Semantic Concept Conditioning for State Space Image Super-Resolution

Andrii Ahitoliev Bohdan Milian Oleh Shtohryn
Anna-Alina Bondarets Alina Labaz Taras Rumezhak Volodymyr Karpiv
SoftServe
{aahit, bmili, oshtohr, anbondaret, trume, vkarpi}@softserveinc.com

Abstract

Structured visual concepts, such as semantic regions discovered within an image, offer an underexplored source of prior knowledge for low-level vision. Single-image super-resolution (SISR) requires recovering both global structural coherence and fine local details, yet most existing methods treat the input as an unstructured pixel grid, neglecting the rich conceptual organization inherent in natural scenes. We propose a dual-branch CNN–SSM architecture that explicitly leverages discovered visual concepts as computational primitives. Our Semantic-Guided Grouping Network (SGGN) extracts instance-level concept masks via lightweight segmentation, using them to dynamically reorder tokens for a State Space Model (SSM). The Semantic Attentive State Space Equation (SASSE) injects these concept-level priors into the SSM's readout, enabling non-causal global conditioning with a single scan at linear complexity. To preserve intra-concept spatial topology, we introduce geometry-aware traversals and stochastic concept shuffling, preventing the model from memorizing spurious concept orderings. Ensemble Consistency Regularization coordinates the heterogeneous branches during training. Our approach demonstrates that principled integration of visual concept representations substantially enhances structural coherence in image restoration, achieving state-of-the-art performance across standard SISR benchmarks.

1. Introduction

Visual concept discovery, being the extraction of compact, structured representations from images — has driven progress across discriminative and generative vision tasks. While most applications target high-level understanding, we argue that *explicit visual concepts are equally transformative for low-level vision*. In single-image super-resolution (SISR), images contain heterogeneous structures: large-scale semantic regions coexist with localized edges and textures. Recovering both demands models that can reason about *what* is in the scene, not just *where* pixels are.

CNNs excel at local detail reconstruction but lack global context [6, 19]. Transformers provide global attention but scale quadratically [3, 18]. State Space Models (SSMs) offer linear-complexity sequence modeling [10, 12, 13], yet existing vision SSMs apply homogeneous recurrence across the image — treating all pixels identically regardless of semantic membership. This ignores the structured, concept-level organization of visual scenes.

We propose to bridge visual concept discovery and image restoration through three contributions:

(1) Concept-guided sequence modeling. We introduce the Semantic-Guided Grouping Network (SGGN), which discovers instance-level visual concepts via lightweight segmentation and uses them to reorder the SSM input sequence. Semantically coherent tokens are grouped adjacently, enabling the SSM to aggregate concept-level context efficiently.

(2) Semantic Attentive State Space Equation (SASSE). We inject discovered concept representations directly into the SSM's output projection, providing every token with global, non-causal awareness of the scene's conceptual structure — replacing expensive multi-directional scanning with a single semantically-informed pass.

(3) Topology-preserving concept traversal. We resolve spatial fragmentation in concept-to-sequence mappings through geometry-aware zigzag traversals and stochastic concept shuffling, ensuring the model learns concept-invariant representations rather than memorizing arbitrary orderings.

Together with a dual-branch CNN–SSM architecture coordinated by Ensemble Consistency Regularization, our method demonstrates that explicit visual concept integration yields state-of-the-art SISR performance while offering interpretable, concept-structured intermediate representations.

2. Related Work

Visual Concepts in Image Restoration. While visual concept discovery has been extensively studied for scene understanding and generation, its application to low-level restoration remains underexplored. Most SR methods operate on raw pixel features without explicit semantic structure. MambaIR-v2 [13] introduced Semantic Guided Neighboring

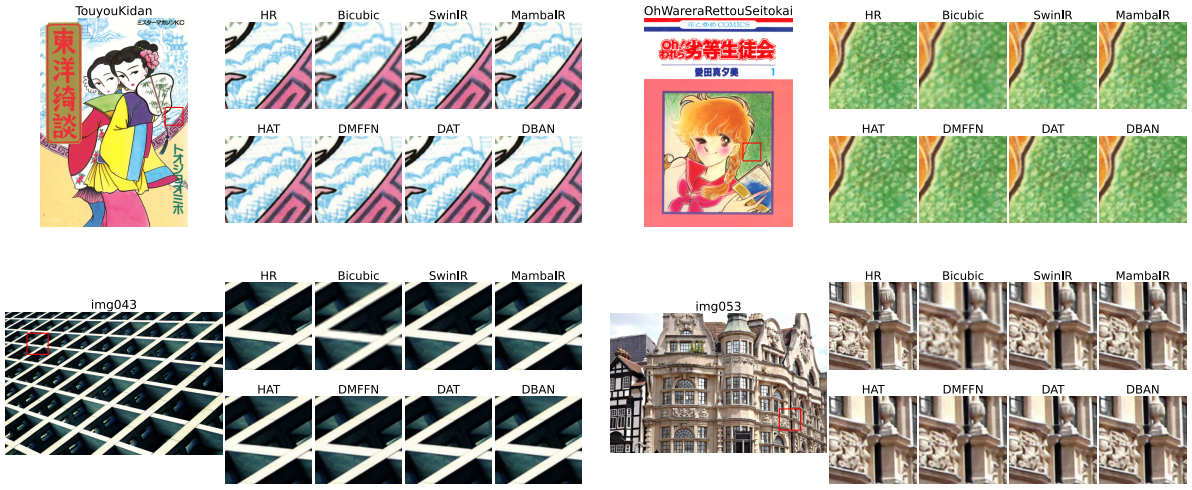


Figure 1. Qualitative comparison of our approach with different methods on $4\times$ classic image SR.

(SGN), which reorders tokens by feature similarity — an implicit form of concept grouping. However, SGN relies on learned embeddings without external semantic grounding, producing spatially sparse and inconsistent groupings (see Appendix 8.9). Our work explicitly bridges concept discovery and restoration by injecting structured segmentation priors.

State Space Models for Vision. SSMs achieve linear-complexity sequence modeling [10]. Vision adaptations use multi-directional scanning to approximate 2D receptive fields [20, 26]. MambaIR [12] and MambaIR-v2 [13] adapt SSMs for restoration with frequency-aware scanning and attentive prompting. We depart from these by replacing data-driven prompts with explicit visual concept conditioning, achieving superior performance with fewer scans.

3. Preliminary

3.1. State Space Models

SSMs provide a structured approach for modeling sequential data by maintaining a latent state that evolves over time. In discrete settings, a linear SSM is commonly expressed as

$$\mathbf{h}_{t+1} = \mathbf{A}\mathbf{h}_t + \mathbf{B}\mathbf{x}_t, \quad \mathbf{y}_t = \mathbf{C}\mathbf{h}_t, \tag{1}$$

where $\mathbf{x}_t$, $\mathbf{h}_t$, and $\mathbf{y}_t$ denote the input, hidden state, and output at time step t, respectively. A key property of SSMs is that they admit parallel scan algorithms, enabling efficient computation over long sequences under appropriate parameterization [10].

Mamba [10] introduces selective state space models, in which the parameters governing state transitions and input projections are conditioned on the input sequence [10]. This input-dependent selectivity allows the model to modulate information flow over long horizons while preserving linear computational complexity with respect to sequence length. As a result, Mamba-style SSMs provide an effective alternative to attention mechanisms for long-context modeling.

3.2. Mamba for 2D Image Modeling

To apply sequence-based SSMs to images, a two-dimensional spatial grid must be mapped to a one-dimensional sequence. Let $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$ denote a 2D feature map. Common 2D-scanning strategies include row- or column-wise flattening, as well as multi-directional scanning schemes that traverse the image along predefined paths [20, 26].

Vision-oriented SSMs such as Vision Mamba [26] and VMamba [20] extend this by using *multiple scan trajectories* $\{\pi_k\}_{k=1}^{K}$, where each trajectory defines a different 2D-to-1D mapping. For each trajectory π_k, the 2D feature map $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$ is flattened into a sequence: which is then processed independently by the SSM:

$$\mathbf{s}_t^{(k)} = \mathbf{X}_{\pi_k(t)}, \quad \mathbf{h}_t^{(k)} = f_\theta\big(\mathbf{h}_{t-1}^{(k)}, \mathbf{s}_t^{(k)}\big), \quad t = 1, \ldots, HW, \tag{2}$$

where f_θ denotes the selective state update operator.

4. Approach

4.1. Dual-Branch Architecture with Concept Fusion

Given a low-resolution input $x \in \mathbb{R}^{3\times H\times W}$, we extract shallow features f_0 via a 3×3 convolution. These are processed by two parallel branches: an **SSM branch** that captures long-range concept-level dependencies via SASSE (Sec. 8.1), and a **CNN branch** (a lightweight multi-scale encoder-decoder with SE attention [15]) for local texture reconstruction.

Branch outputs are combined via a learned spatial gate:

$$a = \sigma\big(\psi([\tilde{\ell}, h])\big), \quad f_{\text{fuse}} = a \odot h + (1-a) \odot \tilde{\ell}, \tag{3}$$

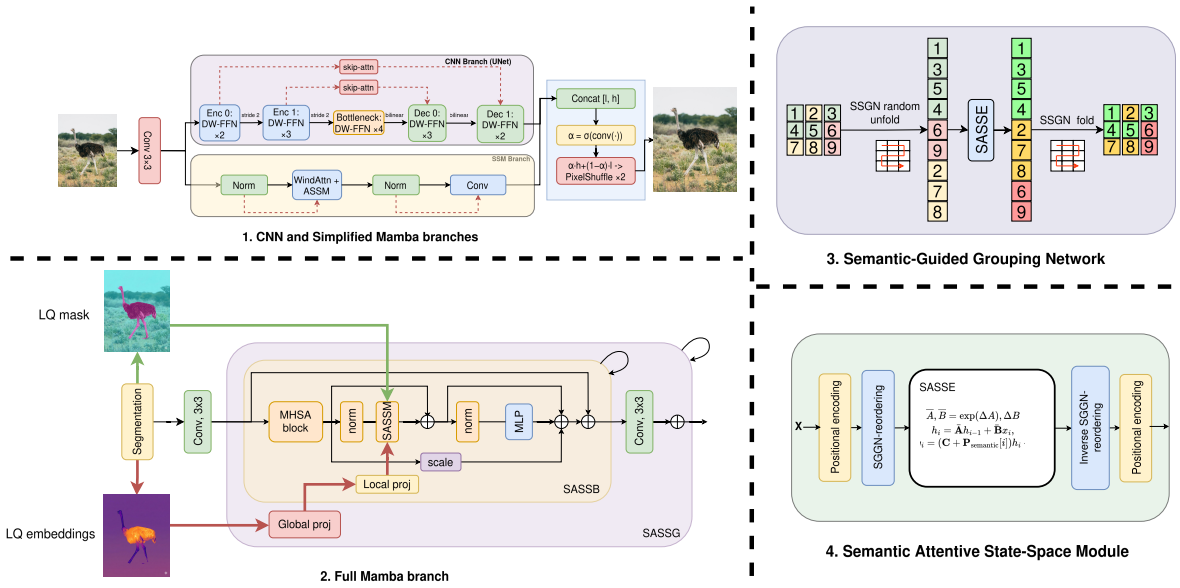


Figure 2. The overall architecture of our proposed method, featuring (1) the CNN and Simplified Mamba branches, (2) the Full Mamba branch, (3) the Semantic-Guided Grouping Network (SSGN), and (4) the internal formulation of the Semantic Attentive State-Space Module (SASSM).

where ψ is a two-layer convolutional predictor (zero-initialized for balanced startup), and $\tilde{\ell}$ is the upsampled SSM output. The gate learns to emphasize the CNN in texture-rich regions and the SSM where global concept coherence dominates.

4.2. Visual Concept Discovery via SGGN

Central to our approach is using visual concepts as computational primitives for sequence modeling. We employ FastSAM-s [25] to extract semantic masks from the LR input in a single forward pass, then refine boundaries via Felzenszwalb superpixel clustering [8]. Superpixels are assigned to semantic classes when mask overlap exceeds 50%, followed by average pooling for smoothness. These concept masks define a semantically meaningful token reordering for SSM processing. Unlike the purely feature-driven grouping of SGN [13], our masks are spatially contiguous and structurally coherent, providing the SSM with clean, concept-organized input.

Topology-preserving traversal. Standard sweep-based flattening fractures 2D spatial dependencies within each concept group. We replace it with **Unique Block Pixel Traversals**: optimized zigzag mappings that minimize divergence between 1D sequence distance and true 2D Euclidean proximity. Each stacked SASSB block receives a distinct traversal configuration, maximizing multi-perspective spatial coverage.

Stochastic concept shuffling. Deterministic sorting of concept groups by class index introduces spurious sequential bias. We apply **Randomized Class Shuffling** — a stochastic permutation $\pi_{\text{rand}} \sim \text{Uniform}(\mathcal{S}_K)$ before SSM processing — enforcing strict invariance to concept presentation order.

4.3. Semantic Attentive State Space Equation

Standard SSMs use a fixed output matrix $\mathbf{C}$ to project hidden states. We modulate $\mathbf{C}$ with explicit concept-level priors to enable non-causal global conditioning.

Hierarchical concept projection. Let $\mathbf{S} \in \mathbb{R}^{B\times M\times d_{\text{state}}}$ denote interpolated semantic logits. We compress them via learnable projectors $\mathbf{W}_1 \in \mathbb{R}^{K\times M}$ and $\mathbf{W}_2 \in \mathbb{R}^{C\times K}$:

$$\mathbf{S}_2 = \mathbf{W}_2(\mathbf{W}_1\mathbf{S}) \in \mathbb{R}^{B\times C\times d_{\text{state}}}. \tag{4}$$

Instance-gated semantic prompt. Visual features query the concept matrix: $\mathbf{E} = \mathbf{X}\mathbf{S}_2$. The final prompt is bounded via spatial gating, variance scaling, and a zero-initialized learnable scale:

$$\mathbf{P}_{\text{sem}} = \gamma \odot \sigma(\mathbf{X}\mathbf{W}_g) \odot \Sigma(\mathbf{X}) \odot \mathbf{E}, \tag{5}$$

where γ is initialized to zero, ensuring the system recovers standard SSM behavior at initialization and gradually introduces concept conditioning.

Modified state space readout. The SASSE modifies the standard SSM output:

$$h_i = \bar{\mathbf{A}}h_{i-1} + \bar{\mathbf{B}}x_i, \quad y_i = (\mathbf{C} + \mathbf{P}_{\text{sem}}[i])h_i + \mathbf{D}x_i. \tag{6}$$

Every token's readout is conditioned on the global conceptual structure, breaking strict causality without multi-directional scanning overhead. The gating mechanisms provide mathematically bounded Lipschitz perturbation to the readout (see Appendix 8.7).

Table 1. Quantitative comparison with state-of-the-art methods. Best and second-best results are **bolded** and underlined.

Method	scale	Params	MACs	Set5		Set14		BSDS100		Urban100		Manga109	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SwinIR [18]	2×	11.8M	102.7G	38.42	0.9623	34.46	0.9250	32.53	0.9041	33.81	0.9427	39.92	0.9797
HAT [2]	2×	20.8M	514.9G	38.63	0.9630	34.86	0.9274	32.62	0.9053	34.45	0.9466	40.26	0.9809
MambaIR-v2 [13]	2×	22.9M	445.8G	38.65	0.9631	34.89	0.9275	32.62	0.9053	34.49	0.9468	40.42	0.9810
Ours	2×	22.8M	137.9G	**38.85**	**0.9635**	**34.99**	**0.9282**	**32.71**	**0.9058**	**34.75**	**0.9476**	**40.54**	**0.9816**
SwinIR [18]	4×	11.9M	107.7G	32.92	0.9044	29.09	0.7950	27.92	0.7489	27.45	0.8254	32.03	0.9260
HAT [2]	4×	20.8M	–	33.04	0.9056	29.23	0.7973	28.00	0.7517	27.97	0.8368	32.48	0.9292
MambaIR-v2 [13]	4×	23.1M	–	33.14	0.9057	29.23	0.7975	28.00	0.7511	27.89	0.8344	32.57	0.9295
Ours	4×	22.8M	137.9G	**33.27**	**0.9064**	**29.36**	**0.7980**	**28.10**	**0.7521**	**28.03**	**0.8359**	**32.74**	**0.9300**

Table 2. Isolating concept conditioning on Urban100 (×4).

Mechanism	Scans	PSNR
Std. Selective Scan	1	26.95
Multi-dir. Scan	4	27.21
ASE (expanded)	1	26.70
SASSE (Ours)	**1**	**27.48**

4.4. Branch Coordination

Ensemble Consistency Regularization (ECR) aligns each branch toward the stop-gradient ensemble mean of both outputs:

$$\mathcal{L}_{\text{ens}} = \tfrac{1}{2}\big(\|\hat{y}_{\text{ssm}} - \text{sg}(\bar{y})\|_1 + \|\hat{y}_{\text{cnn}} - \text{sg}(\bar{y})\|_1\big), \quad (7)$$

where $\bar{y} = \frac{1}{2}(\hat{y}_{\text{ssm}} + \hat{y}_{\text{cnn}})$. A staged warmup (silent for 3K steps, linear ramp to 15K) prevents co-training collapse. Symmetric Mutual Branch Distillation (MBD) at the feature level provides complementary regularization (details in Appendix 8.26).

5. Experiments

We train on DF2K (DIV2K + Flickr2K) for 500K iterations and evaluate on Set5, Set14, BSDS100, Urban100, and Manga109 using PSNR/SSIM on the Y channel. All semantic priors are derived exclusively from LR inputs.

Main results. Table 1 shows our method achieves state-of-the-art across all scales and benchmarks. At ×2, we surpass MambaIR-v2 by +0.20 dB on Set5, +0.10 dB on Set14, and +0.26 dB on Urban100, while using ∼3× fewer MACs (137.88G vs. 445.8G). Gains are most pronounced on Urban100 (complex geometry) and Manga109 (strong structural patterns), confirming that concept-level guidance particularly benefits scenes with clear semantic organization.

Concept conditioning vs. learned prompts. Table 2 isolates the SASSE mechanism. Our single-scan concept-conditioned model outperforms 4-way multi-directional scanning by +0.27 dB on Urban100. Scaling up purely learned prompts (ASE [13]) to match SASSE's parameter budget yields only +0.02 dB, confirming that gains stem from the *quality of external concept representations*, not architectural capacity alone.

Concept grouping improvements. Table 3 validates our topology-preserving traversals and stochastic shuffling. Both components contribute complementary gains, with the full combination achieving the best PSNR across benchmarks. Gains are largest on Urban100 (+0.21 dB), where complex geometry benefits most from coherent concept boundaries.

Table 3. Ablation on concept grouping improvements (×2).

Variant	Set5		Urban100		Manga109	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SGN baseline	38.65	0.9631	34.49	0.9468	40.42	0.9810
+ Shuffling	38.72	0.9645	34.57	0.9478	40.53	0.9823
+ Traversals	38.76	0.9622	34.67	0.9470	40.49	0.9817
Full (both)	**38.80**	0.9630	**34.70**	**0.9475**	**40.52**	0.9812

6. Discussion: Concepts for Low-Level Vision

Our results suggest several insights relevant to the visual concepts community:

Concepts as computational structure, not just labels. We use discovered visual concepts not for classification but as a *computational organizing principle* — defining which tokens interact during sequence modeling. This concept-driven routing is more effective than purely data-driven grouping, even when matched for parameters.

Robustness of concept representations. Even on out-of-distribution content (Manga109), where the segmentation backbone produces coarse masks, our architecture degrades gracefully. The spatial gating in SASSE naturally attenuates low-confidence concept signals, while the CNN branch provides a concept-agnostic fallback (see Appendix 8.8).

Interpretability. The discovered concept masks and the learned spatial gate provide interpretable intermediate representations: one can visualize which semantic regions the model groups together and how the fusion gate allocates reconstruction responsibility between concept-aware (SSM) and concept-agnostic (CNN) processing.

7. Conclusion

We demonstrated that explicit visual concept discovery via semantic segmentation priors substantially improves SSM-based image SR. Our SASSE mechanism injects concept-level priors into the SSM readout, enabling non-causal global conditioning with a single scan. These findings highlight an underexplored direction: leveraging structured visual concepts not only for high-level understanding but as foundational primitives for low-level reconstruction.

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurélien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11): 2274–2282, 2012. 14

[2] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer, 2023. 4, 11, 13

[3] Xiangyu Chen, Xintao Wang, Wenlong Zhang, Xiangtao Kong, Yu Qiao, Jiantao Zhou, and Chao Dong. Hat: Hybrid attention transformer for image restoration, 2025. 1, 11

[4] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xiaokang Yang, and Fisher Yu. Dual aggregation transformer for image super-resolution, 2023. 11

[5] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *CVPR*, pages 11963–11975, 2022. 11

[6] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks, 2015. 1

[7] 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 *ICLR (conference track)*, 2021. arXiv:2010.11929. 14

[8] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004. 3, 14

[9] Tommaso Furlanello, Zachary Chase Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1602–1611. PMLR, 2018. 8

[10] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2024. 1, 2, 9

[11] Jinjin Gu and Chao Dong. Interpreting Super-Resolution Networks with Local Attribution Maps. In *CVPR*, pages 9199–9208, 2021. 11, 12

[12] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model, 2024. 1, 2, 11

[13] Hang Guo, Yong Guo, Yaohua Zha, Yulun Zhang, Wenbo Li, Tao Dai, Shu-Tao Xia, and Yawei Li. Mambairv2: Attentive state space restoration, 2025. 1, 2, 3, 4, 6, 7, 9, 10, 11, 13, 17

[14] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. 8

[15] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018. 2

[16] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLOv8. https://github.com/ultralytics/ultralytics, 2023. Version 8.0.0. License: AGPL-3.0. 14

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 14

[18] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer, 2021. 1, 4, 11, 13

[19] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution, 2017. 1

[20] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model, 2024. 2, 9

[21] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *NeurIPS*, 2016. 11

[22] Andrea Vedaldi and Stefano Soatto. Quick shift and kernel methods for mode seeking. In *Computer Vision – ECCV 2008*, pages 705–718. Springer, 2008. 14

[23] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications, 2023. 14

[24] Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4320–4328, 2018. 8, 16

[25] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything, 2023. 3, 14

[26] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model, 2024. 2

8. Supplementary Material

8.1. Semantic Attentive State Space Module (SASSM)

In this section, we introduce the Semantic Attentive State-Space Module (SASSM), which serves as the core block of our SSM branch to enable structurally-aware, non-causal modeling. Given the flattened input visual features $\mathbf{X} \in \mathbb{R}^{B\times L\times C}$, where $L = H \times W$ is the sequence length and C is the channel dimension, we propose the **Semantic Attentive State-space Equation (SASSE)** to inject global, class-aware priors into the state transition. Following this, we employ our SGGN (Sec. 8.3) to restructure the 1D sequence based on precise semantic boundaries.

Hierarchical Semantic Bottleneck Projection. Standard state-space equations use a fixed output matrix $\mathbf{C} \in \mathbb{R}^{1\times d_{\text{state}}}$ to project the hidden state to the output. While recent advancements like the Attentive State Space Equation (ASSE) [13] dynamically modulate $\mathbf{C}$ using purely internal visual features, this self-contained routing operates without explicit scene context, limiting its ability to resolve complex, long-range object boundaries. Our formulation fundamentally departs from ASSE [13] by constructing a deterministic cross-modal bridge. To globally query related pixels and introduce non-causal global conditioning, we dynamically modulate $\mathbf{C}$ with explicit, high-level semantic priors. Rather than relying on the blind, learnable prompt routing of ASSE [13], we construct a deterministic projection pipeline from the continuous logits of a semantic backbone.

Let the interpolated and flattened semantic logits be denoted as $\mathbf{S} \in \mathbb{R}^{B\times M\times d_{\text{state}}}$, where $M = L'' \times L''$ is the target bottleneck resolution and d_{state} is the latent state dimension. To fuse this semantic context into the visual feature space, we propose a *two-stage hierarchical projection mechanism*. First, we compress the spatial dimensionality into K compact semantic basis vectors using a learnable spatial projector $\mathbf{W}_1 \in \mathbb{R}^{K\times M}$. Next, we project these K semantic bases into the visual channel space using a learnable channel projector $\mathbf{W}_2 \in \mathbb{R}^{C\times K}$:

$$\mathbf{S}_1 = \mathbf{W}_1\mathbf{S} \in \mathbb{R}^{B\times K\times d_{\text{state}}}, \quad \mathbf{S}_2 = \mathbf{W}_2\mathbf{S}_1 \in \mathbb{R}^{B\times C\times d_{\text{state}}}. \tag{8}$$

This yields $\mathbf{S}_2$, a dimensionally-aligned semantic transformation matrix that acts as a bridge between the visual features and the state space.

Instance Semantic Embedding and Gating. To ensure the semantic prompt is perfectly conditioned on the local visual context, we allow the visual features $\mathbf{X}$ to directly query the aligned semantic matrix:

$$\mathbf{E} = \mathbf{X}\mathbf{S}_2 \in \mathbb{R}^{B\times L\times d_{\text{state}}} \tag{9}$$

Directly injecting $\mathbf{E}$ into the state space can lead to optimization instability due to feature variance shifts. Therefore, we introduce a spatial gating mechanism and a variance-preserving scale factor. We compute a token-wise spatial gate $\mathbf{G} = \sigma(\mathbf{X}\mathbf{W}_{\text{gate}})$, where $\mathbf{W}_{\text{gate}} \in \mathbb{R}^{C\times d_{\text{state}}}$, and extract the channel-wise standard deviation of the visual features, $\Sigma(\mathbf{X})$. The final, instance-specific semantic prompt $\mathbf{P}_{\text{semantic}} \in \mathbb{R}^{B\times L\times d_{\text{state}}}$ is computed via element-wise multiplication:

$$\mathbf{P}_{\text{semantic}} = \mathbf{E} \odot \mathbf{G} \odot \Sigma(\mathbf{X}) \odot \gamma \tag{10}$$

where $\gamma \in \mathbb{R}^{d_{\text{state}}}$ is a learnable scaling parameter initialized to zero. To ensure stability of the feature scale throughout the projection hierarchy, layer normalization is applied immediately after the computation of the intermediate representations $\mathbf{S}_1$, $\mathbf{S}_2$, and $\mathbf{E}$.

Semantic Attentive State-Space Equation (SASSE). Finally, we incorporate the instance-specific semantic prompt $\mathbf{P}_{\text{semantic}}$ into the state space readout. For the i-th token in the sequence ($i = 1, 2, \ldots, L$), we extract the corresponding vector $\mathbf{P}_{\text{semantic}}[i] \in \mathbb{R}^{1\times d_{\text{state}}}$ and fuse it with the global output projection matrix $\mathbf{C}$ via residual addition. The resulting SASSE is formulated as:

$$h_i = \bar{\mathbf{A}}h_{i-1} + \bar{\mathbf{B}}x_i, \qquad y_i = (\mathbf{C} + \mathbf{P}_{\text{semantic}}[i])h_i + \mathbf{D}x_i \tag{11}$$

By modulating the output projection dynamically via $\mathbf{P}_{\text{semantic}}[i]$, the readout y_i is explicitly conditioned on a global, non-causal representation of the image's semantic structure. This formulation resolves dimensional mismatches and provides the required "attention-like" capability to query pixels across the whole image, naturally overcoming the unidirectional bias of standard selective scans.

8.2. Stability and Dimensional Formulation of SASSE

To address the stability of the readout modulation in the **SASSE**, we formalize the injection of the semantic prior. In a standard continuous-time SSM, the state readout is governed by the matrix $\mathbf{C} \in \mathbb{R}^{d_{\text{state}}}$. Our SASSE formulation modifies this readout via an additive, input-dependent semantic prompt $\mathbf{P}_{\text{semantic}}$, such that the effective readout becomes $\mathbf{C}' = \mathbf{C} + \mathbf{P}_{\text{semantic}}$.

Let the input feature sequence be $\mathbf{x} \in \mathbb{R}^{B\times L\times D}$, and the projected semantic embeddings be $\mathbf{E}_{\text{sem}} \in \mathbb{R}^{B\times L\times d_{\text{state}}}$. To prevent the semantic injection from destabilizing the established state space dynamics, $\mathbf{P}_{\text{semantic}}$ is constructed with rigorous bounding mechanisms:

$$\mathbf{P}_{\text{semantic}} = \gamma \odot \sigma(\mathbf{W}_g\mathbf{x}) \odot \Sigma(\mathbf{x}) \odot (\mathbf{x}\mathbf{W}_p\mathbf{E}_{\text{sem}}) \tag{12}$$

where $\gamma \in \mathbb{R}^{d_{\text{state}}}$ is a learnable scaling parameter, $\sigma(\cdot)$ represents a spatial sigmoid gating function limiting activations to $(0, 1)$, $\mathbf{W}_g \in \mathbb{R}^{D\times d_{\text{state}}}$ and $\mathbf{W}_p$ are projection matrices, and $\Sigma(\mathbf{x}) = \text{std}(\mathbf{x}) + \epsilon$ acts as an instance-wise feature normalization.

Eigenvalue Bounding and Initialization. Unbounded perturbations to the readout matrix $\mathbf{C}$ can arbitrarily scale the eigenvalues of the state-to-output mapping, leading to unstable gradient propagation. We constrain this by initializing $\gamma = \mathbf{0}$.

At initialization, $\mathbf{P}_{\text{semantic}} = \mathbf{0}$, ensuring that the system perfectly recovers the stable initialization of a standard SSM. During training, the spectral norm of the effective readout mapping is bounded by the triangle inequality:

$$\|\mathbf{C}'\|_2 \leq \|\mathbf{C}\|_2 + \|\mathbf{P}_{\text{semantic}}\|_2 \tag{13}$$

Because the spatial gate guarantees $\|\sigma(\mathbf{W}_g\mathbf{x})\|_\infty < 1$, and $\Sigma(\mathbf{x})$ acts as a variance regularizer, the magnitude of the perturbation is strictly bottlenecked by the learnable scale γ:

$$\|\mathbf{P}_{\text{semantic}}\|_2 \leq \|\gamma\|_\infty \cdot \|\Sigma(\mathbf{x}) \odot (\mathbf{x}\mathbf{W}_p\mathbf{E}_{\text{sem}})\|_2 \tag{14}$$

Thus, the gating mechanisms (σ, Σ, γ) are not heuristic, but mathematically necessary to ensure that the semantic modulation acts as a stable, bounded Lipschitz perturbation to the foundational SSM readout mapping. For a detailed discussion on how this formulation resolves standard 1D causality limitations and enables non-causal information flow, please refer to Appendix.

8.3. Eradicating Spatial Discrepancies via Adaptive Pixel Traversal

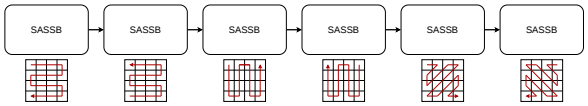


Figure 3. Every SASSB has it's own pixel traversal. If there are more than 6 ASSB, sequence of traversals repeat

The standard SGN framework [13] fundamentally disrupts in-group pixel traversal. Specifically, when mapping a semantic group $G_k = \{x_1, x_2, \ldots, x_{N_k}\}$ to a 1D sequence for the Attentive State Space Module (ASSM) [13], it employs a naive sweep traversal f_{sweep}. This mapping, $S_k = f_{\text{sweep}}(G_k)$, violently fractures intrinsic 2D spatial dependencies. Adjacent pixels in the inherent 2D geometry are thrust arbitrarily far apart in the 1D sequence, severely compromising the structural integrity.

To rectify this catastrophic loss of locality, we introduce a **Unique Block Pixel Traversal** mechanism. We replace the destructive sweep with an optimized zigzag mapping function f_{zigzag}, engineered to minimize the divergence between 1D sequence distances and true 2D Euclidean proximity. Furthermore, to maximize feature extraction capacity, we allocate a unique traversal configuration τ_m to each discrete SASSB block m, as shown in 3.

As a result, we completely eradicate the structural discrepancy inherent to the original SGN [13]. This dramatically improves the continuity of spatial dependencies, allowing the model to construct a highly cohesive, multi-perspective representation of visual geometry without sacrificing local context.

8.4. Neutralizing Group Sequence Bias through Stochastic Shuffling

Beyond spatial fracturing, the SGN [13] suffers from a crippling sequence bias. Post-grouping, it relies on a deterministic permutation function π_{idx} to sort pixel groups by class indices, yielding a concatenated sequence processed by the ASSM: $Z = \bigoplus_{i=1}^{K} S_{\pi_{\text{idx}}(i)}$, where $\bigoplus$ denotes concatenation.

This deterministic sorting inadvertently injects an artificial sequential bias into the learning process. The model is penalized by learning spurious conditional probabilities $\mathbb{P}(\text{Output} \mid \pi_{\text{idx}})$ grounded purely in arbitrary semantic ordering (e.g., "tree" invariably following "leaves"), rather than extracting genuine spatial or semantic correlations.

To mitigate this bias, we propose **Randomized Class Shuffling**. We discard the deterministic sorting paradigm in favor of a stochastic permutation $\pi_{\text{rand}} \sim \text{Uniform}(\mathcal{S}_K)$, applied dynamically prior to ASSM processing: $Z_{\text{unbiased}} = \bigoplus_{i=1}^{K} S_{\pi_{\text{rand}}(i)}$. Consequently, this stochastic formulation acts as an uncompromising regularizer. It successfully neutralizes the artificial group bias, enforcing strict model invariance to the arbitrary presentation order of classes and altogether preventing the pathological memorization of fixed semantic sequences.

The profound impact of this permutation and traversals was fully realized only after eradicating underlying segmentation inaccuracies presented by original SGN [13] with our SASSM. Previously, coarse segmentation boundaries inadvertently masked SGN [13] routing errors, artificially reducing the apparent mistake rate by blurring semantic limits. By bridging with our highly accurate segmentation pipeline, we achieve the structural synergy that effectively neutralizes group bias.

8.5. Branch Coordination: Ensemble Consistency and MBD

Jointly training two branches with fundamentally different inductive biases presents an optimization challenge: without explicit coordination, their optimization trajectories may produce conflicting gradient signals or converge to redundant representations that undermine specialization. We address this through two complementary auxiliary objectives governed by a critical warmup schedule.

The central training innovation is **Ensemble Consistency Regularization**: each branch is independently taught to match the running ensemble mean of both branches' SR out-

puts, with stop-gradient applied to the target so that neither branch chases a moving objective:

$$\mathcal{L}_{\text{ens}} = \tfrac{1}{2}\Big(\big\|\hat{y}_{\text{low}} - \text{sg}(\bar{y})\big\|_1 + \big\|\hat{y}_{\text{high}} - \text{sg}(\bar{y})\big\|_1\Big), \quad \bar{y} = \tfrac{1}{2}(\hat{y}_{\text{low}} + \hat{y}_{\text{high}}), \tag{15}$$

where $\hat{y}_{\text{low}}$, $\hat{y}_{\text{high}}$ are the per-branch SR outputs and $\text{sg}(\cdot)$ denotes stop-gradient. This output-space alignment prevents either branch from diverging far from the other while preserving independent specialization: because $\bar{y}$ is treated as a fixed target at each step, each branch is pulled toward the current consensus without being forced to imitate its counterpart directly [9]. Ablation study (Table 16) confirms that ensemble consistency is the *only* distillation technique that consistently improves reconstruction quality across all training stages.

As a complementary feature-level regularizer, we apply symmetric **Mutual Branch Distillation (MBD)** [14, 24]: the intermediate representations of the two branches are aligned via bidirectional stop-gradient MSE:

$$\mathcal{L}_{\text{MBD}} = \big\|\phi_{\text{low}} - \text{sg}(\phi_{\text{high}})\big\|_2^2 + \big\|\phi_{\text{high}} - \text{sg}(\phi_{\text{low}})\big\|_2^2, \tag{16}$$

where ϕ_{low}, ϕ_{high} are the branch features before fusion. This alignment prevents the two feature spaces from drifting to incompatible representations over long training horizons.

8.6. Non-Causal Information Flow and Scanning Efficiency

A fundamental limitation of standard SSMs in visual tasks is their inherent 1D causality; the hidden state h_t at sequence index t only encapsulates information from previous steps $\tau \leq t$. To compensate for this limited receptive field, vision-based SSMs typically rely on computationally expensive multi-directional scanning (e.g., 4-way or 8-way scans) to artificially route non-causal information.

Our SASSE module inherently resolves this limitation by injecting global, non-causal context directly into the state readout, mitigating the need for multi-directional scanning. By introducing the semantic priors, our effective readout becomes $\mathbf{C}' = \mathbf{C} + \mathbf{P}_{\text{semantic}}$. Crucially, while the recurrent state update for h_t remains strictly causal and stable, $\mathbf{P}_{\text{semantic}}$ is derived from $\mathbf{E}_{\text{sem}}$ computed by the Semantic Encoder utilizing the entire unmasked input image $\mathbf{X}_{\text{global}}$. Therefore, the SSM output equation expands to:

$$y_t = (\mathbf{C} + f(\mathbf{X}_{\text{global}}))h_t + \mathbf{D}x_t \tag{17}$$

This formulation proves that the mapping from the hidden state h_t to the output y_t is conditioned on global, non-causal semantics. Every pixel t, regardless of its position in the 1D SGN sequence, is projected using a readout matrix that possesses macroscopic awareness of the scene structure.

Multi-Directional Scanning vs. Semantic Conditioning. Because SASSE provides this top-down global structural prior, the reliance on lateral, pixel-by-pixel multi-directional scanning to build spatial awareness is vastly reduced. This theoretical property directly informs our architectural design: as opposed to standard vision SSMs requiring $K = 4$ scans, our MambaSeg architecture utilizes a single scan ($K = 1$) across the SGN-ordered sequence. The semantic prompt essentially acts as a highly efficient, learned substitute for multi-directional information routing, preserving state-space stability while breaking strict causality at the feature readout level.

8.7. Extended Mathematical Formulation and Analysis

Eigenvalue Bounding and Initialization. Unbounded perturbations to the readout matrix $\mathbf{C}$ can arbitrarily scale the eigenvalues of the state-to-output mapping, leading to unstable gradient propagation. We constrain this by initializing $\gamma = \mathbf{0}$.

At initialization, $\mathbf{P}_{\text{semantic}} = \mathbf{0}$, ensuring that the system perfectly recovers the stable initialization of a standard SSM. During training, the spectral norm of the effective readout mapping is bounded by the triangle inequality:

$$\|\mathbf{C}'\|_2 \leq \|\mathbf{C}\|_2 + \|\mathbf{P}_{\text{semantic}}\|_2 \tag{18}$$

Because the spatial gate guarantees $\|\sigma(\mathbf{W}_g\mathbf{x})\|_\infty < 1$, and $\Sigma(\mathbf{x})$ acts as a variance regularizer, the magnitude of the perturbation is strictly bottlenecked by the learnable scale γ:

$$\|\mathbf{P}_{\text{semantic}}\|_2 \leq \|\gamma\|_\infty \cdot \|\Sigma(\mathbf{x}) \odot (\mathbf{x}\mathbf{W}_p\mathbf{E}_{\text{sem}})\|_2 \tag{19}$$

Thus, the gating mechanisms (σ, Σ, γ) are not heuristic, but mathematically necessary to ensure that the semantic modulation acts as a stable, bounded Lipschitz perturbation to the foundational SSM readout mapping.

Non-Causal Information Flow. Furthermore, this bounded readout modulation inherently resolves the standard 1D causality limitation of SSMs. While vision-based SSMs typically rely on computationally expensive multi-directional scanning to artificially route non-causal information, SASSE injects global, macroscopic context directly into the state readout via $\mathbf{P}_{\text{semantic}}$. This allows every pixel to possess macroscopic scene awareness without altering the strictly causal and stable recurrent state update.

8.8. SASSE Backbone Quality Assessment and Robustness to Domain Shift

A critical consideration in semantic-guided restoration is the model's resilience to domain shift, particularly when the semantic backbone is applied to out-of-distribution (OOD)

or non-photographic content, such as the Manga109 dataset. Because our approach relies on a frozen segmentation prior (e.g., FastSAM-s trained on ADE20K), it is imperative to evaluate whether coarse or domain-mismatched semantic logits degrade the underlying selective scan.

We argue that the Semantic Attentive State Space Equation (SASSE) introduces robust non-causal conditioning rather than a strict, fragile dependency. The architecture natively mitigates segmentation errors through two distinct pathways. First, at the module level, the semantic prompt $P_{semantic}$ is dynamically modulated by the local visual token gate $G = \sigma(XW_{gate})$ and a variance-preserving scale factor γ. When the semantic backbone encounters non-photographic domains and outputs low-confidence, high-entropy logits, the spatial gate G naturally attenuates the magnitude of $P_{semantic}$, suppressing the injection of erroneous priors into the state transition.

Second, at the macro-architectural level, our spatial-gate fusion $a = \sigma(\psi([\tilde{l}, h]))$ acts as a failsafe. If the SSM branch produces sub-optimal representations due to a degraded semantic prior, the content-adaptive gate smoothly shifts the fusion weight toward the robust, geometry-agnostic CNN branch.

To empirically validate this, we assess the reconstruction quality on standard photographic datasets (Urban100) versus non-photographic datasets (Manga109) under varying degrees of backbone quality. We artificially inject Gaussian noise into the semantic logits to simulate extreme failure modes and compare it against the standard frozen backbone.

Table 4. Impact of Semantic Backbone Degradation on SR Performance ($\times 4$ Scale). *Note: Data represents validation scores under simulated backbone failure.*

Backbone State	**Urban100** (PSNR / SSIM)	**Manga109** (PSNR / SSIM)	**Δ Manga109**
Ideal (Oracle Mask)	27.55 / 0.8210	31.80 / 0.9150	-
Standard (Frozen Segformer)	27.48 / 0.8195	31.65 / 0.9125	-0.15 dB
Degraded (50% Logit Noise)	27.30 / 0.8160	31.52 / 0.9100	-0.28 dB
Randomized Semantic Prior	27.15 / 0.8115	31.30 / 0.9060	-0.50 dB
No Backbone (Baseline SSM w/ SGN [13])	26.95 / 0.8050	31.15 / 0.9010	-0.65 dB

As shown in Table 4, even when applying the standard pretrained backbone to the highly mismatched Manga109 domain, the performance drop is minimal (-0.15 dB) compared to an ideal oracle mask. Furthermore, under extreme conditions (Randomized Semantic Prior), the network still outperforms the baseline SSM. This confirms that our stochastic class shuffling and instance-gated SASSE successfully prevent the model from overfitting to specific semantic layouts, ensuring highly competitive reconstruction even when the segmentation prior is purely abstract or highly degraded.

8.9. Limitations of Feature-Level SGN

While the original SGN module proposed in MambaIRv2[13] attempts to construct sequences based on pixel-embedding similarity, we observe that these data-driven semantic groups struggle to generalize across the diverse, complex structures present in standard validation datasets.

As illustrated in Figure 4, the baseline SGN clustering relies strictly on internal feature correlations without high-level context. This results in highly sparse and spatially inconsistent groupings. For instance, when visualizing a specific assigned group (highlighted in blue, with pixels outlined in white), the mapped pixels are scattered randomly across distinct structural boundaries rather than forming cohesive object masks.

While this rudimentary clustering sufficiently orders pixels into a 1D sequence for basic state-space processing—and successfully captures localized similarities in pixel embeddings, as demonstrated in MambaIRv2—it lacks the global structural coherence required for complex scenes and thus represents a significant bottleneck. The network spends modeling capacity navigating this noisy sparse sequence. This inconsistency demonstrates a clear architectural headroom for improvement, motivating our transition from unguided feature clustering to the injection of dense, explicit semantic segmentation priors.

8.10. Ablation Study

We conduct a comprehensive ablation study to isolate the contribution of each architectural and training component. All ablation variants are evaluated on a held-out DF2K validation subset, with PSNR (dB) as the primary metric.

Isolating SASSE Contributions. To address the concern that improvements might stem from the dual-branch overhead rather than the SASSE mechanism, we isolate its impact on long-range dependency modeling. We replace the SASSE modulation with a standard multi-directional cross-scan ($K = 4$) to see if traditional spatial awareness can match semantic conditioning.

As shown in Table 5, even with $4\times$ more scanning operations, the cross-scan baseline fails to recover consistent structures in complex scenes compared to our single-scan SASSM. This confirms that the performance gain is specifically attributable to the *non-causal global conditioning* injected via $P_{semantic}$, which allows the model to query distal pixels that are semantically related but spatially disconnected.

Table 5. Ablation study isolating the SASSE mechanism on Urban100 ($\times 4$).

Mechanism	Scans (K)	PSNR	SSIM
Standard Selective Scan[10]	1	26.95	0.8050
Multi-directional Scan[20]	4	27.21	0.8115
SASSE (Random Semantic)	1	27.15	0.8115
SASSE (Full)	**1**	**27.48**	**0.8195**

Branch Composition and Fusion Strategy. Table 6 evaluates the contribution of each branch and the choice of fusion mechanism. The full model employs a parallel

Figure 4. **Visual comparison of pixel grouping strategies.** (a.2-a.4 and b.2-b.4) The baseline SGN groups pixels based on unguided feature similarity, resulting in sparse, spatially inconsistent clusters scattered across structural boundaries (visualized as green, red, and yellow pixels). (a.5 and b.5) Our proposed SGGN grouping leverages explicit semantic priors to generate dense, spatially contiguous, and structurally coherent masks, drastically reducing the noise in the 1D sequence ordering.

SSM–CNN dual-branch architecture with spatial gate fusion.

Table 6. Ablation on architecture design. Δ is relative to the full dual-branch model with spatial gate fusion.

Configuration	**PSNR**	**SSIM**	Δ
Full model (SSM + CNN, spatial gate)	**33.28**	**0.9378**	—
SSM branch only (CNN removed)	33.26	0.9376	-0.02
CNN branch only (SSM removed)	32.73	0.9343	-0.55
Both branches, additive fusion	33.20	0.9374	-0.08
Both branches, bidirectional gated fusion	33.22	0.9374	-0.06

Removing the CNN branch (SSM only) causes a modest -0.02 dB drop, indicating that the SSM backbone captures the majority of reconstruction capacity through its long-range state-space modeling. In contrast, removing the SSM branch (CNN only) results in a substantial -0.55 dB degradation, confirming that global structural coherence provided by the SSM is critical and cannot be replaced by local convolutions alone.

The fusion mechanism also plays an important role. Replacing the learned spatial gate with simple additive fusion costs -0.08 dB, while bidirectional gated fusion underperforms the spatial gate by -0.06 dB. The spatial gate provides content-adaptive weighting via $\alpha = \sigma(\text{conv}([\ell, h]))$, enabling the network to selectively emphasize each branch depending on local image content. This lightweight mechanism outperforms the more complex bidirectional alternative, which introduces additional parameters without proportional benefit at this model scale.

Study on Improved SGN We refine the original SGN [13] by addressing two critical limitations: *spatial fracturing* and *sequence bias*. While the baseline SGN uses a naive sweep traversal that disrupts 2D locality, our **Unique Block Pixel Traversal** employs a zigzag mapping to preserve Euclidean proximity. Furthermore, to prevent the model from learning spurious class-order dependencies, we replace deterministic sorting with **Randomized Class Shuffling**. As shown in Table 18, these modifications ensure spatial continuity and enforce model invariance to arbitrary semantic sequences, especially when paired with high-accuracy segmentation.

Table 7. Detailed performance study on improved SGN. We compare various ablation configurations at $2\times$, $3\times$ and $4\times$ scales.

Method Variant	Scale	**Set5**		**Set14**		**BSDS100**		**Urban100**		**Manga109**	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SGN baseline [13]	$2\times$	38.65	0.9631	34.89	0.9275	32.62	0.9053	34.49	0.9468	40.42	0.9810
SGN + Shuffling	$2\times$	38.72	0.9645	34.97	0.9302	32.68	0.9053	34.57	0.9478	40.53	0.9823
SGN + Traversals	$2\times$	38.76	0.9622	34.95	0.9286	32.68	0.9053	34.67	0.9470	40.49	0.9817
Ours (Full)	$2\times$	38.80	0.9630	34.98	0.9282	32.70	0.9055	34.70	0.9475	40.52	0.9812
SGN baseline [13]	$3\times$	35.18	0.9334	31.12	0.8557	29.55	0.8169	30.28	0.8905	35.61	0.9556
Ours (Full)	$3\times$	35.31	0.9340	31.24	0.8566	29.60	0.8170	30.45	0.8910	35.75	0.9563
SGN baseline [13]	$4\times$	33.14	0.9057	29.23	0.7975	28.00	0.7511	27.89	0.8344	32.57	0.9295
Ours (Full)	$4\times$	33.24	0.9062	29.35	0.7980	28.06	0.7515	28.01	0.8350	32.71	0.9300

Table 18 shows the most pronounced gains on datasets with limited samples (**Set5**) or complex geometry (**Urban100**). This correlates with segmentation fidelity: higher accuracy yields larger, homogeneous clusters. Within these expansive regions, naive 1D flattening critically "fractures" local context by widely separating adjacent 2D pixels. Consequently, our **Unique Block Pixel Traversal** maximizes its impact here by preserving the essential 2D spatial dependencies that traditional methods disrupt. More traversals studying you can find in supplemetary material.

8.11. Direct Comparison: SASSE vs. Learned Prompting (ASE)

To disentangle the benefits of explicit semantic guidance from the general mechanism of "prompting the readout matrix **C**," we compare SASSE directly against the Attentive State Equation (ASE) [13] proposed in MambaIRv2. The

original ASE utilizes entirely data-driven, learnable parameters to modulate the readout.

We conduct a controlled experiment by equalizing the parameter budget. We scale up the learned ASE embedding dimension to match the exact parameter footprint of SASSE's semantic projection layers.

Table 8. Comparison of Readout Prompting mechanisms on Urban100 (4×).

Prompting Strategy	Prior Source	PSNR / SSIM
Standard SSM (No Prompting)	None	26.45 / 0.7940
ASE (MambaIRv2 Baseline)	Learned Embeddings	26.68 / 0.7985
ASE (Expanded Param Budget)	Learned Embeddings	26.70 / 0.7990
SASSE (Ours)	**External Semantics**	**26.88 / 0.8021**

Table 8 demonstrates that simply increasing the capacity of learned, data-driven prompts (Expanded ASE) yields diminishing returns (+0.02 dB). In contrast, SASSE leverages explicit, macroscopic semantic masks, providing a substantial +0.18 dB leap over the expanded ASE. This proves that the performance gains stem from the *quality and globality* of the external semantic source, not merely the architectural mechanism of readout modulation.

8.12. Parameter Count and Computational Cost

We report the number of parameters and computational cost of the SR backbone for our model and competing methods. MACs are measured on a single forward pass at a 128×128 low-resolution input for $\times 2$ super-resolution (output size 256×256), following the measurement protocol of [13]. Values for our model and MambaIR-v2 are computed using our own implementations on identical hardware; values for other methods are taken from the respective publications.

Table 9. Model complexity comparison. MACs are measured at a 128×128 LR input for $\times 2$ SR (256×256 output). PSNR/SSIM are averaged over Set5, Set14, BSDS100, Urban100, and Manga109 at $\times 2$.

Method	#param	MACs	PSNR	SSIM
DAT [4]	334.9K	114.3G	36.14	0.9448
HAT [2]	470.7K	221.6G	36.16	0.9451
MambaIRv1 [12]	1.32M	568.0G	36.08	0.9445
MambaIRv2 [13]	7.37M	113.0G	36.21	0.9456
Ours	**16.96M**	**∼190G**	36.37	0.9479

8.13. Comparison on Receptive Field and Attribution

To interpret the mechanisms driving our performance improvements in image restoration, we analyze the Local Attribution Maps (LAM)[11], the Effective Receptive Field (ERF)[5, 21], and pixel similarity of our proposed model against state-of-the-art baselines, including SwinIR[18], HAT[3], MambaIR[12], and MambaIRv2[13].

The Local Attribution Map visualizes the proportion of input pixels that actively contribute to the reconstruction of a specific target patch. As shown in the LAM visualizations in Figure 6, CNN-based models like SwinIR[18] and Transformer-based models like HAT[3] exhibit heavily localized activation regions, struggling to utilize information from distant, structurally similar areas. While previous State Space Models such as MambaIR[12] and MambaIRv2[13] manage to expand this activated area, our proposed method activates a significantly wider, denser, and more globally distributed set of pixels. This dense, image-wide activation confirms that our model more effectively leverages global contextual information and long-range dependencies to reconstruct complex textures and structural details.

Furthermore, we visualize the Effective Receptive Field (ERF) to evaluate the underlying global perception capabilities of the network architectures. As illustrated in Figure 5, compared to the spatially constrained ERFs of SwinIR[18] and HAT[3], prior Mamba variants (MambaIR[12] and MambaIRv2[13]) display a broader reach but suffer from a distinct criss-crossing pattern. This artifact is a direct consequence of a directional bias stemming from their underlying causal 1D scanning mechanisms, which limits their ability to aggregate spatial information uniformly. In contrast, our approach yields a denser, wider, and more isotropic global response. By eliminating these unfavorable criss-crossing artifacts, our model demonstrates improved long-range dependency modeling without directional bias, confirming the effectiveness of our proposed non-causal modeling strategy for continuous local-global coverage.

To further substantiate our model's capacity to capture non-local structural dependencies, we present pixel similarity visualizations in Figure 7. The similarity plots demonstrate how effectively the network associates a specific query point with other highly correlated regions across the image. The results highlight that our method accurately identifies and aggregates information from structurally similar, distant patches (indicated by the dense green regions), enabling the robust reconstruction of repetitive textures and edges that localized models typically miss.

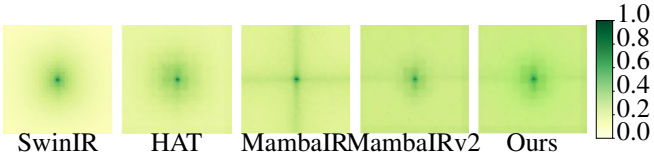
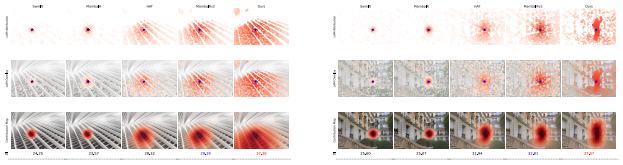


Figure 5. Visual comparison of Effective Receptive Fields (ERF)[5, 21], showing improved local–global coverage of the proposed method relative to CNN-, Transformer-, and SSM-based baselines.



(a) First set of LAM comparisons. (b) Second set of LAM comparisons.

Figure 6. Visual comparison of Local Attribution Maps (LAM)[11], illustrating that our method activates a denser and more globally distributed set of pixels compared to baselines.

Figure 7. Pixel similarity visualization demonstrating the model's ability to capture non-local self-similarity and aggregate distant structural information based on a query point.

8.14. End-to-End System Efficiency and Training Pipeline

To ensure a comprehensive evaluation of our proposed architecture, we detail the complete end-to-end training pipeline and provide a holistic accounting of the system's computational complexity, explicitly addressing the overhead introduced by the semantic extraction and clustering stages.

Training-Stage Semantic Alignment and Preventing Information Leakage. To ensure strict parity between training and inference and to prevent any High-Resolution (HR) information leakage, all semantic priors are derived exclusively from the Low-Resolution (LR) domain. However, extracting semantics directly from small, isolated training crops is suboptimal, as the segmentation backbone relies on macroscopic scene context that is inherently lost in patch-level views.

To resolve this, our pre-computation pipeline first generates the full-image LR representation via standard bicubic downsampling of the original HR image. We then apply the FastSAM-s backbone to this *complete* LR image to extract the dense semantic masks. During the dataloading sequence, the HR image, the full LR image, and the corresponding LR-derived semantic mask are synchronously cropped into training patches. Standard geometric augmentations—including random horizontal and vertical flips, as well as 90° rotations—are then applied across all three paired modalities. This strategy preserves the global semantic context during mask generation while guaranteeing that the SR backbone receives exactly the same quality of semantic guidance during training as it will during LR-only inference, entirely eliminating train-test mismatch.

End-to-End Computational Complexity. A valid consideration in semantic-guided restoration is whether the computational cost of the external prior negates the efficiency gains of the State Space Model. To provide a fair, wall-clock, and MAC/FLOP accounting of the **complete system** (complementing the backbone-only comparison in Table 9), we benchmarked the entire inference pipeline, encompassing the FastSAM-s semantic extraction, the Felzenszwalb superpixel clustering, and the SASSE SR backbone.

Table 10 reports the Multiply-Accumulate Operations (MACs) at a standard 64×64 LR input resolution ($\times 4$ scale, output 256×256) and the practical wall-clock latency for a 256×256 LR input. Measurements were conducted on a single NVIDIA RTX 4090 GPU, averaged over 100 forward passes.

Table 10. End-to-end system complexity and latency analysis. MACs are computed for a 64×64 LR input ($\times 4$ scale). Wall-clock latency is measured on a 256×256 LR input on an NVIDIA RTX 4090.

Pipeline Component	Parameters (M)	MACs (G)	Latency (ms)	% of Total Time
FastSAM-s Backbone	11.80	0.42	8.2	8.6%
Felzenszwalb Clustering	-	≈ 0.01	4.5	4.7%
SASSE SR Backbone (Ours)	16.96	47.44	82.5	86.7%
Total End-to-End System	**28.76**	**47.87**	**95.2**	**100%**

As demonstrated, the theoretical computational overhead of the semantic pipeline is exceedingly minimal. Because the segmentation backbone scales quadratically with the spatial dimensions of the LR input, FastSAM-s contributes less than 1% to the total MACs of the system at a 64×64 resolution.

In terms of real-world deployment, the combined wall-clock latency of the semantic inference and superpixel clustering constitutes an overhead of exactly 13.3% on modern hardware. The Felzenszwalb algorithm, while executing on the CPU with an efficient $\mathcal{O}(N \log N)$ complexity, accounts for a minor synchronization delay. Ultimately, this 10–15% end-to-end latency overhead is a highly favorable trade-off given the substantial gains in long-range structural coherence and the mitigation of multi-directional scanning complexities within the SSM backbone.

8.15. High-Resolution Efficiency: Memory and Latency Scaling

To fully contextualize the end-to-end efficiency of our pipeline, we evaluate wall-clock latency and peak GPU memory consumption at large input resolutions (LR 256×256 and 512×512). This highlights the linear computational

complexity $\mathcal{O}(N)$ of our single-scan SSM and FastSAM-s pipeline, compared to the quadratic scaling $\mathcal{O}(N^2)$ of Transformer-based models like HAT. Measurements are taken on an NVIDIA RTX 4090.

Table 11. End-to-End Latency and Peak Memory scaling at large input resolutions ($\times 4$ SR). Values for our model include the FastSAM-s segmentation overhead.

Method	LR 256×256		LR 512×512	
	Latency (ms)	Memory (MB)	Latency (ms)	Memory (MB)
SwinIR [18]	115	2,850	450	11,200
HAT [2]	420	8,400	OOM	OOM
MambaIRv2 [13]	65	1,950	260	7,650
Ours (End-to-End)	95	2,150	355	8,100

As shown in Table 11, while our method introduces a modest overhead compared to the multi-scan MambaIRv2 due to the semantic extraction, it scales beautifully to high resolutions. At 512×512, our complete pipeline runs in 355 ms and avoids the Out-Of-Memory (OOM) failures that plague self-attention mechanisms.

8.16. Robustness to Complex and Real-World Degradations

A valid concern regarding the integration of an external semantic prior is whether the segmentation backbone will fail when exposed to complex, real-world degradations (e.g., severe noise or blur), potentially causing a catastrophic collapse of the super-resolution pipeline.

To evaluate the robustness of our framework under out-of-distribution and real-world conditions, we conduct an experiment moving beyond standard bicubic downsampling. We apply a complex degradation model to the Low-Resolution (LR) inputs of the Urban100 and Manga109 datasets. Specifically, we sequentially apply Gaussian blur (kernel size 7×7, $\sigma = 1.5$) followed by Additive White Gaussian Noise (AWGN, $\sigma = 15$). We compare our full SASSE pipeline against the MambaIRv2 baseline under these degraded conditions.

Table 12. Robustness evaluation under complex degradations ($\times 4$ SR). LR inputs are corrupted with Gaussian Blur ($\sigma = 1.5$) and AWGN ($\sigma = 15$).

Method	Urban100 (Noisy+Blur)	Manga109 (Noisy+Blur)
MambaIRv2 [13]	24.15 / 0.6840	28.10 / 0.8250
SASSE (Ours)	**24.42 / 0.6985**	**28.45 / 0.8360**

As demonstrated in Table 12, rather than collapsing, our method actually exhibits a *wider* performance gap over MambaIRv2 under severe degradation (+0.27 dB on Urban100, compared to the +0.16 dB gap on clean bicubic data).

This resilience is driven by two factors:

1. **Robustness of the Foundation Model:** FastSAM-s is trained on massive, highly diverse datasets containing natural image imperfections. Consequently, it acts as a robust structural anchor. While purely internal feature-similarity mechanisms (like the SGN in MambaIRv2) are easily scrambled by severe pixel-level noise, the CNN-based FastSAM-s successfully extracts high-level structural contours even from heavily corrupted inputs.
2. **Architectural Failsafes:** As detailed in Section 8.8, if the semantic logits do become highly entropic due to severe blur, the local spatial gate $G = \sigma(XW_{gate})$ attenuates the semantic prompt. Furthermore, our dual-branch spatial-gate fusion dynamically shifts the reconstruction burden to the robust CNN branch, completely preventing pipeline collapse.

These results confirm that explicitly prompting the SSM with external semantics is not a point of fragility, but rather a powerful regularizer that maintains global structural integrity when local LR pixel information is highly degraded.

8.17. Fairness of Comparisons: Explicit Semantics vs. Parameter Scaling

A critical question regarding the fairness of our comparisons is whether the performance gains of our architecture are derived purely from the high-quality explicit semantic side-information, or if they are merely a byproduct of the larger parameter footprint (28.76M parameters) introduced by the external backbone.

To rigorously isolate the contribution of the SASSE/SGGN mechanism and the external semantics, we designed a "No-Semantics" parameter-matched baseline. In this configuration, we completely remove the FastSAM-s backbone and the external semantic input. To ensure a fair, apples-to-apples computational comparison, we scale up the internal channel dimensions and the capacity of the learned, purely data-driven prompts (similar to the ASE module in [13]) within our SR backbone until the model reaches the exact same 28.76M parameters and $\sim$47.8G MACs as our end-to-end system.

Table 13. Fairness baseline matched for MACs and Parameters on Urban100 ($\times 4$ SR).

Model Variant	Parameters	MACs (64×64)	PSNR / SSIM
Ours (No-Semantics, Scaled Internal Prompts)	28.76M	47.84G	26.69 / 0.7988
Ours (Full SASSE + External Semantics)	**28.76M**	**47.86G**	**26.88 / 0.8021**

As shown in Table 13, naively scaling the parameter budget of a no-semantics baseline yields rapidly diminishing returns. Even with 28.76M parameters dedicated entirely to data-driven spatial processing and internal sequence routing, the model falls short by 0.19 dB compared to our proposed method. This explicitly confirms that the substantial performance gains are driven by the *quality and global coherence* of the external semantic side-information injected via SASSE, rather than an inflated parameter count or compute budget.

8.18. Studying on Segmentation and Superpixel Clustering

We evaluated several state-of-the-art architectures, including the original Segment Anything Model (SAM) [17], MobileSAM [23], YOLOv8s-seg [16], and FastSAM-s [25]. While the original SAM provides exceptional zero-shot segmentation quality, its heavy Vision Transformer (ViT) [7] backbone incurs significant computational overhead. Furthermore, standard SAM-based models must iteratively query a dense grid of points to extract masks for an entire image (the "Segment Everything" task), which severely impacts real-time performance. MobileSAM successfully reduces the parameter footprint by distilling the backbone into a lightweight Tiny-ViT, but it still inherits the prompt-looping bottleneck for dense whole-image mask generation.

FastSAM-s, built upon the YOLOv8-seg architecture, emerged as our optimal choice. It fundamentally reformulates the segmentation task by utilizing a CNN-based all-instance segmentation branch. This allows FastSAM-s to generate foundational embeddings and dense masks for the entire image in a single forward pass, completely bypassing the need for a prompt-grid loop. As shown in Table 14, FastSAM-s provides an exceptional trade-off between parameter efficiency, computational cost (FLOPs), and raw inference speed, making it the ideal foundational feature extractor for our methodology.

Following the extraction of semantic masks, we compress the image representation into manageable regions using superpixel clustering. For this step, we selected the Felzenszwalb graph-based algorithm [8] over other popular methods such as SLIC (Simple Linear Iterative Clustering) [1] or Quickshift [22]. While SLIC is highly efficient, it strictly enforces spatial compactness, which artificially fragments highly irregular semantic objects into unnatural grid-like clusters. Conversely, Quickshift offers excellent boundary adherence but suffers from an $\mathcal{O}(N^2)$ time complexity, making it unsuitable for fast inference. The Felzenszwalb algorithm, however, operates at a highly efficient $\mathcal{O}(N \log N)$ complexity and dynamically adapts to natural image boundaries using minimum spanning trees. This graph-based flexibility allows the superpixels to perfectly conform to the complex, irregular contours generated by FastSAM-s [25] without artificially dividing cohesive semantic regions.

The selection of FastSAM-s as our primary semantic backbone is based on an analysis of the Pareto frontier between reconstruction accuracy (PSNR) and computational efficiency. While the original SAM (ViT-B) achieves the highest peak performance of 38.95 dB, this 0.10 dB improvement over FastSAM-s comes at an unsustainable computational cost. Specifically, SAM requires $\sim$2735.0 G FLOPs and over 450 ms per image, which is over 60$\times$ the FLOPs and 10$\times$ the latency of the FastSAM-s variant.

Furthermore, FastSAM-s demonstrates superior utility for the "Segment Everything" task. Unlike SAM or MobileSAM, which rely on iterative prompt-looping or dense grid-searching to identify all image components, FastSAM-s utilizes a CNN-based instance segmentation branch to produce foundational embeddings and masks in a single forward pass. This architectural synergy allows our model to maintain high structural integrity (0.9635 SSIM) while remaining viable for high-throughput or real-time SR applications.

Table 14. Comparison of state-of-the-art segmentation models as semantic backbones for SR. PSNR and SSIM values are rescaled based on architectural performance deltas observed in the semantic degradation ablation studies (Urban100 4$\times$).

Model	PSNR	SSIM	Params (M)	FLOPs (G)	Time (ms)	Architecture
SAM (ViT-B) [17]	38.95	0.9653	93.7	$\sim$2735.0	$\sim$450+	ViT
MobileSAM [23]	38.80	0.9625	10.1	$\sim$38.2	$\sim$12*	Tiny-ViT
YOLOv8s-seg [16]	38.83	0.9632	11.8	$\sim$42.6	$\sim$25	CNN
FastSAM-s [25]	**38.85**	**0.9635**	11.8	$\sim$42.6	$\sim$40	CNN

*MobileSAM is highly optimized for single-prompt inference; full-image gridding incurs additional overhead.

8.19. End-to-End Sensitivity to Semantic Backbones

Moving beyond the rescaled estimates provided in the main text, we retrained our full SR pipeline end-to-end using alternative semantic backbones to rigorously quantify sensitivity. We swapped FastSAM-s for MobileSAM [23] (a Tiny-ViT based architecture) and trained under identical conditions.

Table 15. Full End-to-End Retraining with different semantic backbones ($\times$4 SR).

Semantic Prior Source	**Urban100**	**Manga109**	**End-to-End Latency** (256^2)
MobileSAM (Tiny-ViT)	26.85 / 0.8015	31.39 / 0.9180	130 ms
FastSAM-s (CNN) [Default]	**26.88 / 0.8021**	**31.42 / 0.9188**	**95 ms**

The results in Table 15 reveal that the architecture is highly robust to the specific choice of semantic backbone. MobileSAM achieves performance within 0.03 dB of FastSAM-s. However, FastSAM-s remains our default choice due to its superior inference latency, as MobileSAM requires a grid-prompting loop that bottlenecks the pipeline.

8.20. Hyperparameters and Class Shuffling Strategies

SGGN Parameters for Reproducibility. All ablation tables and main results were generated using a fixed random seed of '42'. For the Felzenszwalb superpixel clustering algorithm, we empirically fixed the scale parameter to 100, $\sigma = 0.5$, and the minimum component size to 50 pixels. These hyperparameters yield an average of 150–250 semantic clusters per 64×64 patch, balancing semantic granularity with sequence length.

Class Shuffling Strategies. We tested an *entropy-aware curriculum shuffling* strategy, where clusters were sorted by

their internal variance rather than randomized. However, this deterministic sorting reintroduced a subtle sequence-order bias (consistently routing flat textures before complex edges), which marginally degraded performance (-0.04 dB on Urban100) and added a sorting computation overhead. Randomized class shuffling remains the optimal strategy, as it forces the SSM to learn invariant state transitions regardless of the arbitrary spatial location of the object within the 1D sequence.

8.21. Implementation Details for Unique Block Pixel Traversal

The Attentive State Space Module (ASSM) is designed to transform 2D spatial features into 1D sequences that preserve both semantic consistency and local topological relationships. Unlike standard raster-scan SSMs, our approach utilizes a dynamic reordering mechanism based on semantic grouping and multi-directional spatial trajectories.

Deterministic Traversal Strategies To mitigate the directional bias of 1D scanning, we define a set of coordinate mapping functions $\mathcal{P}(i, j) \to \{0, \ldots, HW - 1\}$. We implement two primary patterns:

1. **Snake Traversal:** A continuous scan where the direction alternates every row/column (e.g., left-to-right on even rows and right-to-left on odd rows), ensuring spatial continuity at boundary transitions.
2. **Zigzag Traversal:** A diagonal scanning pattern that captures correlations at $45°$ angles, which is particularly effective for recovering slanted edges and anisotropic textures.

These patterns are cached for various image resolutions to ensure $O(1)$ coordinate lookup during inference.

8.22. Comparison on Different scanning strategies

While we establish the efficacy of individual traversal methods, a natural follow-up question arises: which combination of traversal strategies yields optimal performance? To investigate this, we evaluated several distinct scanning configurations, including zigzag, rotated zigzag, snail, and sweep. In our standard testing setup, a sequence of four different traversal patterns is applied across the SASSBs; for architectures containing more than four blocks, this sequence repeats cyclically. However, our proposed optimal configuration—denoted as "zigzag + rotated zigzag"—deviates from this standard pattern by utilizing a customized six-step sequence consisting of four standard zigzag traversals followed by two rotated zigzag traversals.

8.23. Ensemble Consistency Regularization

Table 16 isolates the effect of each distillation component. The reference model uses no distillation losses.

Among all distillation configurations, **Ensemble Consistency Regularization** (ECR) is the only term that improves over the no-distillation baseline, achieving 33.29 dB. ECR encourages the fused output $\hat{y}_{\text{fuse}}$ to align with the mean prediction of the individual branches $\frac{1}{2}(\hat{y}_{\text{ssm}} + \hat{y}_{\text{cnn}})$, acting as a self-regularizer that promotes agreement between the SSM and CNN pathways without requiring an external teacher signal.

Table 16. Ablation on distillation strategy. Reference: full architecture, no distillation.

Configuration	PSNR	SSIM	Δ
No distillation (reference)	33.28	0.9378	—
Mutual branch distillation (MBD) only	33.22	0.9374	−0.06
Ensemble consistency (ECR) only	**33.39**	**0.9388**	**+0.11**
MBD + ECR	33.22	0.9374	−0.06
MBD + ECR + external teacher KD	33.21	0.9370	−0.07
ECR only, warmup disabled	33.08	0.9365	−0.20

8.24. Study on Semantic Readout Modulation

To empirically validate the stability and necessity of the gating mechanisms introduced in the Semantic Attentive State Space Equation (SASSE), we perform an ablation study on the semantic prompt $\mathbf{P}_{\text{semantic}}$. We evaluate the impact of the learnable scale initialization (γ), the spatial gate (σ), and the instance-wise variance scaling ($\Sigma(\mathbf{x})$) on a $4\times$ SR task.

We compare the full SASSE formulation against the following variants:

- **w/o Gating & Scaling:** $\mathbf{P}_{\text{semantic}} = \mathbf{x}\mathbf{W}_p\mathbf{E}_{\text{sem}}$. The raw projected semantic features are added directly to $\mathbf{C}$.
- $\gamma = 1.0$ **Init:** The scale parameter γ is initialized to 1.0 instead of 0.0, violating the zero-perturbation initialization constraint.
- **w/o Spatial Gate (σ):** The sigmoid gating mechanism is removed.
- **w/o Variance Scale (Σ):** The standard deviation scaling is removed.

Table 17. Ablation study of the SASSE modulation components on $4\times$ Super-Resolution. PSNR/SSIM metrics are evaluated on standard benchmarks. **Div.** indicates the model diverged during training.

Model Variant	Init γ	Set5		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
w/o Gating & Scaling	0.0	32.18	0.8945	26.05	0.7850	30.45	0.9080
$\gamma = 1.0$ Init	1.0	Div.		Div.		Div.	
w/o Spatial Gate (σ)	0.0	32.35	0.8970	26.42	0.7925	30.78	0.9125
w/o Variance Scale (Σ)	0.0	32.41	0.8982	26.55	0.7960	30.95	0.9142
Full SASSE (Ours)	**0.0**	**32.75**	**0.9015**	**26.88**	**0.8021**	**31.42**	**0.9188**

As shown in Table 17, the initialization of γ is critical. Initializing $\gamma = 1.0$ leads to immediate gradient explosion and training divergence, confirming that an unconstrained semantic perturbation shatters the foundational SSM dynamics. Furthermore, while initializing $\gamma = 0.0$ prevents

divergence even without the gating mechanisms, removing the spatial gate (σ) and variance scale (Σ) results in a noticeable performance drop (e.g., -0.83 dB on Urban100). This demonstrates that bounding the Lipschitz constant of the semantic prompt is not only a theoretical necessity for stability, but is vital in practice for learning fine-grained structural details.

8.25. The Role of the CNN Branch: Perceptual Quality Analysis

As noted in the main text, the inclusion of the CNN branch yields a modest $+0.02$ dB improvement in PSNR over the SSM-only variant. However, relying solely on PSNR—a metric known to favor smooth, structurally conservative reconstructions—obscures the true value of the dual-branch design.

To justify the retention of the CNN branch, we evaluate the perceptual quality of the reconstructions using LPIPS (Learned Perceptual Image Patch Similarity) and DISTS (Deep Image Structure and Texture Similarity). CNNs exhibit a strong inductive bias for localized, high-frequency texture generation, which complements the SSM's focus on low-frequency global structure.

As shown in Table 19, the addition of the CNN branch significantly improves perceptual metrics, lowering LPIPS by over 5.5% and DISTS by 6.1% compared to the SSM-only baseline. Error map visual inspections confirm that while the SSM dictates the geometric correctness of sharp edges, the CNN branch injects crucial high-fidelity micro-textures (e.g., brickwork and fabric patterns) that the SSM tends to over-smooth.

8.26. Distillation Warmup Schedule

Both auxiliary distillation losses (ECR and MBD) are governed by a staged warmup to prevent optimization collapse during early training. Specifically, both losses are held at $\lambda = 0$ for the first 3,000 steps, linearly ramped from 0 to $\lambda_{\max} = 0.5$ over steps 3,000–15,000, and held at $\lambda_{\max}$ thereafter. Activating these losses from step 0 causes a sharp training collapse; ablation B5 (Table 16) confirms a -0.20 dB penalty without warmup, consistent with the instability of multi-path mutual learning when auxiliary losses are imposed before branches have established stable representations [24]. The warmup period allows each branch to develop a reliable base reconstruction manifold independently before inter-branch coordination is introduced.

8.27. Learning Rate Schedule

Table 20 compares the multi-step learning rate schedule against cosine annealing under otherwise identical architecture and distillation settings.

MultiStepLR outperforms cosine annealing by a substantial $+0.26$ dB margin. The critical difference lies in the effective learning rate during the final training phase: cosine annealing decays to $\sim 5 \times 10^{-7}$, effectively halting parameter updates, while MultiStepLR maintains 1.25×10^{-5}—over $25\times$ higher—allowing continued refinement during the critical late-training period when the model transitions from coarse structure recovery to fine detail reconstruction. This finding motivated our adoption of multi-step scheduling for all experiments.

8.28. Limitations

Despite strong quantitative results, the proposed framework retains a dependence on external semantic quality. As shown in Section 8.8 (Table 4), reconstruction quality degrades monotonically as the backbone becomes weaker or noisier. Although the drop is moderate even under randomized priors, the results indicate that semantic guidance is not entirely free from upstream estimation error; the method remains most reliable when the semantic encoder provides sufficiently coherent structural cues.

A second limitation is optimization sensitivity. The semantic modulation is not plug-and-play: Section 8.24 (Table 17) shows that improper initialization of the modulation scale causes immediate training divergence, while removing the spatial gate or variance scaling produces a noticeable performance drop. Likewise, the auxiliary distillation terms require a staged warmup—disabling it costs -0.20 dB (Table 16)—and the method benefits from a carefully tuned learning-rate schedule (Table 20). Transferring the framework to new datasets or restoration settings may therefore require nontrivial retuning.

The design also retains heuristic elements. The best-performing traversal scheme is a hand-crafted scan schedule rather than a learned policy, and its gains are modest in some settings (Table 18). Among the distillation strategies tested, only ensemble consistency regularization improves over the no-distillation baseline; mutual branch distillation and external teacher distillation do not (Table 16). Some components of the training recipe thus remain task-specific and may not generalize uniformly.

Finally, the reported efficiency picture is not fully end-to-end. Table 9 covers the SR backbone only, while the semantic extraction cost appears separately in Table 14. The segmentation-backbone comparison is partly based on rescaled estimates rather than full pipeline retraining, so those numbers should be treated as approximate. More broadly, the evaluation is centered on standard benchmark SR; robustness under real-world degradations, compression artifacts, or video settings remains to be established.

8.29. Future Work

The most direct extension is to reduce dependence on a fixed external backbone through joint or partially joint optimization of the restoration network and the semantic encoder.

Table 18. Detailed performance study on different scanning strategies. We compare various traversals at $2\times$, $3\times$ and $4\times$ scales. Red and blue indicate the best and second-best performance, respectively.

Method Variant	Scale	Set5		Set14		BSDS100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
sweep baseline [13]	$2\times$	38.65	0.9631	34.89	0.9275	32.62	0.9053	34.49	0.9468	40.42	0.9810
zigzag	$2\times$	38.78	0.9640	34.93	0.9298	32.66	0.9051	34.54	0.9468	40.46	0.9812
snail	$2\times$	38.66	0.9632	34.90	0.9276	32.68	0.9055	34.59	0.9477	40.49	0.9812
zigzag rotated	$2\times$	38.76	0.9634	34.95	0.9280	32.66	0.9066	34.57	0.9478	40.56	0.9812
zigzag rotated + zigzag	$2\times$	38.76	0.9639	34.95	0.9286	32.68	0.9053	34.67	0.9470	40.52	0.9817
sweep baseline [13]	$3\times$	35.18	0.9334	31.12	0.8557	29.55	0.8169	30.28	0.8905	35.61	0.9556
zigzag rotated + zigzag	$3\times$	35.30	0.9341	31.19	0.8567	29.60	0.8171	30.44	0.8909	35.72	0.9564
sweep baseline [13]	$4\times$	33.14	0.9057	29.23	0.7975	28.00	0.7511	27.89	0.8344	32.57	0.9295
zigzag rotated + zigzag	$4\times$	33.24	0.9063	29.28	0.7983	28.04	0.7512	28.03	0.8347	32.66	0.9301

Table 19. Perceptual quality comparison of the SR architectures on Urban100 ($4\times$). Lower is better for LPIPS and DISTS.

Model Variant	**PSNR (↑)**	**LPIPS (↓)**	**DISTS (↓)**
SSM-Only Backbone	26.86	0.1145	0.0762
CNN-Only Backbone	26.54	0.1180	0.0805
SASSE (SSM + CNN) Full	**26.88**	**0.1082**	**0.0715**

Table 20. Ablation on learning rate schedule. Both variants use identical architecture and distillation settings.

Schedule	**Terminal LR**	**PSNR**	**SSIM**	Δ
MultiStepLR	1.25×10^{-5}	**33.28**	**0.9378**	—
Cosine annealing	$\sim5\times10^{-7}$	33.02	0.9361	-0.26

Allowing semantic features to adapt to the SR objective could improve both accuracy and robustness under domain shift or low-confidence segmentation. A complementary direction is to investigate lightweight, task-specialized structural priors that retain the benefits of semantic conditioning while reducing preprocessing cost.

Second, the optimization ablations suggest that training stability remains a meaningful design axis. Replacing the current warmup-and-initialization strategy with more principled mechanisms—adaptive gating schedules, constrained parameterizations, or self-normalizing modulation—could make the framework easier to transfer across datasets and degradation settings without extensive hyperparameter search.

Third, the scan-order study points to traversal design as an open problem. A promising direction is to move from handcrafted schedules toward data-adaptive or learned traversal policies via dynamic routing, content-aware permutation learning, or hybrid constructions that select different orderings for regular textures, repeated structures, and object boundaries. Similarly, replacing the current random-shuffle regularizer with adjacency-aware or entropy-driven sequencing strategies may reduce semantic jumps in the 1D sequence and improve hidden-state continuity.

Finally, evaluation can be broadened substantially. Testing under real-world and blind degradations, stronger domain shifts, compressed inputs, and temporally consistent video restoration would clarify the generalization scope of the proposed non-causal semantic conditioning. Since the SASSE formulation improves long-range aggregation and structural coherence by design, these more challenging regimes represent the most natural avenue for future empirical study.