TPR: Topology-Preserving Reservoirs for Generalized Zero-Shot Learning

Hui Chen[1], Yanbin Liu[2], Yongqiang Ma[1], Nanning Zheng[1], Xin Yu[3]

[1]Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University

[2]Auckland University of Technology

[3]The University of Queensland

Zero-Shot Learning - ZSL

Problem Statement

seen classes

Attributes

- orange beak, white head…
- black beak, red face…
- …

Classnames…

Textual descriptions…

ZSL model

probability distribution

unseen classes

Attributes

- orange beak, white head…
- black beak, red face…
- …

Classnames…

Textual descriptions…

learned model

select the most similar one

Generalized Zero-Shot Learning - GZSL

Institute of Artificial Intelligence and Robotics

Problem Statement

seen classes

Attributes

- orange beak, white head…
- black beak, red face…
- …

Classnames…

Textual descriptions…

ZSL model

probability distribution

seen & unseen classes

Attributes

- orange beak, white head…
- black beak, red face…
- …

Classnames…

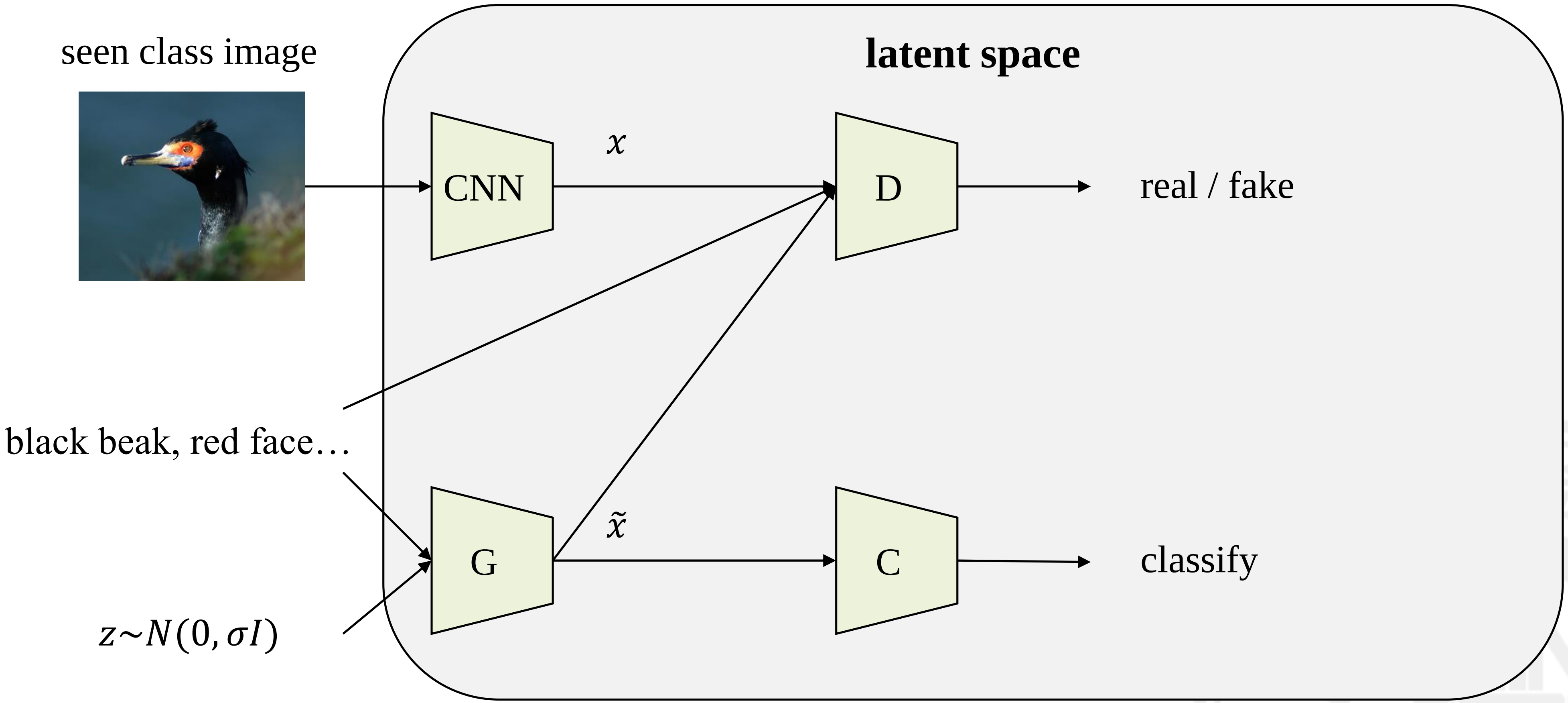
Textual descriptions…

learned model

select the most similar one

Institute of Artificial Intelligence and Robotics

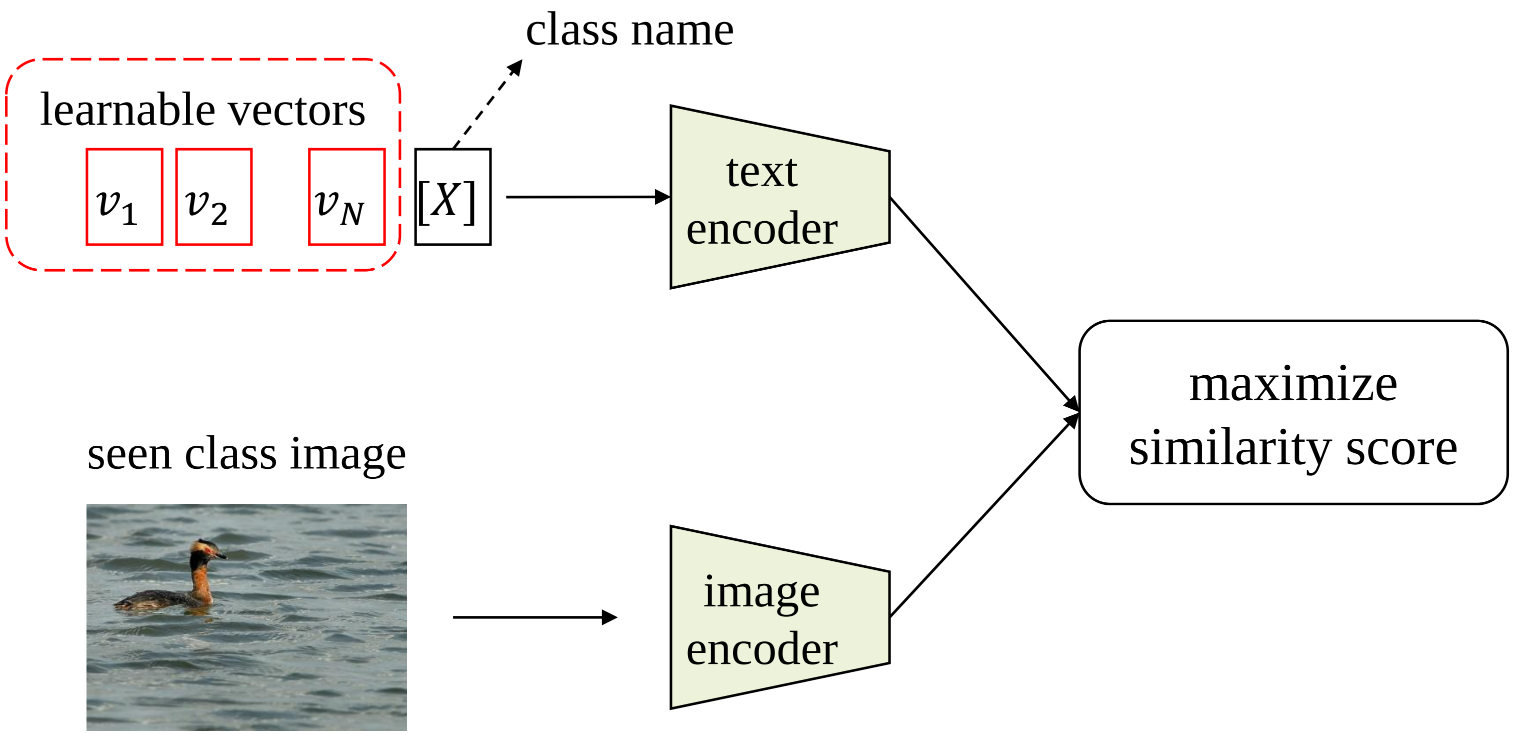
Existing Methods - Generative



Xian Y, et al. Feature Generating Networks for Zero-Shot Learning. CVPR’18

GZSL

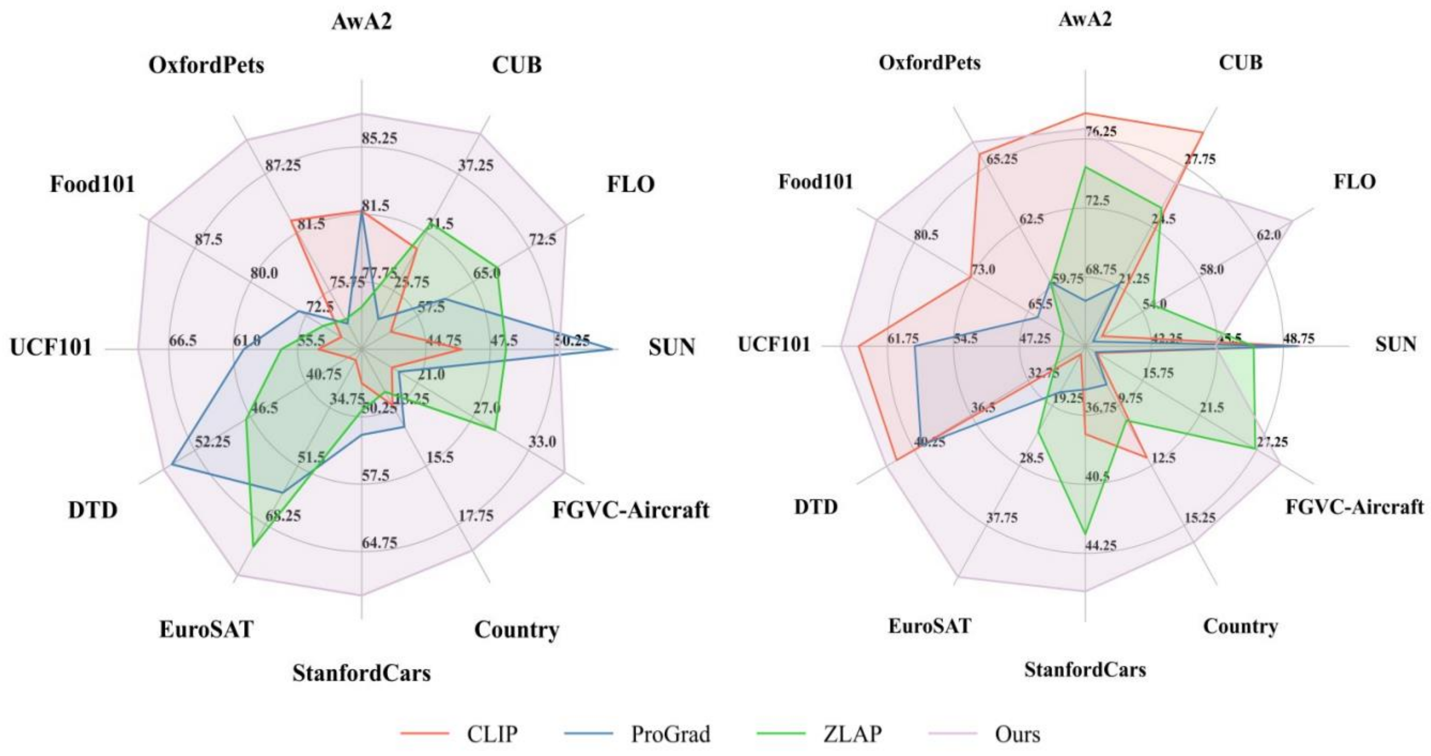
Existing Methods – Prompt Learning



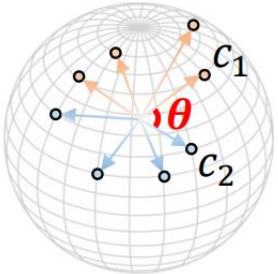
GZSL

Challenge

- A single latent space fails to capture complex and fine-grained patterns for GZSL (a).
- Finetuning CLIP leads to the weak generalization / domain bias problem on unseen classes (b-d).

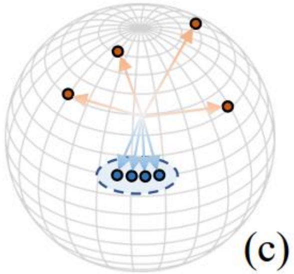


(a) Performance comparison on seen (left) and unseen (right) classes

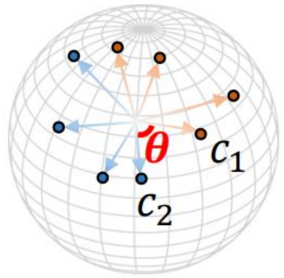


seen features
unseen features

(b) CLIP features



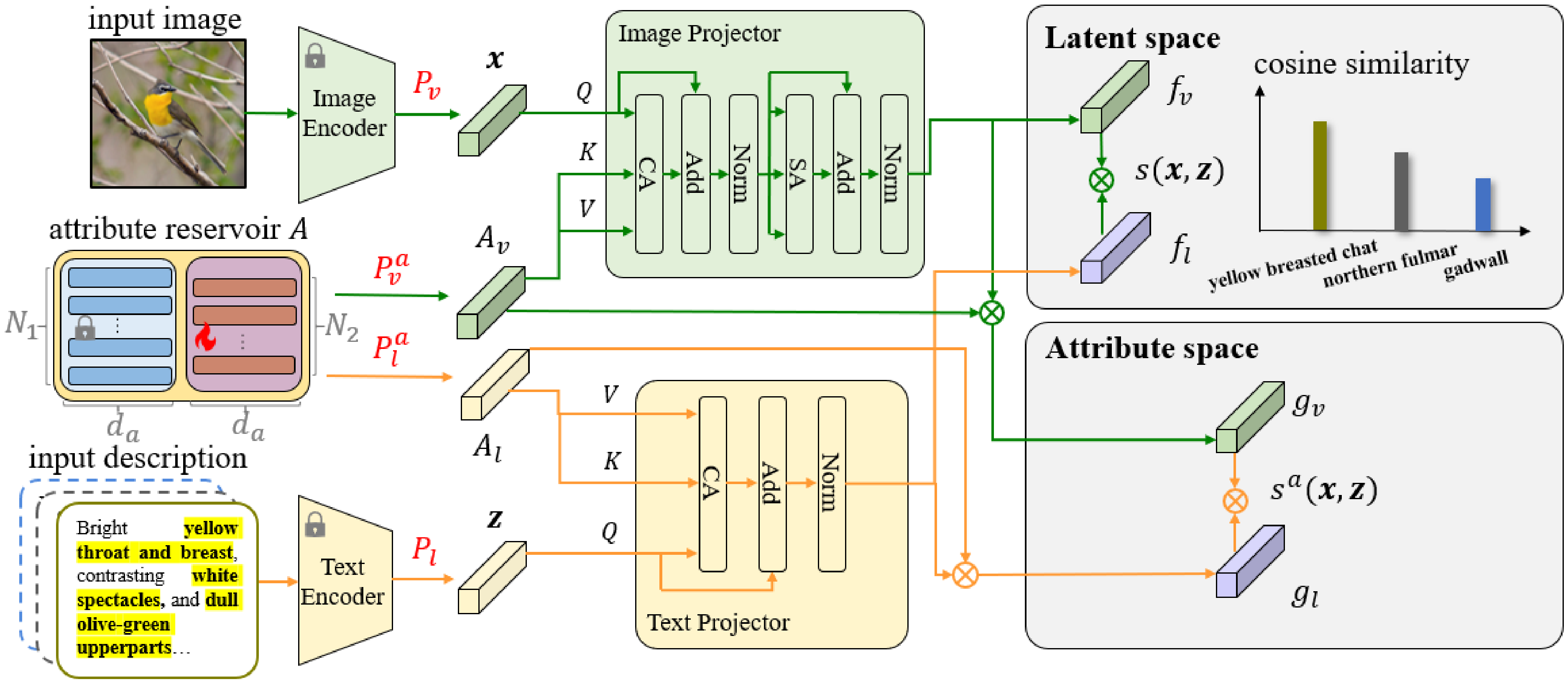
(c) Finetuned features



(d) Our features

Dual-Space Alignment

- Enhance the single latent space with a representative attribute space, which is constructed from a well-devised attribute reservoir. Each dimension of the space corresponds to an attribute concept.
- The reservoir is designed to contain both static and learnable vocabulary tokens.

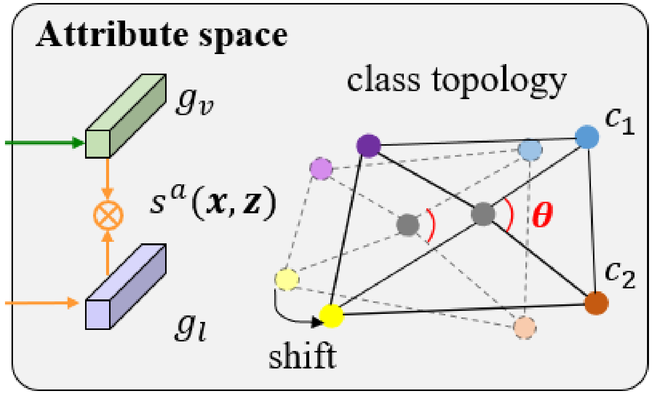


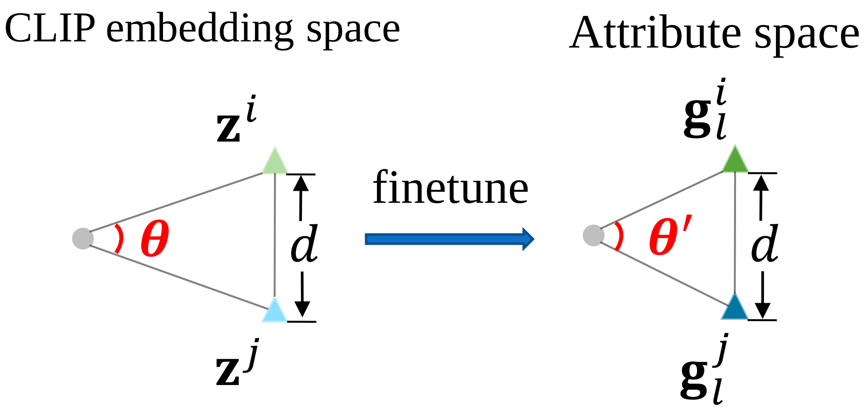
P_v, P_l, P_v^a, P_l^a: learnable projection matrices   CA: cross attention   SA: self-attention $\otimes$: dot product

Topology-Preserving Objective

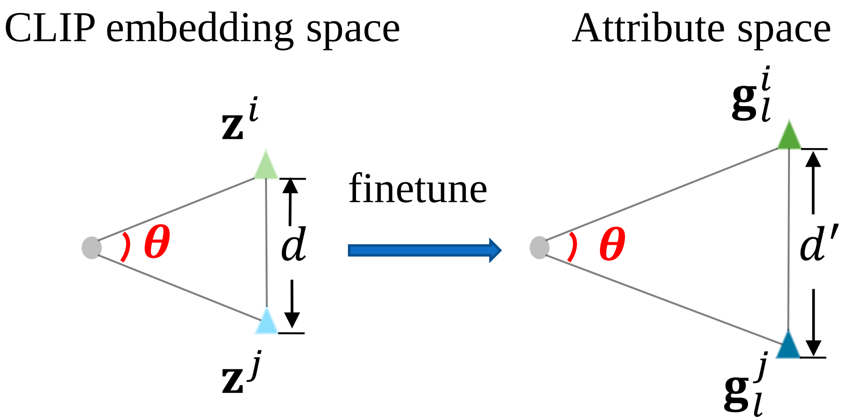
- Maintain the semantic topology structure of the combined seen and unseen classes by referring to the original VLMs embeddings.

a shift of class topology
after finetuning





keep absolute distance



keep angle
more flexible

TPR

Baselines

- Generative methods: CE, LSA, ZLAP
- Prompt learning methods: CoOp, CoCoOp, MaPLe, PromptSRC, ProGrad
- CLIP

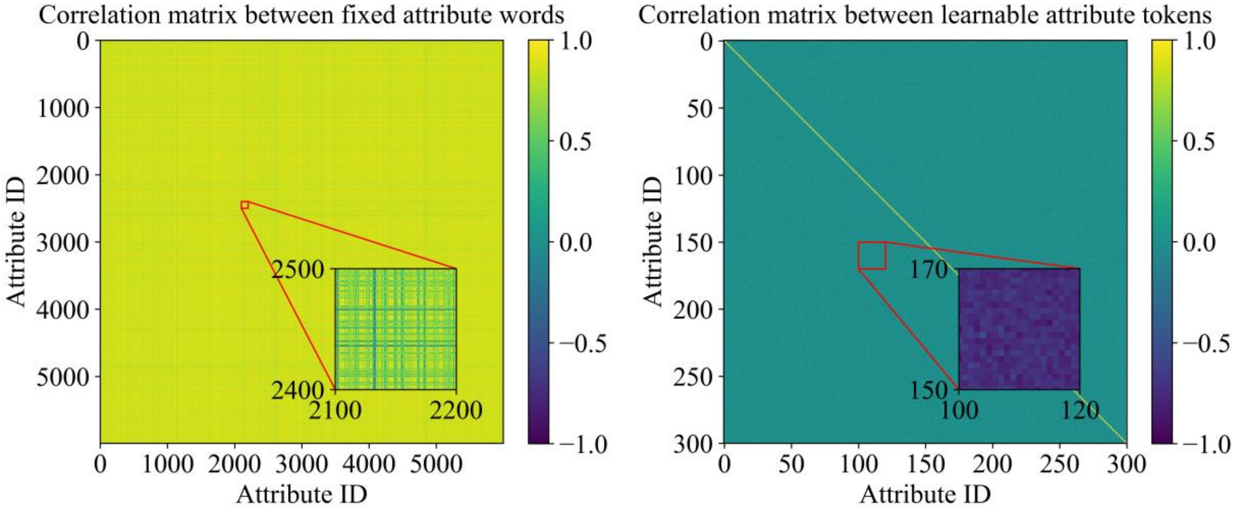
Model	AwA2			CUB*			FLO*			SUN			FGVC-Aircraft*			Country		
	S	*U*	*H*	*S*	*U*	*H*	*S*	*U*	*H*	*S*	*U*	*H*	*S*	*U*	*H*	*S*	*U*	*H*
CLIP [9]	81.69	**77.66**	79.62	29.88	**29.61**	29.74	53.91	51.16	52.50	46.28	**49.51**	47.84	18.25	11.15	13.84	13.16	12.13	12.62
CoOp [40]	81.36	69.42	74.92	22.23	18.23	20.03	56.27	50.65	53.31	49.85	49.31	49.57	17.13	12.10	14.18	12.86	9.73	11.08
CoCoOp [8]	78.53	73.81	76.10	23.53	19.81	21.51	60.21	50.22	54.76	49.53	**49.51**	49.52	18.81	13.60	15.79	13.59	8.03	10.09
MaPLe [10]	78.04	71.25	74.49	22.46	20.66	21.52	59.88	48.39	53.52	46.82	48.68	47.73	21.75	15.20	17.89	12.96	9.54	10.99
PromptSRC [11]	84.04	70.73	76.82	30.92	16.32	21.37	60.68	54.45	57.40	47.83	49.24	48.52	23.44	13.10	16.81	14.42	6.87	9.30
ProGrad [12]	81.73	67.46	73.91	22.97	21.38	22.15	61.21	50.53	55.36	**52.71**	49.44	**51.03**	19.00	11.00	13.93	13.99	8.77	10.78
CE [5]	76.69	67.80	71.97	31.80	19.01	23.80	63.02	44.09	51.88	44.11	47.15	45.58	28.63	25.25	26.83	12.80	8.07	9.90
LSA [6]	77.16	65.87	71.07	37.35	19.54	25.66	77.51	41.03	53.66	45.66	48.19	46.89	29.44	27.85	28.62	12.21	7.51	9.30
ZLAP [7]	76.35	74.74	75.54	32.41	25.51	28.55	68.22	54.77	60.76	48.18	47.29	47.73	29.38	27.10	28.19	12.64	10.42	11.32
TPR	**87.10**	76.81	**81.63**	**41.22**	26.87	**32.53**	**77.58**	**64.52**	**70.45**	50.47	45.40	47.80	**36.88**	**29.65**	**32.87**	**18.75**	**16.03**	**17.28**
TPR†	80.52	71.70	75.86	42.42	25.97	32.22	82.62	62.99	71.48	50.08	45.49	47.67	34.63	31.25	32.85	20.18	15.68	17.65
TPR‡	95.60	78.81	86.39	53.10	32.55	40.36	83.75	64.65	72.97	58.29	52.08	55.01	43.50	31.30	36.41	27.82	23.31	25.37

Model	StanfordCars*			EuroSAT			DTD			UCF101*			Food101*			OxfordPets*		
	S	*U*	*H*	*S*	*U*	*H*	*S*	*U*	*H*	*S*	*U*	*H*	*S*	*U*	*H*	*S*	*U*	*H*
CLIP [9]	46.65	37.78	41.75	21.13	11.25	14.68	36.39	41.39	38.73	53.72	64.92	58.79	67.74	73.05	70.29	82.67	65.83	73.29
CoOp [40]	49.86	38.47	43.43	29.89	12.27	17.40	44.34	36.56	40.07	62.13	47.41	53.78	71.82	64.64	68.04	73.47	57.66	64.61
CoCoOp [8]	51.93	37.84	43.78	52.64	18.34	27.21	42.19	35.94	38.82	58.29	60.62	59.43	72.55	60.42	65.93	72.53	58.97	65.05
MaPLe [10]	55.29	35.67	43.36	30.72	19.52	23.87	42.25	39.72	40.95	55.37	62.51	58.73	72.16	71.47	71.82	75.87	56.01	64.45
PromptSRC [11]	55.56	39.85	46.41	28.71	14.40	19.18	51.30	**42.56**	46.52	61.92	59.89	60.89	77.06	56.31	65.07	78.60	52.99	63.30
ProGrad [12]	52.20	35.36	42.16	59.14	17.12	26.55	54.62	39.78	46.03	60.10	58.76	59.42	73.48	64.26	68.56	72.53	59.95	65.64
CE [5]	56.62	40.94	47.52	61.61	32.88	42.88	44.79	29.61	35.65	54.78	34.66	42.46	70.48	53.88	61.07	71.07	59.54	64.80
LSA [6]	59.19	41.41	48.73	55.10	24.89	34.29	45.64	27.72	34.49	51.76	37.22	43.30	69.10	53.82	60.51	73.73	59.27	65.71
ZLAP [7]	49.60	43.22	46.19	74.59	23.22	35.41	46.94	30.94	37.30	56.89	42.65	48.75	70.20	60.86	65.20	72.93	59.89	65.77
TPR	**69.48**	**46.33**	**55.59**	**82.78**	**45.73**	**58.91**	**55.47**	42.06	**47.84**	**69.14**	**66.88**	**67.99**	**93.67**	**85.41**	**89.35**	**90.60**	**66.39**	**76.63**
TPR†	88.09	70.84	78.53	75.60	61.43	67.78	62.70	45.39	52.66	74.20	61.81	67.44	82.37	74.22	78.08	84.07	71.32	77.17
TPR‡	87.25	73.57	79.83	76.07	56.41	64.78	68.95	46.39	55.46	75.01	74.57	74.79	88.87	79.97	84.18	92.27	70.17	79.72

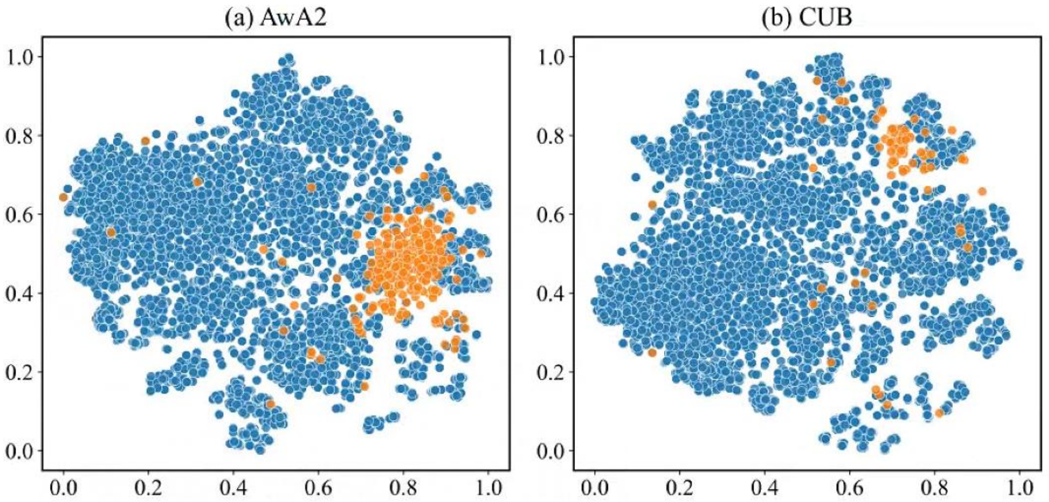
TPR

Correlation matrix between attributes

- high correlations between static attribute vocabulary
- low correlations between learnable attribute tokens
- complementary to each other



static learnable



TPR

Conclusion

- The attribute space and latent space are complementary to each other
- The latent space provides a general representation and the attribute space offers a more structured and interpretable representation
- The static vocabulary and learnable tokens are complementary to each other
- The static vocabulary learns prior knowledge and learnable tokens captures task-specific information
- Topology-preserving objective effectively keep the generalization capability of VLMs
- TPR achieves SOTA performances on both seen and unseen classes across multiple benchmarks

Thank you for your attention!