

A. Datasets Details

Flickr30K [50] dataset contains 31,000 images collected from the Flickr website. These images mostly depict humans performing various activities. Each image is described by five different sentences, and there are 155,000 sentences. Following the settings [2, 38], this dataset is split into 29,783 training images, 1,000 validation images and 1,000 testing images.

COCO [4] dataset is a challenging large-scale dataset containing 123,287 images. These images are collected by searching 80 object categories and 40 scene types from the Flickr website. Each image is paired with five sentences, resulting in a total of 616,435 sentences. We follow the dataset split in works [2, 38], namely, 113, 287 images for training, 5,000 images for validation, and 5,000 images for testing.

B. Implementation Details

In all experiments, we use GPT-3.5 as the LLM and CLIP as the pre-trained model. We use three popular pre-trained backbones of CLIP: ViT-B/32, ViT-B/16, and ViT-L/14-336. During training, we use SGD optimization with an initial learning rate of $1e-5$, a maximum of 4 epochs, and a batch size of 128.

C. More Ablation Studies

Effect of action-aware multi-modal prompting. To evaluate the effectiveness of action-aware multi-modal prompting, we design several variants of our method for comparison. From the results in Table 6, we observe that our method outperforms the other variants on the Flickr30K dataset. These results highlight the importance of both the action triplet prompt and the action state prompt in enhancing matching performance, as each captures fine-grained action semantics. Moreover, our method incorporates the visual prompt to enhance the model’s visual perception.

Action Triplet	Hand-Craft Triplet	Action State	Vis	Image-to-Text			Text-to-Image		
				R@1	R@5	R@10	R@1	R@5	R@10
✓				86.1	97.6	98.6	73.4	92.7	95.4
	✓			85.3	97.2	98.2	72.7	92.5	95.1
		✓		86.7	97.2	98.6	74.0	92.9	95.8
✓		✓		87.5	97.8	98.8	74.3	92.9	95.6
✓		✓	✓	88.1	98.0	99.2	74.7	93.1	95.8

Table 6. Ablation study over types of prompts on Flickr30K 1K test set.

Effect of action knowledge. To evaluate the effectiveness of incorporating action knowledge in prompting the pre-trained CLIP, we replace the action-aware multi-modal prompts with learnable visual prompts, denoted as “w/o action knowledge”. As shown in Table 7, the performance of our method with only visual prompt tuning drops significantly, indicating that incorporating prompts sourced from action-related external knowledge is crucial for enhancing

the fine-grained visual perception ability of the pre-trained model.

Method	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
Baseline	81.2	96.4	98.5	62.2	85.7	91.8
w/o action knowledge	86.8	97.2	98.9	69.7	90.3	93.1
replace AIM with CAT	87.2	97.9	99.1	72.9	91.2	95.6
replace PA with FC	87.9	98.0	99.1	73.2	92.0	95.9
Ours	88.1	98.0	99.2	74.7	93.1	95.8

Table 7. Ablation analysis of different components on Flickr30K 1K test set.

Effect of action-aware adaptive interaction. In Table 7, we evaluate the effect of replacing the action-aware adaptive interaction module (denoted as “AIM”) with a concatenation operation (denoted as “CAT”) on the Flickr30K dataset. We observe a decline in performance, indicating that AIM helps the model mitigate the disturbance caused by irrelevant or noisy information retained in prompts.

Effect of prompt adapter. To evaluate the effectiveness of the prompt adapter (denoted as “PA”) in adapting the prompted knowledge, we replace it with a simple fully connected layer (denoted as “FC”). As shown in Table 7, the prompt adapter performs best. This is reasonable since prompt tuning and adapter play different roles in improving performance, in which the prompt adapter further bridges the feature gap between the pre-trained model and the downstream task.

Effect of Hyper-parameter λ . We evaluate the impact of the trade-off hyper-parameter λ in Eq. 9 by varying its values among 0.1, 0.3, 0.5, 0.7, and 0.9. As depicted in Figure. 6, our method achieves the best performance when λ is set to 0.7 on both datasets.

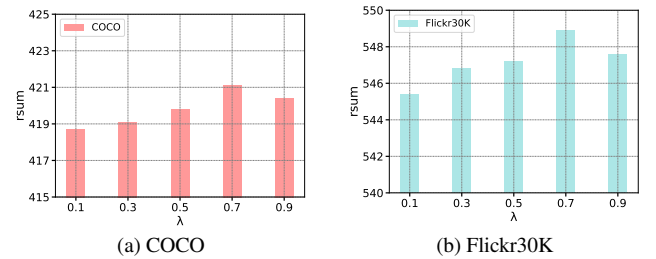


Figure 6. Comparison results of different trade-off hyper-parameters on COCO 5K test set and Flickr30K 1K test set.

D. More Qualitative Results

Qualitative results of the ablation study. In Figure 7, we show a typical example of our method and its variants, including “w/o action triplets”, “w/ action triplets”, both “w/ action triplets” and “w/ action state descriptions”, and “w/

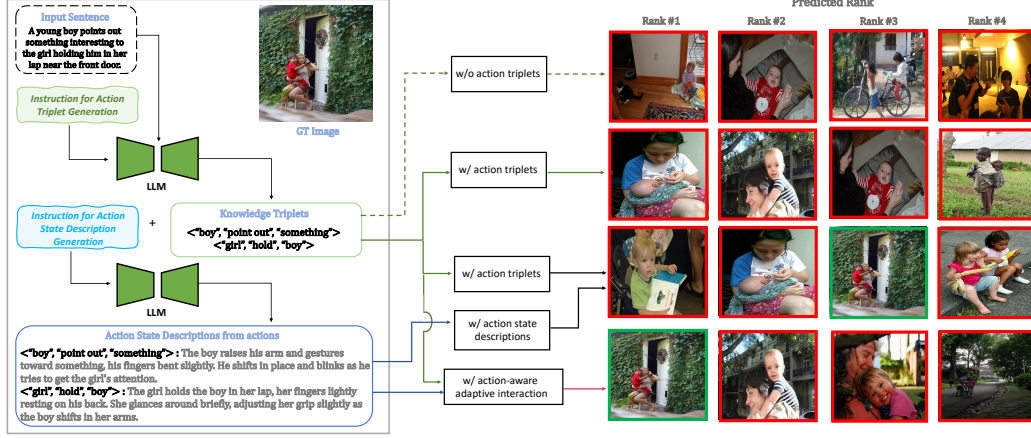


Figure 7. An example from the ablation study in text-to-image retrieval, comparing various variants (*i.e.*, “w/o action triplets,” “w/ action triplets,” “w/ action triplets and action state descriptions,” and “w/ action-aware adaptive interaction”).

















action-aware adaptive interaction”.

As shown in Figure 7, compared with “w/o action triplet”, “w/ action triplet” retrieves images that depict the action of “holding”, which partially aligns with the intent of the query. Moreover, jointly using both “w/ action triplets” and “w/ action state descriptions” retrieves candidate images that align more closely with the query intent than “w/ action triplet”. This is because each action state provides more details and comprehensive descriptions, supplementing the information contained in action triplets “<boy, point out, something >” and “<girl, hold, boy >”. However, the action knowledge of “<boy, point out, something >” generated by LLM contains some irrelevant action information, which misleads CLIP’s fine-grained action-aware visual understanding. “w/ action-aware adaptive interaction” helps mitigate the interference from action-irrelevant contents, enabling the model to retrieve the corresponding ground truth image.












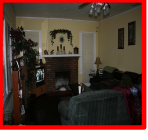




Qualitative results compared with CLIP. To better understand the effectiveness of our method, we visualize some examples of text-to-image retrieval results on Flickr30K and COCO datasets, as shown in Figure 8. For each text query, the top-4 ranked images from our method and baseline model CLIP are listed. The ground truth images are outlined in green boxes while incorrect ones are in red boxes. It can be seen that our method is more robust in complex scenes compared to CLIP, achieving better retrieval results. For example, in the two examples of Figure 8 (a), our method successfully retrieves the ground truth image with the precise description of “playing” and “catch”.

In the first example of Figure 8 (b), CLIP ranks the top-3 images incorrectly, as they do not depict the action of “sitting” from the text query. In contrast, our method can enhance the ground truth image with precise state descriptions of both “flying” and “sitting”. However, in the second example of Figure 8 (b), our method ranks incorrectly for a

given text query, since the action knowledge generated by the LLM for “<woman, use, video game controls >” fails to precisely describe the action state of “use” in the original sentence context, resulting in some noise information is still retained in action-aware prompts and leading to an incorrect retrieval. Thus, using a more advanced LLM in future work may help correct this error in some content.

Query sentence: Two women on the street, one is playing the guitar and the other is playing violin.			Action Knowledge
Retrieved Images	Ours	   	<p>Action Triplets</p> <ul style="list-style-type: none"> ■ <woman, play, guitar> ■ <woman, play, violin> <p>Action State Descriptions</p> <ul style="list-style-type: none"> ■ <woman, play, guitar>: Her right hand strums the guitar strings with controlled movements, while her left hand presses down on the frets, smoothly transitioning between chords. ■ <woman, play, violin>: Her left hand moves between the strings, adjusting finger positions, as her right hand pulls the bow over consistent pressure.
	CLIP	   	
Query sentence: A boy in a red uniform is attempting to avoid getting out at home plate, while the catcher in the blue uniform is attempting to catch him.			Action Knowledge
Retrieved Images	Ours	   	<p>Action Triplets</p> <ul style="list-style-type: none"> ■ <boy, avoid, tag> ■ <catcher, catch, boy> <p>Action State Descriptions</p> <ul style="list-style-type: none"> ■ <boy, avoid, tag>: The boy adjusts his speed and angle, keeping a close watch on the catcher's glove as he attempts to stay out of reach and avoid the tag at home plate. ■ <catcher, catch, boy>: The catcher shifts his weight, swiftly lowering his stance as he stretches out his glove toward the boy, attempting to catch him just inches from home plate.
	CLIP	   	

(a) Examples on the Flickr30K dataset.

Query sentence: A person that is flying a kite that is sitting on the ground.			Action Knowledge
Retrieved Images	Ours	   	<p>Action Triplets</p> <ul style="list-style-type: none"> ■ <person, sit, ground> ■ <person, fly, kite> <p>Action State Descriptions</p> <ul style="list-style-type: none"> ■ <person, sit, ground>: The person sits cross-legged on the ground, one hand resting on their knee while the other is placed beside them for support, with their back upright. ■ <person, fly, kite>: The person sits on the ground, holding the kite string in one hand while gently pulling to lift the kite, which remains grounded, as they prepare to catch a favorable wind.
	CLIP	   	
Query sentence: Woman standing in living room using video game controls.			Action Knowledge
Retrieved Images	Ours	   	<p>Action Triplets</p> <ul style="list-style-type: none"> ■ <woman, stand, living room> ■ <woman, use, video game controls> <p>Action State Descriptions</p> <ul style="list-style-type: none"> ■ <woman, stand, living room>: She stands with her feet slightly apart, her arms relaxed at her sides, maintaining an upright and balanced posture. ■ <woman, use, video game controls>: She holds the video game controls, her hands shifting from the usual buttons to other parts of the controller, interrupting the usual sequence.
	CLIP	   	

(b) Examples on the COCO dataset.

Figure 8. Visual comparisons of text-to-image retrieval examples between our method and baseline CLIP on Flickr30K and COCO datasets. The ground-truth images are outlined in green boxes, and the incorrect ones are outlined in red boxes.