NEURAL INFORMATION PROCESSING SYSTEMS

北京交通大学 BEIJING JIAOTONG UNIVERSITY

TOMCAT: Test-time Comprehensive Knowledge Accumulation for Compositional Zero-Shot Learning

Presenter：Xudong Yan

Date：2025.11.3

03 Task Introduction

Compositional Zero-shot Learning

● Definition

Compositional Zero-Shot Learning (CZSL) aims to recognize novel attribute-object compositions based on the knowledge learned from seen ones during training.

● Key Scientific Problem

(1) **Compositional**: How to model the relationship between attributes and objects.

(2) **Zero-Shot Learning**: How to bridge the gap between seen and unseen compositions.

● Existing Challenges

(1) There exists label space shift between seen and unseen compositions. Using test-time images remains a big challenge.

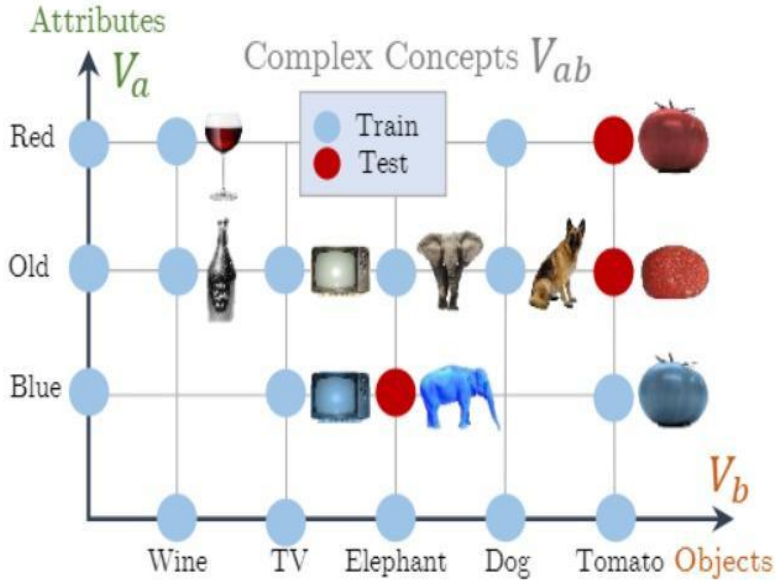
(2) The seen and unseen compositions in CZSL share strong information, but the recombination of attibutes and objects results in label changes.

● How can we solve them?

(1) Using training data to obtain knowledge of seen compositions.

(2) Using test data to optimize the prototypes in the label space.

(3) Recording historical samples to obtain visual knowledge.



TOMCAT: Test-time Comprehensive Knowledge Accumulation for Compositional Zero-Shot Learning

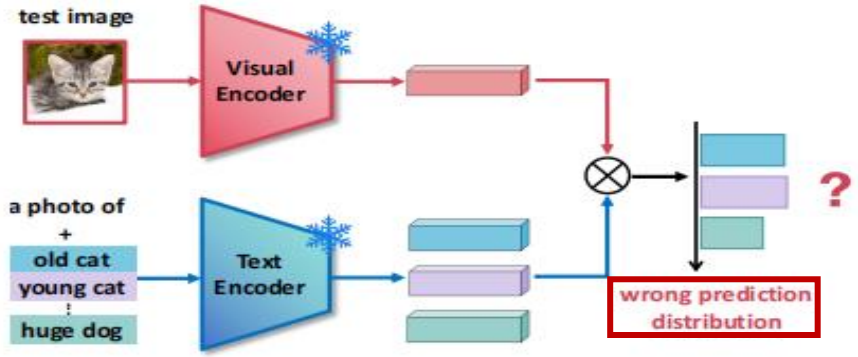
Motivation

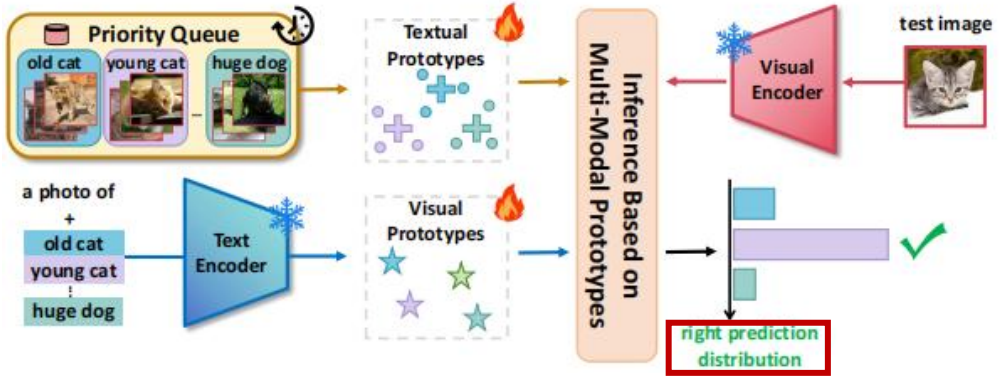
Existing CZSL methods focus on training phase, and they fail to do this:

1. Model keeps frozen at test-time, so they fail to leverage test images to overcome label space shift.

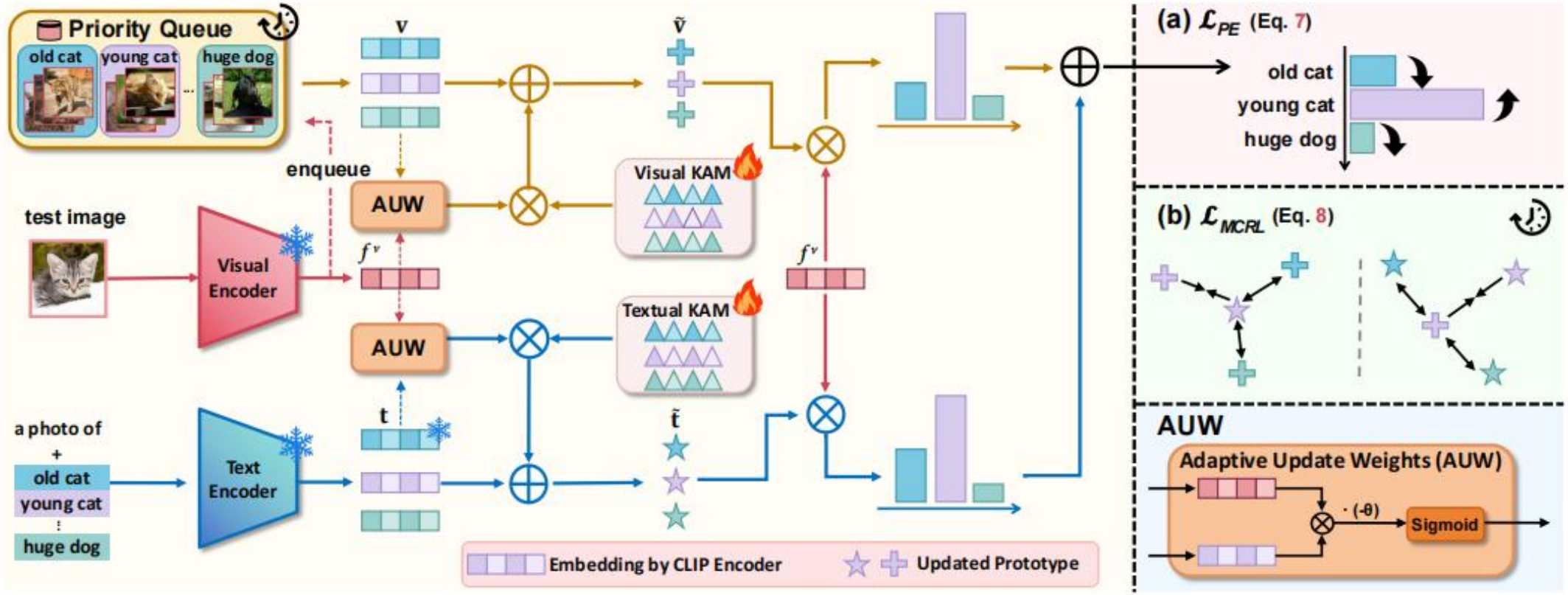
Key Points

1. Using entropy loss to optimize seen and unseen compositions.
2. Using historical images and label to get multimodal prototypes.
3. Updating prototypes based on similarity
4. Using multimodal collaborative learning for constraint





04 Research Method

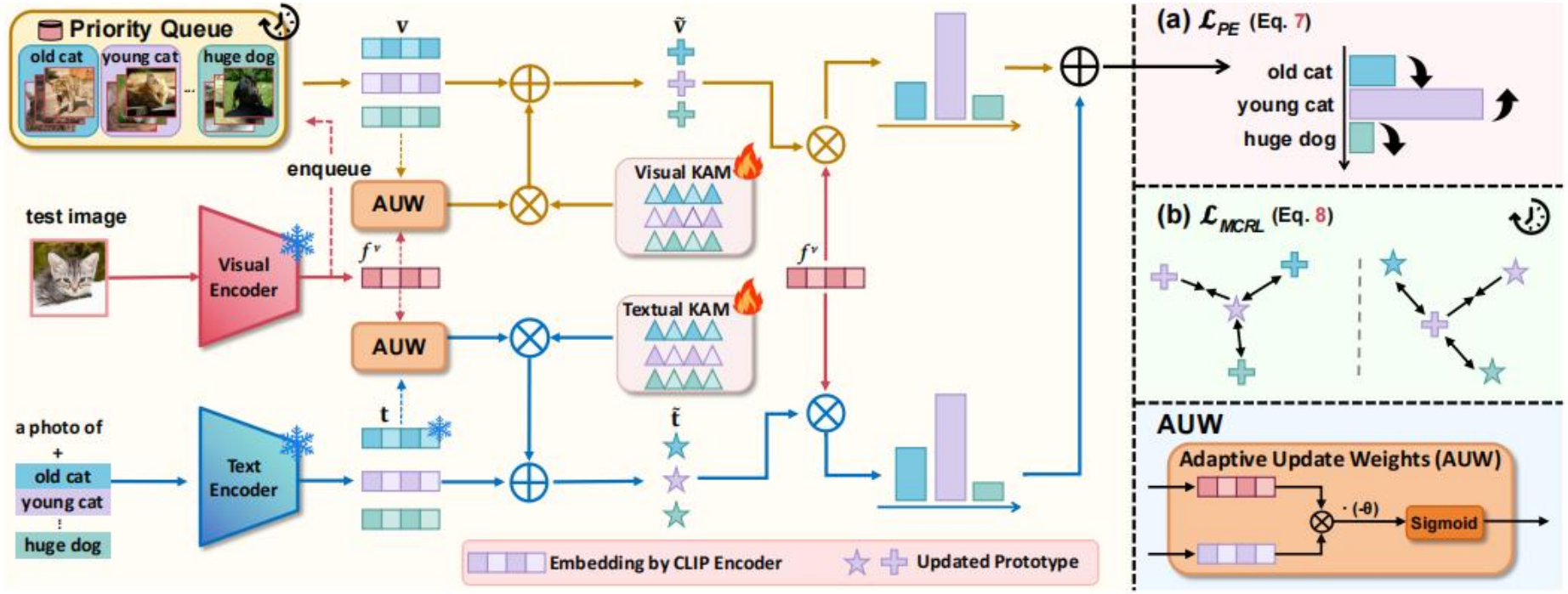


- **Calculating Multimodal Prototypes:**

$$\mathbf{t} = [\mathbf{t}_{c_1}, \mathbf{t}_{c_2}, \ldots, \mathbf{t}_{c_{|C^{te}|}}]^\top \quad \Delta\mathbf{t} = [\Delta\mathbf{t}_{c_1}, \Delta\mathbf{t}_{c_2}, \ldots, \Delta\mathbf{t}_{c_{|C^{te}|}}]^\top \quad w_c = \sigma(-\theta \cdot s_c), \quad s_c = \cos(f^v, \mathbf{t}_c)$$

$$\mathbf{v} = [\mathbf{v}_{c_1}, \mathbf{v}_{c_2}, \ldots, \mathbf{v}_{c_{|C^{te}|}}]^\top \quad \Delta\mathbf{v} = [\Delta\mathbf{v}_{c_1}, \Delta\mathbf{v}_{c_2}, \ldots, \Delta\mathbf{v}_{c_{|C^{te}|}}]^\top \quad \tilde{\mathbf{t}} = [\tilde{\mathbf{t}}_{c_1}, \tilde{\mathbf{t}}_{c_2}, \ldots, \tilde{\mathbf{t}}_{c_{|C^{te}|}}]^\top, \quad \tilde{\mathbf{t}}_c = \frac{\mathbf{t}_c + w_c\Delta\mathbf{t}_c}{||\mathbf{t}_c + w_c\Delta\mathbf{t}_c||}$$

04 Research Method



- **Entropy loss and multimodal collaborative loss:**

$$p(c|x,\tilde{\mathbf{t}},\tilde{\mathbf{v}}) = \frac{\exp\left(f^v \cdot \tilde{\mathbf{t}}_c + \alpha\mathcal{A}(f^v,\tilde{\mathbf{v}}_c)\right)}{\sum_{c'\in C^{te}} \exp\left(f^v \cdot \tilde{\mathbf{t}}_{c'} + \alpha\mathcal{A}(f^v,\tilde{\mathbf{v}}_{c'})\right)}$$

$$\mathcal{A}(f^v,\tilde{\mathbf{v}}_c) = \exp\left(-\beta(1 - f^v \cdot \tilde{\mathbf{v}}_c)\right)$$

$$\mathcal{L}_{PE} = -\sum_{c\in C^{te}} p(c|x,\tilde{\mathbf{t}}_c,\tilde{\mathbf{v}}_c)\log p(c|x,\tilde{\mathbf{t}}_c,\tilde{\mathbf{v}}_c)$$

$$\mathcal{L}_{MCRL} = -\frac{1}{2|C^{te}|}\sum_{c\in C^{te}}\left(\log\frac{\exp\left(\cos(\tilde{\mathbf{t}}_c,\tilde{\mathbf{v}}_c)/\tau\right)}{\sum_{c'\in C^{te}}\exp\left(\cos(\tilde{\mathbf{t}}_c,\tilde{\mathbf{v}}_{c'})/\tau\right)} + \log\frac{\exp\left(\cos(\tilde{\mathbf{t}}_c,\tilde{\mathbf{v}}_c)/\tau\right)}{\sum_{c'\in C^{te}}\exp\left(\cos(\tilde{\mathbf{t}}_{c'},\tilde{\mathbf{v}}_c)/\tau\right)}\right)$$

04 Experiments

Methods	UT-Zappos				MIT-States				C-GQA			
	AUC	HM	Seen	Unseen	AUC	HM	Seen	Unseen	AUC	HM	Seen	Unseen
Closed-world Results												
CLIP [43] *(ICML'21)*	5.0	15.6	15.8	49.1	11.0	26.1	30.2	46.0	1.4	8.6	7.5	25.0
CoOp [62] *(IJCV'22)*	18.8	34.6	52.1	49.3	13.5	29.8	34.4	47.6	4.4	17.1	20.5	26.8
Co-CGE [32] *(TPAMI'22)*	36.3	49.7	63.4	71.3	17.0	33.1	46.7	45.9	5.7	18.9	34.1	21.2
CSP [37] *(ICLR'23)*	33.0	46.6	64.2	66.2	19.4	36.3	46.6	49.9	6.2	20.5	28.8	26.8
DFSP [30] *(CVPR'23)*	36.0	47.2	66.7	71.7	20.6	37.3	46.9	52.0	10.5	27.1	38.2	32.0
GIPCOL [51] *(WACV'24)*	36.2	48.8	65.0	68.5	19.9	36.6	48.5	49.6	7.1	22.5	31.9	28.4
Troika [12] *(CVPR'24)*	41.7	54.6	66.8	73.8	22.1	39.3	49.0	**53.0**	**12.4**	**29.4**	**41.0**	**35.7**
CDS-CZSL [25] *(CVPR'24)*	39.5	52.7	63.9	**74.8**	22.4	39.2	**50.3**	52.9	11.1	28.1	38.3	34.2
PLID [3] *(ECCV'24)*	38.7	52.4	67.3	68.8	22.1	39.0	49.7	52.4	11.0	27.9	38.8	33.0
TOMCAT (Ours)	**48.3**	**60.2**	**74.5**	72.8	**22.6**	**39.5**	50.2	**53.0**	11.1	28.5	37.7	34.2
Open-world Results												
CLIP [43] *(ICML'21)*	2.2	11.2	15.7	20.6	3.0	12.8	30.1	14.3	0.3	4.0	7.5	4.6
CoOp [62] *(IJCV'22)*	13.2	28.9	52.1	31.5	2.8	12.3	34.6	9.3	0.7	5.5	21.0	4.6
Co-CGE [32] *(TPAMI'22)*	28.4	45.3	59.9	56.2	5.6	17.7	38.1	20.0	0.9	5.3	33.2	3.9
CSP [37] *(ICLR'23)*	22.7	38.9	64.1	44.1	5.7	17.4	46.3	15.7	1.2	6.9	28.7	5.2
DFSP [30] *(CVPR'23)*	30.3	44.0	66.8	60.0	6.8	19.3	47.5	18.5	2.4	10.4	38.3	7.2
GIPCOL [51] *(WACV'24)*	23.5	40.1	65.0	45.0	6.3	17.9	48.5	16.0	1.3	7.3	31.6	5.5
Troika [12] *(CVPR'24)*	33.0	47.8	66.4	61.2	7.2	20.1	48.8	18.7	2.7	10.9	**40.8**	7.9
CDS-CZSL* [25] *(CVPR'24)*	32.1	48.0	64.7	60.4	-	-	-	-	2.6	10.9	38.2	8.0
PLID [3] *(ECCV'24)*	30.8	46.6	67.6	55.5	7.3	20.4	49.1	18.7	2.5	10.6	39.1	7.5
TOMCAT (Ours)	**43.7**	**57.9**	**74.1**	**65.8**	**8.2**	**21.7**	**49.2**	**21.0**	**2.9**	**11.9**	37.7	**9.1**

04 Experiments

Table 3: Abaltion study of our proposed modules on UT-Zappos and MIT-States. Queue means visual priority queue. T- and V- KAM denote textual and visual KAM. AUW is adaptive update weights.

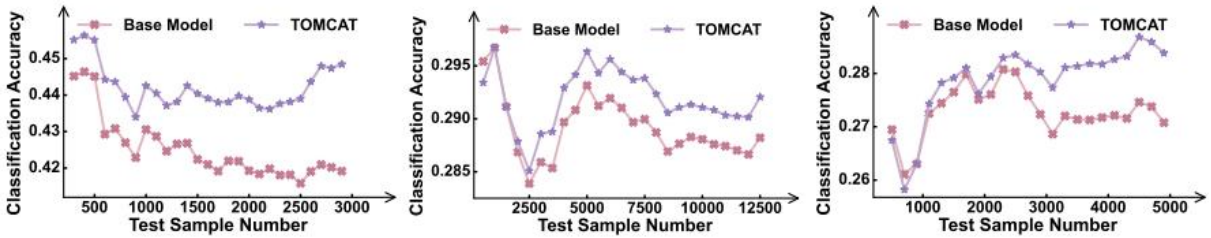
Module				UT-Zappos				MIT-States			
Queue	T-KAM	V-KAM	AUW	AUC	HM	Seen	Unseen	AUC	HM	Seen	Unseen
				43.57	55.54	68.72	**74.30**	22.12	38.97	49.50	52.61
✓				43.28	55.54	68.43	74.14	22.12	39.15	49.45	52.88
	✓			45.68	58.22	74.39	69.01	22.18	39.22	49.50	**52.95**
✓		✓		43.28	55.54	68.43	74.14	22.32	39.32	49.54	52.93
✓	✓	✓		46.64	58.83	76.44	68.38	22.34	39.39	49.58	52.93
✓	✓	✓	✓	**48.31**	**60.18**	**74.99**	72.77	**22.55**	**39.45**	**50.32**	**52.95**

Table 4: Ablation study of our designed loss on UT-Zappos and MIT-States.

$\mathcal{L}oss$		UT-Zappos		MIT-States	
$\mathcal{L}_{PE}$	$\mathcal{L}_{MCRL}$	AUC	HM	AUC	HM
		43.57	55.54	22.12	38.97
✓		44.59	57.29	22.35	39.32
	✓	42.46	53.97	22.29	39.42
✓	✓	**48.31**	**60.18**	**22.55**	**39.45**

Table 5: Influence of initialization strategies of KAMs on UT-Zappos and MIT-States.

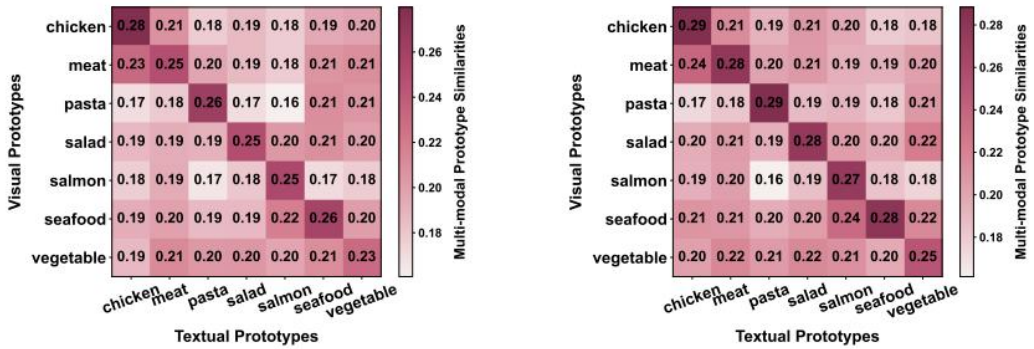
Initialization	UT-Zappos		MIT-States	
	AUC	HM	AUC	HM
Uniform Random	43.49	56.23	21.81	38.59
Normal Random	45.50	57.98	21.32	38.36
Random Walking	47.08	59.12	22.14	39.16
All Zeros	**48.31**	**60.18**	**22.55**	**39.45**



(a) UT-Zappos (b) MIT-States (c) C-GQA

Figure 5: Trend of top-1 classification accuracy with increasing test sample size on three datasets.

Figure 6: Case study on UT-Zappos and MIT-States. We compare TOMCAT (Ours) with the base model (BM) after training. The successful and failure results are marked in green and red, respectively.



(a) Test sample size of 2500. (b) Test sample size of 12000.

Figure 7: Similarity heatmap of multimodal prototypes on MIT-States. All unseen compositions consisting of the attribute `cooked` and its corresponding objects (*e.g.*, `chicken`, `meat`...) are selected.

NEURAL INFORMATION PROCESSING SYSTEMS

北京交通大學 BEIJING JIAOTONG UNIVERSITY

Thanks for Listening!

Persenter：Xudong Yan

Date：2025.11.3