

# Supplementary Material for Efficient Adaptive Human-Object Interaction Detection with Concept-guided Memory

Ting Lei<sup>1</sup> Fabian Caba<sup>2</sup> Qingchao Chen<sup>3</sup> Hailin Jin<sup>2</sup> Yuxin Peng<sup>1</sup> Yang Liu<sup>1\*</sup>

<sup>1</sup>Wangxuan Institute of Computer Technology, Peking University

<sup>2</sup>Adobe Research <sup>3</sup>National Institute of Health Data Science, Peking University

{ting\_lei, qingchao.chen, pengyuxin, yangliu}@pku.edu.cn

{caba, hljin}@adobe.com

This supplementary material offers further information on the implementation of ADA-CM and presents additional ablation studies. Firstly, we present additional implementation details regarding ADA-CM. Next, we conduct an empirical investigation to assess how different strategies for constructing concept-guided memory affect the performance of the proposed method. Finally, we perform a thorough ablation study, evaluating the effectiveness of various prior knowledge information used in the fine-tuning process.

## 1. Implementation Details

In this section, we present a comprehensive description of the implementation details of ADA-CM. We fine-tune the detector DETR prior to training and then freeze its weights. Specifically, for HICO-DET, we fine-tune DETR on HICO-DET with its weights initialized from the publicly available model pre-trained on MS COCO [2]. For V-COCO, we pre-train DETR from scratch on MS COCO, excluding those images in the test set of V-COCO. For interaction prediction, we first filter out detections whose score is lower than 0.2 and perform non-maximum suppression with a threshold of 0.5. Then, we reserve at least 3 and at most 15 boxes for humans and objects each for every image. We employ two ViT variants as our backbone architectures: ViT-B/16 and ViT-L/14, where "B" and "L" refer to base and large, respectively. ViT-B has 12 encoder layers and ViT-L has 24 encoder layers, both preceded by a single 2D convolutional layer. To boost the performance of human-object pairs including small objects [4], we leverage ViT-L/14-336px to extract high-resolution feature maps. The input resolution for ViT-B and ViT-L is 224 pixels and 336 pixels, respectively.  $\gamma_{IC}$ ,  $\gamma_{CA}$  and  $\gamma_T$  are set to be 0.5, 0.5 and 1.0, respectively. Both  $L_{IC}$  and  $L_{CA}$  represent multi-hot *interaction* labels, with dimensions of  $\mathbf{R}^{C \times N}$ , where  $C$  is the number of interaction categories and  $N$  is the mem-

ory size. They have identical content as they serve as the value (label) for the corresponding key (each memory slot in each branch is designed to store information about the same interactive human-object pair). The prior knowledge  $P \in \mathbf{R}^{d \times N}$  is unique to each image and contains information about all  $N$  detected instances in it. For the  $i$ -th instance, we combine its spatial configuration  $b_i$ , semantic information  $e_i$ , and confidence score  $c_i$  into a concatenated form  $p_i \in \mathbf{R}^d$ . Additionally, we apply the augmentation method described in DETR [1], resizing the input images such that the shortest side is at least 480 and at most 800 pixels while the longest side at most is 1333. We unfreeze the keys of the HOI cache memories, semantic embeddings, last projection layer and the positional embeddings of CLIP visual encoder and instance-aware adapter modules.  $\lambda$  is set to 1 during training and 2.8 during inference [5, 6]. We use AdamW [3] as the optimizer with an initial learning rate of 1e-3 and train ADA-CM for 15 epochs. The model is trained on a single NVIDIA A100 device with an efficient batch size of 8.

## 2. Ablation on Concept-guided Memory

This section presents an empirical investigation of how different concept-guided memory construction strategies affect the performance of the proposed method. Algorithm 1 outlines the sample selection strategy adopted in the memory construction procedure, where *memshot* is a hyperparameter that controls the number of samples in the memory per concept. The performance of using different concept sets, such as "Verb," "Object," and "HOI," is compared in Table 1. Selecting HOI as the concept type allows the model to explicitly choose samples for each HOI concept, resulting in the best performance, as shown in the first three rows in Table 1. Opting for "Verb" or "Object" as the concept type yields poor performance, as many HOI concepts may be overlooked.

Additionally, we empirically studied which distribution

\*Corresponding author

Concept Type	Sampling Choice	Memory Size	mAP	Rare	Non-rare
Verb	Uniform	8766	6.47	4.46	7.07
Object	Uniform	7807	6.63	7.24	6.45
HOI	Uniform	7690	<b>25.19</b>	<b>27.24</b>	<b>24.58</b>
HOI	Origin	7290	24.06	24.68	23.87
HOI	Origin	14608	24.46	24.88	24.33

Table 1. **Ablation on strategy of constructing concept-guided memory(Training-Free)**. "Uniform" refers to the method of sampling where we select samples from all available classes in equal proportions. "Origin" indicates that we obtain the samples following the original distribution of the dataset. Results are on HICO-DET.

---

**Algorithm 1** Strategy for Building Concept-guided Memory

---

- 1: Define a concept set  $C$ .
  - 2: Build a dictionary  $D$  with  $|C|$  categories.
  - 3: **for** every HOI sample  $(x, y)$  **do**
  - 4:   **if**  $y$  contains the concept  $C_i$  **then**
  - 5:     insert  $x$  into  $D[C_i]$
  - 6:   **end if**
  - 7: **end for**
  - 8: Denote  $memshot$  as the number of samples in the memory per concept.
  - 9: Build concept-guided memory  $M$ .
  - 10: **for** every concept  $C_i$  in  $C$  **do**
  - 11:   Allocate  $\min(memshot, |D[C_i]|)$  slots for  $C_i$  in concept-guided memory  $M$ .
  - 12:   **if**  $memshot \leq |D[C_i]|$  **then**
  - 13:     Select  $memshot$  samples from  $D[C_i]$  and transfer them to the allocated slots in  $M$ .
  - 14:   **else**
  - 15:     Transfer all samples in  $D[C_i]$  to the allocated slots in  $M$
  - 16:   **end if**
  - 17: **end for**
- 

e	c	b	Full	Rare	Non-rare
			32.81	31.80	33.11
✓			37.04	36.12	37.32
✓	✓		37.20	37.60	37.08
✓	✓	✓	<b>38.40</b>	<b>37.52</b>	<b>38.66</b>

Table 2. **Components of prior knowledge ablation (Fine-Tuning Setting, ViT-L backbone)**. This table studies the effect of different components of prior knowledge on the fine-tuning setting. e: semantic information, c: confidence score, b: spatial configuration. Results are on HICO-DET.

the HOI samples should be selected from. Following the original distribution results in a long-tailed distribution in the concept-guided memory, reducing the performance of rare classes, as shown in the last three rows in Table 1. Our sampling strategy, which uniformly selects samples for each

HOI class and uses HOI as the concept type, proves to be the most efficient. We adopt this design choice in all experiments presented in the submission, unless otherwise specified.

### 3. Components of Prior Knowledge

As stated in the method section, the prior knowledge consists of three components: spatial configuration, semantic information of extracted objects, and a confidence score. In practice, we formulate the above three components using the detected instances' bounding box coordinates, semantic embeddings, and logits, respectively. In this section, we verify the effectiveness of different types of prior knowledge. As shown in Table 2, simply adding semantic information boosts the performance by a margin of 4.23% mAP compared with the model tuned without prior knowledge. This demonstrates the effectiveness of semantic embeddings fused into the visual encoder through instance-adaptor. By further adding spatial configurations and confidence scores of detected instances, our model becomes more spatial-aware and thus achieves the best performance.

### 4. Qualitative Results

As shown in Figure 1, we present several qualitative results of successful HOI detections. For example, in Figure 1(a), our model detects and predicts the human carrying a skateboard with a high confidence score. We also observe that in Figure 1(e), where our model becomes much less confident when the detected human wearing white clothes is far away from the detected sports ball. To illustrate our model's limitations, we also visualize five failure cases as shown in Figure 2. In Figure 2(a), our model fails to predict  $\langle \text{human, throw, frisbee} \rangle$  triplet due to the similar poses of a human throwing a frisbee and a human catching a fresbee. As shown in Figure 2(e), our model may predict interactions between irrelevant human and objects when there're multiple humans and objects in an image.

### References

- [1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-

