# BEYOND LINEAR SCALING: RECURRENT CALIBRATION FOR CROSS-RESOLUTION TRANSFER IN PRETRAINED VISUAL MAMBA MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Mamba has emerged as a leading paradigm for long-sequence modeling and has been widely extended to vision. Despite its linear complexity and appeal for high-resolution representations, pretrained visual Mamba models often degrade severely at unseen resolutions. In this paper, we reveal that resolution scaling in visual Mamba is not merely a one-dimensional sequence extension, but entails a fundamental change in the two-dimensional *spatial rate* of pretrained selective dynamics. The resulting structural mismatch propagates through state-dependent residual updates and hierarchical transitions. To quantify this resolution-induced trajectory drift, we formulate *Cross-Resolution Trajectory Discrepancy* (CRTD). Pre-calibration analysis reveals a strong association between raw trajectory discrepancy and cross-resolution accuracy degradation. Based on these insights, we propose *Architecture-Native Uniform Residual/Update Calibration* (Uniform-RU), a lightweight and label-free framework that calibrates recurrent dynamics and residual trajectories using only a small fraction of the training images while keeping the pretrained backbone frozen. Across diverse visual Mamba architectures and resolutions, Uniform-RU substantially restores cross-resolution accuracy on image classification and generalizes well to detection and segmentation tasks. These results demonstrate that recurrent calibration can effectively extend pretrained visual Mamba models, offering a dynamics-centric perspective for resolution transfer beyond sequence scaling. Code is available at `https://anonymous.4open.science/r/Uniform-RU/`.

## 1 INTRODUCTION

State-space models (SSMs) enable efficient linear or near-linear sequence scaling (Gu et al., 2022; 2020), while Mamba further introduces input-dependent selective dynamics (Gu & Dao, 2024; Dao & Gu, 2024), inspiring a diverse family of visual Mamba architectures (Zhu et al., 2024; Liu et al., 2024; Hatamizadeh & Kautz, 2025; Shi et al., 2024; Pei et al., 2025). This efficiency is attractive for high-resolution vision, where Transformer self-attention over patch-token sequences (Vaswani et al., 2023; Dosovitskiy et al., 2021) becomes increasingly costly. However, linear complexity does not guarantee resolution generalization. As shown in Figure 1, pretrained visual Mamba models degrade substantially at unseen resolutions across architectures and scales.

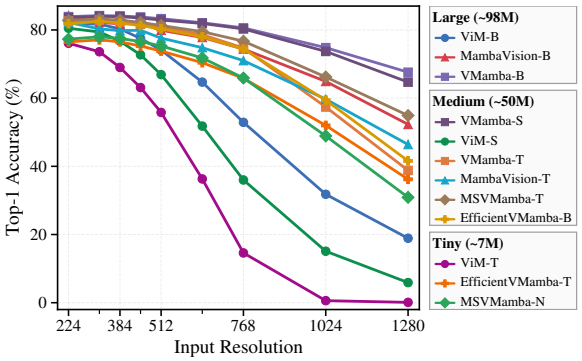



Figure 1: Cross-resolution degradation of representative pretrained visual Mamba models on ImageNet-1K.

Recent studies have shown that Mamba's favorable sequence-length scaling alone does not ensure reliable extrapolation beyond training lengths (Ben-Kish et al.,



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.

2025; Ye et al., 2025; Azizi et al., 2025; Lu et al., 2025). Cross-resolution transfer in vision, however, is more than one-dimensional length extrapolation: with fixed patch or feature construction, higher resolution samples the same two-dimensional domain on a denser lattice, increasing recurrent steps while reducing the spatial distance represented by each step. For scan-based visual Mamba models, this changes the *spatial rate* at which pretrained selective dynamics are executed (Nguyen et al., 2022; Li et al., 2026; Yu & Park, 2026). Resolution extrapolation therefore changes both trajectory length and the spatial discretization of the learned dynamics.

We show that this change induces a local recurrent-dynamics mismatch that propagates through state-dependent residual updates and can re-emerge across hierarchical transitions, forming a **structural trajectory mismatch**. To quantify this behavior, we introduce *Cross-Resolution Trajectory Discrepancy* (CRTD), which measures drift in propagated states and additive updates. Before calibration, raw CRTD-State and CRTD-Update increase monotonically with Top-1 degradation across all architectures from $320^2$ to $1280^2$.

Guided by this analysis, we propose *Architecture-Native Uniform Residual/Update Calibration* (Uniform-RU), a lightweight, label-free framework that keeps the pretrained backbone frozen. *Spatial Dynamics Transport* (SDT) compensates for the geometry-induced recurrent-scale mismatch, while *Carried-State Correction* (CSC) and *Additive-Update Correction* (UC) calibrate carried residual states and newly produced updates according to each backbone's native spatial operators and residual topology. Across diverse architectures, model scales, and resolutions, Uniform-RU substantially restores classification accuracy while reducing CRTD, and the same principle generalizes to semantic segmentation, object detection, and instance segmentation.

**Contributions.** (1) We identify and empirically characterize a structured recurrent-trajectory mismatch associated with cross-resolution degradation, arising from resolution-induced changes in the spatial execution rate of pretrained selective dynamics. (2) We introduce CRTD to quantify state/update drift and propose Uniform-RU, an architecture-native, label-free calibration framework integrating SDT, CSC, and UC with the pretrained backbone frozen. (3) We validate Uniform-RU across heterogeneous visual Mamba architectures, scales, resolutions, and dense prediction tasks using lightweight calibration.

## 2 Related Work

### 2.1 Visual State-Space Models

Mamba extends structured state-space models with input-dependent selective dynamics while retaining linear sequence scaling (Gu et al., 2022; Gu & Dao, 2024; Gu et al., 2020; Dao & Gu, 2024), motivating their rapid adoption as visual backbones. Vision Mamba (Vim) (Zhu et al., 2024) introduces bidirectional state-space modeling for visual sequences, VMamba (Liu et al., 2024) adapts Mamba to 2D features through selective scanning (SS2D), and MambaVision (Hatamizadeh & Kautz, 2025) combines Mamba with self-attention (Vaswani et al., 2023; Dosovitskiy et al., 2021). Subsequent variants further explore spatial organization: PlainMamba (Yang et al., 2024) and LocalMamba (Huang et al., 2024) study continuous 2D and local scanning, MSVMamba (Shi et al., 2024) introduces multi-scale hierarchical modeling, EfficientVMamba (Pei et al., 2025) employs sparse atrous scanning, and FractalMamba (Tang et al., 2024) improves 2D neighborhood preservation. Collectively, these works highlight the importance of spatial serialization, scanning topology, and hierarchical organization in visual Mamba design.

### 2.2 Cross-Resolution Visual Modeling

Visual models are often sensitive to training resolution and can degrade at unseen resolutions (Touvron et al., 2019; Tian et al., 2023; Li et al., 2026). FixRes (Touvron et al., 2019) studies train–test resolution discrepancies, while ResFormer (Tian et al., 2023) improves ViT's (Dosovitskiy et al., 2021) robustness through scale-consistency learning. FlexiViT (Beyer et al., 2023), NaViT (Dehghani et al., 2023), and ViTAR (Fan et al., 2024) support flexible input resolutions through variable patch sizes, native-resolution processing, or flexible positional representations. However, support for arbitrary resolutions does not ensure reliable cross-resolution generalization (Sakarvadia et al., 2026; Wang & Zhang, 2026). For state-space models, S4ND (Nguyen et al., 2022) studies cross-sampling

generalization from continuous multidimensional signals, while FractalMamba++ (Li et al., 2026) improves resolution scalability through redesigned 2D serialization and state propagation. Existing approaches typically rely on multi-resolution training, continuous representations, or architectural redesign, leaving post-hoc cross-resolution transfer largely unexplored.

### 2.3 Length Generalization and Dynamics Calibration in Mamba

Mamba's linear complexity does not naturally guarantee that learned dynamics generalize to unseen lengths; efficient long-sequence execution and correct state extrapolation remain distinct problems (Ben-Kish et al., 2025; Ye et al., 2025; Azizi et al., 2025; Lu et al., 2025; Malach et al., 2025; Ma et al., 2025; Shen et al., 2026). DeciMamba (Ben-Kish et al., 2025) connects length degradation to the effective receptive field, LongMamba (Ye et al., 2025) analyzes receptive-field saturation and memory decay, MambaExtend (Azizi et al., 2025) calibrates out-of-distribution discretization, and Mamba Modulation (Lu et al., 2025) studies the relation between transition spectra and length generalization. These works mainly address one-dimensional sequence extension in LLMs, whereas visual resolution shifts alter two-dimensional sampling relationships and spatial propagation modes (Li et al., 2026; Yu & Park, 2026; Xiao et al., 2025). Therefore, cross-resolution generalization in visual Mamba goes *beyond linear scaling*.

## 3 Resolution-Induced Recurrent Dynamics Mismatch of Visual Mamba

Using VMamba as a representative backbone, we analyze how resolution-dependent changes in the spatial execution rate of selective recurrence induce local mismatch and propagate through residual and hierarchical trajectories.

### 3.1 Preliminaries

Mamba models a sequence with input-dependent selective state-space dynamics. With the learned step bias $b_\Delta$, the effective discretization step and recurrent update are

$$\bar{\Delta}_t = \mathrm{softplus}(\Delta_t + b_\Delta), \qquad h_t = \exp(\bar{\Delta}_t A) \odot h_{t-1} + \bar{\Delta}_t B_t z_t, \qquad y_t = C_t h_t + D z_t. \quad (1)$$

Here, $A$ governs state propagation, while the input-dependent $B_t$ and $C_t$ control state injection and readout. This recurrence is executed once per scan step, linking the recurrent dynamics to the spatial distance represented by each step.

VMamba extends selective scanning to 2D features through SS2D, which has been a representative visual Mamba paradigm. A VSS Block contains two consecutive additive updates,

$$\widetilde{X}_{s,b} = X_{s,b} + U^{\mathrm{S}}_{s,b}, \qquad X_{s,b+1} = \widetilde{X}_{s,b} + U^{\mathrm{M}}_{s,b}, \quad (2)$$

where $U^{\mathrm{S}}$ and $U^{\mathrm{M}}$ denote the SS2D and MLP updates, respectively. These blocks are organized hierarchically across spatial resolutions, combining spatial scanning with additive propagation.

### 3.2 Resolution-Induced Dynamics Mismatch

**Resolution changes the spatial rate of selective recurrence.** In VMamba, changing the input resolution preserves the CrossScan topology but changes the feature lattice on which the same learned selective recurrence is executed. Consequently, the same image-space span is traversed by more recurrent steps at higher resolution. Figure 2 summarizes this cross-resolution propagation. For stage $s$, the corresponding horizontal and vertical geometry ratios are

$$\kappa^w_s = \frac{W^0_s}{W^R_s}, \qquad \kappa^h_s = \frac{H^0_s}{H^R_s}. \quad (3)$$

These ratios describe the nominal lattice-spacing change along each principal scan axis, rather than the Euclidean displacement of every rasterized transition. Let $\bar{\Delta}_{r,t}$ denote the post-softplus effective step. For route $r$, jointly transporting

$$A^R_r = \kappa^r_s A_r, \qquad B^R_{r,t} = \kappa^r_s B_{r,t} \quad (4)$$

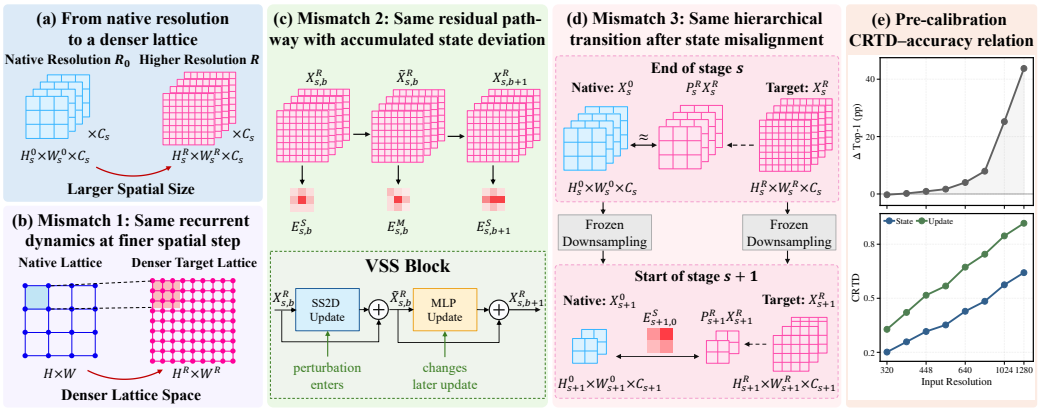



Figure 2: **Cross-resolution behavior of pretrained visual Mamba.** Resolution-induced recurrent mismatch propagates through residual and hierarchical pathways, while pre-calibration CRTD increases monotonically with accuracy degradation.

gives

$$\exp(\bar{\Delta}_{r,t}A_r^R) = \exp((\kappa_s^r\bar{\Delta}_{r,t})A_r), \qquad \bar{\Delta}_{r,t}B_{r,t}^R z_{r,t} = (\kappa_s^r\bar{\Delta}_{r,t})B_{r,t}z_{r,t}.$$

Thus, without transport, the pretrained selective dynamics retain their native discretization scale on a finer spatial lattice, changing the recurrent evolution per unit image-space distance. The geometry-derived factor $\kappa_{s,r}$ provides the corresponding first-order correction of this recurrent scale. The input-dependent quantities of the pretrained model, however, are not structurally constrained to provide this compensation automatically.

**Dynamics mismatch propagates through the residual trajectory.** Correcting this first-order recurrent scale is not sufficient, because each subsequent SS2D or MLP update is evaluated from the residual state produced by all preceding updates. Within a fixed stage $s$, let $P_s^R$ denote a fixed parameter-free restriction from the target-resolution lattice to the corresponding pretraining-resolution lattice. We define the pre-SS2D and pre-MLP state discrepancies as $E_{s,b}^{\mathrm{S}} = P_s^R X_{s,b}^R - X_{s,b}^0$ and $E_{s,b}^{\mathrm{M}} = P_s^R X_{s,b}^{e,R} - X_{s,b}^{e,0}$, respectively, and the additive-update discrepancy as $D_{s,b}^q = P_s^R U_{s,b}^{R,q} - U_{s,b}^{0,q}$ for $q \in \{\mathrm{S},\mathrm{M}\}$. By linearity, the actual VSS Block ordering directly gives

$$E_{s,b}^{\mathrm{M}} = E_{s,b}^{\mathrm{S}} + D_{s,b}^{\mathrm{S}}, \qquad E_{s,b+1}^{\mathrm{S}} = E_{s,b}^{\mathrm{M}} + D_{s,b}^{\mathrm{M}}. \tag{5}$$

Because each update discrepancy is evaluated from an already shifted state, a local perturbation of the selective dynamics changes the realized SS2D update and the states seen by subsequent branches. The initial recurrent-dynamics mismatch is therefore propagated and reshaped into a state-dependent residual-trajectory drift. Consequently, the first-order geometry transport alone does not guarantee trajectory or task-level recovery, motivating explicit correction of the remaining state- and update-level discrepancies.

**Hierarchy can reintroduce discrepancy at stage entry.** VMamba changes spatial representation across stages through down-sampling. Even if the native- and target-resolution trajectories are aligned at the end of one stage, the target-to-native restriction need not commute with the pretrained down-sampling transition. Within-stage alignment therefore does not guarantee alignment at the next stage entry, motivating a new state correction after each hierarchy transition.

These observations suggest a structured propagation mechanism rather than a purely feature-level mismatch: resolution changes first induce a first-order recurrent-dynamics mismatch, which is then propagated through state-dependent residual updates and hierarchical transitions. This analysis motivates correcting both the recurrent-scale mismatch and the trajectory along which the remaining discrepancies propagate. A local counterfactual analysis in Appendix A.6 further isolates this geometry-induced rate change, showing that the analytic SDT transport removes most of the corresponding local transition- and injection-rate discrepancy.

### 3.3 New Metrics: Cross-Resolution Trajectory Discrepancy

To quantify resolution-induced drift along the recurrent residual trajectory, we introduce Cross-Resolution Trajectory Discrepancy (CRTD) with two probes: CRTD-State measures drift in propagated residual states entering the SS2D and MLP branches, while CRTD-Update measures drift in their corresponding additive updates.

For an image $x$, let $Z^R_{x,i}$ and $Z^0_{x,i}$ denote the matched internal representations at trajectory site $i$ under a target resolution $R$ and the pretraining resolution $R^0$, respectively. For $R > R^0$, we define

$$\mathrm{CRTD}_x(R; Z) = \frac{1}{N_Z} \sum_{i=1}^{N_Z} d\big(\mathcal{P}_i^{R\to 0}\big(Z^R_{x,i}\big), Z^0_{x,i}\big), \tag{6}$$

where $\mathcal{P}_i^{R\to 0}$ is a parameter-free area-overlap restriction to the matched pretraining-resolution grid, and $d(\cdot,\cdot)$ equally combines normalized MSE and mean spatial cosine distance. Instantiating $Z$ with propagated states or the corresponding SS2D/MLP updates yields CRTD-State or CRTD-Update, respectively. Lower values indicate smaller cross-resolution drift, and both are set to zero at $R = R_0$ by convention. Figure 2(e) shows that, before calibration, both CRTD-State and CRTD-Update increase monotonically with cross-resolution accuracy degradation in the frozen pretrained VMamba-T. Exact trajectory sites and discrepancy definitions are provided in Appendix A.2.1.

## 4 Architecture-Native Uniform Residual/Update Calibration

Guided by the analysis above, we propose Architecture-Native Uniform Residual/Update Calibration (Uniform-RU), which applies a shared calibration principle through each backbone's native recurrent and residual topology. Spatial Dynamics Transport (SDT) compensates for the nominal lattice-spacing change of the native recurrence, while Carried-State Correction (CSC) and Additive-Update Correction (UC) align carried states and additive updates at their corresponding residual sites. The pretrained backbone remains frozen throughout. An overview is depicted in Figure 3.

### 4.1 Architecture-Native Correction

**Spatial Dynamics Transport.** For the four Cross2D routes in stage $s$, we assign the geometry-derived scales

$$k_r \in \{\kappa_s^w, \kappa_s^h, \kappa_s^w, \kappa_s^h\},$$

for horizontal-forward, vertical-forward, horizontal-reverse, and vertical-reverse scans, respectively, and modify the state-space parameters at the selective-scan boundary as

$$A_r^R = k_r A_r, \qquad B_r^R = k_r B_r. \tag{7}$$

Substituting equation 7 into equation 1 yields state propagation $\exp((k_r\bar{\Delta}_{r,t})A_r)$ and input injection $(k_r\bar{\Delta}_{r,t})B_{r,t}z_{r,t}$. Thus, for the implemented selective recurrence, jointly scaling $A$ and $B$ is algebraically equivalent to rescaling the post-softplus effective step $\bar{\Delta}_{r,t} \mapsto k_r\bar{\Delta}_{r,t}$. SDT therefore compensates for the nominal route-axis lattice-spacing change encoded by equation 3, without learning resolution-specific scale parameters. Because it corrects only the selective recurrence while the remaining pretrained operators are unchanged, SDT is not a stand-alone task-level correction, the residual trajectory mismatch is addressed by CSC and UC.

**Carried-State and Additive-Update Corrections.** For each stage, fixed restriction and lift operators, $\mathcal{P}_s^R$ and $\mathcal{L}_s^R$, predict corrections in the native stage coordinate system and inject them back into the target grid.

At stage entry, a *Carried-State Correction* adjusts the residual state already carried into the stage:

$$X^R_{s,0} \leftarrow X^R_{s,0} + \mathcal{L}_s^R g_s^A\left(\mathcal{P}_s^R X^R_{s,0}\right). \tag{8}$$

CSC adjusts the starting point of each hierarchical trajectory rather than an individual branch.

Within a stage, for an additive branch $q \in \{\mathrm{S}, \mathrm{M}\}$, an *Additive-Update Correction* is predicted from both the current residual state and the branch increment:

$$U^q_{s,b} \leftarrow U^q_{s,b} + \mathcal{L}_s^R g_{s,b}^{U,q}\left(\mathcal{P}_s^R X,\, \mathcal{P}_s^R U^q_{s,b}\right). \tag{9}$$

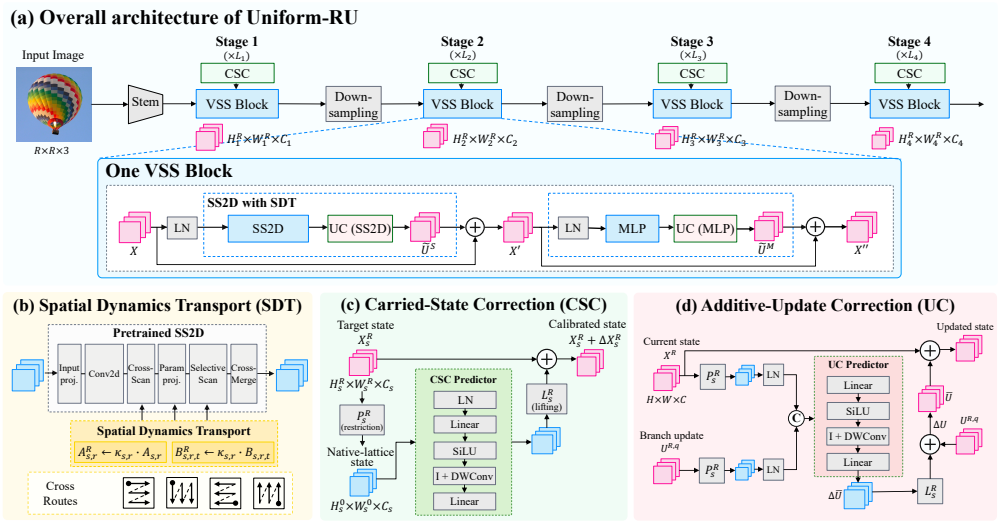



Figure 3: **Overview of Architecture-Native Uniform Residual/Update Calibration (Uniform-RU).** SDT, CSC, and UC operate on the pretrained backbone without altering its native architecture.

The two corrections address different trajectory terms: CSC modifies the state carried into the current stage, whereas UC modifies the update about to be added to the residual stream. In VMamba, the same UC principle is applied to both SS2D and MLP additive branches, which jointly form a single top-level UC component.

All correction predictors are lightweight spatial bottlenecks with zero-initialized output projections, so the model starts exactly from the deterministic SDT configuration. Predictor architectures and parameterization details are provided in Appendix A.1.

### 4.2 Label-Free Calibration and Inference

Uniform-RU is calibrated on unlabeled images with the pretrained backbone frozen. For each image, native- and target-resolution views pass through the same frozen model, with the native-resolution pass providing a fixed reference for trajectory and prediction consistency. Only the correction parameters are optimized.

The native-resolution reference is used only during calibration. At inference, Uniform-RU uses a single target-resolution input, with all corrections predicted from its current trajectory and no auxiliary resolution branch. The calibration protocol is shared across architectures, while the internal consistency terms follow each backbone's native residual topology.

## 5 Experiments

### 5.1 Image Classification

We first evaluate the cross-resolution classification performance of Uniform-RU on ImageNet-1K (Deng et al., 2009). All experiments start from officially pretrained models and directly deploy them at target resolutions different from the pretraining scale. The original backbone remains frozen throughout calibration. SDT deterministically transports architecture-native recurrent operators according to the target spatial scale, while only CSC and UC introduce lightweight learnable corrections. The entire calibration process uses unlabeled training images only. Full implementation and training details are provided in Appendix A.3.

As shown in Table 1, directly changing the input resolution causes substantial performance degradation across different visual Mamba architectures and model scales, whereas Uniform-RU generally recovers the lost accuracy, with particularly clear gains under larger resolution shifts. The recovery is generally more pronounced under larger shifts. This observation is consistent with our mecha-

Table 1: ImageNet-1K Top-1 accuracy (%) under resolution extension. w/ RU denotes Uniform-RU; Params gives backbone or added calibration parameters. FLOPs are calculated with a $320^2$ input rather than $224^2$.

| Method | Params | FLOPs | Input Resolution | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $224^2$ | $320^2$ | $384^2$ | $448^2$ | $512^2$ | $640^2$ | $768^2$ | $1024^2$ | $1280^2$ |
| VMamba-B | 89M | 31.3G | 83.9 | 84.2 | 84.1 | 83.8 | 83.3 | 82.1 | 80.6 | 74.8 | 67.6 |
| VMamba-B w/ RU | +6.6M | 32.5G | 83.9 | 84.0 | 84.0 | 83.8 | 83.6 | 83.3 | 82.9 | 82.0 | 81.0 |
| VMamba-S | 50M | 17.8G | 83.6 | 84.0 | 83.8 | 83.5 | 82.9 | 81.9 | 80.3 | 73.7 | 64.7 |
| VMamba-S w/ RU | +3.7M | 18.4G | 83.6 | 83.8 | 83.6 | 83.5 | 83.3 | 82.8 | 82.4 | 81.4 | 80.3 |
| VMamba-T | 30M | 10.0G | 82.6 | 82.9 | 82.4 | 81.7 | 80.9 | 78.6 | 74.7 | 57.4 | 38.8 |
| VMamba-T w/ RU | +2.7M | 10.4G | 82.6 | 82.7 | 82.6 | 82.4 | 82.0 | 81.7 | 81.0 | 79.4 | 78.4 |
| ViM-B | 98M | 38.6G | 81.9 | 81.7 | 80.3 | 77.5 | 74.1 | 64.7 | 52.9 | 31.8 | 18.9 |
| ViM-B w/ RU | +8.0M | 39.4G | 81.9 | 81.8 | 80.6 | 80.8 | 80.2 | 76.9 | 74.9 | 63.6 | 53.3 |
| ViM-S | 26M | 10.8G | 80.3 | 79.4 | 76.7 | 72.7 | 66.9 | 51.8 | 36.0 | 15.1 | 5.9 |
| ViM-S w/ RU | +2.0M | 11.0G | 80.3 | 80.6 | 79.6 | 80.0 | 79.3 | 77.1 | 75.3 | 68.4 | 62.1 |
| ViM-T | 7M | 3.3G | 76.1 | 73.6 | 69.0 | 63.1 | 55.8 | 36.3 | 14.6 | 0.6 | 0.1 |
| ViM-T w/ RU | +0.5M | 3.3G | 76.1 | 75.0 | 73.7 | 73.7 | 72.1 | 68.2 | 63.9 | 52.2 | 43.2 |
| MambaVision-T | 32M | 9.1G | 82.3 | 80.3 | 79.8 | 79.9 | 77.6 | 74.8 | 71.0 | 59.7 | 46.4 |
| MambaVision-T w/ RU | +2.6M | 9.4G | 82.3 | 82.9 | 82.9 | 82.9 | 82.7 | 82.0 | 81.6 | 80.1 | 78.8 |
| MSVMamba-T | 33M | 9.4G | 82.8 | 83.2 | 82.9 | 82.2 | 81.3 | 79.5 | 76.7 | 66.2 | 54.9 |
| MSVMamba-T w/ RU | +2.8M | 9.8G | 82.8 | 82.5 | 82.3 | 82.0 | 81.7 | 80.9 | 80.1 | 77.9 | 75.7 |
| EfficientVMamba-B | 33M | 8.1G | 81.8 | 82.5 | 82.0 | 81.3 | 80.3 | 78.0 | 74.4 | 59.4 | 41.6 |
| EfficientVMamba-B w/ RU | +1.5M | 8.4G | 81.8 | 82.2 | 81.9 | 81.3 | 81.0 | 79.8 | 78.4 | 75.2 | 71.8 |
| SpatialMamba-T | 27M | 9.0G | 83.5 | 83.3 | 83.0 | 81.6 | 80.2 | 77.4 | 74.4 | 65.7 | 53.7 |
| SpatialMamba-T w/ RU | +1.8M | 9.2G | 83.5 | 83.6 | 83.5 | 83.2 | 82.8 | 81.6 | 80.8 | 77.5 | 75.1 |
| VSSD-Micro | 14M | 4.7G | 82.5 | 82.5 | 81.1 | 78.0 | 72.0 | 48.1 | 20.7 | 2.9 | 0.8 |
| VSSD-Micro w/ RU | +1.0M | 4.8G | 82.5 | 82.8 | 82.8 | 82.6 | 82.3 | 81.4 | 80.5 | 78.5 | 76.3 |

nism analysis: resolution changes not only alter the spatial sampling density, but also perturb the recurrent dynamics established at the pretraining scale and the subsequent state propagation. SDT first restores dynamics that are compatible with the target spatial scale, while CSC and UC further correct the residual discrepancies in the carried state and additive update, respectively. They allow Uniform-RU to mitigate cross-resolution degradation without retraining the backbone, while retaining single-pass inference at the target resolution.

### 5.2 Semantic Segmentation

To evaluate the generalization of Uniform-RU to dense prediction, we conduct semantic segmentation experiments on ADE20K dataset (Zhou et al., 2019) using the UperNet (Xiao et al., 2018) framework while keeping the original task model frozen. As shown in Table 2, directly changing the test resolution leads to clear performance degradation, whereas Uniform-RU consistently recovers the lost accuracy across resolutions and remains close to the native operating point even under large resolution shifts. For example, VMamba-S achieves 50.6 mIoU at its native resolution $512^2$, while Uniform-RU reaches 50.8 mIoU at $896^2$, slightly surpassing the native-resolution performance. This observation suggests that higher resolution does not inherently lead to worse performance: once the internal representation mismatch induced by resolution changes is mitigated, the additional spatial information can still be converted into useful prediction gains. As an optional extension, we also provide Margin-Aware Prediction Selection (MAPS) to retain the higher-margin prediction from Raw and Uniform-RU at each pixel, further improving performance in some settings.

### 5.3 Object Detection and Instance Segmentation

To further evaluate the generalization of Uniform-RU to object-level structured prediction, we conduct object detection and instance segmentation experiments on the MS COCO 2017 dataset (Lin et al., 2015) while keeping the Mask R-CNN (He et al., 2018) with FPN task model frozen. As shown in Table 3, Uniform-RU consistently recovers the losses in both box AP and mask AP caused

Table 2: ADE20K semantic segmentation under resolution extension using UPerNet. w/ RU denotes Uniform-RU and MAPS its optional inference extension. FLOPs are calculated with a $640 \times 2560$ input.

| Method | Params | FLOPs | Input Resolution | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | $512^2$ | $640^2$ | $768^2$ | $896^2$ | $1024^2$ | $1280^2$ |
| VMamba-B | 122M | 1829G | 51.0 | 51.2 | 50.7 | 51.0 | 50.6 | 49.6 |
| VMamba-B w/ RU | +6.6M | 2027G (1.1×) | – | 50.9 | 50.6 | 50.5 | 50.3 | 49.9 |
| + MAPS | +0M | 3856G (2.1×) | – | 51.2 | 50.9 | 51.3 | 51.1 | 50.7 |
| VMamba-S | 82M | 1607G | 50.6 | 50.7 | 50.3 | 50.8 | 50.4 | 49.8 |
| VMamba-S w/ RU | +3.7M | 1752G (1.1×) | – | 50.7 | 50.5 | 50.8 | 50.5 | 50.3 |
| + MAPS | +0M | 3359G (2.1×) | – | 50.9 | 50.7 | 51.3 | 51.1 | 50.9 |
| VMamba-T | 62M | 1482G | 47.9 | 48.1 | 47.5 | 47.8 | 46.8 | 45.4 |
| VMamba-T w/ RU | +2.7M | 1608G (1.1×) | – | 48.1 | 48.0 | 48.3 | 47.4 | 46.9 |
| + MAPS | +0M | 3091G (2.1×) | – | 48.3 | 48.0 | 48.4 | 47.7 | 47.2 |
| MambaVision-T | 55M | 1547G | 46.0 | 44.8 | 40.5 | 38.6 | 36.8 | 33.5 |
| MambaVision-T w/ RU | +2.6M | 1575G (1.0×) | – | 46.3 | 45.5 | 46.0 | 45.2 | 44.9 |
| + MAPS | +0M | 3122G (2.0×) | – | 46.1 | 44.8 | 45.0 | 44.4 | 43.5 |
| MSVMamba-Micro | 42M | 1367G | 45.1 | 45.1 | 45.0 | 43.8 | 42.7 | 40.0 |
| MSVMamba-Micro w/ RU | +1.0M | 1422G (1.0×) | – | 44.8 | 45.0 | 44.8 | 44.1 | 43.5 |
| + MAPS | +0M | 2789G (2.0×) | – | 45.1 | 45.2 | 44.8 | 44.2 | 43.0 |
| SpatialMamba-T | 65M | 1461G | 48.6 | 48.4 | 47.9 | 48.1 | 47.0 | 44.0 |
| SpatialMamba-T w/ RU | +1.8M | 1558G (1.1×) | – | 48.4 | 47.9 | 48.2 | 47.7 | 46.6 |
| + MAPS | +0M | 3019G (2.1×) | – | 48.6 | 48.4 | 48.8 | 48.2 | 47.0 |
| VSSD-Micro | 42M | 1398G | 45.4 | 43.4 | 40.3 | 36.2 | 28.7 | 14.6 |
| VSSD-Micro w/ RU | +1.0M | 1460G (1.0×) | – | 45.6 | 45.1 | 44.9 | 44.1 | 42.7 |
| + MAPS | +0M | 2858G (2.0×) | – | 44.6 | 43.3 | 42.4 | 40.3 | 37.4 |

by resolution changes, while maintaining clear gains under larger resolution shifts. For example, VMamba-S achieves $48.7/43.7$ $\mathrm{AP}^b/\mathrm{AP}^m$ at its native $R800$ setting, whereas Uniform-RU reaches $49.2/44.2$ at $R1024$, surpassing the native-resolution performance on both detection and instance segmentation. Compared with image classification, these results further indicate that the correction introduced by Uniform-RU is not confined to the final prediction scores, but remains effective throughout the structured prediction pipeline, including proposal generation, object classification, localization, and mask prediction. As an optional extension, Proposal-Aligned Logit Fusion (PALF) fuses Raw and Uniform-RU classification logits over shared proposals, further improving performance in most settings and sometimes surpassing the native-resolution baseline.

### 5.4 Ablation Studies

We ablate the three components of Uniform-RU on ViM-T and VMamba-T under identical training and evaluation settings. As shown in Table 4, SDT alone is insufficient to resolve cross-resolution degradation. This is expected because SDT corrects only the local first-order rate mismatch rather than the full residual trajectory, consistent with its role as a local recurrent-scale transport rather than a stand-alone task-level correction. CSC and UC are also complementary, as their combination remains effective even without SDT.

The full model achieves the lowest CRTD-State and CRTD-Update across both backbones and resolutions, supporting the complementary roles of deterministic spatial-dynamics transport and learned trajectory correction. We further evaluate a matched output-only control with identical parameterization and training in Appendix A.3.4, using only native-to-target prediction consistency without internal trajectory supervision. Uniform-RU yields lower CRTD and stronger internal alignment, while the corresponding task-level diagnostic shows consistent gains across resolutions.

Table 3: COCO object detection and instance segmentation under resolution extension using Mask R-CNN. w/ RU denotes Uniform-RU and PALF its optional inference extension. Results are $AP^b/AP^m$. FLOPs are calculated with a $1638 \times 1024$ input.

| Method | Params | FLOPs | Short-side Size | | | |
|---|---|---|---|---|---|---|
| | | | 800 | 1024 | 1280 | 1600 |
| VMamba-B | 108M | 756G | 49.2 / 44.1 | 49.6 / 44.4 | 48.5 / 43.5 | 44.6 / 40.6 |
| VMamba-B w/ RU | +6.6M | 925G (1.2×) | – | 49.4 / 44.4 | 48.3 / 43.7 | 46.6 / 42.7 |
| + PALF | +0M | 1544G (2.0×) | – | 49.5 / 44.6 | 48.8 / 44.0 | 47.6 / 43.2 |
| VMamba-S | 70M | 531G | 48.7 / 43.7 | 49.3 / 44.1 | 48.4 / 43.4 | 44.7 / 40.4 |
| VMamba-S w/ RU | +3.7M | 653G (1.2×) | – | 49.2 / 44.2 | 48.1 / 43.4 | 46.1 / 42.2 |
| + PALF | +0M | 1048G (2.0×) | – | 49.3 / 44.2 | 48.6 / 43.6 | 47.0 / 42.6 |
| VMamba-T | 50M | 402G | 47.3 / 42.7 | 47.5 / 42.9 | 46.4 / 42.0 | 42.7 / 38.9 |
| VMamba-T w/ RU | +2.7M | 509G (1.3×) | – | 47.6 / 43.1 | 46.4 / 42.3 | 44.3 / 40.2 |
| + PALF | +0M | 775G (1.9×) | – | 47.6 / 43.0 | 46.9 / 42.5 | 45.4 / 41.2 |
| MSVMamba-M | 32M | 287G | 43.8 / 39.9 | 44.4 / 40.5 | 42.6 / 39.2 | 38.4 / 35.5 |
| MSVMamba-M w/ RU | +1.0M | 333G (1.2×) | – | 44.2 / 40.3 | 42.9 / 39.6 | 40.4 / 37.6 |
| + PALF | +0M | 484G (1.7×) | – | 44.4 / 40.5 | 43.2 / 39.7 | 40.9 / 37.9 |
| SpatialMamba-T | 46M | 394G | 47.6 / 42.9 | 47.5 / 42.9 | 45.4 / 41.2 | 40.7 / 37.4 |
| SpatialMamba-T w/ RU | +1.8M | 466G (1.2×) | – | 47.6 / 43.0 | 45.8 / 42.0 | 43.8 / 40.3 |
| + PALF | +0M | 724G (1.8×) | – | 47.7 / 43.1 | 46.1 / 41.9 | 44.2 / 40.3 |
| VSSD-Micro | 33M | 321G | 45.4 / 41.3 | 45.2 / 41.1 | 41.8 / 38.4 | 32.7 / 30.2 |
| VSSD-Micro w/ RU | +1.0M | 373G (1.2×) | – | 45.2 / 41.2 | 42.8 / 39.3 | 37.1 / 34.6 |
| + PALF | +0M | 557G (1.7×) | – | 45.3 / 41.2 | 42.5 / 39.0 | 35.9 / 32.6 |

Table 4: Component ablation of SDT, CSC, and UC on ViM-T and VMamba-T across three target resolutions. Higher Top-1 is better, while lower CRTD-State and CRTD-Update are better.

| Backbone | SDT | CSC | UC | Top-1 | | | CRTD-State | | | CRTD-Update | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 448 | 640 | 1024 | 448 | 640 | 1024 | 448 | 640 | 1024 |
| ViM-T | × | × | × | 63.1 | 36.4 | 0.7 | 0.555 | 0.780 | 1.300 | 0.739 | 0.995 | 2.026 |
| | ✓ | × | × | 67.3 | 43.7 | 9.7 | 0.458 | 0.636 | 0.836 | 0.577 | 0.780 | 0.978 |
| | ✓ | ✓ | × | 73.4 | 66.3 | 47.4 | 0.394 | 0.520 | 0.691 | 0.488 | 0.628 | 0.813 |
| | ✓ | × | ✓ | 73.5 | 67.2 | 49.0 | 0.287 | 0.391 | 0.665 | 0.421 | 0.546 | 0.776 |
| | × | ✓ | ✓ | 72.4 | 63.9 | 38.2 | 0.373 | 0.540 | 0.768 | 0.530 | 0.700 | 0.899 |
| | ✓ | ✓ | ✓ | **73.7** | **68.2** | **52.2** | **0.274** | **0.375** | **0.588** | **0.403** | **0.524** | **0.706** |
| VMamba-T | × | × | × | 81.7 | 78.6 | 57.4 | 0.315 | 0.427 | 0.574 | 0.516 | 0.672 | 0.845 |
| | ✓ | × | × | 80.3 | 75.1 | 47.4 | 0.314 | 0.433 | 0.591 | 0.513 | 0.674 | 0.854 |
| | ✓ | ✓ | × | 82.1 | 80.6 | 75.8 | 0.225 | 0.319 | 0.393 | 0.379 | 0.513 | 0.617 |
| | ✓ | × | ✓ | 82.3 | 81.6 | 79.3 | 0.167 | 0.205 | 0.253 | 0.297 | 0.365 | 0.442 |
| | × | ✓ | ✓ | 82.4 | 81.7 | 79.2 | 0.160 | 0.202 | 0.254 | 0.280 | 0.350 | 0.436 |
| | ✓ | ✓ | ✓ | **82.4** | **81.7** | **79.4** | **0.156** | **0.195** | **0.239** | **0.275** | **0.345** | **0.418** |

## 6 Conclusion

This paper reveals that resolution scaling in visual Mamba models changes the spatial rate of pretrained selective dynamics, inducing a recurrent-dynamics mismatch that propagates through residual updates and hierarchical transitions. We propose CRTD and Uniform-RU to respectively expose this drift along internal trajectories and restore cross-resolution performance without modifying the pretrained backbone. Across heterogeneous architectures, model scales, resolutions, and task settings, the pre-calibration association between trajectory discrepancy and accuracy degradation, together with the consistent reduction in trajectory discrepancy and performance recovery after calibration supports both the identified mechanism and the generality of Uniform-RU. We hope our findings will provide a dynamics-centric perspective for resolution scaling in visual Mamba.

### AI USE STATEMENT

Generative AI tools were used as assistive tools for language editing and refinement, drafting limited parts of the manuscript, brainstorming and discussing research ideas, and providing limited support for code development and experimental troubleshooting. All research questions, methodological choices, experimental settings, result interpretation, and final claims were determined and verified by the authors. All reported experimental results were obtained from the authors' own implementations and experiments, and all AI-assisted content was reviewed before inclusion in the paper.

### ETHICS STATEMENT

This work studies cross-resolution generalization in Visual Mamba models and does not involve human subjects, personal or sensitive information, or high-risk applications. All experiments use publicly available datasets and pretrained models under their corresponding research licenses and usage terms. We do not identify ethical concerns beyond those typically associated with general computer vision research.

### REPRODUCIBILITY STATEMENT

We provide model configurations, calibration protocols, data splits, training and inference settings, and major hyperparameters in the main paper and appendix. Architecture-specific implementation details and additional results are also documented. An anonymized code repository containing the implementation, experiment scripts, and configurations is also provided.

## REFERENCES

Seyedarmin Azizi, Souvik Kundu, Mohammad Erfan Sadeghi, and Massoud Pedram. Mambaextend: A training-free approach to improve long context extension of mamba. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=LgzRo1RpLS.

Assaf Ben-Kish, Itamar Zimerman, Shady Abu-Hussein, Nadav Cohen, Amir Globerson, Lior Wolf, and Raja Giryes. Decimamba: Exploring the length extrapolation potential of mamba. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=iWSl5Zyjjw.

Lucas Beyer, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornblith, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim Alabdulmohsin, and Filip Pavetic. Flexivit: One model for all patch sizes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 14496–14506. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01393. URL https://doi.org/10.1109/CVPR52729.2023.01393.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10041–10071. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/dao24a.html.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Peter Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim Alabdulmohsin, Avital Oliver, Piotr Padlewski, Alexey A. Gritsenko, Mario Lucic, and Neil Houlsby. Patch n' pack: Navit, a vision transformer for any aspect ratio and resolution. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=VpGFHmI7e5.

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 Computer Society Conference on Computer Vision and*

*Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pp. 248–255. IEEE Computer Society, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://doi.org/10.1109/CVPR.2009.5206848.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Qihang Fan, Quanzeng You, Xiaotian Han, Yongfei Liu, Yunzhe Tao, Huaibo Huang, Ran He, and Hongxia Yang. Vitar: Vision transformer with any resolution. *CoRR*, abs/2403.18361, 2024. doi: 10.48550/ARXIV.2403.18361. URL https://doi.org/10.48550/arXiv.2403.18361.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2024. URL https://arxiv.org/abs/2312.00752.

Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Re. Hippo: Recurrent memory with optimal polynomial projections, 2020. URL https://arxiv.org/abs/2008.07669.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=uYLFoz1vlAC.

Ali Hatamizadeh and Jan Kautz. Mambavision: A hybrid mamba-transformer vision backbone. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 25261–25270. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02352. URL https://openaccess.thecvf.com/content/CVPR2025/html/Hatamizadeh_MambaVision_A_Hybrid_Mamba-Transformer_Vision_Backbone_CVPR_2025_paper.html.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2018. URL https://arxiv.org/abs/1703.06870.

Tao Huang, Xiaohuan Pei, Shan You, Fei Wang, Chen Qian, and Chang Xu. Localmamba: Visual state space model with windowed selective scan. In Alessio Del Bue, Cristian Canton, Jordi Pont-Tuset, and Tatiana Tommasi (eds.), *Computer Vision - ECCV 2024 Workshops - Milan, Italy, September 29-October 4, 2024, Proceedings, Part XI*, volume 15633 of *Lecture Notes in Computer Science*, pp. 12–22. Springer, 2024. doi: 10.1007/978-3-031-91979-4_2. URL https://doi.org/10.1007/978-3-031-91979-4_2.

Bo Li, Haoke Xiao, and Lv Tang. Fractalmamba++: Scaling vision mamba across resolutions via hilbert fractal geometry, 2026. URL https://arxiv.org/abs/2505.14062.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

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 Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/baa2da9ae4bfed26520bb61d259a3653-Abstract-Conference.html.

Peng Lu, Jerry Huang, Qiuhao Zeng, Xinyu Wang, Boxing Chen, Philippe Langlais, and Yufei Cui. Mamba modulation: On the length generalization of mamba models. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu,

Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/1bfc9f74afa91b9b8add5a97a97001a1-Abstract-Conference.html.

Jun-Yu Ma, Tianqing Fang, Zhisong Zhang, Hongming Zhang, Haitao Mi, and Dong Yu. Recall with reasoning: Chain-of-thought distillation for mamba's long-context memory and extrapolation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 4714–4720. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.235. URL https://doi.org/10.18653/v1/2025.emnlp-main.235.

Eran Malach, Omid Saremi, Sinead Williamson, Arwen Bradley, Aryo Lotfi, Emmanuel Abbe, Josh Susskind, and Etai Littwin. To infinity and beyond: Tool-use unlocks length generalization in state space models, 2025. URL https://arxiv.org/abs/2510.14826.

Eric Nguyen, Karan Goel, Albert Gu, Gordon W. Downs, Preey Shah, Tri Dao, Stephen A. Baccus, and Christopher Ré. S4ND: modeling images and videos as multidimensional signals using state spaces. *CoRR*, abs/2210.06583, 2022. doi: 10.48550/ARXIV.2210.06583. URL https://doi.org/10.48550/arXiv.2210.06583.

Xiaohuan Pei, Tao Huang, and Chang Xu. Efficientvmamba: Atrous selective scan for light weight visual mamba. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 6443–6451. AAAI Press, 2025. doi: 10.1609/AAAI.V39I6.32690. URL https://doi.org/10.1609/aaai.v39i6.32690.

Mansi Sakarvadia, Kareem Hegazy, Amin Totounferoush, Kyle Chard, Yaoqing Yang, Ian Foster, and Michael W. Mahoney. The false promise of zero-shot super-resolution in machine-learned operators, 2026. URL https://arxiv.org/abs/2510.06646.

Haochen Shen, Davis Wertheimer, Zheng Wang, Garrett Goon, Derrick Liu, Naigang Wang, Mudhakar Srivatsa, Raghu K. Ganti, and Minjia Zhang. From collapse to control: Understanding and extending context length in emerging hybrid models via universal position interpolation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=MjmORKLHUI.

Yuheng Shi, Minjing Dong, and Chang Xu. Multi-scale vmamba: Hierarchy in hierarchy visual state space model. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/2d69e771d9f274f7c624198ea74f5b98-Abstract-Conference.html.

Yuheng Shi, Mingjia Li, Minjing Dong, and Chang Xu. VSSD: vision mamba with non-causal state space duality. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 10819–10829. IEEE, 2025. doi: 10.1109/ICCV51701.2025.01007. URL https://doi.org/10.1109/ICCV51701.2025.01007.

Lv Tang, Haoke Xiao, Peng-Tao Jiang, Hao Zhang, Jinwei Chen, and Bo Li. Scalable visual state space model with fractal scanning. *CoRR*, abs/2405.14480, 2024. doi: 10.48550/ARXIV.2405.14480. URL https://doi.org/10.48550/arXiv.2405.14480.

Rui Tian, Zuxuan Wu, Qi Dai, Han Hu, Yu Qiao, and Yu-Gang Jiang. Resformer: Scaling vits with multi-resolution training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 22721–22731. IEEE, 2023. doi: 10.1109/CVPR52729.2023.02176. URL https://doi.org/10.1109/CVPR52729.2023.02176.

Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. Fixing the train-test resolution discrepancy. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 8250–8260, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/d03a857a23b5285736c4d55e0bb067c8-Abstract.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Wenshuo Wang and Fan Zhang. Breaking scale anchoring: Frequency representation learning for accurate high-resolution inference from low-resolution training, 2026. URL https://arxiv.org/abs/2512.05132.

Chaodong Xiao, Minghan Li, Zhengqiang Zhang, Deyu Meng, and Lei Zhang. Spatial-mamba: Effective visual state space models via structure-aware state fusion. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=iDe1mtxqK5.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part V*, volume 11209 of *Lecture Notes in Computer Science*, pp. 432–448. Springer, 2018. doi: 10.1007/978-3-030-01228-1_26. URL https://doi.org/10.1007/978-3-030-01228-1_26.

Chenhongyi Yang, Zehui Chen, Miguel Espinosa, Linus Ericsson, Zhenyu Wang, Jiaming Liu, and Elliot J. Crowley. Plainmamba: Improving non-hierarchical mamba in visual recognition. In *35th British Machine Vision Conference, BMVC 2024, Glasgow, UK, November 25-28, 2024*. BMVA Press, 2024. URL https://bmvc2024.org/proceedings/755/.

Zhifan Ye, Kejing Xia, Yonggan Fu, Xin Dong, Jihoon Hong, Xiangchi Yuan, Shizhe Diao, Jan Kautz, Pavlo Molchanov, and Yingyan Celine Lin. Longmamba: Enhancing mamba's long-context capabilities via training-free receptive field enlargement. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=fMbLszVO1H.

Jin Yu and Juyoun Park. Norm or direction? decoding vision mambas for high-resolution vision. *CoRR*, abs/2607.18625, 2026. doi: 10.48550/ARXIV.2607.18625. URL https://doi.org/10.48550/arXiv.2607.18625.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 12360–12371, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/1e8a19426224ca89e83cef47f1e7f53b-Abstract.html.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20K dataset. *Int. J. Comput. Vis.*, 127(3): 302–321, 2019. doi: 10.1007/S11263-018-1140-0. URL https://doi.org/10.1007/s11263-018-1140-0.

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 Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62429–62442. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/zhu24f.html.

# A Appendix

## A.1 Detailed Formulation and Derivations

This section complements Sec. 4 with the detailed formulation of Spatial Dynamics Transport (SDT), Carried-State Correction (CSC), and Additive-Update Correction (UC). We focus on effective-step transport, native-lattice alignment, and closed-loop residual propagation.

**Spatial Dynamics Transport (SDT).** Following equation 1, let $\bar{\Delta}_{r,t}$ denote the post-softplus effective step of route $r$. For stage $s$, whose native and target lattices are $H_s^0 \times W_s^0$ and $H_s^R \times W_s^R$, respectively, we define

$$\kappa_s^w = \frac{W_s^0}{W_s^R}, \qquad \kappa_s^h = \frac{H_s^0}{H_s^R}, \qquad \boldsymbol{\kappa}_s = [\kappa_s^w,\, \kappa_s^h,\, \kappa_s^w,\, \kappa_s^h]. \tag{10}$$

The four entries correspond to the horizontal-forward, vertical-forward, horizontal-reverse, and vertical-reverse Cross2D routes.

Rather than modifying the pretrained $\Delta$ predictor, we apply the route-wise transport directly at the selective-scan boundary:

$$A_{s,r}^R = \kappa_{s,r} A_{s,r}, \qquad B_{s,r,t}^R = \kappa_{s,r} B_{s,r,t}. \tag{11}$$

The step-dependent terms of the selective recurrence then satisfy

$$\begin{aligned} \exp\big(\bar{\Delta}_{r,t} A_{s,r}^R\big) &= \exp\big((\kappa_{s,r}\bar{\Delta}_{r,t}) A_{s,r}\big)\,, \\ \bar{\Delta}_{r,t} B_{s,r,t}^R z_{r,t} &= (\kappa_{s,r}\bar{\Delta}_{r,t}) B_{s,r,t} z_{r,t}. \end{aligned} \tag{12}$$

Thus, for the implemented selective-state recurrence, jointly transporting $A$ and $B$ is algebraically equivalent to rescaling the effective step as $\bar{\Delta}_{r,t} \mapsto \kappa_{s,r}\bar{\Delta}_{r,t}$. This equivalence concerns the post-softplus effective step and does not assume that scaling commutes with the softplus parameterization.

The factors in equation 10 describe the nominal change in lattice spacing along the corresponding spatial axis. They are not interpreted as the exact Euclidean displacement of every transition in a rasterized scan, which also contains row- or column-wrap transitions. Accordingly, SDT compensates for this deterministic nominal spatial-step scale change rather than imposing exact resolution invariance on the full nonlinear network. Only $A$ and $B$ are transported, the local convolution and all remaining pretrained operators stay unchanged.

**Native-lattice restriction and lifting.** Uniform-RU predicts corrections on the native stage lattice. Consider one spatial axis with native length $n$ and target length $m \geq n$. Let

$$I_j = \Big[j\frac{m}{n}, (j+1)\frac{m}{n}\Big], \qquad J_i = [i, i+1].$$

We define the restriction matrix $P \in \mathbb{R}^{n\times m}$ by normalized cell overlap,

$$P_{ji} = \frac{|I_j \cap J_i|}{m/n}, \tag{13}$$

and use its minimum-norm right inverse

$$L = P^\top (PP^\top)^{-1}. \tag{14}$$

For the lattice pairs used in our experiments, $P$ has full row rank and therefore

$$PL = I. \tag{15}$$

Applying these mappings independently along height and width yields $\mathcal{P}_{s,R}$ and $\mathcal{L}_{s,R}$, satisfying $\mathcal{P}_{s,R}\mathcal{L}_{s,R} = I$ up to numerical precision.

This identity gives a useful exact-correction construction. Let $Q^0$ and $Q^R$ denote the corresponding quantities produced at the native and target resolutions. Define

$$c^\star = Q^0 - \mathcal{P}_{s,R} Q^R, \qquad Q^{R,+} = Q^R + \mathcal{L}_{s,R} c^\star. \tag{16}$$

Then

$$\mathcal{P}_{s,R} Q^{R,+} = Q^0. \tag{17}$$

The restriction and lifting operations are used only for the forward-only oracle diagnostics. Both mappings are fixed and contain no learnable parameters.

**Closed-loop CSC and UC corrections.** Let $x_s^{\mathrm{in},R}$ denote the target-resolution state entering stage $s$. CSC first gives

$$x_{s,0}^{R,\mathrm{preS}} = x_s^{\mathrm{in},R} + \mathcal{L}_{s,R}\mathcal{A}_s\left(\mathcal{P}_{s,R}x_s^{\mathrm{in},R}\right). \tag{18}$$

For block $b$, let $u_{s,b}^{R,\mathrm{S}}$ denote the complete SS2D update computed from $x_{s,b}^{R,\mathrm{preS}}$ using the SDT-transported dynamics. Its corrected update is

$$\widetilde{u}_{s,b}^{R,\mathrm{S}} = u_{s,b}^{R,\mathrm{S}} + \mathcal{L}_{s,R}\mathcal{C}_{s,b}^{\mathrm{S}}\left(\mathcal{P}_{s,R}x_{s,b}^{R,\mathrm{preS}}, \mathcal{P}_{s,R}u_{s,b}^{R,\mathrm{S}}\right), \tag{19}$$

which immediately forms the input to the MLP branch,

$$x_{s,b}^{R,\mathrm{preM}} = x_{s,b}^{R,\mathrm{preS}} + \widetilde{u}_{s,b}^{R,\mathrm{S}}. \tag{20}$$

The MLP is evaluated on this corrected intermediate state and subsequently corrected as

$$\widetilde{u}_{s,b}^{R,\mathrm{M}} = u_{s,b}^{R,\mathrm{M}} + \mathcal{L}_{s,R}\mathcal{C}_{s,b}^{\mathrm{M}}\left(\mathcal{P}_{s,R}x_{s,b}^{R,\mathrm{preM}}, \mathcal{P}_{s,R}u_{s,b}^{R,\mathrm{M}}\right), \tag{21}$$

giving

$$x_{s,b+1}^{R,\mathrm{preS}} = x_{s,b}^{R,\mathrm{preM}} + \widetilde{u}_{s,b}^{R,\mathrm{M}}. \tag{22}$$

The resulting correction is closed-loop: each corrected branch changes the input to all subsequent pretrained operations. In particular, the MLP update is evaluated after the SS2D correction rather than being corrected independently in parallel.

The additive topology further gives a stage-local closure property. If $\mathcal{P}_{s,R}X^R = X^0$ before a branch and its corrected update satisfies $\mathcal{P}_{s,R}U^{R,+} = U^0$, then

$$\mathcal{P}_{s,R}\left(X^R + U^{R,+}\right) = X^0 + U^0. \tag{23}$$

Hence, within a fixed stage, aligning the incoming state and the subsequent additive updates preserves the aligned native-lattice trajectory recursively. We do not assume that this property automatically crosses hierarchy transitions, thus, a new CSC is applied at each stage entry.

**From exact correction to Uniform-RU.** The exact construction in equation 16 requires both native- and target-resolution quantities of the same input and is therefore used only to diagnose the correction structure. The deployable Uniform-RU modules instead predict corrections solely from quantities available on the current target-resolution trajectory:

$$\begin{aligned} a_s &= \mathcal{A}_s\left(\mathcal{P}_{s,R}x_s^{in,R}\right), \\ c_{s,b}^q &= \mathcal{C}_{s,b}^q\left(\mathcal{P}_{s,R}x,\, \mathcal{P}_{s,R}u^q\right), \qquad q \in \{\mathrm{S},\mathrm{M}\}. \end{aligned} \tag{24}$$

During calibration, native- and target-resolution views provide cross-resolution consistency supervision for the correction parameters. At inference, the calibrated model operates directly at the target resolution without requiring native-resolution guidance.

**Calibrator parameterization.** CSC uses a single-input spatial bottleneck. UC independently normalizes the state and branch update, concatenates them, and applies the same bottleneck form. The spatial bottleneck consists of LayerNorm, a linear down-projection, SiLU, an identity residual augmented with a $3 \times 3$ depthwise convolution, and a linear up-projection. UC predictors normalize the state and branch-update inputs independently before concatenation. Both the weight and bias of every output projection are initialized to zero. ViM further uses an independent normalized MLP bottleneck for the CLS-token correction. Their widths are fixed as

$$r_s^A = \frac{C_s}{12}, \qquad r_s^U = \frac{C_s}{6}. \tag{25}$$

For VMamba-Tiny, $C_s = (96, 192, 384, 768)$, giving $r^A = (8, 16, 32, 64)$ and $r^U = (16, 32, 64, 128)$. With stage depths $(2, 2, 8, 2)$, the architecture instantiates four CSC modules, fourteen SS2D UC predictors, and fourteen MLP UC predictors, the latter two are treated jointly as the top-level UC component.

All predictor output projections are zero-initialized. Consequently, before calibration the learned correction path is exactly inactive, and the model reduces to the deterministic SDT configuration.

## A.2 Additional Mechanistic Analyses

### A.2.1 Details of Cross-Resolution Trajectory Discrepancy

We provide the exact construction of the Cross-Resolution Trajectory Discrepancy (CRTD) probe introduced in Sec. 3.3. For a target resolution $R > R^0$, CRTD is computed independently for each image $x$ as

$$\mathrm{CRTD}_x(R; Z) = \frac{1}{N_Z} \sum_{i=1}^{N_Z} d\big(\mathcal{P}_i^{R\to 0}\big(Z_{x,i}^R\big), Z_{x,i}^0\big), \tag{26}$$

where $R^0$ is the model's pretraining resolution, $R$ is a target resolution, and $Z$ denotes a selected internal trajectory. The index $i$ enumerates the corresponding trajectory sites and $N_Z$ is their total number. Below we specify the spatial restriction $\mathcal{P}_i^{R\to 0}$, the feature discrepancy $d(\cdot,\cdot)$, and the state/update trajectory sites used in our experiments.

**Spatial alignment.** Different input resolutions induce different spatial lattices at each stage, so their internal feature maps cannot in general be compared pointwise. Suppose trajectory site $i$ belongs to stage $s$. At this stage, let the spatial grid at the pretraining resolution $R^0$ be $H_s^0 \times W_s^0$ and that at target resolution $R$ be $H_s^R \times W_s^R$. We define $\mathcal{P}_i^{R\to 0} \equiv \mathcal{P}_s^{R\to 0}$ as a parameter-free area-overlap restriction from the target grid to the corresponding pretraining-resolution grid.

For the height axis, the contribution of target-grid cell $i_h$ to reference-grid cell $j_h$ is

$$P^H_{s,j_h i_h} = \frac{\left|\left[j_h \frac{H_s^R}{H_s^0}, (j_h+1)\frac{H_s^R}{H_s^0}\right) \cap [i_h, i_h+1)\right|}{H_s^R / H_s^0}, \tag{27}$$

where $j_h = 0, \ldots, H_s^0 - 1$ and $i_h = 0, \ldots, H_s^R - 1$. The width-axis matrix $P_s^W$ is defined analogously using $W_s^R$ and $W_s^0$.

For $Z \in \mathbb{R}^{C\times H_s^R \times W_s^R}$, the two-dimensional restriction is

$$\big[\mathcal{P}_s^{R\to 0}(Z)\big]_{c,j_h,j_w} = \sum_{i_h=0}^{H_s^R-1} \sum_{i_w=0}^{W_s^R-1} P^H_{s,j_h i_h} P^W_{s,j_w i_w} Z_{c,i_h,i_w}. \tag{28}$$

Thus, each reference-grid location is the area-weighted average of the corresponding region on the target-resolution lattice. The operator contains no learned parameter and does not use any calibration component of our method.

**Per-feature discrepancy.** After spatial alignment, let $a, b \in \mathbb{R}^{C\times H_s^0 \times W_s^0}$ be the target- and pretraining-resolution representations of the same image at a corresponding trajectory site. We define their feature discrepancy as

$$d(a,b) = \frac{1}{2}\left[d_{\mathrm{nMSE}}(a,b) + d_{\cos}(a,b)\right], \tag{29}$$

where the normalized mean-squared error is

$$d_{\mathrm{nMSE}}(a,b) = \frac{\mathrm{mean}\big[(a-b)^2\big]}{\max\big(\mathrm{mean}\big[b^2\big], \epsilon_{\mathrm{m}}\big)}, \tag{30}$$

and the mean spatial cosine distance is

$$d_{\cos}(a,b) = \frac{1}{H_s^0 W_s^0} \sum_{p=1}^{H_s^0 W_s^0} \left[1 - \frac{a_p^\top b_p}{\max(\|a_p\|_2, \epsilon_{\mathrm{c}})\max(\|b_p\|_2, \epsilon_{\mathrm{c}})}\right]. \tag{31}$$

Here, $p$ indexes spatial locations and $a_p, b_p \in \mathbb{R}^C$ are their corresponding channel vectors. We use $\epsilon_{\mathrm{m}} = 10^{-6}$ and $\epsilon_{\mathrm{c}} = 10^{-8}$. The first term measures relative squared feature discrepancy, while the second measures directional changes in feature space. Their weights are fixed to be equal in all experiments.

**State and update trajectories in VMamba.** We now instantiate $Z$ in equation 26 using the VSS Block notation introduced in Sec. 3.1. For block $(s, b)$, $X_{s,b}$ denotes the residual state immediately before the SS2D branch, $\widetilde{X}_{s,b}$ the state immediately before the MLP branch, and $U^{\mathrm{S}}_{s,b}$ and $U^{\mathrm{M}}_{s,b}$ the corresponding SS2D and MLP additive updates:

$$\widetilde{X}_{s,b} = X_{s,b} + U^{\mathrm{S}}_{s,b}, \qquad X_{s,b+1} = \widetilde{X}_{s,b} + U^{\mathrm{M}}_{s,b}. \tag{32}$$

For **CRTD-State**, the trajectory $Z$ consists of the two propagated states in every VSS Block,

$$Z^{\mathrm{state}} = \left\{ X_{s,b}, \widetilde{X}_{s,b} \right\}_{s,b}. \tag{33}$$

Accordingly, equation 26 averages the aligned cross-resolution discrepancies of the states immediately before the SS2D and MLP branches over all VSS Blocks. CRTD-State therefore measures drift of the residual-state trajectory.

For **CRTD-Update**, we instead instantiate

$$Z^{\mathrm{update}} = \left\{ U^{\mathrm{S}}_{s,b}, U^{\mathrm{M}}_{s,b} \right\}_{s,b}. \tag{34}$$

The same CRTD definition then averages the discrepancies of the SS2D and MLP additive updates over all VSS Blocks. CRTD-Update therefore measures how the transformations applied along the residual trajectory change with resolution.

Importantly, CRTD-State and CRTD-Update use the same spatial restriction and the same feature discrepancy in equation 29; they differ only in the internal trajectory $Z$ to which the common CRTD definition is applied. For a VMamba with $N_{\mathrm{blk}}$ VSS Blocks, both trajectories contain $2N_{\mathrm{blk}}$ sites.

**Reference trajectory.** The trajectory induced at the pretraining resolution $R^0$ serves as the fixed reference. For either trajectory instantiation, we define the native-resolution reference discrepancy to be zero:

$$\mathrm{CRTD}_x\big(R^0; Z\big) = 0 \tag{35}$$

This convention fixes the reference point independently of the numerical stabilization in the cosine term. At target resolutions $R > R^0$, lower CRTD indicates closer agreement with the pretraining-resolution trajectory. We use the term discrepancy, rather than error, because the pretraining-resolution trajectory is a reference trajectory rather than a ground-truth hidden trajectory.

**Relation to the mismatch analysis in Sec. 3.2.** The quantities $E^{\mathrm{S}}_{s,b}$, $E^{\mathrm{M}}_{s,b}$, $D^{S}_{s,b}$, and $D^{M}_{s,b}$ in Sec. 3.2 describe how resolution-induced state and update discrepancies propagate through the additive VSS Block recursion. CRTD provides a directly measurable scalar magnitude of these internal deviations from the realized network features: CRTD-State measures the discrepancy of the propagated states, while CRTD-Update measures the discrepancy of the additive updates that drive their evolution. Thus, CRTD operationalizes the trajectory mismatch analyzed in Sec. 3.2 without introducing an additional learned model.

**Per-image computation.** CRTD is defined independently for each image, and ground-truth labels do not enter its computation. When reporting CRTD over an evaluation set, we first evaluate equation 26 separately for every image and only then aggregate the resulting scalar values. Consequently, the metric itself is independent of evaluation batch size.

**Extension to other Visual Mamba architectures.** For the cross-architecture analysis, we retain the same per-image discrepancy in equation 29, the same parameter-free area-overlap alignment, and the same principle of uniform averaging over architecture-native additive trajectory sites. Only the concrete state and update sites are adapted to the residual organization of each architecture.

For ViM, the state trajectory is given by the propagated residual states and the update trajectory by the corresponding mixer updates. Spatial patch tokens are used for the comparison, while the middle class token is excluded because it has no spatially corresponding counterpart in the other architectures. For MambaVision, state sites are taken before the mixer and MLP sublayers, and update sites are their corresponding mixer and MLP updates. For MSVMamba, the analogous sites are the states before SS2D and ConvFFN together with their SS2D and ConvFFN updates. For

EfficientVMamba, each complete residual block contributes its input as a state site and its output-minus-input residual as an update site.

The low-level CRTD definition therefore remains unchanged across architectures. However, because the number and semantics of architecture-native trajectory sites differ, absolute CRTD magnitudes are not intended to rank architectures by intrinsic resolution robustness. The cross-architecture analysis instead tests whether resolution-induced trajectory discrepancy exhibits a consistent trend across heterogeneous Visual Mamba designs.

### A.2.2 Spatial-Rate Interpretation of Resolution-Induced Recurrent Dynamics

The analysis in Sec. 3.2 identifies the change of spatial discretization as the first source of cross-resolution mismatch. Here we make its recurrent-dynamics interpretation explicit. The key distinction is that increasing image resolution does not enlarge the underlying image domain; it samples the same spatial domain with a denser feature lattice. Consequently, along an ordinary within-row or within-column segment of a CrossScan route, the same normalized image-space span is represented by more recurrent steps.

Consider stage $s$ and route $r$, and let $\mathcal{I}^0_{s,r}$ and $\mathcal{I}^R_{s,r}$ denote two matched local scan segments at the native and target resolutions, respectively. We restrict this local argument to segments that do not cross a raster-wrap transition. If $N^0_{s,r} = |\mathcal{I}^0_{s,r}|$ and $N^R_{s,r} = |\mathcal{I}^R_{s,r}|$, then, up to endpoint discretization,

$$\frac{N^R_{s,r}}{N^0_{s,r}} \simeq \frac{1}{\kappa_{s,r}}, \tag{36}$$

where $\kappa_{s,r}$ is the route-axis spacing ratio defined in Sec. 3.2: $\kappa_{s,r} = \kappa^w_s$ for horizontal routes and $\kappa_{s,r} = \kappa^h_s$ for vertical routes. Thus, a smaller spatial cell pitch corresponds to a proportionally larger number of recurrent updates over the same local image-space span.

To see how this changes the state dynamics, let $\bar{\Delta}^{(R)}_{r,t}$ denote the post-softplus effective step produced at target resolution $R$. Over the local segment $\mathcal{I}^R_{s,r}$, the homogeneous state-transition factor of the untransported recurrence is

$$\mathcal{T}^{R,\mathrm{raw}}_{s,r} = \prod_{t\in\mathcal{I}^R_{s,r}} \exp\Big(\bar{\Delta}^{(R)}_{r,t} A_{s,r}\Big) = \exp\big(\Lambda^{R,\mathrm{raw}}_{s,r} A_{s,r}\big), \qquad \Lambda^{R,\mathrm{raw}}_{s,r} = \sum_{t\in\mathcal{I}^R_{s,r}} \bar{\Delta}^{(R)}_{r,t}. \tag{37}$$

The equality follows because all transitions on a fixed route use the same pretrained $A_{s,r}$, when $\bar{\Delta}$ is channel-wise, the expression is understood channel-wise, consistently with equation 1. Therefore, the relevant quantity is not only the effective step of an individual token, but also its accumulation over a fixed spatial extent.

SDT replaces $A_{s,r}$ by $\kappa_{s,r}A_{s,r}$ while leaving the pretrained effective-step predictor unchanged. Its accumulated transition over the same target segment is therefore

$$\mathcal{T}^{R,\mathrm{I}}_{s,r} = \exp\big(\Lambda^{R,\mathrm{I}}_{s,r} A_{s,r}\big), \qquad \Lambda^{R,\mathrm{I}}_{s,r} = \kappa_{s,r} \sum_{t\in\mathcal{I}^R_{s,r}} \bar{\Delta}^{(R)}_{r,t}. \tag{38}$$

This gives a useful first-order interpretation of the geometry transport. To isolate only the change in sampling density, suppose that the mean effective step over two matched local spans is approximately preserved, i.e., $\frac{1}{N^R_{s,r}} \sum_{t\in\mathcal{I}^R_{s,r}} \bar{\Delta}^{(R)}_{r,t} \approx \frac{1}{N^0_{s,r}} \sum_{t\in\mathcal{I}^0_{s,r}} \bar{\Delta}^{(0)}_{r,t}$. Combining this local approximation with equation 36 gives

$$\Lambda^{R,\mathrm{raw}}_{s,r} \approx \frac{1}{\kappa_{s,r}} \Lambda^0_{s,r}, \qquad \Lambda^{R,\mathrm{I}}_{s,r} \approx \Lambda^0_{s,r}, \tag{39}$$

where $\Lambda^0_{s,r} = \sum_{t\in\mathcal{I}^0_{s,r}} \bar{\Delta}^{(0)}_{r,t}$. Hence, before transport, a denser lattice nominally causes excessive recurrent evolution over a fixed spatial span, whereas the geometry-derived factor $\kappa_{s,r}$ compensates for this first-order change in recurrent rate.

The same interpretation applies to state injection. As shown in equation 12, jointly transporting $A$ and $B$ is algebraically equivalent to replacing the effective step by $\kappa_{s,r}\bar{\Delta}_{r,t}$ in both the state-transition and input-injection terms. Scaling only $A$ would alter the memory dynamics without

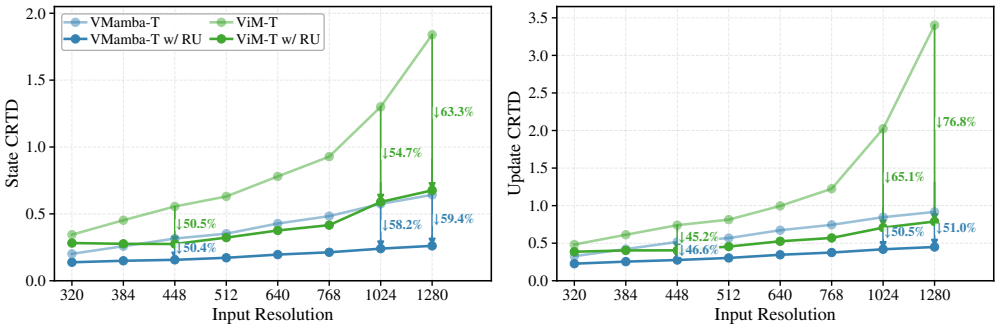



Figure 4: Cross-resolution CRTD-State and CRTD-Update of ViM-T and VMamba-T before and after Uniform-RU calibration. The raw trajectory discrepancy grows with increasing input resolution, whereas Uniform-RU substantially suppresses this growth for both propagated states and additive updates. Percentages indicate the relative CRTD reduction at selected resolutions.

applying the corresponding spatial-rate correction to the injected input, while scaling only $B$ would leave the transition dynamics unchanged. The joint $A/B$ transport therefore preserves the structure of the implemented selective recurrence under this effective-step interpretation.

Importantly, equation 39 is a geometry-isolating first-order interpretation rather than an exact cross-resolution invariance statement. The target-resolution $\bar{\Delta}_{r,t}^{(R)}$ is itself input-dependent and is evaluated from a trajectory that can already differ from the native one; the input-dependent $B_{r,t}$ and the branch inputs can also change with resolution. Moreover, $\kappa_{s,r}$ describes the nominal principal-axis cell-pitch change and does not represent the exact Euclidean displacement of raster-wrap transitions. Finally, local convolution, nonlinear branch transformations, residual propagation, and hierarchy transitions further reshape the target-resolution trajectory. SDT therefore corrects the deterministic first-order spatial-rate mismatch of the selective recurrence, while the remaining state-dependent deviations motivate the CSC and UC corrections and are quantified by CRTD.

### A.2.3 Cross-Architecture Recurrent Trajectory Results and Analysis

In Table 5, we report the CRTD-State and CRTD-Update metrics before and after the full calibration on the same ImageNet-1K probe set respectively.

Figure 4 further visualizes how the internal trajectory discrepancy evolves as the resolution shift becomes more severe. For both ViM-T and VMamba-T, the raw CRTD-State and CRTD-Update increase steadily with input resolution, showing that the deviation from the pretraining-resolution trajectory accumulates as the spatial lattice becomes progressively denser. Uniform-RU consistently suppresses this growth across the entire evaluated range, and the separation from the raw trajectory remains substantial at the largest resolutions. This behavior is observed for both propagated states and additive updates, supporting the view that cross-resolution degradation is not confined to a single internal feature but propagates along the residual trajectory.

The trend is particularly pronounced for ViM-T, whose raw trajectory discrepancy rises rapidly at high resolutions while the calibrated trajectory grows much more slowly. VMamba-T exhibits a milder raw increase, but Uniform-RU likewise maintains a consistently lower trajectory discrepancy throughout the resolution range. These results indicate that the proposed correction remains effective under increasingly large changes in spatial discretization rather than only near the pretraining resolution. Because the number and semantics of the measured trajectory sites differ across architectures, however, the absolute CRTD magnitudes should not be used to rank ViM-T and VMamba-T against each other; the relevant observation is the consistent Raw-to-Uniform-RU reduction within each architecture.

Table 5: Cross-architecture recurrent trajectory discrepancy before and after full calibration. State Red. and Update Red. denote relative CRTD reductions from Raw to Full.

| Architecture | R | Raw State | Full State | State Red. | Raw Update | Full Update | Update Red. |
|---|---|---|---|---|---|---|---|
| VMamba-T | 320 | 0.20 | 0.14 | 31.41 | 0.33 | 0.23 | 30.56 |
| | 384 | 0.26 | 0.15 | 42.13 | 0.42 | 0.25 | 39.59 |
| | 448 | 0.32 | 0.16 | 50.42 | 0.52 | 0.28 | 46.56 |
| | 512 | 0.35 | 0.17 | 51.20 | 0.57 | 0.30 | 46.53 |
| | 640 | 0.43 | 0.20 | 54.27 | 0.67 | 0.35 | 48.59 |
| | 768 | 0.48 | 0.21 | 55.95 | 0.74 | 0.38 | 49.54 |
| | 1024 | 0.57 | 0.24 | 58.23 | 0.85 | 0.42 | 50.48 |
| | 1280 | 0.64 | 0.26 | 59.36 | 0.92 | 0.45 | 50.95 |
| ViM-T | 320 | 0.34 | 0.28 | 18.11 | 0.48 | 0.39 | 19.82 |
| | 384 | 0.45 | 0.28 | 38.98 | 0.61 | 0.40 | 34.04 |
| | 448 | 0.56 | 0.27 | 50.50 | 0.74 | 0.41 | 45.24 |
| | 512 | 0.63 | 0.32 | 48.81 | 0.81 | 0.46 | 44.06 |
| | 640 | 0.78 | 0.38 | 51.79 | 1.00 | 0.53 | 47.23 |
| | 768 | 0.93 | 0.42 | 55.19 | 1.22 | 0.57 | 53.46 |
| | 1024 | 1.30 | 0.59 | 54.71 | 2.02 | 0.71 | 65.06 |
| | 1280 | 1.84 | 0.68 | 63.26 | 3.40 | 0.79 | 76.82 |
| MambaVision-T | 320 | 0.23 | 0.14 | 41.67 | 0.50 | 0.23 | 53.18 |
| | 384 | 0.28 | 0.12 | 56.03 | 0.53 | 0.23 | 55.46 |
| | 448 | 0.32 | 0.12 | 62.89 | 0.55 | 0.24 | 57.45 |
| | 512 | 0.38 | 0.14 | 62.96 | 0.65 | 0.26 | 59.64 |
| | 640 | 0.42 | 0.15 | 64.95 | 0.68 | 0.28 | 58.35 |
| | 768 | 0.47 | 0.16 | 65.32 | 0.74 | 0.30 | 59.06 |
| | 1024 | 0.53 | 0.18 | 66.14 | 0.78 | 0.33 | 57.38 |
| | 1280 | 0.57 | 0.19 | 66.12 | 0.80 | 0.35 | 55.98 |
| MSVMamba-T | 320 | 0.18 | 0.13 | 24.33 | 0.31 | 0.23 | 26.34 |
| | 384 | 0.21 | 0.14 | 32.76 | 0.38 | 0.25 | 32.98 |
| | 448 | 0.24 | 0.15 | 40.07 | 0.45 | 0.27 | 39.00 |
| | 512 | 0.27 | 0.16 | 39.01 | 0.47 | 0.30 | 35.88 |
| | 640 | 0.31 | 0.18 | 42.44 | 0.54 | 0.34 | 36.95 |
| | 768 | 0.34 | 0.19 | 43.72 | 0.58 | 0.37 | 36.68 |
| | 1024 | 0.40 | 0.22 | 45.70 | 0.64 | 0.41 | 36.85 |
| | 1280 | 0.44 | 0.24 | 44.58 | 0.69 | 0.45 | 34.91 |
| EfficientVMamba-B | 320 | 0.15 | 0.11 | 26.75 | 0.22 | 0.16 | 27.66 |
| | 384 | 0.19 | 0.12 | 37.33 | 0.28 | 0.18 | 34.32 |
| | 448 | 0.23 | 0.13 | 44.97 | 0.34 | 0.21 | 38.09 |
| | 512 | 0.26 | 0.14 | 46.69 | 0.37 | 0.23 | 37.80 |
| | 640 | 0.31 | 0.15 | 50.74 | 0.41 | 0.25 | 38.55 |
| | 768 | 0.35 | 0.16 | 52.44 | 0.45 | 0.27 | 38.38 |
| | 1024 | 0.40 | 0.19 | 53.21 | 0.48 | 0.30 | 37.00 |
| | 1280 | 0.43 | 0.20 | 54.00 | 0.51 | 0.32 | 36.63 |

## A.3 Implementation Details and Additional Results

### A.3.1 Calibration Configuration

All classification experiments follow the same label-free calibration protocol. The pretrained model remains frozen, while only CSC and UC are optimized for each target resolution. For each unlabeled image, the frozen model is evaluated at the native resolution to provide reference internal trajectories and predictions, against which the target-resolution branch is calibrated. All correction outputs are initialized to zero, so epoch 0 exactly recovers the corresponding deterministic SDT configuration. The common optimization settings are summarized in Table 6.

For each ImageNet-1K image, the native- and target-resolution views are generated from the same source image using the same canonical evaluation transform at their respective resolutions. Specifically, the shorter side is resized according to the standard crop ratio of $0.875$ using bicubic in-

Table 6: Calibration configuration used for the classification experiments.

| Calibration configuration | Uniform-RU |
|---|---|
| Calibration images | 8,192 |
| Model-selection images | 1,024 |
| Epochs | 8 |
| Optimizer | AdamW |
| Learning rate | $5 \times 10^{-4}$ |
| Weight decay | $10^{-4}$ |
| Effective batch size | 4 |
| Gradient clip norm | 1.0 |
| Output-consistency temperature | 2.0 |
| Loss coefficients | 1.0 |
| Pretrained parameters | Frozen |
| Trainable modules | CSC and UC |
| Class labels | Not used |

terpolation, followed by a center crop to $R \times R$ and standard ImageNet normalization. The same pre-processing is used for calibration, model selection, and final evaluation. Unless otherwise stated, a separate calibrator is fitted for each target resolution, the native-resolution branch is used only as a fixed reference during calibration and is removed at inference.

The calibration objective combines architecture-native trajectory consistency with output consistency. The output-consistency term is the batch-mean KL divergence from the frozen native prediction to the calibrated target prediction. VMamba uses CSC, SS2D-UC, MLP-UC, recurrence, and KD terms; ViM uses CSC, mixer-UC, recurrence, and KD terms; MambaVision uses CSC, mixer-UC, MLP-UC, recurrence, and KD terms; MSVMamba uses CSC, SS2D-UC, ConvFFN-UC, recurrence, and KD terms; EfficientVMamba uses CSC, complete-block UC, recurrence, and KD terms. Spatial-Mamba (Xiao et al., 2025) and VSSD (Shi et al., 2025) follow their corresponding mixer/MLP residual topology described in Appendix A.4.

The final state is selected from epochs $0$–$8$ on the disjoint unlabeled model-selection subset using a fixed ordering of output consistency, recurrent-state consistency, architecture-native update consistency, CSC consistency, and finally the earlier epoch. At inference, the native-resolution branch is removed, only the target-resolution input, deterministic SDT, the frozen pretrained model, and the locked CSC/UC parameters are retained.

### A.3.2 Task-Specific Uniform-RU Instantiation

Uniform-RU is instantiated independently for each vision task, no learned calibrator parameters are transferred across tasks. For a given pretrained task model, the architecture-specific SDT, CSC, and UC construction described in Sec. A.4 is applied directly to that model, while all original task-model parameters remain frozen. The task outputs are used only to define label-free consistency objectives appropriate to the corresponding prediction structure.

**Semantic segmentation.** For ADE20K, we start from the officially trained UPerNet models and freeze the complete task network, including the backbone, decode head, and auxiliary head. The native task resolution is $512$. During calibration, the frozen native-resolution branch and the target-resolution branch receive matched deterministic center-square views of the same training image, resized to $512 \times 512$ and $R \times R$, respectively. This provides a fixed spatial correspondence for both the internal trajectories and dense predictions. SDT follows the stage geometry and native operators of the corresponding backbone, while CSC and UC retain their architecture-specific locations and rank rules but are newly initialized for the segmentation model.

The calibration objective combines CSC consistency, architecture-native UC consistency, recurrent-state consistency, and dense prediction consistency. No segmentation labels are loaded during calibration or model selection. VMamba, MSVMamba, and Spatial-Mamba use 4,096 calibration images, 512 model-selection images, and 512 additional held-out images, whereas MambaVision and

VSSD use 1,024/128/128 images, respectively; the three subsets are disjoint in each setting. Checkpoint selection follows the fixed ordering of dense prediction consistency, recurrent-state consistency, mean architecture-native update consistency, CSC consistency, and the earlier epoch.

The matched square views are used only for calibration. Final evaluation retains the original whole-image, keep-ratio segmentation pipeline. For each test image, SDT and the restriction/lifting operators are instantiated from the actual non-square feature-grid geometry produced by the frozen task model, the native-resolution branch is absent at inference.

**Object detection and instance segmentation.** For COCO, we start from the officially trained Mask R-CNN models and freeze the complete detector, including the backbone, FPN, RPN, and RoI heads. At a target resolution $R$, only the short-side scale is changed while the same aspect-ratio convention is retained. SDT and the restriction/lifting operators are therefore instantiated from the actual per-image, per-stage feature grids. The original detector geometry, including anchor settings and the RoI canonical scale, is left unchanged.

Detection outputs form unordered sets rather than spatially aligned dense predictions, so the segmentation consistency objective cannot be used directly. Instead, we construct task consistency from predictions of the frozen detector. Final source- and target-resolution detections are partitioned into same-class spatial matches, source-only hypotheses, target-only hypotheses, and spatial matches with class disagreement, using an IoU threshold of $0.5$. Same-class matches provide shared support across resolutions, source-only hypotheses provide recovery support, and target-only hypotheses preserve predictions that emerge at the higher resolution. Class-disagreement matches are excluded from RoI consistency supervision.

At the RPN level, target-resolution RAW predictions provide same-grid preservation targets, while source-only detections mapped to the target image provide recovery targets through the corresponding official anchors. We do not impose cross-resolution RPN feature matching. At the RoI level, matched detections constrain classification while retaining target-resolution RAW geometry for box and mask predictions; source-only detections provide source-recovery targets, and target-only detections preserve the corresponding RAW target predictions.

The resulting task objective contains five detector-consistency components: RPN objectness, RPN box regression, RoI classification, RoI box regression, and RoI mask prediction. Their scales are normalized by fixed per-component statistics estimated before optimization; under the formal 4,096/512/512 calibration/model-selection/held-out protocol, these statistics are estimated from the first 256 calibration images and then kept fixed. These task terms are optimized jointly with CSC, architecture-native UC, and recurrent-state consistency, without using ground-truth detection or instance segmentation annotations. Model selection follows the fixed ordering of task consistency, recurrent-state consistency, mean architecture-native update consistency, CSC consistency, and the earlier epoch.

### A.3.3 Optional Inference Extensions

Beyond the standard single-path Uniform-RU, we introduce two optional task-specific inference extensions that combine the direct target-resolution baseline with its Uniform-RU counterpart for semantic segmentation, object detection and instance segmentation. Both extensions operate on an already calibrated Uniform-RU model and do not modify the underlying task model or the calibration procedure.

**Margin-Aware Prediction Selection (MAPS).** For semantic segmentation, the baseline and Uniform-RU predictions are naturally aligned at the pixel level. We therefore perform local selection directly in the prediction space. Specifically, MAPS compares the classification margin of the two predictions at each pixel and retains the one with stronger class separation. Unlike logit averaging, MAPS does not explicitly mix the two predictive distributions, but adaptively selects between them at each spatial location. The selection relies only on model predictions and requires neither additional supervision nor retraining.

**Proposal-Aligned Logit Fusion (PALF).** For object detection and instance segmentation, the final predictions form an unordered set of instances and therefore do not admit direct pixel-wise selection.

Table 7: Task-level comparison with the matched output-only control on ViM-T. Both variants use the same calibration architecture, parameterization, data, and optimization protocol; they differ only in the calibration objective. Results are ImageNet-1K Top-1 accuracy (%).

| Method | $R = 640$ | $R = 1024$ | $R = 1280$ |
|---|---|---|---|
| Output-only | 66.7 | 49.5 | 40.8 |
| Full Uniform-RU | **68.2** | **52.2** | **43.2** |
| Gain | **+1.5** | **+2.8** | **+2.5** |

PALF instead establishes correspondence between the baseline and Uniform-RU RoI predictions through a shared proposal set, and then combines their classification evidence at the aligned RoI-logit level. Localization and mask prediction retain the baseline prediction path. Fusion is therefore performed only after instance correspondence has been established, while the original Mask R-CNN detection and post-processing structure remains unchanged. PALF is applied only at inference time and requires no additional training.

**Common properties.** MAPS and PALF are both used as inference-time compositions rather than additional training stages. Neither changes the backbone, task head, or Uniform-RU calibration procedure, and neither introduces additional learnable parameters. Their additional computational cost mainly arises from evaluating both the baseline and Uniform-RU prediction paths, we therefore report their complexity separately from standard Uniform-RU.

### A.3.4 Additional Quantitative and Qualitative Results

This section complements the main experiments with additional quantitative, qualitative comparisons, and expanded evaluation results. We include supplementary results that further characterize the behavior of Uniform-RU, together with the complete COCO evaluation metrics for object detection and instance segmentation. Unless otherwise stated, all results use the same locked calibrators and evaluation protocols as the corresponding main experiments, with no additional calibration or model selection.

**Task-level effect on ViM-T.** We further examine whether explicit trajectory supervision provides task-level information beyond final-output consistency in ViM-T, where resolution extension induces particularly severe degradation. Under the same calibration protocol and parameterization, the full trajectory-aware objective consistently outperforms the matched output-only control across all three evaluated target resolutions, as shown in Table 7. The gain increases from $+1.5$ pp at $R = 640$ to $+2.8$ pp at $R = 1024$, and remains substantial ($+2.5$ pp) at $R = 1280$. Together with the lower CRTD reported above, these results show that explicit trajectory supervision can provide complementary task-level information beyond final-output matching under severe cross-resolution shifts.

**Matched Output-Only Objective Control.** To isolate the contribution of trajectory-aware supervision, we compare Uniform-RU with a matched *output-only* control on ViM-T and VMamba-T at $R \in \{640, 1024, 1280\}$. The control retains exactly the same SDT, CSC, and UC pathway and trainable parameterization as the full method, while optimizing only the final-output consistency term. The pretrained backbone, data splits, initialization, optimizer, training schedule, and other calibration settings are unchanged.

As shown in Table 8, final-output consistency alone does not recover the same internal cross-resolution trajectories. Across all six backbone–resolution settings, the full objective consistently reduces both CRTD-State and CRTD-Update relative to the matched output-only control. The reductions are $11.4$–$35.7\%$ and $23.2$–$34.6\%$ on ViM-T, and $51.1$–$53.4\%$ and $50.6$–$51.4\%$ on VMamba-T, for CRTD-State and CRTD-Update, respectively. Paired bootstrap confidence intervals for the Full-Output differences remain strictly below zero in all settings. These results show that trajectory-aware calibration restores internal cross-resolution dynamics beyond final-output consistency alone.

Table 8: Matched output-only objective control on ViM-T and VMamba-T. CRTD is evaluated on the fixed reporting subset. Lower CRTD is better.

| Backbone | $R$ | CRTD-State | | CRTD-Update | |
|---|---|---|---|---|---|
| | | Full | Output | Full | Output |
| ViM-T | 640 | **0.3762** | 0.5851 | **0.5294** | 0.8098 |
| | 1024 | **0.5895** | 0.7145 | **0.7104** | 1.0053 |
| | 1280 | **0.6772** | 0.7646 | **0.7919** | 1.0312 |
| VMamba-T | 640 | **0.1956** | 0.3996 | **0.3462** | 0.7009 |
| | 1024 | **0.2400** | 0.5129 | **0.4190** | 0.8591 |
| | 1280 | **0.2610** | 0.5604 | **0.4494** | 0.9249 |

Table 9: Detailed COCO detection and instance segmentation results across target resolutions.

| Backbone | Method | $R = 1024$ | $R = 1280$ | $R = 1600$ |
|---|---|---|---|---|
| VMamba-B | Raw | 49.6 / 71.7 / 54.4<br>44.4 / 68.7 / 48.1 | 48.5 / 70.7 / 53.2<br>43.5 / 68.0 / 47.0 | 44.6 / 66.3 / 48.9<br>40.6 / 64.2 / 43.6 |
| | Uniform-RU | 49.4 / 71.9 / 53.8<br>44.4 / 68.9 / 47.9 | 48.3 / 71.2 / 52.9<br>43.7 / 68.2 / 46.8 | 46.6 / 69.7 / 51.3<br>42.7 / 67.1 / 45.8 |
| | + PALF | 49.5 / 71.9 / 54.2<br>44.6 / 68.9 / 48.2 | 48.8 / 71.4 / 53.7<br>44.0 / 68.6 / 47.3 | 47.6 / 70.3 / 52.5<br>43.2 / 67.7 / 46.3 |
| VMamba-S | Raw | 49.3 / 71.1 / 54.3<br>44.1 / 68.2 / 47.8 | 48.4 / 70.0 / 53.5<br>43.4 / 67.5 / 47.0 | 44.7 / 66.0 / 48.6<br>40.4 / 63.7 / 43.9 |
| | Uniform-RU | 49.2 / 70.9 / 54.0<br>44.2 / 67.9 / 48.0 | 48.1 / 70.0 / 52.9<br>43.4 / 67.3 / 47.0 | 46.1 / 68.5 / 50.9<br>42.2 / 65.8 / 45.9 |
| | + PALF | 49.3 / 71.0 / 54.0<br>44.2 / 68.1 / 47.9 | 48.6 / 70.4 / 53.7<br>43.6 / 67.5 / 47.3 | 47.0 / 69.3 / 51.7<br>42.6 / 66.6 / 46.3 |
| VMamba-T | Raw | 47.5 / 69.9 / 51.9<br>42.9 / 67.0 / 46.0 | 46.4 / 68.7 / 51.0<br>42.0 / 65.9 / 45.1 | 42.7 / 64.5 / 46.5<br>38.9 / 62.1 / 41.6 |
| | Uniform-RU | 47.6 / 69.9 / 52.3<br>43.1 / 67.0 / 46.3 | 46.4 / 69.1 / 50.6<br>42.3 / 66.2 / 45.2 | 44.3 / 66.4 / 48.2<br>40.2 / 63.7 / 42.9 |
| | + PALF | 47.6 / 70.0 / 52.3<br>43.0 / 67.2 / 46.1 | 46.9 / 69.3 / 51.2<br>42.5 / 66.5 / 45.5 | 45.4 / 67.7 / 49.5<br>41.2 / 65.1 / 44.0 |
| MSVMamba-Micro | Raw | 44.4 / 66.8 / 48.6<br>40.5 / 64.0 / 43.2 | 42.6 / 65.3 / 46.5<br>39.2 / 62.5 / 41.8 | 38.4 / 60.4 / 42.1<br>35.5 / 57.9 / 37.8 |
| | Uniform-RU | 44.2 / 66.7 / 48.5<br>40.3 / 63.9 / 43.1 | 42.9 / 65.8 / 47.0<br>39.6 / 63.0 / 42.5 | 40.4 / 63.3 / 43.9<br>37.6 / 60.6 / 40.3 |
| | + PALF | 44.4 / 66.8 / 48.8<br>40.5 / 64.0 / 43.3 | 43.2 / 66.1 / 47.1<br>39.7 / 63.2 / 42.5 | 40.9 / 63.8 / 44.6<br>37.9 / 61.2 / 40.5 |
| SpatialMamba-T | Raw | 47.5 / 69.9 / 52.1<br>42.9 / 66.8 / 46.4 | 45.4 / 68.1 / 49.8<br>41.2 / 65.2 / 44.2 | 40.7 / 62.3 / 44.6<br>37.4 / 60.2 / 39.9 |
| | Uniform-RU | 47.6 / 70.0 / 52.1<br>43.0 / 67.1 / 46.4 | 45.8 / 68.9 / 50.5<br>42.0 / 66.4 / 45.0 | 43.8 / 66.9 / 48.4<br>40.3 / 64.4 / 43.1 |
| | + PALF | 47.7 / 70.1 / 52.2<br>43.1 / 67.2 / 46.7 | 46.1 / 69.0 / 50.8<br>41.9 / 66.3 / 45.1 | 44.2 / 67.5 / 48.5<br>40.3 / 64.4 / 43.3 |
| VSSD-Micro | Raw | 45.2 / 67.6 / 49.4<br>41.1 / 64.6 / 44.0 | 41.8 / 64.9 / 45.3<br>38.4 / 62.1 / 40.8 | 32.7 / 54.0 / 34.5<br>30.2 / 51.6 / 31.2 |
| | Uniform-RU | 45.2 / 67.9 / 49.6<br>41.2 / 64.8 / 44.1 | 42.8 / 65.9 / 46.5<br>39.3 / 63.2 / 42.1 | 37.1 / 59.6 / 39.5<br>34.6 / 57.4 / 36.4 |
| | + PALF | 45.3 / 67.9 / 49.6<br>41.2 / 64.7 / 44.1 | 42.5 / 65.8 / 46.2<br>39.0 / 62.8 / 41.5 | 35.9 / 58.6 / 38.1<br>32.6 / 55.5 / 33.7 |

**Detailed COCO detection and instance segmentation results.** Beyond the overall box and mask AP reported in the main paper, Table 9 provides the corresponding $AP_{50}$ and $AP_{75}$ results across target resolutions. All results use the same locked predictions and evaluation protocol as Table 3, without additional calibration or model selection. The expanded metrics provide a more detailed view of recovery under different IoU criteria.

### A.4 Architecture-Specific Instantiations across Visual Mamba Models

We use VMamba as the canonical instantiation because its hierarchical Cross2D backbone directly exposes the three structures addressed by SDT, CSC, and UC: spatial recurrent dynamics, carried residual states, and additive updates. The same formulation is instantiated in other visual Mamba architectures according to their native scan operators and residual topologies. For each target resolution, SDT is fixed by the architecture and target-grid geometry, while CSC and UC are calibrated on unlabeled images.

Table 10: Architecture-specific instantiations of Uniform-RU. SDT follows the native spatial operator of each backbone, while CSC and UC follow its residual topology.

| Backbone | SDT | CSC/UC topology |
|---|---|---|
| VMamba | Cross2D route-axis $A/B$ transport; local Conv2d unchanged | 1 CSC per stage; 1 UC per SS2D and MLP additive branch |
| ViM | Effective-$\Delta$ scaling and local Conv1d support transport | 1 mid-depth CSC; 13 mixer UCs |
| MambaVision | Token-window transport, Mamba $A/B$ transport, and Conv1d support transport | 2 token-stage CSCs; 12 mixer UCs; 12 MLP UCs |
| MSVMamba | Stage-wise MS-SS2D $A/B$ transport; local convolution and multi-scale topology unchanged | 4 stage CSCs; 14 SS2D UCs; 14 ConvFFN UCs |
| EfficientVMamba | Route/position-aware effective-$\Delta$ transport; native softplus preserved | 4 stage CSCs; 15 complete residual-block UCs |
| SpatialMamba | Stage-wise SSM $A/B$ transport and StateFusion support transport | 4 stage CSCs; 18 StructureAwareSSM UCs; 18 MLP UCs |
| VSSD | Stage-wise Mamba2 transition transport in S0–S2; standard attention in S3 unchanged | 4 stage CSCs; 16 mixer UCs; 16 MLP UCs |

**VMamba.** For stage $s$, let the native and target feature grids be $H_s^0 \times W_s^0$ and $H_s^R \times W_s^R$. Following Appendix A.1, we define

$$\kappa_s^w = \frac{W_s^0}{W_s^R}, \qquad \kappa_s^h = \frac{H_s^0}{H_s^R},$$

and assign the four Cross2D routes the scales $[\kappa_s^w, \kappa_s^h, \kappa_s^w, \kappa_s^h]$. At each SS2D selective-scan call, the corresponding route scale is applied jointly to $A$ and $B$. The local Conv2d and all other pretrained operators remain unchanged.

Each of the four stages receives one CSC at its entry. Here, $\Phi_{s,b}^{\mathrm{S}}$ and $\Phi_{s,b}^{\mathrm{M}}$ denote the complete SS2D and MLP additive branches of the original VSS Block, respectively, so RU is applied to the realized additive increment without altering the pretrained branch ordering. Within every VSS Block, Uniform-RU corrects the complete SS2D update and then the complete MLP update in the original forward order. The two branch families jointly instantiate the top-level UC component. Algorithms 2–7 show only the architecture-specific Uniform-RU insertions within the original target-resolution forward; all unlisted pretrained operations are retained in their original order.

**ViM.** Unlike VMamba, ViM forms a flat sequence of pre-normalized residual Mamba blocks with RMSNorm (Zhang & Sennrich, 2019), without explicit hierarchical stage boundaries. A preliminary

**Algorithm 1** Canonical Uniform-RU Instantiation for VMamba

**Input:** Frozen pretrained VMamba $f_\theta$; native resolution $R_0 = 224$; target resolution $R$; unlabeled $\mathcal{D}_{\text{fit}}, \mathcal{D}_{\text{sel}}$
**Output:** Locked target-specific correction parameters $\phi_R^\star$

**for** stage $s = 0, \dots, 3$ **do**
  $\kappa_s^w \leftarrow W_s^0/W_s^R, \quad \kappa_s^h \leftarrow H_s^0/H_s^R$
  $\boldsymbol{\kappa}_s \leftarrow [\kappa_s^w, \kappa_s^h, \kappa_s^w, \kappa_s^h]$
  At each SS2D scan, use $(\widetilde{A}_r, \widetilde{B}_r) \leftarrow (\kappa_{s,r} A_r, \kappa_{s,r} B_r)$
**end for**
Keep all other pretrained operators unchanged
Initialize one CSC per stage and one UC per SS2D/MLP branch with zero-output projections
Evaluate $\phi_R^{(0)}$ on $\mathcal{D}_{\text{sel}}$
**for** $e = 1, \dots, 8$ **do**
  **for** micro-batch $\mathcal{B} \subset \mathcal{D}_{\text{fit}}$ **do**
    Obtain native references and logits from the unmodified frozen pass $f_\theta(\mathcal{B}^{R_0})$
    $h \leftarrow \mathrm{Stem}_\theta(\mathcal{B}^R)$
    **for** stage $s = 0, \dots, 3$ **do**
      $h \leftarrow h + \mathcal{L}_s^R \mathcal{A}_s(\mathcal{P}_s^R h)$
      **for** VSS Block $(s, b)$ in forward order **do**
        $u^{\mathrm{S}} \leftarrow \Phi_{s,b}^{\mathrm{S}}(h); \quad u^{\mathrm{S}} \leftarrow u^{\mathrm{S}} + \mathcal{L}_s^R \mathcal{U}_{s,b}^{\mathrm{S}}(\mathcal{P}_s^R h, \mathcal{P}_s^R u^{\mathrm{S}})$
        $h \leftarrow h + u^{\mathrm{S}}$
        $u^{\mathrm{M}} \leftarrow \Phi_{s,b}^{\mathrm{M}}(h); \quad u^{\mathrm{M}} \leftarrow u^{\mathrm{M}} + \mathcal{L}_s^R \mathcal{U}_{s,b}^{\mathrm{M}}(\mathcal{P}_s^R h, \mathcal{P}_s^R u^{\mathrm{M}})$
        $h \leftarrow h + u^{\mathrm{M}}$
      **end for**
      **if** $s < 3$ **then**
        $h \leftarrow \mathrm{Downsample}_{\theta,s}(h)$
      **end if**
    **end for**
    $z^R \leftarrow \mathrm{Head}_\theta(h)$
    Accumulate gradients to effective batch size 4; update only $\phi_R$
  **end for**
  Evaluate $\phi_R^{(e)}$ on $\mathcal{D}_{\text{sel}}$
**end for**
Select $e^\star \in \{0, \dots, 8\}$ by output $\rightarrow$ recurrence $\rightarrow$ update $\rightarrow$ CSC $\rightarrow$ earlier epoch
$\phi_R^\star \leftarrow \phi_R^{(e^\star)}$; lock it before official validation
**return** $\phi_R^\star$

label-free phase diagnosis identified a correctable mid-to-late residual regime beginning at L11. We therefore place one CSC at L11 and apply mixer UCs from L11 through the final block L23, following the additive residual recurrence. This placement is fixed before the multi-resolution study and is reused for all target resolutions without further layer search.

For patch size 16, let $G_R = R/16$. The locked SDT uses

$$\alpha_R = \frac{14}{G_R}, \qquad d_R = \mathrm{round}\left(\frac{G_R}{14}\right),$$

where $\alpha_R$ rescales the recurrent step and $d_R$ transports the local Conv1d support.

Calibration treats spatial patch tokens and the middle CLS token through separate correction paths. For the patch tokens, the fixed calibration restriction/lifting maps use the implementation's native-grid center-assignment aggregation and replication; the CLS token is preserved one-to-one. The patch correction uses the spatial bottleneck described in Sec. A.1, whereas the CLS token uses an independent normalized MLP bottleneck. Both output projections are zero-initialized. The L11 CSC acts on ViM's effective carried state (hidden plus residual), and mixer UCs are applied from L11 through L23.

**Algorithm 2** Uniform-RU Instantiation for ViM

**Input:** Frozen pretrained ViM $f_\theta$; target resolution $R$

$G_R \leftarrow R/16$, $\alpha_R \leftarrow 14/G_R$, $d_R \leftarrow \mathrm{round}(G_R/14)$
In each BiMamba mixer, use $\widetilde{\bar{\Delta}} \leftarrow \alpha_R \bar{\Delta}$ and set the local Conv1d dilation to $d_R$
Initialize one CSC at L11 and mixer UCs at L11–L23, with zero-output patch/CLS correction paths
$h_{11} \leftarrow h_{11} + \mathcal{L}_R \mathcal{A}_{11}(\mathcal{P}_R h_{11})$
**for** $\ell = 11, \ldots, 23$ **do**
    $u_\ell \leftarrow \Phi_\ell(h_\ell)$
    $u_\ell \leftarrow u_\ell + \mathcal{L}_R \mathcal{U}_\ell\left(\mathcal{P}_R h_\ell, \mathcal{P}_R u_\ell\right)$
    Propagate $(h_\ell, u_\ell)$ through the original ViM residual recurrence
**end for**
Calibrate and select $\phi_R$ using the common protocol of Algorithm 1

**Algorithm 3** Uniform-RU Instantiation for MambaVision

**Input:** Frozen pretrained MambaVision $f_\theta$; target resolution $R$

$G_2^R \leftarrow R/16, \quad G_3^R \leftarrow R/32, \quad \kappa_R \leftarrow 224/R$
$d_R \leftarrow \max(1, \mathrm{round}(R/224))$
Set token-stage windows to $G_2^R$ and $G_3^R$
For each Mamba mixer, use $(\widetilde{A}, \widetilde{B}) \leftarrow (\kappa_R A, \kappa_R B)$ and $x/z$ Conv1d dilation $d_R$
Initialize one CSC per token stage and one UC per mixer/MLP branch with zero-output projections
**for** token stage $s \in \{2, 3\}$ **do**
    $h \leftarrow h + \mathcal{L}_s^R \mathcal{A}_s(\mathcal{P}_s^R h)$
    **for** block $b$ in stage $s$ **do**
        $u^{\mathrm{mix}} \leftarrow \Phi_{s,b}^{\mathrm{mix}}(h)$
        $u^{\mathrm{mix}} \leftarrow u^{\mathrm{mix}} + \mathcal{L}_s^R \mathcal{U}_{s,b}^{\mathrm{mix}}\left(\mathcal{P}_s^R h, \mathcal{P}_s^R u^{\mathrm{mix}}\right)$
        $h \leftarrow h + u^{\mathrm{mix}}$
        $u^{\mathrm{mlp}} \leftarrow \Phi_{s,b}^{\mathrm{mlp}}(h)$
        $u^{\mathrm{mlp}} \leftarrow u^{\mathrm{mlp}} + \mathcal{L}_s^R \mathcal{U}_{s,b}^{\mathrm{mlp}}\left(\mathcal{P}_s^R h, \mathcal{P}_s^R u^{\mathrm{mlp}}\right)$
        $h \leftarrow h + u^{\mathrm{mlp}}$
    **end for**
    **if** $s = 2$ **then**
        $h \leftarrow \mathrm{Downsample}_{\theta,2}(h)$
    **end if**
**end for**
Calibrate and select $\phi_R$ using the common protocol of Algorithm 1

**MambaVision.** MambaVision contains two token stages composed of Mamba and attention mixers. For target resolution $R$, the locked all-resolution SDT uses

$$G_2^R = \frac{R}{16}, \qquad G_3^R = \frac{R}{32}, \qquad \kappa_R = \frac{224}{R},$$

together with

$$d_R = \max\left(1, \mathrm{round}\left(\frac{R}{224}\right)\right).$$

The token-stage windows are set to $G_2^R$ and $G_3^R$. Only Mamba mixers receive the $A/B$ transport and the $x/z$ Conv1d dilation change; attention mixers retain their native operator.

Uniform-RU places one CSC at each token-stage entry and corrects every mixer and MLP additive update. The 12 mixer UCs consist of six Mamba and six attention branches. For arbitrary target/native grid ratios, the fixed restriction/lifting maps use normalized physical-cell assignment and contain no learnable parameters.

**MSVMamba.** MSVMamba retains a hierarchical two-dimensional selective-scan structure, while each MS-SS2D block contains both full-resolution multi-scale scan branches. For the square-

**Algorithm 4** Uniform-RU Instantiation for MSVMamba

**Input:** Frozen pretrained MSVMamba $f_\theta$; target resolution $R$

**for** stage $s = 0, \dots, 3$ **do**
  $\kappa_s \leftarrow G_s^0 / G_s^R$
  For each MS-SS2D scan in stage $s$, use $(\widetilde{A}, \widetilde{B}) \leftarrow (\kappa_s A, \kappa_s B)$
**end for**
Keep local convolution and the original multi-scale topology unchanged
Initialize one CSC per stage and one UC per MS-SS2D/ConvFFN branch with zero-output projections
**for** stage $s = 0, \dots, 3$ **do**
  $h \leftarrow h + \mathcal{L}_s^R \mathcal{A}_s(\mathcal{P}_s^R h)$
  **for** MSVSS block $(s, b)$ in forward order **do**
    $u^{\mathrm{S}} \leftarrow \Phi_{s,b}^{\mathrm{S}}(h)$
    $u^{\mathrm{S}} \leftarrow u^{\mathrm{S}} + \mathcal{L}_s^R \mathcal{U}_{s,b}^{\mathrm{S}} \left(\mathcal{P}_s^R h, \mathcal{P}_s^R u^{\mathrm{S}}\right)$
    $h \leftarrow h + u^{\mathrm{S}}$
    $u^{\mathrm{F}} \leftarrow \Phi_{s,b}^{\mathrm{F}}(h)$
    $u^{\mathrm{F}} \leftarrow u^{\mathrm{F}} + \mathcal{L}_s^R \mathcal{U}_{s,b}^{\mathrm{F}} \left(\mathcal{P}_s^R h, \mathcal{P}_s^R u^{\mathrm{F}}\right)$
    $h \leftarrow h + u^{\mathrm{F}}$
  **end for**
  **if** $s < 3$ **then**
    $h \leftarrow \mathrm{Downsample}_{\theta,s}(h)$
  **end if**
**end for**
Calibrate and select $\phi_R$ using the common protocol of Algorithm 1

resolution setting considered here, SDT uses the parent-stage spacing ratio

$$\kappa_s = \frac{G_s^0}{G_s^R}.$$

The same $\kappa_s$ is applied jointly to $A$ and $B$ in every selective-scan call associated with stage $s$, including both full and stride-2 multi-scale branches. The fixed stride does not introduce an additional cardinality-based scale. Local convolution and the original multi-scale scan topology remain unchanged.

**EfficientVMamba.** EfficientVMamba requires a different SDT because EfficientScan partitions the feature lattice into four sparse rasterized routes. The physical step therefore depends on both the route and the transition position, particularly at raster-wrap transitions.

For route $r$ and sequence position $t$, the locked geometry defines

$$\kappa_{r,t}(R) = \frac{d_{r,t}^R}{d_{r,\tau(t)}^0},$$

where $\tau(t)$ denotes the corresponding native transition type. For ordinary local transitions, the factor reduces to the local cell-pitch ratio; wrap transitions use the exact normalized raster-wrap distance ratio. Here, $\tau(t)$ denotes the corresponding transition type rather than a one-to-one token correspondence. At the first sequence position, for which no predecessor displacement exists, the implementation assigns the corresponding local cell-pitch ratio.

The production implementation preserves the repository's original softplus path. With

$$\Delta_{\mathrm{eff}} = \mathrm{softplus}(\delta_{\mathrm{raw}} + b),$$

we construct

$$\widetilde{\Delta}_{\mathrm{eff}} = \kappa_{r,t}(R) \Delta_{\mathrm{eff}}, \qquad \widetilde{\delta}_{\mathrm{raw}} = \mathrm{softplus}^{-1}(\widetilde{\Delta}_{\mathrm{eff}}) - b,$$

and invoke the original selective-scan kernel with the unchanged bias and its internal softplus enabled. At the native resolution, the transport path is bypassed. We evaluate the inverse softplus with the numerically stable form $y + \log(-\mathrm{expm1}(-y))$ for $y > 0$.

**Algorithm 5** Uniform-RU Instantiation for EfficientVMamba

**Input:** Frozen pretrained EfficientVMamba $f_\theta$; native resolution $R_0 = 224$; target resolution $R$

**for** each EfficientScan call **do**
  **if** $R = R_0$ **then**
    Run the original selective-scan kernel
  **else**
    Construct the locked factor tensor $\boldsymbol{\kappa}(R)$ over all routes and positions
    $\Delta_{\text{eff}} \leftarrow \text{softplus}(\delta_{\text{raw}} + b)$
    $\widetilde{\delta}_{\text{raw}} \leftarrow \text{softplus}^{-1}(\boldsymbol{\kappa}(R) \odot \Delta_{\text{eff}}) - b$
    Run the original kernel once with $\widetilde{\delta}_{\text{raw}}$, original $b$, and internal softplus $\widetilde{\delta}_{\text{raw}}$
  **end if**
**end for**
Initialize one CSC per stage and one UC per complete residual block with zero-output projections
**for** stage $s = 0, \ldots, 3$ **do**
  $h \leftarrow h + \mathcal{L}_s^R \mathcal{A}_s(\mathcal{P}_s^R h)$
  **for** residual block $\ell$ in stage $s$ **do**
    $y_\ell \leftarrow \text{Block}_{\theta,\ell}(h)$
    $\bar{h}_\ell \leftarrow \mathcal{P}_s^R h, \quad \bar{u}_\ell \leftarrow \mathcal{P}_s^R y_\ell - \bar{h}_\ell$
    $h \leftarrow y_\ell + \mathcal{L}_s^R \mathcal{U}_\ell(\bar{h}_\ell, \bar{u}_\ell)$
  **end for**
  **if** $s < 3$ **then**
    $h \leftarrow \text{Transition}_{\theta,s}(h)$
  **end if**
**end for**
Calibrate and select $\phi_R$ using the common protocol of Algorithm 1

Uniform-RU follows the hybrid residual topology rather than individual internal branches. The locked model contains four Efficient-Mamba residual blocks and eleven InvertedResidual blocks, yielding fifteen complete residual-block UC sites.

**SpatialMamba.** SpatialMamba couples selective state-space dynamics with structure-aware state fusion. For stage $s$, let $G_s^0$ and $G_s^R$ denote the native- and target-resolution feature-grid sizes. We use

$$\kappa_s = \frac{G_s^0}{G_s^R}.$$

Within each StructureAwareSSM, SDT jointly transports the state-transition and input-injection terms as

$$\widetilde{A}_s = \kappa_s A_s, \qquad \widetilde{B}_s = \kappa_s B_s.$$

SpatialMamba additionally fuses recurrent states through depthwise convolutions with native dilations $d \in \{1, 3, 5\}$. To preserve their spatial support under resolution scaling, we use

$$\widetilde{d}_{s,d} = \max\left(1, \text{round}\left(d\,\frac{G_s^R}{G_s^0}\right)\right).$$

The $\Delta$ projection, $C$, $D$, CPE, MLP, stem, and down-sampling operators remain unchanged.

Uniform-RU places one CSC at each of the four stage entries and applies UC to the realized StructureAwareSSM and MLP additive updates in every block, while preserving the original CPE and residual ordering. For SpatialMamba-T with stage depths $(2, 4, 8, 4)$, this gives four CSCs, 18 StructureAwareSSM UCs, and 18 MLP UCs.

**Algorithm 6** Uniform-RU Instantiation for SpatialMamba

**Input:** Frozen pretrained Spatial-Mamba $f_\theta$; target resolution $R$

**for** stage $s = 0, \ldots, 3$ **do**
  $\kappa_s \leftarrow G_s^0 / G_s^R$
  For each StructureAwareSSM, use $(\widetilde{A}, \widetilde{B}) \leftarrow (\kappa_s A, \kappa_s B)$
  Transport StateFusion dilations as $\widetilde{d} \leftarrow \max(1, \mathrm{round}(d/\kappa_s))$, $d \in \{1, 3, 5\}$
**end for**
Keep all other pretrained operators unchanged
Initialize one CSC per stage and one UC per StructureAwareSSM/MLP additive branch with zero-output projections
**for** stage $s = 0, \ldots, 3$ **do**
  $h \leftarrow h + \mathcal{L}_s^R \mathcal{A}_s(\mathcal{P}_s^R h)$
  **for** block $b$ in stage $s$ **do**
    $u^S \leftarrow \Phi_{s,b}^S(h)$
    $u^S \leftarrow u^S + \mathcal{L}_s^R \mathcal{U}_{s,b}^{\mathrm{S}}(\mathcal{P}_s^R h, \mathcal{P}_s^R u^S)$
    $h \leftarrow h + u^S$
    $u^M \leftarrow \Phi_{s,b}^M(h)$
    $u^M \leftarrow u^M + \mathcal{L}_s^R \mathcal{U}_{s,b}^{\mathrm{M}}(\mathcal{P}_s^R h, \mathcal{P}_s^R u^M)$
    $h \leftarrow h + u^M$
  **end for**
  **if** $s < 3$ **then**
    $h \leftarrow \mathrm{Downsample}_{\theta,s}(h)$
  **end if**
**end for**
Calibrate and select $\phi_R$ using the common protocol of Algorithm 1

**VSSD.** VSSD uses Mamba2 mixers in the first three stages and standard attention in the final stage. For a square feature grid, we define

$$\rho_s = \left( \frac{G_s^0}{G_s^R} \right)^2, \qquad s \in \{0, 1, 2\}.$$

For the positive-$dA$ Mamba2 path used by VSSD-Micro, the locked SDT is implemented at the native mixer boundary by replacing the call-time transition parameter with

$$\widetilde{A}_s = \rho_s A_s,$$

while leaving $B$, $C$, $D$, and the remaining pretrained operators unchanged. The standard-attention fourth stage is not modified by SDT.

CSC is inserted at all four stage entries. UC follows the native residual topology and corrects the realized mixer and MLP increments after the architecture's native scaling operations. Thus, for stage depths $(2, 2, 8, 4)$, VSSD-Micro contains four CSCs, 16 mixer UCs, and 16 MLP UCs; the mixer UCs comprise 12 Mamba2 sites and four standard-attention sites.

**Algorithm 7** Uniform-RU Instantiation for VSSD

**Input:** Frozen pretrained VSSD $f_\theta$; target resolution $R$

**for** stage $s = 0, 1, 2$ **do**
  $\rho_s \leftarrow (G_s^0/G_s^R)^2$
  For each Mamba2 mixer, use $\widetilde{A} \leftarrow \rho_s A$
**end for**
Keep stage-3 standard attention and all other pretrained operators unchanged
Initialize one CSC per stage and one UC per mixer/MLP additive branch with zero-output projections
**for** stage $s = 0, \ldots, 3$ **do**
  $h \leftarrow h + \mathcal{L}_s^R \mathcal{A}_s(\mathcal{P}_s^R h)$
  **for** block $b$ in stage $s$ **do**
    $(y, u^{\mathrm{mix}}) \leftarrow \Phi_{s,b}^{\mathrm{mix}}(h)$
    $h \leftarrow y + \mathcal{L}_s^R \mathcal{U}_{s,b}^{\mathrm{mix}}(\mathcal{P}_s^R h, \mathcal{P}_s^R u^{\mathrm{mix}})$
    $(y, u^{\mathrm{mlp}}) \leftarrow \Phi_{s,b}^{\mathrm{mlp}}(h)$
    $h \leftarrow y + \mathcal{L}_s^R \mathcal{U}_{s,b}^{\mathrm{mlp}}(\mathcal{P}_s^R h, \mathcal{P}_s^R u^{\mathrm{mlp}})$
  **end for**
  **if** $s < 3$ **then**
    $h \leftarrow \mathrm{Downsample}_{\theta,s}(h)$
  **end if**
**end for**
Calibrate and select $\phi_R$ using the common protocol of Algorithm 1

### A.5 Zero-Shot Cross-Resolution Transfer of the Learned Calibrator

Here, we investigate an additional property of the learned correction: whether a calibrator fitted at one resolution remains useful when directly transferred to unseen resolutions. This analysis is supplementary to our main setting and is not used for any headline result in the main paper.

**Protocol.** We choose a single calibration resolution $R_c = 512$ and obtain a calibrator $\phi_{512}$ using the same label-free calibration procedure as in the main experiments. After calibration, all learned parameters in $\phi_{512}$ are frozen. We then directly evaluate this same calibrator at target resolutions $R \neq 512$, without target-resolution optimization, model selection, or parameter updates.

Throughout this section, *zero-shot* refers specifically to zero-shot transfer across calibration resolutions. It does not refer to the training of the underlying ImageNet classifier. The pretrained VMamba backbone remains unchanged, and validation labels are used only for final accuracy evaluation.

For a target resolution $R$, the UC predictor continues to operate on features mapped to the fixed native-resolution stage grids. For an update site indexed by stage $s$ and block $b$, the transferred correction can be written as

$$c_{s,b}^R = \mathcal{U}_{s,b;\phi_{512}}\left(P_s^R x_{s,b}^R, P_s^R u_{s,b}^R\right), \qquad \widetilde{u}_{s,b}^R = u_{s,b}^R + L_s^R c_{s,b}^R, \tag{40}$$

where $P_s^R$ restricts the target-resolution representation to the corresponding native-resolution stage grid and $L_s^R$ lifts the predicted correction back to the target grid. Thus, the learned parameters $\phi_{512}$ remain fixed across resolutions, whereas the deterministic operators $P_s^R$ and $L_s^R$ are instantiated according to the runtime feature-grid geometry. In the full model, the deterministic SDT is also instantiated from the runtime resolution.

**Zero-shot transfer on ImageNet validation.** Table 11 reports zero-shot transfer results for VMamba-T. Raw and transferred results in this table are obtained within the same evaluation run. Although the calibrator is fitted only at $512 \times 512$, it remains useful at substantially different resolutions. At $R = 768$, the frozen $R512$ calibrator improves Top-1 accuracy from $74.666\%$ to $78.350\%$, corresponding to a gain of $3.684$ percentage points. At $R = 1280$, it improves Top-1 accuracy from $38.840\%$ to $56.896\%$, corresponding to a gain of $18.056$ percentage points.

These results should not be interpreted as showing that a single calibrator can replace target-specific calibration. In the main experiments, dedicated calibration at each target resolution remains substan-

Table 11: Zero-shot cross-resolution transfer of a calibrator fitted only at $R_c = 512$ on VMamba-T. No target-resolution calibration, model selection, or parameter update is performed. Raw and transferred results are obtained in the same ImageNet validation evaluation.

| Target resolution | Raw Top-1 | $R512$ calibrator Top-1 | Gain |
|---|---|---|---|
| 768 | 74.666 | 78.350 | +3.684 |
| 1280 | 38.840 | 56.896 | +18.056 |

Table 12: Inference-time interaction between SDT and the frozen $R512$ calibrator on the FRESH1000 diagnostic split. These are diagnostic results rather than the ImageNet validation results reported in Table 11. Output KL is the temperature-scaled distillation divergence to the frozen native-resolution prediction.

| | Top-1 (%) | | | | Output KL | |
|---|---|---|---|---|---|---|
| Resolution | Raw | SDT | RU-only | Full | RU-only | Full |
| 768 | 81.40 | 73.10 | 87.90 | 88.60 | 0.619 | 0.153 |
| 1280 | 39.30 | 28.50 | 59.50 | 59.60 | 1.097 | 0.720 |

tially stronger, particularly under extreme resolution shifts. Instead, the results show that the learned residual correction is not completely tied to the resolution at which it was fitted: a non-trivial component of the correction remains useful after a large change in spatial discretization. These results should not be interpreted as showing that a single calibrator can replace target-specific calibration. Instead, they provide empirical evidence that corrections learned at $R_c = 512$ retain partial utility after a large resolution change, consistent with, but not uniquely identifying, a partially resolution-shared correction structure. The remaining gap to target-specific calibration shows that substantial resolution-specific adjustment is still required.

**Inference-time interaction with spatial-step transport.** We next examine whether the observed transfer requires the deterministic spatial-step transport to remain active at inference time. As shown in Table 12, using the same frozen $R512$ calibrator, we compare four inference configurations on the predefined diagnostic split: the raw model, SDT only, the residual-update pathway with spatial-step transport disabled (RU-only), and the full model.

Importantly, RU-only does *not* remove all resolution-dependent geometry. It disables SDT while retaining the restriction and lifting operators $P_s^R$ and $L_s^R$ in equation 40. Therefore, this experiment isolates the inference-time contribution of SDT rather than removing the resolution-normalized residual-update construction itself.

For Top-1 accuracy, most of the zero-shot improvement remains when the spatial-step transport is disabled at inference time. At $R = 768$, RU-only achieves $87.90\%$ Top-1 compared with $88.60\%$ for the full model. At $R = 1280$, the corresponding accuracies are $59.50\%$ and $59.60\%$. Thus, in this diagnostic, most of the transferable Top-1 correction is already carried by the resolution-normalized residual-update pathway $P_s^R \to \phi_{512} \to L_s^R$.

The two configurations are nevertheless not equivalent. Restoring spatial-step transport substantially improves agreement with the native-resolution distribution: output KL decreases from $0.619$ to $0.153$ at $R = 768$ and from $1.097$ to $0.720$ at $R = 1280$. Therefore, the spatial-step transport still has a pronounced effect on the predicted distribution even when its additional Top-1 gain is small in this particular zero-shot diagnostic.

Finally, this inference-time experiment does not establish that spatial-step transport is unnecessary during calibration. The reused calibrator $\phi_{512}$ was originally learned using the full calibration pipeline, with spatial-step transport enabled. Here, that component is removed only after the calibrator has been learned. The experiment therefore characterizes an inference-time transfer property of the learned correction and should not be interpreted as a training-time ablation of the full method.

**Summary.** Although our main method uses target-specific calibration to maximize accuracy, the learned residual correction exhibits non-trivial empirical cross-resolution transferability. A calibrator fitted only at $R_c = 512$ remains useful when directly transferred as far as $R = 1280$ without any target-resolution optimization. We treat this observation as an additional empirical property of the proposed calibration framework rather than as a separate method or primary experimental setting.

### A.6 Comparison with Long-Context Extension in Language Mamba Models

**Common ground: dynamics extrapolation beyond the training horizon.** Long-context extension in language Mamba and cross-resolution transfer in visual Mamba share a common high-level issue: linear-time execution does not guarantee that pretrained recurrent dynamics extrapolate correctly beyond the training regime. Prior larger language model studies analyze this problem through effective receptive fields, memory decay, out-of-distribution discretization, and transition spectra (Ben-Kish et al., 2025; Ye et al., 2025; Azizi et al., 2025; Lu et al., 2025). In particular, MambaExtend calibrates the discretization dynamics for unseen context lengths (Azizi et al., 2025), while Mamba Modulation relates length generalization to the spectrum and convergence behavior of the recurrent state transition (Lu et al., 2025).

To illustrate the latter perspective, consider a simplified stable linear recurrence

$$h_t = \Lambda h_{t-1} + \xi_t, \tag{41}$$

where $\xi_t$ is a zero-mean driving process with stationary second-order statistics. For an eigenmode with $|\lambda_i| < 1$, the contribution of the driving process to its finite-horizon state energy contains the geometric accumulation factor

$$\sum_{k=0}^{T-1} |\lambda_i|^{2k} = \frac{1 - |\lambda_i|^{2T}}{1 - |\lambda_i|^2}. \tag{42}$$

Thus, increasing the recurrent horizon changes how strongly modes with different transition eigenvalues accumulate, decay, or approach their steady-state behavior. Equation 42 is an intentionally simplified spectral view rather than an assumption that the full selective Mamba is linear time invariant, it isolates the horizon-dependent mechanism studied by language-side length-generalization analyses.

**Resolution scaling additionally changes the spatial unit of recurrence.** Visual resolution scaling contains a longer-horizon component, but cannot be reduced to it. Increasing the resolution does not enlarge the normalized image domain. Instead, it samples the same two-dimensional domain with a denser feature lattice. Consequently, the same normalized spatial span is traversed by more recurrent steps.

Consider an ordinary local transition along route $r$ of stage $s$, excluding raster-wrap transitions. Let $s^0_{s,r}$ and $s^R_{s,r}$ denote the normalized feature-cell pitches at the pretraining resolution and target resolution, respectively, and define

$$\kappa_{s,r}(R) = \frac{s^R_{s,r}}{s^0_{s,r}}, \qquad s^R_{s,r} = \kappa_{s,r}(R) s^0_{s,r}. \tag{43}$$

For VMamba, the local selective recurrence uses an effective positive step $\bar{\Delta}^R_{t,c}$ and a stable state parameter $A_{c,j} < 0$. The corresponding local transition mode is

$$\lambda^R_{t,c,j} = \exp\left(\bar{\Delta}^R_{t,c} A_{c,j}\right). \tag{44}$$

We therefore define its log-decay rate per unit normalized spatial distance as

$$\begin{aligned} g^R_{t,c,j,\mathrm{raw}} &= \frac{-\log \lambda^R_{t,c,j}}{s^R_{s,r}} \\ &= \frac{1}{\kappa_{s,r}(R)} \underbrace{\frac{-\bar{\Delta}^R_{t,c} A_{c,j}}{s^0_{s,r}}}_{\widetilde{g}^R_{t,c,j}}. \end{aligned} \tag{45}$$

which separates a deterministic geometry factor $\kappa_{s,r}^{-1}$ from the target-resolution selective dynamics $\widetilde{g}^R$. Importantly, this is a multiplicative factorization of the local spatial rate, not an additive decomposition of the full network error.

Our VMamba Spatial Dynamics Transport (SDT) applies the route-wise transport

$$A \leftarrow \kappa_{s,r} A, \qquad B \leftarrow \kappa_{s,r} B. \tag{46}$$

For the transition-rate diagnostic, this gives

$$g^R_{t,c,j,\mathrm{SDT\text{-}CF}} = \frac{-\bar{\Delta}^R_{t,c}\kappa_{s,r}A_{c,j}}{\kappa_{s,r}s^0_{s,r}} = \widetilde{g}^R_{t,c,j}, \tag{47}$$

so the nominal lattice-spacing factor is cancelled exactly within this local diagnostic.

The same factorization holds for the magnitude of the state-injection term. For the implemented recurrence, define

$$\begin{aligned} j^R_{t,c,j,\mathrm{raw}} &= \frac{\left|\bar{\Delta}^R_{t,c}B^R_{t,j}u^R_{t,c}\right|}{s^R_{s,r}} \\ &= \frac{1}{\kappa_{s,r}(R)} \underbrace{\frac{\left|\bar{\Delta}^R_{t,c}B^R_{t,j}u^R_{t,c}\right|}{s^0_{s,r}}}_{\widetilde{j}^R_{t,c,j}}. \end{aligned} \tag{48}$$

Scaling $B$ by $\kappa_{s,r}$ therefore removes the same deterministic cell-pitch factor from the local injection rate. These exact cancellations apply to the local transition and injection rates defined above, they do not imply that SDT alone makes the complete hierarchical network resolution invariant.

**A residual input-conditioned shift remains after geometry transport.** Removing the deterministic spatial-rate factor does not require the target-resolution selective quantities to match their native-resolution counterparts. In VMamba, $\bar{\Delta}_t$ and $B_t$ are input conditioned and directly control the recurrent transition/injection, while $C_t$ is an input-conditioned state-readout quantity. The scan input $u_t$ is likewise resolution dependent. We therefore measure the magnitude statistics of $\bar{\Delta}$, $B$, $C$, and $u$ rather than assuming that a statistical shift follows mathematically from resolution scaling.

We evaluate this diagnostic on 1,000 ImageNet-1K training images, using one image from each class only to construct a balanced diagnostic set. No labels enter either the frozen model or the discrepancy metric. For each image, resolution, stage, and diagnostic quantity, we compare the target and native empirical distributions using a quantile-based normalized Wasserstein distance,

$$\mathrm{nW}_1(P_R, P_0) = \frac{W_1(P_R, P_0)}{\max\{Q_{0.95}(P_0) - Q_{0.05}(P_0), |Q_{0.50}(P_0)|, \epsilon\}}, \tag{49}$$

where $W_1$ is approximated using 101 uniformly spaced empirical quantiles. The four stage-level discrepancies are first averaged within each image, all reported summary statistics therefore use the 1,000 images as the independent experimental units.

Table 13 shows two distinct effects. First, the deterministic spatial-rate factor dominates the raw local dynamics mismatch. For log-decay rate, geometry-only transport reduces the median discrepancy by $90.1\%$, $93.7\%$, and $95.9\%$ at $448^2$, $640^2$, and $1024^2$, respectively, all 1,000 images improve at every target resolution. For the injection rate, the corresponding reductions are $75.6\%$, $85.0\%$, and $91.0\%$, with improvement on $99.1\%$, $99.7\%$, and $100\%$ of the images. The actual SDT forward closely follows the local counterfactual: its median log-decay discrepancies are $0.024$, $0.029$, and $0.037$, while its injection discrepancies are $0.157$, $0.173$, and $0.184$. This agreement indicates that the dominant first-order correction is explained by the analytic route-wise transport itself.

Second, the residual selective statistics remain systematically resolution dependent after the deterministic geometry factor is isolated. The raw discrepancies of $\bar{\Delta}$, $|B|$, $|C|$, and $|u|$ all increase from $448^2$ to $1024^2$. Because the same 1,000 images are used at every resolution, we additionally perform paired image-level comparisons. From $448^2$ to $1024^2$, the discrepancy increases for $100\%$, $100\%$, $100\%$, and $82.2\%$ of the images for $\bar{\Delta}$, $|B|$, $|C|$, and $|u|$, respectively. Thus, after removing the dominant deterministic lattice-rate factor, a smaller but systematic resolution-dependent shift remains in the input-conditioned selective statistics.

Table 13: Image-level cross-resolution dynamics diagnostic for frozen VMamba-T. Values are median normalized Wasserstein discrepancies from the native $224^2$ statistics over 1,000 images after aggregating the four stages within each image. *Local SDT-CF* applies the deterministic route-wise $A/B$ transport only inside the diagnostic while preserving the same raw target-resolution selective quantities. Reduction is the median per-image relative reduction from Raw to Local SDT-CF.

| Diagnostic | Condition | $448^2$ | $640^2$ | $1024^2$ |
|---|---|---|---|---|
| *Local physical-rate discrepancy* | | | | |
| Log-decay rate | Raw | 0.266 | 0.499 | 0.948 |
| | Local SDT-CF | 0.026 | 0.032 | 0.039 |
| | Reduction (%) | 90.1 | 93.7 | 95.9 |
| State-injection rate | Raw | 0.601 | 1.066 | 1.973 |
| | Local SDT-CF | 0.153 | 0.167 | 0.184 |
| | Reduction (%) | 75.6 | 85.0 | 91.0 |
| *Raw input-conditioned statistics* | | | | |
| $\bar{\Delta}$ | Raw vs. native | 0.025 | 0.031 | 0.038 |
| $\lvert B\rvert$ | Raw vs. native | 0.042 | 0.060 | 0.077 |
| $\lvert C\rvert$ | Raw vs. native | 0.041 | 0.060 | 0.083 |
| $\lvert u\rvert$ | Raw vs. native | 0.049 | 0.052 | 0.059 |

**Implication for cross-resolution calibration.** The above analysis clarifies the distinction from language-side long-context extension. Language length generalization can be naturally analyzed through how recurrent modes extrapolate over a longer horizon, including their spectral accumulation and state-convergence behavior. Visual resolution extrapolation contains this longer-horizon issue, but additionally changes the spatial unit over which the pretrained recurrence is executed. The latter induces a deterministic spatial-rate factor that is absent from a pure one-dimensional horizon extension. After this factor is removed, the target-resolution selective statistics still differ measurably from their native-resolution counterparts.

This residual statistical shift should not be interpreted as a complete explanation of the remaining end-to-end error. It is one measurable source of the post-transport trajectory mismatch, which can be further reshaped by nonlinear residual updates and hierarchical transitions. Accordingly, SDT addresses the predictable geometry-induced rate change, whereas CSC and UC are designed to correct the remaining carried-state and additive-update trajectory discrepancies rather than any individual selector statistic. Visual cross-resolution transfer is therefore not merely a longer-context version of language-side Mamba extrapolation, but a coupled problem of recurrent-horizon extension, geometry-induced spatial-rate change, and residual input-conditioned dynamics.

### A.7 Discussion and Future Directions

**From resolution-specific calibration to continuous adaptation.** Uniform-RU is formulated around the observation that changing the input resolution alters the spatial execution geometry of a pretrained visual Mamba model. By separating the geometry-determined transport in SDT from the learned residual corrections in CSC and UC, the framework provides a structured way to adapt the pretrained computation without modifying the backbone parameters. In the current formulation, calibration is performed for a target operating resolution. The cross-resolution transfer results in Appendix A.5, however, suggest that the learned correction captures structure that is shared across resolutions. This motivates a natural extension in which CSC and UC are conditioned directly on the stage-wise geometry $\boldsymbol{\kappa}_s$, enabling a single calibrator to operate continuously over a range of deployment resolutions.

**Beyond isotropic resolution changes.** Our analysis focuses on isotropic changes in image resolution, for which the induced spatial transformation can be characterized compactly through the stage-wise geometry ratios. The same perspective can be extended to more general changes in spatial layout. For example, anisotropic resizing and aspect-ratio variation induce direction-dependent changes in the execution geometry and can be represented through separate horizontal and vertical

factors. More broadly, the transport-and-correction view developed here provides a useful basis for studying how pretrained visual state-space models respond when their deployment geometry differs from that encountered during pretraining.

**Toward broader architecture and task coverage.** The experiments in this work cover multiple visual Mamba families and extend from image classification to dense prediction, while preserving the same overall calibration principle. Different architectures nevertheless realize spatial scanning, local mixing, hierarchy, and residual interaction in different forms. An interesting direction is therefore to derive a more unified calibration interface that abstracts these implementation-specific details while retaining architecture-aware transport where necessary. Such a formulation could further extend resolution calibration toward a general deployment-time adaptation framework for pretrained visual state-space models.