Learning Dynamic Evidence Routes for Vision Transformer Probing

Zice Wang*
School of Future Technology
Northeastern University
Shenyang, China

Zhenyu Zhang†
School of Computer Science and Technology
Beijing Institute of Technology
Beijing, China

Abstract

Probing frozen vision transformers typically uses permutation-invariant aggregation (GAP or `[CLS]`), treating patch tokens as an unstructured set. Content-dependent probes such as self-attention are useful accuracy controls, but they do not expose a fixed token schedule or fixed position weights for auditing. We introduce **SSM-Probe**, a probe of frozen vision transformers that learns a Sinkhorn patch route and reads it with a diagonal S4 decoder. The decoder's causal linear time-invariant (LTI) recurrence yields fixed position coefficients, exposing how patch arrangement interacts with sequential weighting. Held-out spatial predictability measures route organization beyond a location-bias null. On CUB and Cars, MAE, BEiT, DINOv2, and supervised ViT yield spatially organized routes with distinct readout profiles: DINOv2 is `[CLS]`-dominant, while MAE benefits from learned routing over fixed scans. On MAE/ImageNet, Sinkhorn reaches 70.33% accuracy, 5.55 percentage points above the strongest fixed scan.

1 Introduction

Lightweight probes evaluate information accessible in frozen vision transformers (ViTs). Global Average Pooling (GAP) and `[CLS]` use global representations, while self-attention and deep MLPs learn content-dependent aggregation [Dosovitskiy et al., 2020, Caron et al., 2021]. We study how patch evidence is arranged for sequential readout.

Pre-training objectives such as DINOv2 [Oquab et al., 2024], MAE [He et al., 2022], and BEiT [Bao et al., 2021] support different global and patch readouts. An inspectable patch sequence connects these accuracy profiles to the spatial organization of selected evidence.

SSMProbe learns an *evidence route* with a Sinkhorn router and reads it with a diagonal S4 decoder [Gu et al., 2021]. S4's native causal linear time-invariant (LTI) recurrence supplies fixed position coefficients after training. The resulting route–readout decomposition supports spatial measurements and analytical sequence perturbations. Appendix B.3 reports parameters and computation budgets.

Figure 1 visualizes routes on a shared image. We quantify their organization by held-out spatial predictability beyond marginal location bias, then measure foreground alignment, seed reproducibility, and sensitivity to sequence perturbations.

On CUB and Cars, all four backbones yield positive held-out spatial information. MAE benefits from learned routing; BEiT and ViT support strong patch aggregation; DINOv2 retains a `[CLS]`-dominant profile. Decoder controls show that route geometry also depends on the trained readout.

*Email: wangzc1@mails.neu.edu.cn
†Email: charlie@bit.edu.cn

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

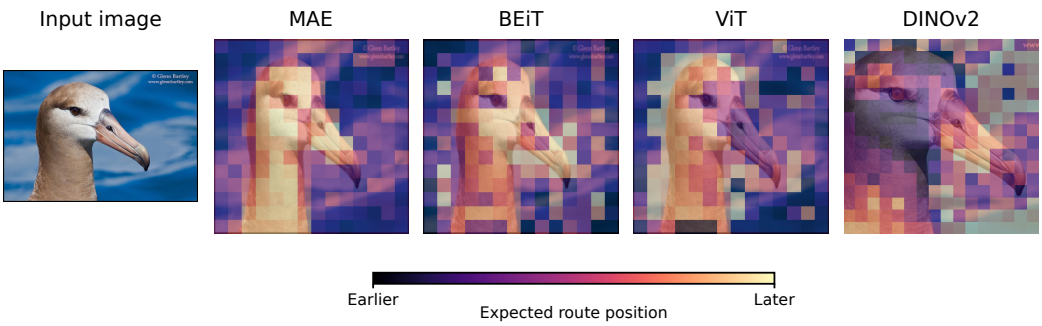


Figure 1: **Patch evidence routes.** The same CUB image is shown alongside the SSMProbe route for four frozen backbones. Color encodes each source patch's normalized expected position in the learned sequence, from earlier (dark) to later (bright). Each overlay uses its backbone's image preprocessing and patch grid. Spatial compactness and bird-foreground alignment are measured by pairwise patch distances and official-mask coverage, respectively.

Our contributions are: (1) **SSMProbe**, a Sinkhorn+S4 diagnostic that exposes a learned route and fixed positional coefficients; (2) **evidence-route geometry**, a compact-vs-dispersed analysis across ViT pre-training objectives; and (3) **analytical readout audits** through decomposability, spectrum, perturbation, and skip-path measurements.

2 Related Work

Information Organization in Vision Transformers. Pre-training objectives structure token information differently [Oquab et al., 2024]: contrastive and supervised models often make `[CLS]` linearly useful [Caron et al., 2021], whereas MAE [He et al., 2022] emphasizes dense patch reconstruction. DeepSets [Zaheer et al., 2017] and NetVLAD [Arandjelovic et al., 2016] aggregate patch sets; pruning methods [Wu et al., 2023, Bolya et al., 2023] reduce token-processing costs. SSMProbe measures spatial evidence through an explicit route and fixed positional readout.

State Space Models in Vision. S4 [Gu et al., 2021], Mamba [Gu and Dao, 2023], and visual SSMs [Zhu et al., 2024, Liu et al., 2024] are usually studied as efficient backbones. SSMProbe uses S4 to read frozen features: its linear time-invariant form gives an explicit decomposition of ordering effects (Proposition 1).

Probing and Differentiable Routing. Linear probing evaluates representation quality via feature separability [Alain and Bengio, 2016], but ViT patch probes also face an ordering choice. Our router builds on differentiable sorting and Sinkhorn optimal transport [Mena et al., 2018, Cuturi et al., 2019], including NeuralSort [Grover et al., 2019] and SoftSort [Prillo and Eisenschlos, 2020]. We use these tools to interrogate frozen patch features by ranking them before sequential aggregation.

Spatial and Token-Level Structure. Spatial pretext tasks such as Jigsaw-ViT [Chen et al., 2023] and PuzzleTuning [Zhang et al., 2023] explicitly train image representations to recover spatial arrangements. Focal self-attention [Yang et al., 2021] instead enriches ViT attention with local-global interactions. These approaches modify training objectives or backbone computation; SSMProbe complements them as a post-hoc diagnostic of the route induced over frozen patch features by a specified downstream readout.

3 Method

3.1 Problem Setup

Given an image $x \in \mathbb{R}^{H\times W\times C}$, a frozen Vision Transformer encoder [He et al., 2022, Bao et al., 2021, Oquab et al., 2024, Dosovitskiy et al., 2020] (e.g., MAE, BEiT, DINOv2, or ViT) processes it into non-overlapping patches $x_p \in \mathbb{R}^{N\times(P^2C)}$, where $N = HW/P^2$ for images whose spatial

dimensions are divisible by the patch size P. The encoder produces final-layer hidden states:

$$H = [h_{\texttt{cls}}, h_1, \ldots, h_N] \in \mathbb{R}^{(N+1)\times d},$$

where d is the latent dimension. The frozen backbone supplies patch tokens $T = [h_1, \ldots, h_N] \in \mathbb{R}^{N\times d}$ to the S4 probe and $h_{\texttt{cls}}$ to the CLS baseline.

3.2 Structured State Space Sequence Models (S4)

Our probing mechanism uses a diagonal S4 layer [Gu et al., 2021], a discrete Linear Time-Invariant (LTI) system with state transition $x_k = \overline{\mathbf{A}}x_{k-1} + \overline{\mathbf{B}}u_k$ and output $y_k = \mathbf{C}x_k + \mathbf{D}u_k$. Our `torch-s4` implementation parameterizes continuous modes as $\lambda_i = -\mathrm{softplus}(a_i) + \mathrm{i}\,\omega_i$ and applies a bilinear transform, placing discrete modes inside the unit disk (Appendix A.1). S4 provides a native causal recurrence with a content-independent kernel: ordering effects, spectral attenuation, and the skip path are directly measurable.

3.3 SSMProbe Head Formulation

For each sample, we apply a permutation matrix $\mathbf{P} \in \{0,1\}^{N\times N}$ with $\mathbf{P}\mathbf{1} = \mathbf{1}$ and $\mathbf{P}^\top\mathbf{1} = \mathbf{1}$ (or its Sinkhorn soft relaxation) to the token sequence T, yielding $\tilde{T} = \mathbf{P}^\top T \in \mathbb{R}^{N\times d}$. We then process the d-dimensional features across the sequence length using a diagonal S4 block applied channel-wise:

$$Z = \mathrm{S4}(\tilde{T}) \in \mathbb{R}^{N\times d}, \quad z_{\text{out}} = Z_N.$$

The final sequence output z_{out} aggregates information along the chosen traversal path. A linear projection produces logits $\hat{y} = W z_{\text{out}} + b$. Downstream supervision trains $\theta_{\text{probe}} = \{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \Delta, W, b\}$ and the router scorer when present.

3.4 The SSM as a Diagnostic Instrument

Evidence-route position. Given the Sinkhorn assignment $\mathbf{P}^{(K)}$ defined in Section 3.5, the expected S4 position of source patch i is

$$\tau_i = \sum_{t=1}^{N} t\, \mathbf{P}^{(K)}_{i,t}.$$

We inspect each patch through this position and the coefficient applied after routing (Proposition 1).

Spatial information in routed evidence. Let $\mathcal{G}(i) \in \{1, \ldots, \sqrt{N}\}^2$ be the original 2D grid coordinate of patch token i, and let $\pi_{1:k}$ be the top-k source-token route induced by the trained scorer. We measure spatial organization by held-out route predictability beyond a location-bias null. We estimate an exchangeable null $p_0(i) \propto m(i)$ from a calibration split, where $m(i)$ is the empirical marginal frequency with which grid position i appears in top-k routes. We compare it to a one-parameter spatial route model

$$p_{\text{geo}}(i_t \mid i_{<t}; \sigma) \propto m(i_t) \exp\left(-\frac{d(i_t, \{i_1, \ldots, i_{t-1}\})^2}{2\sigma^2}\right),$$

where d is the distance to the nearest previously routed patch; both models sample without replacement. The spatial scale σ is selected on the calibration split and evaluated on held-out images. Our primary geometry diagnostic is the excess held-out code length saved by the spatial model,

$$\mathrm{ERSI}_k = \frac{1}{k \log 2}\left[\log p_{\text{geo}}(\pi_{1:k}) - \log p_0(\pi_{1:k})\right] \quad \text{bits per routed patch.}$$

ERSI can be negative on held-out data, but our σ grid includes a very large value that makes $p_{\text{geo}} \approx p_0$; reported zeros therefore indicate that the selected spatial model collapses to the null within rounding. Positive values indicate spatial organization beyond marginal location bias, and small fitted σ indicates local evidence routes. We report task accuracy alongside ERSI to relate spatial organization to classification.

We also report the pairwise compactness ratio $R_k = D_k / D_k^{\text{rand}}$, where D_k is the mean pairwise grid distance among top-k routed patches. Connected components, convex hull area, minimum-spanning-tree length, and coarse-bin entropy describe complementary aspects of route shape (Appendix A.4).

Controlled interventions. Fixed scans, random permutations, and reversal measure the decoder's dependence on route scheduling. Cross-seed router swaps measure route reuse. We report within-checkpoint accuracy changes and compare learned routes with deterministic traversals.

Fixed-coefficient decomposition. S4's recurrence defines the position coefficients below. A fixed-position linear decoder parameterizes them directly (Appendix D); self-attention uses input-dependent coefficients and offers positional-encoding design choices.

Proposition 1 (Decomposability gap). Let $f_\theta(T, \pi)$ denote the logits produced by sequential decoder θ on patch tokens T arranged in order π.

1. **S4 decoder.** Since our classifier reads the final hidden/output vector $z_{\text{out}} = Z_N$, the forward pass simplifies to $f_{\text{S4}}(T, \pi) = \sum_{k=1}^{N} \alpha_k \cdot t_{\pi(k)} + b$, where the fixed coefficients are $\alpha_k = W\mathbf{C}\,\bar{A}^{N-k}\bar{B} + \mathbf{1}_{\{k=N\}} W\mathbf{D}$. Here W is the classifier and $\mathbf{C}$ the S4 state-to-output projection. The direct feedthrough term contributes at the final routed position. The fixed coefficients determine logits for every alternative permutation π' analytically.
2. **Transformer.** A self-attention probe computes $f_{\text{TF}}(T, \pi) = g\big(\sum_j a_j(T, \pi) \cdot (W_V\, t_{\pi(j)})\big)$, where the attention coefficients $a_j(T, \pi)$ depend on the input tokens and their arrangement.

Analytical permutation logits match direct S4 forwards within 5×10^{-6} absolute error (Appendix C.1.2). A position-free Transformer is permutation-equivariant before pooling [Zaheer et al., 2017].

3.5 Differentiable Sinkhorn Token Routing

We structure token ordering as a 1D optimal transport problem. A lightweight linear scorer assigns each token h_i a standardized scalar score $s_i = \mathbf{w}^\top h_i$, which is matched to evenly spaced rank targets by a squared-distance cost. Sinkhorn-Knopp [Cuturi, 2013] (K=20, τ=0.1) yields a differentiable doubly-stochastic permutation $\mathbf{P}^{(K)}$ (Appendix A.3), and $\tilde{T} = {\mathbf{P}^{(K)}}^\top T$. Joint optimization of $\mathbf{w}$ and the probe induces a task-conditioned route. Cross-dataset scorer transfer measures reuse of this learned ranking.

3.6 Static 2D Traversal Families

For our fixed-scan baselines, we evaluate deterministic heuristics that linearize the 2D patch grid: simple raster (single row-major), 4-dir VMamba (row/column sweeps following SS2D [Liu et al., 2024]), 4-dir diagonal, and 4-dir snake mappings. The 4-dir layout families average their directional variants. Figure 3 visualizes these heuristic strategies (detailed definitions in Appendix A.5).

4 Experimental Setup

Backbones and datasets. We evaluate four frozen ViT backbones spanning distinct pre-training paradigms: MAE (`facebook/vit-mae-base`), BEiT (`microsoft/beit-base-patch16-224-pt22k-ft22k`), DINOv2 (`facebook/dinov2-base`), and supervised ViT (`google/vit-base-patch16-224`). Our ImageNet-1K study focuses on MAE, and the cross-backbone comparison uses CUB-200-2011 [Wah et al., 2011] and Stanford Cars [Krause et al., 2013] as fine-grained transfer benchmarks.

Probe families and optimization. Beyond GAP and CLS, we compare attention pooling, content-weighted pooling, Top-k pooling, Transformer, DeepSets, Bi-GRU, fixed scans, and the learned Sinkhorn route; full head details are deferred to Appendix A.6. All heads are trained with AdamW at learning rate 10^{-3}, cosine schedule, batch size 256, and no weight decay for 5 epochs on ImageNet-1K and 100 epochs on CUB/Cars. The backbone is always frozen, and each head uses an independent optimizer (Appendix A.1, Algorithm 1).

5 Results

We compare readout accuracy, source-grid route geometry, foreground alignment, and decoder dependence across frozen backbones.

5.1 MAE: Task-Conditioned Evidence Routes

ImageNet-1K: route reliability. Table 1 compares MAE readouts under the five-epoch protocol.

Table 1: Frozen MAE probing on ImageNet-1K (five-seed summaries). Scan/permutation and Bi-GRU controls use final checkpoints and source-grid patch order; their error bars are sample standard deviations.

Tier	Method	Top-1 Acc. (%)
Floor	GAP	58.10 ± 0.03
	CLS	56.65 ± 0.03
Order only	Raster scan	63.64 ± 0.05
	4-dir (VMamba-style)	64.68 ± 0.07
	4-dir (Snake)	64.76 ± 0.06
	4-dir (Diagonal)	64.78 ± 0.07
	Random-Fixed Perm + S4	63.42 ± 0.31
	Random-Dynamic Perm + S4	63.95 ± 0.06
Content only	Attention Pool	67.70 ± 0.04
	Content-Weighted Pool	67.68 ± 0.06
Full method	**Sinkhorn (learned)**	$\mathbf{70.33 \pm 0.05}$
Ablation	TopK Pool (k=16)	50.24 ± 2.57
	Transformer	68.10 ± 0.15
	DeepSets	68.48 ± 0.05
	Sinkhorn + 1D-Conv	66.60 ± 0.07
	Bi-GRU (Raster)	69.31 ± 0.12
Reference	*Transformer ref. (3M params)*	71.61 ± 0.14

Sinkhorn reaches 70.33%, gaining 5.55 points over the strongest fixed scan and 6.38 over random-dynamic ordering. Across five seeds, source-grid routes give $\mathrm{ERSI}_{20} = .311 \pm .006$ bits/patch and $R_{20} = .581 \pm .002$ (Appendix D.6).

Fine-grained transfer: spatially organized learned routes. On CUB and Cars, fixed S4 scans fit the training labels closely but reach 6.7–10.5% and 5.2–7.0% test accuracy. Dynamic random ordering reaches 28.4% and 29.4%, while Sinkhorn reaches 56.7% and 54.9% (Table 2). The source-aligned 512-image audits yield positive ERSI and compact routes on both datasets (Table 3). Thus, MAE supports spatially organized learned routes where deterministic traversals generalize weakly.

5.2 DINOv2: CLS-Dominant Readout

On CUB, DINOv2 [Oquab et al., 2024] CLS reaches 89.49% against GAP at 68.55%, the largest CLS–GAP separation among the four backbones. Fixed scans span 35.45–56.03%, random-dynamic ordering reaches 68.91%, and Sinkhorn reaches 82.28% (Table 2).

On Cars, CLS reaches 87.54% and Sinkhorn 84.83%, while fixed scans span 23.05–36.83%. DINOv2 also has the highest CUB ERSI (.484 bits/patch), showing spatially predictable patch routes alongside a CLS-dominant accuracy profile.

5.3 ViT and BEiT: Compact Patch Evidence

ViT (supervised): compact and balanced. CLS and GAP have similar accuracy on both datasets for supervised ViT [Dosovitskiy et al., 2020]. Sinkhorn is close to these readouts on CUB and exceeds

them on Cars. Positive ERSI and compactness ratios $R_k = .669$ and $.561$ indicate predictable, localized patch selections (Table 3).

BEiT: patch-dominant and compact. BEiT [Bao et al., 2021] has stronger GAP than CLS readouts. Sinkhorn is close to GAP on CUB and exceeds it by 5.80 points on Cars. Its routes have positive ERSI and $R_k \approx .60$ on both datasets, pairing strong patch aggregation with compact evidence selection.

5.4 Cross-Backbone Diagnostic Comparison

Table 2: Top-1 accuracy (mean ± std, %; five seeds except †). MAE, BEiT, and DINOv2 use self-supervised pretraining; ViT is supervised. Colors rank the top three capacity-controlled readouts in each column. The wider Transformer reference appears in its own group.

Method	MAE		BEiT	
	CUB	Cars	CUB	Cars
CLS	28.88 ± 0.10	16.64 ± 0.11	80.80 ± 0.13	57.28 ± 0.28
GAP	19.40 ± 0.10	17.62 ± 0.05	88.27 ± 0.13	76.41 ± 0.10
Raster scan	6.72 ± 0.12	5.34 ± 0.21	75.90 ± 0.16	54.64 ± 0.35
4-dir (Snake)	9.84 ± 0.23	6.02 ± 0.12	80.57 ± 0.40	51.65 ± 0.59
4-dir (Diagonal)	10.51 ± 0.20	6.97 ± 0.16	80.86 ± 0.46	54.50 ± 0.40
4-dir (VMamba)	8.91 ± 0.09	5.21 ± 0.10	80.12 ± 0.08	51.72 ± 0.21
Random-Fixed Perm + S4	6.66 ± 1.43	5.90 ± 1.72	81.68 ± 0.95	58.22 ± 4.99
Random-Dynamic Perm + S4	28.39 ± 0.28	29.41 ± 0.15	87.48 ± 0.15	75.08 ± 0.22
Sinkhorn (learned)	56.67 ± 7.54	54.86 ± 4.35	88.12 ± 0.32	**82.21 ± 0.64**
Transformer	**61.03 ± 0.65**	**64.13 ± 0.57**	86.03 ± 0.37	79.71 ± 0.42
DeepSets	34.69 ± 0.44	39.96 ± 0.44	**88.60 ± 0.11**	78.47 ± 0.22
Sinkhorn + 1D-Conv	31.57 ± 0.53	33.86 ± 1.92	87.73 ± 0.20	71.64 ± 0.51
Bi-GRU (Raster)	18.76 ± 0.89	15.29 ± 0.72	79.15 ± 0.69	27.48 ± 0.23
Transformer ref.	*70.08 ± 0.57*	*71.11 ± 0.65*	*88.03 ± 0.14*	*85.71 ± 0.17*

Method	ViT		DINOv2	
	CUB	Cars	CUB	Cars
CLS	**85.45 ± 0.10**	57.01 ± 0.21	**89.49 ± 0.21**	**87.54 ± 0.11**
GAP	85.14 ± 0.13	56.89 ± 0.10	68.55 ± 0.23	75.13 ± 0.09
Raster scan	79.90 ± 0.24	40.51 ± 0.21	35.45 ± 0.32	23.05 ± 0.39
4-dir (Snake)	82.17 ± 0.21	41.85 ± 0.18	56.03 ± 0.22	36.64 ± 0.40
4-dir (Diagonal)	82.32 ± 0.16	43.90 ± 0.15	50.71 ± 0.21	32.97 ± 0.46
4-dir (VMamba)	81.73 ± 0.17	41.84 ± 0.16	50.40 ± 0.15†	36.83 ± 0.59
Random-Fixed Perm + S4	81.36 ± 0.58	44.48 ± 2.08	37.24 ± 5.49	35.52 ± 6.50
Random-Dynamic Perm + S4	83.85 ± 0.20	55.24 ± 0.29	68.91 ± 0.18	74.20 ± 0.14
Sinkhorn (learned)	85.22 ± 0.19	**62.38 ± 0.15**	82.28 ± 1.60	84.83 ± 0.60
Transformer	81.22 ± 0.26	58.40 ± 0.63	83.56 ± 0.19	85.97 ± 0.46
DeepSets	84.03 ± 0.25	56.98 ± 0.34	76.62 ± 0.29	81.06 ± 0.16
Sinkhorn + 1D-Conv	83.88 ± 0.30	55.34 ± 0.35	71.19 ± 0.45	74.24 ± 0.52
Bi-GRU (Raster)	82.12 ± 0.46	40.87 ± 0.52	59.76 ± 0.80	50.52 ± 1.07
Transformer ref.	*85.17 ± 0.07*	*68.70 ± 0.39*	*87.31 ± 0.15*	*89.84 ± 0.16*

S4 four-direction controls use seeds 42–46 and final checkpoints, with sample standard deviations. Each path covers all 196 MAE/BEiT/ViT patches or 256 DINOv2 patches. MAE scan/permutation and Bi-GRU controls restore source-grid order; Bi-GRU uses Raster order. †: four valid seeds and one numerical failure (seed 44, NaN loss at epoch 96), retained without replacement and excluded from the accuracy summary. Appendix A.5 specifies the paths.

Table 3 pairs route geometry with readout accuracy; Appendix A.4 defines the grid descriptors.

All four backbones yield positive ERSI and $R_k < 1$ on both datasets. Their differences lie in readout profiles and the degree of spatial organization.

Table 3: Trained evidence-route geometry. ERSI_{20} measures held-out spatial-information gain over the location-bias null (bits/patch); σ is the fitted spatial scale and R_k the pairwise compactness ratio. Accuracy provides readout reliability context.

Measure	MAE	BEiT	ViT	DINOv2
Spatial information ERSI_{20} (CUB)	0.318	0.175	0.113	0.484
Spatial information ERSI_{20} (Cars)	0.336	0.276	0.296	0.328
Fitted σ (CUB / Cars)	3 / 2	3 / 2	3 / 2	3 / 3
Compactness ref. R_k (CUB / Cars)	0.687 / 0.567	0.609 / 0.599	0.669 / 0.561	0.697 / 0.680
Sinkhorn accuracy (CUB)	56.67	88.12	85.22	82.28
Sinkhorn accuracy (Cars)	54.86	82.21	62.38	84.83
Characteristic pattern	Organized routing-sensitive	Organized compact patch	Organized compact local	Organized, CLS-leaning

Object foreground. Official CUB masks connect route geometry to bird foreground (Table 4). All four backbones exceed both random selection and the location-bias null. MAE reaches 55.40% foreground coverage against 12.60% random coverage and a 33.25% location-bias null, an excess of 22.15 percentage points.

Table 4: Top-20 routed patches on bird foreground (%), over all 5,794 CUB test images. Mask preprocessing matches each backbone. DINOv2 uses 256 patches; the others use 196.

Selection	MAE	BEiT	ViT	DINOv2
Sinkhorn route	55.40	43.09	32.73	58.81
Uniform random	12.60	12.60	12.60	21.64
Location-bias null	33.25	31.43	26.60	31.83

Route and backbone attention. Figure 2 compares expected route positions with backbone attention rollout on the same CUB image. The route is learned for the downstream S4 readout. Rollout multiplies head-averaged layer attention with the residual correction $\frac{1}{2}A_\ell + \frac{1}{2}I$, then normalizes the CLS row over patches. Table 5 reports agreement on the first 256 CUB test images. MAE rollout has mean patch variance 1.094×10^{-6}, indicating a relatively flat relevance field; its rank agreement is interpreted alongside this variation.

Table 5: Mean per-image agreement between Sinkhorn scores and backbone attention on 256 shared CUB test images. Heads are averaged. MAE/BEiT/ViT use 196 patches; DINOv2 base uses 256 patches and zero register tokens. Top-20 IoU is intersection size divided by union size.

	Rollout		Last-layer CLS	
Backbone	Spearman	Top-20 IoU	Spearman	Top-20 IoU
MAE	.013	.160	.185	.233
BEiT	.201	.193	.526	.296
ViT	.129	.129	.380	.066
DINOv2	.081	.109	.111	.167

6 Evidence-Route Validation

Decoder dependence. A learned image-independent scalar α_k at each routed position gives the fixed-weight readout $W \sum_k \alpha_k \tilde{h}_k + b$. On DINOv2/CUB it reaches $80.62 \pm 1.11\%$, compared with $81.86 \pm 1.23\%$ for S4 under a matched five-seed protocol. MAE fixed weights reach $46.49 \pm 18.17\%$, with substantial seed variation. Table 6 quantifies route geometry: BEiT is compact in five of five S4 seeds and three of five fixed-weight seeds; DINOv2 is compact in all five seeds with either decoder. S4's causal LTI recurrence and the directly parameterized fixed coefficients induce different strengths of spatial concentration. Appendix D gives per-seed results and training-family conventions.

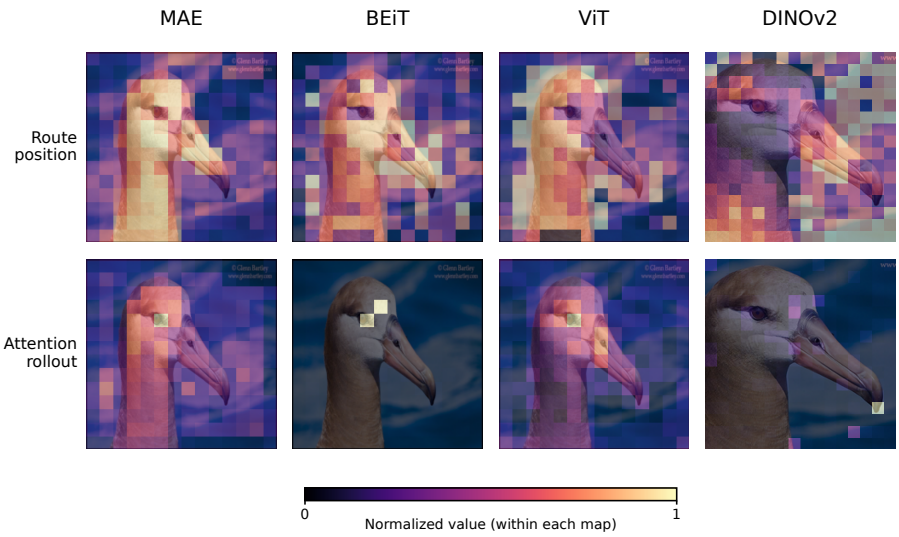


Figure 2: Same-image CUB comparison. Top: normalized expected route position (later positions are hotter). Bottom: attention rollout. Each column uses the corresponding backbone's preprocessed image and full patch grid; MAE patches retain source-grid order. Colors are normalized within each map.

Table 6: Decoder-dependent top-20 compactness on the same 256 CUB test images. Mean ± sample std over seeds 42–46; the final column counts seeds with $R_{20} < 1$ (smaller pairwise grid distance than random patches). These controls use training families distinct from Table 3.

Backbone	Decoder	R_{20}	Compact seeds
MAE	Fixed weights	$.772 \pm .173$	5/5
BEiT	Fixed weights	$.806 \pm .246$	3/5
DINOv2	Fixed weights	$.834 \pm .115$	5/5
BEiT	S4	$.639 \pm .071$	5/5
DINOv2	S4	$.530 \pm .074$	5/5

Ordering and readout interventions. Across five MAE/ImageNet checkpoints, own-router and cross-seed-router accuracies are 71.18% and 71.21%. A separate ordering study gives 71.21% with the learned route and 34.70% after scrambling. On a 512-image subset, pairwise route Spearman is .994 and top-20 overlap is 94.3%; reversal reduces accuracy by 66.3 points, while adjacent swaps reduce it by 0.2. These interventions distinguish global sequencing from local changes. The diagonal S4 has stable modes ($\max_i |\lambda_i| = .98996 \pm .00030$) and a broad coefficient profile ($n_{90} = 105.0 \pm 2.1$ positions). On the 512-image subset, dropping 20 Sinkhorn-ranked tokens reduces accuracy by 19.9 points, compared with 3.1 for CLS-cosine selection and 0.9 for random selection. Appendix C.1 and Appendix B report each protocol, full spectral statistics, saliency comparisons, and perturbation tables.

7 Discussion

Learned ordering improves MAE task accuracy over deterministic traversals, linking sequence utility to both patch selection and positional weighting.

Fixed weights retain most DINOv2 accuracy, while BEiT compactness varies more across seeds than with S4. A shared decoder exposes cross-backbone differences in route geometry; decoder swaps measure geometry's dependence on positional weighting.

CUB masks quantify object-foreground selection. Coordinate-aligned perturbations show stable route ranks under mild noise and larger changes under cropping; attention comparisons measure agreement with backbone relevance maps (Appendix D).

8 Limitations

Task and decoder dependence. The linear Sinkhorn scorer is learned jointly with the downstream readout. Its route reflects frozen features, task supervision, and decoder coefficients. BEiT's five-seed compactness changes from 0.639 ± 0.071 under S4 to 0.806 ± 0.246 under fixed weights, including two fixed-weight seeds above the random-distance baseline. Cross-dataset scorer transfer retains reusable signal but falls 12–21 pp below full retraining (Appendix B.6).

Evaluation scope. Experiments cover classification on ImageNet-1K, CUB, and Cars, with object-mask alignment on CUB. Extension to dense prediction and other downstream tasks requires task-specific evaluation. Router hyperparameter sweeps cover accuracy sensitivity; geometry sensitivity across routing configurations remains an open evaluation question.

Input sensitivity. On MAE/CUB, Gaussian noise of standard deviation .005 yields route Spearman .995 and top-20 overlap 96.91%, while accuracy changes from 59.38% to 57.81% on 128 images. A crop retaining 90% of each dimension yields Spearman .783 and top-20 overlap 69.30% measured by bilinear coordinate alignment over 144 common source-grid centers (Appendix D.4).

9 Conclusion

SSMProbe combines Sinkhorn routing and a causal S4 decoder to measure task-conditioned patch arrangement and fixed positional weighting. Across CUB and Cars, spatially organized routes yield distinct readout profiles. Object masks, spatial nulls, decoder controls, and input perturbations ground this route–readout analysis.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

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

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. arXiv:2104.14294.

Yingyi Chen, Xi Shen, Yahui Liu, Qinghua Tao, and Johan A. K. Suykens. Jigsaw-ViT: Learning jigsaw puzzles in vision transformer. *Pattern Recognition Letters*, 168:53–60, 2023. doi: 10.1016/j.patrec.2022.12.023.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

Marco Cuturi, Olivier Teboul, and Jean-Philippe Vert. Differentiable ranking and sorting using optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Aditya Grover, Eric Wang, Aaron Zweig, and Stefano Ermon. Stochastic optimization of sorting networks via continuous relaxations. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=H1eSS3CcKX.

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

Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. Hippo: Recurrent memory with optimal polynomial projections. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. URL https://arxiv.org/abs/2111.00396. arXiv:2111.00396.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2013.

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

Gonzalo E. Mena, David Belanger, Scott Linderman, and Jasper Snoek. Learning latent permutations with gumbel-sinkhorn networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Michaël Gharbi, Fabian Pedregosa, Armand Joulin, and Pierre Stock. Dinov2: Learning robust visual features without supervision. In *Transactions on Machine Learning Research (TMLR)*, 2024. arXiv:2304.07193.

Sebastian Prillo and Julian Martin Eisenschlos. Softsort: A continuous relaxation for the argsort operator. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research, pages 7793–7802. PMLR, 2020. URL http://proceedings.mlr.press/v119/prillo20a.html.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Xinjian Wu, Fanhu Zeng, Xiudong Wang, Yunhe Wang, and Xinghao Chen. Ppt: Token pruning and pooling for efficient vision transformers. *arXiv preprint arXiv:2310.01812*, 2023. URL https://arxiv.org/abs/2310.01812.

Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal self-attention for local-global interactions in vision transformers. In *Advances in Neural Information Processing Systems*, 2021. arXiv:2107.00641.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Tianyi Zhang, Shangqing Lyu, Yanli Lei, Sicheng Chen, Nan Ying, Yufang He, Yu Zhao, Yunlu Feng, Hwee Kuan Lee, and Guanglei Zhang. Puzzletuning: Explicitly bridge pathological and natural image with puzzles. *arXiv preprint arXiv:2311.06712*, 2023.

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

A Theory and Method Details

A.1 S4 Dynamics and Routed Position Coefficients

The diagonal S4 maps each input channel through a continuous-time state $x(t) \in \mathbb{R}^Q$:

$$x'(t) = \mathbf{A}x(t) + \mathbf{B}u(t), \tag{1}$$

$$y(t) = \mathbf{C}x(t) + \mathbf{D}u(t). \tag{2}$$

The continuous modes are initialized from a stable S4/HiPPO-inspired spectrum [Gu et al., 2020, 2021] and parameterized as

$$\lambda_i = -\mathrm{softplus}(a_i) + \mathrm{i}\,\omega_i,$$

with negative real parts. The bilinear discretization is

$$\bar{\mathbf{A}} = (I - \Delta\mathbf{A}/2)^{-1}(I + \Delta\mathbf{A}/2), \qquad \bar{\mathbf{B}} = (I - \Delta\mathbf{A}/2)^{-1}\Delta\mathbf{B},$$

placing the discrete modes inside the unit disk. Our `torch-s4` implementation uses $d = 768$, $Q = 128$, and dropout 0. The learnable step size is initialized at .05 and clamped to $[10^{-3}, 10^{-1}]$; $\mathbf{D}$ is initialized to zero. Channels share S4 parameters.

For a routed sequence $\tilde{T} = \mathbf{P}^\top T$ of length L, the discrete recurrence and final emitted output are

$$x_k = \bar{\mathbf{A}}x_{k-1} + \bar{\mathbf{B}}\tilde{t}_k, \tag{3}$$

$$z_L = \sum_{k=1}^{L} \mathbf{C}\bar{\mathbf{A}}^{L-k}\bar{\mathbf{B}}\tilde{t}_k + \mathbf{D}\tilde{t}_L. \tag{4}$$

With classifier $Wz_L + b$, define the fixed position coefficient

$$\alpha_k = W\mathbf{C}\bar{\mathbf{A}}^{L-k}\bar{\mathbf{B}} + \mathbf{1}_{\{k=L\}}W\mathbf{D}.$$

For the full route ($L = N$), substituting $\tilde{t}_k = \sum_i P_{i,k}t_i$ gives

$$f(T, \mathbf{P}) = \sum_{i=1}^{N}\left(\sum_{k=1}^{N} P_{i,k}\alpha_k\right)t_i + b.$$

The router assigns patch features to these coefficients through downstream supervision. Slow modes retain evidence over longer horizons; faster modes attenuate earlier contributions. Spectral, skip-path, and sequence perturbation audits measure this interaction.

Ordering properties. GAP satisfies $f(\Pi T) = f(T)$ for any permutation matrix Π; CLS uses the global token. The S4 decoder generally changes when its routed inputs are permuted because α_k varies with position. The full Sinkhorn+S4 probe canonicalizes input permutations through its router. Accordingly, sequence perturbations are applied after routing to measure decoder order dependence.

A.2 Joint Training Protocol with Gradient Isolation

Probe heads share one frozen-backbone forward per batch and use independent optimizers:

Each head uses AdamW (lr=10^{-3}, cosine schedule) on its own parameters; detached backbone features are shared across heads.

A.3 Full Sinkhorn OT Formulation

The per-token scorer $s_i = \mathbf{w}^\top h_i$ is a bias-free $\texttt{Linear}(d, 1)$ layer with default PyTorch initialization. Scores are standardized within each image:

$$\tilde{s}_i = \frac{s_i - \mu(s)}{\sigma(s) + \epsilon}.$$

Canonical positions are $p_j = j/(N-1)$ for $j \in \{0, \ldots, N-1\}$, with cost $\mathbf{C}_{i,j} = (\tilde{s}_i - p_j)^2$. Joint supervision of the scorer and S4 readout yields task-conditioned patch ordering; scorer transfer measures reuse across tasks.

Algorithm 1 Joint Training with Per-Head Gradient Isolation

Input: frozen backbone B, training set $\mathcal{D}$, list of probe heads $\mathcal{H}$
for each head $h \in \mathcal{H}$ **do**
  optimizer[h] $\leftarrow$ AdamW(h.parameters(), lr=10^{-3})
for epoch $= 1$ **to** N_{epochs} **do**
  for (images, labels) $\in \mathcal{D}$ **do**
    features $\leftarrow$ `detach`(B(images)) {no-grad, frozen}
    for head $h \in \mathcal{H}$ **do**
      logits[h] $\leftarrow$ h(features)
      loss[h] $\leftarrow$ CrossEntropy(logits[h], labels)
      loss[h].backward() {gradient isolated per head}
      optimizer[h].step()
      optimizer[h].zero_grad()

The entropy-regularized OT plan:

$$\mathbf{P}_\tau = \underset{\mathbf{P}\in\mathbb{R}_+^{N\times N}}{\operatorname{argmin}} \langle \mathbf{P}, \mathbf{C}\rangle - \tau\mathcal{H}(\mathbf{P}) \quad \text{s.t.} \quad \mathbf{P}\mathbf{1} = \mathbf{1},\ \mathbf{P}^\top\mathbf{1} = \mathbf{1},$$

where $\mathcal{H}(\mathbf{P}) = -\sum_{i,j} \mathbf{P}_{i,j} \log \mathbf{P}_{i,j}$ is elementwise entropy for a matrix of total mass N. Initialize $\mathbf{G} = \exp(-\mathbf{C}/\tau)$ and iterate:

$$\mathbf{u}^{(r+1)} = \mathbf{1} \oslash (\mathbf{G}\mathbf{v}^{(r)}), \tag{5}$$

$$\mathbf{v}^{(r+1)} = \mathbf{1} \oslash (\mathbf{G}^\top\mathbf{u}^{(r+1)}), \tag{6}$$

After R=20 Sinkhorn iterations, $\mathbf{P}^{(R)} = \operatorname{diag}(\mathbf{u}^{(R)})\mathbf{G}\operatorname{diag}(\mathbf{v}^{(R)})$ acts as a soft permutation with temperature τ=0.1.

A.4 Formal Definitions for Evidence-Route Spatial Geometry

This section defines the spatial-information diagnostic and the random-null-calibrated descriptive statistics used in Table 3. Let the N patch tokens form a square grid of side length $G = \sqrt{N}$. Let $\gamma(i) = (r_i, c_i) \in \{0, \dots, G-1\}^2$ denote the original 2D coordinate of source token i. For a trained route score or contribution field, let

$$\mathcal{T}_k \subset \{1, \dots, N\}, \qquad |\mathcal{T}_k| = k,$$

be the set of top-k source-token indices. All metrics are computed per image and averaged over the evaluation subset. For each statistic $m(\mathcal{T}_k)$, we also sample random top-k sets $\mathcal{R}_k$ uniformly without replacement from the same $G \times G$ grid and report a ratio to the random-null expectation when appropriate:

$$\rho_m = \frac{m(\mathcal{T}_k)}{\mathbb{E}_{\mathcal{R}_k}[m(\mathcal{R}_k)]}.$$

For metrics where larger values mean more concentration (largest connected component), compactness corresponds to $\rho_m > 1$; for metrics where smaller values mean more concentration (distance, component count, hull area, MST length, entropy), compactness corresponds to $\rho_m < 1$.

Evidence-route spatial information. Let $\pi_{1:k} = (i_1, \dots, i_k)$ be the ordered top-k route for one image. We split evaluation images into calibration and held-out subsets. Calibration routes provide a smoothed marginal location prior $m(i)$ capturing dataset-level and probe-level location bias. The exchangeable null samples without replacement according to

$$p_0(i_t \mid i_{<t}) = \frac{m(i_t)\mathbf{1}\{i_t \notin i_{<t}\}}{\sum_{j \notin i_{<t}} m(j)}.$$

The spatial alternative uses the same marginal prior but additionally favors locations near previously routed evidence:

$$p_{\text{geo}}(i_t \mid i_{<t}; \sigma) = \frac{m(i_t)\exp\{-d(i_t, i_{<t})^2/(2\sigma^2)\}\mathbf{1}\{i_t \notin i_{<t}\}}{\sum_{j \notin i_{<t}} m(j)\exp\{-d(j, i_{<t})^2/(2\sigma^2)\}},$$

where $d(i_t, i_{<t}) = \min_{q<t} \|\gamma(i_t) - \gamma(i_q)\|_2$. For $t = 1$, the spatial factor is omitted, so both models use the same marginal prior. We select σ from a grid on the calibration split and report the held-out log-likelihood gain

$$\mathrm{ERSI}_k = \frac{1}{Mk \log 2} \sum_{m=1}^{M} \left[\log p_{\mathrm{geo}}(\pi_{1:k}^{(m)}; \hat{\sigma}) - \log p_0(\pi_{1:k}^{(m)}) \right],$$

in bits per routed patch over M held-out images. Positive values indicate spatial predictability beyond marginal location bias; negative values indicate lower held-out likelihood than the null. The grid includes a large-σ endpoint approximating p_0, so selection of this endpoint yields a gain near zero.

Mean pairwise grid distance. The descriptive compactness ratio uses

$$D_k(\mathcal{T}_k) = \frac{2}{k(k-1)} \sum_{\substack{i,j \in \mathcal{T}_k \\ i<j}} \|\gamma(i) - \gamma(j)\|_2, \qquad R_k = \frac{D_k(\mathcal{T}_k)}{\mathbb{E}_{\mathcal{R}_k}[D_k(\mathcal{R}_k)]}.$$

Values $R_k \approx 1$ indicate random-like dispersion; values below one indicate that routed evidence is closer than random top-k patches.

4-neighbor connected components. Construct an undirected graph $\mathcal{G}_k = (\mathcal{T}_k, E_k)$ with edges between horizontally or vertically adjacent selected patches:

$$E_k = \big\{(i,j) : i, j \in \mathcal{T}_k,\ \|\gamma(i) - \gamma(j)\|_1 = 1\big\}.$$

Let $\mathcal{C}_1, \ldots, \mathcal{C}_{M_k}$ be the connected components of $\mathcal{G}_k$. We report the largest-component fraction

$$L_k = \frac{1}{k} \max_{1 \le q \le M_k} |\mathcal{C}_q|$$

and the number of components M_k. A compact route should have larger L_k and smaller M_k than a random top-k set.

Convex hull area. Let $\mathrm{Hull}(\mathcal{T}_k)$ be the Euclidean convex hull of the point set $\{(c_i, r_i) : i \in \mathcal{T}_k\} \subset \mathbb{R}^2$. We normalize its polygon area by the full grid area:

$$A_k = \frac{\mathrm{area}\big(\mathrm{Hull}(\mathcal{T}_k)\big)}{(G-1)^2}.$$

Smaller A_k indicates that top-ranked evidence occupies a smaller spatial envelope.

Euclidean minimum-spanning-tree length. Let $\mathcal{K}_k$ be the complete graph over selected coordinates with edge weight $w_{ij} = \|\gamma(i) - \gamma(j)\|_2$. Let $\mathrm{MST}(\mathcal{T}_k)$ denote its Euclidean minimum spanning tree. We use the normalized tree length

$$S_k = \frac{1}{\sqrt{k}(G-1)} \sum_{(i,j) \in \mathrm{MST}(\mathcal{T}_k)} \|\gamma(i) - \gamma(j)\|_2.$$

This graph-length statistic is less sensitive to a single outlier pair than the full pairwise distance while still measuring spatial spread.

Entropy over coarse spatial bins. Partition the $G \times G$ grid into $B \times B$ coarse bins $\{\mathcal{B}_b\}_{b=1}^{B^2}$; in our audit $B = 4$. Define

$$p_b = \frac{1}{k} \sum_{i \in \mathcal{T}_k} \mathbf{1}\{\gamma(i) \in \mathcal{B}_b\}.$$

The normalized spatial-bin entropy is

$$H_k = -\frac{1}{\log(B^2)} \sum_{b=1}^{B^2} p_b \log(p_b + \epsilon),$$

with a small ϵ only for numerical stability. Lower entropy indicates that routed evidence occupies fewer coarse image regions.

Descriptive robustness checks. The descriptive audit uses the random-null thresholds $R_k < 0.9$, $L_k/\mathbb{E}[L_k^{\text{rand}}] > 1.5$, $M_k/\mathbb{E}[M_k^{\text{rand}}] < 0.85$, $A_k/\mathbb{E}[A_k^{\text{rand}}] < 0.85$, $S_k/\mathbb{E}[S_k^{\text{rand}}] < 0.9$, and $H_k/\mathbb{E}[H_k^{\text{rand}}] < 0.9$ to compare concentration across geometric descriptors.

A.5 Fixed Traversal Families

We evaluate four families of deterministic scan orders designed to linearize the 2D grid of patches (assumed to form an $\sqrt{N} \times \sqrt{N}$ grid):

Raster scan. A single row-major traversal (left-to-right, top-to-bottom).

4-dir (VMamba). Four scans across rows and columns, including row-major forward and reverse, and column-major forward and reverse [Liu et al., 2024]. This follows the SS2D (2D Selective Scan) strategy introduced in VMamba.

4-dir (Diagonal). Concatenate anti-diagonals in increasing order of $r+c$, visiting rows in increasing order within each anti-diagonal. A horizontal reflection gives a second complete traversal. Each traversal and its reverse form the four paths.

4-dir (Snake). Alternating directions at each boundary (e.g., left-to-right for even rows, right-to-left for odd rows), performed both row-wise and column-wise.

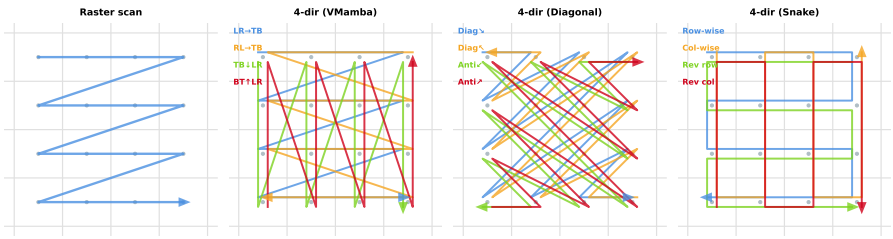


Figure 3: Four scan strategies: (1) raster (row-major), (2) 4-dir VMamba (row/column sweeps), (3) 4-dir diagonal, (4) 4-dir snake (alternating directions).

For the MAE results in Tables 1 and 2, source-grid patch order is restored with `ids_restore`. Each four-direction path visits all $N = 196$ patches once. The same S4 layer processes each path, and its four final states are averaged before classification. Raster uses one complete row-major path. Four complete paths process $4N$ tokens per image.

BEiT and supervised ViT use four complete paths of 196 patches; DINOv2 uses four complete paths of 256 patches. Each path is a permutation of the full source grid. The four paths share one S4 layer and average their final states before classification, giving 784 patch visits for MAE/BEiT/ViT and 1,024 for DINOv2.

A.6 Probe Implementation Details

Checkpoint and reporting protocol. Each head trains for the full fixed budget: 5 epochs on ImageNet-1K and 100 epochs on CUB/Cars. We report accuracy from the final checkpoint at the end of training. Intermediate evaluations track training progress.

Seeding. All random seeds (Python, NumPy, PyTorch CUDA) set to a common value per run. Data order fixed by the dataset loader's default shuffling with a seeded generator.

Baselines. (i) **Attention pooling**: single learned query token attending to all patches via dot-product attention (single head, no FFN). (ii) **Content-weighted pooling**: softmax token-wise weights from the per-token linear scorer without reordering. (iii) **Top-k pooling**: average of top-k highest-scored tokens. These disentangle dynamic token weighting from dynamic token ordering.

Bi-GRU. The fine-grained Bi-GRU baseline receives source-grid-ordered patches in Raster order. It uses a single bidirectional GRU with 128 hidden units per direction and a linear classifier on the concatenated terminal states.

B Extended Results

Protocol map. Table 7 identifies the comparison and checkpoint family for each study. Accuracy changes are compared within each family.

Table 7: Result families and evaluation settings.

Table	Setting
Table 1	MAE/ImageNet main comparison
Table 2	CUB/Cars cross-backbone main comparison
Table 8	Separate MAE/CUB routing family
Table 11	Separate MAE/CUB temperature/iteration sweep
Table 13	Separate MAE/CUB state-size sweep
Table 9	Separate ImageNet ordering-sensitivity family

B.1 Layer-wise Representation Evolution

We evaluate Sinkhorn, CLS, and GAP readouts at layers 3, 6, 9, and 12 of frozen supervised ViT, DINOv2, BEiT, and MAE on CUB-200-2011 (Figure 4).

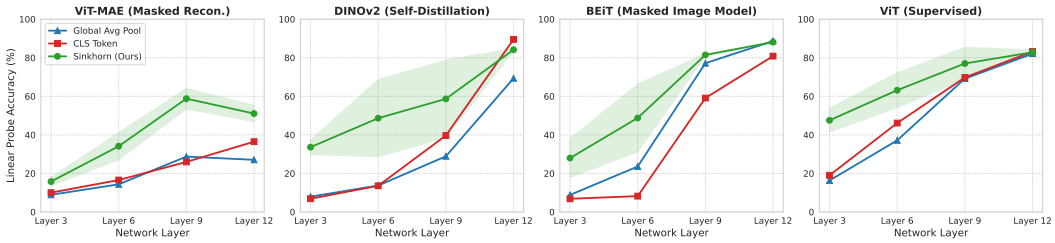


Figure 4: Layer-wise evaluation of probing architectures. We extract token representations at intermediate layers (3, 6, 9, 12) of four distinct Vision Transformer families and evaluate them using Sinkhorn, CLS-based, and GAP-based linear probes. Sinkhorn routing (green) often extracts useful patch evidence earlier than CLS-based probing (red), indicating that route-based readouts can reveal intermediate-layer patch structure.

At DINOv2 layer 3, Sinkhorn reaches 33.62% against 6.87% for CLS, showing task-useful patch evidence at intermediate depth. MAE reaches 58.79% with Sinkhorn against 36.46% with CLS in this layerwise protocol. Sinkhorn accuracy generally increases with depth across the four backbones.

B.2 Routing Validation and Ordering Dependence

We compare Sinkhorn, NeuralSort, and SoftSort on frozen MAE/CUB features, holding the linear scorer, S4 decoder, and optimization settings fixed.

Readout accuracy. Sinkhorn reaches 60.32%, SoftSort 54.07%, and NeuralSort 29.51%, compared with GAP at 19.57% and CLS at 29.01% (Table 8).

Assignment coverage. We measure the fraction of target positions selected by row-wise argmax, the normalized entropy of their histogram, and boundary mass (Figure 5). Sinkhorn covers 112/196 positions (.571) with entropy .868 and row-max p95 .047. NeuralSort selects one position (.005 coverage, entropy 0), with mean row maximum .0077. SoftSort covers 24/196 positions (.122), with entropy .235, mass .883 in the outer 10% of ranks, and row-max p95 .127.

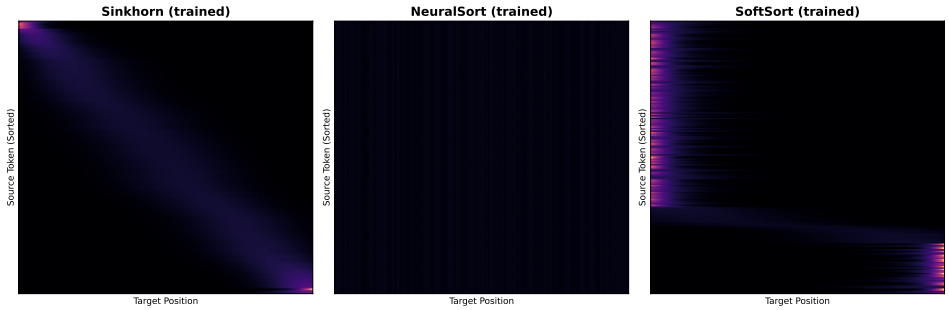


Figure 5: Differentiable assignment matrices, with tokens sorted by each method's argmax target position. Sinkhorn distributes assignments across ranks; NeuralSort has low target coverage; SoftSort concentrates mass near boundary ranks.

Table 8: Differentiable routing ablation on frozen MAE/CUB features.

Routing Mechanism	Method Type	Perm. Invariant?	Top-1 Acc. (%)
GAP	Global Pooling	✓	19.57
CLS	Token Extraction	✓	29.01
NeuralSort	Deterministic Soft-Rank	×	29.51
SoftSort	Continuous Sorting	×	54.07
Sinkhorn (Ours)	Doubly-Stochastic OT	×	**60.32**

For the ImageNet ordering study, we randomly permute the soft-routed sequence $\tilde{X} = \mathbf{P}^\top X$ before S4. Table 9 reports five-seed results for this checkpoint family.

Table 9: ImageNet ordering sensitivity: five-seed accuracy and percentage-point change from the learned route.

Configuration	ImageNet-1K Acc.	Drop from Normal
Normal routing (Sinkhorn)	$71.21 \pm 0.02\%$	–
Scramble after routing	$34.70 \pm 0.60\%$	-36.51%
No routing (S4 on raw tokens)	$21.12 \pm 0.78\%$	-50.09%
Random permutation before S4	$20.80 \pm 0.41\%$	-50.41%

Scrambling reduces accuracy by 36.51 percentage points. Raw-token and random-permutation inputs reach 21.12% and 20.80%, respectively, measuring the trained decoder's dependence on the learned sequence.

B.3 Probe Parameters and Computational Budgets

Table 10 reports exact parameter counts and forward-operation estimates obtained from the implemented probe heads using `fvcore` JIT tracing with custom handlers for FFTs, complex arithmetic, Sinkhorn normalization, and GRU gates. Profiling uses batch size one, 196 patch tokens of width 768, 1,000 output classes, and evaluation mode; the frozen backbone is excluded. We count a real multiply–add as two operations and a complex multiply–accumulate as eight, with a radix-2 real-FFT estimate of $2.5L \log_2 L$ per transform and inverse normalization counted separately. Pointwise arithmetic and reductions are included; indexing, sorting, and memory traffic are excluded. S4 constructs its complex convolution kernel on each forward pass. Four complete paths share parameters, make 784 patch visits, and classify their averaged final states once. Table 23 separately reports measured latency. The main Transformer row uses width 96; Transformer ref. uses width 768.

Table 10: Exact probe parameter counts and estimated forward FLOPs for 196 patch tokens of width 768 and 1,000 output classes (batch size one; frozen backbone excluded). Four-path scans cover the complete grid on each path.

Probe Head	Params (M)	Est. FLOPs (M)
GAP (linear)	0.769	1.69
CLS (linear)	0.769	1.54
Attention Pool	0.770	2.14
Content-Weighted Pool	0.770	2.14
S4 Raster (state dim 128)	0.967	184.97
VMamba / Diagonal / Snake (four paths)	0.967	735.28
Sinkhorn (S4 + scorer, I=20)	0.968	250.63
Sinkhorn (S4 + scorer, I=10)	0.968	247.55
Sinkhorn (S4 + scorer, I=5)	0.968	246.01
Transformer	0.992	75.70
Transformer ref.	3.430	1170.05
DeepSets	1.066	118.53
Random-Fixed Perm + S4	0.967	184.97
Random-Dynamic Perm + S4	0.967	184.97
Sinkhorn + 1D-Conv	0.931	355.43
Bi-GRU	0.947	271.41

B.4 Sinkhorn Hyperparameter Grid Search

We sweep Sinkhorn iterations $I \in \{1, 5, 10, 20\}$ and temperature $\tau \in \{.1, .2, .5, 1.0\}$ on frozen MAE/CUB features (Table 11). Five-seed mean accuracy spans 59.79–61.71% at $\tau \in \{.1, .2\}$ and 49.86–55.70% at $\tau = 1$. The default $I = 20$, $\tau = .1$ gives the highest mean in this grid.

Table 11: Sinkhorn iteration and temperature sweep on frozen MAE/CUB features (five-seed mean $\pm$ std).

	Temperature τ			
Sinkhorn Iterations I	0.1	0.2	0.5	1.0
1	60.12 ± 0.86	60.98 ± 0.72	59.76 ± 1.57	55.70 ± 1.11
5	59.92 ± 1.85	60.64 ± 2.15	56.79 ± 2.07	49.86 ± 1.56
10	60.82 ± 1.03	60.22 ± 1.29	54.91 ± 2.62	50.04 ± 1.37
20	$\mathbf{61.71 \pm 1.00}$	59.79 ± 0.71	56.81 ± 1.94	50.24 ± 1.64

B.5 Token Sensitivity Analysis

We evaluate seed-42 final MAE/ImageNet probe checkpoints on the first 2,048 validation images. The backbone uses bfloat16 with masking disabled, restores source-grid order with `ids_restore`, and supplies float32 features to the heads. Complete four-path traversals retain the relative order of their remaining source patches after removal. Random removal uses per-image seed-42 subsets nested across k.

Token dropping. We rank patches by CLS cosine and remove the top-k, bottom-k, or random k tokens (Figure 6). At $k = 100$, top-ranked removal reduces Raster, Snake, Sinkhorn, Bi-GRU, and Transformer accuracy by 10.89, 24.51, 13.53, 14.16, and 11.72 percentage points, respectively. Random removal reduces the same readouts by 4.10, 15.14, 2.05, 3.96, and 2.15 points. Bottom-ranked removal gives intermediate decreases of 9.72, 20.70, 6.88, 9.67, and 7.76 points.

Token permutation. We shuffle, reverse, or reorder tokens by ascending/descending CLS similarity before inference (Figure 7). Shuffling reduces Raster, Snake, and Bi-GRU accuracy by 1.86, 6.30, and 5.76 points; reversal changes them by -2.78, $-.63$, and -17.87 points. GAP, Sinkhorn, and the

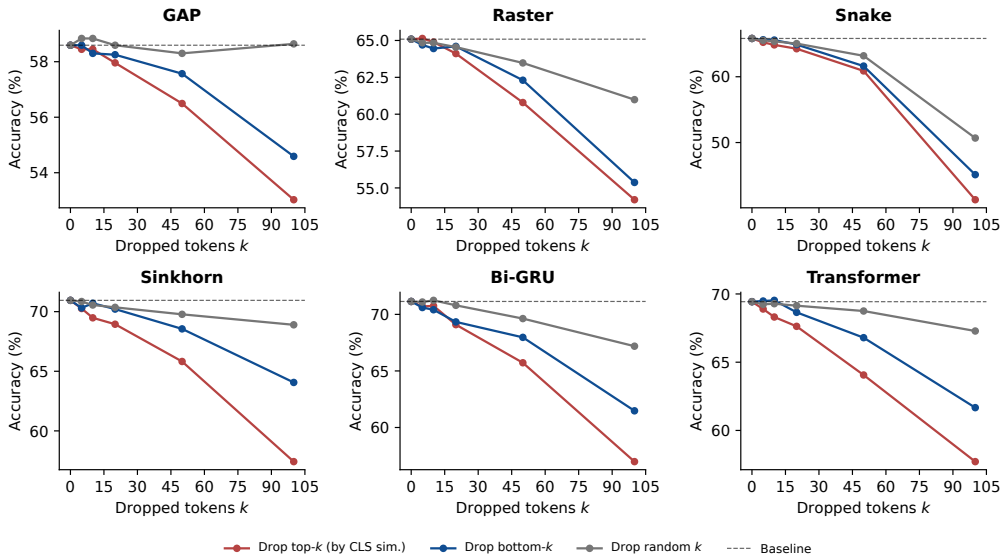


Figure 6: Token dropping on 2,048 MAE/ImageNet validation images, using seed-42 final checkpoints. Six panels show GAP, Raster, complete-path Snake, Sinkhorn, Raster Bi-GRU, and the width-96 Transformer. Top/bottom ranking uses CLS cosine similarity; random removal uses per-image subsets. Dashed lines show intact-token accuracy on the same images.

Transformer retain their accuracy under all four input reorderings. VMamba-style and diagonal path sets are closed under reversal, so their averaged readout preserves accuracy under that intervention. Random-dynamic order is sampled anew with a reset seed for each intervention.

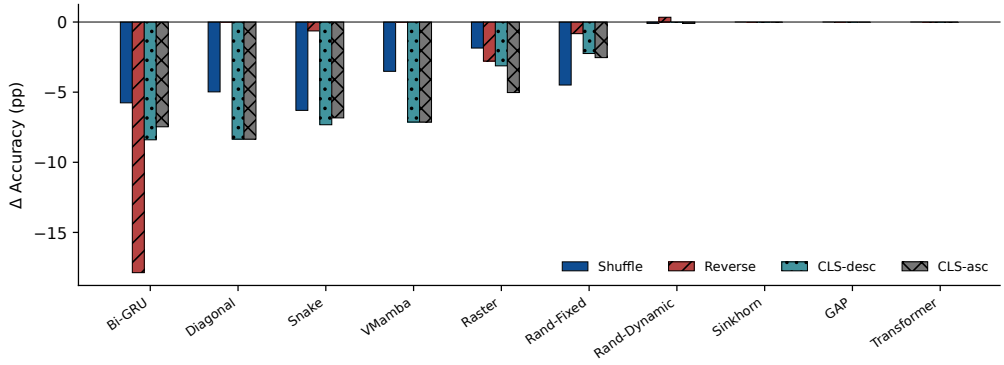


Figure 7: Input-order sensitivity for ten MAE/ImageNet readouts on the same 2,048 validation images. Bars show accuracy changes in percentage points relative to each seed-42 final checkpoint's intact-token baseline. Four-direction scans use complete source-grid paths; Transformer denotes the width-96 control.

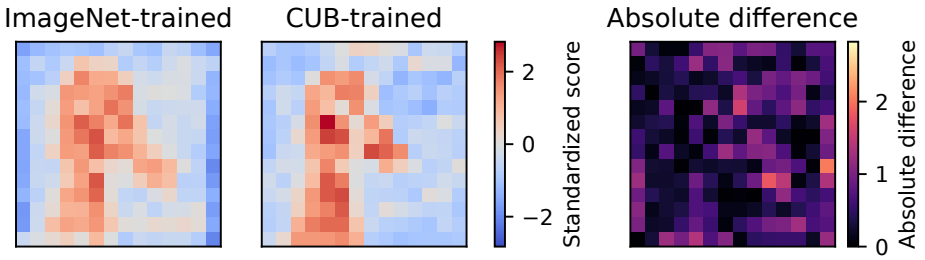
B.6 Cross-Dataset Scorer Transfer

We freeze the bias-free MAE scorer $s_i = \mathbf{w}^\top h_i$ learned on CUB or Cars and train the S4 decoder and classifier on the other dataset. Table 12 compares both transfer directions with full scorer/decoder training and a frozen random scorer.

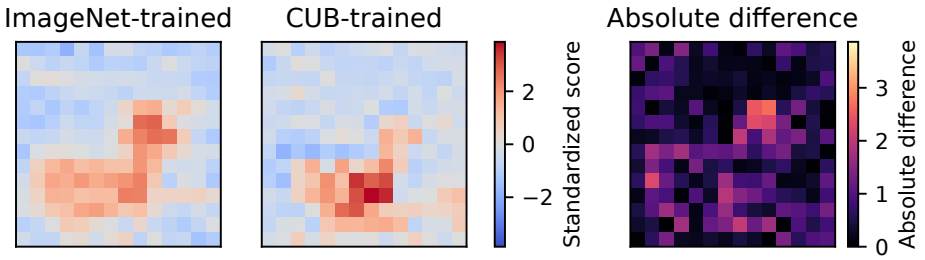
Transfer exceeds the random scorer by 11.5 points on Cars and 19.4 on CUB, while falling 12.2 and 21.0 points below full training. The 768-parameter scorer retains task-useful signal across both directions, with a measurable benefit from target-task adaptation.

Table 12: Cross-dataset scorer transfer on frozen MAE features. "Transfer": scorer loaded from source checkpoint, only S4+classifier retrained. "Full retrain": all parameters (including scorer) trained from scratch on target. "Random": random scorer direction, frozen.

Source	Target	Scorer	Test Acc. (%)
CUB	Cars	Transfer (CUB frozen)	41.2
CUB	Cars	Full retrain	53.4
CUB	Cars	Random (frozen)	29.7
Cars	CUB	Transfer (Cars frozen)	40.2
Cars	CUB	Full retrain	61.2
Cars	CUB	Random (frozen)	20.8



CUB sample 000



CUB sample 001

Figure 8: Task-adapted scorer heatmaps on two fixed CUB test images, with MAE patches restored to source-grid order. Each triplet shows standardized patch scores from an ImageNet-trained scorer, a CUB-trained scorer, and their absolute difference. Signed fields share a diverging scale; absolute differences use a separate nonnegative scale.

B.7 State-Size Scaling

We vary S4 state dimension on frozen MAE/CUB features with Sinkhorn settings $I = 20$, $\tau = .1$. Mean accuracy rises from 55.22% at dimension 1 to 60.42% at dimension 4, then spans 59.19–60.79% across dimensions 8–256 (Table 13).

C Validation and Reproducibility Details

C.1 Eigenvalue Spectrum of the Trained S4 Decoder

Protocol. We extract the trained diagonal $\bar{A}$ and compute its 128 complex discrete modes $\{\lambda_i\}$. We report the minimum, maximum, and mean $|\lambda_i|$, the number of slow modes near the unit circle, and the position-wise recurrent coefficient norm $\|\alpha_k\|$ where $\alpha_k = W \cdot \Re(\sum_i C_i \lambda_i^{N-k} \bar{B}_i) + \mathbf{1}_{\{k=N\}} W\mathbf{D}$ for $k \in \{1, \ldots, N\}$ under the final-emitted-state notation. Here C_i is the output coefficient of mode i and $\bar{B}_i$ its discretized input coefficient. We also audit the learned D skip term by evaluating checkpoints with and without the skip path.

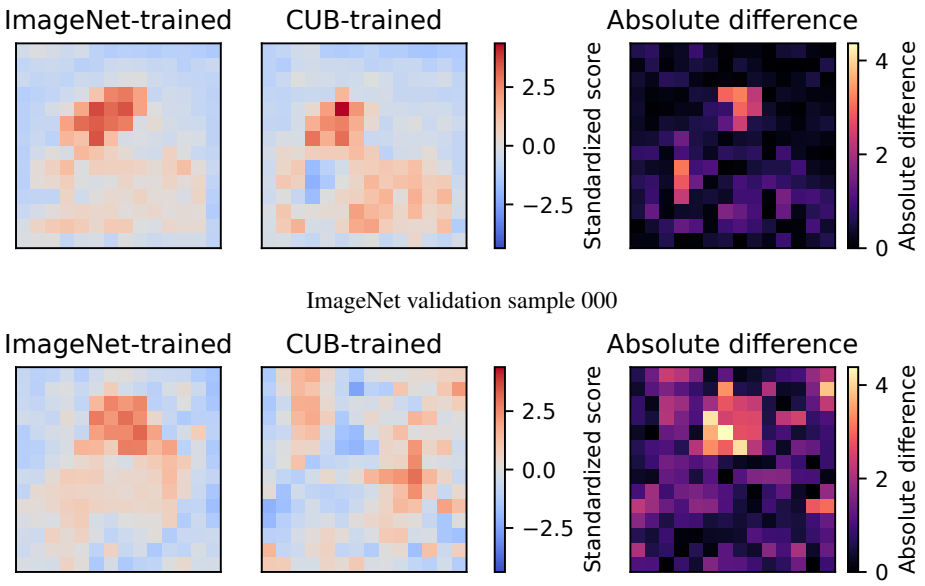


Figure 9: Task-adapted scorer fields on the first two ImageNet validation images, using source-aligned MAE patches. Each row compares the ImageNet-trained and CUB-trained scorers on the same image, followed by the absolute difference. Signed fields share a diverging scale within each row; differences use a separate nonnegative scale.

Table 13: S4 state-size sweep on frozen MAE/CUB features with Sinkhorn routing ($I = 20$, $\tau = .1$; five-seed mean $\pm$ std).

State Dim. N	Top-1 Acc. (%)	Δ vs. GAP
1	55.22 $\pm$ 1.72	+35.65
2	57.51 $\pm$ 2.07	+37.94
4	60.42 $\pm$ 0.64	+40.85
8	60.55 $\pm$ 1.49	+40.98
16	60.71 $\pm$ 0.83	+41.14
32	59.58 $\pm$ 2.19	+40.01
64	60.79 $\pm$ 2.14	+41.22
128	60.19 $\pm$ 2.01	+40.62
256	59.19 $\pm$ 0.87	+39.62
Permutation-invariant baselines (from table 8)		
GAP	19.57	–
CLS	29.01	–

Results. The trained S4 decoders have stable discrete modes. On the five MAE/ImageNet-1K Sinkhorn checkpoints, the dominant mode satisfies $\max_i |\lambda_i| \approx 0.990$ while many modes decay faster (mean $|\lambda_i| \approx 0.306$). This gives a stable LTI kernel combining slow long-memory modes and faster attenuation.

C.1.1 Post-Hoc Router Interventions

Protocol. We evaluate frozen Sinkhorn checkpoints on 2,048 ImageNet validation images (main seeds 42–46) and 1,024 CUB test images. Sequence interventions include reversal, shuffling, fixed random raw-token order, and ordering by CLS-patch cosine similarity or a random scorer. Cross-seed reuse feeds one scorer's routed sequence to another checkpoint's S4/classifier. Token removal selects

the top or bottom 20 patches by Sinkhorn score, CLS cosine, or token norm, alongside random removal.

Table 14: Post-hoc router interventions. ImageNet uses five MAE checkpoints on 2,048 images; CUB uses one MAE/CUB checkpoint and two ImageNet-trained routers on 1,024 images. Changes are percentage points; Spearman coefficients and accuracy rows use their stated units.

Quantity	MAE/ImageNet	MAE/CUB
Own learned order acc.	71.18 ± 0.36	55.86
Alternate router order acc.	71.21 ± 0.30	41.21
Shuffle after routing drop	-52.14 ± 0.31	-43.75
Random raw-order drop	-50.46 ± 0.41	-50.10
Drop top-20 Sinkhorn tokens	-16.34 ± 0.75	-41.89
Drop top-20 CLS-cosine tokens	-2.02 ± 0.14	-8.69
Drop random 20 tokens	-0.41 ± 0.34	-1.66
Sinkhorn vs. CLS-cosine Spearman	-0.04 ± 0.01	-0.20
Sinkhorn vs. token-norm Spearman	-0.19 ± 0.00	$+0.01$

Results. Cross-seed router replacement changes ImageNet accuracy by +.03 points on average, with score-rank Spearman .994 and top-20 overlap 94.5%. ImageNet scorers paired with the CUB decoder reach 41.21%, compared with 55.86% for its own scorer. Shuffling the routed sequence reduces accuracy by 52.14 points on ImageNet and 43.75 on CUB. Removing the top 20 Sinkhorn patches produces larger drops than CLS-cosine or random removal (Table 14), connecting scorer selection to the trained readout.

512-image perturbations. On 512 validation images with the same five ImageNet checkpoints, global rearrangements reduce accuracy by 48–66 points, while adjacent swaps change it by $.16 \pm .15$ points (Table 15). Cross-seed Spearman is .994 with top-20 overlap 94.3%. Sinkhorn scores have Spearman $-.041$ and overlap 15.5% with CLS cosine, and $-.200$ and 7.5% with token norm. Table 16 compares token removal on this subset.

Table 15: Routed-sequence perturbations on 512 ImageNet validation images (five-seed mean $\pm$ std). Changes are paired percentage-point differences from clean accuracy.

Sequence	Accuracy (%)	Change (pp)
Learned route	$73.87 \pm .70$	–
Reversed	$7.62 \pm .49$	$-66.25 \pm .50$
Half-sequence cyclic shift	$23.28 \pm .38$	$-50.59 \pm .65$
Four-block shuffle	$25.70 \pm .62$	$-48.16 \pm .26$
Full shuffle	$18.20 \pm .65$	$-55.66 \pm .85$
Adjacent swaps	$73.71 \pm .62$	$-.16 \pm .15$

Table 16: Accuracy decrease after removing 20 patches on the same 512 ImageNet images (five-seed mean $\pm$ std, percentage points).

Selection	Sinkhorn	CLS cosine	Token norm	Random
Decrease (pp)	19.9 ± 1.2	$3.1 \pm .5$	$.0 \pm .4$	$.9 \pm .6$

C.1.2 True-Class Logit under Permutation Controls

Protocol. We use the seed-44 MAE/ImageNet checkpoint and the first five validation images, selected before inspecting routes. Soft routing and hard permutations use identical source tokens and sequence lengths. The percentile is the fraction of hard permutations whose true-class logit is at most the soft readout's logit. Analytical S4 logits match direct forwards on 32 random permutations per image and length, with maximum absolute error below 5×10^{-6}.

Scale 1: $N = 9$**.** We subsample source-grid rows and columns 5–7 (zero-based indices 75, 76, 77, 89, 90, 91, 103, 104, 105) and enumerate all $9! = 362{,}880$ permutations. This restricted central-region readout has 4/5 correct predictions and a mean percentile of 69.30%.

Table 17: Exact enumeration of hard permutations on nine central source patches. "Higher logits" counts permutations exceeding the soft Sinkhorn true-class logit.

Image	True Class	Correct?	Higher logits	Percentile
1	91 (lizard)	✓	67,018 / 362,880	81.53%
2	171 (greyhound)	×	109,104 / 362,880	69.93%
3	980 (volcano)	✓	157,922 / 362,880	56.48%
4	218 (Komondor)	✓	120,960 / 362,880	66.67%
5	19 (bulbul)	✓	102,095 / 362,880	71.87%
Mean		4/5	111,419.8	**69.30%**

Scale 2: $N = 196$**.** We sample 7.2×10^6 random permutations per image (36×10^6 total) from the 196! space, using sampling seeds 42–46 in image order. The full-grid soft readout has 5/5 correct predictions and an aggregate percentile of 97.43%.

Table 18: Random permutation sampling on full $N = 196$ patches (MAE backbone, ImageNet-1K).

Image	True Class	Correct?	Higher logits	Percentile
1	91 (lizard)	✓	136,680 / 7.2M	98.10%
2	171 (greyhound)	✓	85,185 / 7.2M	98.82%
3	980 (volcano)	✓	177,012 / 7.2M	97.54%
4	218 (Komondor)	✓	153,158 / 7.2M	97.87%
5	19 (bulbul)	✓	373,304 / 7.2M	94.82%
Aggregate (36M total)		5/5	925,339 / 36M	**97.43%**

D Semantic Alignment and Decoder Controls

Evaluation conventions. CUB controls use frozen backbones, the official train/test split, batch size 256, and 100 epochs. Five-seed summaries use seeds 42–46 and sample standard deviations. Settings are matched within each training family. The MAE temperature-control family has baseline $61.64 \pm 1.11\%$, while the main comparison has $56.67 \pm 7.54\%$ (Table 2); each retains its own seed summary. Cars and ImageNet audits specify their image sets below.

D.1 Foreground alignment

We match all 5,794 CUB test images to the official foreground segmentation masks. Each mask undergoes the spatial preprocessing of its corresponding backbone image processor. A patch is foreground when more than half its transformed pixels belong to the bird mask. We select the 20 source patches with the largest expected routed positions. The uniform random-patch baseline measures foreground coverage on the same grid; the location-bias null retains the route's marginal preference for grid locations while removing image-specific selection. Table 4 reports the resulting foreground fractions. MAE, BEiT, and supervised ViT use 14×14 grids (196 patches); DINOv2 uses a 16×16 grid (256 patches). The different processors explain the differing random foreground coverage.

D.2 Fixed-position decoder and seed variation

Section 6 defines the fixed-position readout and reports aggregate decoder comparisons. The fixed-position MAE decoder reaches $46.49 \pm 18.17\%$ (per-seed accuracies: 22.28, 59.48, 31.53, 58.96, and 60.18%).

Table 19: Per-seed top-20 compactness on the same 256 CUB test images, ordered by seeds 42–46. Aggregate summaries appear in Table 6.

Backbone	Decoder	Per-seed R_{20}
MAE	Fixed weights	.700, .933, .744, .533, .950
BEiT	Fixed weights	.687, 1.075, 1.069, .606, .591
DINOv2	Fixed weights	.910, .916, .646, .800, .898
BEiT	S4	.681, .631, .733, .598, .550
DINOv2	S4	.654, .535, .470, .517, .477

Per-seed coefficient flatness and test accuracy show no consistent separation between compact and noncompact BEiT fixed-weight runs.

D.3 Agreement with backbone attention

We compare maps on the same first 256 CUB test images across all four backbones. Attention is averaged over heads. For rollout, each layer contributes $\frac{1}{2}A_\ell + \frac{1}{2}I$; the product of these matrices supplies the CLS-to-patch row, renormalized over patches. Last-layer attention uses the head-averaged final CLS-to-patch row. Spearman correlation compares route scores with each attention map; top-20 IoU is intersection size divided by union size.

Table 5 reports all four backbones using their full patch grids.

MAE masking is disabled. Its rollout patch variance has mean 1.094×10^{-6} and minimum 1.806×10^{-7}, indicating a relatively flat relevance field. Route and attention agreement is reported alongside this within-image variation.

D.4 Input perturbations

On the same first 128 CUB test images, the MAE/S4 checkpoint yields 59.375% clean accuracy and 57.8125% with Gaussian noise of standard deviation .005 added after preprocessing. Noise uses a dedicated seed-42 generator, shared between accuracy/geometry and rank-agreement audits. Clean/noisy R_{20} is .6702/.6695, and ERSI is .4895/.4883 bits per patch. Clean-to-noisy route Spearman is .9953, top-20 overlap (intersection divided by 20) is 96.9141%, and normalized expected-position displacement is .0109. Two independent clean forwards give Spearman 1.000, top-20 overlap 100%, and displacement below 5×10^{-7}.

A center crop retaining 90% of each spatial dimension (81% of image area) gives Spearman .6639, top-20 overlap 62.6563%, and displacement .1546 when compared at relative resized-grid positions. For a source-coordinate comparison, we map the cropped route field to 144 common interior source-grid patch centers using bilinear sampling. Within that shared domain, Spearman is .7832, top-20 overlap is 69.2969%, and displacement is .1241 on the original 196-position route scale. The crop changes image content and preprocessing; the aligned measurements also include interpolation.

D.5 Source-grid shape descriptors

On the first 128 test images of each dataset, we select the 20 largest MAE route scores after restoring patch tokens to the source 14×14 grid. Table 20 reports complementary geometric descriptors. Connected components use four-neighbor adjacency; hull area is divided by the squared grid span; MST length is divided by $\sqrt{20}$ times the grid span; entropy uses a 4×4 spatial binning. The random column averages 2,048 uniform patch subsets with seed 42. The 128-image compactness summaries and the 512-image main geometry audit use distinct image counts.

For the fixed seed-selected CUB test image at index 5238, MAE's top-20 source patches form three components, with 18 patches in the largest component. The top-20 score and absolute-contribution sets overlap by 97.23% on the 128-image CUB subset, linking the score-based descriptors to the coefficient-based readout audit.

Table 20: MAE top-20 source-grid shapes (128 images per dataset).

Descriptor	CUB	Cars	Random
Mean pairwise grid distance	4.910	4.010	7.321
R_{20}	.671	.548	1.000
Largest component fraction	.588	.675	.133
Number of components	6.547	3.477	16.482
Convex hull area fraction	.354	.176	.729
Normalized MST length	.492	.383	.748
Normalized spatial entropy	.632	.567	.847

D.6 MAE/ImageNet source-grid geometry

We evaluate the five main MAE/ImageNet Sinkhorn checkpoints (seeds 42–46) on the same first 512 validation images. Patch tokens are restored with `ids_restore` before assigning source-grid coordinates. A seed-42 split assigns 256 routes to calibration and 256 to held-out evaluation; each checkpoint's spatial scale is selected on calibration routes. The random compactness reference averages 2,048 uniform top-20 subsets. Frozen features use bfloat16 and readouts use float32.

Table 21: MAE/ImageNet source-grid geometry on 512 validation images. Error bars are sample standard deviations across the five checkpoints.

Seed	ERSI (bits/patch)	σ	R_{20}
42	.3088	2	.5806
43	.3196	2	.5791
44	.3073	2	.5834
45	.3043	2	.5828
46	.3137	2	.5784
Mean $\pm$ std	$.3107 \pm .0060$	2	$.5809 \pm .0022$

A separate shape audit on the first 128 validation images gives a largest-component fraction of $.571 \pm .005$, $5.917 \pm .087$ components, hull-area fraction $.241 \pm .002$, normalized MST length $.427 \pm .001$, and spatial entropy $.630 \pm .002$ across seeds. Their random references are .133, 16.482, .729, .748, and .847, respectively. The 128-image compactness ratio is $.575 \pm .003$.

D.7 Source-ordered scan interventions and computation

We evaluate seed-42 final CUB and Cars checkpoints on the first 256 test images. Each MAE multi-direction head averages four S4 final outputs, with each path traversing all 196 source patches. Reversal and shuffling change the input sequence before applying each head's specified traversal. Dropping removes the 20 patches with largest cosine similarity to the frozen CLS feature and filters each original path by the retained source indices. Random-dynamic evaluation uses seed 42 for each condition. These subset accuracies describe within-checkpoint interventions; the five-seed, full-test accuracies appear in Table 2.

Table 22: Source-grid MAE scan interventions. Each cell gives CUB / Cars accuracy (%).

Readout	Clean	Reverse	Shuffle	Drop top-20
Raster	14.45 / 13.67	5.86 / 3.91	5.47 / 3.12	14.45 / 14.84
VMamba-style	12.89 / 18.36	12.89 / 18.36	13.28 / 11.72	12.50 / 17.97
Diagonal	19.92 / 23.05	19.92 / 23.05	14.45 / 16.41	19.53 / 19.92
Snake	18.75 / 17.97	11.72 / 16.02	11.72 / 8.20	17.58 / 17.19
Random-fixed	11.72 / 19.53	8.98 / 8.98	7.81 / 12.89	10.16 / 16.41
Random-dynamic	32.81 / 42.97	35.55 / 45.31	33.98 / 45.31	29.69 / 44.14
Bi-GRU (Raster)	28.12 / 31.25	14.06 / 10.94	22.66 / 15.62	25.00 / 26.95

The VMamba-style and diagonal path sets are closed under sequence reversal, so averaging their four outputs preserves the clean prediction under reversal. Random-dynamic interventions include fresh random traversals and consequently vary with that sampling.

Table 23: MAE/CUB scan head budgets and CPU latency. Batch size 1, float32, four CPU threads, Intel Xeon Gold 6530, PyTorch 2.5.1; median of 30 forwards after five warmups. Latencies include traversal and head computation, with the frozen backbone excluded.

Readout	Parameters	Patch visits	Median (ms)
Raster	351,689	196	9.44
VMamba-style	351,689	784	35.57
Diagonal	351,689	784	34.81
Snake	351,689	784	34.49
Random-fixed	351,689	196	9.50
Random-dynamic	351,689	196	7.32
Bi-GRU (Raster)	741,064	196	10.22

Four-direction heads share their parameters across paths and process four times the single-scan patch visits. CUB heads have 351,689 parameters and Cars heads 348,613. MAE/BEiT/ViT make 784 patch visits per image; DINOv2 makes 1,024. Table 10 reports forward-operation estimates.

Table 24: Full-grid scan CPU latency (ms), shown as CUB / Cars. Batch 1, float32, four CPU threads, Intel Xeon Gold 6530, PyTorch 2.5.1. Each value is the median of 30 forwards after five warmups, with seed-42 randomized interleaving across the 18 heads. Frozen backbone computation is excluded.

Backbone	VMamba-style	Diagonal	Snake
BEiT	39.66 / 31.18	39.44 / 30.92	39.32 / 31.17
ViT	50.69 / 50.85	49.87 / 49.18	50.50 / 46.39
DINOv2	46.98 / 49.80	48.30 / 53.05	46.84 / 48.78

Table 25: Complete-path scan interventions on the first 256 test images per dataset. Each cell gives CUB / Cars accuracy (%). Seed-42 final checkpoints use the full source grid. Dropping removes the 20 source patches with largest CLS cosine similarity from each traversal, preserving the order of the remaining patches.

Readout	Clean	Reverse	Shuffle	Drop top-20
BEiT / VMamba	82.42 / 65.62	82.42 / 65.62	87.89 / 63.67	81.25 / 66.02
BEiT / Diagonal	82.03 / 68.36	82.03 / 68.36	87.11 / 66.41	79.69 / 67.19
BEiT / Snake	82.03 / 66.41	83.20 / 65.23	87.50 / 66.80	80.08 / 66.80
ViT / VMamba	80.86 / 55.08	80.86 / 55.08	81.25 / 52.73	80.47 / 56.64
ViT / Diagonal	82.42 / 56.25	82.42 / 56.25	82.81 / 56.64	80.47 / 57.03
ViT / Snake	81.64 / 57.81	82.03 / 54.30	84.38 / 51.95	80.86 / 58.20
DINOv2 / VMamba	63.67 / 43.75	63.67 / 43.75	56.25 / 32.42	62.89 / 42.58
DINOv2 / Diagonal	59.77 / 46.88	59.77 / 46.88	54.69 / 28.52	58.20 / 46.09
DINOv2 / Snake	63.67 / 48.44	57.81 / 32.42	43.75 / 19.92	61.72 / 49.22

VMamba-style and diagonal path sets preserve their accuracy under sequence reversal. Shuffling changes the source-grid adjacency encountered by the fixed traversals; token removal changes both the available evidence and the effective sequence length. These subset measurements characterize each checkpoint separately from the full-test five-seed summaries.

D.8 Routing hyperparameters and matched readouts

These sweeps report accuracy sensitivity; geometry sensitivity to routing hyperparameters is listed in Section 8.

Table 27 reports the matched DINOv2/CUB readout family under a common training budget.

Table 26: CUB accuracy controls (five-seed mean ± sample std). Temperature and iteration settings are matched within each backbone's sweep. Scorer weight decay uses a separate matched pair.

Backbone	Configuration	Accuracy (%)
DINOv2	$\tau = .05$, 20 iterations	80.64 ± 1.79
DINOv2	$\tau = .10$, 5 iterations	81.22 ± 3.22
DINOv2	$\tau = .10$, 20 iterations	81.19 ± 3.67
DINOv2	$\tau = .20$, 20 iterations	82.72 ± 1.07
MAE	$\tau = .10$, 20 iterations	61.64 ± 1.11
MAE	$\tau = .20$, 20 iterations	$59.76 \pm .78$
DINOv2	Scorer weight decay 0	81.86 ± 1.23
DINOv2	Scorer weight decay 10^{-3}	81.40 ± 3.01

Table 27: Matched DINOv2/CUB readouts from the independent control family (five-seed mean ± sample std). Complete four-path scan results appear in Table 2.

Readout	Accuracy (%)
CLS	$89.45 \pm .21$
GAP	$68.53 \pm .24$
Sinkhorn/S4	81.17 ± 2.03
Larger Transformer	$87.15 \pm .17$
DeepSets	$76.93 \pm .17$
Sinkhorn/1D-convolution	$71.01 \pm .64$
Raster	$35.36 \pm .31$
Random-dynamic	$69.04 \pm .09$
Random-fixed	38.40 ± 7.49
Bi-GRU	59.64 ± 1.57

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state that SSMProbe is an inspectable Sinkhorn+S4 diagnostic for evidence routing, and that route geometry differs across frozen ViT pre-training objectives. These claims are supported by the ImageNet-1K MAE results in Table 1, the cross-backbone fine-grained results in Table 2, and the evidence-route geometry summary in Table 3.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 8 discusses task and decoder dependence, evaluation scope, and input sensitivity.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The appendix provides formal derivations of the LTI dynamics, position-dependent modulation mechanism, and the information scheduling interpretation of the learned Sinkhorn permutation. Assumptions (discrete LTI, causal scanning, permutation sensitivity) are explicitly stated.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4 specifies the backbones, datasets, probe families, optimization (AdamW, lr=10^{-3}, cosine schedule, batch size 256), and epoch counts. Appendix A.6 provides S4 and Sinkhorn implementation details.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code and instructions for reproducing the main experimental results will be released publicly after the camera-ready submission. ImageNet-1K is a standard benchmark with access subject to its terms. Experiment configurations are detailed in the Experimental Setup section.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 provides hyperparameter specifications including backbone, dataset, probe head families, optimizer type (AdamW), learning rate (10^{-3}), schedule (cosine), batch size (256), and number of epochs (5 for ImageNet-1K, 100 for CUB/Cars).

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Multi-seed evaluations use the fixed seed set {42, 43, 44, 45, 46}. Summaries report mean and standard deviation with the number of valid runs and numerical failures stated explicitly. The DINOv2/CUB full-grid VMamba-style scan at seed 44 develops NaN loss during epoch 96 of 100. This run is retained as a numerical failure without replacement or protocol changes; its final NaN-output accuracy is excluded from valid-run accuracy summaries. Random-fixed permutations also exhibit substantial seed variation.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Experiments use a single H100 80GB GPU. Seven-head MAE training shares the frozen backbone forward and takes 19.8–20.2 minutes per CUB seed, 36.5–39.2 minutes per Cars seed (100 epochs), and 2.49–3.26 hours per ImageNet seed (5 epochs). The BEiT/ViT/DINOv2 complete-path controls run as three concurrent jobs, with three heads per job: CUB takes 18.8–24.6 minutes per seed and Cars 37.4–45.9 minutes. Timings include final test evaluation. Combined GPU memory during the concurrent jobs was observed at approximately 32 GiB; this is a spot measurement. Frozen backbones use bfloat16 and heads use float32. Appendix B.3 reports probe parameters and forward-operation estimates.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics?

   Answer: [Yes]

   Justification: This work presents fundamental research on representation probing without direct societal applications. All experiments use established benchmarks and publicly available model checkpoints.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is foundational representation learning research focused on probing methodology. It does not propose new applications with direct societal impact.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse?

    Answer: [N/A]

    Justification: This work uses standard pretrained models and benchmark datasets without release risks. No high-risk dual-use models are involved.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: ViT-MAE-base is credited to He et al. (2022) under Apache 2.0. ImageNet-1K is cited as the standard benchmark. S4 implementation is credited to Gu et al. (2021). DINOv2 is credited to Oquab et al. (2023/2024). CUB-200-2011 (Caltech-UCSD Birds-200-2011) is cited as a standard fine-grained visual classification benchmark. Stanford Cars is cited as a standard fine-grained vehicle classification benchmark.