This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

LLM-enhanced Action-aware Multi-modal Prompt Tuning for Image-Text Matching

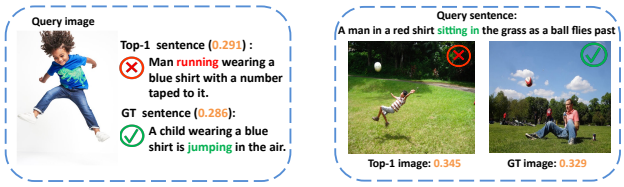
Mengxiao Tian[1,2,3], Xinxiao Wu[1,2], Shuo Yang[2] *
[1] Beijing Key Laboratory of Intelligent Information Technology,
School of Computer Science & Technology, Beijing Institute of Technology, China
[2] Guangdong Laboratory of Machine Perception and Intelligent Computing,
Shenzhen MSU-BIT University, China
[3] Beijing Research Center of Intelligent Equipment for Agriculture, China

Abstract

Driven by large-scale contrastive vision-language pre-trained models such as CLIP, recent advancements in the image-text matching task have achieved remarkable success in representation learning. Due to image-level visual-language alignment, CLIP falls short in understanding fine-grained details such as object attributes and spatial relationships between objects. Recent efforts have attempted to compel CLIP to acquire structured visual representations by introducing prompt learning to achieve object-level alignment. While achieving promising results, they still lack the capability to perceive actions, which are crucial for describing the states or relationships between objects. Therefore, we propose to endow CLIP with fine-grained action-level understanding by introducing an LLM-enhanced action-aware multi-modal prompt-tuning method, incorporating the action-related external knowledge generated by large language models (LLMs). Specifically, we design an action triplet prompt and an action state prompt to exploit compositional semantic knowledge and state-related causal knowledge implicitly stored in LLMs. Subsequently, we propose an adaptive interaction module to aggregate attentive visual features conditioned on action-aware prompted knowledge for establishing discriminative and action-aware visual representations, which further improves the performance. Comprehensive experimental results on two benchmark datasets demonstrate the effectiveness of our method. Codes are at https://github.com/Mengxiao-Tian/LAMP.

1. Introduction

Image-text matching has gained much attention in recent years, aiming to match images with the most relevant texts and vice versa. This is a fundamental task in vision and language research, involving multimodal reasoning and alignment of visual and textual concepts at different levels, such as cross-modal retrieval [11, 30], natural language visual reasoning [3, 14], and visual question answering [13, 33]. Recently, large-scale pre-trained models such as CLIP [32] have shown superior generalization and transferability on various downstream tasks [12, 41, 54], making CLIP one of the most widely used pre-trained models for image-text matching.

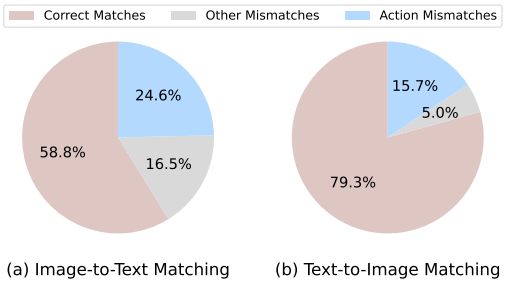


(a) Image-to-Text Matching (b) Text-to-Image Matching

Figure 1. Failure cases of image-text matching (image-to-text and text-to-image) using CLIP on the COCO test set. Similarity scores are shown in orange, and correct matches are marked in ✓ and mismatches in ✗. (a) CLIP incorrectly retrieves the top-1 text for a given image query; (b) CLIP incorrectly retrieves the top-1 image for a given text query.

However, a large number of studies have demonstrated that discarding fine-grained visual information results in subpar performance on downstream tasks involving localization [53], counting [29], and understanding relationships between objects or object attributes [15, 54]. Recent studies [17, 25, 40] focus on object-level understanding by introducing learnable prompts that describe more details of image regions, *i.e.*, objects or object attributes, enabling CLIP to learn fine-grained representations, thereby enhancing fine-grained understanding. Despite the impressive performance, whether CLIP can effectively perceive actions (*i.e.*, states or relationships between objects) in the image-

*Corresponding author



(a) Image-to-Text Matching (b) Text-to-Image Matching

Figure 2. The statistical analysis of inconsistent actions between the query and the candidate in image-text matching using CLIP on the Flickr30K test set.

text matching task remains unresolved.

As shown in Figure 1, the CLIP score of the ground-truth image-text pair is lower than that of a mismatched pair. We further perform a statistical analysis of the action inconsistency between the query and the candidate, as shown in Figure 2. These results indicate that CLIP exhibits inadequacies in accurately perceiving actions. Therefore, we ask: *How to endow CLIP with the fine-grained and action-aware perception to improve its image-text matching capability?* This paper aims to approach this from both textual and visual perspectives.

From the textual perspective, we consider the compositionality and causality of actions. We assume that each action has an initiator and a recipient, and there is a logical relationship between the two regarding the specific action. Therefore, we propose to leverage the world knowledge about actions in LLMs to decompose the original textual descriptions into all possible action triplets in the format of <subject, action, object>. Each triplet captures a pair of potential entities with their action interrelations. For example, as shown in Figure 1, the original description "A man wearing a blue shirt is jumping in the air" is decomposed into "<man, jumping, air>". Furthermore, we believe that each action has its state, called action state, which can be caused or inferred by specific entities. Each action state provides more details and comprehensive descriptions to supplement the information contained in the action triplet. Thus, for each action triplet, we feed the hand-crafted instructions into LLMs to generate state descriptions about the action in the form of language prompts. For example, the action state "lifting both feet off the ground and propelling the body upwards" corresponds to the action triplet "<man, jumping, air>".

From the visual perspective, we also need to consider how to embed the action knowledge obtained from the textual descriptions into model training, that is, how to use action triplets and action states to guide the image-text training of CLIP. To this end, we design action-aware prompts based on different action information during training: action triplet prompts and action state prompts, which will be fed into the image encoder of CLIP to enable its visual representation to better learn action information at different levels. The action triplet prompts guide the image encoder to capture the rich compositional semantics existing in the action relationships between entities, while the action state prompts guide the image encoder to pay more attention to the resulting states of action in the visual content. Furthermore, to reduce the disturbance of irrelevant or noisy action information in the action-aware prompts, we propose an adaptive interaction module, which adaptively attends to the salient action cues relevant to the visual content, facilitating fine-grained alignment with the corresponding textual representations.

The main contributions are summarized as follows: (1) To the best of our knowledge, we are the first to enhance CLIP with fine-grained action-aware perception, and propose an LLM-enhanced action-aware multimodal prompt tuning method that incorporates external action knowledge from LLMs to improve the image-text matching performance. (2) We design two complementary action-aware prompts that use fine-grained action knowledge from compositionality and causality perspectives, and introduce an action-aware adaptive interaction module to aggregate attentive visual features based on action knowledge. (3) Extensive experiments on public datasets show that our method achieves significant performance improvements over existing methods.

2. Related Work

Image-Text Matching. Image-text matching has attracted much attention in recent years, and most existing methods attempt to embed images and texts into a common space to measure their semantic similarity. These methods can be categorized into coarse-grained and fine-grained matching. The coarse-grained matching methods [8, 23, 51] utilize separate encoders to extract global features from images and texts, and then project them into a common embedding space. The fine-grained matching methods [11, 19, 20, 46] focus on learning fine-grained alignments between image regions and textual words. SCAN [19] proposes a stacked cross-attention network to explore fine-grained correspondences between image regions and words, inspiring a range of various sophisticated methods [5, 30, 31]. Recently, many works [7, 9, 42] focus on employing GCN-based networks to learn local region-word alignment, achieving significant performance improvements. Despite the significant progress achieved by exploring diverse fine-grained interaction patterns, these fine-grained methods still heavily rely on expert experience to handcraft networks.

The most relevant method to our work is AME [20], which enhances action-aware representations using the action-similar texts from the memory bank. In contrast, we exploit the powerful in-context learning capability of LLMs to generate fine-grained action knowledge instead of relying

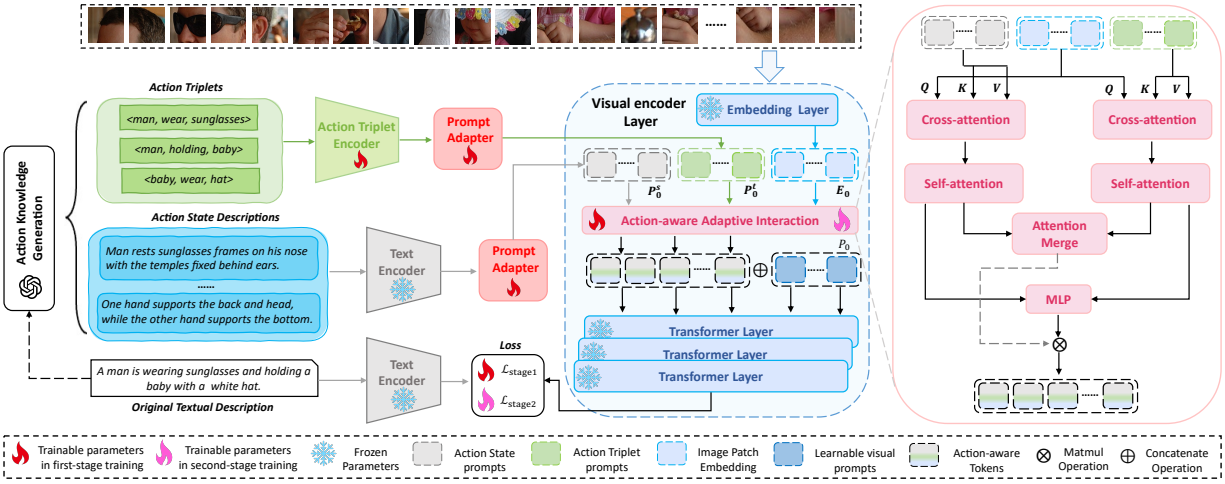


Figure 3. Overview of the proposed method.

solely on fixed original textual descriptions, thus enabling the model to achieve more fine-grained action-aware visual understanding.

Prompt Learning for CLIP. Various methods for prompt learning [18, 39, 40] have emerged as an effective and parameter-efficient approach to enhance the adaptation of CLIP to the downstream tasks, such as few/zero classification [22, 52], object detection [21, 26], and video understanding [39, 41].

A method closely related to ours is actionCLIP [41], which leverages CLIP's joint visual-textual embedding space for action recognition. However, it only uses action names as text prompts with prefixes and suffixes that carry limited semantic meanings. In contrast, we adopt a systematic approach that uses LLM as a knowledge engine to generate knowledge-rich prompts tailored to actions. These prompts incorporate fine-grained knowledge from compositionality and causality perspectives to provide fine-grained guidance for representation learning. In addition, we introduce an additional module to maximize the capture of action-related information.

External Knowledge from LLMs. Recently, the integration of task-relevant external knowledge from LLMs into prompts has been explored to enhance the discriminative ability of the prompt tuning process. KgCoOp [48] constrains the learnable prompt embeddings with general knowledge to improve generalization. Kan *et al.* [17] developed knowledge-aware prompts to assist the model in identifying the correspondences between images and different categories. Wang *et al.* [45] proposed to generate diverse forms of linguistic knowledge and conduct hierarchical prompt tuning for better performance. Menon *et al.* [27] proposed to classify images by querying VLMs with descriptive features extracted from LLMs for interpretable and adaptable recognition. Su *et al.* [34] leveraged LLMs to extract causal commonsense knowledge for reasoning about only the "why" questions in video understanding. Furthermore, Tan *et al.* [36] proposed category-wise and content-wise guidance for prompt optimization, helping the model understand general category information while capturing intraclass variation. Buettner *et al.* [1] distinguished visual descriptions of the same object class across different geographical regions by using LLMs. Niu *et al.* [28] proposed a state-aware reasoning framework with LLM-generated knowledge for dynamic procedure planning in instructional videos.

The above LLMs-based methods typically focus on general descriptions of entities and attributes in the text space. In contrast, we propose leveraging external action knowledge from LLMs to concentrate on the action states and relationships between objects. This knowledge can then be integrated into the prompts learning to enhance CLIP's fine-grained visual perception of actions.

3. Our Method

3.1. Overview

In this paper, we propose an LLM-enhanced action-aware multi-modal prompt-tuning method, which incorporates external action knowledge from LLMs to enhance action-aware perception, with a two-stage training strategy. As illustrated in Figure 7, starting with the original text description, the action knowledge generation module extracts and generates two kinds of knowledge: action triplets that outline the relationships between entities, and action state descriptions that detail the states of entities related to the action. These forms of knowledge serve as prompts to enhance the action-aware perception abilities of CLIP. Furthermore, these prompts are integrated into the proposed action-aware adaptive interaction modules together with visual tokens to effectively

capture various levels of visual information related to the action.

3.2. Generation of Action Knowledge

Figure 4 shows the generation of action knowledge by using the powerful in-context learning capability of LLMs. Our prompt template is generated using GPT-3.5, and minor adjustments do not affect the output. This knowledge includes not only fine-grained action relationships but also precise descriptions of their action states. Specifically, we design a prompt to instruct GPT-3.5 to construct action triplets, denoted as $\mathcal{R}_t$, and we further feed them with another instruction to generate knowledge-rich and fine-grained action state descriptions D about actions in the form of language, serving as valuable informative supplements for $\mathcal{R}_t$. A simple format check is performed afterward to ensure the output meets the desired specifications. It is important to note that the prompts can vary depending on datasets to accommodate different distributions of data.

Action Triplet Generation. We use one of the most powerful LLMs, GPT-3.5 to generate action triplet. Specifically, we first prepare several in-context learning examples of possible inputs along with their expected output triplets for GPT-3.5 to perform in-context learning. Each in-context sample consists of a textual description and corresponding action triplets. By combining multiple in-context samples, the LLM is constrained to generate higher-quality outputs while reducing irrelevant content. Subsequently, we use a well-designed instruction template Q_1 with in-context samples to instruct GPT-3.5 to parse an input caption T into decoupled textual action triplets, denoted as $\mathcal{R}_t$, which include entities and their interrelated action relationships, formulated as

$$\mathcal{R}_t = \text{LLM}([Q_1, T]). \tag{1}$$

Here $\mathcal{R}_t = \{r_i\}_{i=1}^{N}$, and r_i is represented as

$$r_i = \left\{(e_i, r(e_i, e_j), e_j) \mid 1 \le i, j \le N\right\}, \tag{2}$$

where N denotes the valid number of entity pairs in T, and $r(e_i, e_j)$ is the action relationships between entity pair (e_i, e_j). Notably, our action triplets can contain entities along with their fine-grained attributes (*e.g.,* "(girl, is, small)"), since actions inherently depend on both motion dynamics and static appearance. This ensures a comprehensive, contextually nuanced representation while maintaining focus on dynamics.

Action State Description Generation. Relying solely on action triplets may not be sufficient to capture the factual details of actions due to their inherent ambiguity and variability across different scenarios. Therefore, we consider the state of each action, referred to as action states, which can be caused or inferred by specific actions. Each action state contains more detailed and comprehensive descriptions, enabling the model to gain deeper insights into the fine-grained action semantics. As shown in Figure 4, we adopt a refined language instruction Q_2 with in-context samples to query GPT-3.5 to generate state descriptions about actions in the form of language, denoted as D, where the in-context samples consist of the action triplets, the original textual descriptions, and the expected output descriptions, formulated as

$$D = \text{LLM}(Q_2, \mathcal{R}_t). \tag{3}$$

For example, when querying the GPT-3.5 about *"(girl, laptop, study)"*, GPT-3.5 would generate an answer like *"She focuses on the laptop screen, types on the keyboard, and maintains a stable posture"*.

3.3. Action-aware Multi-modal Prompt Tuning

To address this limitation of pre-trained models in action perception, we propose an action-aware multi-modal prompt tuning. Specifically, we design two complementary action-aware prompts based on different action information: action triplet prompts and action state prompts, which will be fed into the image encoder to help CLIP's visual representation better capture different levels of action information. For the early layers of the image encoder, we map the generated prompted action knowledge into visual prompts by employing a lightweight action-aware prompt adapter, meaning that both types of action-aware prompts guided the encoding process of images. Moreover, for the later layers of the image encoder, each layer utilizes an independent prompt, allowing the layer to independently capture distinct visual and semantic features of the image. To further improve fine-grained visual perception, we propose an action-aware adaptive interaction module to adaptively capture the salient action cues relevant to the visual contents.

Action Triplet Prompts. To make action triplet prompts carry sufficient compositional semantics present in actions, we design an action-aware triplet encoder with transformer layers to explicitly model the action relations between paired objects. Specifically, for each action triplet $(e_i, r(e_i, e_j), e_j)$ in $\mathcal{R}_t$, we generate the corresponding word embedding W_y by employing a CLIP-based tokenizer together with a vocabulary embedding, formulated as

$$\mathbf{W}_y = \text{WordEmb}_{\text{CLIP}}(y), \quad y \in \{e_i, r(e_j), e_j\}. \tag{4}$$

Subsequently, we concatenate these embeddings to form a unified representation for each triplet, denoted as e_{Δ_i}. This process involves handling a total of K action triplets to produce K semantic embeddings, which are subsequently fed into multiple transformer layers to generate continuous prompt embedding features $\mathbf{p}_t$, formulated as

$$\mathbf{p}_t = \text{Multi-Transformer}([e_{\Delta_1}, e_{\Delta_2}, .., e_{\Delta_K}]). \tag{5}$$

Action State Prompts. Action state prompts can be constructed by plugging the generated action state descriptions

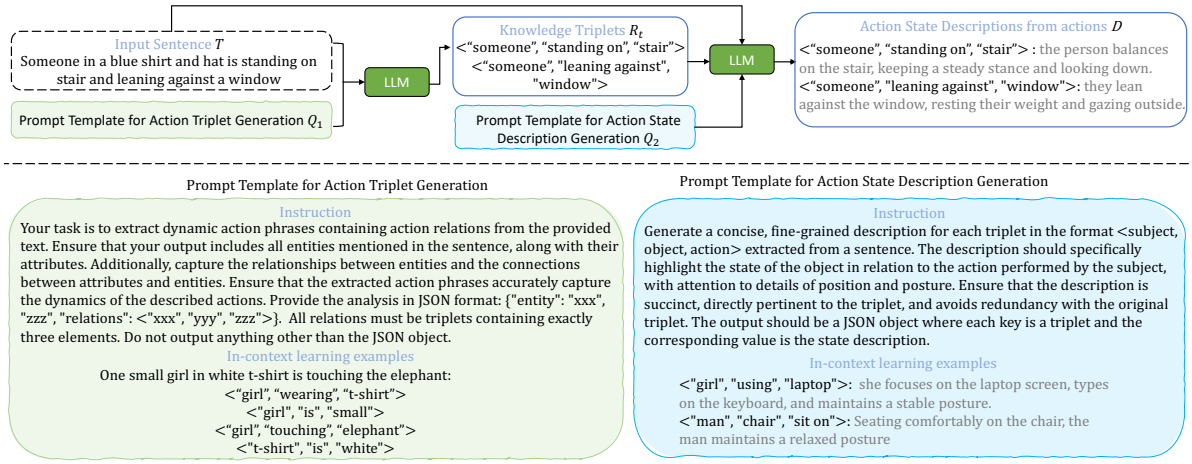


Figure 4. Action knowledge generation using LLM.

from actions in a predefined prompt template like *"Fine-grained state description of action triplet <subject, action, object>is [description]"*. These prompts guide the image encoder to focus on the resulting states of actions in visual content. We use a frozen text encoder to encode these state prompts into K action state prompt embedding features, represented as $\mathbf{p}_s = [\mathbf{p}_{s_1}, \ldots, \mathbf{p}_{s_K}]^\top \in \mathbb{R}^{K\times d}$, where d is the hidden dimension of CLIP.

Prompt Adapter. We feed different types of action-aware prompts $\mathbf{p}_t$ and $\mathbf{p}_s$ into a prompt adapter, which consists of two Multilayer Perceptron (MLP) layers to generate prompt embedding features. Between these layers, a ReLU activation and layer-normalization are applied to generate the action-aware visual prompts:

$$\begin{aligned}\mathbf{p}_0^t &= \text{ReLU}\left(\text{LN}\left(\mathbf{p}_t\right)\cdot\mathbf{W}_{\text{down}}\right)\cdot\mathbf{W}_{\text{up}},\\ \mathbf{p}_0^s &= \text{ReLU}\left(\text{LN}\left(\mathbf{p}_s\right)\cdot\mathbf{W}_{\text{down}}\right)\cdot\mathbf{W}_{\text{up}},\end{aligned} \tag{6}$$

where the feature dimension is first scaled from d to t ($t \ll d$) by the learnable matrix $\mathbf{W}_{\text{down}} \in \mathbb{R}^{d\times t}$ and then expanded back to d by $\mathbf{W}_{\text{up}} \in \mathbb{R}^{t\times d}$.

Action-aware Adaptive Interaction Module. To reduce the disturbance caused by irrelevant or noise information in the prompts, we construct an action-aware adaptive interaction module to adaptively tune the visual features by focusing on key information relevant to actions, thereby allowing the prompted image features to better align with the corresponding textual features, as shown in the right part of Figure 7.

Specifically, we first divide the input image $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$ into a sequence of M image patches, denoted as $\mathbf{X} = \{\mathbf{x}_p^1, \mathbf{x}_p^2, \ldots, \mathbf{x}_p^M\}$, $\mathbf{X} \in \mathbb{R}^{M\times d_v}$. Subsequently, each patch is transformed into an embedding vector and augmented with a learnable position embedding. The pre-processed image patch embedding for the transformer input is represented as $\mathbf{E}_0 = [\mathbf{e}_0^1, \mathbf{e}_0^2, \ldots, \mathbf{e}_0^M]$, where $\mathbf{e}_0^i \in \mathbb{R}^{C_e}$ is the patch embedding corresponding to position i, with C_e being the number of channels for each embedding. We take pre-processed image patch embedding $\mathbf{E}_0$, action triplet prompt embedding features $\mathbf{p}_0^t$ and action state prompt embedding features $\mathbf{p}_0^s$ as query, key, and value, formulated as

$$\begin{aligned}\tilde{\mathbf{p}}_0^t &= \text{Cross-Attn}(\mathbf{p}_0^t, \mathbf{E}_0), \tilde{\mathbf{p}}_0^s = \text{Cross-Attn}(\mathbf{p}_0^s, \mathbf{E}_0),\\ \mathbf{V}_{cs}^t, \mathbf{A}_{cs}^t &= \text{Self-Attn}(\tilde{\mathbf{p}}_0^t), \mathbf{V}_{cs}^s, \mathbf{A}_{cs}^s = \text{Self-Attn}(\tilde{\mathbf{p}}_0^s).\end{aligned} \tag{7}$$

where $\mathbf{A}_{cs}^t$ and $\mathbf{A}_{cs}^s$ are the attention matrix, Cross-Attn$(\cdot)$ and Self-Attn$(\cdot)$ denote the cross-attention and self-attention operation, respectively. In this manner, we make the original patch feature attend to action triplet prompts and action state prompts, resulting in augmented image features. These features are then transformed by MLP and a re-weighting strategy. The whole process with attention is formulated as

$$\mathbf{V} = \text{MLP}([\mathbf{V}_{cs}^t, \mathbf{V}_{cs}^s]), \tilde{\mathbf{V}} = (\lambda\mathbf{A}_{cs}^t + (1-\lambda)\mathbf{A}_{cs}^s)\mathbf{V}, \tag{8}$$

where $[\cdot]$ indicates the concatenation operation, $\tilde{\mathbf{V}} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_L\} \in \mathbb{R}^{L\times d}$ denotes the action-enhanced visual features, and λ is a hyper-parameter with a value of 0.7. Afterward, we randomly initialize learnable visual vector, defined as $\mathbf{P}_v = [\mathbf{P}]_1[\mathbf{P}]_2\ldots[\mathbf{P}]_N$, where $[\mathbf{P}]_n (n \in \{1, \ldots, N\})$ keeps the same dimension as the image patch embedding. We concatenate the learnable visual prompts and action-enhanced visual features interlaced, resulting in the final visual features $\mathbf{z}_l' = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_L, [\mathbf{P}]_1[\mathbf{P}]_2\ldots[\mathbf{P}]_N]$. Subsequently, we feed them into the rest L-layer transformer layers of visual encoder θ for generating the action-aware visual embedding:

$$\mathbf{z}_i' = \theta_i(\mathbf{z}_{i-1}'),\ i \in [l+1, L]. \tag{9}$$

The output of the last layer $\mathbf{z}_l'$ is treated as the prompted image features $\mathbf{z}^{img}$ used for optimization with contrastive loss and triplet loss in Eq. 10 and Eq. 11.

3.4. Training

We introduce a two-stage training procedure to learn an action-aware vision-language model.

The first training stage. In this stage, the objective is to update newly added parameters of action-aware prompts, triplet encoder, and action-aware adaptive interaction module while keeping the original weights of CLIP frozen. Based on the prompted image features and textual features, the contrastive loss is defined by

$$\begin{aligned}\mathcal{L}_{\text{stage1}} &= \mathcal{L}_{\text{i2t}} + \mathcal{L}_{\text{t2i}},\\ \mathcal{L}_{\text{i2t}}(i) &= -\log \frac{\exp(\text{sim}(\mathbf{z}_i^{img}, \mathbf{z}_i^{text})/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\mathbf{z}_i^{img}, \mathbf{z}_j^{text})/\tau)},\\ \mathcal{L}_{\text{t2i}}(i) &= -\log \frac{\exp(\text{sim}(\mathbf{z}_i^{img}, \mathbf{z}_i^{text})/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\mathbf{z}_j^{img}, \mathbf{z}_i^{text})/\tau)},\end{aligned} \tag{10}$$

where $\mathbf{z}^{text}$ denotes the final text embeddings encoded by a frozen text encoder given an input caption T, $\cos(\cdot)$ is the cosine similarity between the inputs, and τ is a learnable temperature parameter. The terms $\mathcal{L}_{\text{i2t}}$ and $\mathcal{L}_{\text{t2i}}$ refer to image-to-text and text-to-image contrastive losses, respectively. By minimizing $\mathcal{L}_{\text{i2t}}$ and $\mathcal{L}_{\text{t2i}}$, the action-aware prompts will be updated during training through gradient backpropagation.

The second training stage. In this stage, only the parameters of the action-aware adaptive interaction module in visual encoder are optimized while keeping the action-aware prompts and prompt adapter frozen. To boost the final performance, we employ the triplet loss in our second training stage for optimization, given by

$$\mathcal{L}_{\text{stage2}} = \max(d_p - d_n + \alpha, 0), \tag{11}$$

where d_p and d_n represent the feature distances of the positive pair and negative pair, respectively, and α is the margin of $\mathcal{L}_{\text{tri}}$.

The whole training process utilizes the action-aware prompts to mine and store the hidden states of the pre-trained image encoder, enabling CLIP to retain its advantages. The first stage focuses on learning effective prompts to fully harness the capabilities of CLIP, while the second stage refines multi-modal interactions within a more fine-grained embedding space.

3.5. Inference

First of all, the original CLIP model is used to pre-select the top-k (*i.e.*, k=20) samples, after which our method is applied to refine features and re-rank results, enabling fast retrieval. Specifically, for image-to-text retrieval, we first extract text features for pre-selected top-k gallery texts and enrich their semantics using LLM-generated action triplets and state descriptions. Then we extract the image feature enhanced by the enriched texts. Finally, we retrieve the target text by calculating the similarities between the enhanced image feature and all enhanced text features. A similar procedure is adopted for text-to-image retrieval.

4. Experiments

4.1. Datasets & Evaluation Metrics

To evaluate the effectiveness of the proposed method, we conduct extensive experiments using two datasets: Flickr30K [50] and COCO [4]. More details about the statistics of these datasets are reported in the supplementary material.

To evaluate the image-text matching performance, we measure the proportion of queries that match the correct item within the top-k results, denoted as R@K, where K takes values of 1, 5, and 10. The sum of all R@K values is calculated to evaluate the overall matching performance, denoted as Rsum. More details about implementation can be found in the supplementary material.

4.2. Comparison with State-of-the-Art Methods

To verify the effectiveness of the proposed method, we compare it with several state-of-the-art methods on the Flickr30K and COCO datasets. For fair comparisons, all these methods are CLIP-based without pre-training.

The comparison results on the two datasets are shown in Table 1. We have the following observations: 1) Our method achieves significant performance improvements over the pre-trained CLIP models in different ViT architectures. This clearly indicates that enriching the fine-grained action semantic information of visual modality helps image-text alignment. 2) Compared to other CLIP-based methods, Our method achieves competitive results on both datasets in terms of most evaluation metrics, demonstrating the effectiveness of incorporating action knowledge into prompt tuning.

It is worth noting that our method achieves significant performance improvements using the ViT-B backbone compared to those methods based on the ViT-L backbone. The possible reason is that the ViT-B architecture has fewer layers, smaller hidden dimensions, and fewer attention heads, and is inherently less capable of capturing and processing fine-grained image details. Our method can address this limitation by injecting action knowledge into CLIP through action-aware multi-modal prompt tuning, enabling the model to achieve a more fine-grained action perception ability.

4.3. Ablation Study

We perform in-depth ablation studies to evaluate each component of our method on the COCO dataset, using pre-trained CLIP with a ViT-B/16 backbone as the baseline. We

Backbone	Method	COCO (5K test images)						Rsum	Flickr30K (1K test images)						Rsum
		Image-to-Text			Text-to-Image				Image-to-Text			Text-to-Image			
		R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
ViT-B/32	CLIP [32]	50.2	74.6	83.6	30.4	56.0	66.8	361.6	79.0	94.3	98.2	58.0	82.9	89.9	502.3
	MGCA [37]	54.5	78.6	86.8	37.7	63.7	74.0	395.3	81.5	93.9	96.8	64.4	86.5	92.0	515.1
	PyramidCLIP [10]	52.8	78.1	-	38.8	64.9	-	-	84.2	96.4	-	69.1	89.8	-	-
	CLIPDualDIS [44]	-	-	-	37.9	63.4	73.4	-	-	-	-	59.0	83.4	90.1	-
	OpenCLIP [6]	-	-	-	39.4	65.5	75.7	-	-	-	-	63.9	87.3	93.2	-
	SaCo [47]	54.6	77.6	-	36.1	62.0	-	-	81.7	95.5	-	64.9	88.1	-	-
	Ours	**55.5**	**79.2**	**87.8**	**39.9**	**66.0**	**76.3**	**404.7**	**85.1**	**97.4**	**98.3**	**69.9**	**90.1**	**94.8**	**535.6**
ViT-B/16	CLIP [32]	52.4	76.9	84.8	33.1	58.5	69.2	374.9	81.2	96.4	98.5	62.2	85.7	91.8	515.8
	MGCA [37]	57.6	80.5	87.8	39.8	65.7	75.3	406.7	82.2	96.1	98.1	67.7	88.5	93.2	525.8
	PyramidCLIP [10]	55.7	80.8	-	42.6	68.6	-	-	85.6	97.7	-	74.5	92.9	-	-
	EVA-02-CLIP [35]	**58.7**	80.7	88.2	42.2	66.9	76.3	413.1	85.7	96.7	98.9	71.2	91.0	94.7	538.2
	FineCLIP [16]	54.5	78.6	85.8	40.2	66.5	76.1	401.7	82.5	96.4	98.6	67.9	89.1	94.1	528.6
	SaCo [47]	57.8	80.0	-	39.8	64.7	-	-	85.5	96.5	-	69.1	90.1	-	-
	Ours	58.4	**81.8**	**89.1**	**43.2**	**69.5**	**79.0**	**421.1**	**88.1**	**98.0**	**99.2**	**74.7**	**93.1**	**95.8**	**548.9**
ViT-L/14-336	CLIP [32]	57.3	80.6	87.8	37.9	63.4	73.4	394.2	86.6	98.0	99.1	67.1	88.9	93.2	532.9
	MGCA [37]	59.7	83.2	89.7	44.3	69.6	78.8	425.3	86.9	97.3	98.6	74.4	91.7	95.4	544.3
	FILIP [49]	61.3	84.3	90.4	45.9	70.6	79.3	431.8	89.8	99.2	**99.8**	75.0	93.4	96.3	553.5
	REACT [24]	63.3	85.1	-	47.5	72.0	-	-	90.4	99.1	-	76.5	93.7	-	-
	MetaCLIP [43]	**65.5**	85.2	91.1	**48.2**	72.3	81.1	**443.4**	89.5	98.8	99.7	**76.8**	93.9	96.6	555.3
	Ours	62.5	**85.7**	**91.6**	44.1	**72.9**	**81.7**	438.5	**91.5**	**99.5**	**99.8**	74.0	**94.4**	**97.6**	**556.7**

Table 1. Comparative results (without pre-training on external training datasets) on the COCO 5K test set and Flickr30K test set. The best results and the second best results are marked by **bold** and underline, respectively.

report the results on the Flickr30K dataset in the supplementary material.

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
✓				56.2	81.0	88.4	42.4	68.9	78.3
	✓			55.5	80.7	88.1	41.5	68.1	77.9
		✓		53.7	80.9	88.5	42.3	69.1	78.9
✓		✓		58.1	81.8	88.5	42.9	69.3	78.7
✓		✓	✓	**58.4**	**81.8**	**89.1**	**43.2**	**69.5**	**79.0**

Table 2. Ablation studies of prompts on the COCO 5K test set. "Hand-Craft Triplet" denotes replacing the action triplet prompt with a hand-crafted prompt template, *i.e.*, "A photo capturing a [subject] performing [action] in relation to [object].", and "Vis" denotes visual prompting.

Effectiveness 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 shown in Table 2, it is obvious that the proposed action-aware multi-modal prompting outperforms other variants. Specifically, removing either the action triplet prompt or the action state prompt leads to a drop in matching performance. The possible reason is that the action triplet prompt is designed to explicitly model action relations between entities, while the action state prompt is designed to provide additional fine-grained information from action states to mitigate ambiguity in action relationships. The combination of these two types of action prompts can further improve the performance. In addition, removing the visual prompt weakens the model's visual perception ability, thereby hurting the performance.

Effectiveness 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, which simplifies the framework to a standard visual prompt tuning for the vision encoder. As shown in Table 3, we observe that the performance drops significantly. This is expected, as incorporating action knowledge enhances the fine-grained visual perception ability of the pre-trained model.

Effective of attribute knowledge. To evaluate the impact of attribute knowledge in promoting pre-trained CLIP, we remove attribute knowledge from action triplets. As shown in Table 3, the results indicate a slight drop in performance, as attribute knowledge provides more fine-grained details about the states of object attributes, enriching the overall semantic understanding of actions.

Method	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
Baseline	52.4	76.9	84.8	33.1	58.5	69.2
w/o action knowledge	55.3	78.8	86.8	41.5	67.3	76.9
w/o attribute knowledge	58.2	81.6	88.6	43.1	68.6	78.2
replace AIM with CAT	57.4	79.9	87.6	41.9	67.8	77.4
Ours	**58.4**	**81.8**	**89.1**	**43.2**	**69.5**	**79.0**

Table 3. Ablation analysis of different components on the COCO 5K test set.

LLMs	version	Image-to-Text			Text-to-Image		
		R@1	R@5	R@10	R@1	R@5	R@10
Llama-2	Vicuna-7B	58.4	81.4	89.0	43.0	69.2	78.8
Llama-2	Vicuna-13B	**58.5**	81.7	**89.1**	43.2	69.4	**79.1**
GPT-3.5	GPT-3.5-Turbo	58.4	**81.8**	**89.1**	**43.2**	**69.5**	79.0

Table 4. Ablation study of different LLMs.

Effectiveness of action-aware adaptive interaction. To evaluate the effectiveness of action-aware adaptive interaction, we replace the action-aware adaptive interaction module (denoted as "AIM") with a concatenation operation (denoted as "CAT"). Specifically, we directly concatenate action triplet prompts, action state prompts, and learnable visual prompts to generate multi-modal prompts for prompt tuning. The results are reported in Table 3, which shows that

using concatenation operation degrades performance. This demonstrates the vital importance of AIM as it can focus on key action information and aggregate it to generate attentive features, thereby reducing the disturbance caused by the irrelevant or noisy information retained in the action triplet prompts and action state prompts.

Method	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
Baseline	52.4	76.9	84.8	33.1	58.5	69.2
combined training	56.5	80.2	87.9	42.1	68.0	77.6
one-stage training	56.8	80.5	87.8	41.9	67.8	77.8
two-stage training (Ours)	**58.4**	**81.8**	**89.1**	**43.2**	**69.5**	**79.0**

Table 5. Ablation study of different training.

Effectiveness of different LLMs. To explore the effect of using different types and sizes of LLMs (*i.e.*, Llama-2-7B, Llama-2-13B, GPT-3.5), we conduct an ablation study on the COCO dataset. As shown in Table 4, the results indicate that using different types and sizes of LLMs has little impact on performance, demonstrating that even smaller open-source models can achieve comparable results while reducing costs and enhancing reproducibility of our method.

Effect of two-stage training. To evaluate the effectiveness of the proposed two-stage training strategy, we design two different training strategies for comparison:

- **One-stage training**: we only use $\mathcal{L}_{\text{stage1}}$ to jointly train the triplet encoder and action-aware adaptive interaction simultaneously.
- **Combined Training**: we optimize the model by combining $\mathcal{L}_{\text{stage1}}$ and $\mathcal{L}_{\text{stage2}}$ into a unified training process rather than separating the training into two stages.

As shown in Table 5, the results demonstrate that both one-stage training and combination training are less effective. In the early stage of training, the learnable visual tokens cannot effectively describe the image, which negatively impacts the optimization of the image encoder. Additionally, combination training struggles to achieve desirable results, as the simultaneous optimization of $\mathcal{L}_{\text{stage1}}$ and $\mathcal{L}_{\text{stage2}}$ leads to interference between objectives, which hinders effective feature learning. In contrast, our proposed two-stage training strategy achieves better performance. During the first stage, the model focus more on the comparison between positive samples and negative samples without explicit fixed distance constraints, which can accelerate training. During the second stage, the model establishes stable intra-class and inter-class relationships, further enlarging feature distances between categories and enhancing sensitivity to subtle feature differences.

4.4. Qualitative Results

We have presented the comparison results between our method and CLIP in Figure 5. As shown in Figures 5 (a) and (b), given a text query or an image query, with the help of action knowledge from LLM, our method retrieves top-2 ranked results matching the query. In Figure 5 (a), the action knowledge of "standing" is obviously different from actions such as "resting" or "playing". In Figure 5 (b), with the precise description of "riding", our method effectively filters out images of bicycles or trains. However, as shown in Figure 5 (c), the action knowledge generated by the LLM for "<tennis player, backhand, ball>" can not accurately capture the action state of "backhand ball", resulting in incorrect retrieval. A more advanced LLM may be able to solve this problem. More visualization results and analyses of retrieval results can be found in the supplementary material.



Figure 5. The qualitative results on the COCO 5K test set, where correct matches are marked in ✓ and mismatches in ✗.

5. Conclusion

We have presented a simple yet effective LLM-enhanced action-aware multi-modal prompt-tuning method that integrates external action knowledge to prompt the pre-trained vision-language model. We propose an action triplet prompt and an action state prompt that use the knowledge of LLM to help CLIP concentrate on detailed action-related visual cues. We also design an action-aware adaptive interaction module to aggregate key information from different prompts, enhancing the action-aware visual representations. Extensive experiments on the COCO and Flickr30K datasets demonstrate the effectiveness of our method. In the future, we plan to apply our method to tasks such as video-text matching and action recognition, where LLMs can assist in perceiving detailed actions. Furthermore, incorporating LLMs into the training process of CLIP-like methods may be a potential direction.

6. Acknowledgments

This work was supported in part by the grants from the Shenzhen Science and Technology Program under Grant No. JCYJ20241202130548062, the Natural Science Foundation of Shenzhen under Grant No. JCYJ20230807142703006, and the Key Research Platforms and Projects of the Guangdong Provincial Department of Education under Grant No. 2023ZDZX1034.

References

[1] Kyle Buettner, Sina Malakouti, Xiang Lorraine Li, and Adriana Kovashka. Incorporating geo-diverse knowledge into prompting for increased geographical robustness in object recognition. In *CVPR*, pages 13515–13524, 2024. 3

[2] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *CVPR*, pages 15789–15798, 2021. 11

[3] Liangyu Chen, Bo Li, Sheng Shen, Jingkang Yang, Chunyuan Li, Kurt Keutzer, Trevor Darrell, and Ziwei Liu. Large language models are visual reasoning coordinators. *NeurIPS*, pages 10–16, 2024. 1

[4] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 6, 11

[5] Yuhao Cheng, Xiaoguang Zhu, Jiuchao Qian, Fei Wen, and Peilin Liu. Cross-modal graph matching network for image-text retrieval. *TOMM*, 18(4):1–23, 2022. 2

[6] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, pages 2818–2829, 2023. 7

[7] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *AAAI*, pages 1218–1226, 2021. 2

[8] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: improving visual-semantic embeddings with hard negatives. In *BMVC*, page 12, 2018. 2

[9] Zheren Fu, Zhendong Mao, Yan Song, and Yongdong Zhang. Learning semantic relationship among instances for image-text matching. In *CVPR*, pages 15159–15168, 2023. 2

[10] Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, Rongrong Ji, and Chunhua Shen. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *NeurIPS*, pages 35959–35970, 2022. 7

[11] Xuri Ge, Fuhai Chen, Songpei Xu, Fuxiang Tao, and Joemon M Jose. Cross-modal semantic enhanced interaction for image-sentence retrieval. In *WACV*, pages 1022–1031, 2023. 1, 2

[12] Shijie Geng, Jianbo Yuan, Yu Tian, Yuxiao Chen, and Yongfeng Zhang. Hiclip: Contrastive language-image pretraining with hierarchy-aware attention. In *ICLR*, 2023. 1

[13] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *CVPR*, pages 10867–10877, 2023. 1

[14] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *CVPR*, pages 14953–14962, 2023. 1

[15] Yufeng Huang, Jiji Tang, Zhuo Chen, Rongsheng Zhang, Xinfeng Zhang, Weijie Chen, Zeng Zhao, Zhou Zhao, Tangjie Lv, Zhipeng Hu, et al. Structure-clip: Towards scene graph knowledge to enhance multi-modal structured representations. In *AAAI*, number 3, pages 2417–2425, 2024. 1

[16] Dong Jing, Xiaolong He, Yutian Luo, Nanyi Fei, Guoxing Yang, Wei Wei, Huiwen Zhao, and Zhiwu Lu. Fineclip: Self-distilled region-based clip for better fine-grained understanding. In *NeurIPS*, pages 2301–2310, 2024. 7

[17] Baoshuo Kan, Teng Wang, Wenpeng Lu, Xiantong Zhen, Weili Guan, and Feng Zheng. Knowledge-aware prompt tuning for generalizable vision-language models. In *ICCV*, pages 15670–15680, 2023. 1, 3

[18] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, pages 19113–19122, 2023. 3

[19] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, pages 201–216, 2018. 2

[20] Jiangtong Li, Li Niu, and Liqing Zhang. Action-aware embedding enhancement for image-text retrieval. In *AAAI*, number 2, pages 1323–1331, 2022. 2

[21] Jiaming Li, Jiacheng Zhang, Jichang Li, Ge Li, Si Liu, Liang Lin, and Guanbin Li. Learning background prompts to discover implicit knowledge for open vocabulary object detection. In *CVPR*, pages 16678–16687, 2024. 3

[22] Wujin Li, Bin-Bin Gao, Bizhong Xia, Jinbao Wang, Jun Liu, Yong Liu, Chengjie Wang, and Feng Zheng. Cross-modal alternating learning with task-aware representations for continual learning. *TMM*, 2023. 3

[23] Zhuoyi Li, Huibin Lu, Hao Fu, and Guanghua Gu. Image-text bidirectional learning network based cross-modal retrieval. *Neurocomputing*, 483:148–159, 2022. 2

[24] Haotian Liu, Kilho Son, Jianwei Yang, Ce Liu, Jianfeng Gao, Yong Jae Lee, and Chunyuan Li. Learning customized visual models with retrieval-augmented knowledge. In *CVPR*, pages 15148–15158, 2023. 7

[25] Xin Liu, Jiamin Wu, Wenfei Yang, Xu Zhou, and Tianzhu Zhang. Multi-modal attribute prompting for vision-language models. *TCSVT*, 2024. 1

[26] Yanxin Long, Jianhua Han, Runhui Huang, Hang Xu, Yi Zhu, Chunjing Xu, and Xiaodan Liang. Fine-grained visual–text prompt-driven self-training for open-vocabulary object detection. *TNNLS*, pages 1–11, 2023. 3

[27] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *ICLR*, 2023. 3

[28] Yulei Niu, Wenliang Guo, Long Chen, Xudong Lin, and Shih-Fu Chang. SCHEMA: state changes matter for procedure planning in instructional videos. In *ICLR*, 2024. 3

[29] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *ICCV*, pages 3170–3180, 2023. 1

[30] Zhengxin Pan, Fangyu Wu, and Bailing Zhang. Fine-grained image-text matching by cross-modal hard aligning network. In *CVPR*, pages 19275–19284, 2023. 1, 2

[31] Leigang Qu, Meng Liu, Jianlong Wu, Zan Gao, and Liqiang Nie. Dynamic modality interaction modeling for image-text retrieval. In *SIGIR*, pages 1104–1113, 2021. 2

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 1, 7

[33] Sahithya Ravi, Aditya Chinchure, Leonid Sigal, Renjie Liao, and Vered Shwartz. Vlc-bert: Visual question answering with contextualized commonsense knowledge. In *WACV*, pages 1155–1165, 2023. 1

[34] Hung-Ting Su, Yulei Niu, Xudong Lin, Winston H Hsu, and Shih-Fu Chang. Language models are causal knowledge extractors for zero-shot video question answering. In *CVPR*, pages 4951–4960, 2023. 3

[35] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 7

[36] Hao Tan, Jun Li, Yizhuang Zhou, Jun Wan, Zhen Lei, and Xiangyu Zhang. Compound text-guided prompt tuning via image-adaptive cues. In *AAAI*, pages 5061–5069, 2024. 3

[37] Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *NeurIPS*, pages 33536–33549, 2022. 7

[38] Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lin Ma. Consensus-aware visual-semantic embedding for image-text matching. In *ECCV*, pages 18–34. Springer, 2020. 11

[39] Hao Wang, Fang Liu, Licheng Jiao, Jiahao Wang, Zehua Hao, Shuo Li, Lingling Li, Puhua Chen, and Xu Liu. Vilt-clip: Video and language tuning clip with multimodal prompt learning and scenario-guided optimization. In *AAAI*, number 6, pages 5390–5400, 2024. 3

[40] Jinpeng Wang, Pan Zhou, Mike Zheng Shou, and Shuicheng Yan. Position-guided text prompt for vision-language pre-training. In *CVPR*, pages 23242–23251, 2023. 1, 3

[41] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 1, 3

[42] Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *WACV*, pages 1508–1517, 2020. 2

[43] Wenxuan Wang, Quan Sun, Fan Zhang, Yepeng Tang, Jing Liu, and Xinlong Wang. Diffusion feedback helps clip see better. In *ICLR*, 2024. 7

[44] Yimu Wang, Xiangru Jian, and Bo Xue. Balance act: Mitigating hubness in cross-modal retrieval with query and gallery banks. In *EMNLP*, pages 10542–10567, 2023. 7

[45] Yubin Wang, Xinyang Jiang, De Cheng, Dongsheng Li, and Cairong Zhao. Learning hierarchical prompt with structured linguistic knowledge for vision-language models. In *AAAI*, pages 5749–5757, 2024. 3

[46] Jie Wu, Leiquan Wang, Chenglizhao Chen, Jing Lu, and Chunlei Wu. Multi-view inter-modality representation with progressive fusion for image-text matching. *Neurocomputing*, 535:1–12, 2023. 2

[47] Sitong Wu, Haoru Tan, Zhuotao Tian, Yukang Chen, Xiaojuan Qi, and Jiaya Jia. Saco loss: Sample-wise affinity consistency for vision-language pre-training. In *CVPR*, pages 27358–27369, 2024. 7

[48] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *CVPR*, pages 6757–6767, 2023. 3

[49] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: fine-grained interactive language-image pre-training. In *ICLR*, 2022. 7

[50] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014. 6, 11

[51] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *TOMM*, 16(2): 1–23, 2020. 2

[52] Zhaoheng Zheng, Jingmin Wei, Xuefeng Hu, Haidong Zhu, and Ram Nevatia. Large language models are good prompt learners for low-shot image classification. *arXiv preprint arXiv:2312.04076*, 2023. 3

[53] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *CVPR*, pages 16793–16803, 2022. 1

[54] Yi Zhu, Zhaoqing Zhu, Bingqian Lin, Xiaodan Liang, Feng Zhao, and Jianzhuang Liu. Relclip: Adapting language-image pretraining for visual relationship detection via relational contrastive learning. In *EMNLP*, pages 4800–4810, 2022. 1