification ablation. To examine whether the component gains persist when patches contain more detailed local morphology, Table 6 evaluates the cumulative configurations on CAMELYON16 at $10\times$. Adding DCGSP increases Dice from 89.84 to 90.72, mIoU from 87.91 to 88.17, and BF1 from 83.45 to 83.97. The complete AWPVC+DCGSP model further increases these metrics to 92.47, 91.09, and 85.74, corresponding to gains of 2.63 Dice, 3.18 mIoU, and 2.29 BF1 points over the shared backbone.

The improvement at $10\times$ shows that discontinuity-aware sequence construction and propagation remain useful when local tissue details are more visible. However, the gain is evaluated through a cumulative comparison and does not separately quantify an AWPVC-only configuration at this scale. The more computationally demanding $20\times$ component factorization was not completed, and no claim is made about the separate module contributions at $20\times$.

Table 6: Additional same-backbone component ablation on CAMELYON16 at $10\times$ (%).

Configuration	Dice	mIoU	BF1
Backbone	89.84	87.91	83.45
+ DCGSP	90.72	88.17	83.97
+ AWPVC + DCGSP	**92.47**	**91.09**	**85.74**

Table 8: Analysis of spatial continuity in pseudo-video construction.

Pseudo-video Construction	Adjacent Transition (Within a Region)	Non-adjacent Transition (Within a Region)	Cross-region Transition
Conventional WSI Traversal	6.4	2.4	5.3
AWPVC (Ours)	10.9 ↑	1.1 ↓	2.1 ↓

Alternative context operators. To determine whether the improvement can be reproduced by replacing the proposed cross-patch mechanism with another context operator, Table 7 compares Spatial-GAT, MAF-style retrieval, a coordinate-window transformer, generic pseudo-video propagation, and PV-WSI on CAMELYON16 at $2.5\times$. The surrounding encoder, fusion module, decoder, loss, data split, and optimization protocol are held fixed, so only the cross-patch context operator changes.

PV-WSI obtains 88.67 Dice and 81.19 BF1. It exceeds Spatial-GAT by 3.14 Dice and 3.60 BF1 points, MAF-style retrieval by 5.95 Dice and 5.41 BF1 points, the coordinate-window transformer by 7.56 Dice and 8.51 BF1 points, and generic pseudo-video propagation by 4.71 Dice and 3.71 BF1 points. These controlled results show that the proposed coupling of coherent sequence construction and selective state renewal is more effective than the tested replacements within the fixed PV-WSI pipeline. They do not establish universal superiority over independently optimized graph, retrieval, or transformer architectures with different capacities or training procedures.

Table 7: Controlled comparison of cross-patch context operators on CAMELYON16 at $2.5\times$ (%).

Context operator	Dice	BF1
Spatial-GAT	85.53	77.59
MAF-style retrieval	82.72	75.78
Coordinate-window transformer	81.11	72.68
Generic pseudo-video	83.96	77.48
PV-WSI	**88.67**	**81.19**

Analysis of Spatial Continuity. To evaluate how the proposed PV-WSI preserves spatial continuity across patches, we compare conventional WSI traversal with AWPVC in Table 8. AWPVC yields more coherent patch transitions by increasing adjacent transitions within the same region from 6.4 to 10.9, indicating stronger local continuity. It also reduces non-adjacent intra-region transitions from 2.4 to 1.1, suggesting fewer unnecessary long-range jumps. Notably, cross-region transitions decrease from 5.3 to 2.1, showing that AWPVC effectively suppresses transitions between spatially separated regions and alleviates region-level mixing. Overall, these results demonstrate that AWPVC constructs pseudo-videos with stronger spatial coherence, better local continuity, and fewer discontinuous transitions than conventional WSI traversal.

Backbone portability. To test whether the benefit of AWPVC depends on the PV-WSI backbone, Table 9 introduces AWPVC into an otherwise unchanged MedSAM2 pipeline. Only the input patch ordering is modified; the downstream MedSAM2 components and optimization procedure remain unchanged. At $1.25\times$, AWPVC improves MedSAM2 from 64.02 to 69.12 Dice and from 62.39 to 67.96 mIoU, corresponding to gains of 5.10 and 5.57 points. At $2.5\times$, the gains increase to 7.06 Dice and 8.34 mIoU points.

Positive improvements are also observed at higher magnifications. At $5\times$, Dice/mIoU increase from 86.58/84.12 to 88.96/87.66, while at $10\times$ they increase from 89.42/86.96 to 90.36/88.61. The improvement at every evaluated scale provides evidence that morphology-aware patch ordering is useful beyond the original PV-WSI backbone. However, this experiment transfers only AWPVC. DCGSP is coupled to the recurrent state update of PV-WSI and is not inserted into MedSAM2, so the experiment should not be interpreted as a portability evaluation of the complete framework.

Table 9: AWPVC transfer to an otherwise unchanged MedSAM2 pipeline (%).

Scale	Method	Dice	mIoU
1.25×	MedSAM2	64.02	62.39
	MedSAM2 + AWPVC	**69.12**	**67.96**
2.5×	MedSAM2	75.69	73.12
	MedSAM2 + AWPVC	**82.75**	**81.46**
5×	MedSAM2	86.58	84.12
	MedSAM2 + AWPVC	**88.96**	**87.66**
10×	MedSAM2	89.42	86.96
	MedSAM2 + AWPVC	**90.36**	**88.61**

Table 10: AWPVC sequence-construction scaling across magnifications.

Scale	Patches	AWPVC	Per patch	Raster
1.25×	21	0.56 s	26.6 ms	0.01 ms
2.5×	91	2.4 s	26.3 ms	0.03 ms
5×	316	8.6 s	27.2 ms	0.06 ms
10×	1,120	30.6 s	27.4 ms	0.16 ms
20×	3,986	108.2 s	27.2 ms	0.41 ms

Efficiency and scaling. To examine the computational cost of pseudo-video construction as the number of patches increases, Table 10 measures AWPVC at all five CAMELYON16 magnifications. The number of patches increases from 21 at $1.25\times$ to 3,986 at $20\times$, while the corresponding construction time increases from 0.56 s to 108.2 s. Despite this large change in sequence length, the measured cost remains between 26.3 and 27.4 ms per patch.

AWPVC computes patch-importance scores independently and visits each patch once, comparing it with at most four spatial neighbors during traversal. Its sequence-construction complexity is therefore $O(4N) = O(N)$ rather than an all-patch $O(N^2)$ search. The measured times produce a linear fit of $R^2 = 0.999993$, supporting approximately linear scaling over the evaluated range. Nevertheless, AWPVC remains substantially slower than raster ordering and introduces additional preprocessing overhead. These measurements characterize the present implementation and hardware and do not constitute a peak-memory analysis or a hardware-independent claim about complete end-to-end complexity.

Controlled transition perturbation. To test whether a trained model is sensitive to the coherence of its input sequence, Table 11 progressively replaces coherent transitions with distant, low-SSIM transitions at inference time. The patch set, annotations, sequence length, model weights, and metric implementation remain fixed. With no perturbed transitions, the model achieves 88.67 Dice and 81.19 BF1. Perturbing 25% of transitions reduces the scores to 88.05 and 80.52, while perturbing 50% reduces them to 87.21 and 79.45.

At the highest perturbation ratio of 75%, Dice decreases to 85.87 and BF1 to 77.54, representing total reductions of 2.80 Dice and 3.65 BF1 points relative to the unperturbed sequence. The monotonic degradation shows that the fixed trained model is sensitive to the coherence of cross-patch transitions, with boundary accuracy declining more strongly as discontinuities accumulate. However, the perturbation also introduces an inference-distribution shift. The experiment therefore supports the importance of sequence order but does not prove that spatial discontinuity is the only possible explanation for the observed degradation.

Table 11: Inference-time transition perturbation on CAMELYON16 at $2.5\times$ (%).

Perturbed transitions	Dice	BF1
0%	**88.67**	**81.19**
25%	88.05	80.52
50%	87.21	79.45
75%	85.87	77.54

Restart and parameter sensitivity. When no acceptable neighboring patch is available, AWPVC must select a new anchor from the remaining patches. Table 12 compares random restart, SSIM-based restart, texture-difference restart, and Lab-histogram restart at $2.5\times$ and $5\times$. At $2.5\times$, Lab-histogram restart achieves the best result, reaching 88.67 Dice and 81.19 BF1, compared with 86.96/78.37 for random restart, 87.95/80.68 for SSIM restart, and 87.27/79.85 for texture-difference restart.

At $5\times$, Lab-histogram restart obtains the highest Dice of 91.69, whereas SSIM restart produces the highest BF1 of 84.64. Lab-histogram restart therefore provides the strongest overlap performance at both scales but is not optimal for every metric. The comparison suggests that color-distribution information is useful for locating a meaningful new region after a continuity break, whereas direct structural similarity can be advantageous for preserving fine boundaries at higher magnification.

Table 12: Restart-strategy comparison (Dice/BF1, %).

Restart	$2.5\times$	$5\times$
Random	86.96/78.37	89.63/81.19
SSIM	87.95/80.68	90.39/**84.64**
Texture difference	87.27/79.85	89.91/84.22
Lab histogram	**88.67/81.19**	**91.69**/83.47

Table 13: Sensitivity of fixed defaults $\lambda = 0.75$ and $\beta = 0.50$ (Dice/BF1, %).

Parameter	Value	$1.25\times$	$2.5\times$
λ	0.00	81.96/75.59	85.68/78.66
	0.25	82.78/76.51	87.20/79.58
	0.50	83.27/77.05	88.29/80.87
	0.75	**83.42/77.18**	**88.67/81.19**
	1.00	82.64/75.73	87.22/80.67
β	0.25	82.37/76.47	86.95/79.98
	0.50	**83.42/77.18**	**88.67/81.19**
	0.75	81.11/74.84	85.29/79.48

Sensitivity to λ, β, and τ. Table 13 examines whether performance is stable around the selected defaults. Increasing the propagation-control parameter λ from 0 to 0.75 progressively improves Dice and BF1 at both $1.25\times$ and $2.5\times$. At $1.25\times$, performance rises from 81.96/75.59 at $\lambda = 0$ to 83.42/77.18 at $\lambda = 0.75$. At $2.5\times$, it rises from 85.68/78.66 to 88.67/81.19. Increasing λ further to 1.0 reduces performance at both scales, indicating that unrestricted historical propagation can reintroduce unreliable context. The intermediate value $\lambda = 0.75$ therefore provides the best balance between retaining useful history and suppressing noisy transitions.

For the loss-balancing parameter β, the intermediate setting $\beta = 0.50$ achieves the strongest Dice and BF1 at both evaluated scales. At $2.5\times$, for example, $\beta = 0.50$ reaches 88.67 Dice and 81.19 BF1, compared with 86.95/79.98 at $\beta = 0.25$ and 85.29/79.48 at $\beta = 0.75$. This result indicates that balancing region-overlap and pixel-wise classification objectives is more effective than strongly emphasizing either term.

At $2.5\times$, the similarity threshold values $\tau = 0.3, 0.4, 0.5, 0.6$, and 0.7 yield Dice scores of 86.16, 86.58, 88.67, 85.19, and 83.61, respectively. Low thresholds allow weakly related neighbors to enter the sequence, whereas high thresholds reject too many candidates and fragment the traversal. The best result at $\tau = 0.5$ reflects a balance between local continuity and sequence coverage. All threshold settings retain the same patch set, so the experiment changes ordering rather than data membership. These defaults were fixed during training and were not reselected according to held-out test performance.

Boundary accuracy on CAMELYON16 and BCSS. To determine whether the overlap improvements are accompanied by more accurate tissue boundaries, Table 14 reports Boundary F1 under the same valid-pixel support used for the region-based metrics. On CAMELYON16, PV-WSI achieves the highest BF1 at every evaluated magnification, increasing from 77.18 at $1.25\times$ to 87.11 at $20\times$. Its five-scale average is 82.94, exceeding the strongest reported baseline average, MedSAM2 at 74.62, by 8.32 points. The particularly large improvement at low magnification suggests that cross-patch

context is useful when individual patches provide limited local detail. However, lesion scale, field of view, and patch count also change with magnification, so the larger low-magnification gain cannot be attributed to magnification alone.

On BCSS, PV-WSI achieves a mean BF1 of 78.15, improving over the H&E branch result of 74.10 by 4.05 points. PV-WSI improves Stroma from 78.75 to 80.93, Inflammatory tissue from 69.67 to 75.70, Necrosis from 72.33 to 77.41, and Other from 62.61 to 69.80. The H&E branch remains slightly better for Tumor, obtaining 87.12 compared with 86.89 for PV-WSI. These class-dependent results indicate that PV-WSI provides its clearest boundary improvements for heterogeneous and context-dependent tissue categories, whereas the visually distinctive Tumor class is already handled effectively by strong local-appearance baselines.

Table 14: Boundary F1 on CAMELYON16 and BCSS; averages are unweighted (%).

(a) CAMELYON16 by magnification

Method	1.25×	2.5×	5×	10×	20×	Avg.
U-Net	46.52	48.48	50.79	59.42	65.55	54.15
Swin-UNet	48.80	51.21	56.25	69.43	73.74	59.89
UNETR	44.76	57.52	59.69	63.89	69.63	59.10
PathSeqSAM	56.33	68.58	79.51	80.57	83.86	73.77
Path-SAM2	56.82	70.98	78.79	80.06	82.29	73.79
SegAnyPath	50.11	68.82	75.98	77.21	81.41	70.71
MedSAM	52.92	65.57	76.55	81.49	82.18	71.74
MedSAM2	58.88	69.84	79.73	81.29	83.37	74.62
MFTrans	53.82	67.02	72.91	76.06	78.44	69.65
PV-WSI	**77.18**	**81.19**	**83.47**	**85.74**	**87.11**	**82.94**

(b) BCSS by tissue class

Method	Tumor	Stroma	Inflam.	Necrosis	Other	Avg.
FCN-8s	78.47	72.85	64.42	65.19	59.79	68.14
HookNet	83.61	73.16	42.78	42.77	31.25	54.71
UNet++	82.32	73.98	66.10	58.19	55.23	67.16
Attention U-Net	84.78	75.62	71.16	65.87	62.57	72.00
Swin-UNet	80.41	71.13	69.24	55.37	61.99	67.63
VM-UNet	81.97	69.54	65.95	58.96	48.88	65.06
CONCH	52.58	64.11	23.73	73.69	46.71	52.16
nnWNet	85.76	76.53	73.58	54.75	63.13	70.75
H&E branch	**87.12**	78.75	69.67	72.33	62.61	74.10
PV-WSI	86.89	**80.93**	**75.70**	**77.41**	**69.80**	**78.15**

C Discussion

Why retain pseudo-video modeling? Pseudo-video modeling provides an efficient recurrent mechanism for transmitting information across WSI patches, although pseudo-time does not imply the natural temporal continuity of ordinary videos. AWPVC addresses this distinction by constructing locally coherent traversal paths, while DCGSP regulates historical-state propagation when spatial distance or appearance changes reduce the reliability of a transition. Spatially Selective Renewal (SSR) is applied before and after merge–shift, thereby regulating state propagation during both within-window modeling and the cross-window context exchange introduced by merge–shift.

The main contribution of PV-WSI lies in the coordinated integration of morphology-aware sequence construction and discontinuity-aware state propagation, rather than in any individual scoring, traversal, or weighting operation alone. The controlled alternative-operator experiment demonstrates the effectiveness of this design within a consistent evaluation pipeline. Graph models, retrieval mechanisms, and coordinate-aware attention offer complementary directions for further cross-patch modeling.

Interpreting the experimental evidence. The ablations and diagnostic experiments address complementary questions. Same-backbone component ablations quantify the improvements associated with AWPVC and DCGSP beyond generic sequential aggregation. Transition statistics characterize how AWPVC changes the spatial structure of the constructed path. Fixed-weight perturbation tests the sensitivity of an already trained model to disruptions of path coherence, while the alternative-operator experiment examines whether other context mechanisms recover the same performance within a fixed surrounding architecture.

Together, these results provide converging evidence that coherent traversal and selective state renewal contribute to segmentation performance, although contextual capacity, feature interaction, and optimization may also influence the observed gains. The qualitative boundary examples complement the quantitative evidence by illustrating how the improvements appear at challenging tissue interfaces.

At low magnification, changes in lesion scale, field of view, and patch count occur together; the corresponding gains should therefore be interpreted as evidence under the resulting imaging setting rather than as an effect of magnification alone.

Architecture and generalization. The MedSAM2 portability experiment shows that AWPVC improves a second pipeline when only its input ordering is changed (Appendix Table 9), supporting the transferability of morphology-aware pseudo-video construction. DCGSP is designed around the recurrent state update used by PV-WSI; adapting it to architectures with different state-update mechanisms is an interesting direction for future work.

The current design uses unidirectional propagation, so each patch receives context from previously visited patches. Bidirectional propagation could provide additional context, but would require an additional state pass, greater memory consumption, and a revised fusion mechanism. The unidirectional formulation therefore represents an empirically effective and computationally efficient design choice; future work may examine whether bidirectional context justifies the additional cost.

Class-dependent behavior. PV-WSI achieves the best average Dice, IoU, and BF1 on BCSS, although the magnitude of its improvement varies across tissue categories. The H&E branch remains marginally stronger for Tumor, whereas PV-WSI provides clearer improvements for Stroma, Inflammatory tissue, Necrosis, and Other. One possible explanation is that Tumor regions are often characterized by relatively distinctive local appearance and boundary cues, with several baselines already achieving strong performance for this category. The remaining classes exhibit greater morphological heterogeneity and may therefore benefit more from cross-patch consistency and discontinuity-aware propagation. This pattern motivates further investigation of class-aware sequence construction and boundary-sensitive refinement.

Scope and limitations. The present study focuses on breast H&E pathology, including ROI-level tissue parsing on BCSS and annotated tumor-positive slides from CAMELYON16. Future studies could further evaluate the framework in broader settings, including normal-slide screening, unannotated whole-slide cohorts, other organ types, external medical centers, different scanners, and alternative staining protocols. The reported results consistently demonstrate the effectiveness of the proposed method under the evaluated settings; repeated-seed experiments and statistical analyses would provide a more comprehensive characterization of performance variability. External validation, false-positive assessment on normal slides, calibration, and uncertainty estimation may further support translation to broader clinical scenarios. Pseudo-video construction also introduces preprocessing cost, and efficiency at high patch counts may vary with implementation and hardware.

Dataset resources. The official BCSS materials are available from the BCSS repository (`https://github.com/PathologyDataScience/BCSS`) and the benchmark page (`https://bcsegmentation.grand-challenge.org/Baseline/`).