Hierarchical Visual Primitive Experts for Compositional Zero-Shot Learning -Supplementary Materials-

Hanjae Kim[1] Jiyoung Lee[2] Seongheon Park[1] Kwanghoon Sohn[1,3,*]

[1]Yonsei University [2]NAVER AI Lab [3]Korea Institute of Science and Technology (KIST)

{incohjk,sam121796,khsohn}@yonsei.ac.kr lee.j@navercorp.com

This document presents supplementary materials for the Anonymous ICCV 2023 submission, "Hierarchical Visual Primitive Experts for Compositional Zero-Shot Learning". In Sec. B, we describe the implementation details and applicability of minority attribute augmentation. We report more experimental results and discussion in Sec. C, and computational analysis in Sec. D.

A. Data statistics

Table 1 shows detailed data statistics of MIT-States [3], C-GQA [6] and VAW-CZSL [8]. Compared to MIT-States, the latest C-GQA and VAW-CZSL have a large number of attribute, object and composition labels, effective for discussing CZSL problems on realistic scenarios.

A.1. Long-tailed distribution

In Fig. 1, we visualize the distributions of composition class ids in the training set. All datasets, especially for C-GQA and VAW-CZSL having a large number of composition classes, show the long-tailed distribution of compositions. This is a natural effect because we can easily guess that 'black dog' is more frequent than 'blue dog' in the real world. Fig. 2 illustrates the imbalanced attribute composition (*e.g*., 'white box' are 8 times more frequent than 'pink box'). However, this phenomenon makes it difficult to predict various and novel compositions. Therefore, we proposed minority attribute augmentation (MAA), which remedies a biased prediction caused by the imbalanced data distribution.

B. Implementation details of MAA

We summarize our training procedure of the proposed MAA in Algorithm 1. An auxiliary image x_B is sampled with a sampling weight κ, which has a different attribute class to a given input x_A. Then, a virtual sample (x_M, y_M) is generated by blending the input with the sampled auxiliary image. We first optimize the object and attribute experts in CoT with the generated virtual samples utilizing object and attribute losses (*i.e*., $\mathcal{L}_{\text{obj}}$ and $\mathcal{L}_{\text{att}}$) in the main paper. To align a virtual visual prototype $\mathbf{p}_{\text{vx}}$ from x_M with a semantic prototype $\mathbf{p}_{\text{vy}} = g([w(o), w(a_M)])$, we simply modify the compositional loss $\mathcal{L}_{\text{comp}}$ with the virtual label as follows:

Algorithm 1: Minority attribute augmentation

Require: Training dataset $\mathcal{D}_{\text{tr}}$.
Initialize: Model parameter θ.

while *Training* **do**
  for $(x_A, a_A, o) \in \mathcal{D}_{tr}$ **do**
    while $a_B \neq a_A$ **do**
      Sample $(x_B, a_B, o) \in \mathcal{D}_{\text{tr}}$ with $\kappa(a_B, o)$
    Get $\mathbf{p}_{x_A}$ and $\mathbf{p}_{x_B}$ from CoT
    Sample $\lambda \sim Beta(1,1)$
    $\mathbf{p}_{x_M} = \lambda \mathbf{p}_{x_A} + (1-\lambda)\mathbf{p}_{x_B}$
    $w(a_M) = \lambda w(a_A) + (1-\lambda) w(a_B)$
    $\theta \leftarrow \theta - \nabla \mathcal{L}_{\text{total}}(\mathbf{p}_{x_M}, w(a_M), w(o))$

$$\mathcal{L}_{\text{comp}} = -\log \frac{\exp\{d(\mathbf{p}_{\text{vx}}, \mathbf{p}_{\text{vy}})/\tau_c\}}{\sum_{y_k \in \mathcal{Y}_M} \exp\{d(\mathbf{p}_{\text{vx}}, \mathbf{p}_{y_k}))/\tau_c\}}, \quad (1)$$

where $\mathcal{Y}_M = \mathcal{Y}_s \cup \{y_M\}$. We empirically find that directly applying the augmentation at the beginning of training leads to under-fitting. To remedy this, we schedule the training procedure for COT and MAA. Specifically, we first train pure COT with a backbone network at 2/3 of the total epochs. Then we freeze the backbone and apply MAA to COT till the end of training.

C. More results

We provide more complementary results to validate CoT on MIT-States, CGQA and VAL-CZSL benchmarks.

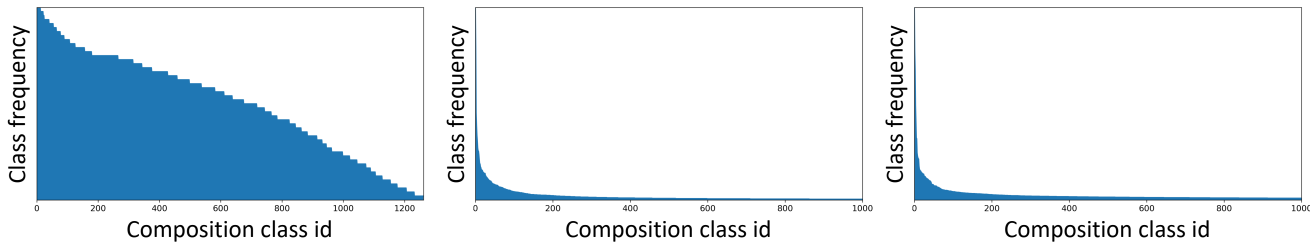
C.1. Hubness effect

We illustrate a distribution of k-occurrences (N_k) [7] to measure a hubness effect of visual features. Note that dif-

*Corresponding author

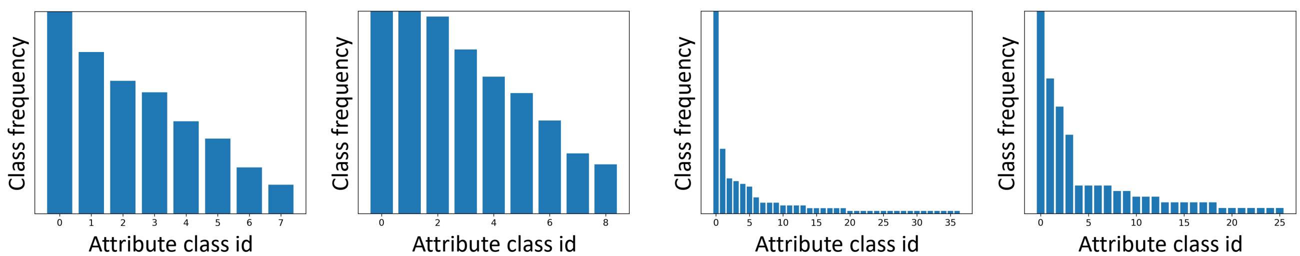
Dataset	#Attribute / #Object	Train		Validation		Test	
		$\mathcal{Y}_s$	#img	$\mathcal{Y}_s$ / $\mathcal{Y}_u$	#img	$\mathcal{Y}_s$ / $\mathcal{Y}_u$	#img
MIT-States [3]	115 / 245	1262	30338	300 / 300	10420	400 / 400	12995
C-GQA [6]	453 / 870	6963	26920	1173 / 1368	7280	1022 / 1047	5098
VAW-CZSL [8]	440 / 541	11175	72203	2121 / 2322	9524	2449 / 2470	10856

Table 1: The statistics of three benchmarks: MIT-States[3], C-GQA[6] and VAW-CZSL[8].



(a) MIT-States (b) C-GQA (c) VAW-CZSL

Figure 1: Composition class distribution on three datasets. x-axis (composition class id) is ordered by decreasing composition frequency. For better visualization, we plot the top 1000 frequent composition classes on (b) and (c).



(a) MIT-States (b) C-GQA

Figure 2: Attribute class distributions on MIT-States and C-GQA. Every attributes in each plot are composed into a fixed object; 'Silk', 'Bucket', 'Street', and 'Box'. (from left to right).

CoT	MAA	AUC	HM
		8.89	22.9
	✓	9.08	23.5
✓		10.26	25.2
✓	✓	**10.54**	**25.8**

(a) MIT-States

CoT	MAA	AUC	HM
		5.25	17.6
	✓	5.58	20.3
✓		7.07	21.2
✓	✓	**7.42**	**22.1**

(b) C-GQA

Table 2: Component analysis on MIT-States and C-GQA dataset.

Frequency type	AUC	HM
$1/(\zeta_{o_i})^{0.5}$	7.19	21.5
$1/(\zeta_{o_i})^{2}$	7.07	21.2
$1/(\zeta_{o_i})$ [default setting in paper]	**7.20**	**21.7**

Table 3: Ablation study for different frequency types of MAA on VAW-CZSL.

ferent from [1], we calculate the nearest neighbors among visual features (queries) to analyze a hubness problem on the **visual domain** in both Fig. 3 and Fig. 4 in the main paper. To be consistent, the proposed CoT and MAA significantly alleviate the hubness problem by enhancing visual discrimination.

C.2. Component analysis

In Table 2, we show the impact of each component (CoT and MAA) for CZSL performance on MIT-States and C-GQA datasets. The ablation results together with Table 3a in the main paper demonstrate that both CoT and MAA consistently give improvements in AUC and HM.

C.3. Other sampling weights

In Table 3, we conduct the ablation study for three different sampling weights leveraging an inverse attribute frequency $1/(\zeta_{o_i})$. Notably, the frequency of $1/(\zeta_{o_i})^2$ yields worse performance. It is under-fitting because sampling few tail class samples with too high probabilities prevents learning with other majority classes. Sampling with square-root frequency, $1/(\zeta_{o_i})^{0.5}$, improves the performance on AUC and HM, but slightly below the result with $1/(\zeta_{o_i})$. We will include the above discussion in the paper.

Methods	S	U	AUC	HM
CompCos [5]	25.4	<u>10.0</u>	8.9	1.6
CGE [6]	32.4	5.1	6.0	1.0
KG-SP [4]	28.4	7.5	7.4	1.3
Ours (CoT)	<u>28.8</u>	**11.3**	**9.5**	**1.8**

Table 4: Open-world CZSL results on MIT-States. All methods use Resnet18 [2] backbone with a fine-tunning setup.

C.4. Open World setting

To further analyze the generalization performance, we evaluate our CoT on Open World setting [5] in Table 4. Following [5], we compute the best seen (S), unseen (U) accuracies, area under curve (AUC) and the best harmonic mean (HM). Our method also performs well in Open world setting, outperforming previous state-of-the-art methods in all metrics except the best seen accuracy. This result demonstrates that enlarging visual discrimination with context modeling could also mitigate unfeasible compositions [5] from a large output space of Open World scenario.

C.5. Object-guided attention maps

We illustrate the object-guided attention maps in Fig. 4 with VAW-CZSL, and Fig. 5 with C-GQA. For visualization, we merge three attention maps from low, middle and high blocks through multiplication. The results demonstrate that the attention module could capture the contextualized regions for each attribute, enhancing the visual discrimination of attribute prototypes and its composition.

C.6. Top-3 prediction results

We visualize top-3 prediction results in Fig. 6 with VAW-CZSL, and Fig. 7 with C-GQA. CoT clearly outperforms the baseline [8] to retrieve the relevant composition labels. As discussed in Sec. 4.4 of the main paper, the qualitative results show the limitation of existing CZSL datasets [6, 8] including multi-label composition and multiple attribute-object interaction.

D. Computational Analysis

In Table 5, we compare computational complexity with the previous state-of-the-art OADis [8] by reporting the number of parameters and GFLOPs. Although CoT has more parameters induced from the ensemble of block features, it has almost the same model complexity in terms of GFLOPs, thanks to the parameter-efficient convolution based attention module.

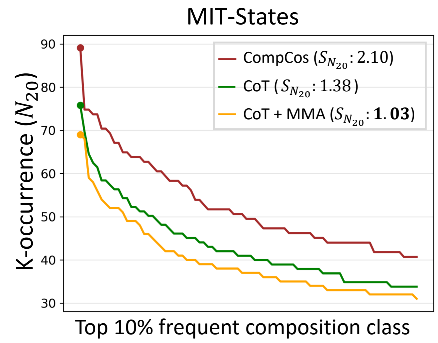


Figure 3: Distribution of k-occurence counts (N_k) of MIT-States test set. We use the same setting with Fig. 4 in main paper.

Methods	# Parameters (M)	GFLOPs
OADis[8]	2.25	11.2
CoT	3.34	11.8

Table 5: Comparison of computational complexity between OADis [8] and CoT. Note that backbone (ViT-B) is excluded to count the parameters.

References

[1] Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*, 2014. 2

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[3] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *CVPR*, 2015. 1, 2

[4] Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. Kg-sp: Knowledge guided simple primitives for open world compositional zero-shot learning. In *CVPR*, 2022. 3

[5] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Open world compositional zero-shot learning. In *CVPR*, 2021. 3

[6] Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning. In *CVPR*, 2021. 1, 2, 3

[7] Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovic. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(sept):2487–2531, 2010. 1

[8] Nirat Saini, Khoi Pham, and Abhinav Shrivastava. Disentangling visual embeddings for attributes and objects. In *CVPR*, 2022. 1, 2, 3, 5

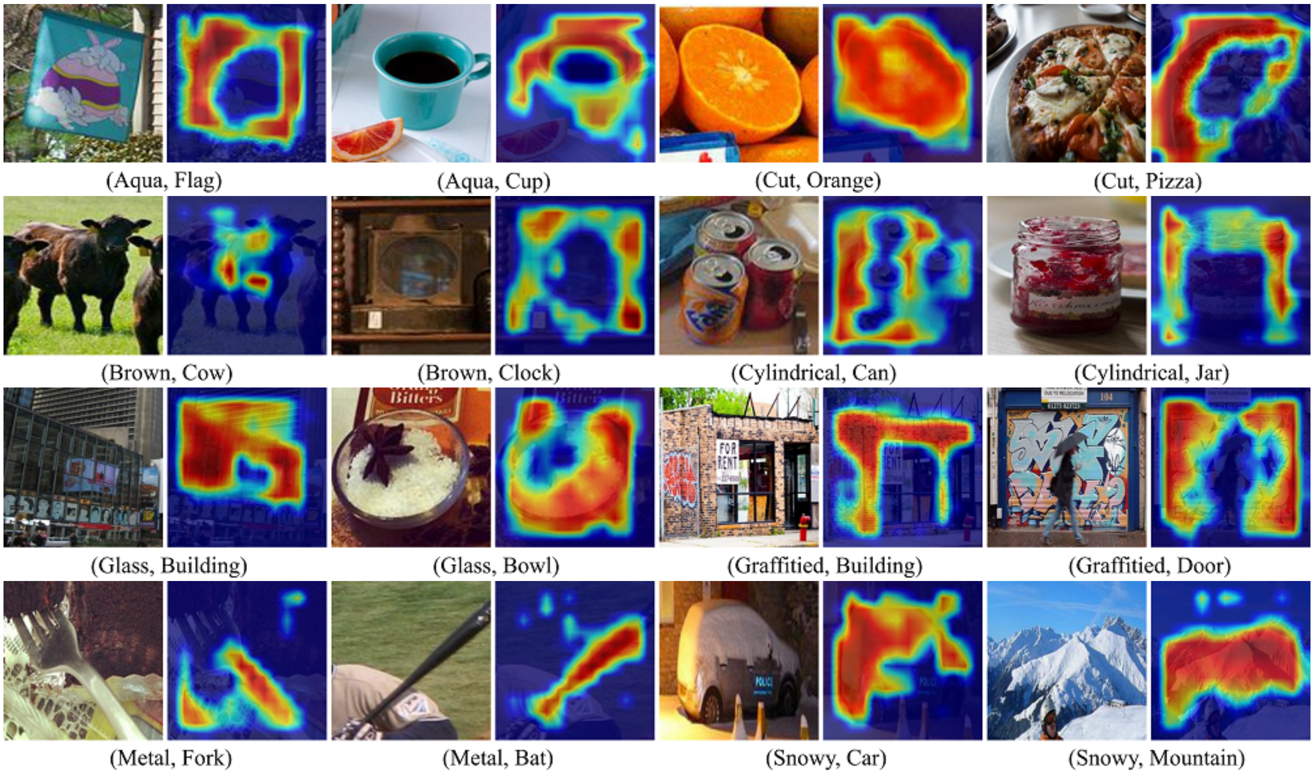


Figure 4: Visualization of object-guided attention maps obtained on **VAW-CZSL**. The input image and attended region by its specific attribute are paired with (attribute, object) labels. (Attention weights: High to Low).

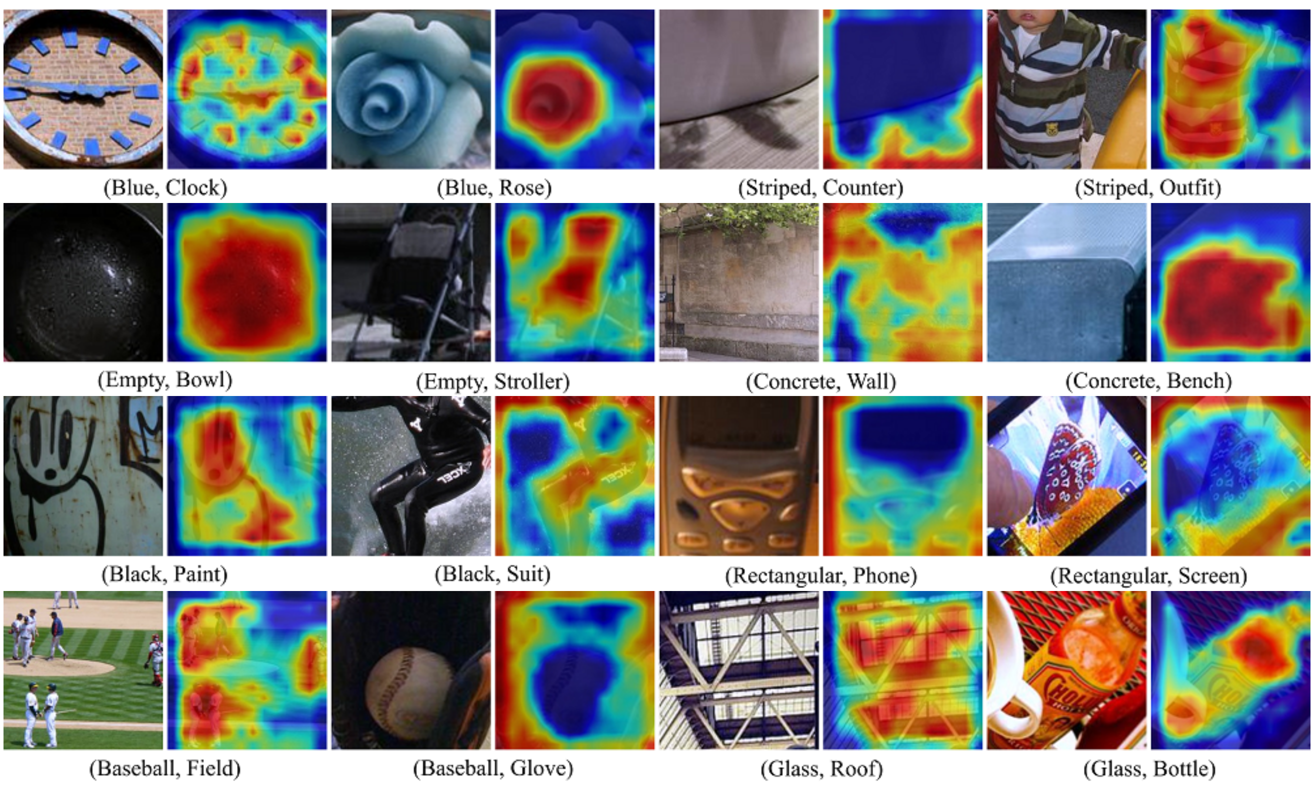


Figure 5: Visualization of object-guided attention maps on **C-GQA**. The input image and attended region by its specific attribute are paired with (attribute, object) labels. (Attention weights: High to Low).

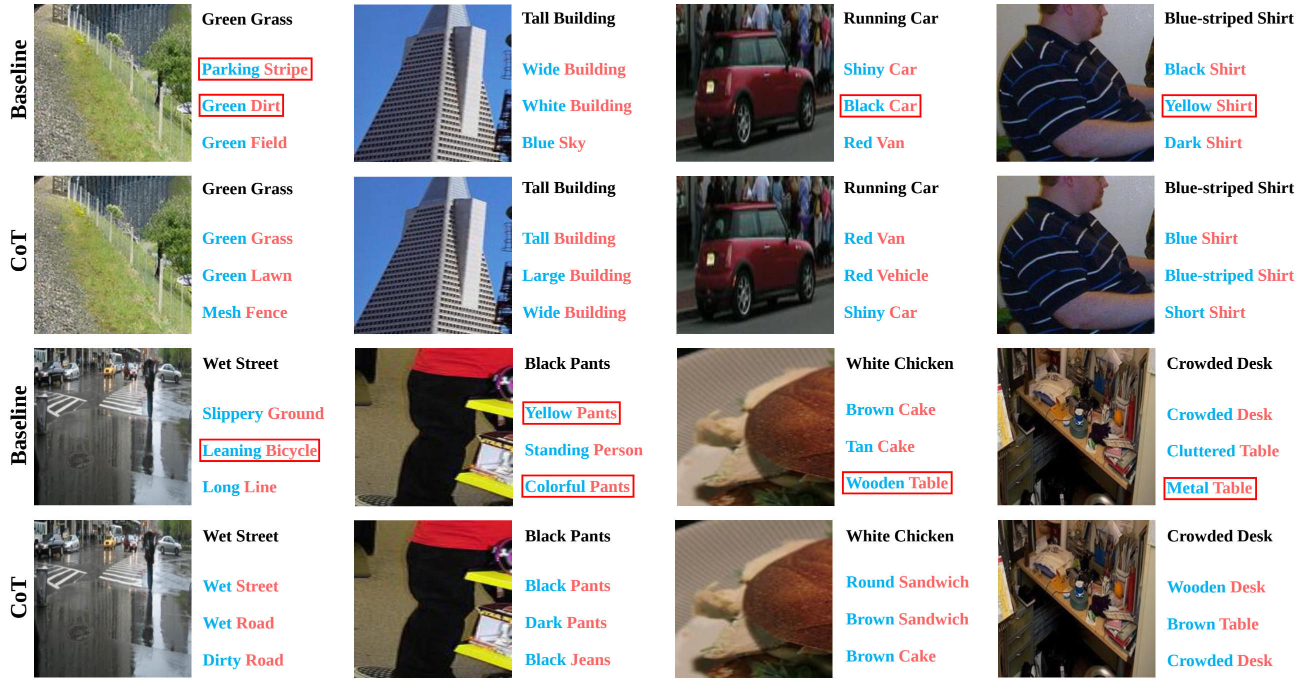


Figure 6: Ground-truth and top-3 prediction results on **VAW-CZSL**. We compare CoT with baseline (OADis) [8]. Red box denotes the false positive, having irrelevant or opposite compared to ground truth.

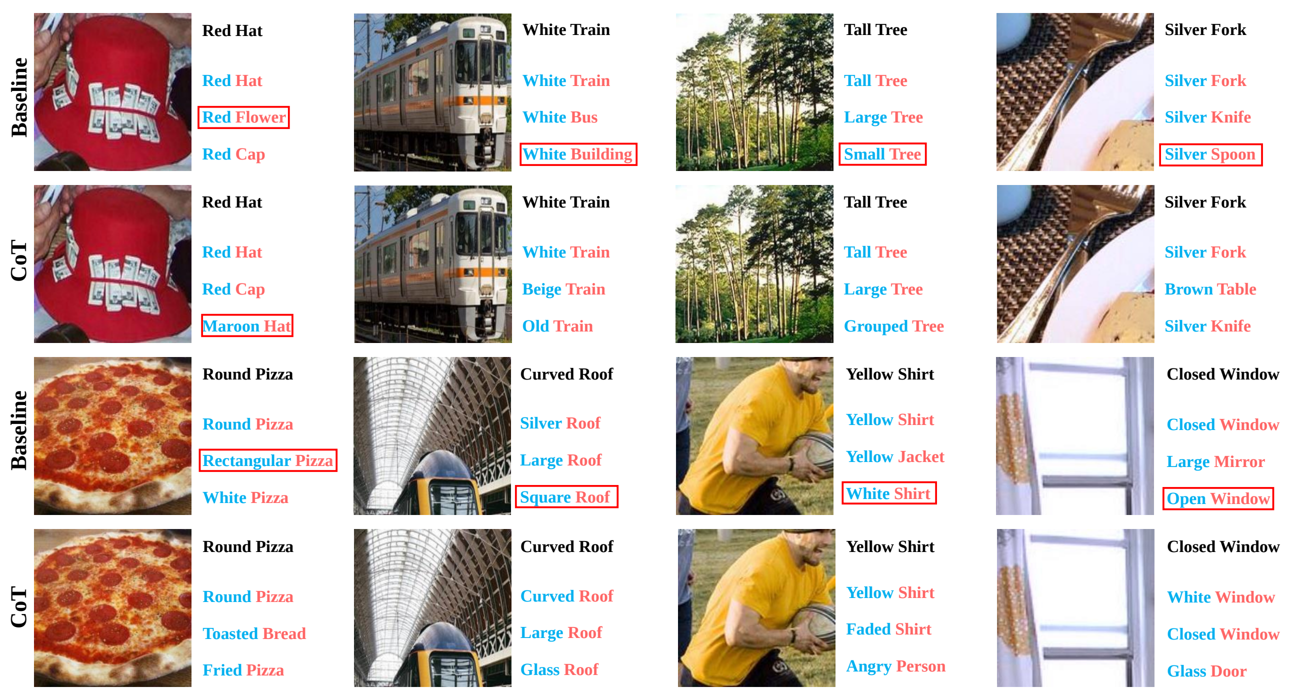


Figure 7: Ground-truth and top-3 prediction results on **C-GQA**. We compare CoT with baseline (OADis) [8]. Red box denotes the false positive, having irrelevant or opposite compared to ground truth.