NEURAL INFORMATION PROCESSING SYSTEMS

Poster ID：93769

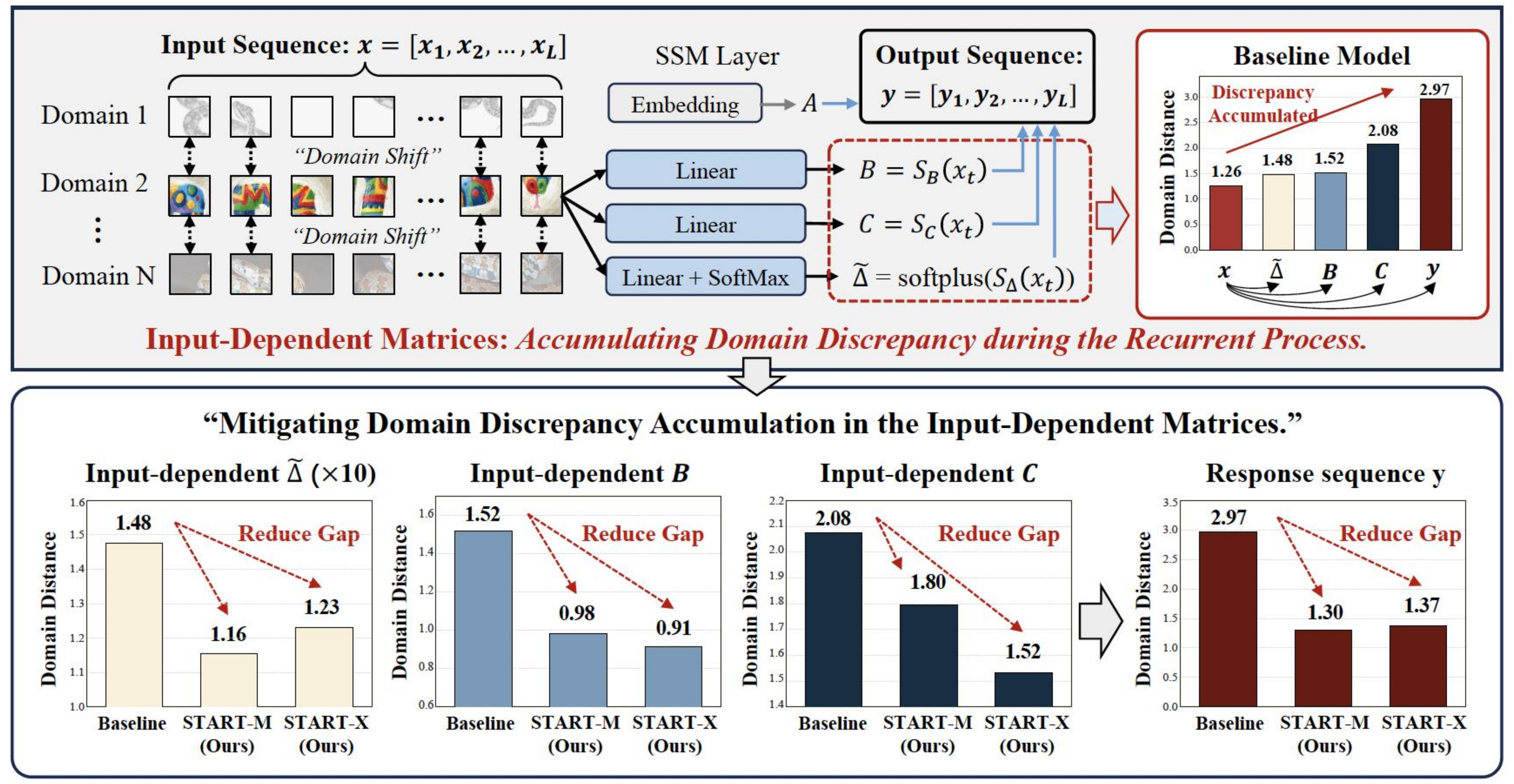
START: A Generalized State Space Model with Saliency-Driven Token-Aware Transformation

Jintao Guo[1] **Lei Qi**[2*] **Yinghuan Shi**[1*] **Yang Gao**[1†]
[1] Nanjing University [2] Southeast University
guojintao@smail.nju.edu.cn, qilei@seu.edu.cn, {syh, gaoy}@nju.edu.cn

南京大學 NANJING UNIVERSITY

東南大學 SOUTHEAST UNIVERSITY

Summary of highlights



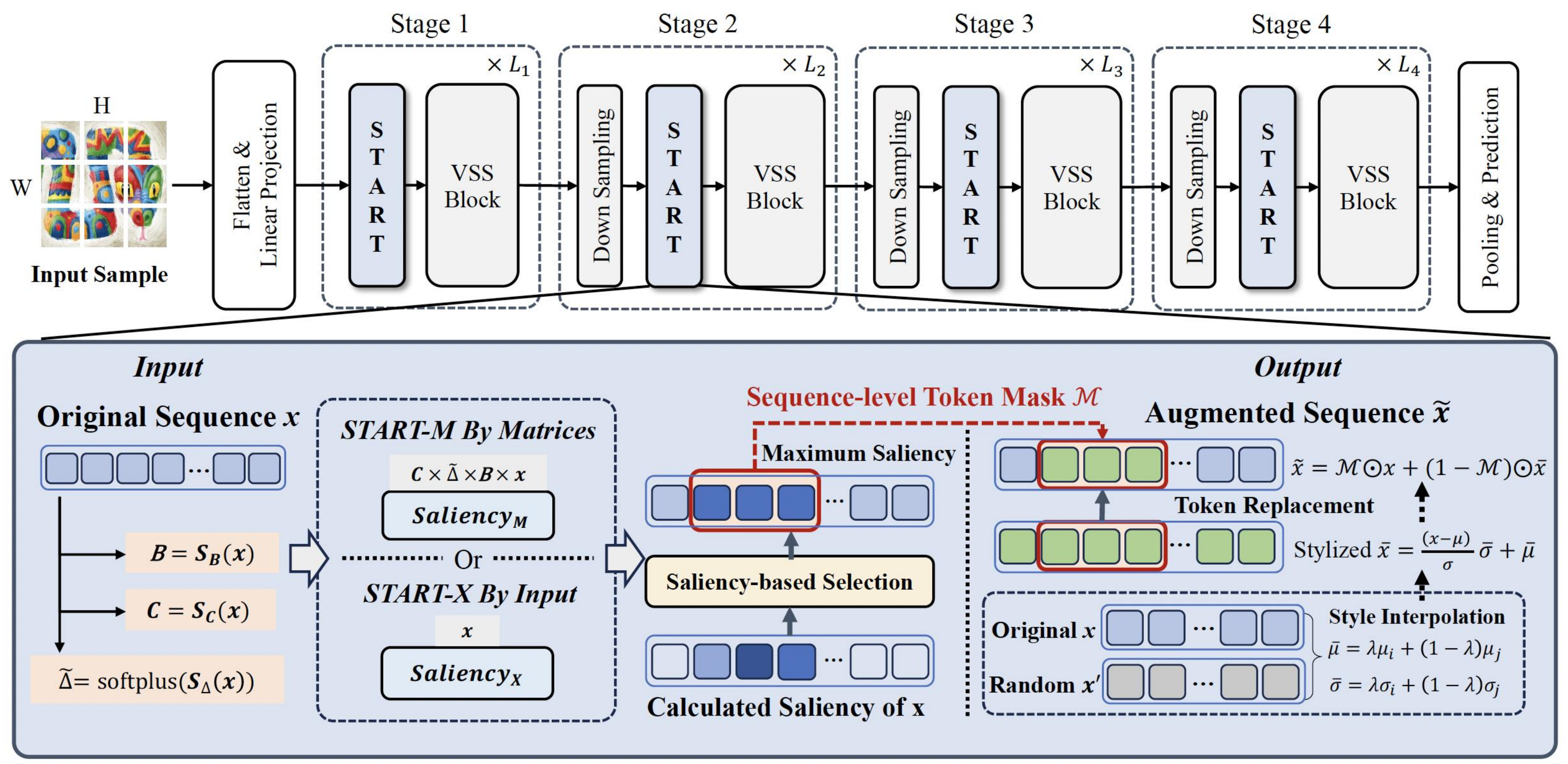
Theorem 1 (Generalization Risk Bound). *With the previous setting and assumptions, let D_S^i and D_T be two sets with M samples independently drawn from $\mathcal{D}_S^n$ and $\mathcal{D}_T$, respectively. For any $\delta \in (0,1)$ with probability of at least $1-\delta$, for all $h \in \mathcal{H}$, the following inequality holds:*

$$R_{D_T}(h) \le \sum_{n=1}^{N} \pi_n R^n_{D_S}(h) + d_{\text{To-MMD}}(D_T, \bar{D}_T) + \sup_{i,j\in[N]} d_{\text{To-MMD}}(D_S^i, D_S^j) + 2\lambda_\pi + \sigma, \qquad (5)$$

where $\lambda_\pi = \frac{1}{M}(\sum_{n=1}^{N} \pi_n \mathbb{E}_{x\sim D_S^n}[\sqrt{tr(K_{D_S^n})}] + \mathbb{E}_{x\sim D_T}[\sqrt{tr(K_{D_T})}]) + \sqrt{\frac{\log(2/\epsilon)}{2M}}$, and σ is the minimum combined error of the ideal hypothesis h^* on both D_S and D_T. Let $\kappa_T = d_{\text{To-MMD}}(D_T, \bar{D}_T)$ and $\kappa_S = \sup_{i,j\in[N]} d_{\text{To-MMD}}(D_S^i, D_S^j)$, respectively.

Proposion 1 (Accumulation of Domain Discrepancy). *Given two distinct domains D_S and D_T, the token-level domain distance $d_{\text{To-MMD}}(D_S, D_T)$ depends on $d_{C\bar{\Delta}Bx}(\bar{x}_i^S, \bar{x}_i^T)$ and $d_{\bar{\Delta}}(\bar{x}_i^S, \bar{x}_i^T)$ for the i-th token. For the entire recurrent process, domain-specific information encoded in S_Δ, S_C, and S_B will accumulate, thereby amplifying domain discrepancy.*

Proposion 2 (Mitigating Domain Discrepancy Accumulation). *Perturbing domain-specific features in tokens focused on by S_Δ, S_C, and S_B can enhance their learning of domain-invariant features, thus effectively mitigating the accumulation issue in these input-dependent matrices.*



(1) A theoretical investigation into generalization ability of Mamba models, revealing that input-dependent matrices in Mamba accumulate domain-specific features during the recurrent process, thus hindering model's generalizability.

(2) A novel SSM-based architecture with saliency-driven token-aware transformation as a competitive alternative to CNNs and ViTs for DG, which performs excellent generalization ability with efficient linear complexity.

(3) For saliency-driven token-aware transformation, we explore two variants to identify and perturb salient tokens in feature sequences, effectively reducing domain-specific information within the input-dependent matrices of Mamba.

Poster ID：93769

NEURAL INFORMATION PROCESSING SYSTEMS

START: A Generalized State Space Model with Saliency-Driven Token-Aware Transformation

Jintao Guo[1] **Lei Qi**[2*] **Yinghuan Shi**[1*] **Yang Gao**[1†]

[1] Nanjing University [2] Southeast University

guojintao@smail.nju.edu.cn, qilei@seu.edu.cn, {syh, gaoy}@nju.edu.cn

南京大学 NANJING UNIVERSITY

東南大學 SOUTHEAST UNIVERSITY

Backgrounds

NEURAL INFORMATION PROCESSING SYSTEMS

Domain Generalization (DG)

- Learn a model from source domains that performs well on arbitrary unseen target domains without re-training.

Problems of existing DG studies

- Existing CNN-based DG works inevitably tend to learn local texture information due to limited receptive fields of local convolutions, leading to overfitting to source domains.

- Recent works have introduced ViTs as the backbone for DG, utilizing global receptive field of self-attention to mitigate local texture bias, but suffer from high complexity that increases quadratically with input length.

Motivation

NEURAL INFORMATION PROCESSING SYSTEMS

State Space Models (SSMs)

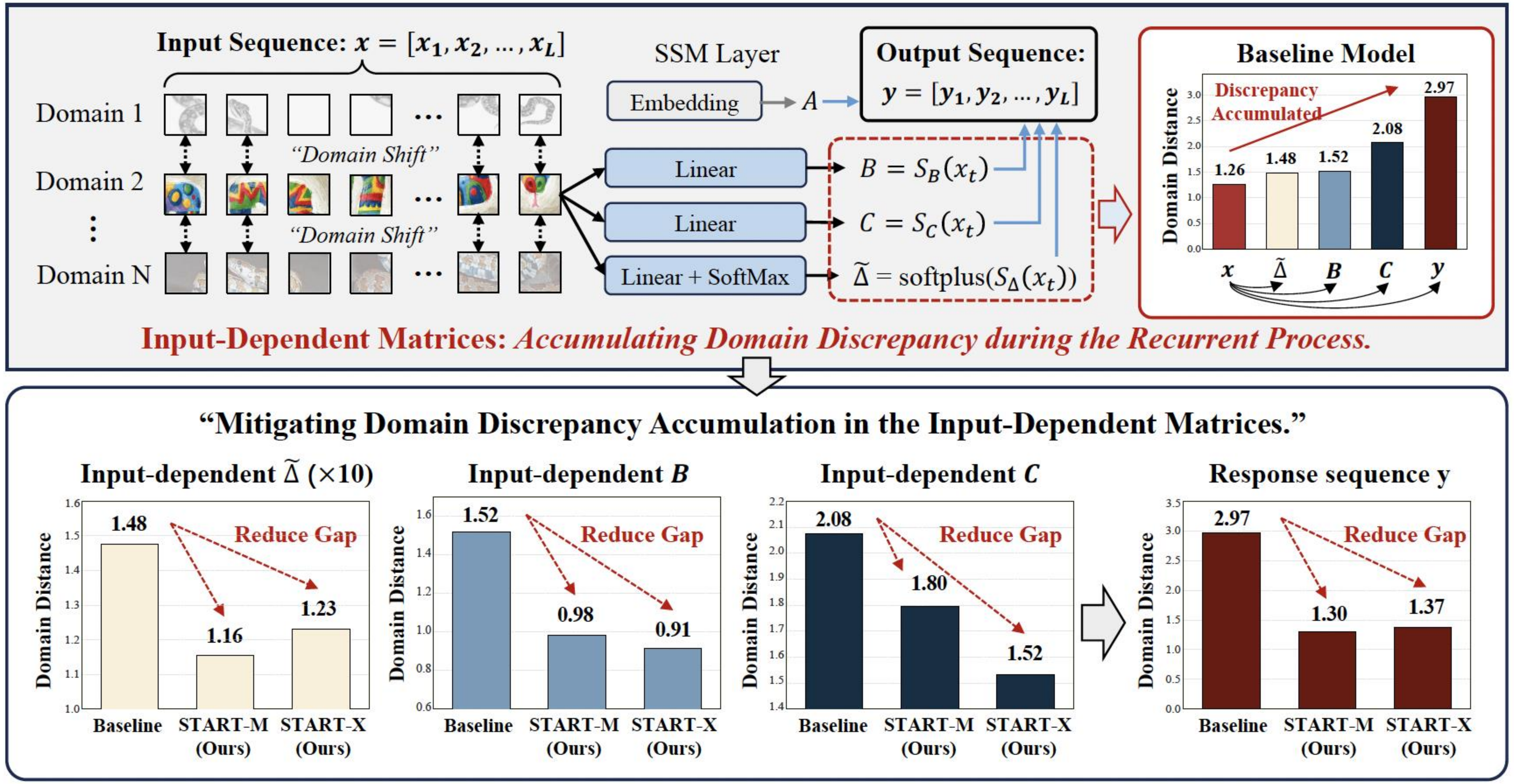
- Represented by Mamba, the advanced state space models (SSMs) have achieved remarkable performance on various supervised learning tasks.
- SSMs rely on input-dependent matrixes to selectively models token dependencies in input sequences in a compressed state space, which achieve linear complexity in sequence length.
- Few existing works have analyzed the generalization ability of Mamba under domain shift.

Whether the Mamba model can achieve excellent performance for DG tasks?

Motivation

Whether the Mamba model can achieve excellent performance for DG tasks?

- Empirical evidence reveals that the key input-dependent matrixes in Mamba could accumulate and amplify domain-specific features during training, which exacerbates overfitting issue of the model to source domains.



Theoretically Analysis

We theoretically explore the generalization error bound of Mamba, proving that perturbing the domain-specific features within the input-dependent matrices of Mamba can effectively diminish the upper bound of the model's generalization risk.

Theorem 1 (Generalization Risk Bound). *With the previous setting and assumptions, let D_S^i and D_T be two sets with M samples independently drawn from $\mathcal{D}_S^n$ and $\mathcal{D}_T$, respectively. For any $\delta \in (0,1)$ with probablity of at least $1-\delta$, for all $h \in \mathcal{H}$, the following inequality holds:*

$$R_{D_T}(h) \leq \sum_{n=1}^{N} \pi_n R_{D_S}^n(h) + d_{\text{To-MMD}}(D_T, \bar{D}_T) + \sup_{i,j\in[N]} d_{\text{To-MMD}}(D_S^i, D_S^j) + 2\lambda_\pi + \sigma, \qquad (5)$$

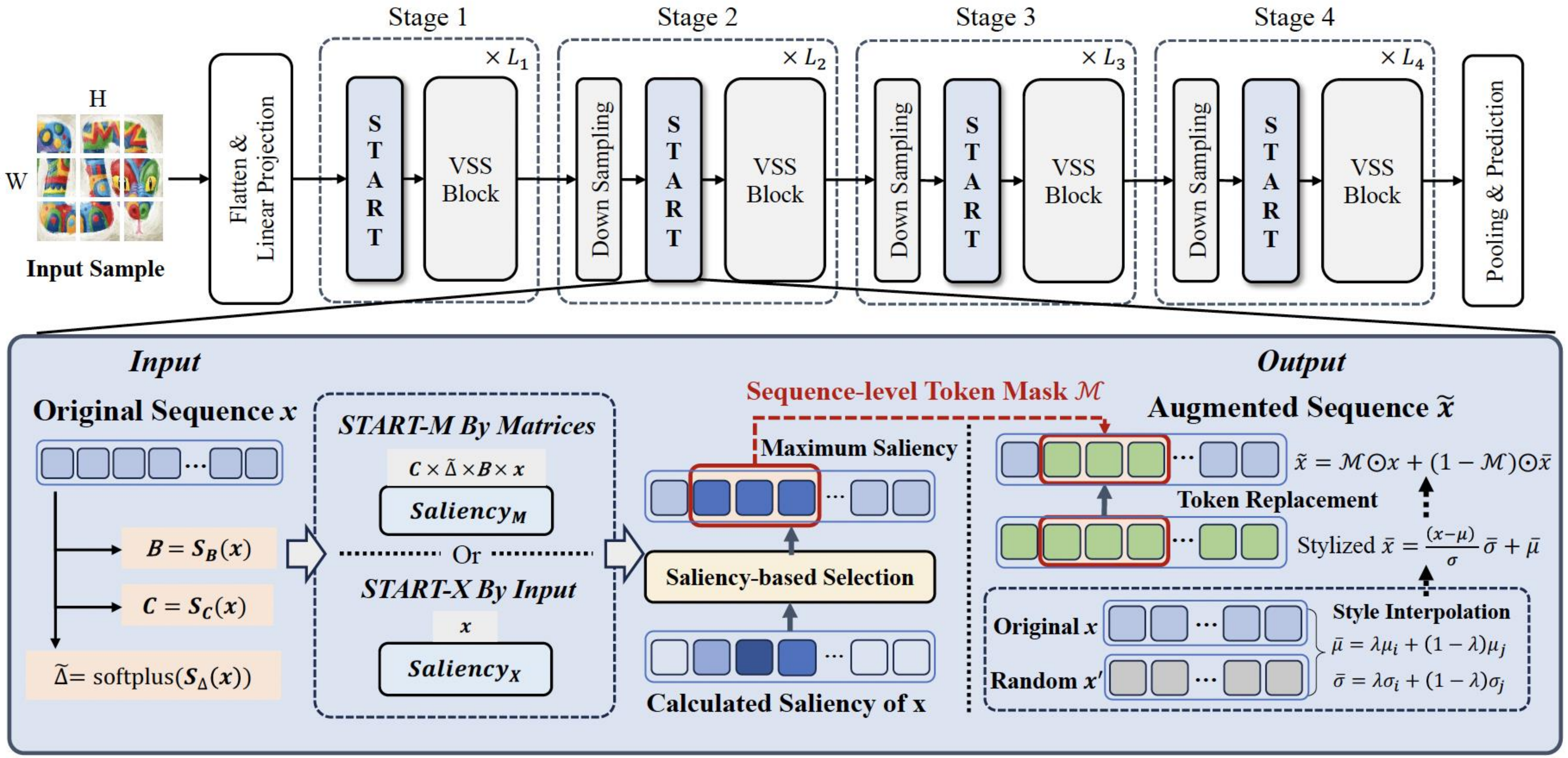
where $\lambda_\pi = \frac{1}{M}(\sum_{n=1}^{N} \pi_n \mathbb{E}_{x\sim D_S^n}[\sqrt{tr(K_{D_S^n})} + \mathbb{E}_{x\sim D_T}[\sqrt{tr(K_{D_T})}]) + \sqrt{\frac{log(2/\epsilon)}{2M}}$, and σ is the minimum combined error of the ideal hypothesis h^* on both D_S and D_T. Let $\kappa_T = d_{\text{To-MMD}}(D_T, \bar{D}_T)$ and $\kappa_S = \sup_{i,j\in[N]} d_{\text{To-MMD}}(D_S^i, D_S^j)$, respectively.

Proposion 1 (Accumulation of Domain Discrepancy). *Given two distinct domains D_S and D_T, the token-level domain distance $d_{\text{To-MMD}}(D_S, D_T)$ depends on $d_{C\tilde{\Delta}Bx}(\bar{x}_i^S, \bar{x}_i^T)$ and $d_{\tilde{\Delta}}(\bar{x}_i^S, \bar{x}_i^T)$ for the i-th token. For the entire recurrent process, domain-specific information encoded in S_Δ, S_C, and S_B will accumulate, thereby amplifying domain discrepancy.*

Proposion 2 (Mitigating Domain Discrepancy Accumulation). *Perturbing domain-specific features in tokens focused on by S_Δ, S_C, and S_B can enhance their learning of domain-invariant features, thus effectively mitigating the accumulation issue in these input-dependent matrices.*

Methodology

Based on the theoretically analysis, we propose a novel Saliency-driven Token-AwaRe Transformation paradigm (START in short), which aims to explicitly suppress domain-related features within the input-dependent matrixes.



Methodology

- START incorporates a saliency-driven token selection scheme to perturb the prominent regions of input-dependent matrices.
- We propose two variants to identify and perturb tokens within salient regions, including START-M that determines saliency using input-dependent matrices, and START-X computing saliency based on input sequences.

START-M: *based on input-dependent matrices*

$$Saliency_M(x_i) = S_C(x_i)\mathrm{softmax}(S_\Delta(x_i))S_B(x_i)x_i$$

START-X: *based on input sequences*

$$Saliency_X(x_i) = x_i$$

Diversify Style Information:

$$\tilde{\mu} = \epsilon\mu(x) + (1-\epsilon)\mu(x'), \quad \tilde{\sigma} = \epsilon\sigma(x) + (1-\epsilon)\sigma(x'),$$

$$\epsilon \sim Beta(0.1, 0.1), \quad \tilde{x} = \frac{x - \mu(x)}{\sigma(x)} \cdot \tilde{\mu} + \tilde{\sigma},$$

Augment Saliency Tokens:

$$x_{\text{aug}} = \mathcal{M}_S \odot x + (1 - \mathcal{M}_S) \odot \tilde{x},$$

Experiments

START achieves SOTA performances on various DG datasets.

Method	Params.	PACS Art	PACS Cartoon	PACS Photo	PACS Sketch	PACS Avg.	Office-Home Art	Office-Home Clipart	Office-Home Product	Office-Home Real	Office-Home Avg.
CNN: ResNet-50											
DeepAll [65] (AAAI'20)	23M	84.70	80.80	97.20	79.30	85.50	61.30	52.40	75.80	76.60	66.50
PCL [66] (CVPR'22)	23M	90.20	83.90	98.10	82.60	88.70	67.30	59.90	78.70	80.70	71.60
EoA [67] (NeurIPS'22)	23M	90.50	83.40	98.00	82.50	88.60	69.10	59.80	79.50	81.50	72.50
EQRM [68] (NeurIPS'22)	23M	86.50	82.10	96.60	80.80	86.50	60.50	56.00	76.10	77.40	67.50
SAGM [69] (CVPR'23)	23M	87.40	80.20	98.00	80.80	86.60	65.40	57.00	78.00	80.00	70.10
iDAG [70] (ICCV'23)	23M	90.80	83.70	98.00	82.70	88.80	68.20	57.90	79.70	81.40	71.80
DomainDrop [60] (ICCV'23)	23M	89.82	84.22	98.02	85.98	89.51	67.33	60.39	79.05	80.22	71.75
CCFP [71] (ICCV'23)	23M	87.50	81.30	96.40	81.40	86.60	63.70	55.50	77.20	79.20	68.90
MADG [72] (NeurIPS'23)	23M	87.80	82.20	97.70	78.30	86.50	67.60	54.10	78.40	80.30	70.10
PGrad [73] (ICLR'23)	23M	87.60	79.10	97.40	76.30	85.10	64.70	56.00	77.40	78.90	69.30
AGFA [74] (ICLR'23)	23M	89.80	85.20	97.60	84.70	89.30	67.50	58.50	79.30	80.70	71.50
GMDG [75] (CVPR'24)	23M	84.70	81.70	97.50	80.50	85.60	68.90	56.20	79.90	82.00	70.70
ViT-based or MLP-like models											
MLP-B [76] (NeurIPS'21)	59M	85.00	77.86	94.43	65.72	80.75	63.45	56.31	77.81	79.76	69.33
SDViT [18] (ACCV'22)	22M	87.60	82.40	98.00	77.20	86.30	68.30	56.30	79.50	81.80	71.50
ResMLP-S [77] (TPAMI'22)	40M	85.50	78.63	97.07	72.64	83.46	62.42	51.94	75.40	77.21	66.74
ViP-S [78] (TPAMI'22)	25M	88.09	84.22	98.38	82.41	88.27	69.55	61.51	79.34	83.11	73.38
GMoE-S [19] (ICLR'23)	34M	89.40	83.90	99.10	74.50	86.70	69.30	58.00	79.80	82.60	72.40
SSM-based models											
DGMamba [54] (ACM MM'24)	22M	91.30	87.00	99.00	87.30	91.20	**76.20**	61.80	83.90	86.10	77.00
Strong Baseline [22]	22M	91.55	85.11	99.14	83.97	89.94±0.52	75.06	60.48	84.71	85.45	76.43±0.15
START-M (Ours)	22M	**93.29**	**87.56**	99.14	87.07	**91.77**±0.40	75.15	62.04	**85.31**	**85.84**	**77.09**±0.16
START-X (Ours)	22M	92.76	87.43	**99.22**	**87.46**	91.72±0.49	75.48	**62.06**	85.24	85.47	77.07±0.07

Experiments

NEURAL INFORMATION PROCESSING SYSTEMS

Ablation studies of each components on multiple datasets.

Method	OfficeHome					TerraIncognita				
	Art	Clipart	Product	Real	Avg.	L100	L38	L43	L46	Avg.
Baseline [22]	75.06	60.48	84.71	85.45	76.43±0.15	66.39	47.27	62.42	48.56	56.16±0.41
w/o. Saliency Guided	75.12	61.06	84.91	85.42	76.63±0.17	69.49	49.10	62.70	47.92	57.30±0.07
w/o. Token Selection	75.11	61.77	84.97	85.26	76.78±0.07	68.97	49.19	62.87	48.74	57.44±0.22
START-M (Ours)	75.15	62.04	**85.31**	**85.84**	**77.09**±0.16	70.13	**49.98**	63.02	**49.49**	58.16±0.79
START-X (Ours)	**75.48**	**62.06**	85.24	85.47	77.07±0.07	**70.70**	49.47	**63.96**	48.95	**58.27**±0.75

START outperforms previous augmentation methods.

Method	Art	Cartoon	Photo	Sketch	Avg.
Baseline [22]	91.55	85.11	99.14	83.97	89.94±0.52
MixStyle [13]	92.05	86.55	98.90	86.35	90.94±0.18
DSU [14]	92.58	85.91	98.98	85.39	90.71±0.22
ALOFT [15]	93.07	86.04	99.16	85.31	90.89±0.24
START-M (Ours)	**93.29**	**87.56**	99.14	87.07	**91.77**±0.40
START-X (Ours)	92.76	87.43	**99.22**	**87.46**	91.72±0.49

START can effectively reduce domain gaps in input-dependent matrices.

Method	$\tilde{\Delta}$ (↓)	**B** (↓)	**C** (↓)	**Feat.** (↓)
Baseline [22]	1.48	1.52	2.08	2.97
MixStyle [13]	1.73	1.36	1.90	1.91
DSU [14]	1.38	1.28	2.18	1.59
ALOFT [15]	1.37	1.25	2.33	1.67
START-M (Ours)	**1.16**	0.98	1.80	**1.30**
START-X (Ours)	1.23	**0.91**	**1.52**	1.37

Experiments

NEURAL INFORMATION PROCESSING SYSTEMS

START introduces no additional inference time, has significantly fewer FLOPs but higher performance than CNNs.

Method	Backbone	Params (M)	GFlops (G)	Time (ms)	Avg. (%)
DeepAll [65] (AAAI'20)	ResNet-50	23	8.26	-	85.50
iDAG [70] (ICCV'23)	ResNet-50	23	8.00	94	88.80
iDAG [70] (ICCV'23)	ResNet-101	41	15.00	495	89.20
GMoE-S [19] (ICLR'23)	DeiT-S	34	5.00	136	88.10
GMoE-B [19] (ICLR'23)	DeiT-B	133	19.00	361	89.20
ViP [78] (TPAMI'22)	ViP-S	25	13.84	-	88.27
GFNet [88] (TPAMI'23)	GFNet-H-Ti	13	4.10	-	87.76
DGMamba [54] (ACM MM'24)	VMamba-T	31	5.00	233	91.20
Strong Baseline [22]	VMamba-T	22	5.68	252	89.94
START-M (Ours)	VMamba-T	22	5.68	252	91.77
START-X (Ours)	VMamba-T	22	5.68	252	91.72

Conclusion

NEURAL INFORMATION PROCESSING SYSTEMS

- In this paper, we conduct a theoretical investigation into the generalization ability of the Mamba model, revealing that the input-dependent matrices in Mamba can accumulate domain-specific features during the recurrent process, thus hindering the model's generalizability.

- Based on theoretical analysis, we propose a novel SSM-based architecture with saliency-driven token-aware transformation as a competitive alternative to CNNs and ViTs for DG, which performs excellent generalization ability with efficient linear complexity.

- For saliency-driven token-aware transformation, we explore two variants to identify and perturb salient tokens in feature sequences, effectively reducing domain-specific information within the input-dependent matrices of Mamba.

Paper

Code