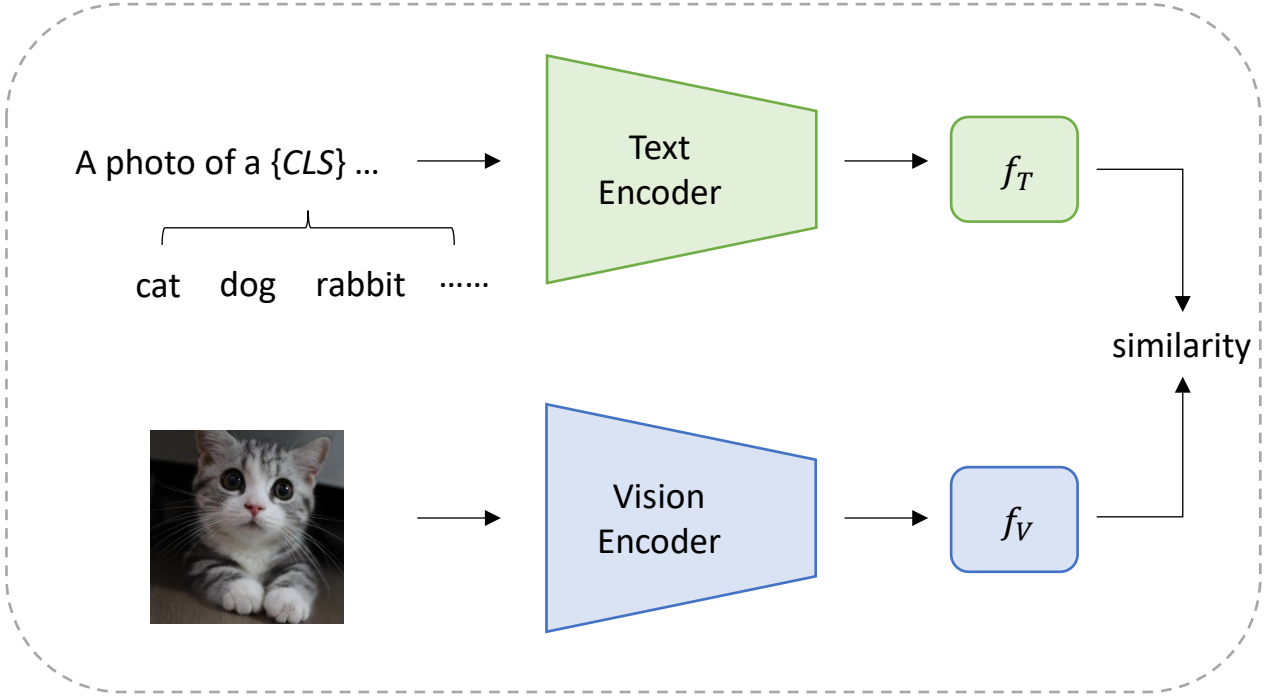
TransAgent: Transfer Vision-Language Foundation Models with Heterogeneous Agent Collaboration

Yiwei Guo[1,2], Shaobin Zhuang[3,4], Kunchang Li[1,2,3], Yu Qiao[3], Yali Wang[1,3]

[1]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Shanghai AI Laboratory
[4]Shanghai Jiao Tong University

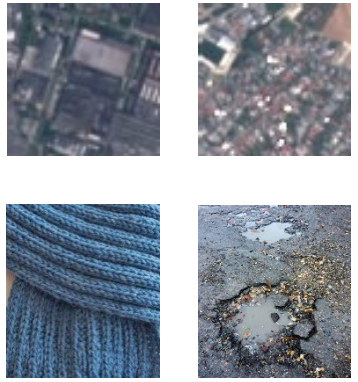
Background

Vision-Language foundation models (*e.g.*, CLIP[1]) have recently shown their power in **transfer learning** owing to large-scale image-text pre-training.



These models are typically pre-trained on **general domain data**, and struggle to generalize well on **diversified target domain data** (e.g., fine-grained) due to large **domain shift**.

General domain data



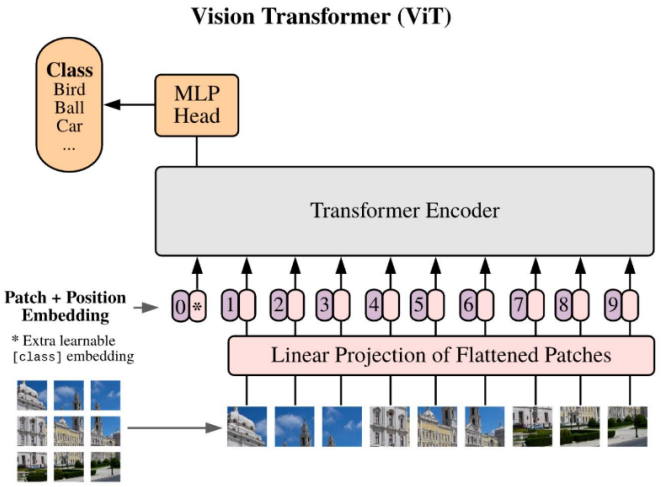
Fine-grained domain data

[1] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. (ICML2021)

Motivation

Existing methods typically adapt the representation of foundation models to downstream tasks. However, it is challenging to achieve good generalization by adopting such a single model, especially under low-shot regime.

Alternatively, expert models from different modalities, tasks, networks and datasets contain **complementary knowledge** with CLIP-like models, and can be utilized to further boost the generalization ability of foundation models.



Vision Model

Large Language Model

Text-to-Image Generative Model

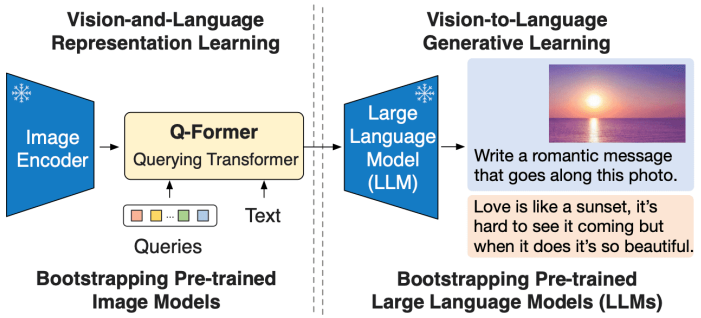
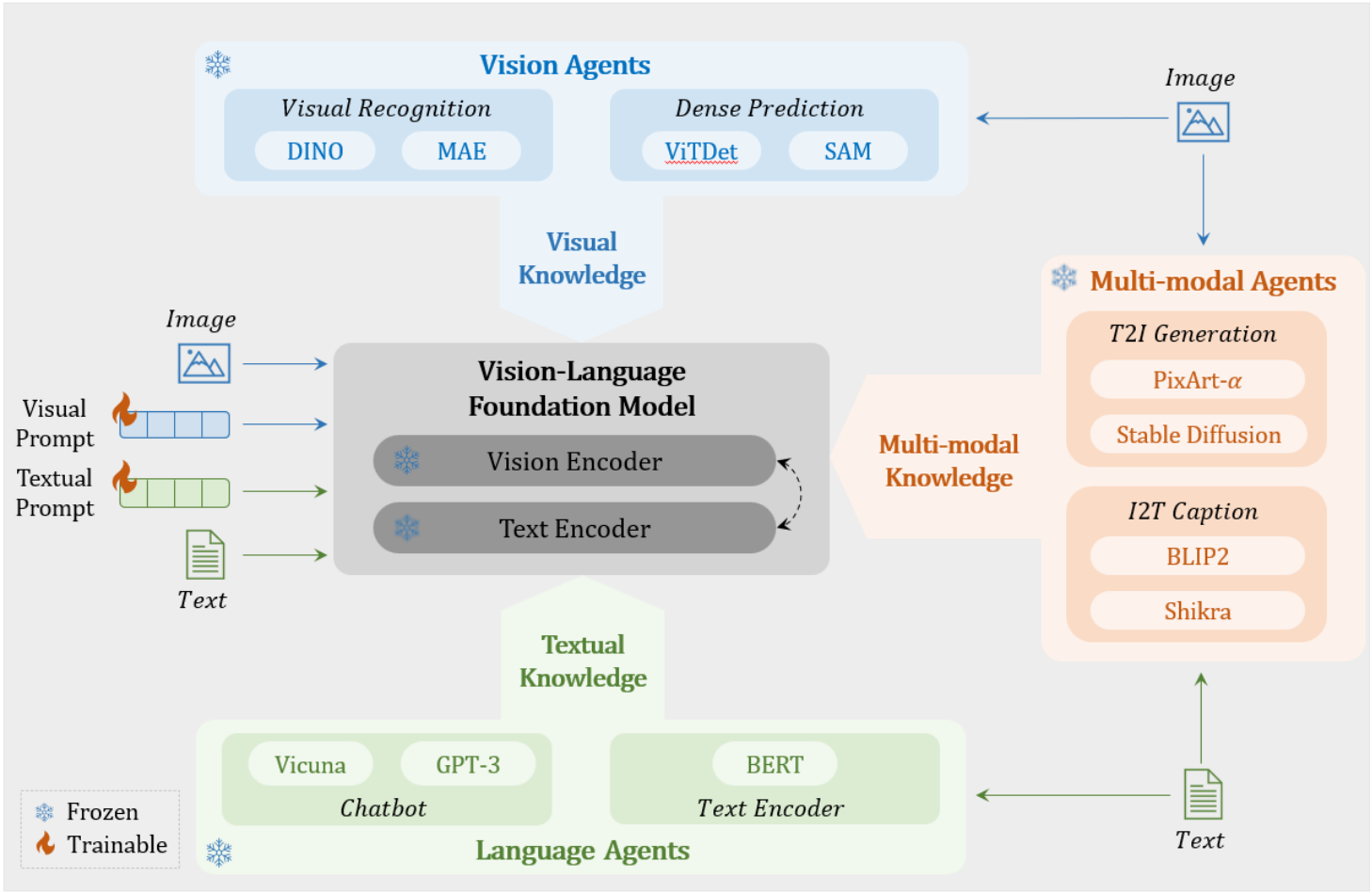


Image-to-Text Captioning Model

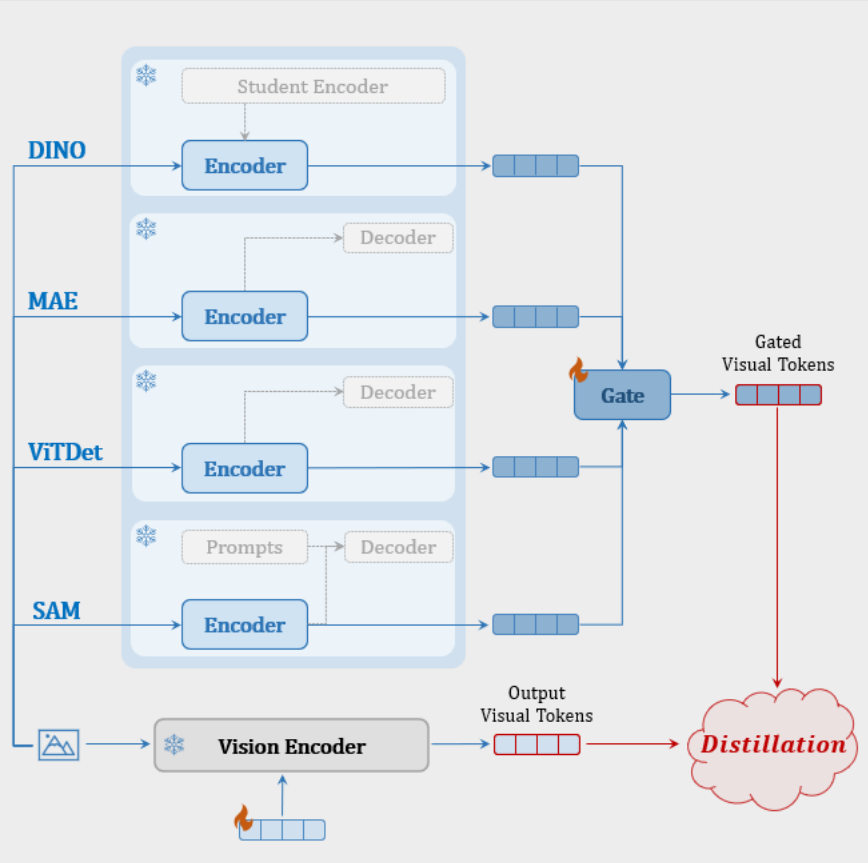
Contributions

- Knowledge Versatility
  - TransAgent leverages **11** heterogeneous agents from vision, language and multi-modal research.
  - The diversified knowledge are **complementary** with CLIP-like models.
- Transfer Flexibility
  - TransAgent leverages the proposed **Mixture-of-agents (MoA)** gating mechanism to **adaptively integrate** external knowledge of different agents in each modality.
- Deployment Efficiency
  - **Multi-source distillation** is applied to transfer knowledge of heterogeneous agents into CLIP.
  - Along with **prompt learning**, TransAgent achieves deployment efficiency without a heavy model ensemble.

Method: Overview



Method: Vision Agent Collaboration (VAC)

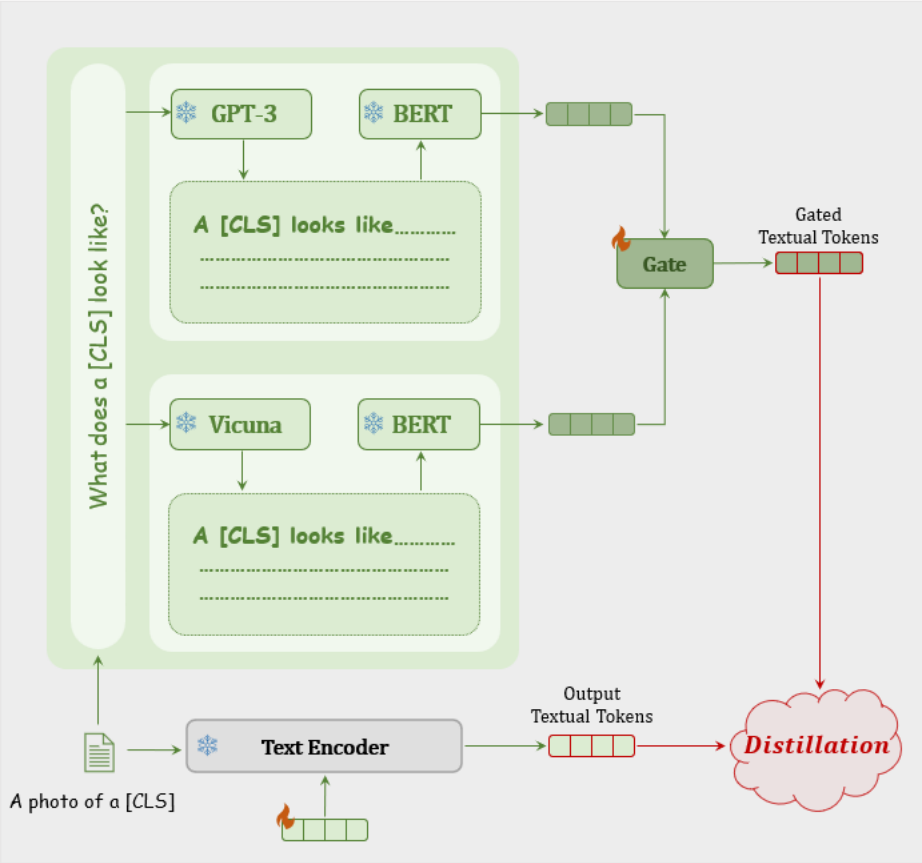


VAC integrates visual knowledge via MoA gating and transfers the knowledge through layer-wise feature distillation:

$$\mathbf{W}_V = \mathrm{MLP}(\mathrm{Concat}(\{\mathbf{V}_A(i)\})), \quad \mathbf{V}_A = \sum_i \mathbf{W}_V(i)\mathbf{V}_A(i)$$

$$\mathcal{L}_{\mathrm{VAC}} = |\mathbf{V} - \mathbf{V}_A|$$

Method: Language Agent Collaboration (LAC)

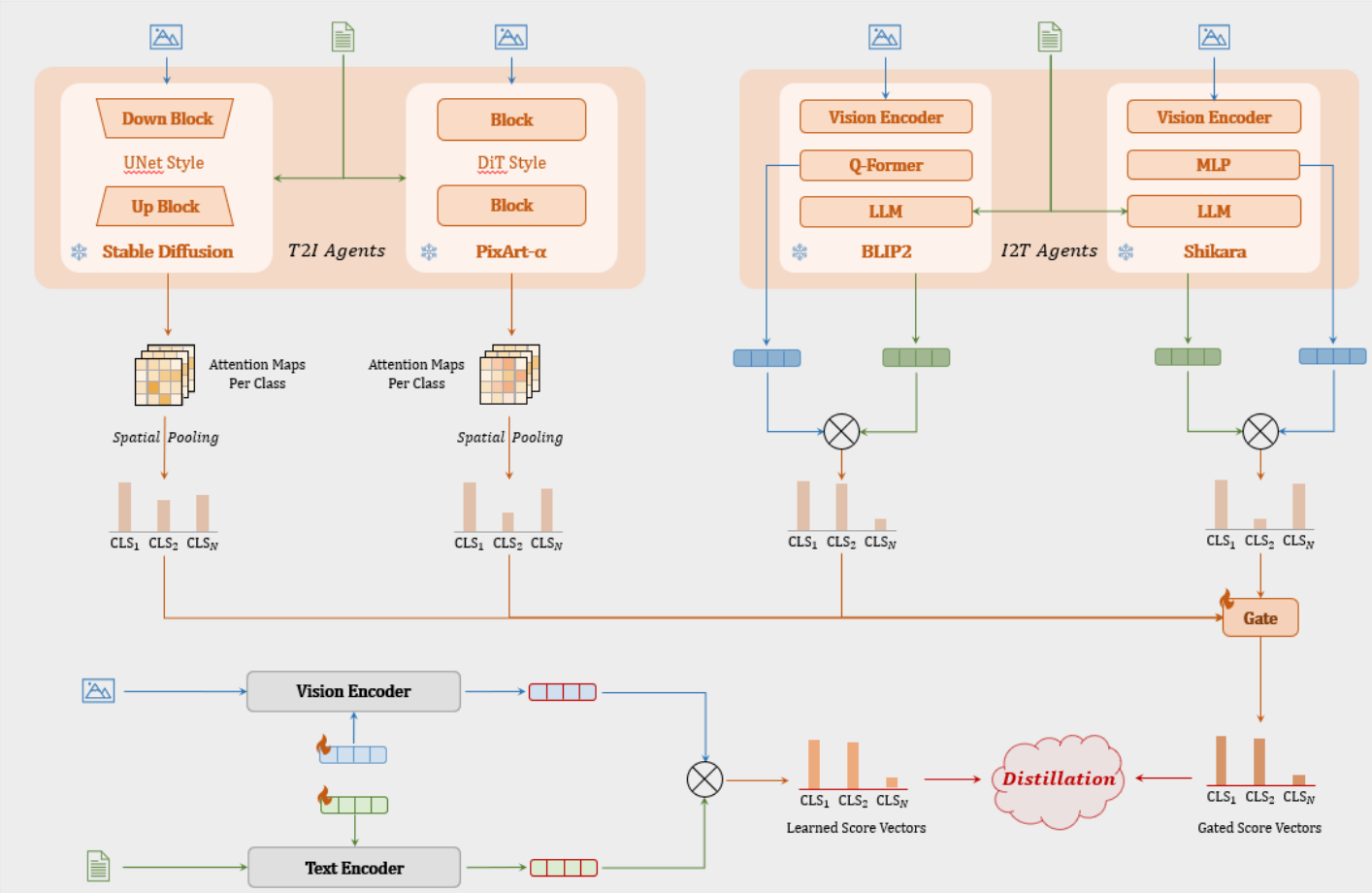


LAC enhances the textual representations through class-specific feature distillation between the prompted textual feature and the gated textual feature:

$$\mathbf{W}_T = \mathrm{MLP}(\mathrm{Concat}(\{\mathbf{T}_A(j)\})), \quad \mathbf{T}_A = \sum_j \mathbf{W}_T(j)\mathbf{T}_A(j)$$

$$\mathcal{L}_{\mathrm{LAC}} = |\mathbf{T} - \mathbf{T}_A|$$

Method: Multi-modal Agent Collaboration (MAC)



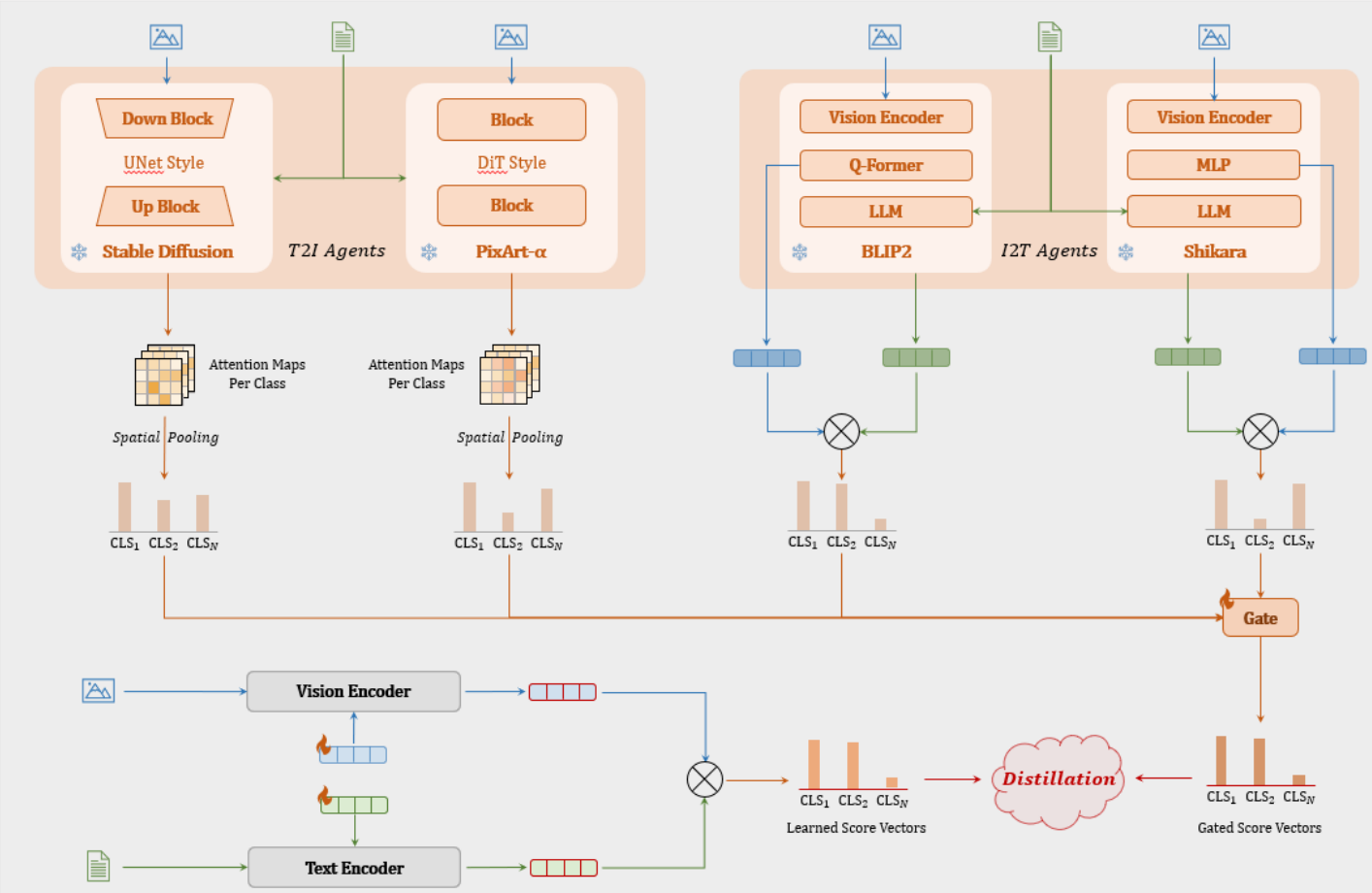
We extract the cross attention maps from the *T2I agents* and then obtain the score vectors through LogSumExp (LSE) pooling:

$$\mathbf{S}_{T2I} = \log(\sum_k \exp(\mathbf{M}_k))$$

We compute the score vectors from the *I2T agents* as the cosine similarity between the projected visual feature and the LLM's textual feature:

$$\mathbf{S}_{I2T} = \frac{\exp(sim(proj(f_V), f_T))}{\sum_{i=1}^{C} \exp(sim(proj(f_V), f_T^i))}$$

Method: Multi-modal Agent Collaboration (MAC)



Finally, we perform score distillation between the learned score vectors and the gated score vectors to further align the learnable prompts:

$$\mathbf{M}_A = \text{Concat}(\{\mathbf{S}_{T2I}, \mathbf{S}_{I2T}\})$$

$$\mathbf{W}_S = \text{MLP}(\mathbf{M}_A), \quad \mathbf{S}_A = \sum_n \mathbf{W}_S(n)\mathbf{M}_A(n)$$

$$\mathcal{L}_{\text{MAC}} = \text{KL}(\text{softmax}(\mathbf{S}_P)||\text{softmax}(\mathbf{S}_A))$$

Method: Multi-source Knowledge Distillation

Finally, we combine all the distillation loss from multiple sources, achieving heterogeneous agent collaboration for knowledge transfer:

$$\mathcal{L}_{\text{TransAgent}} = \mathcal{L}_{\text{CE}} + \lambda_1 \mathcal{L}_{\text{MAC}} + \lambda_2 \mathcal{L}_{\text{MAC}} + \lambda_3 \mathcal{L}_{\text{MAC}}$$

where $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters.

By fine-tune the learnable prompts with distillation strategy, all the agents can be unloaded and the modality-specific gates can be abandoned in the inference phase, which largely boosts deployment efficiency.

Experiments

- Comparisons on ***base-to-novel generalization***

  Our TransAgent exhibits strong generalization ability and outperforms previous SOTA on all datasets. The best results are **bolded**.

Method	Average			ImageNet			Caltech101			OxfordPets		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP	69.34	74.22	71.70	72.43	68.14	70.22	96.84	94.00	95.40	91.17	97.26	94.12
CoOp	82.69	63.22	71.66	76.47	67.88	71.92	98.00	89.81	93.73	93.67	95.29	94.47
CoCoOp	80.47	71.69	75.83	75.98	70.43	73.10	97.96	93.81	95.84	95.20	97.69	96.43
MaPLe	82.28	75.14	78.55	75.40	70.32	72.72	98.27	93.23	95.68	95.43	97.83	96.62
RPO	81.13	75.00	77.78	76.60	**71.57**	74.00	97.97	94.37	96.03	94.63	97.50	96.05
PromptSRC	84.26	76.10	79.97	77.60	70.73	74.01	98.10	94.03	96.02	95.33	97.30	96.30
TransAgent	**85.29**	**77.62**	**81.27**	**78.07**	70.57	**74.13**	**98.90**	**95.23**	**97.03**	**96.33**	**98.13**	**97.22**

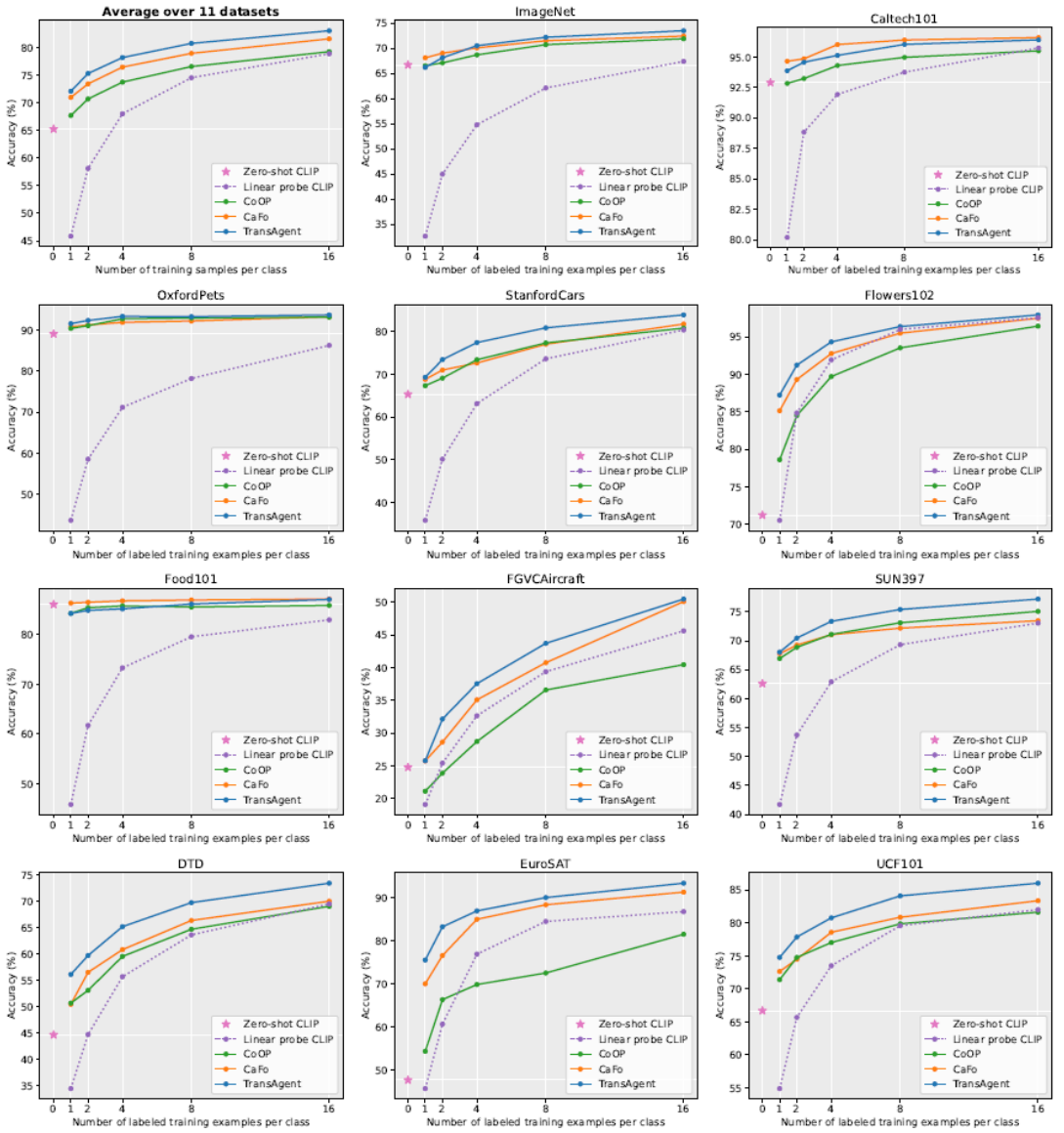
Method	StanfordCars			Flowers102			Food101			FGVCAircraft		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP	63.37	74.89	68.65	72.08	77.80	74.83	90.10	91.22	90.66	27.19	36.29	31.09
CoOp	78.12	60.40	68.13	97.60	59.67	74.06	88.33	82.26	85.19	40.44	22.30	28.75
CoCoOp	70.49	73.59	72.01	94.87	71.75	81.71	90.70	91.29	90.99	33.41	23.71	27.74
MaPLe	74.70	71.20	72.91	97.70	68.68	80.66	90.30	88.57	89.43	36.90	34.13	35.46
RPO	73.87	**75.53**	74.69	94.13	76.67	84.50	90.33	90.83	90.58	37.33	34.20	35.70
PromptSRC	78.27	74.97	76.58	98.07	76.50	85.95	90.67	91.53	91.10	42.73	37.87	40.15
TransAgent	**79.53**	74.73	**77.06**	**98.37**	**77.13**	**86.46**	**90.87**	**92.20**	**91.53**	**43.77**	**39.00**	**41.25**

Method	SUN397			DTD			EuroSAT			UCF101		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP	69.36	75.35	72.23	53.24	59.90	56.37	56.48	64.05	60.03	70.53	77.50	73.85
CoOp	80.60	65.89	72.51	79.44	41.18	54.24	92.19	54.74	68.69	84.69	56.05	67.46
CoCoOp	79.74	76.86	78.27	77.01	56.00	64.85	87.49	60.04	71.21	82.33	73.45	77.64
MaPLe	78.47	76.93	77.79	80.67	56.48	66.44	83.90	66.00	73.88	85.23	71.97	78.04
RPO	80.60	77.80	79.18	76.70	62.13	68.61	86.63	68.97	76.79	83.67	75.43	79.34
PromptSRC	82.67	78.47	80.52	83.37	62.97	71.75	92.90	73.90	82.32	87.10	78.80	82.74
TransAgent	**82.90**	**79.30**	**81.06**	**84.37**	**63.67**	**72.57**	**97.43**	**83.43**	**89.89**	**87.60**	**80.47**	**83.88**

Experiments

- Comparisons on ***few-shot classification***

TransAgent demonstrates SOTA performance for all few-shot settings on different datasets, which proves promising learning capability even under extremely limited supervision.



Experiments

- Comparisons on ***cross-dataset evaluation*** and ***domain generalization***

TransAgent does not overfit on the source dataset and leads to an overall improvement over the previous methods.

TransAgent improves the robustness of VLMs against out-of-distribution data.

Method	Source	Target										
	ImageNet	Caltech 101	Oxford Pets	Standford Cars	Flowers 102	Food101	FGVC Aircraft	SUN397	DTD	Euro SAT	UCF101	**Avg.**
CLIP	66.72	92.94	89.07	65.29	71.30	86.11	**24.87**	62.62	44.56	47.69	66.77	65.12
CoOp	71.51	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
CoCoOp	71.02	**94.43**	90.14	65.32	71.88	86.06	22.94	**67.36**	45.73	45.37	68.21	65.74
MaPLe	70.72	93.53	**90.49**	65.57	**72.23**	86.20	24.74	67.01	46.49	48.06	68.69	66.30
PromptSRC	71.27	93.60	90.25	**65.70**	70.25	86.15	23.90	67.10	**46.87**	45.50	68.75	65.81
TransAgent	**72.00**	94.37	90.33	65.43	71.40	**86.47**	23.20	66.20	45.30	**52.13**	**69.93**	**66.48**

cross-dataset evaluation

Method	Source	Target				
	ImageNet	-V2	-S	-A	-R	Avg.
CLIP	66.73	60.83	46.15	47.77	73.96	57.18
CoOp	71.51	64.20	47.99	49.71	75.21	59.28
CoCoOp	71.02	64.07	48.75	50.63	76.18	59.91
MaPLe	70.72	64.07	49.15	50.90	76.98	60.27
PromptSRC	71.27	64.35	49.55	50.90	**77.80**	60.65
TransAgent	**72.00**	**64.87**	**49.63**	**51.23**	77.53	**60.82**

domain generalization

Conclusion

- We propose a unified framework to transfer vision-language foundation models through heterogeneous agent collaboration.
  - Achieving knowledge versatility by leveraging diversified knowledge from external experts
  - Achieving transfer flexibility by adaptively integrating the external knowledge via MoA gating mechanism
  - Achieving deployment efficiency by multi-source distillation along with prompt learning
- TransAgent achieves state-of-the-art performance on 11 datasets under the low-shot scenarios.

Thank You