

Information-Guided Spatiotemporal Scanning for Efficient Video State-Space Modeling

Tian Ye

The First Affiliated Hospital of Hebei North University

Wang Lu

The First Affiliated Hospital of Hebei North University

Chaowei Yang

The First Affiliated Hospital of Hebei North University

Yahui Kang

Tianjin Medical University

Jun Miao

Tianjin University

Shuxuan Li

`l1sx@jlu.edu.cn`

First Hospital of Jilin University

Research Article

Keywords: State-space models, Information-guided adaptive scanning, Spatiotemporal path planning, Video modeling

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11247752/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Information-Guided Spatiotemporal Scanning for Efficient Video State-Space Modeling

Abstract

State-space models (SSMs) have emerged as an efficient paradigm for video modeling owing to their favorable computational complexity in processing long visual sequences. However, existing SSM-based approaches typically serialize visual tokens using predefined, content-independent scanning patterns, which fail to account for the non-uniform spatial distribution and temporal evolution of information in videos. To address this limitation, we propose **AdaScan**, an information-guided adaptive scanning framework that dynamically determines the regions to be processed, their spatial granularity, and their serialization order according to video content. Specifically, AdaScan comprises three complementary components. First, an information-aware scoring mechanism jointly characterizes static structural information, residual motion, and motion-compensated predictive innovation to identify informative regions across space and time. Second, an adaptive spatial-resolution strategy dynamically adjusts patch granularity according to local video activity, enabling fine-grained representation of informative regions while avoiding redundant computation in less informative areas. Third, a spatiotemporal path-planning strategy organizes the selected regions into coherent and non-redundant scanning trajectories, thereby improving information acquisition while reducing unnecessary traversal. Together, these components replace fixed visual serialization with a content-adaptive sequence that is better aligned with the sequential modeling mechanism of SSMs. Extensive experiments on action recognition, fine-grained video understanding, and video semantic segmentation demonstrate the effectiveness and generalizability of AdaScan. The proposed method achieves **88.5%** Top-1 accuracy on Kinetics-400, **97.6%** accuracy on Breakfast, and **58.7%** mIoU on VSPW. Further ablation and analytical studies confirm the complementary contributions of the proposed components and demonstrate that information-guided path planning achieves a favorable trade-off between informative content acquisition and scanning efficiency compared with conventional scanning strategies.

Keywords: State-space models, Information-guided adaptive scanning, Spatiotemporal path planning, Video modeling

1 Introduction

Video understanding requires effective modeling of both spatial semantics and temporal dynamics while avoiding the substantial redundancy that commonly exists across consecutive frames. Recent state-space models (SSMs), ranging from structured state-space sequence models to Mamba-style selective SSMs [1, 2], have emerged as an efficient alternative to attention-based architectures because of their favorable computational complexity for long-sequence modeling. When extending SSMs to visual data, however, the two-dimensional spatial structure and temporal evolution of videos must first be serialized into sequential representations. Consequently, the order in which visual tokens are processed directly affects how information is propagated through the recurrent state evolution. Existing visual SSMs typically employ predefined serialization strategies, including bidirectional, cross-directional, local, and multi-dimensional scanning patterns [3–6]. Although these strategies provide efficient and regular implementations, their scanning orders are largely independent of the input content and therefore cannot adapt to the heterogeneous information distribution of dynamic video scenes.

Such content-independent scanning is not well aligned with the characteristics of natural videos. Visual information is generally distributed non-uniformly across both space and time: foreground objects may occupy only a limited portion of a frame, large regions in consecutive frames can often be reliably predicted from previous observations, and newly informative regions may emerge as objects move or actions evolve. A fixed scanning order therefore assigns comparable sequential processing priority to regions with substantially different information content, potentially introducing redundant state updates while postponing the integration of newly emerging evidence. A straightforward alternative is to prioritize patches according to saliency or motion magnitude. However, patch importance alone is insufficient for constructing an effective sequence. Independently selected high-score patches may be spatially scattered or temporally inconsistent, resulting in unnecessary long-range transitions, repeated observations, and fragmented spatiotemporal trajectories. These observations motivate an adaptive serialization mechanism that jointly considers regional informativeness, representation granularity, and sequential traversal.

To address these limitations, we propose **AdaScan**, an information-guided adaptive scanning framework that formulates visual token serialization as a spatiotemporal decision problem. AdaScan consists of three complementary components. First, **Information-Aware Weighting and Scoring (IAWS)** estimates the utility of candidate regions by jointly modeling static structural information, residual motion, and *predictive innovation*. Unlike conventional saliency or motion-based criteria, predictive innovation measures the discrepancy between the current observation and its motion-compensated prediction from preceding frames, thereby emphasizing visual evidence that cannot be adequately inferred from previous observations. Second, **Adaptive Spatial Resolution (ASR)** adjusts the spatial granularity of visual tokens according to local video activity, preserving fine-grained information in active regions while using coarser representations for relatively redundant content. Third, **Information-Guided Spatiotemporal Path Planning (IGSPP)** organizes the

candidate regions into coherent scanning trajectories by jointly considering information utility, transition consistency, and revisit redundancy. To avoid the computational burden of global trajectory optimization, a receding-horizon strategy is employed to efficiently approximate the scanning path.

Through these components, AdaScan replaces predefined visual serialization with a content-adaptive sequence whose structure is determined by the evolving information distribution of the input video. Specifically, IAWS identifies informative visual evidence, ASR determines an appropriate spatial representation granularity, and IGSP determines the sequential organization of the selected regions. Unlike approaches that modify the internal state-space operator, AdaScan operates at the serialization stage and is therefore conceptually decoupled from the underlying SSM. By restructuring the input sequence rather than redesigning the recurrent operator, AdaScan enables state evolution to focus on informative and temporally evolving visual evidence while reducing redundant traversal.

We evaluate AdaScan on three complementary video understanding tasks, including action recognition on Kinetics-400, fine-grained video understanding on Breakfast, and video semantic segmentation on VSPW. AdaScan achieves **88.5%** Top-1 accuracy on Kinetics-400, **97.6%** accuracy on Breakfast, and **58.7%** mIoU on VSPW. Extensive ablation studies verify the complementary contributions of information-aware scoring, adaptive spatial resolution, and spatiotemporal path planning. Furthermore, comparisons with random, zigzag, and score-greedy scanning strategies demonstrate that AdaScan provides a favorable balance between informative content acquisition, trajectory coherence, and redundancy reduction. These results demonstrate the effectiveness of adaptive visual serialization for state-space-based video modeling.

The main contributions of this work are summarized as follows:

- We formulate visual token serialization in video state-space models as an *information-guided spatiotemporal scanning* problem, allowing the serialization process to adapt to the non-uniform and temporally evolving information distribution of video content.
- We propose **AdaScan**, a unified adaptive scanning framework that integrates information-aware scoring with motion-compensated predictive innovation, adaptive spatial resolution, and spatiotemporal path planning. The framework jointly determines which visual regions should be processed, at what spatial granularity they should be represented, and how they should be sequentially organized.
- We validate AdaScan across action recognition, fine-grained video understanding, and video semantic segmentation. Comprehensive experiments and ablation studies demonstrate the complementary contributions of the proposed components and the effectiveness of content-adaptive scanning over conventional predefined serialization strategies.

2 Related Work

2.1 Visual State Space Models and Scanning Strategies

State space models (SSMs) provide an efficient alternative to self-attention for long-sequence modeling. Structured State Spaces (S4) [1] established strong long-range sequence modeling with structured state-space layers, while Mamba [2] introduced input-dependent selective state transitions with linear sequence complexity. Its success has motivated a growing family of visual SSMs that serialize two-dimensional visual features into one-dimensional sequences for recurrent processing. Representative approaches such as Vision Mamba, VMamba, LocalMamba, and Mamba-ND explore bidirectional, cross-directional, local, or multidimensional scanning to improve spatial information propagation [3–6]. Video extensions further exploit SSMs for efficient temporal modeling, including VideoMamba and more recent refinements such as VMP and VideoMAP [7–10].

A fundamental design choice in these models is the scan order used to serialize visual tokens. Conventional raster and zigzag scans follow predetermined coordinate-based trajectories, while space-filling curves such as Hilbert and Z-order traversal aim to preserve spatial locality. More recent visual SSMs introduce multiple scan directions, local windows, or multidimensional traversals to alleviate directional bias and preserve spatial locality [3–6]. Despite different geometric designs, these strategies are predominantly *content-independent*: the same traversal is applied regardless of which regions contain informative objects, motion, or temporal changes. As a result, geometrically adjacent but redundant background patches may be processed before distant yet task-relevant regions.

AdaScan departs from this paradigm by treating scan construction as an information-guided optimization problem rather than a fixed architectural convention. The scanning order is dynamically determined from the observed video content and balances three objectives: collecting informative evidence, maintaining spatiotemporal continuity, and avoiding redundant revisits. This makes the trajectory responsive to both spatial structure and temporal evolution instead of being determined solely by patch coordinates.

2.2 Adaptive Spatiotemporal Modeling

Transformer-based visual modeling established token sequences as a general representation for images and videos [11, 12], while non-local networks provided an earlier mechanism for capturing long-range dependencies in video [13].

Efficient video understanding introduces an additional challenge beyond spatial redundancy: the usefulness of visual evidence changes over time. Adjacent frames often contain highly correlated content, while informative events may appear only briefly or within small moving regions. Existing video models capture spatiotemporal dependencies using factorized attention, joint space-time attention, hierarchical attention, or motion-aware aggregation [14–19]. Dynamic token sparsification, token selection, early exiting, and token merging further reduce redundant visual computation [20–24]. Nevertheless, importance estimation alone does not directly determine an effective

sequential traversal. Selecting the highest-scoring patches independently may produce discontinuous transitions, repeatedly process correlated content, or overlook temporal correspondence between moving regions.

Our approach addresses this issue through *predictive innovation*. Rather than measuring temporal importance solely from appearance differences or raw motion magnitude, AdaScan evaluates how much the current representation deviates from a motion-compensated prediction based on previously observed content. Predictable regions therefore receive lower priority, whereas newly appearing or independently changing regions are emphasized. Static structure, predictive innovation, and residual motion are then combined using content-adaptive weights to capture complementary spatial and temporal cues.

AdaScan further couples this information estimate with adaptive spatial resolution and spatiotemporal path planning. The former allocates finer patches to information-rich or highly dynamic videos while using coarser representations for redundant content. The latter formulates scanning as a graph-based sequential decision problem that jointly considers information gain, transition cost, and revisit redundancy. In this respect, AdaScan differs from token pruning, saliency ranking, and fixed scanning strategies: it does not merely determine *what* visual content is important, but jointly determines *where to look*, *at what granularity*, and *in what order* visual evidence should be processed.

3 Methodology

3.1 Overview

Given a video clip $\mathcal{X} = \{X_t\}_{t=1}^T$, our goal is to convert the spatial-temporal visual features into an informative one-dimensional sequence for state-space modeling. A key observation is that the serialization order is not neutral for an SSM: because the hidden state is updated sequentially, the order in which visual regions are visited determines which information is propagated early and which dependencies must be preserved over long state transitions. Conventional raster or predefined multi-directional scans ignore this content dependence and apply the same traversal to every video.

We therefore introduce **AdaScan**, a content-adaptive scanning mechanism embedded in our **Video Temporal Selective Scan (VTSS)** block. As illustrated in Fig. 1, the standard Mamba block first extends to a visual VSS block through SS2D, while our VTSS block replaces the fixed SS2D traversal with three coordinated modules: **Information-Aware Weighting and Scoring (IAWS)**, **Adaptive Spatial Resolution (ASR)**, and **Information-Guided Spatiotemporal Path Planning (IGSPP)**. IAWS estimates *where* useful information lies, ASR determines *at what granularity* it should be represented, and IGSPP decides *in what order* the selected regions should be traversed. The resulting sequence is then processed by the remaining state-space transformation in the VTSS block.

Let $\mathcal{S}_t = \{S_t^1, \dots, S_t^{N_t}\}$ denote the patch set of frame t , and let $\tau = (s_0, s_1, \dots, s_B)$ be a scanning trajectory containing at most B transitions. We formulate adaptive

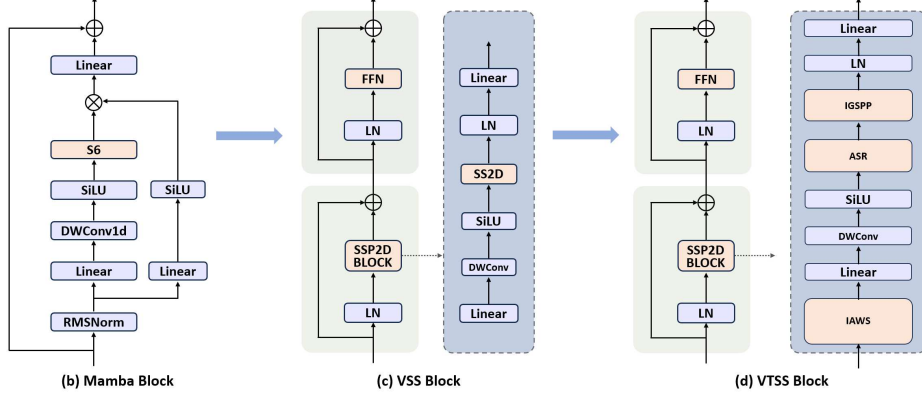


Fig. 1 Evolution from Mamba to our AdaScan-based VTSS block. (b) The standard Mamba block performs selective state-space modeling over a 1D sequence. (c) The VSS block extends Mamba to visual representations through SS2D with predefined spatial scanning routes. (d) Our VTSS block replaces fixed SS2D scanning with AdaScan, consisting of Information-Aware Weighting and Scoring (**IAWS**), Adaptive Spatial Resolution (**ASR**), and Information-Guided Spatiotemporal Path Planning (**IGSP**). The three modules jointly determine *where to look*, *at what granularity*, and *in what order*, yielding content-adaptive and temporally coherent visual sequences for state-space modeling.

scanning using a unified objective

$$J(\tau) = \sum_{k=0}^B \gamma^k \mathcal{I}_{\phi(s_k)}(s_k) - \mu \sum_{k=0}^{B-1} \mathcal{C}_{\text{trans}}(s_k, s_{k+1}) - \eta \sum_{k=1}^B M_k(s_k), \quad (1)$$

where $\phi(s_k)$ is the frame containing s_k , $\mathcal{I}$ measures information utility, $\mathcal{C}_{\text{trans}}$ measures traversal inconsistency, and $M_k(s_k) = \mathbb{I}[s_k \in \{s_0, \dots, s_{k-1}\}]$ penalizes revisits. The discount factor $\gamma \in (0, 1]$ encourages useful evidence to be observed early. Eq. (1) captures the central principle of AdaScan: a useful scan should not simply visit individually salient patches, but should collect information through a coherent and non-redundant trajectory.

3.2 Information-Aware Weighting and Scoring

The first challenge is to estimate which regions contain useful information at the current time. A purely spatial saliency measure may preserve textured objects but cannot distinguish repeated content from new temporal evidence. Conversely, motion magnitude alone may overemphasize camera movement or uninformative displacement. IAWS therefore combines three complementary cues: *static structure*, *predictive innovation*, and *residual motion*.

Static structural information.

For each frame X_t , the visual encoder produces $\mathbf{F}_t \in \mathbb{R}^{H' \times W' \times C}$. For a patch S , we measure its static information by

$$\mathcal{I}_{\text{sta}}(S) = \beta \hat{H}(S) + (1 - \beta) \hat{V}(S), \quad (2)$$

where $\hat{H}(S)$ is normalized color entropy and $\hat{V}(S)$ is normalized local intensity variance. The former measures appearance diversity, while the latter favors local boundaries and structural changes. This term prevents temporally quiet but semantically structured regions from being systematically ignored.

Motion-compensated predictive innovation.

For video scanning, information should reflect not only what is visually complex but also what is *unexpected given the past*. We estimate optical flow [25] $\mathbf{V}_{t-1 \rightarrow t}$ between consecutive frames and remove its dominant global component, producing residual flow $\mathbf{V}_{t-1 \rightarrow t}^{\text{comp}}$. Using this compensated flow, the previous feature map is warped toward the current frame:

$$\hat{\mathbf{F}}_t = \mathcal{W}(\mathbf{F}_{t-1}, \mathbf{V}_{t-1 \rightarrow t}^{\text{comp}}). \quad (3)$$

The predictive innovation of patch S is then

$$\mathcal{I}_{\text{innov}}(S) = \frac{1}{|\Omega'_S|} \sum_{p \in \Omega'_S} \left\| \mathbf{F}_t(p) - \hat{\mathbf{F}}_t(p) \right\|_2^2, \quad (4)$$

where Ω'_S denotes the corresponding feature-space support. A low innovation score indicates that the current region can already be explained by previous observations, whereas a high score indicates newly appearing, changing, or otherwise unpredictable content. Camera-motion compensation is important here because raw inter-frame displacement does not necessarily correspond to semantic novelty.

Residual motion.

Innovation captures prediction error, but explicit motion remains useful for highlighting independently evolving regions. We therefore define

$$\mathcal{I}_{\text{mot}}(S) = \frac{1}{|\Omega_S|} \sum_{p \in \Omega_S} \left\| \mathbf{V}_{t-1 \rightarrow t}^{\text{comp}}(p) \right\|_2, \quad (5)$$

where Ω_S is the image-space support of S . Unlike raw optical-flow magnitude, this residual-motion score suppresses dominant camera motion and better reflects local dynamics.

Content-adaptive weighting.

The usefulness of these cues varies across videos. Static structure is more important in slowly changing scenes, while innovation and motion become more informative for

dynamic actions. Rather than fixing their coefficients globally, IAWS predicts frame-dependent weights:

$$\mathbf{w}_t = [w_t^s, w_t^i, w_t^m]^\top = \text{softmax}(\text{MLP}(\text{GAP}(\mathbf{F}_t))). \quad (6)$$

After normalizing the three score maps within each frame, the final information score becomes

$$\mathcal{I}_t(S) = w_t^s \bar{\mathcal{I}}_{\text{sta}}(S) + w_t^i \bar{\mathcal{I}}_{\text{innov}}(S) + w_t^m \bar{\mathcal{I}}_{\text{mot}}(S). \quad (7)$$

Thus, $\mathcal{I}_t(S)$ is not merely a saliency score: it measures the utility of visiting a region given both its current visual structure and the information already available from preceding frames. The ablation results in Sec. 4.4 show that predictive innovation provides the strongest individual cue, while adaptive weighting further improves the combination of complementary information cues.

3.3 Adaptive Spatial Resolution

Knowing where information lies does not determine the appropriate spatial granularity. Using uniformly small patches preserves local details but increases sequence length and scanning cost; uniformly large patches reduce computation but may merge informative foreground regions with redundant background. We therefore introduce ASR to adapt patch resolution to the activity of the current video.

For a frame subset $\mathcal{F}$, we summarize its information activity as

$$A_{\mathcal{F}} = \mathbb{E}_{t \in \mathcal{F}} \left[\frac{1}{|\mathcal{S}_t|} \sum_{S \in \mathcal{S}_t} \mathcal{I}_t(S) (1 + \omega_S) \right], \quad (8)$$

where ω_S is the normalized residual-motion magnitude. This statistic jointly reflects visual informativeness and temporal dynamics.

We select the patch size N_p by balancing representation loss against execution cost:

$$\mathcal{C}_{\text{patch}}(N_p) = A_{\mathcal{F}} N_p^\beta + \lambda k_p P |\mathcal{F}| N_p^{\alpha-2}, \quad (9)$$

where $P = HW$ is the number of input pixels, $\beta > 0$ controls how rapidly information loss increases with patch size, and $k_p N_p^\alpha$ approximates the measured processing latency of one patch. The first term discourages coarse partitioning for information-rich clips; the second captures the computational benefit of reducing the number of patches.

For $\alpha < 2$, differentiating Eq. (9) yields the continuous optimum

$$\tilde{N}_p^* = \left[\frac{(2 - \alpha) \lambda k_p P |\mathcal{F}|}{\beta A_{\mathcal{F}}} \right]^{\frac{1}{\beta + 2 - \alpha}}. \quad (10)$$

In practice, the architecture supports a finite set of patch sizes $\mathcal{N}$, so we project the continuous solution onto the feasible set:

$$N_p^* = \Pi_{\mathcal{N}}(\tilde{N}_p^*). \quad (11)$$

Eq. (10) makes the behavior of ASR explicit: as $A_{\mathcal{F}}$ increases, the optimal N_p^* decreases, assigning finer spatial resolution to information-rich or highly dynamic videos. Conversely, predictable clips are represented with coarser patches and shorter sequences. Importantly, ASR changes the candidate set supplied to the path planner rather than independently pruning tokens after feature aggregation.

3.4 Information-Guided Spatiotemporal Path Planning

The information score specifies *where* useful evidence lies, but directly sorting patches by $\mathcal{I}_t(S)$ is not an effective scan. High-score regions can be spatially distant, and greedy selection may repeatedly jump between unrelated locations. Such discontinuities are particularly undesirable for state-space models because adjacent sequence positions define the immediate recurrent context. IGSPP therefore converts patch selection into a structured trajectory optimization problem.

Spatiotemporal graph construction.

After ASR determines the spatial partition, we construct a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Each node $u \in \mathcal{V}$ corresponds to a patch and stores its frame index $\phi(u)$, center coordinate $\mathbf{x}_u$, and information score $\mathcal{I}_{\phi(u)}(u)$. We include spatial edges between neighboring patches within the same frame and temporal edges between motion-consistent candidates in adjacent frames. For a temporal transition $u \rightarrow v$, motion inconsistency is

$$d_{\text{motion}}(u, v) = \left\| \mathbf{x}_v - \left(\mathbf{x}_u + \bar{\mathbf{V}}_u^{\text{comp}} \right) \right\|_2, \quad (12)$$

where $\bar{\mathbf{V}}_u^{\text{comp}}$ is the mean compensated displacement within patch u . We use the affinity

$$\mathcal{A}(u, v) = \mathcal{I}_{\phi(u)}(u) \mathcal{I}_{\phi(v)}(v) \exp(-\rho d_{\text{motion}}(u, v)) \quad (13)$$

to retain temporally plausible candidates. This favors correspondences that are simultaneously informative and consistent with observed motion.

Transition regularization.

For each valid edge, the traversal cost is

$$\mathcal{C}_{\text{trans}}(u, v) = \lambda_s d_{\text{spatial}}(u, v) + \lambda_t d_{\text{temporal}}(u, v) + \lambda_m d_{\text{motion}}(u, v), \quad (14)$$

where d_{spatial} measures normalized displacement on the image plane, d_{temporal} measures frame distance, and d_{motion} is given by Eq. (12). The transition term therefore discourages arbitrary long-range jumps while still permitting them when the expected information gain is sufficiently large. Together with the revisit penalty in Eq. (1), this separates AdaScan from score-only token ranking: the quality of a patch depends not only on its individual score but also on how naturally it extends the current trajectory.

Receding-horizon optimization.

Globally maximizing Eq. (1) over all possible trajectories is combinatorial. A purely greedy policy is inexpensive but ignores whether the current choice leads to poor future

transitions. We therefore use a receding-horizon solver with look-ahead depth L . At step k , moving from the current node u to candidate v receives

$$R_k(u, v) = \gamma^{k+1} \mathcal{I}_{\phi(v)}(v) - \mu \mathcal{C}_{\text{trans}}(u, v) - \eta M_k(v). \quad (15)$$

For a horizon of ℓ steps, the value function is

$$V_k^{(0)}(u) = 0, \quad V_k^{(\ell)}(u) = \max_{v \in \mathcal{N}(u)} \left[R_k(u, v) + V_{k+1}^{(\ell-1)}(v) \right]. \quad (16)$$

Only the first action of the locally optimized plan is executed:

$$s_{k+1}^* = \arg \max_{v \in \mathcal{N}(s_k)} \left[R_k(s_k, v) + V_{k+1}^{(L-1)}(v) \right]. \quad (17)$$

The visited set is then updated and the optimization is repeated. When $L = 1$, IGSP reduces to greedy information-aware scanning. For $L > 1$, it can sacrifice a locally optimal patch when another transition provides a better route to future informative regions, where score-greedy scanning collects informative patches but suffers from substantially larger transition cost and long-range jumps.

3.5 VTSS Integration and Sequence Construction

We finally describe how AdaScan is integrated into the visual state-space block. Let $\mathcal{P}_t(N_p^*)$ denote the patch representation of frame t under the ASR-selected resolution. IAWS assigns an information score to each candidate, and IGSP returns an ordered trajectory

$$\tau^* = (s_0^*, s_1^*, \dots, s_B^*). \quad (18)$$

The corresponding features are serialized as

$$\mathbf{Z}_\tau = [\mathbf{z}(s_0^*); \mathbf{z}(s_1^*); \dots; \mathbf{z}(s_B^*)], \quad (19)$$

where $\mathbf{z}(s)$ is the feature associated with patch s . This ordered sequence replaces the predefined SS2D serialization in the conventional VSS block, as shown in Fig. 1.

Given an input feature $\mathbf{X}$, we summarize the VTSS transformation as

$$\mathbf{Y} = \mathbf{X} + \mathcal{P}_{\text{out}}(\text{SSM}(\text{AdaScan}(\mathbf{X}))), \quad (20)$$

where $\text{AdaScan}(\cdot)$ represents the IAWS–ASR–IGSP serialization process and $\mathcal{P}_{\text{out}}$ denotes the output projection. The surrounding normalization, depth-wise convolution, activation, residual connection, and FFN follow the visual SSM backbone and are illustrated in Fig. 1. Thus, our contribution changes the *structure of the sequence presented to the state-space operator*, rather than replacing the state-space operator itself.

This design has two useful consequences. First, AdaScan is conceptually modular: the scoring and trajectory construction can be inserted into a visual SSM while preserving its standard feature transformation and residual architecture. Second, all

three components operate toward the same goal. IAWS estimates information utility, ASR controls the number and granularity of candidate regions, and IGSP converts these candidates into a sequential representation whose ordering explicitly reflects spatial structure and temporal evolution. The resulting VTSS block therefore provides a content-adaptive alternative to fixed visual scanning while retaining the linear sequential modeling advantages of state-space architectures.

4 Experiments

4.1 Experimental Set up

We evaluate AdaScan on three complementary video benchmarks: Kinetics-400 (K400) [26] for action recognition, Breakfast [27] for fine-grained video understanding, and VSPW [28] for video semantic segmentation. We report Top-1 classification accuracy on K400, frame-level accuracy on Breakfast, and mean Intersection-over-Union (mIoU) on VSPW. Unless otherwise specified, all ablation studies are conducted on K400 using the same backbone, training protocol, and evaluation setting to isolate the contribution of each component.

Since the compared methods differ in backbone capacity, number of input frames, and pretraining strategy, Table 1 reports these configurations whenever available. The comparisons are therefore intended to provide a broad assessment across representative video modeling approaches rather than a strictly controlled comparison under identical architectures and pretraining conditions.

4.2 Main Results

Table 1 summarizes the results on the three benchmarks. On K400, AdaScan achieves **88.5%** Top-1 accuracy with 43M parameters and 8 input frames. The result compares favorably with the representative video SSM and Transformer-based methods included in the table, including substantially larger models and methods using more input frames.

On Breakfast, AdaScan obtains **97.6%** frame-level accuracy using 8 input frames. The effectiveness of the proposed scanning strategy also extends to dense video prediction: on VSPW, AdaScan achieves **58.7%** mIoU with an 8-frame input.

Although the three benchmarks differ substantially in task formulation, the results consistently show that content-adaptive scanning can be applied to classification, fine-grained temporal understanding, and dense semantic prediction. This cross-task behavior suggests that the benefit of AdaScan is not restricted to a single video recognition setting.

4.3 Qualitative Analysis

Figure 2 compares the trajectories generated by different scanning strategies at three representative timestamps. Zigzag Scan follows the same coordinate-dependent route regardless of the evolving scene content, resulting in systematic coverage but also substantial traversal of visually redundant regions. Random Scan removes this

Table 1 Main results on Kinetics-400, Breakfast, and VSPW. “–” denotes unavailable or non-comparable entries. † denotes masked pretraining and ‡ denotes large-scale pretraining. Since different methods may use different backbones, pretraining strategies, and numbers of input frames, the results should be interpreted together with the corresponding model configurations. The best reported result within each dataset block is shown in bold.

Dataset	Method	Venue	Params	Frames	Metric (%) ↑
K400	VideoMamba-M [7]	ECCV’24	74M	16	81.9
	VideoMamba-M [†] [7]	ECCV’24	74M	16	83.4
	VMP-M [8]	ICCV’25	72M	32	84.0
	VMP-M [‡] [8]	ICCV’25	72M	32	86.5
	VideoMAP-M [9]	arXiv’25	96M	8	85.5
	VideoMAP-L [9]	arXiv’25	301M	8	87.3
	VideoMamba++ [10]	PR’25	75M	8	84.5
	MoMa (ViT-L/14)	ICML’25	39M	8	87.8
	Ours	–	43M	8	88.5
Breakfast	VideoMamba-S [7]	ECCV’24	26M	64	97.4
	VideoMamba-M [†] [7]	ECCV’24	74M	64	96.9
	VMP-M [‡] [8]	ICCV’25	72M	32	92.4
	VideoMAP-M [9]	arXiv’25	96M	24	96.7
	VideoMAP-L [9]	arXiv’25	301M	8	95.2
	VideoMamba++ [10]	PR’25	75M	8	93.2
	MoMa (ViT-B/16)	ICML’25	11M	64	93.2
	TimeSformer-L [14]	ICML’21	121M	96	95.1
	Ours	–	43M	8	97.6
VSPW	SegFormer (MiT-B2) [29]	NeurIPS’21	24.8M	8	43.9
	CFM (MiT-B5) [30]	CVPR’22	85.5M	8	49.3
	MPVSS (Swin-S) [31]	NeurIPS’23	108.0M	16	40.4
	MPVSS (Swin-B) [31]	NeurIPS’23	147.0M	16	52.6
	MPVSS (Swin-L) [31]	NeurIPS’23	255.4M	8	53.9
	TV3S (MiT-B2) [32]	CVPR’25	28.3M	8	46.3
	TV3S (Swin-S) [32]	CVPR’25	46M	8	50.6
	RS-SSM (MiT-B5) [33]	CVPR’26	88.4M	8	51.6
	RS-SSM (Swin-S) [33]	CVPR’26	55.9M	8	52.2
	Ours	–	43M	8	58.7

deterministic bias but produces fragmented trajectories with frequent long-range transitions.

InfoScan preferentially visits informative regions, but selecting regions primarily according to their individual importance does not guarantee a coherent sequential trajectory. Consequently, transitions between high-score regions can still span distant spatial locations. In contrast, AdaScan jointly considers regional informativeness and transition consistency. As the scene evolves, its trajectory shifts toward informative foreground regions while maintaining substantially more coherent transitions between successive observations. This qualitative behavior is consistent with the quantitative scanning-policy analysis in Table 3.

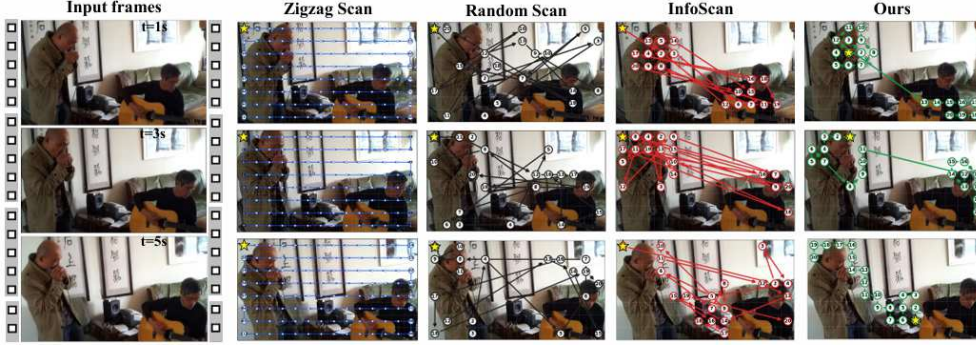


Fig. 2 Visualization of scanning trajectories produced by different strategies at three representative timestamps ($t = 1$ s, 3 s, and 5 s). From left to right: input frames, Zigzag Scan, Random Scan, InfoScan, and AdaScan. Zigzag Scan follows a predefined content-independent route, whereas Random Scan produces irregular traversal. InfoScan prioritizes informative regions but can introduce long-range transitions when high-score regions are spatially separated. AdaScan jointly considers information utility and transition consistency, producing more coherent trajectories while retaining informative visual evidence.

4.4 Ablation Studies and Analysis

We conduct systematic ablation studies on Kinetics-400 to examine the design choices of AdaScan. Unless otherwise specified, all variants use the same backbone and evaluation protocol. We investigate four aspects: (1) the contribution of the major components of AdaScan; (2) the complementarity of the information cues used by IAWS; (3) the formulation of predictive innovation; and (4) the effect of different scanning policies.

Component ablation.

Table 2(a) evaluates the contributions of Information Scoring (IS), Adaptive Resolution (AR), and Path Planning (PP). The fixed-resolution raster baseline achieves 82.1% Top-1 accuracy with 24.8 GFLOPs and a latency of 31.7 ms. Introducing IS improves the accuracy to 83.8%, indicating that prioritizing informative regions is beneficial even without modifying the spatial resolution or traversal order.

Combining IS with AR further improves the accuracy to 84.3%. This configuration reduces the measured latency from 33.1 ms to 27.4 ms relative to IS alone, although its nominal computation increases from 24.8 to 27.2 GFLOPs. In contrast, combining IS with PP increases the accuracy to 86.9%, demonstrating that explicitly organizing informative regions into a structured scanning trajectory provides a substantial benefit. This improvement is accompanied by additional computation when adaptive resolution is disabled, as reflected by the 45.5 GFLOPs of the IS+PP variant.

Integrating all three components yields the complete AdaScan model, which achieves 88.5% Top-1 accuracy with 23.0 GFLOPs and a measured latency of 29.2 ms. Compared with the fixed-resolution raster baseline, the complete model improves accuracy by 6.4 percentage points while reducing the nominal computation from 24.8

to 23.0 GFLOPs and the measured latency from 31.7 to 29.2 ms. These results indicate that the three components are complementary: IS identifies informative content, AR regulates its representation granularity, and PP determines an effective sequential traversal.

Information-cue ablation.

Table 2(b) analyzes the contributions of the information cues used by IAWS. Among the individual cues, predictive innovation (I) achieves the highest Top-1 accuracy of 86.1%, followed by residual motion (M) at 85.3% and static structure (S) at 84.2%. This result indicates that visual evidence that cannot be sufficiently predicted from preceding observations provides a particularly informative signal for adaptive video scanning.

Combining multiple cues consistently improves performance over the corresponding individual cues. Specifically, S+I, S+M, and I+M achieve 87.4%, 86.5%, and 87.9% Top-1 accuracy, respectively. Among the two-cue combinations, I+M performs best, further demonstrating the importance of temporal information for identifying informative video regions. Incorporating all three cues yields 88.2%, confirming that static structure provides complementary spatial information in addition to predictive innovation and residual motion.

Finally, introducing adaptive weighting (AW) further improves the accuracy from 88.2% to **88.5%**. Rather than assigning fixed contributions to the three information cues, AW dynamically adjusts their relative importance according to the current video content. The resulting improvement supports the use of content-adaptive cue integration for handling variations in spatial structure and temporal dynamics across video sequences.

Predictive-innovation design.

Table 2(c) evaluates different formulations of predictive innovation. Direct frame difference achieves 86.8% Top-1 accuracy, while measuring temporal discrepancy in feature space improves the result to 87.5%. Introducing flow-guided feature warping further increases the accuracy to 88.4%, indicating that explicitly establishing temporal correspondence helps distinguish genuine visual innovation from apparent changes caused by object displacement.

Camera-motion compensation (CMC) provides a further improvement to 88.5%. By accounting for dominant global motion before evaluating local temporal discrepancy, the resulting innovation score is less affected by camera movement and better reflects visual evidence that cannot be explained by preceding observations. These results support the use of motion-compensated predictive innovation as the temporal information cue in IAWS.

Scanning-policy analysis.

We further isolate the effect of trajectory construction by comparing different scanning policies under the same backbone, patch partition, and scanning budget. Table 3 reports both task performance and trajectory characteristics. Random scanning achieves 82.7% accuracy but exhibits a relatively high path cost ($C_{\text{path}} = 47.3$)

(a) Component ablation

Table 2 Ablation studies on Kinetics-400. (a) Contributions of Information Scoring (IS), Adaptive Resolution (AR), and Path Planning (PP). (b) Effects of static structure (S), predictive innovation (I), residual motion (M), and adaptive weighting (AW). (c) Alternative formulations of predictive innovation, where FW and CMC denote feature warping and camera-motion compensation, respectively.

Setting	IS	AR	PP	Top-1 (%) $\uparrow$	GFLOPs $\downarrow$	Latency (ms) $\downarrow$
Baseline	–	–	–	82.1	24.8	31.7
+ IS	✓	–	–	83.8	24.8	33.1
+ IS + AR	✓	✓	–	84.3	27.2	27.4
+ IS + PP	✓	–	✓	86.9	45.5	35.0
Full model	✓	✓	✓	88.5	23.0	29.2

(b) Information cues

Variant	S	I	M	AW	Top-1 (%) $\uparrow$
Static only	✓	–	–	–	84.2
Innovation only	–	✓	–	–	86.1
Motion only	–	–	✓	–	85.3
S + I	✓	✓	–	–	87.4
S + M	✓	–	✓	–	86.5
I + M	–	✓	✓	–	87.9
S + I + M	✓	✓	✓	–	88.2
S + I + M + AW	✓	✓	✓	✓	88.5

(c) Predictive innovation

Variant	FW	CMC	Top-1 (%) $\uparrow$
Frame difference	–	–	86.8
Feature difference	–	–	87.5
Flow warping	✓	–	88.4
Compensated warping	✓	✓	88.5

FW: feature warping. CMC: camera-motion compensation.

and frequent long-range transitions ($R_{\text{jump}} = 43.6$), reflecting its lack of spatial and informational guidance.

Zigzag scanning follows a regular spatial route and consequently achieves the lowest path cost (19.0) and eliminates long-range jumps. However, this regularity comes at the cost of the highest redundant-visit ratio (38.7) and relatively limited information gain ($G_{\text{info}} = 7.16$). In contrast, Greedy scanning prioritizes locally high-scoring regions and increases the information gain to 9.73, but produces the largest path cost (55.8) and jump ratio (51.3). This observation indicates that maximizing instantaneous information utility alone does not necessarily produce an effective serialization order.

Table 3 Scanning-policy analysis on K400. All policies use the same backbone, patch partition, and scanning budget. G_{info} denotes accumulated information gain, C_{path} trajectory cost, R_{red} redundant visits, R_{jump} long-range transitions, and E_{info} information efficiency.

Policy	Acc. $\uparrow$	G_{info} $\uparrow$	C_{path} $\downarrow$	R_{red} $\downarrow$	R_{jump} $\downarrow$	E_{info} $\uparrow$
Random	82.7	6.84	47.3	31.8	43.6	0.145
Zigzag	83.5	7.16	19.0	38.7	0.0	0.377
Greedy	86.2	9.73	55.8	20.4	51.3	0.174
Ours	88.5	10.41	24.6	10.8	5.6	0.423

IGSPP jointly considers information utility and transition consistency. It obtains the highest accumulated information gain (10.41) and information efficiency (0.423), while reducing the redundant-visit ratio to 10.8 and the long-range jump ratio to 5.6. Its path cost (24.6) is slightly higher than that of Zigzag scanning but substantially lower than those of Random and Greedy scanning. Together with the highest Top-1 accuracy of 88.5%, these results show that effective scanning requires balancing information acquisition with trajectory coherence rather than optimizing either criterion in isolation.

Summary.

The ablation studies provide complementary evidence for the design of AdaScan. Predictive innovation provides the strongest individual information cue, while its combination with static structure and residual motion yields further improvements. Adaptive weighting provides an additional, albeit modest, gain by allowing the relative importance of these cues to vary with video content. At the system level, adaptive resolution and path planning complement information scoring by controlling representation granularity and sequential organization, respectively. Finally, the scanning-policy comparison shows that information acquisition alone is insufficient: effective serialization also requires controlling trajectory continuity and redundant traversal. Together, these components enable the complete AdaScan model to improve both recognition accuracy and computational efficiency over the fixed-resolution raster baseline.

5 Conclusion

We introduce **AdaScan**, a content-adaptive spatiotemporal scanning framework that jointly determines *where to look*, *at what granularity*, and *in what order* for efficient video state-space modeling. By integrating information-aware scoring, adaptive spatial resolution, and spatiotemporal path planning, AdaScan prioritizes informative visual evidence while reducing redundant and incoherent scanning. Extensive experiments and ablations demonstrate the effectiveness and generalizability of AdaScan across diverse video understanding tasks.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by: Zhangjiakou Science and Technology Plan Project in 2025 (2522236D);Hebei Province Medical Science Research Project Plan in 2025 (20250148);Tianjin Key Medical Discipline Construction Project (Grant No.TJYXZDXK-3-005A);The First Hospital of Jilin University (grant number JDYY-DEP-2024023) and the 16th Youth Project of The First Hospital of Jilin University (grant number JDYY16202515)..

Data availability

This work has not developed any new dataset, it has used three existing public datasets.

References

- [1] Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. In: International Conference on Learning Representations (2022)
- [2] Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
- [3] Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. In: Proceedings of the 41st International Conference on Machine Learning, pp. 62429–62442 (2024)
- [4] Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y.: Vmamba: Visual state space model. In: Advances in Neural Information Processing Systems, vol. 37 (2024)
- [5] Huang, T., Pei, X., You, S., Wang, F., Qian, C., Xu, C.: Localmamba: Visual state space model with windowed selective scan. In: European Conference on Computer Vision Workshops, pp. 12–22 (2024). https://doi.org/10.1007/978-3-031-91979-4_2
- [6] Li, S., Singh, H., Grover, A.: Mamba-nd: Selective state space modeling for multi-dimensional data. arXiv preprint arXiv:2402.05892 (2024)
- [7] Li, K., Li, X., Wang, Y., He, Y., Wang, Y., Wang, L., Qiao, Y.: Video-mamba: State space model for efficient video understanding. In: European

- Conference on Computer Vision, pp. 237–255 (2024). https://doi.org/10.1007/978-3-031-73347-5_14
- [8] Lu, H., Salah, A.A., Poppe, R.: Snakes and ladders: Two steps up for videomamba. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 24234–24244 (2025)
 - [9] Liu, Y., Wu, P., Liang, C., Shen, J., Wang, L., Yi, L.: Videomap: Toward scalable mamba-based video autoregressive pretraining. arXiv preprint arXiv:2503.12332 (2025)
 - [10] Song, X., Tian, W., Zhu, Q., Zhang, X.: Videomamba++: Integrating state space model with dual attention for enhanced video understanding. Image and Vision Computing **161**, 105609 (2025) <https://doi.org/10.1016/j.imavis.2025.105609>
 - [11] Dosovitskiy, A., Beyer, L., Kolesnikov, A., *et al.*: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
 - [12] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers and distillation through attention. In: International Conference on Machine Learning, pp. 10347–10357 (2021)
 - [13] Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)
 - [14] Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: Proceedings of the 38th International Conference on Machine Learning, pp. 813–824 (2021)
 - [15] Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6836–6846 (2021)
 - [16] Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., Feichtenhofer, C.: Multiscale vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6824–6835 (2021)
 - [17] Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3202–3211 (2022)
 - [18] Patrick, M., Campbell, D., Asano, Y.M., Misra, I., Metze, F., Feichtenhofer, C., Vedaldi, A., Henriques, J.F.: Keeping your eye on the ball: Trajectory attention in video transformers. In: Advances in Neural Information Processing Systems, vol. 34 (2021)

- [19] Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6202–6211 (2019)
- [20] Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., Hsieh, C.-J.: Dynamicvit: Efficient vision transformers with dynamic token sparsification. In: Advances in Neural Information Processing Systems, vol. 34, pp. 13937–13949 (2021)
- [21] Ryoo, M.S., Piergiovanni, A.J., Arnab, A., Dehghani, M., Angelova, A.: Token-learner: What can 8 learned tokens do for images and videos? In: Advances in Neural Information Processing Systems, vol. 34 (2021)
- [22] Liang, Y., Ge, C., Tong, Z., Song, Y., Wang, J., Xie, P.: Not all patches are what you need: Expediting vision transformers via token reorganizations. In: International Conference on Learning Representations (2022)
- [23] Yin, H., Vahdat, A., Alvarez, J.M., Mallya, A., Kautz, J., Molchanov, P.: A-vit: Adaptive tokens for efficient vision transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10809–10818 (2022)
- [24] Bolya, D., Fu, C.-Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your vit but faster. In: International Conference on Learning Representations (2023)
- [25] Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European Conference on Computer Vision, pp. 402–419 (2020). https://doi.org/10.1007/978-3-030-58536-5_24
- [26] Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017)
- [27] Kuehne, H., Arslan, A., Serre, T.: The language of actions: Recovering the syntax and semantics of goal-directed human activities. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2014)
- [28] Miao, J., Wei, Y., Wu, Y., Liang, C., Li, G., Yang, Y.: Vspw: A large-scale dataset for video scene parsing in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4133–4143 (2021)
- [29] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. In: Advances in Neural Information Processing Systems (2021)
- [30] Sun, G., Liu, Y., Ding, H., Probst, T., Van Gool, L.: Coarse-to-fine feature mining for video semantic segmentation. In: IEEE/CVF Conference on Computer Vision

and Pattern Recognition (2022)

- [31] Weng, Y., Han, M., He, H., Li, M., Yao, L., Chang, X., Zhuang, B.: Mask propagation for efficient video semantic segmentation. In: Advances in Neural Information Processing Systems, vol. 36 (2023)
- [32] Hesham, S.A.S., Liu, Y., Sun, G., Ding, H., Yang, J., Konukoglu, E., Geng, X., Jiang, X.: 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)
- [33] Zhu, K., Cui, Z., Zang, Z., Zhou, J.: Rs-ssm: Refining forgotten specifics in state space model for video semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10741–10752 (2026)