VTDS: A VISION SEMANTIC TIME-DELAY STATE SPACE MIXER FOR EFFICIENT LONG-RANGE VISUAL MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual state space models (SSMs) provide an efficient alternative to self-attention for long-range visual representation learning. However, existing approaches mainly improve state transition dynamics or spatial scan paths, while the content-dependent semantic distance between a token and its relevant context remains implicit. To address this limitation, we propose VTDS, a Visual Semantic Time-Delay State Space Architecture, which explicitly incorporates token-wise semantic delay modeling into visual state propagation. VTDS learns a differentiable dependency distribution over a bounded historical window to retrieve semantically relevant context, and adaptively integrates the retrieved information with the recurrent state through a bounded fuzzy regulator. A loop-free diagonal recurrence further enables parallel state propagation. For a fixed lag window, VTDS scales linearly with sequence length without constructing a global token-wise attention matrix. Experiments on ImageNet-1K classification, COCO object detection and instance segmentation, and ADE20K semantic segmentation demonstrate competitive accuracy-efficiency trade-offs across diverse visual recognition tasks.

1 INTRODUCTION

The most informative visual evidence is not always local. Parts of the same object, repeated structures, and semantically related regions can be separated by large spatial distances, requiring visual models to connect information beyond their neighborhoods. CNNs model local patterns efficiently but capture long-range interactions only indirectly, whereas Vision Transformers (ViTs) enable direct global communication at quadratic cost in the number of visual tokens. State space models (SSMs), exemplified by Mamba, offer an alternative with linear sequence complexity through recurrent state propagation (Gu & Dao, 2023). Recent visual SSMs have demonstrated competitive performance in image classification and dense prediction tasks (Zhu et al., 2024; Liu et al., 2024).

Visual SSMs typically serialize two-dimensional feature maps into one-dimensional sequences, such that the chosen scan order determines the sequential distance and propagation path between visual tokens. Existing approaches improve this serialization through bidirectional or multi-directional scanning (Zhu et al., 2024; Liu et al., 2024), local windows (Huang et al., 2024), sparse traversal (Pei et al., 2025), deformable paths (Liu et al., 2025), or two-dimensional state propagation (Yu et al., 2026). However, these designs primarily modify *how visual tokens are traversed*, while the contribution of specific historical positions to each token remains largely implicit within recurrent state propagation. This creates a mismatch between serialization-induced sequential distance and token-dependent semantic relevance: semantically related regions may be far apart in the serialized sequence, whereas sequentially adjacent tokens may belong to different visual structures (Fig. 1). We therefore ask whether a visual SSM can explicitly learn which historical positions provide useful context for each token.

To address this question, we propose a Vision Semantic Time-Delay State Space Mixer (**VTDS**). Here, "time delay" denotes a content-dependent look-back relation under a given serialization order, rather than physical time. For each token, VTDS learns a differentiable distribution over candidate positions in a bounded causal history, jointly determined by semantic similarity and an ordered-

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

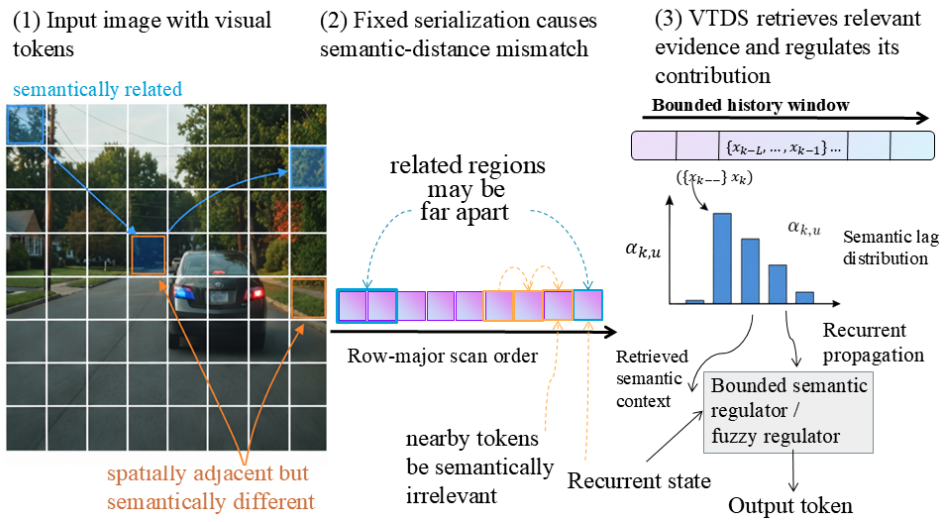


Figure 1: (1) Visual tokens that are semantically related can be spatially separated, whereas adjacent tokens may contain different semantic content. (2) Fixed serialization can therefore distort these relationships in sequence space, placing relevant regions far apart and irrelevant tokens nearby. (3) VTDS mitigates this mismatch by retrieving relevant evidence from a bounded history window and adaptively regulating its contribution while retaining efficient recurrent propagation.

history positional prior, and uses it to retrieve relevant preceding representations. This content-dependent retrieval complements full-sequence recurrent propagation rather than replacing it.

To realize this, within an efficient visual backbone, successive VTDS layers alternate among row-major, reverse row-major, column-major, and reverse column-major serialization orders, evaluating one order per layer to expand spatial coverage without parallel scan branches, and a bounded fuzzy semantic regulator adaptively balances recurrent and retrieved information through token-wise coefficients. A position-independent diagonal state transition enables parallel sequence evaluation. With a fixed history window, VTDS maintains linear complexity in sequence length while avoiding a global token-wise affinity matrix. We evaluate VTDS on different tasks, including ImageNet-1K, COCO, and ADE20K; and controlled ablations examine semantic lag estimation, the ordered-history prior, Gaussian-membership regulation, and alternating serialization.

Our main contributions are threefold: (1) We reveal the *semantic-distance mismatch* between scan-induced sequential distance and token-dependent visual relevance, and introduce VTDS to explicitly model semantic look-back distances within a visual state space architecture. (2) We propose a differentiable semantic lag estimator that retrieves token-specific context from a bounded causal history, complementing full-sequence recurrent propagation with an explicit, interpretable dependency pathway. (3) We develop a bounded Gaussian-membership regulator and a parallelizable diagonal recurrence, forming an efficient state space mixer that is systematically evaluated across image classification, object detection, instance segmentation, and semantic segmentation.

2 Related Work

2.1 Visual State Space Models

State space models (SSMs) have emerged as an efficient alternative to attention-based sequence modeling. Mamba introduces input-dependent selection and hardware-aware computation, enabling content-adaptive sequence modeling with linear complexity in sequence length (Gu & Dao, 2023). Its success has motivated extensive research on adapting selective state space modeling to vision. Vision Mamba (Vim) processes serialized image patches with bidirectional state propagation (Zhu et al., 2024), while VMamba introduces four-directional selective scanning for hierarchical visual modeling (Liu et al., 2024). MambaVision combines redesigned Mamba blocks with later-stage

self-attention to strengthen global interaction (Hatamizadeh & Kautz, 2025). EfficientVMamba integrates atrous selective scanning with convolution (Pei et al., 2025), whereas GroupMamba assigns channel groups to different scan directions (Shaker et al., 2024). These methods mainly improve scan direction, architectural hierarchy, local feature extraction, or computational efficiency. In contrast, VTDS models content-dependent historical dependencies induced by visual serialization, rather than relying solely on recurrent propagation along predefined scan routes.

2.2 Visual Serialization and Spatial Dependency Modeling

Applying a sequential state space operator to images requires arranging two-dimensional features into ordered token sequences. However, sequential distance does not necessarily reflect spatial proximity or semantic relevance: semantically related regions may become widely separated after serialization, while neighboring sequence elements may not be strongly related. LocalMamba performs state propagation within local windows and adopts layer-specific scan configurations (Huang et al., 2024). Multi-scale scanning combines features from different resolutions (Shi et al., 2024). DefMamba makes traversal content adaptive through deformable scanning (Liu et al., 2025), while 2D-CrossScan introduces two-dimensional state propagation with multiple scan paths (Yu et al., 2026). Semantic grouping has also been explored to reorganize visual sequences according to image content, such as in accelerated MRI reconstruction (Meng et al., 2026). These approaches primarily modify *how visual tokens are arranged, grouped, or traversed*. VTDS instead operates under an established serialization order and estimates a soft distribution over candidate positions in a bounded causal history. The resulting context is retrieved from semantically relevant preceding representations without requiring a learned traversal path.

2.3 Content-Adaptive Retrieval and Feature Regulation

Content-adaptive mechanisms have been widely explored to control computation and information flow according to input semantics. DynamicViT progressively removes less informative tokens based on learned token importance (Rao et al., 2021). Mamba-Adaptor augments hidden states with features sampled from learned locations to mitigate long-range forgetting (Xie et al., 2025). Selective Visual Prompting introduces token-wise prompts that modulate update and forget behaviors in pretrained Vision Mamba models (Yao et al., 2025). VTDS differs from token pruning by preserving the complete visual sequence. Unlike methods based on shared learnable memory locations, its historical context is conditioned on each current token and retrieved directly from preceding representations. Specifically, VTDS estimates a differentiable semantic lag distribution over a bounded causal window, allowing different tokens to retrieve information from different historical distances. The reliability of the retrieved context may vary across object interiors, boundaries, occlusions, and cluttered backgrounds. VTDS therefore couples semantic retrieval with a bounded fuzzy semantic regulator. Rather than serving as an independent classifier or reasoning pipeline, the regulator uses learnable memberships to produce continuous token-wise coefficients that balance explicit historical retrieval and recurrent state propagation. This design jointly determines *where* information is retrieved and *how strongly* it influences the current representation.

3 Method

3.1 Overview

We propose the **Visual Semantic Time-Delay State Space Architecture (VTDS)**, which augments recurrent visual state propagation with explicit, content-dependent retrieval from historical token positions. Unlike conventional selective state space models that mainly determine how the current input updates an evolving latent state, VTDS additionally estimates where semantically relevant evidence should be retrieved. Here, *semantic delay* denotes a token-dependent look-back distance under a specified visual serialization order rather than physical time. Given an input image $\mathbf{I} \in \mathbb{R}^{B\times 3\times H_0\times W_0}$, a convolutional patch stem produces a sequence of visual tokens:

$$\mathbf{X}^0 = \mathrm{Dropout}\left[\mathrm{LN}\left(\mathcal{P}(\mathbf{I}) + \mathbf{E}_{\mathrm{pos}}\right)\right], \quad \mathbf{X}^0 \in \mathbb{R}^{B\times T\times D}, \tag{1}$$

Here, T is the number of tokens, and D is the embedding dimension. The patch stem $\mathcal{P}(\cdot)$ returns the token sequence. The learnable positional embedding $\mathbf{E}_{\mathrm{pos}} \in \mathbb{R}^{T\times D}$ is associated with the two-dimensional grid and broadcast across the batch. In the ImageNet-1K instantiation, four stride-2

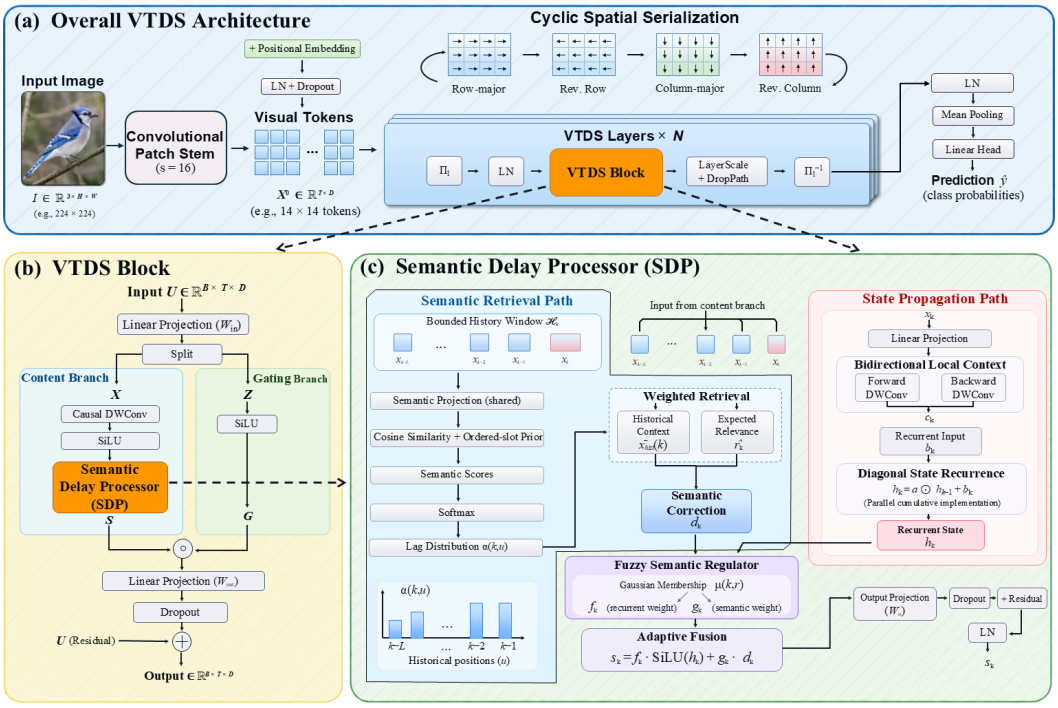


Figure 2: Overview of VTDS. Each layer processes visual tokens under one spatial serialization order. Inside the Semantic Delay Processor, semantic lag estimation retrieves token-specific historical context, a diagonal recurrence propagates full-sequence states, and fuzzy semantic regulation adaptively combines the recurrent and retrieval pathways.

convolutional stages produce a total stride of 16, mapping a 224×224 image to 14×14 tokens. To introduce complementary spatial ordering priors, VTDS cyclically alternates row-major, reverse row-major, column-major, and reverse column-major serialization across layers. Let $\mathbf{\Pi}_l$ denote the permutation used in layer l. A stabilized VTDS layer is formulated as

$$\begin{aligned} &\bar{\mathbf{X}}^l = \mathbf{\Pi}_l \mathbf{X}^l, \mathbf{U}^l = \mathrm{LN}(\bar{\mathbf{X}}^l), \\ &\mathbf{V}^l = \bar{\mathbf{X}}^l + \mathrm{DropPath}\left[\boldsymbol{\eta}_l \odot \left(\mathrm{VTDSBlock}(\mathbf{U}^l) - \mathbf{U}^l\right)\right], \mathbf{X}^{l+1} = \mathbf{\Pi}_l^{-1} \mathbf{V}^l. \end{aligned} \tag{2}$$

where $\boldsymbol{\eta}_l \in \mathbb{R}^D$ is a learnable LayerScale vector. Both $\bar{\mathbf{X}}^l$ and $\mathbf{V}^l$ follow layer l's serialization order, while $\mathbf{X}^{l+1}$ is restored to canonical spatial order. Subtracting $\mathbf{U}^l$ cancels the block's internal identity shortcut, isolating its residual update. Each layer evaluates only one serialization order, so the four-order design requires no four-way parallel scanning.

3.2 VTDS Block

Given a normalized sequence $\mathbf{U} \in \mathbb{R}^{B \times T \times D}$, the VTDS block expands the channel dimension and forms content and gating branches through $[\mathbf{X}, \mathbf{Z}] = \mathrm{Split}(\mathbf{U}\mathbf{W}_{\mathrm{in}})$. Here, $\mathbf{X}, \mathbf{Z} \in \mathbb{R}^{B \times T \times D_e}$, D_e is the expanded hidden dimension, and $\mathbf{W}_{\mathrm{in}} \in \mathbb{R}^{D \times 2D_e}$. The content branch applies a causal depthwise convolution followed by SiLU, yielding $\tilde{\mathbf{X}} = \mathrm{SiLU}[\mathrm{DWConv_c}(\mathbf{X})]$. The Semantic Delay Processor (SDP) then produces $\mathbf{S} = \mathrm{SDP}(\tilde{\mathbf{X}})$, while the gating branch generates the input-dependent gate $\mathbf{G} = \mathrm{SiLU}(\mathbf{Z})$. The block output is

$$\mathrm{VTDSBlock}(\mathbf{U}) = \mathbf{U} + \mathrm{Dropout}\left[(\mathbf{S} \odot \mathbf{G})\mathbf{W}_{\mathrm{out}}\right]. \tag{3}$$

Here, $\mathbf{W}_{\mathrm{out}} \in \mathbb{R}^{D_e \times D}$. The feature gate $\mathbf{G}$ modulates the SDP output channel-wise and is not constrained to $[0, 1]$.

3.3 Differentiable Semantic Lag Estimation

Let $\mathbf{x}_k \in \mathbb{R}^{D_e}$ denote the SDP input at serialized position k. VTDS constructs a bounded causal history window of length L, $\mathcal{H}_k = [\mathbf{x}_{k-L}, \mathbf{x}_{k-L+1}, \ldots, \mathbf{x}_{k-1}]$, with unavailable positions at the beginning of the sequence zero-padded. Candidates are stored from oldest to most recent. Slot $u \in \{1, \ldots, L\}$ contains $\mathbf{x}^{\text{hist}}_{k,u} = \mathbf{x}_{k-L+u-1}$, at serialized look-back distance $\delta_u = L - u + 1$. The current token and historical candidates share a semantic projection of dimension D_s. The current token is mapped to $\mathbf{q}_k = \mathbf{W}_s\mathbf{x}_k + \mathbf{b}_s$, and each historical candidate to $\mathbf{q}^{\text{hist}}_{k,u} = \mathbf{W}_s\mathbf{x}^{\text{hist}}_{k,u} + \mathbf{b}_s$. Here, $\mathbf{W}_s \in \mathbb{R}^{D_s \times D_e}$ and $\mathbf{b}_s \in \mathbb{R}^{D_s}$. Their semantic relevance is measured using cosine similarity:

$$r_{k,u} = \frac{\mathbf{q}_k^\top \mathbf{q}^{\text{hist}}_{k,u}}{\|\mathbf{q}_k\|_2 \|\mathbf{q}^{\text{hist}}_{k,u}\|_2 + \epsilon}. \tag{4}$$

The current implementation applies a fixed prior $\omega_u = \exp(-\lambda u^\rho)$ over ordered history slots and computes the weighted score $e_{k,u} = r_{k,u}\omega_u$, where λ and ρ control the prior shape. Since the slots are stored from oldest to newest, ω_u is an ordered-slot positional prior rather than a monotonic penalty with respect to the physical lag δ_u. A token-specific semantic lag distribution is obtained by

$$\alpha_{k,u} = \frac{\exp(\kappa e_{k,u})}{\sum_{v=1}^{L} \exp(\kappa e_{k,v})}, \tag{5}$$

where κ controls its sharpness. Instead of selecting a single historical position, VTDS retrieves a weighted context and computes its expected semantic relevance:

$$\bar{\mathbf{x}}^{\text{hist}}_k = \sum_{u=1}^{L} \alpha_{k,u}\mathbf{x}^{\text{hist}}_{k,u}, \qquad \hat{r}_k = \sum_{u=1}^{L} \alpha_{k,u} e_{k,u}. \tag{6}$$

To analyze the learned distribution, we additionally characterize it by its expected look-back distance and normalized entropy:

$$\bar{\delta}_k = \sum_{u=1}^{L} \alpha_{k,u}\delta_u, \qquad \mathcal{E}_k = -\frac{1}{\log L} \sum_{u=1}^{L} \alpha_{k,u} \log(\alpha_{k,u} + \epsilon), \tag{7}$$

where δ_u is the serialized look-back distance defined above, and $\mathcal{E}_k \in [0, 1]$: lower entropy indicates more concentrated retrieval, while higher entropy indicates more distributed aggregation. We use these quantities in Section 4.5 to analyze how retrieval behavior changes with representation depth and image content.

3.4 Delay-Regulated State Propagation

Alongside historical retrieval, the SDP constructs a recurrent representation over the serialized sequence. It first projects the input and extracts bidirectional local context:

$$\begin{aligned} \mathbf{p}_k &= \mathbf{W}_p\mathbf{x}_k + \mathbf{b}_p, \\ \mathbf{c}_k &= \text{GELU}(\text{DWConv}_{\rightarrow}(\mathbf{P}))_k + \text{GELU}(\text{DWConv}_{\leftarrow}(\mathbf{P}))_k, \end{aligned} \tag{8}$$

Here, $\mathbf{P} = [\mathbf{p}_0, \ldots, \mathbf{p}_{T-1}]$, $\mathbf{W}_p \in \mathbb{R}^{D_e \times D_e}$, and $\mathbf{b}_p \in \mathbb{R}^{D_e}$. Both depthwise operators are causal in their respective input order; $\text{DWConv}_{\leftarrow}$ is applied to the reversed token sequence, and its output is restored to the original order before indexing position k. The resulting local context therefore contains information from both directions. Consequently, the complete SDP is not causal, although its retrieval window contains only preceding positions in the current layer's input sequence. The recurrent input is then computed as $\mathbf{b}_k = \mathbf{B}(\mathbf{x}_k + \mathbf{c}_k)$, where $\mathbf{B} \in \mathbb{R}^{D_e \times D_e}$ is a learnable projection matrix. VTDS employs a diagonal transition vector $\mathbf{a} \in (0, 1)^{D_e}$:

$$\mathbf{h}_k = \mathbf{a} \odot \mathbf{h}_{k-1} + \mathbf{b}_k, \qquad \mathbf{h}_{-1} = \mathbf{0}. \tag{9}$$

Because $\mathbf{a}$ is diagonal and position-independent within each layer, expansion from $\mathbf{h}_{-1} = \mathbf{0}$ gives the exact-arithmetic identity

$$\mathbf{h}_k = \mathbf{a}^{\circ k} \odot \sum_{i=0}^{k} \mathbf{a}^{\circ(-i)} \odot \mathbf{b}_i. \tag{10}$$

Consequently, the complete state sequence is evaluated through parallel cumulative summation:

$$\mathbf{H} = \mathbf{P}^a \odot \mathrm{CumSum}_T \left[\mathbf{B}_{\mathrm{seq}} \odot (\mathbf{P}^a)^{\circ(-1)} \right], \tag{11}$$

where $\mathbf{P}^a_k = \mathbf{a}^{\circ k}$ and $\mathbf{B}_{\mathrm{seq}} = [\mathbf{b}_0, \ldots, \mathbf{b}_{T-1}]$. This cumulative formulation removes explicit token-wise loops. Transition powers are evaluated in the logarithmic domain for numerical stability. The retrieved historical context forms a parallel semantic correction:

$$\mathbf{d}_k = \gamma \tanh\left(\hat{r}_k \mathbf{C} \bar{\mathbf{x}}^{\mathrm{hist}}_k\right), \qquad \gamma = 0.1 \tanh(\gamma_{\mathrm{raw}}), \tag{12}$$

where $\mathbf{C} \in \mathbb{R}^{D_e \times D_e}$ is a learnable projection matrix. The bounded scale prevents the retrieval pathway from overwhelming the recurrent representation. The correction is fused with the recurrent response rather than inserted into the state-transition equation. Because the reliability of retrieved context varies across tokens, VTDS applies a learnable fuzzy semantic regulator to $\mathbf{z}_k = \mathbf{d}_k/2$. For fuzzy rule center $\mathbf{c}_r$, the normalized Gaussian membership is

$$\mu_{k,r} = \frac{\exp\left[-\frac{1}{2}\left(\frac{\|\mathbf{z}_k - \mathbf{c}_r\|_2}{\sigma}\right)^2\right]}{\sum_{j=1}^{R} \exp\left[-\frac{1}{2}\left(\frac{\|\mathbf{z}_k - \mathbf{c}_j\|_2}{\sigma}\right)^2\right] + \epsilon}. \tag{13}$$

Two bounded token-wise coefficients are generated:

$$f_k = \tanh\left(\sum_{r=1}^{R} \mu_{k,r} w^f_r\right), \qquad g_k = \mathrm{sigmoid}\left(\sum_{r=1}^{R} \mu_{k,r} w^g_r\right). \tag{14}$$

Here, $f_k \in [-1, 1]$ modulates the recurrent response and $g_k \in (0, 1)$ regulates the semantic correction. Both coefficients are token-wise scalars broadcast across channels. The final SDP representation is:

$$\mathbf{s}_k = \mathrm{LN}\left[\mathbf{W}_o \,\mathrm{Dropout}\big(f_k \,\mathrm{SiLU}(\mathbf{h}_k) + g_k \mathbf{d}_k\big) + \mathbf{x}_k\right], \tag{15}$$

where $\mathbf{W}_o \in \mathbb{R}^{D_e \times D_e}$ is an output projection. The resulting sequence $\mathbf{S} = [\mathbf{s}_0, \ldots, \mathbf{s}_{T-1}]$ is returned to the outer VTDS block and modulated by the outer feature gate $\mathbf{G}$.

3.5 Complexity

Let T, D_e, L, R, and K denote the sequence length, expanded dimension, history-window length, fuzzy-rule count, and depthwise convolution kernel size, respectively. The current implementation explicitly materializes $T \times L$ historical candidates. Its dominant computational complexity is $\mathcal{O}\left(TD_e^2 + TLD_e + TRD_e + TKD_e\right)$ where the first term accounts for the input/output projections and the semantic correction projection $\mathbf{C}$, the second for the pairwise similarity computation and weighted sum over the history window, the third for the fuzzy regulation, and the fourth for the depthwise convolutions. The temporary memory required by the history windows is $\mathcal{O}(TLD_e)$. For fixed D_e, L, R, and K, both costs scale linearly with T, and VTDS does not construct a global $T \times T$ affinity matrix.

4 Experiments

4.1 Datasets and evaluation

We evaluate classification on ImageNet-1K (Deng et al., 2009) using Top-1 and Top-5 accuracy. For COCO 2017 (Lin et al., 2014), we integrate the ImageNet-pretrained VTDS into Mask R-CNN (He et al., 2017) with a feature pyramid network and report box AP (AP^b) and mask AP (AP^m) under the standard $1\times$ and $3\times$ schedules. For ADE20K (Zhou et al., 2017), we use UPerNet (Xiao et al., 2018) with 160K training iterations and 512×512 crops, reporting single-scale and multi-scale mean intersection-over-union ($\mathrm{mIoU}_{\mathrm{SS}}$ and $\mathrm{mIoU}_{\mathrm{MS}}$). Downstream learning rates, batch sizes, feature-extraction interfaces, and augmentations follow the corresponding implementations and remain fixed across controlled comparisons.

Table 1: ImageNet-1K classification at 224×224. The best and second-best results under comparable complexity are highlighted in bold and underlined, respectively.

Method	Params (M)	FLOPs (G)	Top-1 (%)	Top-5 (%)
ResNet-50 (He et al., 2016)	25.6	4.1	76.1	92.9
ConvNeXt-T (Liu et al., 2022)	29	4.5	82.1	95.9
DeiT-S (Touvron et al., 2021)	22	4.6	79.9	95.0
Swin-T (Liu et al., 2021b)	28	4.5	81.2	95.5
PVTv2-B2 (Wang et al., 2022)	25.4	4.0	82.0	–
Vim-S (Zhu et al., 2024)	26	5.1	80.5	95.1
VMamba-T (Liu et al., 2024)	30	4.9	82.6	–
SwinMLP-B (Liu et al., 2021b)	61	10.4	81.3	95.3
gMLP-B (Liu et al., 2021a)	73	15.8	81.6	–
PiT-B (Heo et al., 2021)	73.8	12.5	82.0	–
VRWKV-B (Duan et al., 2025)	93.7	18.2	82.0	–
VTDS (Ours)	70.7	70.0	82.5	96

4.2 Experimental configuration

Model configuration VTDS adopts a single-scale, non-hierarchical backbone. A convolutional stem with total stride 16 maps a 224×224 input to a $14 \times 14 = 196$ grid of visual tokens. The backbone stacks 18 TDS blocks with embedding dimension $D = 384$; the content branch expands to $D_e = 768$ (expand factor 2). The semantic lag estimator maintains a bounded causal history of $L = 10$ positions and computes the semantic similarity in a low-dimensional projection space of dimension $D_s = 64$. The fuzzy regulator uses $R = 6$ Gaussian rules. Depthwise convolutions use kernel size $K = 4$. Successive layers alternate row-major, reverse row-major, column-major, and reverse column-major serialization, evaluating one order per layer. The backbone uses learnable 2D position embeddings, Pre-Norm with LayerScale (initialized to 10^{-4}), and stochastic depth increasing linearly to 0.2. Classification is performed by LayerNorm, mean pooling, and a linear head. The state transition is a position-independent diagonal matrix, which together with the bounded history window keeps the per-layer cost linear in T and avoids a global $T \times T$ affinity matrix (Section 3.5).

Training settings. We train VTDS from scratch on ImageNet-1K for 300 epochs at 224×224 using AdamW ($\beta = (0.9, 0.95)$, $\epsilon = 10^{-8}$) with an effective global batch size of 768, weight decay 0.05, and gradient-norm clipping at 1.0. The initial learning rate is 7.5×10^{-4}, with a 20-epoch linear warm-up and cosine decay to 10^{-6}. Data augmentation comprises RandAugment (2 operations, magnitude 9), MixUp ($\alpha = 0.8$, applied with probability 1.0), CutMix ($\alpha = 1.0$, switch probability 0.5), label smoothing (0.1), and random erasing ($p = 0.25$). An exponential moving average (EMA) with decay 0.9999 is maintained, and the checkpoint with the highest EMA validation accuracy is kept. Validation is performed with 256×256 resizing followed by a 224×224 center crop.

4.3 Main Result

Table 1 compares VTDS with representative CNN, Transformer, and visual SSM backbones on ImageNet-1K. VTDS reaches 82.5% Top-1 accuracy with 70.7M parameters. It improves over the visual SSM baseline Vim-S by 2.0 points and over the Transformer baseline DeiT-S by 2.6 points, and is within 0.2 points of the stronger T-tier SSMs VMamba-T (82.6%) and LocalVMamba-T (82.7%). We report these results at a larger parameter budget than the T-tier baselines; the accuracy–efficiency trade-off is analyzed in Table 4. For COCO, we transfer the ImageNet-pretrained backbone to Mask R-CNN and evaluate detection and instance segmentation under the $1\times$ and $3\times$ schedules (Table 2); all compared methods use the same framework and protocol. VTDS improves box AP by 1.3 points over the strongest $1\times$ SSM baseline (Spatial-Mamba-T) and by 2.2 (box) / 1.7 (mask) points over DefMamba-S at $3\times$, with the gap widening under longer training. These results come at a larger budget (70.7M params, 500G FLOPs); the trade-off is analyzed in Appendix A.

4.4 Ablation Studies

Figure 4 summarizes the effects of the key design choices in VTDS. Starting from the recurrent baseline, progressively introducing SLE, OHP, FSR, and AS consistently improves ImageNet-1K

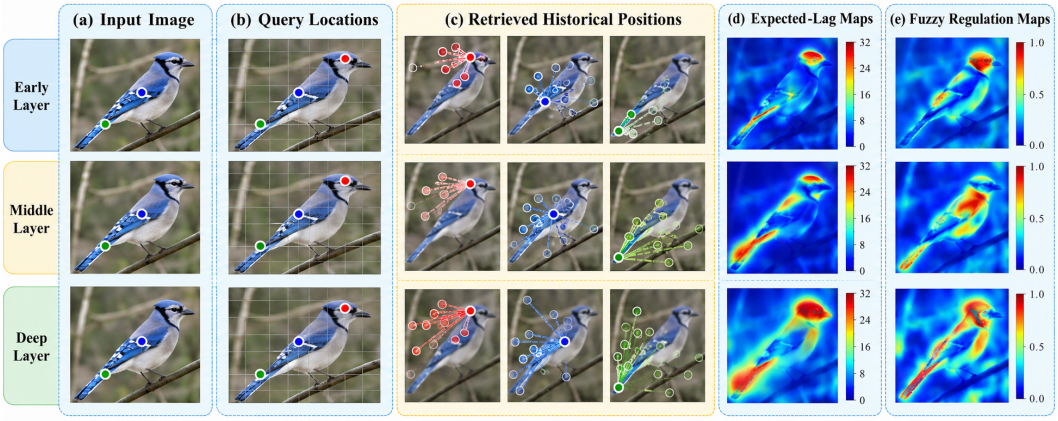


Figure 3: Visualization of semantic-delay modeling. From left to right: input image, query locations, retrieved historical positions, expected-lag maps, and fuzzy regulation maps. Results from early, middle, and deep layers are included.

Table 2: COCO 2017 object detection and instance segmentation with Mask R-CNN. FLOPs are measured at 800×1280.

Backbone	Params (M)	FLOPs (G)	1×		3×	
			AP^b	AP^m	AP^b	AP^m
ConvNeXt-S (Liu et al., 2022)	70	348	45.4	41.8	47.9	42.9
Swin-S (Liu et al., 2021b)	69	354	44.8	40.9	48.2	43.2
InternImage-S (Wang et al., 2023)	69	340	47.8	43.3	49.7	44.5
InternImage-B (Wang et al., 2023)	115	501	48.8	44.0	50.3	44.8
PVTv2-B3 (Wang et al., 2022)	63	397	47.0	42.5	48.4	43.2
VMamba-S (Liu et al., 2024)	70	384	48.7	43.7	49.9	44.2
LocalVMamba-S (Huang et al., 2024)	69	414	48.4	43.2	49.9	44.1
Spatial-Mamba-S (Xiao et al., 2025)	63	315	49.2	44.0	50.5	44.6
VSSD-S (Shi et al., 2025)	59	325	48.4	43.5	50.0	44.6
VTDS (Ours)	70.7	500	48.9	43	52	46.2

accuracy, increasing Top-1 performance from 78.9% to 82.5%. Among these components, SLE provides the largest individual improvement, highlighting the importance of content-dependent historical retrieval. The fusion comparison further shows that the fuzzy regulator achieves higher accuracy than fixed summation, sigmoid gating, and MLP-based gating. Increasing the history length also improves performance up to $L = 10$, after which the gain saturates, suggesting that a moderate history window is sufficient to capture useful long-range dependencies without introducing redundant context. Detailed quantitative results are provided in Appendix Table 5.

4.5 Mechanism Analysis

Figure 3 visualizes the learned lag distributions at three depths on a representative image; the expected-lag maps show the expected look-back distance $\bar{\delta}_k$ defined in Section 3 (Eq. 7) at each position. Retrieved historical positions spread over a larger spatial extent in deeper layers, and the maps shift toward larger values, indicating that early layers retrieve mostly nearby context while deeper layers aggregate more distant tokens. The fuzzy regulation maps also vary across layers, showing that the balance between recurrent and retrieved information is applied token-wise. The differences across query positions indicate that lag selection is conditioned on image content, and its progression across layers reflects representation depth.

Table 3: ADE20K semantic segmentation with UPerNet trained for 160K iterations.

Backbone	Params (M)	mIoU_{SS}	mIoU_{MS}
ResNet-50 (He et al., 2016)	67	42.1	42.8
ConvNeXt-S (Liu et al., 2022)	82	48.7	49.6
InternImage-T (Wang et al., 2023)	59	47.9	48.1
Swin-S (Liu et al., 2021b)	81	47.6	49.5
MambaOut-Tiny (Yu & Wang, 2025)	54	47.4	48.6
VMamba-T (Liu et al., 2024)	62	47.9	48.8
LocalVMamba-T (Huang et al., 2024)	57	47.9	49.1
Spatial-Mamba-T (Xiao et al., 2025)	57	48.6	49.4
EfficientVMamba-B (Pei et al., 2025)	65	46.5	47.3
VTDS (Ours)	70.7	48.6	49.5

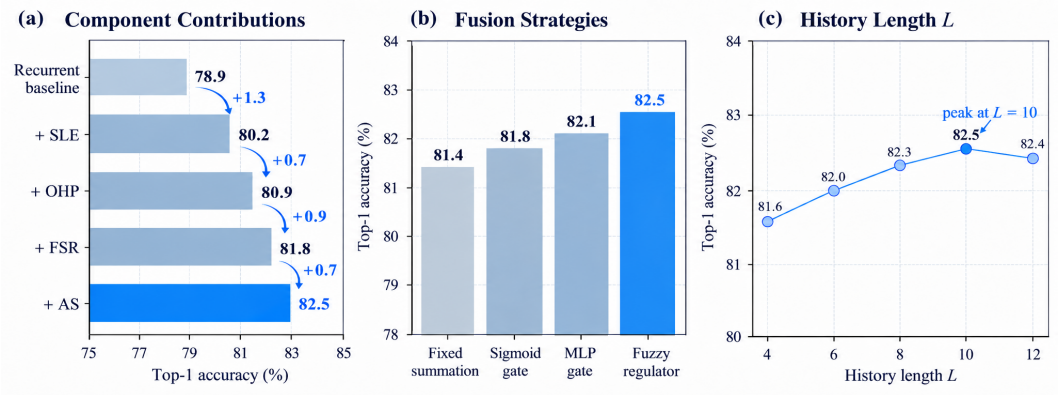


Figure 4: Ablation studies of the key components and design choices in VTDS. **(a)** Cumulative contribution of individual components. **(b)** Comparison of different fusion strategies. **(c)** Influence of the history length L on ImageNet-1K Top-1 accuracy. Full quantitative results are provided in Appendix Table 5.

5 Conclusion

We proposed VTDS to address the semantic-distance mismatch between predefined visual serialization orders and content-dependent dependencies among image regions. VTDS complements recurrent state propagation with differentiable semantic lag estimation, bounded fuzzy semantic regulation, and a parallel diagonal recurrence, enabling explicit retrieval of relevant historical context without constructing a global token-to-token affinity matrix. Experiments on ImageNet-1K, COCO, and ADE20K, together with controlled ablations and mechanism analyses, demonstrate the effectiveness and transferability of the proposed design. Nevertheless, VTDS currently relies on a fixed-length causal history window and explicitly materialized historical candidates, which introduce additional computation and memory overhead. Future work will explore adaptive history ranges, more efficient projected-history reuse, and multiscale semantic retrieval.

AI use statement

In this work, we used generative AI tools to assist with language translation and English-language editing. We did not use generative AI tools to generate experimental results, synthetic training or evaluation data, or to formulate or prove mathematical claims; other required-disclosure tasks are not applicable to this work.

REFERENCES

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Junxian Duan, Yuang Ai, Jipeng Liu, Shenyuan Huang, Huaibo Huang, Jie Cao, and Ran He. Test-time forgery detection with spatial-frequency prompt learning. *International Journal of Computer Vision*, 133(2):672–687, 2025.

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

Ali Hatamizadeh and Jan Kautz. Mambavision: A hybrid mamba-transformer vision backbone. In *2025 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 25261–25270. IEEE, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11916–11925. IEEE, 2021.

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*, pp. 12–22. Springer, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Hanxiao Liu, Zihang Dai, David So, and Quoc V Le. Pay attention to mlps. *Advances in neural information processing systems*, 34:9204–9215, 2021a.

Leiye Liu, Miao Zhang, Jihao Yin, Tingwei Liu, Wei Ji, Yongri Piao, and Huchuan Lu. Defmamba: Deformable visual state space model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8838–8847. IEEE, 2025.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9992–10002. Ieee, 2021b.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11966–11976. IEEE, 2022.

Yucong Meng, Zhiwei Yang, Kexue Fu, Zhijian Song, and Yonghong Shi. Image content matters: An image content aware state space model for accelerated mri reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8025–8033, 2026.

Xiaohuan Pei, Tao Huang, and Chang Xu. Efficientvmamba: Atrous selective scan for light weight visual mamba. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 6443–6451, 2025.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021.

Abdelrahman Shaker, Syed Talal Wasim, Salman Khan, Juergen Gall, and Fahad Shahbaz Khan. Groupmamba: Parameter-efficient and accurate group visual state space model. *arXiv preprint arXiv:2407.13772*, 2, 2024.

Yuheng Shi, Minjing Dong, and Chang Xu. Multi-scale vmamba: Hierarchy in hierarchy visual state space model. *Advances in Neural Information Processing Systems*, 37:25687–25708, 2024.

Yuheng Shi, Mingjia Li, Minjing Dong, and Chang Xu. Vssd: Vision mamba with non-causal state space duality. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10819–10829. IEEE, 2025.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational visual media*, 8(3):415–424, 2022.

Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14408–14419. IEEE, 2023.

Chaodong Xiao, Minghan Li, Deyu Meng, Lei Zhang, et al. Spatial-mamba: Effective visual state space models via structure-aware state fusion. In *International Conference on Learning Representations*, volume 2025, pp. 73777–73795, 2025.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *European conference on computer vision*, pp. 432–448. Springer, 2018.

Fei Xie, Jiahao Nie, Yujin Tang, Wenkang Zhang, and Hongshen Zhao. Mamba-adaptor: State space model adaptor for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20124–20134, 2025.

Yifeng Yao, Zichen Liu, Zhenyu Cui, Yuxin Peng, and Jiahuan Zhou. Selective visual prompting in vision mamba. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22083–22091, 2025.

Longlong Yu, Wenxi Li, Yaoqi Sun, Hang Xu, Chenggang Yan, and Yuchen Guo. 2d-crossscan mamba: Enhancing state space models with spatially consistent multi-path 2d information propagation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 17957–17965, 2026.

Weihao Yu and Xinchao Wang. Mambaout: Do we really need mamba for vision? In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4484–4496. IEEE, 2025.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5122–5130. IEEE, 2017.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.

Algorithm 1 Forward Procedure of the VTDS Block

Require: Input $\mathbf{U} \in \mathbb{R}^{B\times T\times D}$, history length L, model parameters Θ.
Ensure: Output $\mathbf{Y} \in \mathbb{R}^{B\times T\times D}$.

$[\mathbf{X}, \mathbf{Z}] \leftarrow \mathrm{Split}(\mathbf{U}\mathbf{W}_{\mathrm{in}})$
$\mathbf{X} \leftarrow \mathrm{SiLU}(\mathrm{DWConv}_{\mathrm{c}}(\mathbf{X}))$
$\mathbf{G} \leftarrow \mathrm{SiLU}(\mathbf{Z})$
$\mathbf{C}_{\mathrm{loc}} \leftarrow \mathrm{BiDWConv}(\mathrm{Linear}_p(\mathbf{X}))$
$\mathcal{H} \leftarrow \mathrm{HistoryWindow}_L(\mathbf{X})$
$(\boldsymbol{\alpha}, \bar{\mathbf{X}}^{\mathrm{hist}}, \hat{\mathbf{r}}) \leftarrow \mathrm{SLE}(\mathbf{X}, \mathcal{H})$
$\mathbf{B}_{\mathrm{seq}} \leftarrow \mathbf{B}(\mathbf{X} + \mathbf{C}_{\mathrm{loc}})$
$\mathbf{a} \leftarrow \mathrm{clip}(\mathbf{A}_{\mathrm{diag}}, 0.5, 0.99)$
$\mathbf{H} \leftarrow \mathrm{ParallelScan}(\mathbf{B}_{\mathrm{seq}}, \mathbf{a})$
$\gamma \leftarrow 0.1 \tanh(\gamma_{\mathrm{raw}})$
$\mathbf{d}_k \leftarrow \gamma \tanh(\hat{r}_k \mathbf{C}\bar{\mathbf{x}}_k^{\mathrm{hist}})$
$(f_k, g_k) \leftarrow \mathrm{FSR}(\mathbf{d}_k/2)$
$\mathbf{v}_k \leftarrow f_k \,\mathrm{SiLU}(\mathbf{h}_k) + g_k \mathbf{d}_k$
$\mathbf{V} \leftarrow [\mathbf{v}_0, \ldots, \mathbf{v}_{T-1}]$
$\mathbf{S} \leftarrow \mathrm{LN}\left(\mathrm{Linear}_o(\mathrm{Dropout}(\mathbf{V})) + \mathbf{X}\right)$
$\mathbf{Y} \leftarrow \mathbf{U} + \mathrm{Dropout}\left((\mathbf{S} \odot \mathbf{G})\mathbf{W}_{\mathrm{out}}\right)$
return $\mathbf{Y}$

Algorithm 2 VTDS for Image Classification

Require: Image $\mathbf{I} \in \mathbb{R}^{B\times 3\times 224\times 224}$; number of layers N; model parameters Θ.
Ensure: Class logits $\mathbf{Y} \in \mathbb{R}^{B\times C}$.

$\mathbf{X}^0 \leftarrow \mathrm{Dropout}(\mathrm{LN}(\mathcal{P}(\mathbf{I}) + \mathbf{E}_{\mathrm{pos}}))$
for $l = 0, \ldots, N-1$ **do**
  Select $\boldsymbol{\Pi}_l$ from the four alternating serialization orders.
  $\bar{\mathbf{X}}^l \leftarrow \boldsymbol{\Pi}_l \mathbf{X}^l$
  $\mathbf{U}^l \leftarrow \mathrm{LN}(\bar{\mathbf{X}}^l)$
  $\boldsymbol{\Delta}^l \leftarrow \mathrm{VTDSBlock}_l(\mathbf{U}^l) - \mathbf{U}^l$
  $\mathbf{X}^{l+1} \leftarrow \boldsymbol{\Pi}_l^{-1}\left[\bar{\mathbf{X}}^l + \mathrm{DropPath}(\boldsymbol{\eta}_l \odot \boldsymbol{\Delta}^l)\right]$
end for
$\mathbf{z} \leftarrow \mathrm{MeanPool}(\mathrm{LN}(\mathbf{X}^N))$
$\mathbf{Y} \leftarrow \mathrm{Classifier}(\mathrm{Dropout}(\mathbf{z}))$
return $\mathbf{Y}$

A Appendix

A.1 Algorithmic Description of VTDS

Algorithm 1 summarizes the forward procedure of the proposed VTDS block. It integrates semantic lag estimation, parallel state propagation, and fuzzy semantic regulation into a unified representation learning process.

Implementation notes. The SLE operator implements the semantic lag estimation defined in Eqs. 4–6. The ParallelScan operator evaluates the diagonal recurrence using Eq. 11, while FSR generates the bounded coefficients defined in Eqs. 13–14. The BiDWConv operator follows Eq. 8. All token-indexed operations are evaluated in parallel.

A.2 VTDS for Image Classification

A.3 VTDS for Semantic Segmentation

Practical efficiency. Although VTDS has linear complexity in T for a fixed history window, practical performance also depends on history-window materialization and projection overhead. We therefore compare theoretical and measured efficiency under an identical environment. Compared with the recurrent-only variant, semantic retrieval introduces **34.85G** FLOPs and **5.3GB** peak memory while changing Top-1 accuracy by **3.6** points. VTDS reaches **410.0** images/s and **0.62** ms per im-

Algorithm 3 VTDS for Semantic Segmentation

Require: Image $\mathbf{I} \in \mathbb{R}^{B\times 3\times H_0\times W_0}$; pretrained VTDS backbone; feature indices $\mathcal{J}$; segmentation adapter $\mathcal{A}_{\text{seg}}$.
Ensure: Segmentation logits $\mathbf{Y}$.
$\mathbf{F}^0 \leftarrow \mathcal{P}(\mathbf{I})$
Obtain token grid size (h, w) from $\mathbf{F}^0$.
Resize positional embeddings to (h, w).
$\mathbf{X}^0 \leftarrow \text{Dropout}(\text{LN}(\text{Flatten}(\mathbf{F}^0) + \mathbf{E}^{h,w}_{\text{pos}}))$
$\mathcal{F} \leftarrow \emptyset$
for $l = 0, \ldots, N-1$ **do**
    $\mathbf{X}^{l+1} \leftarrow \text{VTDSLayer}_l(\mathbf{X}^l)$
    if $l + 1 \in \mathcal{J}$ **then**
        $\mathcal{F} \leftarrow \mathcal{F} \cup \{\text{Reshape}_{h,w}(\mathbf{X}^{l+1})\}$
    end if
end for
$\mathcal{F}_{\text{pyr}} \leftarrow \mathcal{A}_{\text{seg}}(\mathcal{F})$
$\mathbf{Z} \leftarrow \text{UPerNet}(\mathcal{F}_{\text{pyr}})$
$\mathbf{Y} \leftarrow \text{Interpolate}(\mathbf{Z}, H_0, W_0)$
return $\mathbf{Y}$

age. Together with the task results, these measurements characterize the actual accuracy–efficiency trade-off beyond FLOPs alone.

Table 4: ImageNet-1K accuracy and practical inference efficiency. All efficiency metrics are measured on an A100 GPU with batch size 256 under FP32 precision (native PyTorch, no compilation).

Method	Top-1 (%)	FLOPs (G)	Throughput (img/s)	Latency (ms/img)	Memory (GB)
Swin-T	81.3	4.5	310.0	0.83	3.6
Vim-S	80.5	4.8	285.0	0.90	3.4
VMamba-T	82.6	5.2	295.0	0.87	3.5
ConvNeXt-T	82.1	4.5	480.0	0.53	3.2
VTDS (Ours)	**82.5**	34.85	410.0	0.62	5.3

All ablations use the same ImageNet-1K training protocol and model dimension. Unless specified otherwise, the default configuration uses $L = 10$ historical positions, $R = 6$ fuzzy rules, and alternating four-order serialization. Starting from a recurrent baseline with diagonal recurrence and an outer feature gate, we progressively add SLE, OHP, FSR, and AS (Table 5). Each component yields a positive increment, and together they raise Top-1 by 3.6 points to 82.5%, indicating that the four mechanisms are complementary; the largest single gain comes from SLE, validating the benefit of token-conditioned historical retrieval. In the fusion comparison, the fuzzy regulator outperforms fixed summation and the MLP and sigmoid gates, while remaining bounded and parameter-efficient. As the history length L grows, accuracy improves with more available candidates and saturates at $L = 10$, beyond which a larger window adds computation and potentially irrelevant candidates, making $L = 10$ the best accuracy–efficiency trade-off.

Table 5: Controlled ablations on ImageNet-1K.

Group	Variant	Top-1 (%)	Δ
Components	Recurrent baseline	78.9	–
	+ SLE	80.2	+1.3
	+ OHP	80.9	+0.7
	+ FSR	81.8	+0.9
	+ AS	82.5	+0.7
Fusion	Fixed summation	81.4	–
	Sigmoid gate	81.8	+0.4
	MLP gate	82.1	+0.7
	Fuzzy regulator	82.5	+1.1
History L	4	81.6	–
	6	82.0	–
	8	82.3	–
	10	82.5	–
	12	82.4	–