香 港 大 學
THE UNIVERSITY OF HONG KONG

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

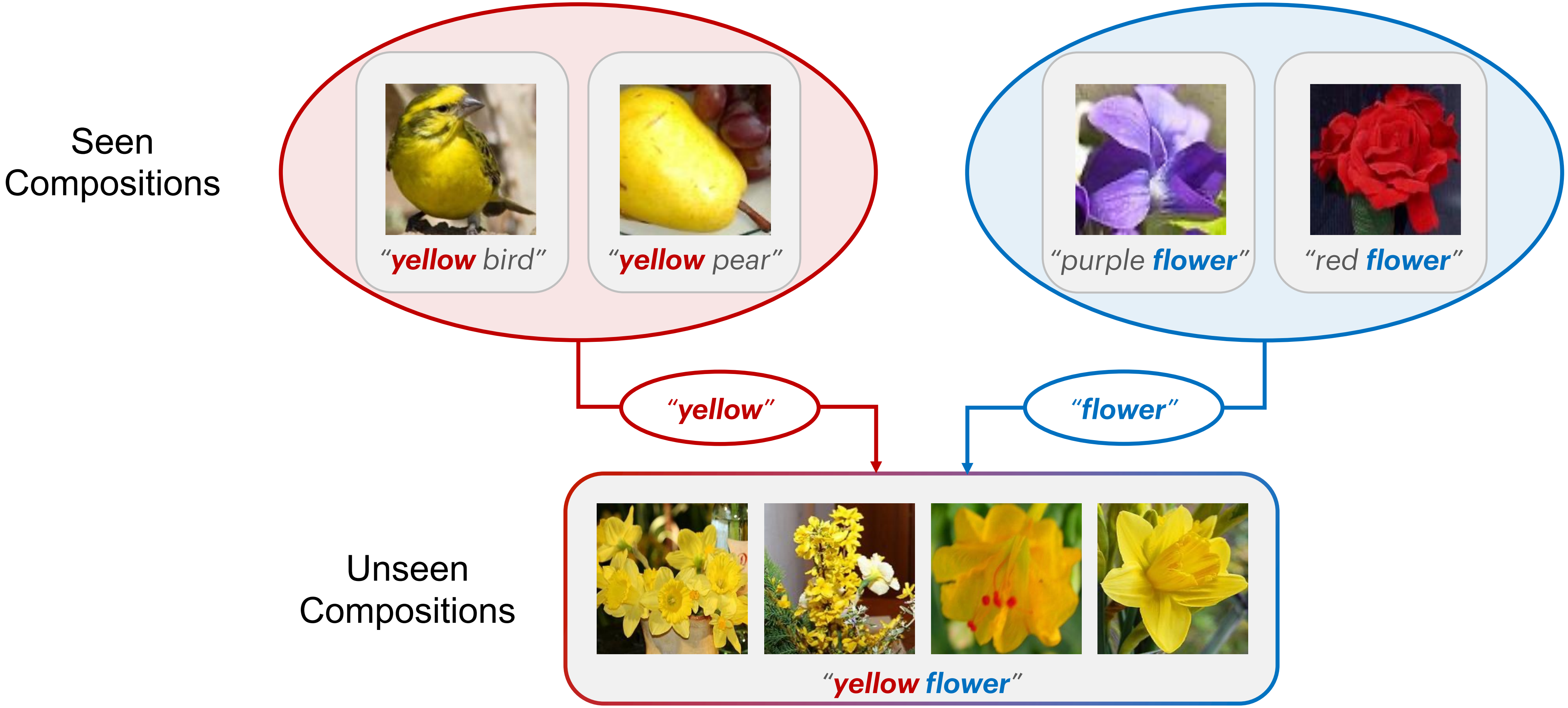
Learning Attention as Disentangler for Compositional Zero-shot Learning

Shaozhe Hao, Kai Han, Kwan-Yee K. Wong

The University of Hong Kong

WED-PM-282

Compositional Zero-shot Learning (CZSL)



Seen Compositions

"***yellow*** *bird*"

"***yellow*** *pear*"

"*purple* ***flower***"

"*red* ***flower***"

"***yellow***"

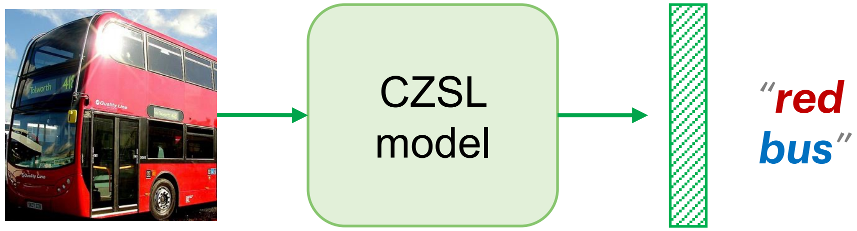
"***flower***"

Unseen Compositions

"***yellow flower***"

Quick Overview

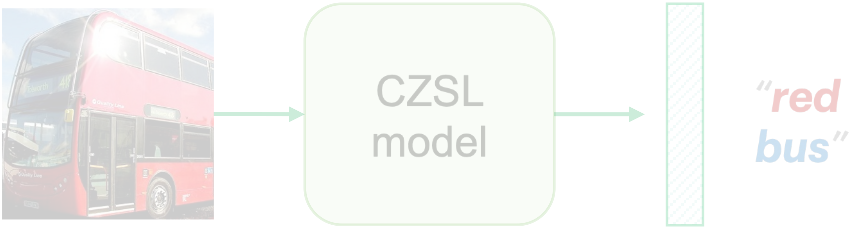
Baseline



learn attribute + object by cross entropy

Quick Overview

Baseline

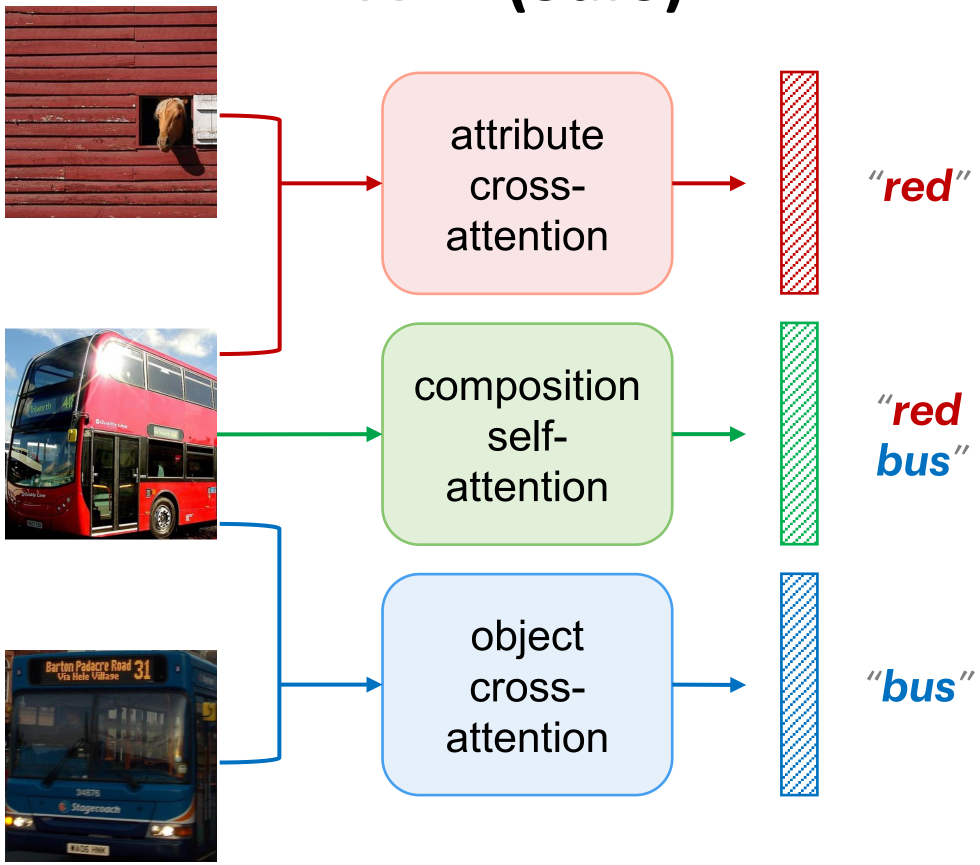


CZSL model

"red bus"

learn attribute + object by cross entropy

ADE (*ours*)



attribute cross-attention

"red"

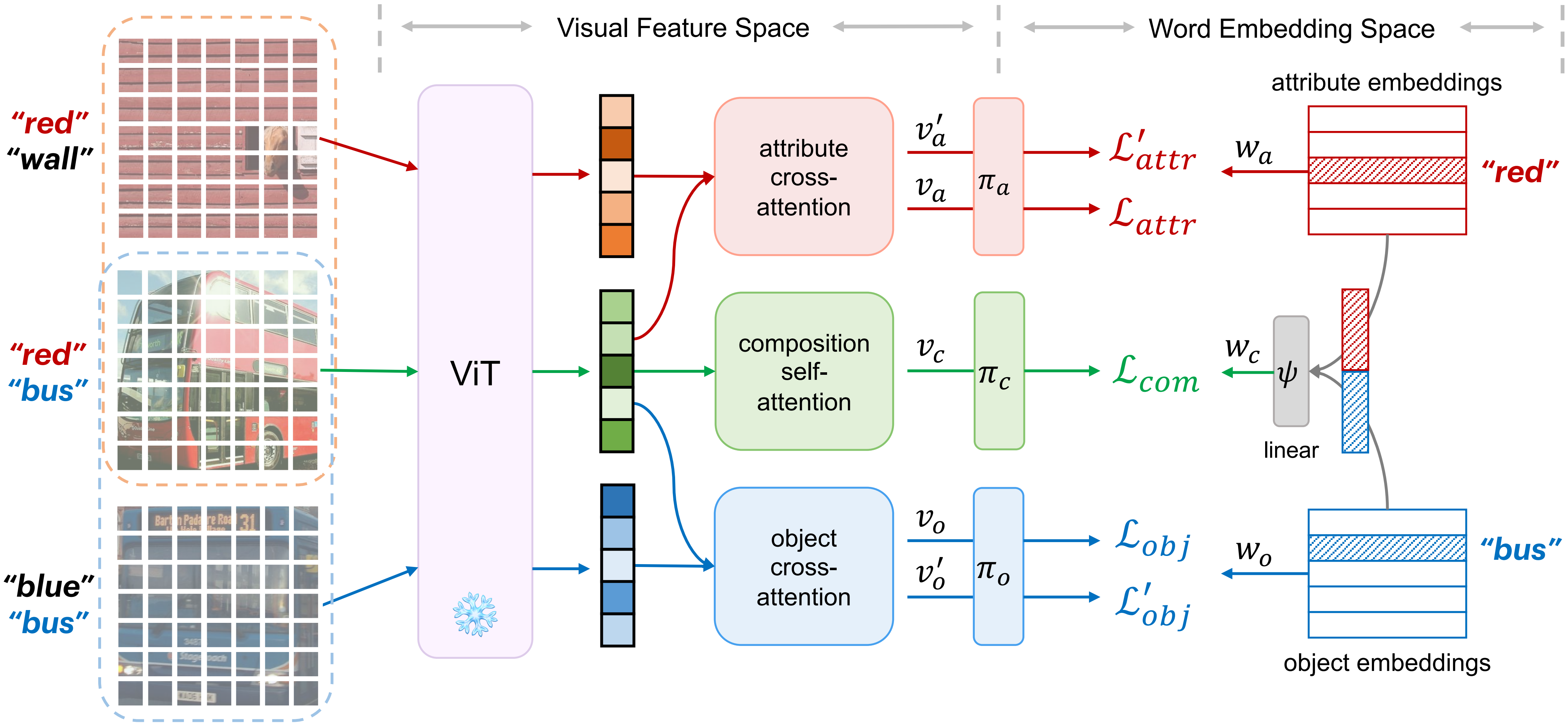
composition self-attention

"red bus"

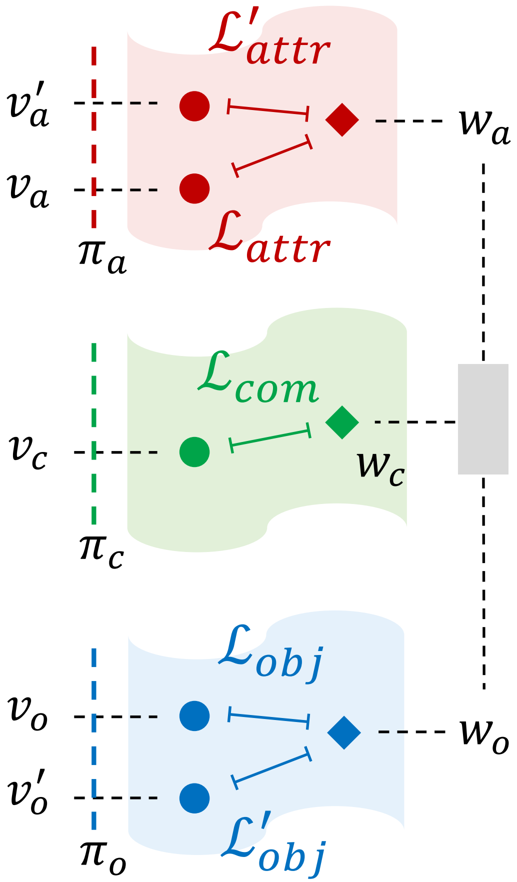
object cross-attention

"bus"

Attention as Disentangler (ADE)



Attention as Disentangler (ADE)

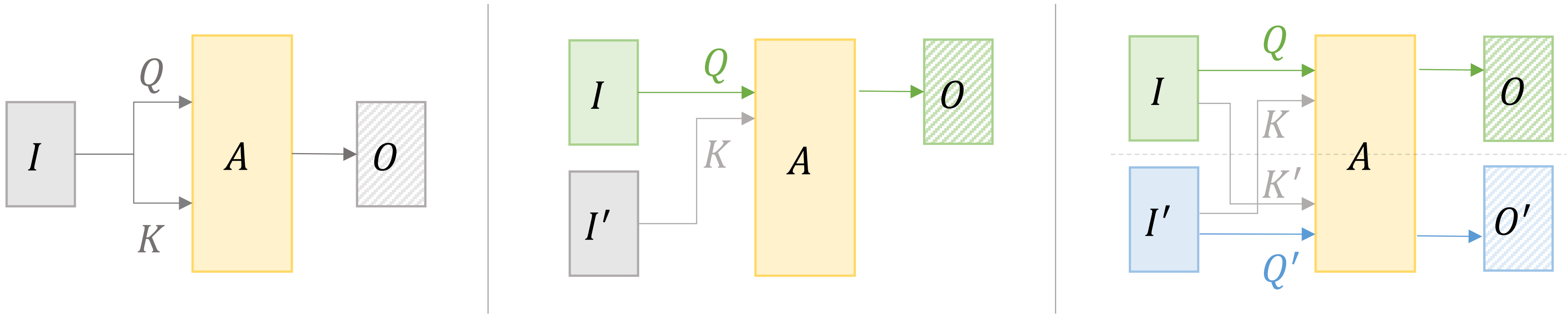


$$\mathcal{L}_{ce} = \underbrace{H_{\pi_a}(v_a, a)}_{\mathcal{L}_{attr}} + \underbrace{H_{\pi_a}(v'_a, a)}_{\mathcal{L}'_{attr}} + \underbrace{H_{\pi_c}(v_c, c)}_{\mathcal{L}_{com}} + \underbrace{H_{\pi_o}(v_o, o)}_{\mathcal{L}_{obj}} + \underbrace{H_{\pi_o}(v'_o, o)}_{\mathcal{L}'_{obj}}$$

Attention as Disentangler (ADE)

Cross-attention with query-key swapping (QKS)

I, I': inputs A: attention module O, O': outputs



self-attention

cross-attention

cross-attention w/ QKS

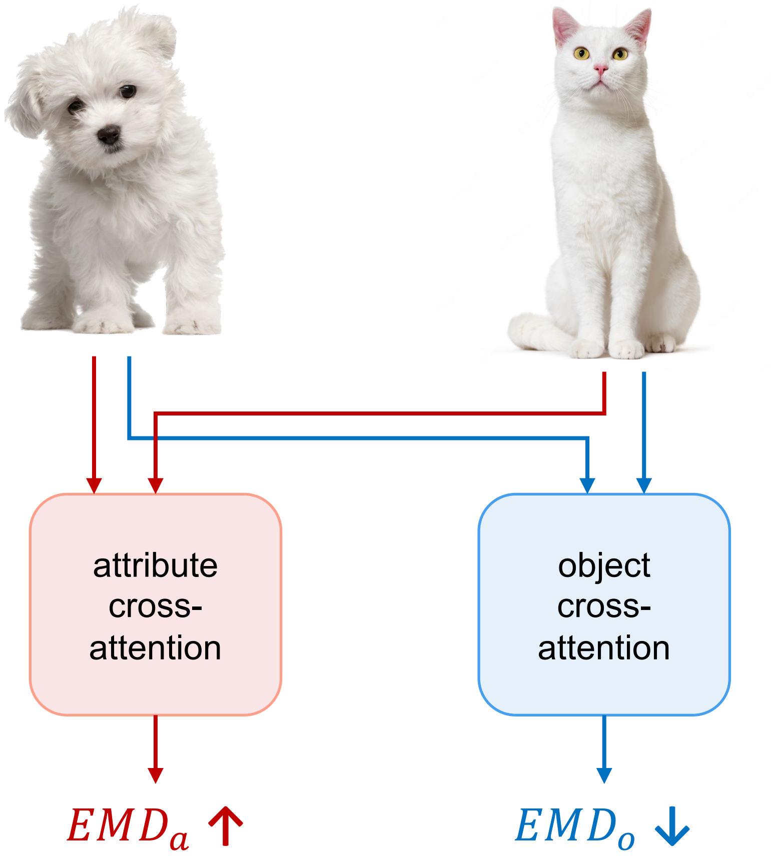
Attention as Disentangler (ADE)

Earth moving distance (EMD)

$$
\begin{aligned}
\underset{f_{ij}}{\text{minimize}} \quad & \sum\nolimits_{i=1}^{n_s} \sum\nolimits_{j=1}^{n_d} c_{ij} f_{ij} \\
\text{subject to} \quad & f_{ij} \geqslant 0,\ i = 1, ..., n_s,\ j = 1, ..., n_d \\
& \sum\nolimits_{j=1}^{n_d} f_{ij} = s_i,\ i = 1, ..., n_s \\
& \sum\nolimits_{i=1}^{n_s} f_{ij} = d_j,\ j = 1, ..., n_d
\end{aligned}
$$

$$\mathrm{EMD}(c_{ij}, s_i, d_j) = (1 - c_{ij})\tilde{f}_{ij}.$$

Greater EMD,
Closer distributions,
More focused on the concept



attribute cross-attention

object cross-attention

EMD_a ↑

EMD_o ↓

same attribute

different objects

$L^a_{reg} = EMD_o - EMD_a$, and vice versa

Attention as Disentangler (ADE)

Earth moving distance (EMD)

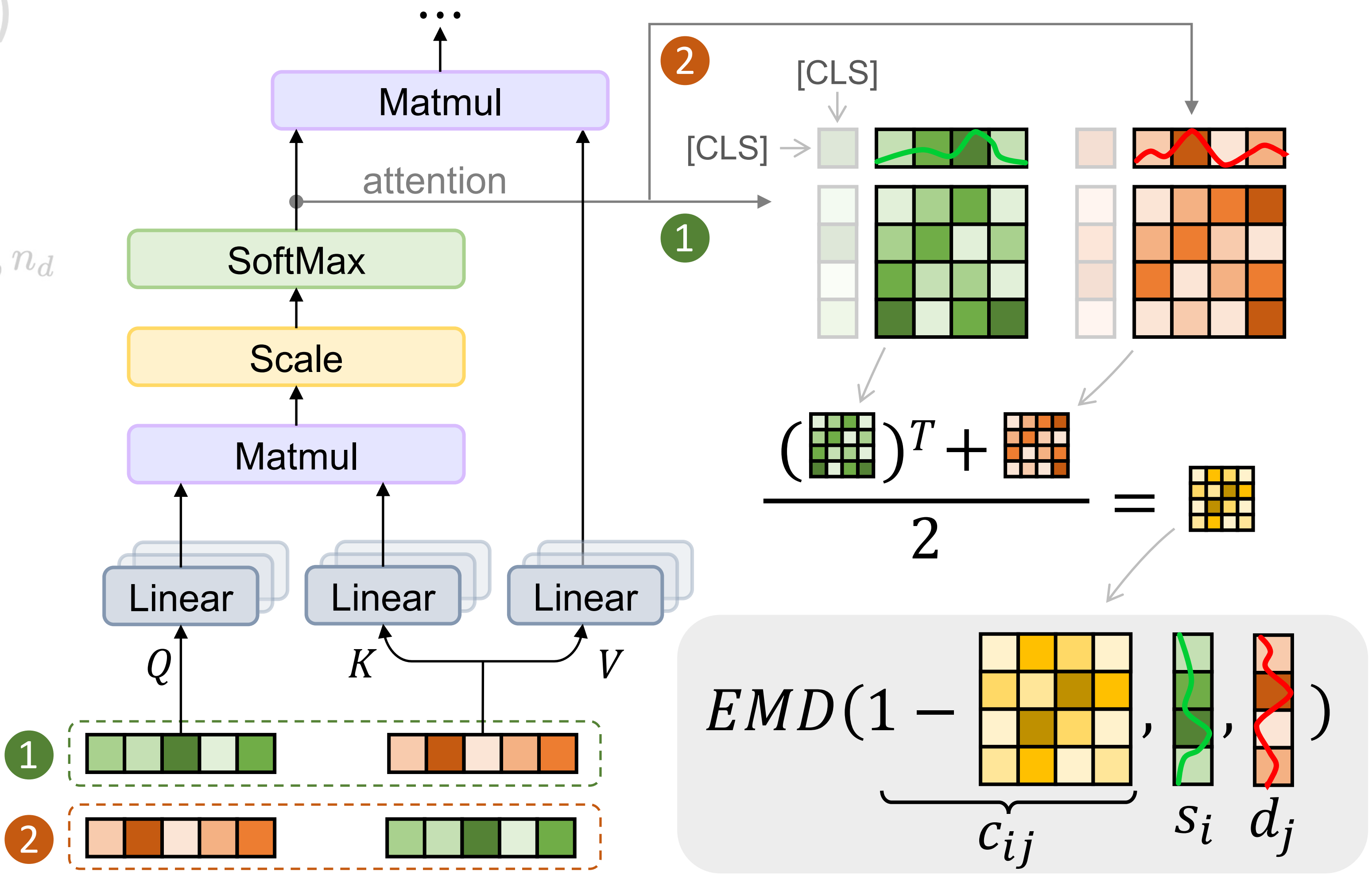
$$\begin{aligned}\underset{f_{ij}}{\text{minimize}} \quad & \sum_{i=1}^{n_s}\sum_{j=1}^{n_d} c_{ij} f_{ij} \\ \text{subject to} \quad & f_{ij} \geqslant 0,\ i=1,\ldots,n_s,\ j=1,\ldots,n_d \\ & \sum_{j=1}^{n_d} f_{ij} = s_i,\ i=1,\ldots,n_s \\ & \sum_{i=1}^{n_s} f_{ij} = d_j,\ j=1,\ldots,n_d\end{aligned}$$

$$\mathrm{EMD}(c_{ij}, s_i, d_j) = (1 - c_{ij})\tilde{f}_{ij}.$$

Greater EMD,

Closer distributions,

More focused on the concept



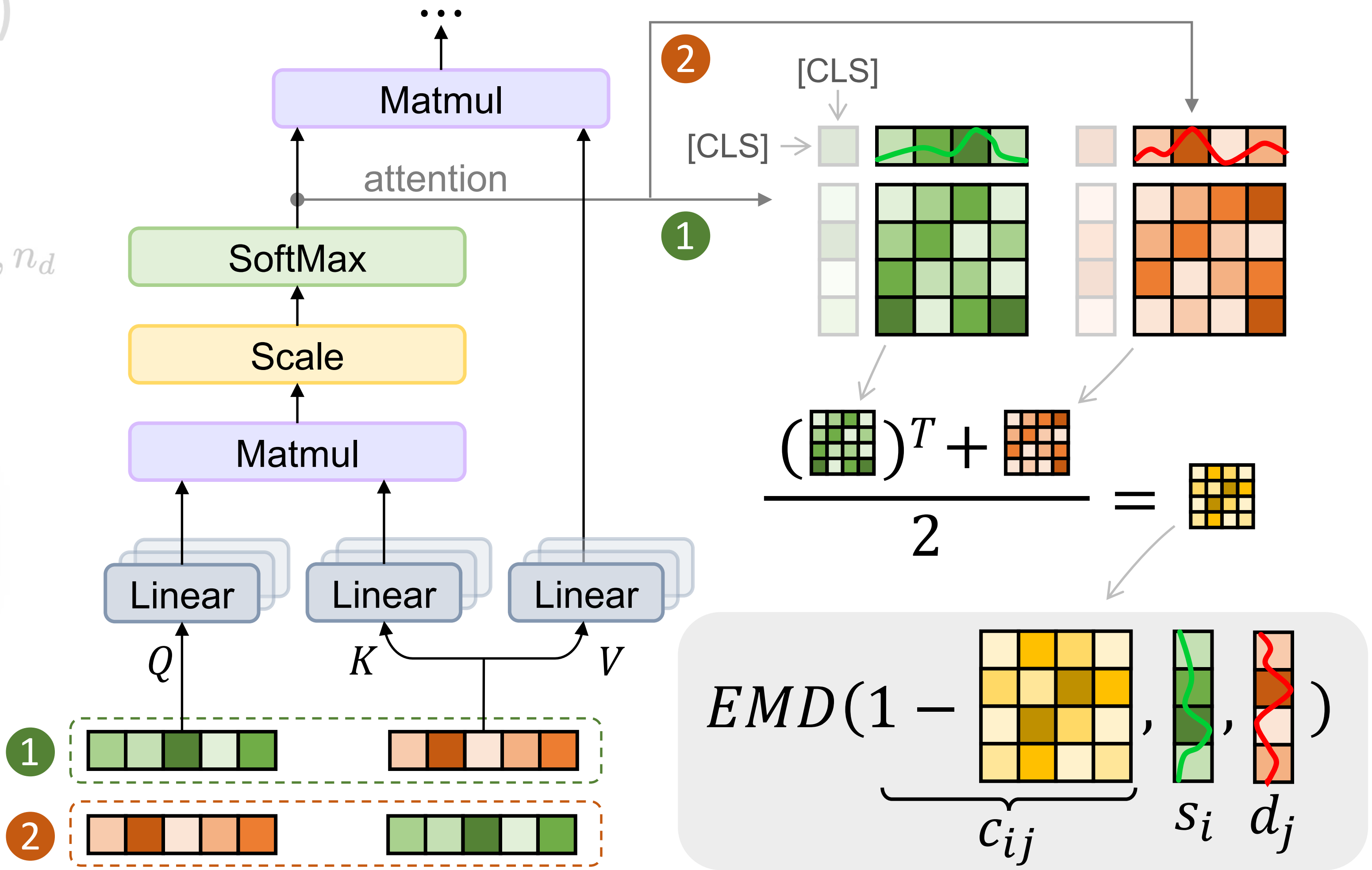
Attention as Disentangler (ADE)

Earth moving distance (EMD)

$$\begin{aligned}
\underset{f_{ij}}{\text{minimize}} \quad & \sum_{i=1}^{n_s}\sum_{j=1}^{n_d} c_{ij} f_{ij} \\
\text{subject to} \quad & f_{ij} \geqslant 0,\ i = 1, \ldots, n_s,\ j = 1, \ldots, n_d \\
& \sum_{j=1}^{n_d} f_{ij} = s_i,\ i = 1, \ldots, n_s \\
& \sum_{i=1}^{n_s} f_{ij} = d_j,\ j = 1, \ldots, n_d
\end{aligned}$$

$$\mathrm{EMD}(c_{ij}, s_i, d_j) = (1 - c_{ij})\tilde{f}_{ij}.$$

Greater EMD,
Closer distributions,
More focused on the concept



Training and Inference

- Training objective

$$\mathcal{L} = \mathcal{L}_{ce} + \mathcal{L}_{reg}$$

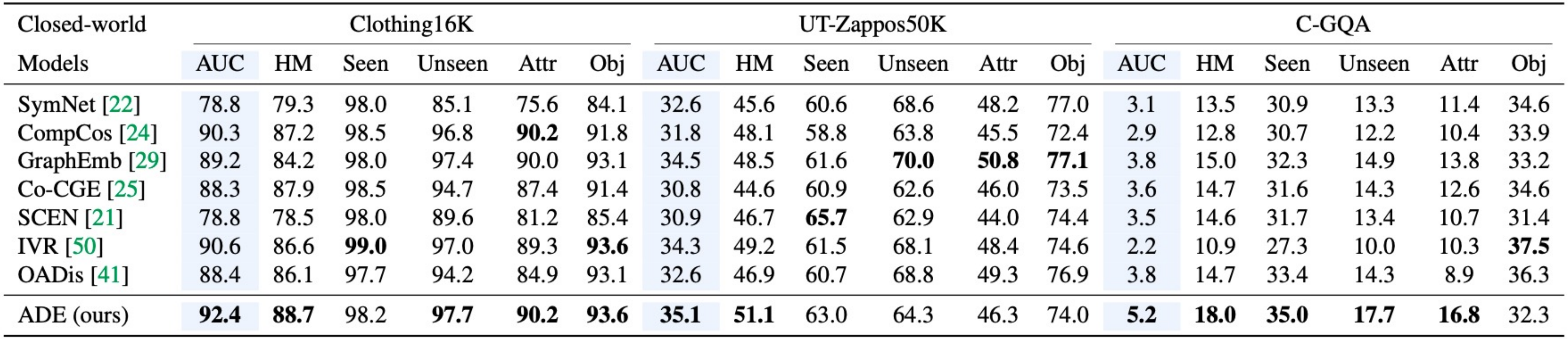
- Inference: score tuning

$$\hat{c} = \underset{c \in \mathcal{C}_{test}}{\arg\max}\ p(c) + \beta \cdot p(a) \cdot p(o)$$

choose the best β on the validation set

Comparison with SOTA methods

➢ Closed-world evaluation

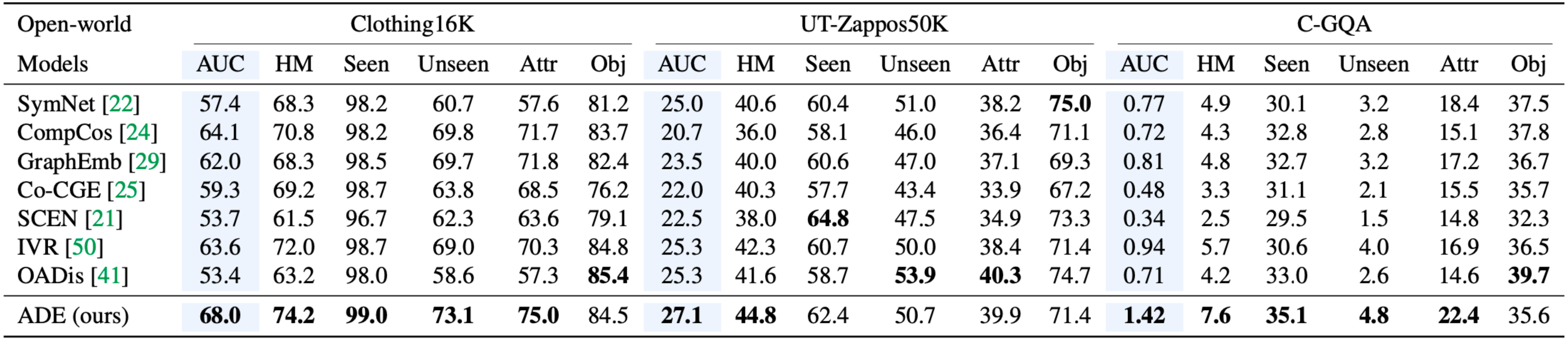


Closed-world	Clothing16K						UT-Zappos50K						C-GQA					
Models	AUC	HM	Seen	Unseen	Attr	Obj	AUC	HM	Seen	Unseen	Attr	Obj	AUC	HM	Seen	Unseen	Attr	Obj
SymNet [22]	78.8	79.3	98.0	85.1	75.6	84.1	32.6	45.6	60.6	68.6	48.2	77.0	3.1	13.5	30.9	13.3	11.4	34.6
CompCos [24]	90.3	87.2	98.5	96.8	**90.2**	91.8	31.8	48.1	58.8	63.8	45.5	72.4	2.9	12.8	30.7	12.2	10.4	33.9
GraphEmb [29]	89.2	84.2	98.0	97.4	90.0	93.1	34.5	48.5	61.6	**70.0**	**50.8**	**77.1**	3.8	15.0	32.3	14.9	13.8	33.2
Co-CGE [25]	88.3	87.9	98.5	94.7	87.4	91.4	30.8	44.6	60.9	62.6	46.0	73.5	3.6	14.7	31.6	14.3	12.6	34.6
SCEN [21]	78.8	78.5	98.0	89.6	81.2	85.4	30.9	46.7	**65.7**	62.9	44.0	74.4	3.5	14.6	31.7	13.4	10.7	31.4
IVR [50]	90.6	86.6	**99.0**	97.0	89.3	**93.6**	34.3	49.2	61.5	68.1	48.4	74.6	2.2	10.9	27.3	10.0	10.3	**37.5**
OADis [41]	88.4	86.1	97.7	94.2	84.9	93.1	32.6	46.9	60.7	68.8	49.3	76.9	3.8	14.7	33.4	14.3	8.9	36.3
ADE (ours)	**92.4**	**88.7**	98.2	**97.7**	**90.2**	**93.6**	**35.1**	**51.1**	63.0	64.3	46.3	74.0	**5.2**	**18.0**	**35.0**	**17.7**	**16.8**	32.3

Evaluate on a predefined composition **subset**

Comparison with SOTA methods

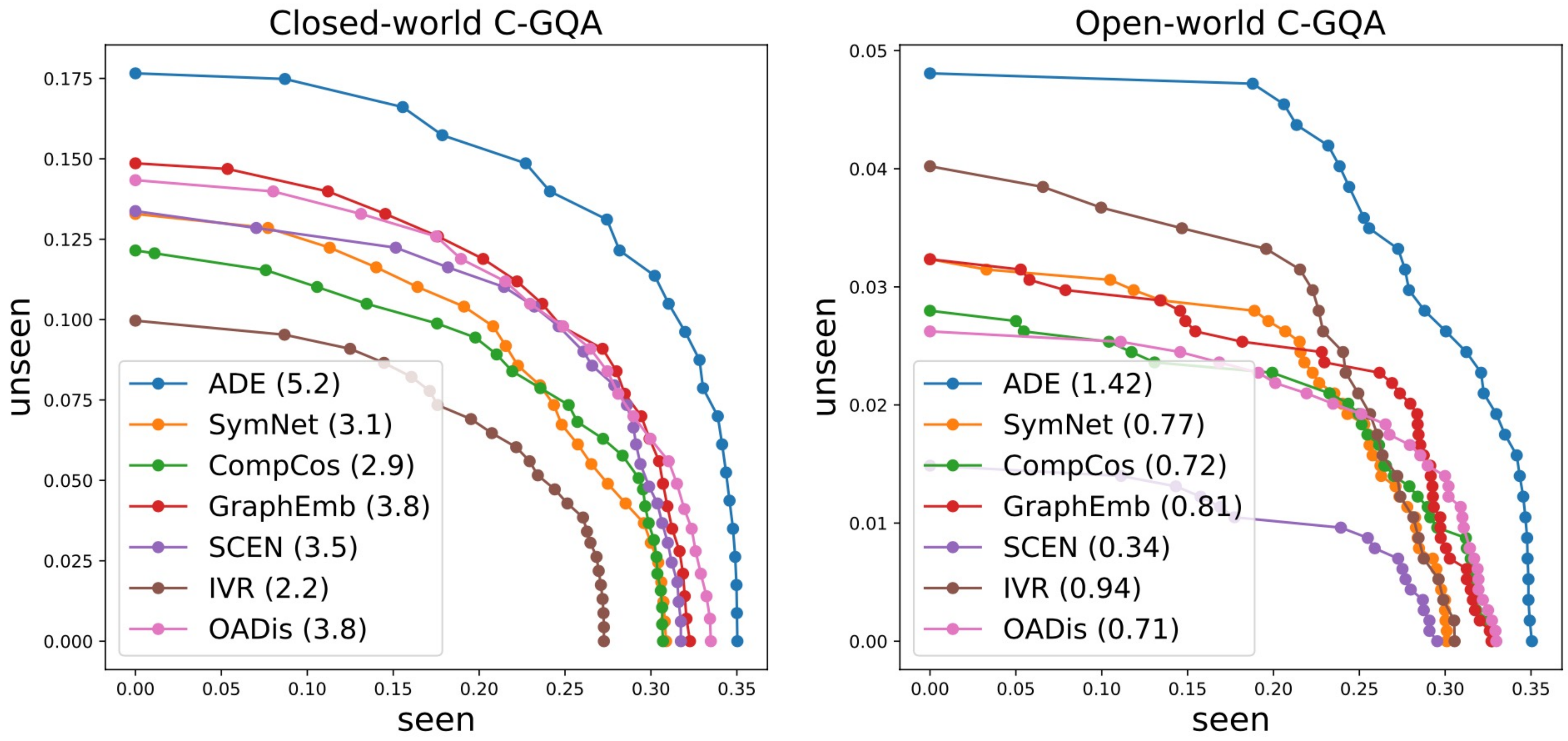
- ➢ Open-world evaluation



Open-world	Clothing16K						UT-Zappos50K						C-GQA					
Models	AUC	HM	Seen	Unseen	Attr	Obj	AUC	HM	Seen	Unseen	Attr	Obj	AUC	HM	Seen	Unseen	Attr	Obj
SymNet [22]	57.4	68.3	98.2	60.7	57.6	81.2	25.0	40.6	60.4	51.0	38.2	**75.0**	0.77	4.9	30.1	3.2	18.4	37.5
CompCos [24]	64.1	70.8	98.2	69.8	71.7	83.7	20.7	36.0	58.1	46.0	36.4	71.1	0.72	4.3	32.8	2.8	15.1	37.8
GraphEmb [29]	62.0	68.3	98.5	69.7	71.8	82.4	23.5	40.0	60.6	47.0	37.1	69.3	0.81	4.8	32.7	3.2	17.2	36.7
Co-CGE [25]	59.3	69.2	98.7	63.8	68.5	76.2	22.0	40.3	57.7	43.4	33.9	67.2	0.48	3.3	31.1	2.1	15.5	35.7
SCEN [21]	53.7	61.5	96.7	62.3	63.6	79.1	22.5	38.0	**64.8**	47.5	34.9	73.3	0.34	2.5	29.5	1.5	14.8	32.3
IVR [50]	63.6	72.0	98.7	69.0	70.3	84.8	25.3	42.3	60.7	50.0	38.4	71.4	0.94	5.7	30.6	4.0	16.9	36.5
OADis [41]	53.4	63.2	98.0	58.6	57.3	**85.4**	25.3	41.6	58.7	**53.9**	**40.3**	74.7	0.71	4.2	33.0	2.6	14.6	**39.7**
ADE (ours)	**68.0**	**74.2**	**99.0**	**73.1**	**75.0**	84.5	**27.1**	**44.8**	62.4	50.7	39.9	71.4	**1.42**	**7.6**	**35.1**	**4.8**	**22.4**	35.6

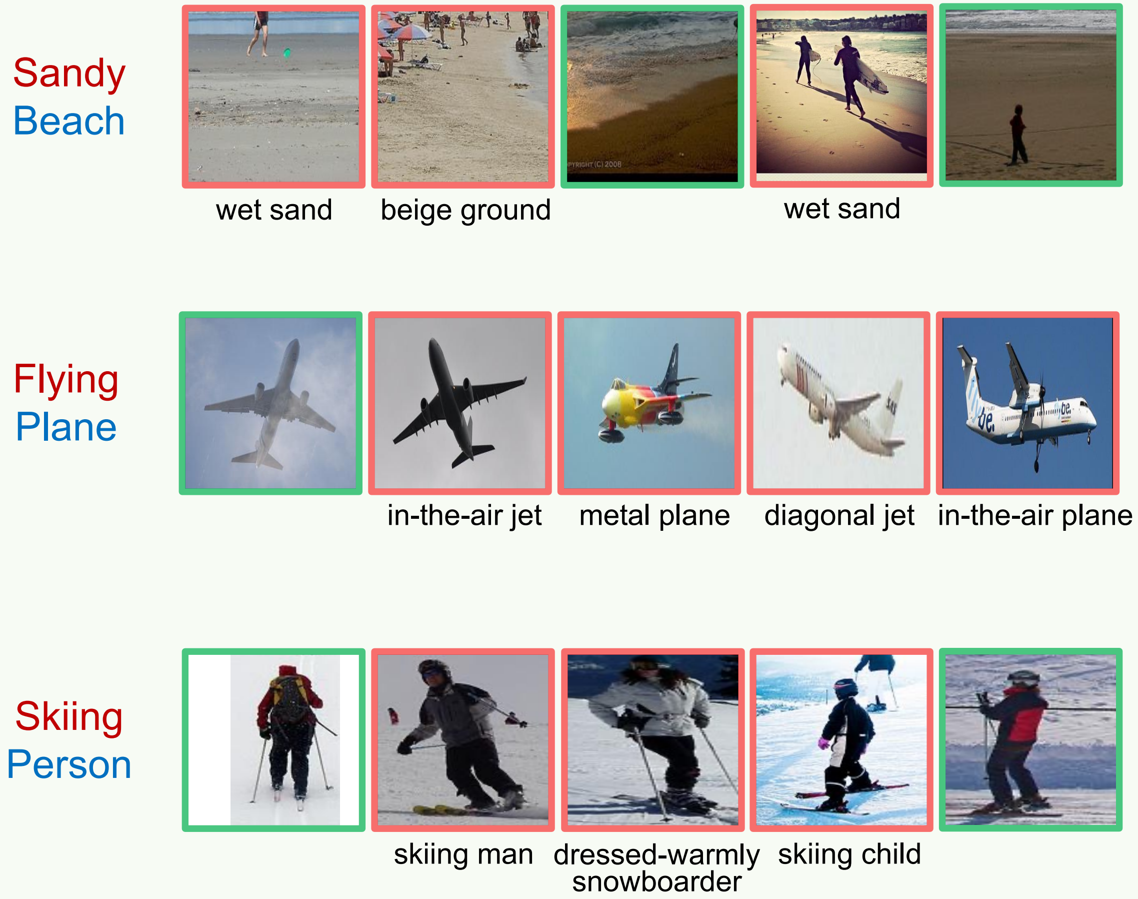
Evaluate on **all** compositions

Seen-Unseen Accuracy Curve on C-GQA

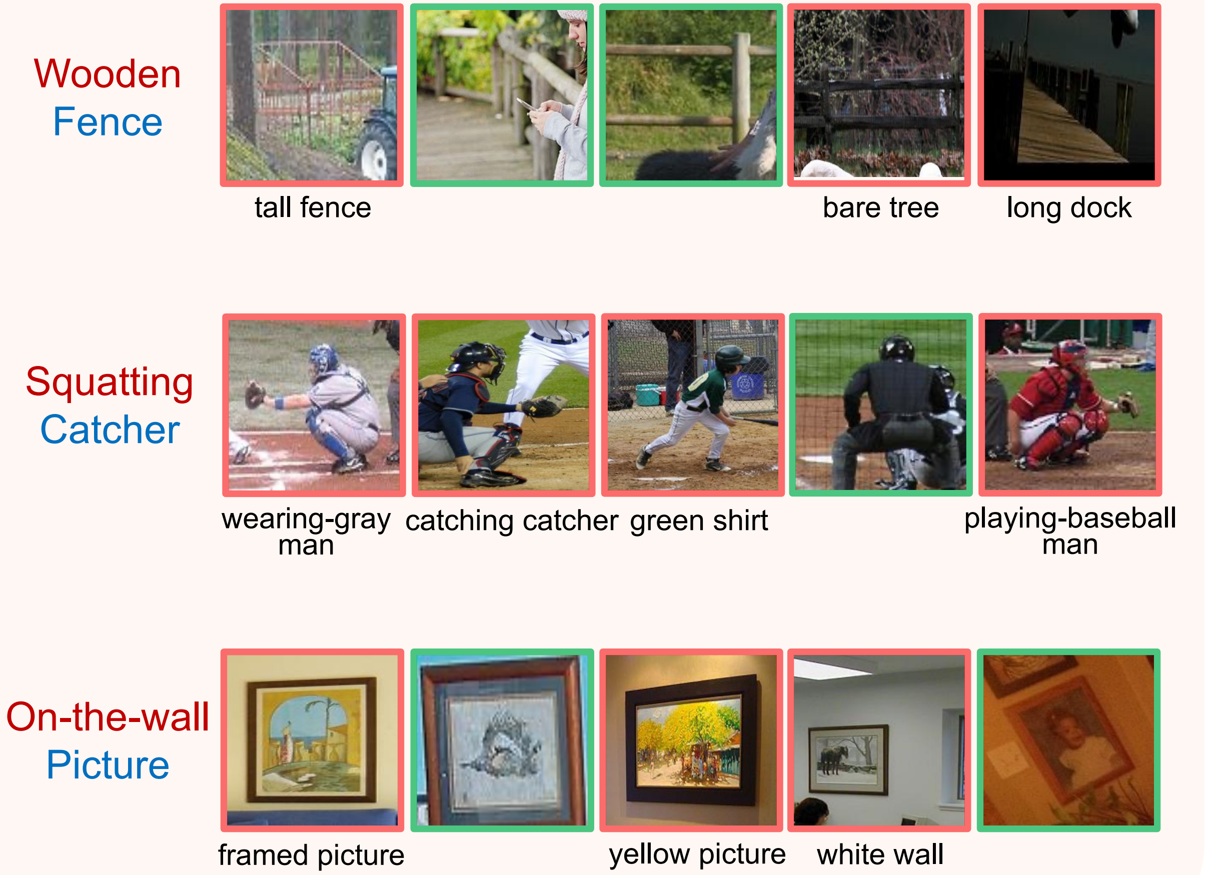


Applications – Text-to-Image Retrieval

Seen compositions

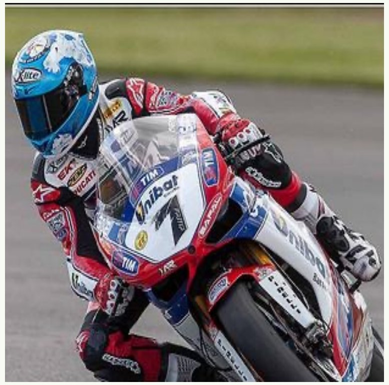


Unseen compositions



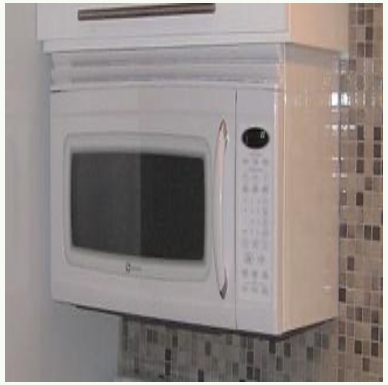
Applications – Image-to-Text Retrieval

Seen compositions



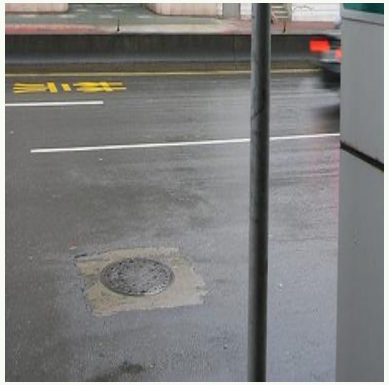
Multicolored Clothing

Colorful Suit
Colorful Clothing
Red Suit
Multicolored Suit
Multicolored Clothing



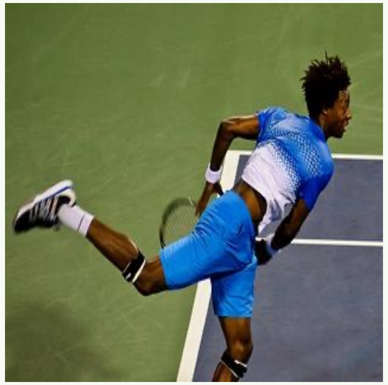
Rectangular Microwave

Rectangular Microwave
Turned-off Microwave
Digital Microwave
Closed Microwave
White Microwave



Wet Road

Asphalt Street
Asphalt Road
Wet Road
Paved Street
Wet Street



Jumping Tennis-player

Jumping Tennis-player
Playing-tennis Tennis-player
Wearing-green Tennis-player
Wearing-blue Tennis-player
Jumping Player

Applications – Image-to-Text Retrieval

Unseen compositions

Brown Carpet

Clean Carpet
Tan Carpet
Beige Carpet
Rectangular Carpet
Brown Carpet

Squatting Umpire

Dressed Umpire
Kneeling Player
Dressed Catcher
Squatting Player
Kneeling Catcher

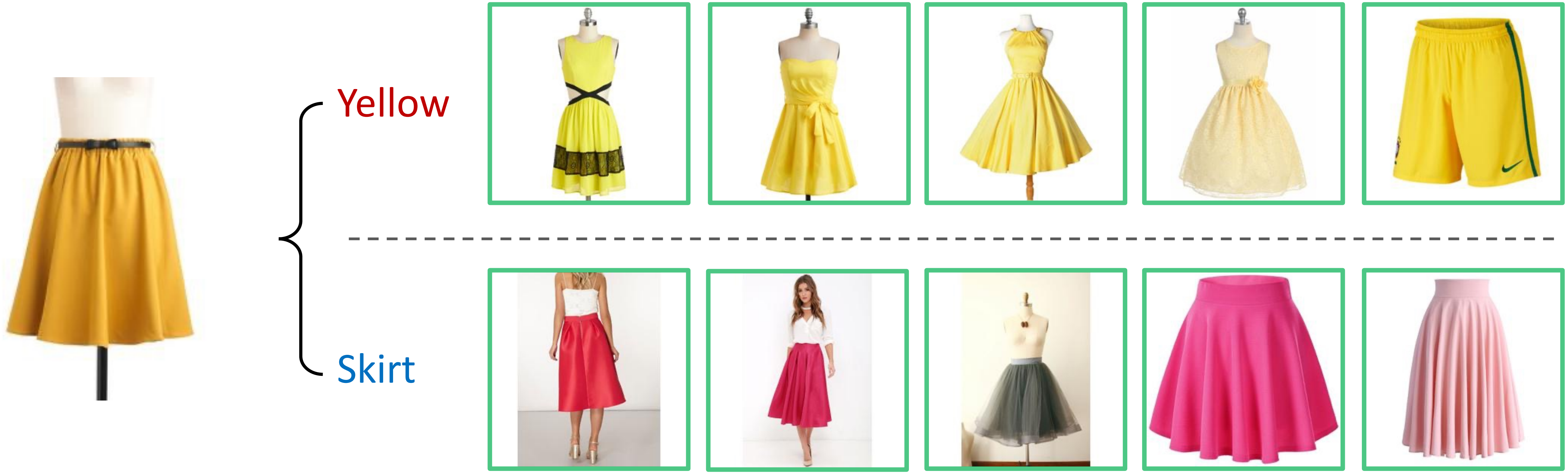
Spotted Neck

Brown Spot
Spotted Fur
Spotted Neck
Long Neck
Brown Fur

Metal Fence

Metal Pole
Gray Fence
Gray Metal
Gray Wire
Metal Leg

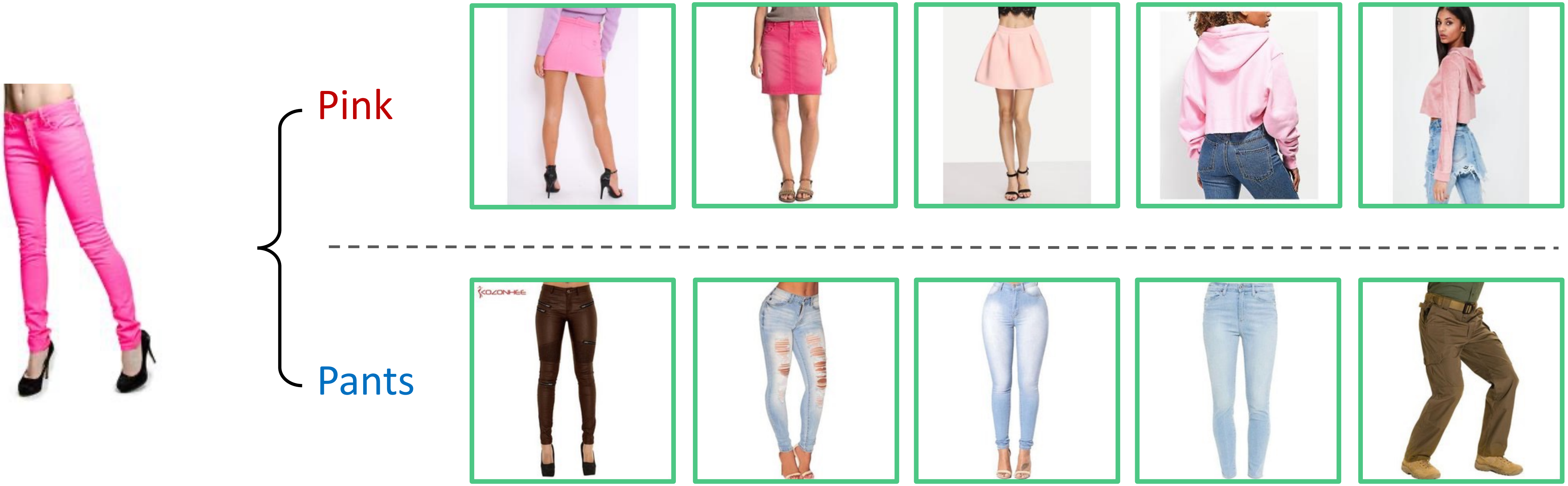
Applications – Visual Concept Retrieval



Yellow

Skirt

Applications – Visual Concept Retrieval



Pink

Pants

Thank you for your listening!

Welcome to our Poster: **WED-PM-282**

Code & Model:
https://github.com/haoosz/ade-czsl