SIMPLE APPROACH TO TOKEN PRUNING FOR VISION STATE-SPACE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Token pruning is a highly effective strategy for accelerating vision architectures, yet adapting it to state-space models presents unique challenges. Although Vision Mamba models offer linear sequence complexity, they remain computationally costly at the massive token counts required for high-resolution images. Existing pruning methods developed for Vision Transformers cannot resolve this without additional adjustment, as they do not directly account for the recurrent, order-dependent computation of state-space models. Removing tokens changes subsequent state updates and can disrupt information propagation along the sequence. We introduce DeltA, a parameter-free token-pruning importance metric designed for vision state-space models. DeltA scores each token using the interaction between its input-dependent discretization parameter Δ_t and the magnitude of its post-block representation. The method uses signals already available in the backbone and requires no learned pruning modules. Experiments on image classification, semantic segmentation, and facial-attribute recognition using ViM and PlainMamba architectures show a favorable accuracy-computation trade-off and further indicate that the same criterion generalizes across different vision Mamba architectures and tasks.

1 INTRODUCTION

Vision Transformers Dosovitskiy (2020) have established a strong performance baseline across computer vision, but their quadratic computational complexity limits their scalability to high-resolution images. Mamba Gu et al. (2021); Gu & Dao (2023); Lahoti et al. (2026) was created as an efficient alternative to Transformers Vaswani et al. (2017) by replacing quadratic self-attention with state-space models (SSMs) whose computational complexity scales linearly with sequence length. Through input-dependent state-space parameters, Mamba maintains content-aware modeling typical of transformer models, while avoiding the quadratic cost of attention. This allows for competitive performance in both natural language processing Fu et al. (2023); Ren et al. (2025); Glorioso et al. (2024) and vision tasks Shi et al. (2024); Yang et al. (2024). However, despite having lower complexity compared to transformers, Vision Mamba (ViM) Liu et al. (2024); Zhu et al. (2024) models can still have substantial computational cost when processing long sequences of visual tokens.

Token pruning is used as an effective tool for reducing this cost in Vision Transformers by removing tokens estimated to contribute little to the final prediction Rao et al. (2021); Kong et al. (2022); Yu et al. (2022); Pan et al. (2021). Applying the same idea to visual SSMs, however, requires additional consideration. Mamba processes tokens through recurrent state-space updates, allowing the state at one position to contribute to the computation of later positions. Removing a token may change both its local contribution and the states computed over the remaining sequence. This implies that pruning criteria for Mamba models should be investigated separately. This includes developing custom approaches, which take the specifics of the architecture, such as incorporating signals from the internal state-space dynamics, into account.

A natural signal for this purpose is already present in Mamba. The input-dependent discretization parameter Δ_t controls the discretized state-space updates for each token and therefore provides an indication of how strongly the current input influences the overall value of state-spaces. Recent work has exploited this property for token pruning in Mamba-based vision-language models Park et al. (2026) and Vision Mamba architectures Huang & Yoon (2026), demonstrating that Δ_t con-

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.

tains useful information for identifying redundant visual tokens. Unlike methods that use learnable token selectors, this signal is obtained directly from the model's existing computation and does not require an auxiliary importance predictor. However, Δ_t alone may provide an incomplete measure of token importance. It characterizes one component of the state-space update but does not directly account for the magnitude of the representation propagated by the block. Tokens with similar discretization signals may therefore produce representations with different magnitudes and, therefore, have different influences on downstream computations.

Motivated by this observation, we introduce DeltA (**Delt**a+**A**ctivation), a token-importance metric that combines the token-level discretization signal Δ_t with the L_2-norm of the corresponding post-SSM block token representation, $\|\mathbf{y}_t\|_2$. The two quantities provide different information: Δ_t is associated with the token's role in the current state-space update, whereas $\|\mathbf{y}_t\|_2$ measures the magnitude of the representation passed to subsequent layers. Because both are obtained from the existing forward pass, DeltA does not require auxiliary token selectors or additional learnable pruning parameters. We fine-tune the backbone with pruning enabled to adapt it to the shortened token sequence.

We evaluate DeltA on ImageNet-1K, ADE20K, and CelebA, spanning image classification, semantic segmentation, and multi-label facial-attribute recognition. Across tasks and model scales, DeltA reduces computation while maintaining strong performance. On ImageNet-1K, DeltA achieves 75.3% and 79.4% Top-1 accuracy with ViM-Tiny and ViM-Small while reducing theoretical FLOPs by 16% and 32%, respectively. These results demonstrate that combining the intrinsic state-space signal with representation magnitude provides an effective pruning metric across different vision tasks and ViM architectures, without introducing learned token selectors, hidden-state alignment modules, or token-merging stages.

Our main contributions are:

- We identify a limitation of one metric token pruning in Vision Mamba: although the discretization parameter provides a useful signal of a token's influence in the recurrent dynamics, it does not capture the magnitude of the resulting token representation, which provides additional information for token importance.
- We introduce DeltA, a parameter-free importance criterion that combines the token-level discretization response with the L_2 magnitude of the corresponding Mamba output representation, providing complementary information without introducing additional learnable pruning components.
- We demonstrate consistent efficiency gains compared to current approaches across image classification, semantic segmentation, and facial-attribute recognition, showing that DeltA generalizes across tasks, architecture variants, and Vision Mamba model scales.

2 Related work

Vision Mamba optimization Most current work on reducing the size and increasing the speed of vision Mamba architectures focuses on designing lightweight variants or modifying computationally expensive operations. For instance, EfficientViM Lee et al. (2025) shifts expensive hidden-state channel mixing operations from the raw image space to a compressed latent space, thereby reducing the total computational load. MobileMamba He et al. (2025) addresses hardware performance bottlenecks associated with the parallel processing of SSMs. It improves throughput by removing a downstream processing stage and introducing custom modules that split intermediate features into parallel paths. On the other hand, the TinyVim Ma et al. (2025) approach targets very small Mamba architectures by introducing a Laplace mixer to decouple the processing of low- and high-frequency components. This enhances the model's ability to capture fine, high-frequency details, an area where smaller SSMs typically struggle. However, none of these structural variations address the fundamental challenge of spatial input redundancy in images.

Token pruning In the parallel domain of Vision Transformers, spatial redundancy is routinely managed via token sparsification or dynamic pruning, where selection modules gradually discard uninformative visual patches to systematically downscale operation volume. DynamicViT Rao et al. (2021) inserts independent, fully learnable neural network blocks between specific transformer layers to predict binary drop/keep masks. IA-RED2 Pan et al. (2021) incorporates a multi-head in-

terpreter module to rank token hierarchies, while SPViT Kong et al. (2022) utilizes a multi-head attention-based selector tailored to real-world hardware latency profiles, which merges discarded patches into a singular soft-pruning 'package token'. Other strategies rely on heavy, multi-stage architecture search loops Chavan et al. (2022) to optimize the pruning mask. These approaches paradoxically expand the model's parameter footprint and require compute-intensive retraining routines before any compression is realized. EViT Liang et al. (2022) offers a lightweight compression alternative and avoids added selector weights by using existing classification tokens. Applying these frameworks directly to visual SSMs, however, introduces severe structural limitations. Whereas transformers process tokens as a permutation-invariant bag-of-features, the continuous scan lines of visual SSMs rely entirely on strict chronological and spatial sequential dependencies. Directly deleting a patch token from an active sequence alters the underlying recurrence, which can disrupt historical context and subsequent latent state updates.

Recent exploratory works addressing Mamba compression have attempted to absorb token-deletion shocks through specialized, pruning-aware hidden-state alignment layers Zhan et al. (2024), or by evaluating the distribution profiles of native state parameters across different network depths in multimodal large language models Park et al. (2026). However, we argue that Mamba token pruning can be efficiently done by a proper selection of an importance metric, without resorting to hidden-state alignments. Our work builds on this direction by designing an intrinsic importance metric tailored to Mamba dynamics, enabling token pruning without auxiliary selection or alignment modules.

3 METHOD

We introduce DeltA, a token-importance criterion for Vision Mamba that combines Mamba's input-dependent discretization signal with the magnitude of the corresponding post-block output. We first describe the role of Δ in the recurrent state-space update and then present the combined scoring function and the differentiable token-selection procedure used during fine-tuning.

3.1 STATE-SPACE TOKEN IMPORTANCE

Mamba builds on a continuous-time state-space model that maps an input $x(t)$ to an output $y(t)$ through a latent state $\mathbf{s}(t) \in \mathbb{R}^N$:

$$\dot{\mathbf{s}}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{B}x(t), \qquad y(t) = \mathbf{C}\mathbf{s}(t), \tag{1}$$

where $\mathbf{A}$ represents the state transition and $\mathbf{B}$ and $\mathbf{C}$ define the input and output mappings. In Mamba, the discretization step Δ, together with $\mathbf{B}$ and $\mathbf{C}$, is conditioned on the input, allowing the state-space dynamics to vary across tokens.

For a discrete sequence, the continuous system is discretized using the input-dependent step size Δ_t. Under zero-order hold, the discretized operators are

$$\overline{\mathbf{A}}_t = \exp\left(\Delta_t \mathbf{A}\right), \qquad \overline{\mathbf{B}}_t = \left(\Delta_t \mathbf{A}\right)^{-1} \left[\exp\left(\Delta_t \mathbf{A}\right) - \mathbf{I}\right] \Delta_t \mathbf{B}_t, \tag{2}$$

which gives the recurrent update

$$\mathbf{s}_t = \overline{\mathbf{A}}_t \mathbf{s}_{t-1} + \overline{\mathbf{B}}_t x_t. \tag{3}$$

For sufficiently small Δ_t, the discretized operators can be approximated as

$$\overline{\mathbf{A}}_t \approx \mathbf{I} + \Delta_t \mathbf{A}, \qquad \overline{\mathbf{B}}_t \approx \Delta_t \mathbf{B}_t. \tag{4}$$

A smaller Δ_t therefore makes the state transition closer to an identity mapping and reduces the immediate contribution of the current input to the state update. Since a larger Δ_t results in a stronger difference from the previous state, it is implied that this parameter could provide a good metric as to how much the corresponding token influences the current layer.

In practice, Mamba predicts a channel-wise discretization value for each token. Let $\widetilde{\Delta}_{t,c}$ denote the projected time-step value for token t and inner channel c. The positive discretization value used by the selective scan is

$$\Delta_{t,c} = \mathrm{softplus}\left(\widetilde{\Delta}_{t,c} + b_{\Delta,c}\right). \tag{5}$$

For a block with a single scan, we obtain a scalar token-level discretization parameter by averaging over the inner channels:

$$\Delta_t = \frac{1}{d_{\text{inner}}} \sum_{c=1}^{d_{\text{inner}}} \Delta_{t,c}. \tag{6}$$

Vision Mamba architectures may process the token sequence using multiple scan directions or scan orders. Let $\mathcal{S}$ denote the set of scan paths and let π_s^{-1} map positions in scan s back to the original token order. The aggregated token-level discretization signal is

$$\Delta_t = \frac{1}{|\mathcal{S}| d_{\text{inner}}} \sum_{s \in \mathcal{S}} \sum_{c=1}^{d_{\text{inner}}} \Delta^{(s)}_{\pi_s^{-1}(t),c}. \tag{7}$$

For bidirectional ViM, $\mathcal{S}$ contains the forward and backward scans, with the backward values realigned to the original token order before aggregation. For architectures in which multiple scan paths share the same discretization values, equation 7 reduces to the channel-wise average in equation 6.

3.2 DeltA: Delta–Activation Token Importance

The discretization parameter Δ_t describes the token's association with the input-dependent state-space update, but it does not directly account for the magnitude of the representation propagated to subsequent layers. We therefore combine Δ_t with the magnitude of the token representation after the Mamba block, in an attempt to extract information about the importance of the token for subsequent layers.

Let $\mathbf{y}_t \in \mathbb{R}^{d_{\text{model}}}$ denote the output of the Mamba block for token t, after any architecture-specific directional aggregation and output projection. We define the post-block output magnitude as

$$A_t = \|\mathbf{y}_t\|_2 . \tag{8}$$

The two signals provide different information about the token. The discretization signal Δ_t is derived from the input-dependent state-space dynamics, whereas A_t measures the magnitude of the representation retained after the current block. We define the DeltA importance score as

$$\mathcal{I}_t = \Delta_t A_t = \Delta_t \|\mathbf{y}_t\|_2. \tag{9}$$

A token receives a high score when both its discretization signal and post-block representation magnitude are large. A small value in either component reduces its priority for retention.

The quantities required by DeltA are obtained from the existing forward pass. The method introduces only a token-wise norm and multiplication and requires no auxiliary token selector or additional learnable pruning parameters.

Differentiable token selection. During fine-tuning, we use $\ell_t = \log(D_t + \epsilon) + \log(A_t + \epsilon)$ as the token-selection logit and apply a Gumbel–Sigmoid relaxation with a straight-through estimator Jang et al. (2017); Maddison et al. (2016). At each pruning stage, the hard mask retains the target fraction of tokens with the largest relaxed scores, while gradients are propagated through the soft decisions. At inference, Gumbel noise is removed, and tokens are selected deterministically according to their DeltA scores.

4 Experiments

We first examine how the signals underlying DeltA evolve across network depth. Following prior work that employs multi-stage pruning in Mamba-based models Zhan et al. (2024); Park et al. (2026), we analyze the distributions of the two components of our importance criterion: the token representation magnitude A_t and the input-dependent discretization response Δ_t, at equally distant layers 6, 12, and 18 of the ViM-Tiny architecture. Figure 1 shows the distributions of these quantities using violin plots.

Both signals show different behaviour based on the depth of the layer. At layers 6 and 12, their values are more uniformly distributed within relatively small ranges across tokens. At layer 18,

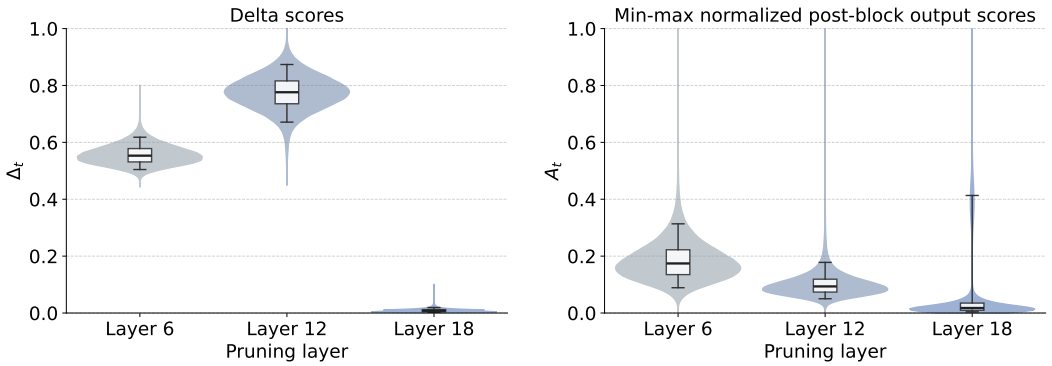


Figure 1: Layer-wise distributions of pruning signals. The left panel shows delta-based token-importance scores Δ_t, while the right panel shows post-block token value norms min–max normalized independently within each pruning layer.

however, the distributions become significantly more asymmetric, with a large fraction of tokens concentrated toward lower-valued responses and a smaller subset extending into a high-valued tail. This pattern is particularly evident in the representation magnitudes, where only a limited subset of tokens retains substantially stronger responses than the remainder.

These observations suggest that token importance becomes increasingly concentrated at greater network depth, indicating that a uniform pruning ratio across all considered layers may be suboptimal. We therefore adopt a multi-stage pruning strategy that preserves a larger fraction of tokens in earlier layers while progressively increasing the pruning strength in later stages.

Next, we evaluate DeltA across three vision tasks to assess both its efficiency and transferability: ImageNet-1K classification, CelebA facial attribute recognition, and ADE20K semantic segmentation. We compare the results to the baseline reported by the authors in Zhan et al. (2024).

4.1 Image Classification on ImageNet-1K.

Experimental setup We initialize the model weights from pretrained ViM-Tiny and ViM-Small checkpoints and apply staged pruning at blocks 6, 12, and 18. Retention rates follow geometric schedules of 0.9, 0.9^2, and 0.9^3 for ViM-Tiny, and 0.75, 0.75^2, and 0.75^3 for ViM-Small, with stronger pruning in deeper layers. These retention rates were chosen specifically because they reduce the network to several FLOPs, which is comparable to the current SOTA approaches. In further text, the phrase retention rate signifies the value of the first pruned layer, while the subsequent layers follow the same geometric schedule. The pruned models are then fine-tuned for 30 epochs. In order to improve on the fine-tuning process, feature-level knowledge distillation from the unpruned baseline is added as an additional loss. We use AdamW with a learning rate of 5×10^{-4}, weight decay of 1×10^{-8}, cosine learning rate decay, and a batch size of 512. FLOPs are measured after physically removing pruned tokens from subsequent Mamba blocks. All of the experiments were run using NVIDIA H100 GPUs.

Results Table 1 compares the classification accuracy and computational complexity (GFLOPs) of our method against contemporary compression frameworks across two Vision Mamba scales. On the ViM-Tiny baseline, traditional token reorganization via EViT Liang et al. (2022) drops Top-1 accuracy by 4.8% down to 71.3%. While the state-space-adapted ToP-ViM Zhan et al. (2024) improves this to 75.1%, our framework achieves the highest accuracy at 75.3%. Crucially, this gain is achieved alongside the largest computational reduction in the Tiny class, reducing the model to 1.26 GFLOPs (a 16.0% reduction).

We observe a similar trend with the larger ViM-Small configuration. At a 32% computational reduction, EViT collapses to 74.8% accuracy. ToP-ViM maintains 78.8% accuracy, but requires complex external hidden-state alignment modules to resolve the numerical shock of patch deletion along the 1D scan lines. In contrast, our framework achieves an accuracy of 79.4% while maintaining a comparable footprint of 3.61 GFLOPs. These results demonstrate that using the joint importance metric

Table 1: Comparison of computational efficiency and classification accuracy against token pruning and compression methods across different Vision Mamba scales.

Model	**GFLOPs** ↓	**Top-1 Acc (%)**↑
ViM-T (Baseline)	1.50	76.1
ViM-T-EViT Liang et al. (2022)	1.28 (-14.30%)	71.3
ViM-T-ToP Zhan et al. (2024)	1.29 (-14.00%)	75.1
Ours	1.26 (-16.00%)	**75.3**
ViM-S (Baseline)	5.30	80.5
ViM-S-EViT Liang et al. (2022)	3.57 (-32.64%)	74.8
ViM-S-ToP Zhan et al. (2024)	3.60 (-32.07%)	78.8
Ours	3.61 (-31.89%)	**79.4**

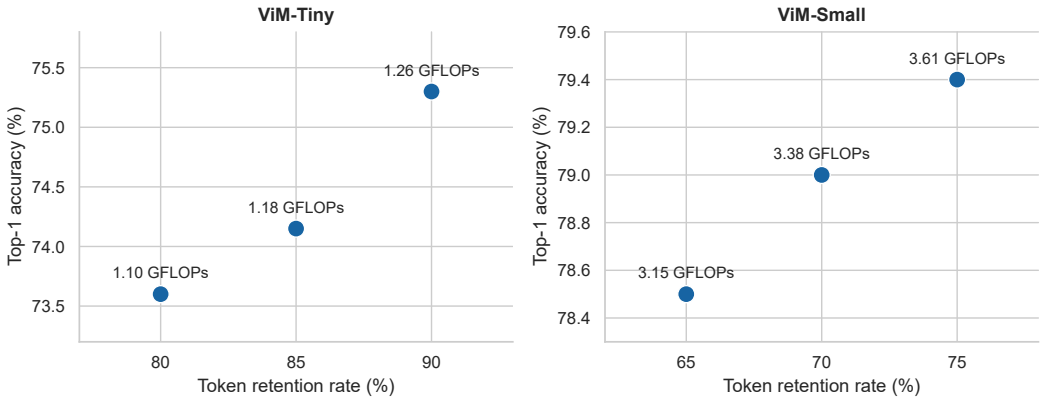


Figure 2: Top-1 accuracy at different token-retention rates for ViM-Tiny and ViM-Small on ImageNet-1K. Labels above the points indicate GFLOPs. The two panels use different vertical scales.

allows the pruned model to outperform existing baselines without introducing auxiliary parameter overhead or token alignment modules.

Figure 2 shows Top-1 accuracy and number of FLOPs for both architectures while changing the percentage of preserved tokens. The results show that ViM-Small maintains accuracy as token retention decreases from 0.75 to 0.65, with a drop of 0.9 percentage points, whereas ViM-Tiny loses 1.7 points as retention decreases from 0.9 to 0.8, where the larger drop for ViM-Tiny could be explained by less information redundancy in the smaller model.

To verify that these computational and accuracy gains translate into physical deployment advantages, we profile the real-world inference throughput of our framework. Table 2 details the classification accuracy, computational footprint, and physical wall-clock acceleration on the ViM-Small architecture. The throughput benchmarks were conducted on a workstation equipped with an NVIDIA RTX 4070 GPU and compared against the reported throughput of the ToP-ViM method. As the network operates at this scale, the reduction in sequence length via token pruning translates into physical hardware acceleration. Our framework achieves a physical speedup of $1.24\times$ over the dense baseline. Under a comparable number of floating-point operations, these results are competitive with the $1.27\times$ hardware speedup reported by ToP-ViM, while maintaining a higher classification accuracy of 79.4%.

4.2 Face Attribute Recognition on CelebA.

Experimental setup To assess whether DeltA transfers beyond conventional image classification and segmentation, we evaluate ViM-Tiny and ViM-Small on CelebA Liu et al. (2015) using the same

Table 2: Physical deployment efficiency and wall-clock acceleration profiling on an NVIDIA RTX 4070 GPU

Pruning Method	Accuracy ↑	GFLOPs ↓	Speedup ↑
TOP-VIM Zhan et al. (2024)	78.8%	3.60	1.27×
Ours	79.4%	3.61	1.24×

progressive pruning schedule as in our ImageNet experiments. We adapt the ImageNet-pretrained backbones to 40-way attribute prediction and apply the same staged pruning configuration. Models are fine-tuned for 30 more epochs using AdamW with a learning rate of 1×10^{-4} and weight decay of 1×10^{-8}.

Results As shown in Table 3, DeltA preserves multi-label recognition performance while reducing computation across both ViM scales. Facial-attribute recognition provides a sensitive evaluation of pruning because precision, recall, and F1 capture changes in the balance between positive and negative predictions that may not be reflected by aggregate accuracy alone. Despite this sensitivity, precision, recall, and F1 remain within approximately two percentage points of their corresponding unpruned baselines, while average accuracy remains at a comparable level. The limited changes in both precision and recall suggest that the model retains a similar prediction profile after pruning. Overall, these results indicate that the same importance criterion can be applied to structured multi-label recognition without a task-specific pruning mechanism.

Table 3: Performance and computational cost across model scales on CelebA.

Config	Avg Acc ↑	Precision ↑	Recall ↑	F1 ↑	GFLOPs ↓
ViM-Tiny					
Full-size	91.2%	77.5%	66.0%	70.1%	1.50
Ours	91.1%	77.1%	65.2%	69.5%	1.26
ViM-Small					
Full-size	91.3%	77.4%	67.0%	70.9%	5.30
Ours	90.8%	75.1%	66.1%	69.5%	3.61

4.3 SEMANTIC SEGMENTATION ON ADE20K

Experimental setup. We evaluate DeltA on semantic segmentation using ADE20K Zhou et al. (2017) and UPerNet Xiao et al. (2018) with PlainMamba Yang et al. (2024) backbones. Whereas the ViM models used in our classification experiments process the sequence bidirectionally, PlainMamba uses a continuous two-dimensional scanning strategy. This setting allows us to examine whether the same importance criterion can be applied to a different vision Mamba architecture and to a dense-prediction task. We consider PlainMamba-L1 and PlainMamba-L2 and follow the original PlainMamba training protocol Yang et al. (2024). The same pruning procedure and retention schedule are used for both backbone scales as in the case of image classification.

Table 4: Semantic-segmentation results on ADE20K.

Model	mIoU (%) ↑
PlainMamba-L1	
Full model	44.24
EViT Liang et al. (2022)	42.20
ToP-ViM Zhan et al. (2024)	44.10
DeltA	**44.58**
PlainMamba-L2	
Full model	46.82
EViT Liang et al. (2022)	44.10
ToP-ViM Zhan et al. (2024)	46.50
DeltA	**46.72**

Results. Table 4 reports the semantic-segmentation results. With PlainMamba-L1, DeltA obtains 44.58% mIoU, beating the performance of the unpruned model, as well as ToP-ViM, and EViT. With PlainMamba-L2, DeltA obtains 46.72% mIoU, 0.10 percentage points below the unpruned model and above both pruning baselines. These results indicate that DeltA preserves dense-prediction performance when transferred from bidirectional ViM models to PlainMamba's two-dimensional

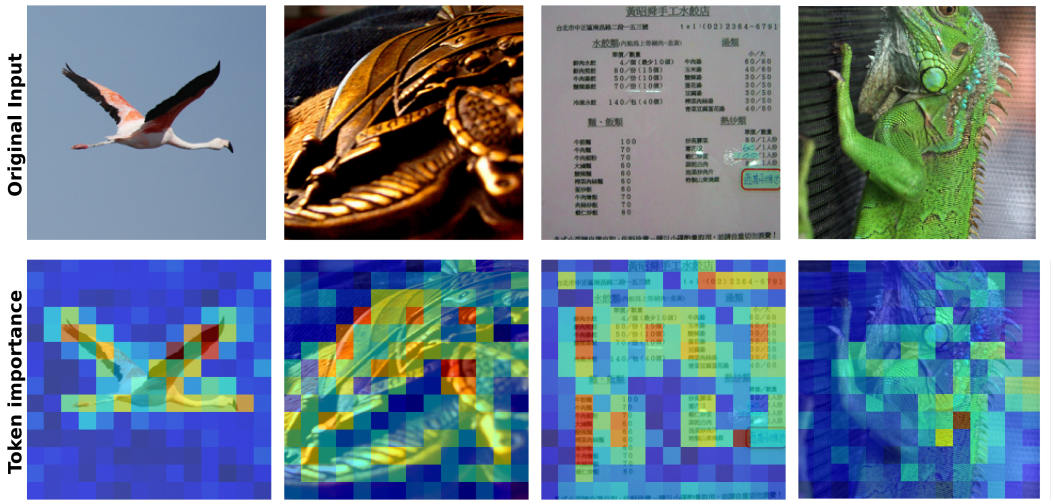


Figure 3: **Spatial distribution of DeltA importance scores prior to pruning. Top:** Representative validation images. **Bottom:** Continuous heatmaps of DeltA token importance scores $\mathcal{I}_t$ at Layer 6 of ViM-Tiny, projected to their corresponding spatial locations. Higher intensity indicates a larger importance score.

scanning architecture. Notably, these results are obtained without introducing a task-specific pruning mechanism or modifying the retention schedule, suggesting that the DeltA importance criterion transfers effectively across both task type and scanning topology.

5 Analysis and Ablation Studies

Beyond accuracy and efficiency metrics, we further analyze how DeltA behaves at the token level and whether its individual components contribute complementary information. We first examine the spatial structure of the learned importance scores and the resulting feature localization after pruning. We then perform a component-wise ablation to isolate the contributions of the discretization response and token representation magnitude across different ViM scales.

5.1 Visualization

Before introducing the pruning procedure, we examine the behavior of the proposed importance criterion in both the spatial and depth dimensions of ViM. We first visualize the continuous DeltA importance scores $\mathcal{I}_t$ prior to any hard pruning decisions. Specifically, we project the scores obtained at layer 6, the first pruning stage of ViM-Tiny, back to their corresponding spatial locations in the input image. As shown in Fig. 3, the resulting maps exhibit clear spatial structure: tokens corresponding to foreground objects and prominent structural boundaries tend to receive higher importance scores, whereas many background regions receive comparatively low scores. This suggests that DeltA captures meaningful spatial variation in token importance rather than assigning scores uniformly across image patches.

Next, we inspect the visual localization profiles of the pruned state-space models before and after zero-shot sequence reduction. This is achieved by mapping the network's internal representation layer topologies using Gradient-weighted Class Activation Mapping (Grad-CAM) Selvaraju et al. (2017). Figure 4 contrasts the qualitative activation distributions for representative ImageNet-1K samples, mapping the raw baseline model against our pruned counterpart. In the examples shown, activation maps after pruning appear more concentrated around the target object, while the unpruned maps show broader responses to background regions. This suggests that DeltA can retain tokens associated with salient image structures while removing some background tokens. These visualizations provide qualitative support for the selection criterion, but do not establish its effect on recurrent state dynamics.

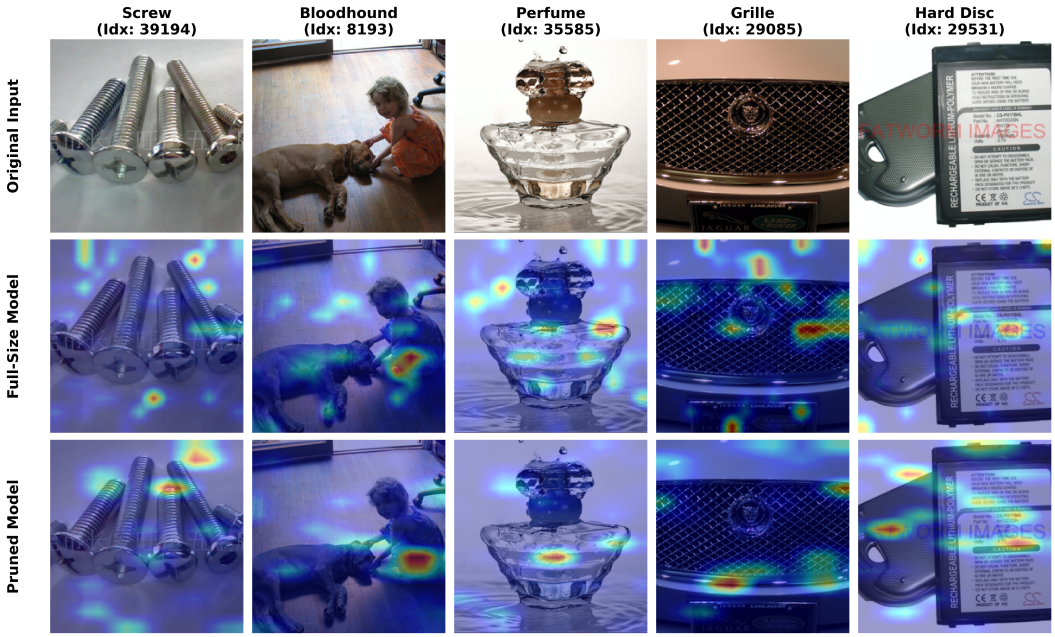


Figure 4: Qualitative feature localization comparison via Grad-CAM.

5.2 Ablation of the DeltA importance criterion.

Table 5 shows the Top-1 accuracy of the fine-tuned pruned models when using either the discretization parameter or the block output magnitude as the importance criterion, compared with the proposed combined metric. The Δ_t-only criterion performs strongly on ViM-Small but is less effective on ViM-Tiny. On the other hand, the activation-only criterion performs best on ViM-Tiny but degrades on ViM-Small. In contrast, DeltA remains consistently competitive across both architectures, suggesting that the two signals capture complementary aspects of token importance. This behavior supports the use of the combined criterion as a more robust importance measure across different ViM scales.

Table 5: Ablation of the DeltA importance metric components on ImageNet-1K.

Backbone	Δ_t only	A_t only	DeltA
ViM-Tiny	74.1%	**75.6%**	75.3%
ViM-Small	**80.0%**	79.1%	79.4%

6 Conclusion

We introduced DeltA, a lightweight token pruning method for vision state-space models that derives token importance directly from quantities already present in the Mamba block. By combining the local state-space influence captured by Δ_t with downstream post-block token representation magnitude, DeltA provides a more complete pruning signal than either component alone, without introducing auxiliary token-selection modules or additional token alignment mechanisms. Across ImageNet-1K, ADE20K, and CelebA, the method consistently reduces computation while preserving strong predictive performance across classification, dense prediction, and multi-label recognition, and outperforms current state-of-the-art approaches. These results suggest that the internal dynamics of state-space models provide a useful basis for efficient token reduction. More broadly, our results indicate that efficient token reduction in visual SSMs can be achieved by exploiting their native information rather than adding external pruning machinery or combining the approach with token alignment mechanisms. It is, however, important to note that Mamba models still face limitations when it comes to deployment to lower-level hardware. Since parallelization of sequentially dependent state-space modules depends highly on specialized kernels, their deployment remains limited by currently available software.

AI USE STATEMENT

In this work, we used generative AI tools for assisting with translation, designing or providing feedback on research methodology or experiments. We have not used generative AI tools for implementing methods, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, or interpreting results. Generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, are not applicable to this work.

Additionally, we used generative AI tools for creating or editing software code, brainstorming, identifying relevant literature, editing a research paper to improve readability, and summarizing or analysing existing literature. We have reviewed all AI-assisted work. All LLM-generated code was checked and tested, as well as LLM-generated text. The text is original, with generative AI used only for polishing and improving the text. All references and literature were manually checked. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Sections 3.1–3.2 provide the complete mathematical definition of DeltA, including the construction of the token-importance score and the token-selection procedures used during training and inference. Section 4 reports the datasets, pretrained backbones, pruning locations and retention schedules, optimization settings, training duration, hardware, and efficiency-measurement protocols used in our experiments. All experiments use publicly available datasets and released pretrained model checkpoints. Following the review decision, we will release the complete implementation in a public GitHub repository, including dependency specifications, experiment configurations, and scripts for training, evaluation, computational-cost measurement, and reproducing the analyses and results reported in the paper.

REFERENCES

Arnav Chavan, Zhiqiang Shen, Zhuang Liu, Zechun Liu, Kwang-Ting Cheng, and Eric P. Xing. Vision transformer slimming: Multi-dimension searching in continuous optimization space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4931–4941, 2022.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Daniel Y Fu, Tri Dao, Khaled Kamal Saab, Armin W Thomas, Atri Rudra, and Christopher Re. Hungry hungry hippos: Towards language modeling with state space models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=COZDy0WYGg.

Paolo Glorioso, Quentin Anthony, Yury Tokpanov, James Whittington, Jonathan Pilault, Adam Ibrahim, and Beren Millidge. Zamba: A compact 7B SSM hybrid model. *arXiv preprint arXiv:2405.16712*, 2024.

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

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

Haoyang He, Jiangning Zhang, Yuxuan Cai, Hongxu Chen, Xiaobin Hu, Zhenye Gan, Yabiao Wang, Chengjie Wang, Yunsheng Wu, and Lei Xie. MobileMamba: Lightweight multi-receptive visual Mamba network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4497–4507, 2025.

Huiyuan Huang and Sang Min Yoon. SSM-aware token-efficient VMamba via adaptive patch pruning and merging for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4363–4372, 2026.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=rkE3y85ee.

Zhenglun Kong, Peiyan Dong, Xiaolong Ma, Xin Meng, Wei Niu, Mengshu Sun, Xuan Shen, Geng Yuan, Bin Ren, Hao Tang, et al. SPViT: Enabling faster vision transformers via latency-aware soft token pruning. In *European Conference on Computer Vision*, pp. 620–640. Springer, 2022.

Aakash Sunil Lahoti, Kevin Li, Berlin Chen, Caitlin Wang, Aviv Bick, Zico Kolter, Tri Dao, and Albert Gu. Mamba-3: Improved sequence modeling using state space principles. In *International Conference on Learning Representations*, volume 2026, pp. 86173–86202, 2026.

Sanghyeok Lee, Joonmyung Choi, and Hyunwoo J. Kim. EfficientViM: Efficient vision Mamba with hidden state mixer based state space duality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14923–14933, 2025.

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. *arXiv preprint arXiv:2202.07800*, 2022.

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.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

Xiaowen Ma, Zhenliang Ni, and Xinghao Chen. TinyViM: Frequency decoupling for tiny hybrid vision Mamba. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23519–23529, 2025.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

Bowen Pan, Rameswar Panda, Yifan Jiang, Zhangyang Wang, Rogerio Feris, and Aude Oliva. IA-RED^2: Interpretability-aware redundancy reduction for vision transformers. *Advances in Neural Information Processing Systems*, 34:24898–24911, 2021.

Seong Yeol Park, Min Jung Kwon, Xianghua Piao, and Yeong Hyeon Gu. DTP: Delta-guided two-stage pruning for Mamba-based multimodal large language models. In *The Fourteenth International Conference on Learning Representations*, 2026.

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.

Liliang Ren, Yang Liu, Yadong Lu, Chen Liang, Weizhu Chen, et al. Samba: Simple hybrid state space models for efficient unlimited context language modeling. In *International Conference on Learning Representations*, 2025.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision*, pp. 418–434, 2018.

Chenhongyi Yang, Zehui Chen, Miguel Espinosa, Linus Ericsson, Zhenyu Wang, Jiaming Liu, and Elliot J. Crowley. PlainMamba: Improving non-hierarchical Mamba in visual recognition. *arXiv preprint arXiv:2403.17695*, 2024.

Fang Yu, Kun Huang, Meng Wang, Yuan Cheng, Wei Chu, and Li Cui. Width and depth pruning for vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 3143–3151, 2022.

Zheng Zhan, Zhenglun Kong, Yifan Gong, Yushu Wu, Zichong Meng, Hangyu Zheng, Xuan Shen, Stratis Ioannidis, Wei Niu, Pu Zhao, et al. Exploring token pruning in vision state space models. *Advances in Neural Information Processing Systems*, 37:50952–50971, 2024.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 633–641, 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. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=YbHCqn4qF4.