PREDICTIVE SCANNING FOR EFFICIENT SPATIOTEMPORAL REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

State-space models (SSMs) provide an efficient alternative for video modeling, but most existing approaches serialize visual tokens using fixed, content-independent scanning orders. Such predefined traversals are poorly matched to videos, where informative evidence is spatially sparse, temporally varying, and often highly redundant across frames. We propose **VScan**, an information-guided adaptive scanning framework that jointly determines *where to look, at what granularity, and in what order* visual regions should be processed. VScan consists of three complementary components: an information-aware scoring mechanism that combines static structure, residual motion, and motion-compensated predictive innovation; an adaptive spatial-resolution strategy that allocates finer representations to more informative content; and an information-guided spatiotemporal path planner that organizes selected regions into coherent and non-redundant trajectories. Rather than modifying the underlying state-space operator, VScan adapts the sequence presented to it according to the spatial and temporal information distribution of each video. Experiments on action recognition, fine-grained video understanding, and video semantic segmentation show consistent improvements, achieving **88.5%** Top-1 accuracy on Kinetics-400, **97.6%** accuracy on Breakfast, and **65.8%** mIoU on VSPW. Extensive ablations further demonstrate the complementary roles of information scoring, adaptive resolution, and path planning, while controlled scanning-policy comparisons show that VScan achieves a better trade-off between information acquisition, traversal continuity, and redundant computation than conventional scanning strategies.

1 INTRODUCTION

Video understanding requires models to capture both spatial semantics and temporal evolution while avoiding substantial redundancy across consecutive frames. Recent state-space models (SSMs), including structured sequence models and Mamba-style selective SSMs Gu et al. (2022); Gu and Dao (2023), have emerged as an efficient alternative to attention-based architectures due to their favorable scaling with sequence length. However, applying sequential state-space modeling to visual data introduces a fundamental design choice: *in what order should visual tokens be processed?* Existing visual SSMs typically rely on predefined serialization strategies, such as bidirectional, cross-directional, local, or multi-dimensional scans Zhu et al. (2024); Liu et al. (2024); Huang et al. (2024); Li et al. (2024b). Although computationally convenient, these strategies are largely content-independent and impose the same traversal pattern regardless of how information is distributed across a video.

This fixed-scanning paradigm is poorly matched to natural video. Informative evidence is often spatially sparse, temporally evolving, and highly redundant across adjacent frames: salient objects may occupy only a small fraction of the scene, large portions of the video can be predictable from preceding observations, and new evidence may emerge only briefly as an action develops. A predefined traversal can therefore allocate sequential modeling capacity to redundant regions while delaying access to more informative content. Simply ranking patches by saliency or motion is not sufficient either, because individually important patches may still form a poor sequence, leading to abrupt spatial jumps, repeated visits to correlated regions, or temporally inconsistent trajectories. We therefore argue that effective video scanning should jointly determine *where to look, at what spatial granularity, and in what order* visual evidence should be processed.

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.

To this end, we propose **VScan**, an information-guided adaptive scanning framework that formulates visual serialization as a spatiotemporal decision problem. VScan consists of three complementary components. First, **Information-Aware Weighting and Scoring (IAWS)** estimates patch utility using static structure, residual motion, and *predictive innovation*. Unlike conventional saliency or raw motion magnitude, predictive innovation measures how much the current representation cannot be explained by a motion-compensated prediction from previous observations, thereby emphasizing newly emerging information. Second, **Adaptive Spatial Resolution (ASR)** adjusts patch granularity according to video content, allocating finer representations to more informative regions while avoiding unnecessary computation on redundant content. Third, **Information-Guided Spatiotemporal Path Planning (IGSPP)** organizes candidate regions into coherent trajectories by jointly considering information gain, transition consistency, and revisit redundancy, with a receding-horizon strategy used to approximate the trajectory optimization efficiently.

Together, these components replace fixed visual serialization with an input-dependent sequence tailored to the spatial and temporal information distribution of each video. IAWS determines *where* useful evidence lies, ASR determines *how finely* it should be represented, and IGSPP determines *how* informative regions should be traversed. Importantly, VScan does not modify the underlying state-space operator. Instead, it changes the structure of the sequence presented to the operator, allowing recurrent state propagation to follow a more informative and less redundant visual trajectory.

We evaluate VScan on three complementary video benchmarks: Kinetics-400 for action recognition, Breakfast for fine-grained video understanding, and VSPW for video semantic segmentation. VScan achieves **88.5%** Top-1 accuracy on Kinetics-400, **97.6%** accuracy on Breakfast, and **65.8%** mIoU on VSPW. Extensive ablations show that information scoring, adaptive resolution, and path planning provide complementary benefits. Controlled comparisons with random, zigzag, and score-greedy scanning further show that VScan achieves a stronger balance between information acquisition, traversal continuity, and redundant computation than conventional scanning policies.

Our contributions are summarized as follows: ❶We formulate visual token serialization for video state-space models as an *information-guided spatiotemporal scanning* problem, replacing fixed traversal patterns with input-dependent scanning decisions that adapt to the evolving information distribution of video content. ❷We introduce VScan, a unified framework that jointly determines *where to look, at what granularity, and in what order* by combining information-aware scoring with motion-compensated predictive innovation, adaptive spatial resolution, and spatiotemporal path planning. ❸ We demonstrate the effectiveness of adaptive scanning across action recognition, fine-grained video understanding, and video semantic segmentation, together with controlled ablations and policy analyses that validate the contributions of its information modeling and trajectory-design components.

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) Gu et al. (2022) established strong long-range sequence modeling with structured state-space layers, while Mamba Gu and Dao (2023) 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 Zhu et al. (2024); Liu et al. (2024); Huang et al. (2024); Li et al. (2024b). Video extensions further exploit SSMs for efficient temporal modeling, including VideoMamba and more recent refinements such as VMP and VideoMAP Li et al. (2024a); Lu et al. (2025); Liu et al. (2025); Song et al. (2025).

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 Zhu et al. (2024); Liu et al. (2024); Huang et al.

(2024); Li et al. (2024b). 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.

VScan 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 Dosovitskiy et al. (2021); Touvron et al. (2021), while non-local networks provided an earlier mechanism for capturing long-range dependencies in video Wang et al. (2018).

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 Bertasius et al. (2021); Arnab et al. (2021); Fan et al. (2021); Liu et al. (2022); Patrick et al. (2021); Feichtenhofer et al. (2019). Dynamic token sparsification, token selection, early exiting, and token merging further reduce redundant visual computation Rao et al. (2021); Ryoo et al. (2021); Liang et al. (2022); Yin et al. (2022); Bolya et al. (2023). 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, VScan 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.

VScan 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, VScan 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 **VScan**, 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

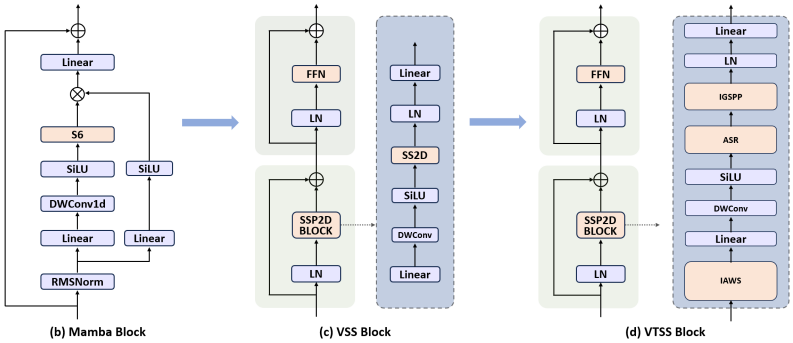

Figure 1: **Evolution from Mamba to our VScan-based VTSS block.** (b) The standard Mamba block performs selective state-space modeling over a 1D sequence. (c) The VSS block extends Mamba to visual representations through SS2D with predefined spatial scanning routes. (d) Our VTSS block replaces fixed SS2D scanning with VScan, consisting of Information-Aware Weighting and Scoring (**IAWS**), Adaptive Spatial Resolution (**ASR**), and Information-Guided Spatiotemporal Path Planning (**IGSPP**). 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.

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

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

where $\phi(s_k)$ is the frame containing s_k, $\mathcal{I}$ measures information utility, $\mathcal{C}_{\text{trans}}$ measures traversal inconsistency, and $M_k(s_k) = \mathbb{I}[s_k \in \{s_0, \ldots, s_{k-1}\}]$ penalizes revisits. The discount factor $\gamma \in (0, 1]$ encourages useful evidence to be observed early. Eq. (1) captures the central principle of VScan: 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

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

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 Teed and Deng (2020) $\mathbf{V}_{t-1\to t}$ between consecutive frames and remove its dominant global component, producing residual flow $\mathbf{V}^{\text{comp}}_{t-1\to t}$. Using this compensated flow, the previous feature map is warped toward the current frame:

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

The predictive innovation of patch S is then

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

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

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

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:

$$\mathbf{w}_t = [w^s_t, w^i_t, w^m_t]^\top = \text{softmax}\left(\text{MLP}\left(\text{GAP}(\mathbf{F}_t)\right)\right). \tag{6}$$

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

$$\mathcal{I}_t(S) = w^s_t\overline{\mathcal{I}}_{\text{sta}}(S) + w^i_t\overline{\mathcal{I}}_{\text{innov}}(S) + w^m_t\overline{\mathcal{I}}_{\text{mot}}(S). \tag{7}$$

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 ablations in Sec. **??** show that predictive innovation is the strongest individual cue and that adaptive weighting further improves their combination.

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

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

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:

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

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

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

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

$$N_p^{\star} = \Pi_{\mathcal{N}} \left(\widetilde{N}_p^{\star} \right). \tag{11}$$

Eq. (10) makes the behavior of ASR explicit: as $A_{\mathcal{F}}$ increases, the optimal $N_p^{\star}$ 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 $u \to v$, motion inconsistency is

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

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

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

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

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

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 VScan 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

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

For a horizon of ℓ steps, the value function is

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

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

$$s^{\star}_{k+1} = \arg \max_{v \in \mathcal{N}(s_k)} \left[R_k(s_k, v) + V^{(L-1)}_{k+1}(v) \right]. \tag{17}$$

The visited set is then updated and the optimization is repeated. When $L = 1$, IGSPP 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. This property explains the behavior observed in Sec. **??**, 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 VScan is integrated into the visual state-space block. Let $\mathcal{P}_t(N^{\star}_p)$ denote the patch representation of frame t under the ASR-selected resolution. IAWS assigns an information score to each candidate, and IGSPP returns an ordered trajectory

$$\tau^{\star} = (s^{\star}_0, s^{\star}_1, \ldots, s^{\star}_B). \tag{18}$$

The corresponding features are serialized as

$$\mathbf{Z}_{\tau} = [\mathbf{z}(s^{\star}_0); \mathbf{z}(s^{\star}_1); \ldots; \mathbf{z}(s^{\star}_B)], \tag{19}$$

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

$$\mathbf{Y} = \mathbf{X} + \mathcal{P}_{\text{out}} \left(\text{SSM} \left(\text{VScan}(\mathbf{X}) \right) \right), \tag{20}$$

where $\text{VScan}(\cdot)$ represents the IAWS–ASR–IGSPP 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, VScan 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 IGSPP 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 Setup

We evaluate VScan on three complementary video understanding benchmarks: Kinetics-400 (K400) Carreira and Zisserman (2017) for action recognition, Breakfast Kuehne et al. (2014) for fine-grained video understanding, and VSPW Miao et al. (2021) for video semantic segmentation. We report Top-1 accuracy on K400, frame-level accuracy on Breakfast, and mIoU on VSPW. The three benchmarks cover complementary settings, ranging from clip-level recognition to temporally fine-grained understanding and dense frame-level prediction, allowing us to examine whether adaptive scanning generalizes beyond a single video task. For readability, the main paper reports representative SSM- and Transformer-based competitors, while additional comparisons and implementation details are provided in the Appendix.

4.2 Main Results

Table 1 summarizes the results across the three benchmarks.

Action recognition. On Kinetics-400, VScan achieves **88.5%** Top-1 accuracy using 43M parameters and only 8 input frames. Under this compact observation budget, it outperforms the representative video SSM and Transformer-based models listed in Table 1. In particular, VScan improves over VideoMAP-L by 1.2 percentage points while using substantially fewer parameters (43M vs. 301M), and also surpasses MoMa (ViT-L/14), which uses the same number of input frames. These results suggest that reorganizing the visual sequence according to content-dependent information can improve state-space video modeling without requiring a larger model or a longer temporal input.

Fine-grained video understanding. On Breakfast, VScan obtains **97.6%** frame-level accuracy with 8 input frames. The result is competitive with, and in most cases higher than, methods that process considerably longer clips. For example, VideoMamba-S reaches 97.4% using 64 frames, whereas VScan achieves 97.6% with an eight-frame input. This result indicates that the proposed scanning strategy can prioritize temporally informative evidence even under a relatively compact observation budget.

Video semantic segmentation. The improvement also transfers to dense prediction. On VSPW, VScan achieves **65.8%** mIoU with 43M parameters and 8 input frames, outperforming the representative segmentation baselines reported in Table 1. The gain is notable because dense prediction requires preserving both spatial detail and temporal consistency rather than only aggregating clip-level semantics. The strong performance on VSPW therefore provides complementary evidence that content-adaptive serialization can benefit tasks with different output granularities.

Overall, the results across K400, Breakfast, and VSPW show that the benefit of VScan is not restricted to a particular prediction objective. The same principle—adapting where to look, at what granularity, and in what order—is effective for recognition, fine-grained temporal understanding, and dense video prediction.

4.3 Ablation and Analysis

We next analyze the three design choices of VScan on Kinetics-400 under the same backbone and evaluation protocol. Unless otherwise specified, all variants use the same training setting. We study (i) the contribution of information scoring, adaptive resolution, and path planning; (ii) the information cues used by IAWS; and (iii) the construction of predictive innovation. We further isolate the scanning policy by comparing alternative trajectories under the same backbone, partition, and scanning budget.

Contribution of individual components. Table 2(a) evaluates Information Scoring (IS), Adaptive Resolution (AR), and Path Planning (PP). The fixed-resolution raster baseline obtains 82.1% Top-1 accuracy. Introducing IS improves the result to 83.8%, showing that prioritizing informative regions is already beneficial without changing the scan resolution or traversal policy. Adding AR on top of IS further improves accuracy to 84.3% while reducing latency from 33.1 ms to 27.4 ms.

Path planning provides a larger accuracy gain: combining IS and PP reaches 86.9%, although at a substantially higher computational cost of 45.5 GFLOPs. The full model combines all three components and achieves the best result, **88.5%**, while requiring only **23.0 GFLOPs** and 29.2 ms latency. Compared with the baseline, this corresponds to a 6.4-point Top-1 gain together with a reduction in both GFLOPs (24.8→23.0) and latency (31.7→29.2 ms). This result suggests that the gain of VScan does not arise from simply spending more computation: adaptive resolution compensates for the additional path-planning machinery by allocating computation selectively.

Information cues. Table 2(b) studies the three signals used by IAWS: static structure (S), predictive innovation (I), and residual motion (M). Among individual cues, predictive innovation is the strongest, reaching 86.1%, compared with 85.3% for residual motion and 84.2% for static structure. Pairwise combinations consistently improve over individual signals, with $I + M$ reaching 87.9%. Using all three cues without adaptive weighting achieves 88.1%, indicating that static appearance, motion, and temporal novelty contain complementary information. Finally, content-adaptive weighting further improves the result to **88.5%**. Although the gain over the unweighted three-cue combination is moderate, it consistently favors adjusting the cue mixture according to the current video rather than applying a fixed combination.

Table 1: **Main results on Kinetics-400, Breakfast, and VSPW.** "–" denotes unavailable or non-comparable entries. † denotes masked pretraining and ‡ denotes large-scale pretraining. The best result in each dataset is shown in bold.

Dataset	Method	Venue	Params	Frames	Metric (%) ↑
K400	VideoMamba-M Li et al. (2024a)	ECCV'24	74M	16	81.9
	VideoMamba-M† Li et al. (2024a)	ECCV'24	74M	16	83.4
	VMP-M Lu et al. (2025)	ICCV'25	72M	32	84.0
	VMP-M‡ Lu et al. (2025)	ICCV'25	72M	32	86.5
	VideoMAP-M Liu et al. (2025)	arXiv'25	96M	8	85.5
	VideoMAP-L Liu et al. (2025)	arXiv'25	301M	8	87.3
	VideoMamba++ Song et al. (2025)	PR'25	75M	8	84.5
	MoMa (ViT-L/14)	ICML'25	39M	8	87.8
	Ours	–	43M	8	**88.5**
Breakfast	VideoMamba-S Li et al. (2024a)	ECCV'24	26M	64	97.4
	VideoMamba-M† Li et al. (2024a)	ECCV'24	74M	64	96.9
	VMP-M‡ Lu et al. (2025)	ICCV'25	72M	32	92.4
	VideoMAP-M Liu et al. (2025)	arXiv'25	96M	24	96.7
	VideoMAP-L Liu et al. (2025)	arXiv'25	301M	8	95.2
	VideoMamba++ Song et al. (2025)	PR'25	75M	8	93.2
	MoMa (ViT-B/16)	ICML'25	11M	64	93.2
	TimeSformer-L Bertasius et al. (2021)	ICML'21	121M	96	95.1
	Ours	–	43M	8	**97.6**
VSPW	SegFormer (MiT-B2) Xie et al. (2021)	NeurIPS'21	24.8M	8	43.9
	CFFM (MiT-B5) Sun et al. (2022)	CVPR'22	85.5M	16	49.3
	MPVSS (Swin-S) Weng et al. (2023)	NeurIPS'23	108.0M	16	40.4
	MPVSS (Swin-B) Weng et al. (2023)	NeurIPS'23	147.0M	16	52.6
	MPVSS (Swin-L) Weng et al. (2023)	NeurIPS'23	255.4M	8	53.9
	TV3S (MiT-B2) Hesham et al. (2025)	CVPR'25	28.3M	8	46.3
	TV3S (Swin-S) Hesham et al. (2025)	CVPR'25	76.5M	8	50.6
	RS-SSM (MiT-B5) Zhu et al. (2026)	CVPR'26	88.4M	8	51.6
	RS-SSM (Swin-S) Zhu et al. (2026)	CVPR'26	55.9M	8	52.2
	Ours	–	43M	8	**65.8**

Predictive innovation. Table 2(c) examines how temporal novelty is estimated. A simple raw frame difference reaches 86.8%, while computing the difference in feature space improves the result to 87.5%. Introducing motion-guided feature warping further raises accuracy to 88.4%, indicating that temporal novelty is more informative after accounting for local motion. Finally, compensating global camera motion achieves the best result of **88.5%**. These results support the intended role of predictive innovation: regions should receive high priority when their current content cannot be well explained by temporally aligned observations, rather than merely because they exhibit large raw pixel or motion changes.

Scanning policy. We finally isolate the contribution of IGSPP by comparing different trajectory-construction policies under the same backbone, spatial partition, and scanning budget. As shown in Table 3, the policies exhibit substantially different trade-offs.

Random scanning obtains only 82.7% accuracy and produces both high transition cost and frequent long-range jumps. Zigzag scanning preserves strong spatial continuity, achieving the lowest path cost (19.0) and no long-range jumps, but suffers from substantial redundant traversal ($R_{\text{red}} = 38.7$) and relatively low information gain. In contrast, score-greedy scanning collects more informative regions ($G_{\text{info}} = 9.73$) and reaches 86.2% accuracy, but its locally myopic decisions result in the largest path cost and jump ratio.

IGSPP achieves **88.5%** accuracy together with the highest information gain (10.41) and information efficiency (0.423). At the same time, it reduces redundant visits to **10.8** and limits long-range jumps to 5.6. Although its path cost is not as low as that of Zigzag, it provides a substantially better balance between information acquisition and traversal continuity. This comparison supports the use

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) Design choices for predictive innovation, where FW and CMC denote feature warping and camera-motion compensation, respectively.

(a) Component ablation

Setting	IS	AR	PP	Top-1 (%) ↑	GFLOPs ↓	Latency (ms) ↓
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	Top-1 (%) ↑
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	✓	✓	✓	88.1
S + I + M + AW	✓	✓	✓	**88.5**

(c) Predictive innovation

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

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

Table 3: **Scanning-policy analysis on K400.** All policies use the same backbone, spatial partition, and scanning budget.

Policy	Acc. ↑	G_{info} ↑	C_{path} ↓	R_{red} ↓	R_{jump} ↓	E_{info} ↑
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**

of trajectory-level planning rather than either fixed geometric traversal or independent score-based patch selection.

Summary. Taken together, the ablations provide complementary evidence for the three components of VScan. IAWS improves the identification of useful visual evidence, with predictive innovation emerging as the most informative individual temporal cue. ASR allocates spatial resolution according to video content and helps offset the computational overhead introduced by adaptive traversal. Finally, IGSPP converts patch-level importance into a coherent sequence, avoiding the excessive redundancy of fixed scans and the discontinuous transitions produced by greedy selection. Their combination yields the strongest recognition performance while maintaining a favorable computational profile.

5 Conclusion

We introduce **VScan**, 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, VScan prioritizes informative visual evidence while reducing redundant and incoherent scanning. Extensive experiments and ablations demonstrate the effectiveness and generalizability of VScan across diverse video understanding tasks.

AI USE STATEMENT

In this work, we used generative AI tools for code completion, unit test generation, and language editing. We have not used generative AI tools for formulating research hypotheses or drafting core scientific arguments. All generated code snippets were independently executed, unit-tested, and verified for correctness by the authors. We take full responsibility for the final content and technical artifacts presented in this work.

ETHICS STATEMENT

This work focuses on foundational algorithms and is evaluated entirely on standard, publicly available benchmark datasets. To the best of our knowledge, this work does not involve human participants, proprietary personal data, or direct applications with potential negative societal impacts. All scientific sources and baseline works have been properly cited.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed descriptions of the algorithmic formulation and experimental protocol of **VScan**. The three core components of VScan—information-aware patch scoring, adaptive spatial resolution, and information-guided spatiotemporal path planning. In particular, we specify the computation of static structure, predictive innovation, and residual motion cues, the content-adaptive weighting mechanism, the resolution allocation strategy, and the receding-horizon optimization used to construct the scanning trajectory. Additional derivations and implementation details are provided in the Appendix.

Our experimental setup is described in Section 4. We evaluate VScan on three standard video benchmarks covering complementary tasks: Kinetics-400 for action recognition, Breakfast for fine-grained video understanding, and VSPW for video semantic segmentation. We report the corresponding standard evaluation metrics and use consistent experimental settings for comparisons whenever applicable. Component-wise ablations and additional analyses of information scoring, predictive innovation, adaptive resolution, and scanning policies are provided in Section **??** and the Appendix. The complete source code, training and evaluation configurations, and scripts required to reproduce the reported results will be released upon publication.

REFERENCES

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021. 3

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning*, pages 813–824, 2021. 3, 9

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. In *International Conference on Learning Representations*, 2023. 3

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 7

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021. 3

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6202–6211, 2019. 3

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 1, 2

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022. 1, 2

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*, pages 24211–24221, 2025. 9

Tao Huang, Xiaohuan Pei, Shan You, Fei Wang, Chen Qian, and Chang Xu. Localmamba: Visual state space model with windowed selective scan. In *European Conference on Computer Vision Workshops*, pages 12–22, 2024. doi: 10.1007/978-3-031-91979-4_2. 1, 2

Hilde Kuehne, Ali Arslan, and Thomas Serre. 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. 7

Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *European Conference on Computer Vision*, pages 237–255, 2024a. doi: 10.1007/978-3-031-73347-5_14. 2, 9

Shufan Li, Harkanwar Singh, and Aditya Grover. Mamba-nd: Selective state space modeling for multi-dimensional data. *arXiv preprint arXiv:2402.05892*, 2024b. 1, 2, 3

Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. In *International Conference on Learning Representations*, 2022. 3

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. In *Advances in Neural Information Processing Systems*, volume 37, 2024. 1, 2

Yunze Liu, Peiran Wu, Cheng Liang, Junxiao Shen, Limin Wang, and Li Yi. Videomap: Toward scalable mamba-based video autoregressive pretraining. *arXiv preprint arXiv:2503.12332*, 2025. 2, 9

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022. 3

Hui Lu, Albert A. Salah, and Ronald Poppe. Snakes and ladders: Two steps up for videomamba. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24234–24244, 2025. 2, 9

Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. 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*, pages 4133–4143, 2021. 7

Mandela Patrick, Dylan Campbell, Yuki M. Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and João F. Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. In *Advances in Neural Information Processing Systems*, volume 34, 2021. 3

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. In *Advances in Neural Information Processing Systems*, volume 34, pages 13937–13949, 2021. 3

Michael S. Ryoo, A. J. Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: What can 8 learned tokens do for images and videos? In *Advances in Neural Information Processing Systems*, volume 34, 2021. 3

Xin Song, Wang Tian, Qiqi Zhu, and Xianglong Zhang. Videomamba++: Integrating state space model with dual attention for enhanced video understanding. *Image and Vision Computing*, 161: 105609, 2025. doi: 10.1016/j.imavis.2025.105609. 2, 9

Guolei Sun, Yun Liu, Henghui Ding, Thomas Probst, and Luc Van Gool. Coarse-to-fine feature mining for video semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 9

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

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers and distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357, 2021. 3

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 3

Yuetian Weng, Mingfei Han, Haoyu He, Mingjie Li, Lina Yao, Xiaojun Chang, and Bohan Zhuang. Mask propagation for efficient video semantic segmentation. In *Advances in Neural Information Processing Systems*, volume 36, 2023. 9

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, 2021. 9

Hongxu Yin, Arash Vahdat, Jose M. Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. A-vit: Adaptive tokens for efficient vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10809–10818, 2022. 3

Kai Zhu, Zhenyu Cui, Zehua Zang, and Jiahuan Zhou. 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*, pages 10741–10752, 2026. 9

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *Proceedings of the 41st International Conference on Machine Learning*, pages 62429–62442, 2024. 1, 2