Learning a Cross-Modal Schrödinger Bridge for Visual Domain Generalization

Hao Zheng, Jingjun Yi, Qi Bi, Huimin Huang, Haolan Zhan,

Yawen Huang, Yuexiang Li, Xian Wu, Yefeng Zheng

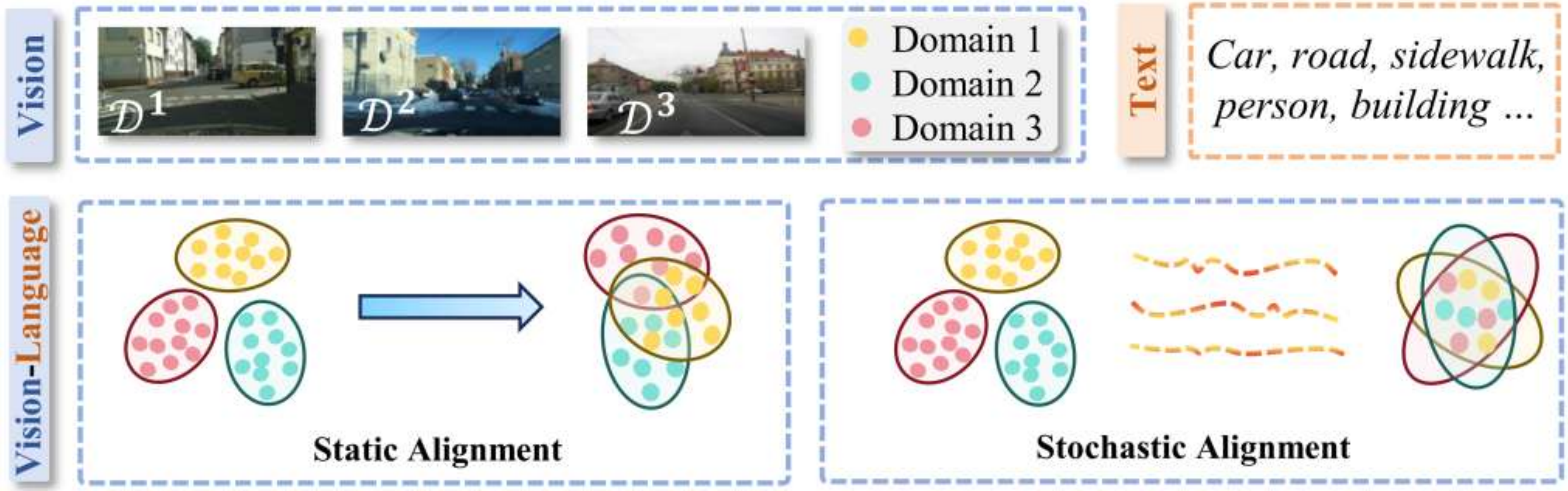
1. Westlake University, China 2. Jarvis Research Center, Tencent, China

Problem Statement

- Domain generalization aims to train models that perform robustly on unseen target domains without access to target data.
- The realm of vision-language foundation model has opened a new venue owing to its inherent out-of-distribution generalization capability.
- However, the static alignment to class-level textual anchors remains insufficient to handle the dramatic distribution discrepancy from diverse domain-specific visual features.

Motivation

- We propose a novel cross-domain Schrödinger Bridge (SB) method, namely SBGen, to handle this challenge.
- It explicitly formulates the stochastic semantic evolution, to gain better generalization to unseen domains.



Background

- What's Schrödinger Bridge (SB)?

Problem Definition. Let $\mathcal{X}$ and $\mathcal{Y}$ denote the space of input images and the space of structured labels from a certain task (*e.g.*, classification). Given a set of labeled source domains $\mathcal{D}^S = \{(x_n^S, y_n^S)\}_{n=1}^{N_S}$ with $x_n^S \in \mathcal{X}, y_n^S \in \mathcal{Y}$, and a set of unseen target domains $\mathcal{D}^U = \{(x_m^U, y_m^U)\}_{m=1}^{N_U}$, the objective is to train a model on source domains that generalizes to these unseen domains.

Definition 1. Optimal Transport (OT). *Let P^S and P^U be two probability distributions over $\mathbb{R}^C$, respectively. The classical OT problem seeks a deterministic transport map $M : \mathbb{R}^C \to \mathbb{R}^C$ minimizing a transport cost:*

$$\min_{M: M_\# P^S = P^U} \mathbb{E}_{z^S \sim P^S} \left[\| \boldsymbol{z}^S - M(\boldsymbol{z}^S) \|^2 \right], \tag{1}$$

where $M_\# P^S$ denotes the pushforward measure of P^S through M.

Definition 2. Entropy-Regularized OT. *A stochastic coupling $\pi(\boldsymbol{z}^S, \boldsymbol{z}^U)$ with marginal constraints $\pi \in \Pi(P^S, P^U)$ is introduced to improve the robustness of OT, minimizing:*

$$\min_{\pi \in \Pi(P^S, P^U)} \int \| \boldsymbol{z}^S - \boldsymbol{z}^U \|^2 \, d\pi(\boldsymbol{z}^S, \boldsymbol{z}^U) + \varepsilon \cdot \mathrm{KL}(\pi \| \mathcal{R}), \tag{2}$$

where $\mathcal{R}$ is a reference measure and $\varepsilon > 0$ controls the regularization strength, enabling stochastic transport but lacks a notion of dynamics over time.

Definition 3. Schrödinger Bridge (SB). *OT is extended to the dynamic setting by introducing a continuous-time stochastic process $\{P_t\}_{t \in [0,1]}$ that evolves from P^S to P^U, while being minimally deviated from a prior diffusion process $\mathbb{P}$ (e.g., Brownian motion). The SB formulation is:*

$$\min_{\mathbb{Q}} \mathrm{KL}(\mathbb{Q} \| \mathbb{P}) \quad \text{subject to} \quad \mathbb{Q}_{t=0} = P^S, \quad \mathbb{Q}_{t=1} = P^U, \tag{3}$$

where $\mathbb{Q}$ denotes the law of the interpolating process over latent features. This yields a family of time-indexed distributions P_t modeling the optimal evolution of visual features across domains.

Method Design

Technically, the proposed SBGen consists of three key components:

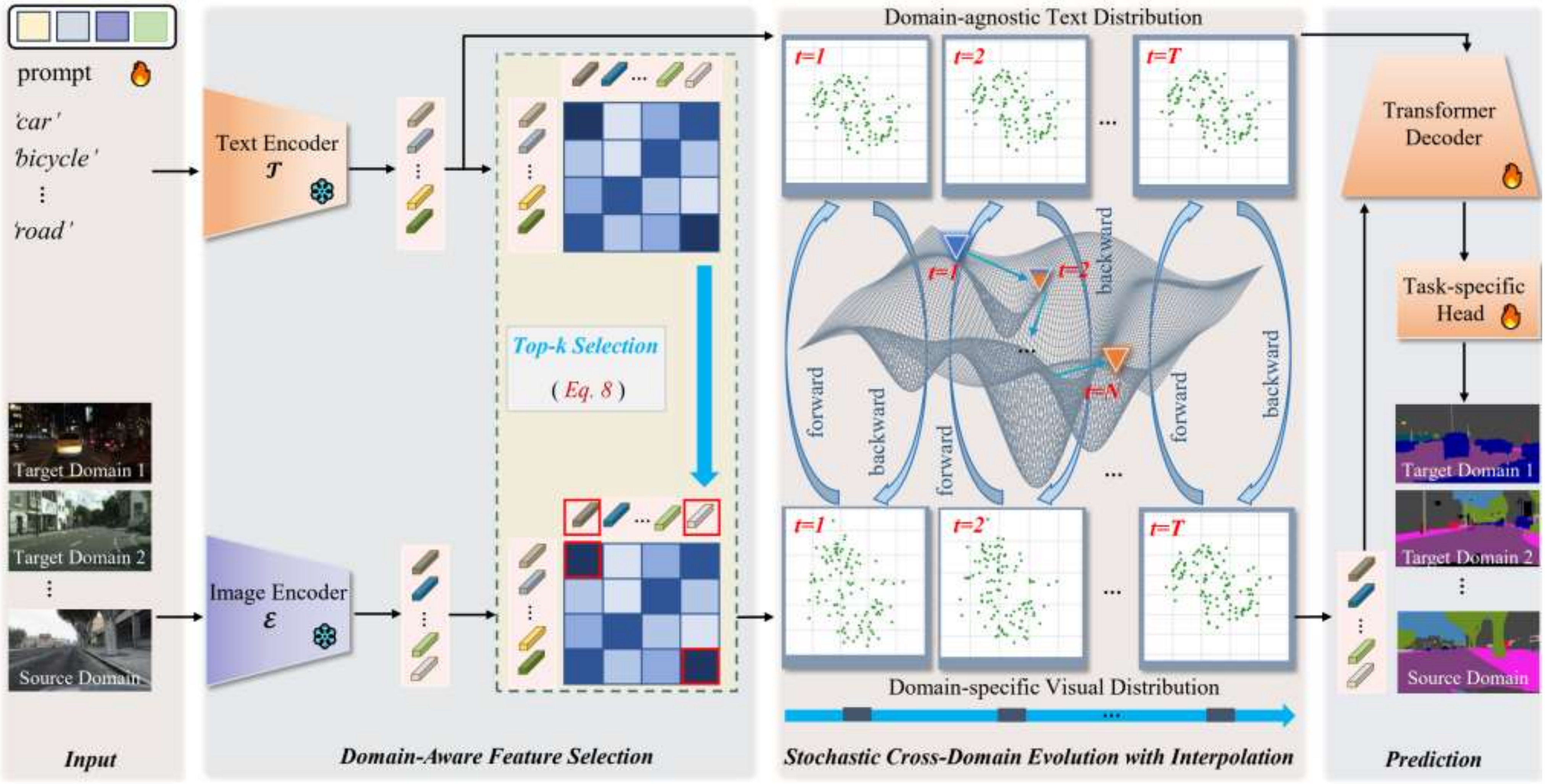
(1) text-guided domain-aware feature selection to isolate semantically aligned image tokens;

(2) stochastic cross-domain evolution to simulate the SB dynamics via a learnable time-conditioned drift;

(3) stochastic domain-agnostic interpolation to construct semantically grounded feature trajectories.

Framework Overview

Cross-modal Schrödinger Bridge for visual domain Generalization (SBGen)



Experiment

- Experiment 1: Domain Generalization in Classification

Table 1: Comparison with the state-of-the-art methods on PACS, VLCS, OfficeHome, DomainNet and TerraInc. By default the results are cited from [15, 65, 16, 36, 79]. Evaluation metric is classification accuracy (in %). Top three results are highlighted as **best**, second and third, respectively.

Method	Venue	PACS	VLCS	OfficeHome	DomainNet	TerraInc	Avg
ResNet-50 Pre-trained by ImageNet:							
DANN [27]	IJCAI'2016	83.6	78.6	65.9	38.3	46.4	65.6
Fish [63]	ICML'2022	85.5	77.8	68.6	42.7	45.1	63.9
DAC-SC [37]	CVPR'23	87.5	78.7	70.3	44.9	46.5	65.6
SAGM [73]	CVPR'2023	86.6	80.0	70.1	45.0	48.8	66.1
ViT-B/16 Pre-trained by CLIP:							
SWAD [12]	NIPS'2021	91.3	79.4	76.9	51.7	45.4	68.9
CLIP [59]	ICML'2021	96.2	81.7	82.0	57.5	33.4	70.2
SMA [3]	NIPS'2022	92.1	79.7	78.1	55.9	48.3	70.8
DUPRG [50]	ICLR'2023	97.1	83.9	83.6	59.6	42.0	73.2
CoOp [86]	IJCV'2022	96.2	77.6	83.9	59.8	48.8	73.3
MIRO [13]	ECCV'2022	95.6	82.2	82.5	54.0	54.3	73.7
SEDGE [43]	ArXiv'2022	96.1	82.2	80.7	54.7	56.8	74.1
DPL [82]	TAI'2023	97.3	84.3	84.2	56.7	52.6	75.0
CLIPOOD [65]	ICML'2023	97.3	85.0	87.0	63.5	60.4	78.6
Promptstyler [16]	ICCV'2023	97.2	82.9	83.6	59.4	-	-
KAdaptaion [36]	WACV'2025	**97.5**	83.0	**90.3**	62.7	51.9	77.1
GESTUR [39]	ICCV'2023	96.0	82.8	84.2	58.9	55.7	75.5
DPR [15]	CVPR'2024	**97.5**	86.4	86.1	62.1	57.1	77.8
CLIPCEIL++ [79]	NeurIPS'2024	97.2	85.2	87.7	63.6	62.0	79.1
Ours	2025	97.4	**86.7**	89.9	**64.4**	**63.5**	**80.4**

Experiment

- Experiment 2: Domain Generalization in Segmentation

Table 2: Performance comparison between the proposed method and existing DGSS methods. C: CityScapes [18]; B: BDD-100K [78]; M: Mapillary [47]; S: SYNTHIA [62]; G: GTA5 [61]. '-': results were not reported and official source code is not available; '*': only reported one decimal official results; '†': re-implementation with official source code under default settings. Evaluation metric is mIoU in %. Top three results are highlighted as **best**, second and third, respectively.

Method	Venue	Encoder	G → C	G → B	G → M	Avg.	C → B	C → M	Avg.
ImageNet Pretrained:									
ISW [17]	CVPR'2021	ResNet-101	36.58	35.20	40.33	-	50.73	58.64	-
GTR [56]	TIP'2021	ResNet-101	37.53	33.75	34.52	-	50.75	57.16	-
SHADE [83]	ECCV'2022	ResNet-101	44.65	39.28	43.34	-	50.95	60.67	-
SAW [57]	CVPR'2022	ResNet-101	39.75	37.34	41.86	-	52.95	59.81	-
WildNet [38]	CVPR'2022	ResNet-101	44.62	38.42	46.09	-	50.94	58.79	-
AdvStyle [85]	NeurIPS'2022	ResNet-101	39.62	35.54	37.00	-	-	-	-
SPC [30]	CVPR'2023	ResNet-101	44.10	40.46	45.51	-	-	-	-
BlindNet [2]	CVPR'2024	ResNet-101	45.72	41.32	47.08	-	51.84	60.18	-
HGFormer* [21]	CVPR'2023	Swin-T	-	-	-	-	53.4	66.9	-
CMFormer [9]	AAAI'2024	Swin-B	55.31	49.91	60.09	-	59.27	71.10	-
VLM Pretrained:									
DIDEX* [49]	WACV'2024	Stable Diffusion	62.0	54.3	63.0	59.7	-	-	
VLTSeg* [32]	ACCV'2024	CLIP-L	55.6	52.7	59.6	56.0	-	-	-
REIN* [74]	CVPR'2024	EVA02-L	65.3	60.5	64.9	63.6	64.1	69.5	66.8
SET* [77]	MM'2024	EVA02-L	66.4	61.8	65.6	64.6	-	-	-
FADA* [8]	NeurIPS'2024	EVA02-L	66.7	61.9	66.1	64.9	-	-	-
tqdm [52]	ECCV'2024	EVA02-L	68.88	59.18	70.10	66.05	64.72	76.15	70.44
MGRNet [67]	AAAI'2025	EVA02-L	69.53	61.14	69.97	66.88	64.70	76.43	70.56
Ours		EVA02-L	**71.24**	**62.26**	**71.91**	**68.74**	**66.03**	**77.90**	**71.97**
			↑1.71	↑1.12	↑1.94	↑1.59	↑1.33	↑1.47	↑1.41

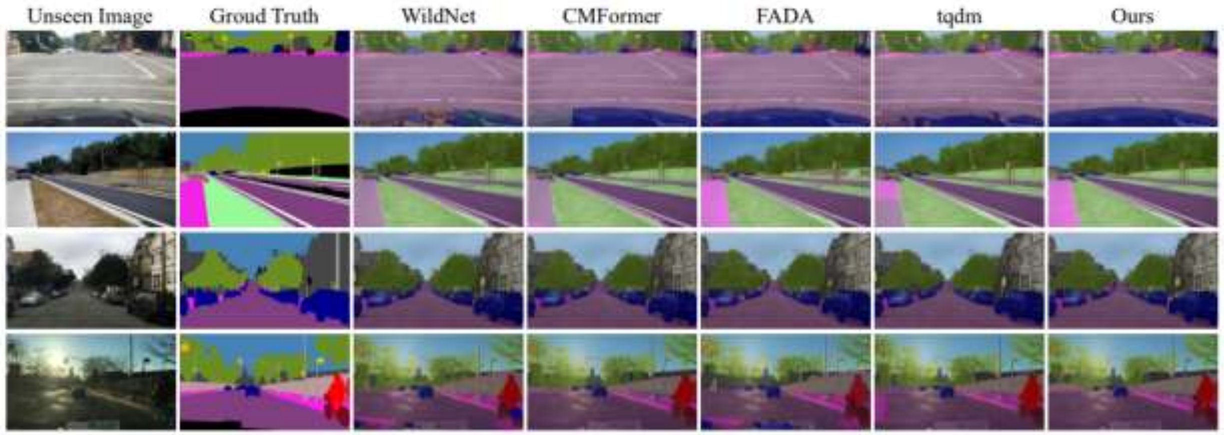


Figure 3: Exemplar segmentation results of existing DGSS methods (WildNet [38], CMFormer [9], FADA [8], tqdm [52]), and the proposed SBGen on unseen target domains.

Thanks for your attention!