Multimodal Semantic Decoupled Prompt for Zero-Shot Referring Expression Comprehension

Yuxuan Zhang[a,*], **Longfei Huang**[a,1] **and Yang Yang**[a,**]

[a]Nanjing University of Science and Technology

Abstract. Large-scale Vision-Language Models (VLMs) have demonstrated impressive zero-shot performance in sample-level downstream tasks (e.g., image classification), driven by their powerful generalization ability. However, they still struggle in instance-level tasks, e.g., zero-shot Referring Expression Comprehension (REC), which requires precisely locating the target instance in an image based on a provided text caption. To address this issue, we propose Multimodal Semantic Decoupled Prompting (MSDP), a simple yet effective prompt engineering approach that contains both textual- and visual-focused instance-level understanding prompting. Specifically, we first propose a novel textual restructure strategy to eliminate the impact of task-irrelevant semantic information, steering the model's attention at the textual understanding level. Meanwhile, we design a united visual prompt at the visual understanding level that maximally activates the instance-level understanding capabilities of VLMs. Experiments on several benchmarks reveal that the proposed approach outperforms state-of-the-art (SOTA) methods. The code is available at repository.

1 Introduction

The development of large-scale vision-language models has enabled various sample-level downstream vision tasks [3, 4, 9, 17, 18, 26, 31], such as image captioning [1, 5, 24, 30] and visual grounding [19, 37], to achieve exceptional zero-shot performance, powered by the advanced generalization capabilities. However, the lack of an explicit instance-level process continues to limit the performance of existing VLMs in zero-shot instance-level tasks, e.g., zero-shot REC [2, 10, 39, 38], which seeks to match an arbitrary given text caption with the corresponding target proposal from multiple candidates within an image.

To solve these problems, existing studies can be broadly classified into two main pathways: 1). finetuning-based methods; 2). visual prompt-based methods. The finetuning-based methods seek to improve instance-level understanding capabilities by incorporating additional instance-level finetuning tasks [6, 11, 28]. Specifically, existing works typically construct scene graphs or triplets for both visual and textual modalities, intending to explicitly map these elements and learn the relationships and attributes between different entities, thereby enhancing instance-level understanding capabilities. Unlike the aforementioned method, which requires additional task design and retraining resources, the visual prompt-based methods [20, 22, 27] focus on leveraging the inherent instance-level understanding abilities of VLMs by incorporating visual prompts (e.g., boxes [22], circles [20], and attention matrix [40]). Early attempts [20, 22] aim to leverage bounding boxes of varying shapes and colors to guide the model's attention. Recent studies [27, 40] have begun to explore fine-grained visual prompts and investigate their relationship with the attention mechanisms of VLMs.

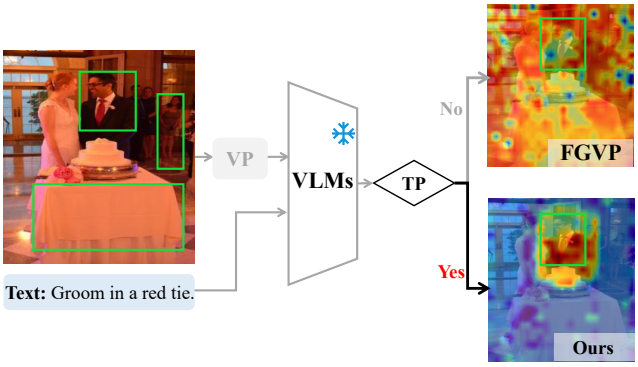

Figure 1. Comparison of the SOTA visual prompt-based method FGVP with the proposed approach in activating VLMs' instance-level understanding capabilities. VP: visual prompt-based methods. TP: Textual Prompt (e.g., Textual Restructure). Deeper red shows more focus on that specific area.

However, existing visual prompt-based methods ignore the textual modality, which may fail to fully leverage VLMs' inherent instance-level understanding capabilities. To verify the intention, we first analyze the SOTA visual prompt-based method FGVP's [27] attention to the target region, as depicted in Fig. 1. The results indicate that existing visual prompt-based methods fail to focus fully on the target region, suggesting that the potential of instance-level understanding capabilities remains to be explored. Conversely, by integrating the textual modality, our approach enables the model to focus more effectively on the target region.

Therefore, we propose a simple yet effective approach called Multimodal Semantic Decoupled Prompting (MSDP) for zero-shot REC. Unlike existing methods focusing solely on visual perception, the proposed MSDP emphasizes the activation of multimodal instance-level understanding capabilities and introduces a textual restructure strategy along with the united visual prompts to steer the model's attention in tandem. Specifically, the textual restructure module aims to eliminate the influence of irrelevant semantic information, thereby directing the model's attention away from non-relevant textual semantics. Meanwhile, the united visual prompt seeks to harness the benefits of multi-granularity visual prompts, thereby optimizing the

* Email: xuan_yuzhang@njust.edu.cn
** Corresponding Author. Email: yyang@njust.edu.cn.
[1] Equal contribution. Email: hlf@njust.edu.cn

activation of instance-level understanding abilities. Additionally, our experimental results demonstrate that when the inherent instance-level understanding capabilities of the model are effectively and fully activated through the proposed MSDP framework, it consistently outperforms existing state-of-the-art methods, even those that rely on fine-tuning in zero-shot REC.

2 Related Work

2.1 Large-scale Vision-Language Models

Leveraging the generalization capabilities of large-scale pretraining, VLMs, e.g., CLIP [17], have demonstrated impressive zero-shot performance on sample-level tasks like image classification [25, 31] and image-text retrieval [29, 33, 32]. This advancement further inspires researchers to leverage large language models to develop more powerful VLMs [13, 15, 21]. FLAVA [21] incorporates both unimodal and multimodal pretraining tasks, including cross-modal alignment and multimodal fusion objectives, to develop a unified model that spans all modalities. BLIP-2 [13] employs a two-stage pretraining strategy that leverages a frozen image encoder and a powerful large language model, enabling efficient language-image pertaining. LLaVA [15] trains a connector to align image-text features and performs end-to-end finetuning on multimodal instruction data.

Nevertheless, for instance-level tasks (e.g., zero-shot REC) requiring instance-level understanding capabilities (i.e., identifying specific instances of visual scenes and complex text), current methods still require the design of specialized modules [6, 11] for finetuning and enhancement, due to the lack of clear guidance for instance-level process in VLMs.

2.2 Zero-shot Referring Expression Comprehension

Zero-shot REC harnesses the generalization ability of VLMs to understand entity relationships in a provided text description and choose the corresponding proposal from the candidate bounding boxes. To better leverage the instance-level understanding capabilities of VLMs, existing methods can be categorized into two pathways: finetuning-based methods for instance-level tasks [6, 11, 28], and visual prompt-based methods incorporating additional visual cues [20, 23, 22, 27, 40]. The finetuning-based methods aim to design tasks capable of modeling the relationships between instances. For example, REC_SS [6] disentangles both image and text into triplets and further introduces a triplet-matching task to facilitate the understanding of relationships among the disentangled entities.

Meanwhile, visual prompt-based methods leverage manually crafted visual prompts to direct the model's attention. Typically, RedCircle [20] replaces the traditional bounding box with a red circle to further direct the model's attention. Furthermore, recent studies aim to explore the impact of fine-grained visual prompts and investigate the relationship between visual prompts and model attention. FGVP [27] takes advantage of SAM [12] to generate the fine-grained visual prompt and achieve a better performance. FALIP [40] posits that different visual prompts essentially increase the weights of the corresponding attention modules, thereby enhancing the model's instance-level understanding capabilities by explicitly weighting the target regions. Given that existing methods with visual prompts focus on visual perception, whether multimodal prompts could further enhance instance-level understanding abilities remains to be explored.

3 Proposed Method

This section begins by defining zero-shot REC and providing an overview of the pipeline. The proposed MSDP framework, illustrated in Figure 2, consists of two key components: the textual restructure and the united visual prompt.

3.1 Problem Statement

The zero-shot REC task takes an image with a set of proposals and textual captions as input, aiming to achieve the best alignment between the proposals and the given caption. Specifically, given an image, we define the box proposal set generated by the image as $\mathcal{P} = \{\boldsymbol{p}_m\}_{m=1}^{M}$, where M denotes the number of proposals. Subsequently, the textual caption set $\mathcal{T} = \{\boldsymbol{t}_n\}_{n=1}^{N}$ corresponding to the image is also available, where N denotes the number of captions. Following the setting of [20, 27], we adopt an image encoder and a text encoder to obtain features: $\boldsymbol{u}_m = \mathtt{E}_v(\boldsymbol{p}_m), \boldsymbol{v}_n = \mathtt{E}_t(\boldsymbol{t}_n)$, where $\boldsymbol{u}_m$ and $\boldsymbol{v}_n$ represents the visual and textual feature, respectively. Then we can calculate the matching score by:

$$s(\boldsymbol{p}_m, \boldsymbol{t}_n) = \mathtt{sim}(\mathtt{E}_v(\boldsymbol{p}_m), \mathtt{E}_t(\boldsymbol{t}_n)) = \frac{\boldsymbol{u}_m^\top \cdot \boldsymbol{v}_n}{\|\boldsymbol{u}_m\|\|\boldsymbol{v}_n\|},$$

where $s(\cdot,\cdot)$ denotes the score function. Then, the best match proposal of the given caption $\boldsymbol{t}_n$ can be chosen by: $\boldsymbol{p}_n^* \triangleq \mathrm{argmax}_{\boldsymbol{p}\in\mathcal{P}}\ \{s(\boldsymbol{p}, \boldsymbol{t}_n)\}$. Furthermore, the overall objective can be defined as identifying the optimal proposal set $\mathcal{P}^*$ for each textual caption $\boldsymbol{t}_n \in \mathcal{T}$, which can be expressed as:

$$\mathcal{P}^* = \left\{ \boldsymbol{p}_n^* \mid \boldsymbol{p}_n^* \triangleq \underset{\boldsymbol{p}\in\mathcal{P}}{\mathrm{argmax}}\ s(\boldsymbol{p}, \boldsymbol{t}_n) \right\}_{n=1}^{N}.$$

3.2 Textual Restructure

Existing VLMs mitigate the modality gap between image and text by pretraining on large-scale image-text pairs. To further reduce the gap, they introduce aligned textual expressions before the task-relevant text $\boldsymbol{t}_n^{(tr)}$ in the input [17]. To illustrate, in the frequent example "A photo of {*text*}", "A photo of" is associated with the task-irrelevant text $\boldsymbol{t}^{(ti)}$, while the "{*text*}" represents the $\boldsymbol{t}_n^{(tr)}$ (i.e., $\boldsymbol{t}_n$). This paradigm also appears in most image-level downstream tasks, facilitating improved zero-shot performance [13, 15]. When it comes to instance-level tasks, e.g., zero-shot REC, introducing prompting or finetuning [40, 6] to activate the model's instance-level understanding abilities has become a consensus. Since existing visual prompt-based methods aim to attract the model's attention solely from a visual perspective, it is natural to investigate whether refocusing the model's attention from the textual modality can further enhance its capabilities. Inspired by [8], we argue that concise textual descriptions, which emphasize specific aspects of a scene without delving into extraneous details, are more effective for feature representation.

To validate this hypothesis, we first propose a textual restructure strategy. Specifically, we define the total textual caption $\boldsymbol{t}_n^{total}$ for each $\boldsymbol{t}_n \in \mathcal{P}$ containing a task-irrelevant caption $\boldsymbol{t}^{(ti)}$ and a task-relevant caption $\boldsymbol{t}_n^{(tr)}$. To refocus the model's attention, we aim to remove the impact of task-irrelevant text $\boldsymbol{t}^{(ti)}$ in the total textual caption $\boldsymbol{t}_n^{total}$. The formulation of the original textual feature vector as utilized in existing methodologies for zero-shot REC can be comprehensively described by the following equation:

$$\boldsymbol{v}_n^{total} = \mathtt{E}_t(\boldsymbol{t}_n^{total}) = \mathtt{E}_t(\boldsymbol{t}^{(ti)} \oplus \boldsymbol{t}_n^{(tr)}), \tag{1}$$

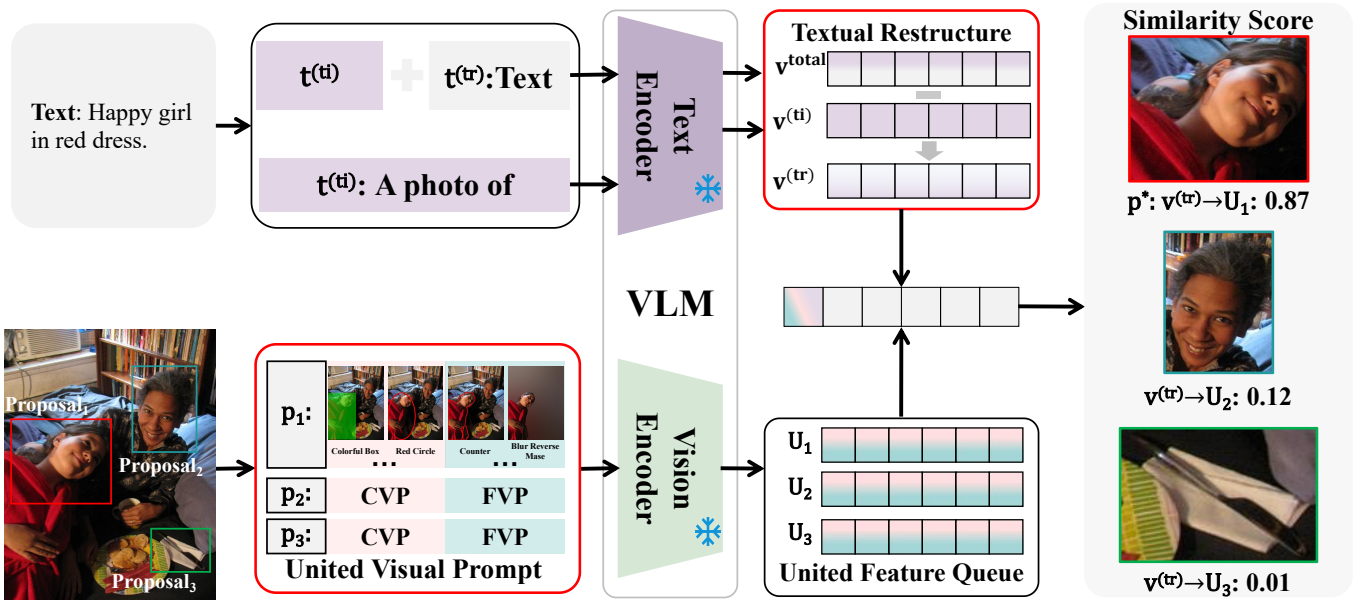

Figure 2. Framework of the proposed approach. $\boldsymbol{p}$: proposal; $t^{(ti)}$: task-irrelevant text. $t^{(tr)}$: task relevant text; CVP: coarse-grained visual prompt; FVP: fine-grained visual prompt. Initially, we apply a textual restructure strategy to remove task-irrelevant semantic features at the textual understanding level, while the united visual prompt steers the model's attention at the visual understanding level. Finally, we pair each text with its most relevant proposal $\boldsymbol{p}^*$.

where $\oplus$ represents the operation of the connection. We can observe that any target information in the $\boldsymbol{t}_n^{total}$ will be diluted by the presence of $\boldsymbol{t}^{(ti)}$ (i.e., task-irrelevant text). In a similar vein to existing methods [20, 27], we perceive $\boldsymbol{t}^{(ti)}$ as useless information in the textual domain. To minimize the impact of $\boldsymbol{t}^{(ti)}$ on the model's performance, we aim to reduce its influence at the feature level. Specifically, we subtract the textual feature vector $\boldsymbol{v}^{(ti)} = \mathtt{E}_t(\boldsymbol{t}^{(ti)})$ corresponding to $\boldsymbol{t}^{(ti)}$ from Eq (1). The entire process can be formalized as follows:

$$\boldsymbol{v}_n^{(tr)} = \boldsymbol{v}_n^{total} - \boldsymbol{v}^{(ti)}, \tag{2}$$

where $\boldsymbol{v}_n^{(tr)}$ denotes the semantic decoupled textual feature.

3.3 *United Visual Prompt*

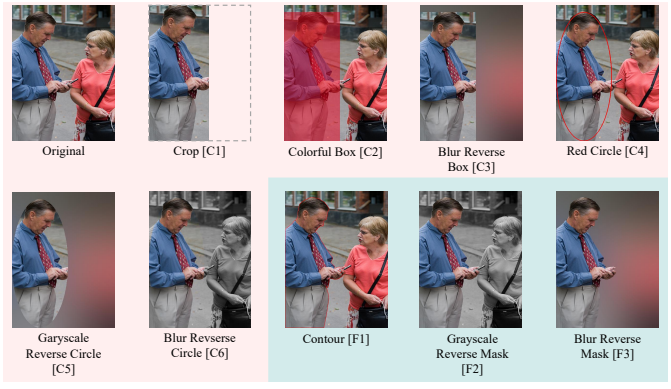

Figure 3. A summary of the united visual prompts used in this paper with the caption "man on the left". Note that visual prompts highlighted in pink represent CVP, while those highlighted in green represent FVP.

Existing visual prompt-based methods utilize the initial proposals generated by a pre-trained detector to produce specialized visual prompts. Prior works can be categorized into coarse-grained and fine-grained visual prompts (i.e., CVP, FVP), depending on the utilization of SAM. We define the different CVPs as [C1]-[C6], and the FVPs as [F1]-[F3], as illustrated in Fig. 3. The CVP utilizes the handmade visual cue to steer the model's attention. Meanwhile, the FVP utilizes SAM to obtain the boundary contours of the target region, thereby achieving fine-grained representation, which could preserve more global information than CVP. Since visual prompts of varying

Algorithm 1: The algorithm of the proposed MSDP.

Input : An image with M box proposals $\mathcal{P} = \{\boldsymbol{p}_m\}_{m=1}^{M}$;
N texts captions $\mathcal{T} = \{\boldsymbol{t}_n\}_{n=1}^{N}$;
Vision and Text Encoder $\mathtt{E}_v(\cdot)$, $\mathtt{E}_t(\cdot)$.
Output: The optimal proposal set $\hat{\mathcal{P}}^*$ for each t_n.
INIT : Generate $K + L$ kinds visual prompts for each $\boldsymbol{p}_m \in \mathcal{P}$.

for $n \leftarrow 1$ **to** N **do**
  /* Textual Restructure */
  Compute the original text feature $\boldsymbol{v}_n^{total}$ according to Eq. (1);
  Compute the feature of task-irrelevant text prompt according to $\boldsymbol{v}^{(ti)} = \mathtt{E}_t(\boldsymbol{t}^{(ti)})$;
  Compute the semantic decoupled text feature $\boldsymbol{v}_n^{(tr)}$ according to Eq. (2);
  /* United Visual Prompt */
  for $m \leftarrow 1$ **to** M **do**
    for $k \leftarrow 1$ **to** K **do**
      Compute $\boldsymbol{u}_{km}^{(c)}$ according to Eq. (3);
    for $l \leftarrow 1$ **to** L **do**
      Compute $\boldsymbol{u}_{lm}^{(f)}$ according to Eq. (3);
    Get $\boldsymbol{U}_m$ according to Eq. (4);
    Compute $\hat{s}(\boldsymbol{p}_m, \boldsymbol{t}_n)$ according to Eq. (5);
Get the optimal proposal set $\hat{\mathcal{P}}^*$ according to Eq. (6).

granularities steer the model's attention differently [40], it is natural to extend the original proposal with multi-grained prompts for richer visual representations.

Thus, we propose the united visual prompt strategy. Specifically, we extend the original proposal set $\mathcal{P}$ to $\mathcal{P}^{(\mathrm{c})} \cup \mathcal{P}^{(\mathrm{f})}$ by leveraging the various visual prompts, where $\mathcal{P}^{(\mathrm{c})}$ and $\mathcal{P}^{(\mathrm{f})}$ represent the images set after using CVP and FVP. We then define i-th CVP and FVP as $\phi_i(\cdot)$ and $\psi_i(\cdot)$, respectively. Hence, we can obtain the coarse- and fine-grained features for proposal $\boldsymbol{p}_m$ by:

$$\begin{aligned} \boldsymbol{u}_{im}^{(\mathrm{c})} &= \mathtt{E}_v(\phi_i(\boldsymbol{p}_m)), \\ \boldsymbol{u}_{im}^{(\mathrm{f})} &= \mathtt{E}_v(\psi_i(\mathtt{SAM}(\boldsymbol{p}_m))). \end{aligned} \tag{3}$$

Since the proposed united visual prompt is a united framework extendable by any new visual prompt, we define it to contain K coarse-grained visual prompts and L fine-grained visual prompts. Then we get the set of visual features $\boldsymbol{U}_m$ by:

$$\boldsymbol{U}_m = \{\boldsymbol{u}_{1m}^{(\mathrm{c})}, \cdots, \boldsymbol{u}_{Km}^{(\mathrm{c})}\} \cup \{\boldsymbol{u}_{1m}^{(\mathrm{f})}, \cdots, \boldsymbol{u}_{Lm}^{(\mathrm{f})}\}. \tag{4}$$

3.4 Overall Objective

Based on the module described above, the revised matching score $\hat{s}(\cdot,\cdot)$ between each proposal and the caption can be represented as follows:

$$\hat{s}(\boldsymbol{p}_m, \boldsymbol{t}_n) = \sum_{\boldsymbol{u} \in \boldsymbol{U}_m} \mathtt{sim}(\boldsymbol{u}, \boldsymbol{v}_n^{(tr)}), \tag{5}$$

where $\boldsymbol{u}$ represents the visual features of different visual prompts. The equation shows that the matching score of a proposal is determined by the sum of similarities between the vision features obtained from united visual prompting and the text features $\boldsymbol{v}_n^{(tr)}$ obtained after textual restructure. Consequently, the best match proposal of the given caption can be chosen by the following equation: $\hat{\boldsymbol{p}}_n^* \triangleq \mathrm{argmax}_{\boldsymbol{p} \in \mathcal{P}} \{\hat{s}(\boldsymbol{p}, \boldsymbol{t}_n)\}$. Thereby, the overall objective can be rewritten as:

$$\hat{\mathcal{P}}^* = \left\{ \hat{\boldsymbol{p}}_n^* \mid \hat{\boldsymbol{p}}_n^* \triangleq \underset{\boldsymbol{p} \in \mathcal{P}}{\mathrm{argmax}}\ \hat{s}(\boldsymbol{p}, \boldsymbol{t}_n) \right\}_{n=1}^{N}. \tag{6}$$

To better understand our algorithmic process, we present the detailed flow of MSDP in Algorithm 1.

4 Experiment

4.1 Experimental Setting

Datasets. We evaluate the effectiveness of the proposed MSDP on widely used REC benchmarks, including RefCOCO [35], RefCOCO+ [35], and RefCOCOg [16] following the standard categorization of zero-shot REC methods [27, 22]. The three datasets mentioned above are subsets of the COCO [14] dataset, which consists of bounding boxes and masks associated with captioned instances. RefCOCO+ is specifically designed to exclude spatial relations, focusing solely on appearance-based expressions. In contrast, the RefCOCO and RefCOCOg datasets encompass both appearance-based and relation-based expressions. This distinction allows for a comprehensive evaluation of different types of expression, enabling a deeper analysis of the challenges posed by spatial relations in referring expression comprehension. RefCOCO and RefCOCO+ test sets are divided into two subsets: "TestA" comprises only people, while "TestB" includes non-people.

Baselines and Performance Criteria. We conduct a comparison between MSDP and three distinct categories of methods: 1). CVP-based methods, including CPT [34], ReCLIP [22], and RedCircle [20]. 2). FVP-based method, FGVP [27], FALIP [40]. 3). Finetuning-based method, REC_SS [6]. The evaluation focuses on the accuracy of text caption and proposals.

Implementation Details. We utilize the CLIP [17] trained by OpenAI, namely ViT-B/16, ViT-B/32, ViT-L/14@336px, and RN50×16 backbones, following [27, 12, 40]. Following [6], we also conduct experiments on FLAVA [21] to further assess the performance of MSDP across other VLMs. For the fine-grained visual prompt, we employ SAM-ViT-H, a variant of the SAM [6]. To optimize performance, we have established a line thickness of 2 pixels for line-based visual prompts, while maintaining the mask precision of 1.0. All experiments are conducted on a single RTX 3090 GPU. More details can be found in the supplementary material.

4.2 Performance Comparison

The results in Table 1 present the comparison of the accuracy for MSDP and other zero-shot REC methods. To ensure fairness, we utilize the proposals from MAttNet [36] to evaluate the robustness of the proposed MSDP while ensuring that different proposal selections only affect the box candidates that are equitably shared among all the comparison prompting methods, following by [22, 27, 40]. Note that, we adopt the same baseline as various state-of-the-art methods to ensure a fair comparison. Besides the rows marked with $*$, all results are sourced from the original papers. The results in Table 1 show that: 1). Under the backbones of various state-of-the-art methods, the proposed MSDP consistently achieves superior performance, with an average improvement of at least 3.9%. In particular, on the TestA subset of RefCOCO with the backbone of ViT-L, RN50, it outperforms the SOTA method FGVP by 7.4%, indicating that the proposed MSDP can better activate the VLM's instance-level understanding capability. 2). Compared to CVP-based methods, such as ReCLIP and RedCircle, FVP-based methods demonstrate superior performance. For instance, the FGVP achieves a significant improvement of 5.4% over RedCircle on average using the ViT-B/32, RN50 baseline. The results may indicate the FVP makes more contributions to activate the instance-level understanding abilities. 3). We supervised find that even under identical conditions, the proposed approach achieves an average improvement of 10.8% over the finetuning-based state-of-the-art method REC_SS with the ViT-B/32 baseline. Specifically, on the TestA subset of RefCOCO with the backbone of ViT-B/32, it outperforms the SOTA method REC_SS by 20.1%, indicating that the inherent instance-level understanding can achieve strong performance when effectively activated through appropriate methods. 4). The results using FLAVA demonstrate that the proposed method exhibits strong generalization across different VLMs. Furthermore, the MSDP outperforms the REC_SS in most cases, with an average performance improvement of 5.3%.

4.3 Ablation Study

4.3.1 Impact of the Textual Restructure

To further validate the contribution of textual restructure, we compare the performance of existing visual prompt-based methods on each dataset before and after incorporating textual restructure, as illustrated in Table 2. The results indicate that existing visual prompt-based methods can significantly benefit from the incorporation of

Table 1. The accuracy (%) of zero-shot REC on RefCOCO, RefCOCO+, and RefCOCOg datasets. The best results are highlighted in **bold**, and the second-best results are underlined. $*$ is our reproduction.

Method	Backbone	RefCOCO			RefCOCO+			RefCOCOg		AVG
		Val	TestA	TestB	Val	TestA	TestB	Val	Test	
CPT[Arxiv2022]	ViT-B/16	32.2%	36.1%	30.3%	31.9%	35.2%	28.8%	36.7%	36.5%	33.5%
ReCLIP[ACL2022]		38.2%	40.5%	37.0%	41.5%	42.9%	41.3%	55.2%	55.2%	44.0%
RedCircle[ECCV2023]		45.3%	52.7%	36.5%	49.4%	57.7%	40.6%	53.7%	53.3%	48.7%
FALIP[ECCV2024]		46.7%	51.7%	38.3%	51.9%	57.1%	43.0%	54.2%	54.9%	49.7%
MSDP		**61.6**%	**68.8**%	**54.6**%	**59.9**%	**69.0**%	**49.5**%	**63.0**%	**63.5**%	**61.2**%
CPT[Arxiv2022]	ViT-B/32	23.8%	22.9%	26.0%	23.5%	21.7%	26.3%	21.8%	22.8%	23.6%
ReCLIP[ACL2022]		40.7%	44.0%	37.6%	45.0%	48.2%	41.7%	55.3%	54.4%	45.8%
RedCircle[ECCV2023]		38.7%	45.1%	33.5%	42.9%	49.5%	36.5%	45.8%	45.6%	42.2%
REC_SS[CVPR2024]		48.2%	48.4%	49.2%	45.6%	47.6%	42.8%	57.6%	56.6%	49.5%
MSDP		**60.5**%	**68.5**%	**53.4**%	**59.7**%	**68.2**%	**48.4**%	**61.2**%	**62.0**%	**60.3**%
CPT[Arxiv2022]	ViT-B/32, RN50	41.3%	40.6%	44.0%	41.3%	41.8%	41.1%	51.3%	51.2%	44.1%
ReCLIP[ACL2022]		42.0%	43.5%	39.0%	47.4%	50.1%	43.9%	57.8%	57.2%	47.6%
RedCircle[ECCV2023]		45.6%	54.0%	37.1%	50.7%	60.5%	41.7%	54.0%	53.8%	49.7%
FGVP*[NeurIPS2023]		52.0%	55.9%	48.8%	53.3%	60.4%	46.7%	62.1%	61.9%	55.1%
MSDP		**63.5**%	**72.1**%	**54.3**%	**62.9**%	**73.3**%	**50.7**%	**64.5**%	**64.8**%	**63.3**%
RedCircle[ECCV2023]	ViT-L, RN50	49.8%	58.6%	40.0%	55.3%	63.9%	45.4%	59.4%	58.9%	53.9%
FGVP[NeurIPS2023]		59.6%	65.0%	52.0%	60.0%	66.8%	49.7%	63.3%	63.4%	60.0%
MSDP		**64.0**%	**72.4**%	**54.9**%	**63.5**%	**73.3**%	**51.6**%	**65.5**%	**65.8**%	**63.9**%
REC_SS+FLAVA	FLAVA	49.4%	47.8%	51.7%	48.9%	50.0%	46.9%	**61.0**%	60.0%	51.9%
MSDP+FLAVA		**57.3**%	**62.8**%	**53.1**%	**55.2**%	**62.5**%	**47.1**%	59.6%	**60.2**%	**57.2**%

Table 2. Ablation study of text restructure conducted on RefCOCO, RefCOCO+, and RefCOCOg. TR: text restructure. The w/ TR symbol denotes the performance after applying text restructure to different visual prompt-based methods.

Method	Backbone	RefCOCO			RefCOCO+			RefCOCOg		AVG
		Val	TestA	TestB	Val	TestA	Testb	Val	Test	
FALIP	ViT-B/16	46.7%	51.7%	38.3%	51.9%	57.1%	43.0%	54.2%	54.9%	49.7%
FALIP w/ TR		47.0%	52.5%	40.1%	52.5%	57.9%	45.5%	57.0%	56.4%	51.1%
RedCircle	ViT-L, RN50	49.8%	58.6%	40.0%	55.3%	63.9%	45.4%	59.4%	58.9%	53.9%
RedCircle w/ TR		50.0%	58.7%	41.4%	55.5%	64.6%	45.6%	59.9%	58.9%	54.3%
FGVP	ViT-L, RN50	59.6%	65.0%	52.0%	60.0%	66.8%	49.7%	63.3%	63.4%	60.0%
FGVP w/ TR		59.3%	67.2%	51.7%	60.1%	68.8%	49.6%	63.3%	64.2%	60.5%
MSDP	ViT-L, RN50	**64.0**%	**72.4**%	**54.9**%	**63.5**%	**73.3**%	**51.6**%	**65.5**%	**65.8**%	**63.9**%

our proposed textual restructure module. This module enhances the model's focus on the textual modality, thereby enhancing instance-level understanding and overall performance. As shown in Table 2, all methods show a minimum performance improvement of 0.5% on average in all three datasets when textual restructure is applied.

Table 3. Ablation study of united visual prompt on RefCOCO.

Method	Visual Prompt	RefCOCO		
		Val	TestA	TestB
MSDP	C1	56.1%	64.2%	47.8%
	C1 | C3 | C4	57.8%	66.6%	49.4%
	C1 | F1	58.8%	62.8%	50.6%
	C1 | F1 | F2 | F3	59.3%	67.2%	51.7%
	C1 | C3 | C4 | F1 | F2	62.9%	71.0%	53.2%
	C1 | C3 | C4 | F1 | F2 | F3	**64.0**%	**72.4**%	**54.9**%

4.3.2 *Impact of the United Visual Prompt*

To better understand the impact of the united visual prompt, we conduct an in-depth performance evaluation on the RefCOCO dataset. The results presented in Table 3 reveal a proportional relationship between the inclusion of visual prompts at varying granularities and method performance, with fine-grained visual prompts yielding substantial improvements.

4.4 *Impact of Different Text Prompts*

Considering the distinct impact of different text prompts (i.e., $t^{(ti)}$), we select 3 different types of task-irrelevant text (i.e., text prompts) for the experiment based on length. Note that apart from changing the text prompt, all other settings remain the same. The results in Table 4 demonstrate that longer text prompts can provide better visual alignment ability.

Table 4. Influence of different text prompts on RefCOCO.

Method	Text Prompt	RefCOCO		
		Val	TestA	TestB
MSDP	-	36.6%	35.5%	54.8%
	This is {*text*}	62.8%	70.9%	54.4%
	A photo of {*text*}	64.0%	72.4%	54.9%
	A photo of a {*text*}	**64.1**%	**72.7**%	**55.1**%

4.5 *Impact of the Post Process*

Note that some methods adopt other strategies including Relations [22] and Subtraction [20] to obtain the best match proposal as the post process. Relations focus on enhancing performance by considering positional relationships between every two proposals.

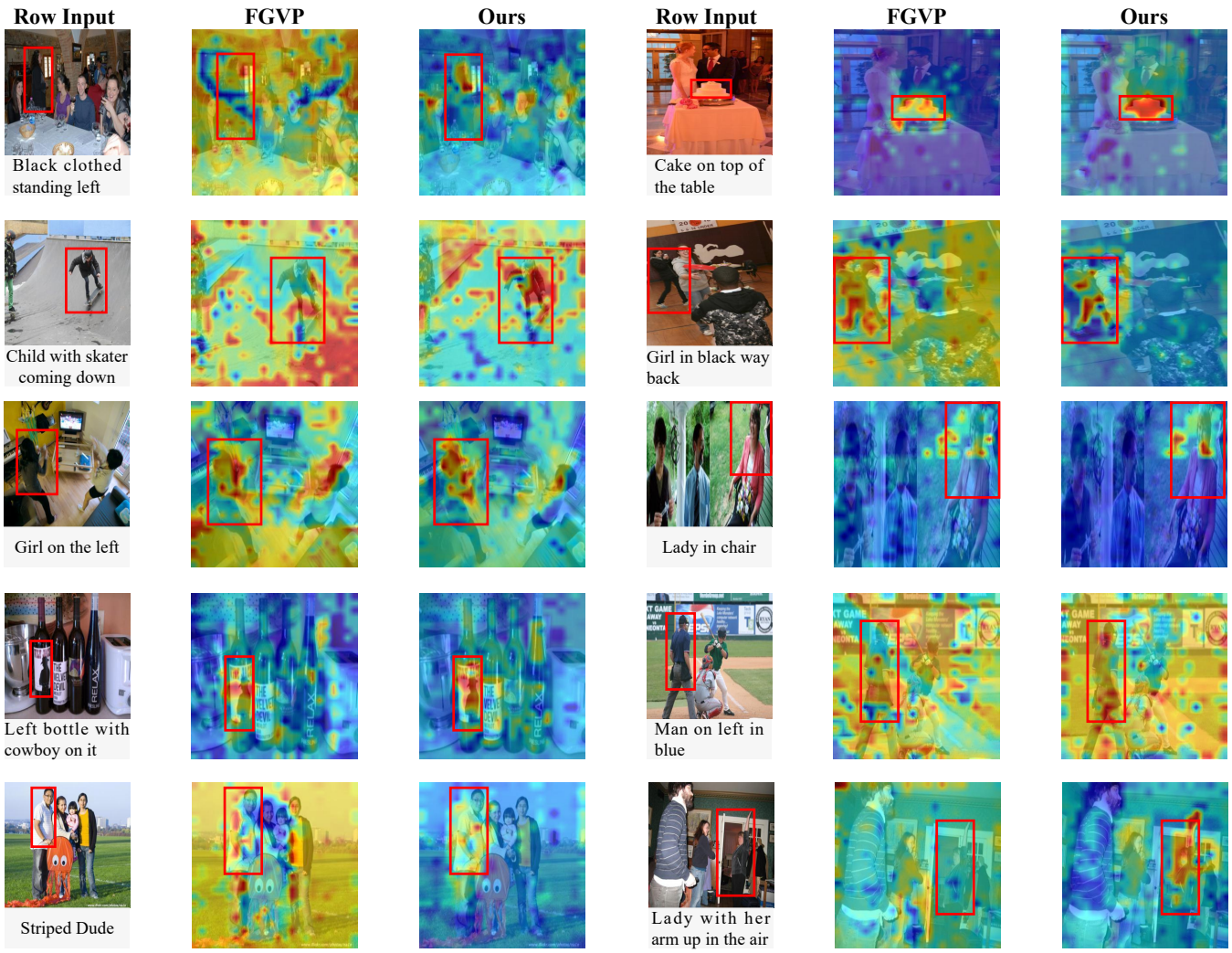

Figure 4. Visualization of the model's attention for the SOTA visual prompt-based method FGVP and the proposed MSDP. The intensity of red indicates a higher level of attention to the corresponding target region, while greener areas indicate less attention.

Meanwhile, Subtraction utilizes negative text to penalize matching scores. To investigate the impact and effectiveness of post process, we conduct a comparative analysis of different post-processing combinations on accuracy. Throughout the evaluation, we ensure that the optimal visual prompt combination remained unchanged. Table 5 presents the results in the RefCOCO datasets. Although the approach score subtraction approach (denoted as **S**) achieves better performance, it comes with a higher computational cost. Furthermore, we observe that utilizing textual restructure alone can achieve performance on par with approach **S** while requiring fewer computational resources. Finally, we find that combining approach spatial relations (denoted as **R**) with textual restructure maximizes the effectiveness of referring expressions while minimizing computational costs.

Table 5. Ablation study of post-process technique on RefCOCO with unchanged optimal visual prompts. PP: post-process. TR: Text Restructure. **R**: Relation. **S**: Subtraction.

Method	PP	TR	Time Cost	RefCOCO		
				Val	TestA	TestB
MSDP	R		1.6h	52.3%	56.6%	49.5%
	S		3.7h	54.9%	60.6%	44.8%
		✓	1.5h	54.9%	63.5%	44.0%
	R	✓	1.7h	**64.0**%	**72.4**%	**54.9**%
	S	✓	6.0h	54.6%	60.3%	44.4%
	RS		7.6h	60.0%	64.6%	52.5%
	RS	✓	13.8h	59.9%	64.5%	52.5%

4.6 Impact of the Larger Text Encoder

To further investigate the influence of the text encoder on spatial relations, we conduct a comparison of performance and computational costs using a larger text encoder implemented by spaCy [7], following the setting of ReCLIP [22] on the RefCOCO dataset. The results in Table 6 demonstrate that larger text encoders improve VLMs' instance-level understanding capability with minimal impact on storage and time.

4.7 Impact of Line Thickness

Taking into account that our approach incorporates various line-based visual prompts, we conduct a detailed parameter study of line thickness. Specifically, compared to the previous method [27], we conduct an ablation study on a larger scale by varying the thickness with a set $\{1, 2, 4, 6, 8\}$, keeping the color fixed as red. The results on the RefCOCO dataset depicted in Figure 6 demonstrate the consistent performance of the proposed MSDP across different parameter settings. This phenomenon highlights the text features after the

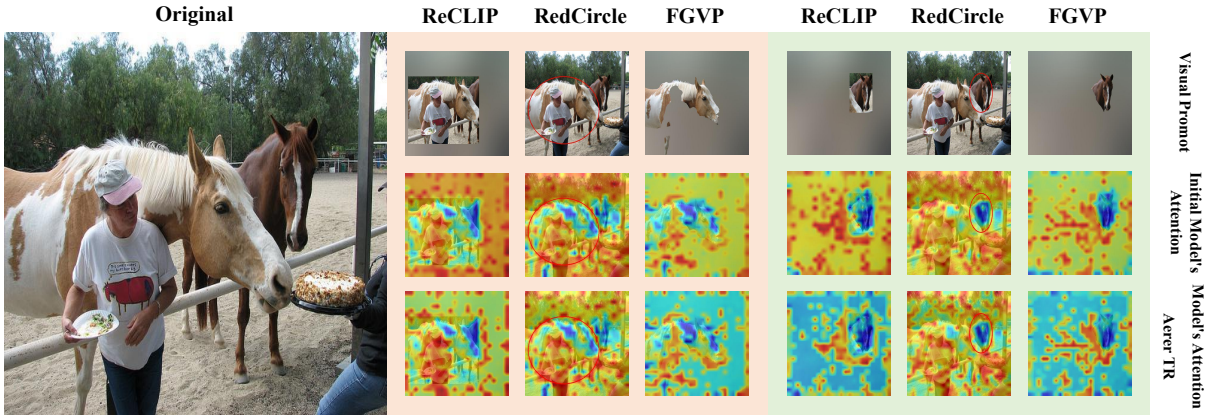

Figure 5. Visualization of the model's attention to background noise in existing visual prompt-based methods, both before and following the implementation of textual restructure.

Table 6. Ablation study of larger text encoder to post-process on RefCOCO while keeping the optimal visual prompt combination, textual restructure, and post-process unchanged.

Method	GPU Memory	Time Cost	RefCOCO		
			Val	TestA	TestB
Normal	14GB	1.7h	64.0%	72.4%	54.9%
Large	17GB	1.9h (+0.2)	**64.1**%	**72.7**%	**55.5**%

semantic decoupling that not only aids in aligning with visual information but also enhances the overall effectiveness of the model.

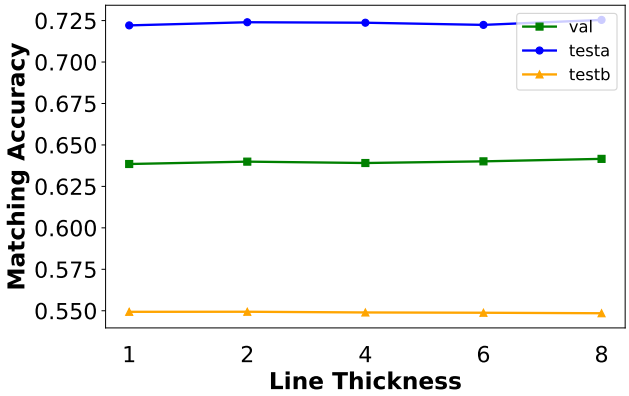

Figure 6. The impact of different line thicknesses on RefCOCO.

4.8 Impact of Mask Preciseness

To demonstrate the impact of different mask precisions on zero-shot performance, we conduct an additional parameter study about mask precision. The parameter "mask preciseness" enables us to adjust the size of the mask around the target by expanding or shrinking it. When the mask preciseness exceeds 1, it indicates an outward expansion. The results in Figure 7 show the consistent performance of the proposed MSDP with different parameter settings.

4.9 Visualization of Model's Attention

We present visual results that compare the model's attention between the existing SOTA visual prompt-based method and the proposed MSDP, as illustrated in Figure 4. These results demonstrate that the simultaneous activation of both text and visual modalities at the instance level improves the model's ability to direct its attention to the target region.

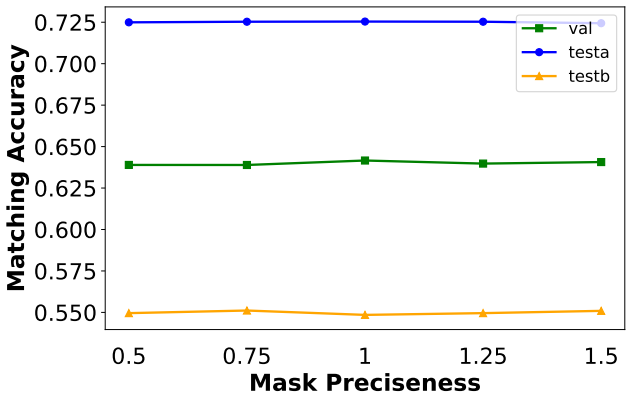

Figure 7. The impact of different mask preciseness on RefCOCO.

4.10 Visualization of the Textual Restructure

To further investigate the impact of textual restructure on the model's attention to background noise, we provide visual results that illustrate the model's attention before and after the addition of textual restructure. The results shown in Figure 5 demonstrate that existing methods that rely solely on visual prompts show noticeable attention to background noise regardless of the size of the proposal. However, the incorporation of textual restructure in any method effectively diminishes the model's attention to background noise.

5 Conclusion

In this paper, we explore the effectiveness of multimodal prompts in activating the instance-level understanding capabilities of VLMs. We propose MSDP, a simple yet effective approach for zero-shot REC that considers the activation of instance-level understanding abilities for each modality. The proposed MSDP integrates two strategies, namely textual restructure and united visual prompts, to activate textual and visual understanding aspects simultaneously. The experiments achieve SOTA performance across multiple datasets, demonstrating that the inherent potential of VLMs is fully realized for instance-level tasks. We aim to explore the activation of VLMs' instance-level understanding in low-quality or occluded scenes for tasks like part detection in future studies.

Acknowledgements

This work is supported by the National Key RD Program of China (2022YFF0712100), NSFC (62276131), Natural Science Foundation of Jiangsu Province of China under Grant (BK20240081).

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: visual question answering. In *ICCV*, pages 2425–2433, Santiago, Chile, 2015.

[2] H. Bangalath, M. Maaz, M. U. Khattak, S. H. Khan, and F. Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. In *NeurIPS*, Orleans, LA, 2022.

[3] B. Bowman, A. Achille, L. Zancato, M. Trager, P. Perera, G. Paolini, and S. Soatto. À-la-carte prompt tuning (APT): combining distinct data via composable prompting. In *CVPR*, pages 14984–14993, BC, Canada, 2023.

[4] D. Chao, Y. Zhang, L. Zhou, and Y. Yang. Enriching category representations with llms towards robust zero-shot out-of-distribution detection. In *ECML-PKDD*, Porto, Portugal, 2025.

[5] Z. Fu, K. Song, L. Zhou, and Y. Yang. Noise-aware image captioning with progressively exploring mismatched words. In *AAAI*, pages 12091–12099, Vancouver, Canada, 2024.

[6] Z. Han, F. Zhu, Q. Lao, and H. Jiang. Zero-shot referring expression comprehension via structural similarity between images and captions. In *CVPR*, pages 14364–14375, Seattle, WA, 2024.

[7] M. Honnibal and M. Johnson. An improved non-monotonic transition system for dependency parsing. In *CEMNLP*, pages 1373–1378, Lisbon, Portugal, 2015.

[8] Z. Huang, A. Zhou, Z. Lin, M. Cai, H. Wang, and Y. J. Lee. A sentence speaks a thousand images: Domain generalization through distilling CLIP with language guidance. In *ICCV*, pages 11651–11661, Paris, France, 2023.

[9] C. Jia, Y. Yang, Y. Xia, Y. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916, Virtual, 2021.

[10] R. Jiang, L. Liu, and C. Chen. Clip-count: Towards text-guided zero-shot object counting. In A. El-Saddik, T. Mei, R. Cucchiara, M. Bertini, D. P. T. Vallejo, P. K. Atrey, and M. S. Hossain, editors, *ACM MM*, pages 4535–4545, Ottawa, Canada, 2023.

[11] J. Ke, J. Wang, J. Chen, I. Jhuo, C. Lin, and Y. Lin. CLIPREC: graph-based domain adaptive network for zero-shot referring expression comprehension. *IEEE Transactions on Multimedia*, 26:2480–2492, 2024.

[12] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W. Lo, P. Dollár, and R. B. Girshick. Segment anything. In *ICCV*, pages 3992–4003, Paris, France, 2023.

[13] J. Li, D. Li, S. Savarese, and S. C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742, Honolulu, Hawaii, 2023.

[14] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. In *ECCV*, pages 740–755, Zurich, Switzerland, 2014.

[15] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In *NeurIPS*, Orleans, LA, 2023.

[16] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, Las Vegas, NV, 2016.

[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, Baltimore, MD, 2021.

[18] R. Schaeffer, B. Miranda, and S. Koyejo. Are emergent abilities of large language models a mirage? In *NeurIPS*, Orleans, LA, 2023.

[19] H. Shen, T. Zhao, M. Zhu, and J. Yin. Groundvlp: Harnessing zero-shot visual grounding from vision-language pre-training and open-vocabulary object detection. In *AAAI*, pages 4766–4775, Vancouver, Canada, 2024.

[20] A. Shtedritski, C. Rupprecht, and A. Vedaldi. What does CLIP know about a red circle? visual prompt engineering for vlms. In *ICCV*, pages 11953–11963, Paris, France, 2023.

[21] A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, and D. Kiela. FLAVA: A foundational language and vision alignment model. In *CVPR*, pages 15617–15629, New Orleans, LA, 2022.

[22] S. Subramanian, W. Merrill, T. Darrell, M. Gardner, S. Singh, and A. Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. In *ACL*, pages 5198–5215, Dublin, Ireland, 2022.

[23] S. Wang, F. Lyu, W. Feng, and S. Wang. Mutatt: Visual-textual mutual guidance for referring expression comprehension. In *ICME*, pages 1–6, London, UK, 2020.

[24] Y. Wang, J. Hu, and L. Shang. Accurate and complete captions for question-controlled text-aware image captioning. In *ICME*, pages 2795–2800, Brisbane, Australia, 2023.

[25] X. Wu, Q. Jiang, Y. Yang, Y. Wu, Q. Chen, and J. Lu. TAI++: text as image for multi-label image classification by co-learning transferable prompt. In *IJCAI*, pages 5226–5234, Jeju, South Korea, 2024.

[26] X. Xu, C. Wu, S. Rosenman, V. Lal, W. Che, and N. Duan. Bridgetower: Building bridges between encoders in vision-language representation learning. In *AAAI*, pages 10637–10647, Washington D.C., 2023.

[27] L. Yang, Y. Wang, X. Li, X. Wang, and J. Yang. Fine-grained visual prompting. In *NeurIPS*, Orleans, LA, 2023.

[28] S. Yang, G. Li, and Y. Yu. Dynamic graph attention for referring expression comprehension. In *ICCV*, pages 4643–4652, Seoul, Korea, 2019.

[29] Y. Yang, C. Zhang, Y. Xu, D. Yu, D. Zhan, and J. Yang. Rethinking label-wise cross-modal retrieval from A semantic sharing perspective. In *IJCAI*, Montreal, Canada, 2021.

[30] Y. Yang, H. Wei, H. Zhu, D. Yu, H. Xiong, and J. Yang. Exploiting cross-modal prediction and relation consistency for semisupervised image captioning. *IEEE Transactions on Cybernetics*, 54(2):890–902, 2022.

[31] Y. Yang, Y. Zhang, X. Song, and Y. Xu. Not all out-of-distribution data are harmful to open-set active learning. In *NeurIPS*, Orleans, LA, 2023.

[32] Y. Yang, J. Guo, G. Li, L. Li, W. Li, and J. Yang. Alignment efficient image-sentence retrieval considering transferable cross-modal representation learning. *Frontiers of Computer Science*, 18(3):181335, 2024.

[33] Y. Yang, W. Xi, L. Zhou, and J. Tang. Rebalanced vision-language retrieval considering structure-aware distillation. *IEEE Transactions on Image Processing*, 33:6881–6892, 2024.

[34] Y. Yao, A. Zhang, Z. Zhang, Z. Liu, T. Chua, and M. Sun. CPT: colorful prompt tuning for pre-trained vision-language models. *CoRR*, abs/2109.11797, 2021.

[35] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85, Amsterdam, Netherlands, 2016.

[36] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, and T. L. Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, pages 1307–1315, Salt Lake City, UT, 2018.

[37] A. Zareian, K. D. Rosa, D. H. Hu, and S. Chang. Open-vocabulary object detection using captions. In *CVPR*, pages 14393–14402, Virtual, 2021.

[38] A. Zeng, M. Attarian, B. Ichter, K. M. Choromanski, A. Wong, S. Welker, F. Tombari, A. Purohit, M. S. Ryoo, V. Sindhwani, J. Lee, V. Vanhoucke, and P. Florence. Socratic models: Composing zero-shot multimodal reasoning with language. In *ICLR*, Kigali, Rwanda, 2023.

[39] S. Zhao, Z. Zhang, S. Schulter, L. Zhao, B. Vijay Kumar, A. Stathopoulos, M. Chandraker, and D. N. Metaxas. Exploiting unlabeled data with vision and language models for object detection. In *ECCV*, pages 159–175, Tel Aviv, Israel, 2022.

[40] J. Zhuang, J. Hu, L. Mu, R. Hu, X. Liang, J. Ye, and H. Hu. FALIP: visual prompt as foveal attention boosts CLIP zero-shot performance. In *ECCV*, pages 236–253, Milan, Italy, 2024.