AI Open 5 (2024) 30–38

Contents lists available at ScienceDirect

AI Open

journal homepage: www.keaipublishing.com/en/journals/ai-open

Full length article

CPT: Colorful Prompt Tuning for pre-trained vision-language models

Yuan Yao [a,1], Ao Zhang [b,1], Zhengyan Zhang [a], Zhiyuan Liu [a,*], Tat-Seng Chua [b], Maosong Sun [a]

[a] Department of Computer Science and Technology, Institute for Artificial Intelligence, Tsinghua University, Beijing, China
[b] Sea-NExT Joint Lab, Singapore, School of Computing, National University of Singapore, Singapore

ARTICLE INFO

Keywords:
Vision-language pre-training models
Prompt tuning

ABSTRACT

Vision-Language Pre-training (VLP) models have shown promising capabilities in grounding natural language in image data, facilitating a broad range of cross-modal tasks. However, we note that there exists a significant gap between the objective forms of model pre-training and fine-tuning, resulting in a need for large amounts of labeled data to stimulate the visual grounding capability of VLP models for downstream tasks. To address the challenge, we present **C**olor-based **P**rompt **T**uning (CPT), a novel paradigm for tuning VLP models, which reformulates visual grounding into a fill-in-the-blank problem with color-based co-referential markers in image and text, maximally mitigating the gap. In this way, CPT enables strong few-shot and even zero-shot visual grounding capabilities of VLP models. Comprehensive experimental results show that CPT achieves state-of-the-art performance on zero/few-shot visual grounding (e.g., 75.1 zero-shot accuracy in RefCOCO evaluation), outperforming fine-tuned and other prompt-tuned models by a large margin. Moreover, CPT can also be easily extended to achieve promising zero/few-shot performance on other vision-language tasks, such as visual relation detection, visual commonsense reasoning and visual question answering. We make the data and codes publicly available at https://github.com/thunlp/CPT.

1. Introduction

Grounding natural language in fine-grained image regions is fundamental to a broad range of vision-language (VL) tasks, such as robotic navigation (Tellex et al., 2011; Anderson et al., 2018b), visual question answering (Antol et al., 2015; Anderson et al., 2018a), visual dialogue (Das et al., 2017), and visual commonsense reasoning (Zellers et al., 2019). Recently Vision-Language Pre-training (VLP) models have shown promising capabilities in visual grounding. Typically, generic cross-modal representations are first pre-trained on large-scale image–text data in a self-supervised fashion, and then fine-tuned to adapt to downstream tasks (Lu et al., 2019; Chen et al., 2020; Zhang et al., 2021). This *pre-training-then-fine-tuning* paradigm has greatly pushed forward the state-of-the-art of many cross-modal tasks.

Despite the success, we note that there exists a significant gap between the objective forms of pre-training and fine-tuning of VLP models. As illustrated in Fig. 1, during pre-training, most VLP models are optimized based on the masked language modeling objective, trying to recover the masked token from the cross-modal context. However, during fine-tuning, downstream tasks are usually conducted by classifying unmasked tokens into semantic labels, where task-specific parameters are introduced. The gap hinders the effective adaptation of VLP models to downstream tasks. As a result, a large amount of labeled data is typically required to stimulate the visual grounding capabilities of VLP models for downstream tasks.

In this work, inspired by the recent progress in prompting pre-trained language models (Schick and Schütze, 2021; Liu et al., 2021), we present **C**olor-based **P**rompt **T**uning (CPT), a novel paradigm for tuning VLP models. The key insight is that by adding color-based co-referential markers in both image and text, visual grounding can be reformulated into a fill-in-the-blank problem, maximally mitigating the gap between pre-training and fine-tuning. As shown in Fig. 1, to ground natural language in image data, CPT consists of two components: (1) a *visual sub-prompt* that uniquely marks image regions with colored blocks, and (2) a *textual sub-prompt* that puts the query text into a color-based query template. Explicit grounding to image regions can then be naturally achieved by recovering the corresponding color text from the masked token in the query template. To search for high-quality cross-modal prompt configurations (i.e., visual appearances and texts of colors), we present a principled approach that probes the strongly activated cross-modal signals in VLP models for prompt construction.

By mitigating the gap from pre-training, CPT enables strong few-shot and even zero-shot visual grounding capabilities of VLP models. Experimental results show that CPT achieves state-of-the-art performance on zero/few-shot visual grounding, outperforming fine-tuned,

* Corresponding author.
E-mail address: liuzy@tsinghua.edu.cn (Z. Liu).
[1] Indicates equal contribution.

https://doi.org/10.1016/j.aiopen.2024.01.004
Received 6 August 2023; Received in revised form 14 December 2023; Accepted 29 January 2024
Available online 1 February 2024

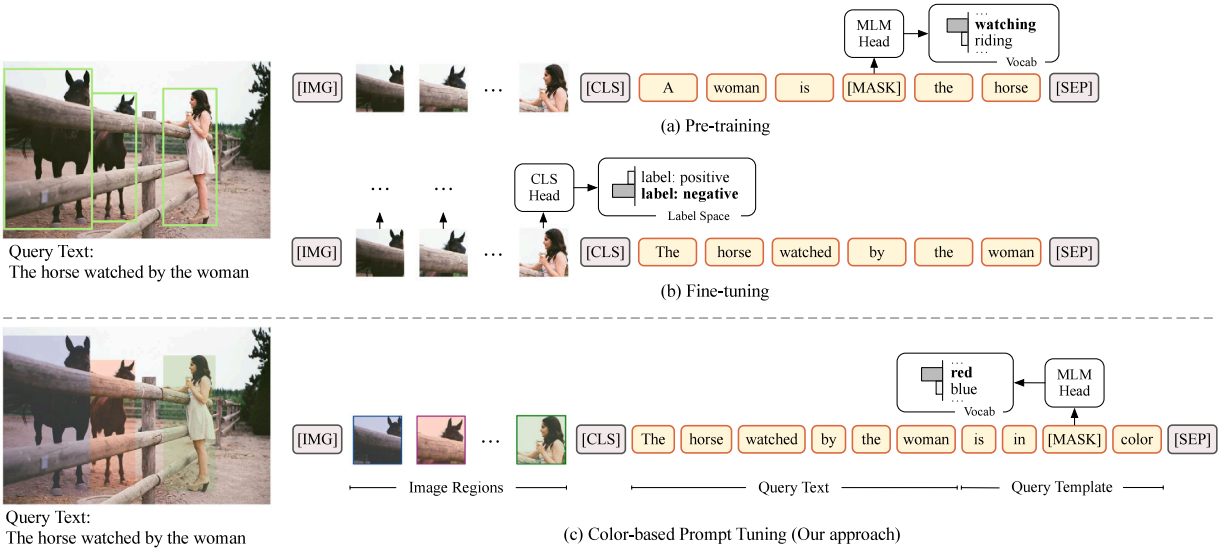


Fig. 1. Illustration of (a) pre-training for VLP models with masked language modeling (MLM) head, (b) vanilla fine-tuning with new classification (CLS) head, and (c) our color-based prompt tuning (CPT) framework that reformulates visual grounding into a fill-in-the-blank problem with reused MLM head.

other prompt-tuned and even task-specific models by a large margin. The consistent tuning approach of CPT can also bring more stable adaptation performance. For example, CPT achieves 73.8% reduction of standard deviation over fine-tuning with one-shot in RefCOCO evaluation. Moreover, for other VL tasks that require or benefit from grounded inputs, CPT can also be useful in prompting VLP models to explicitly indicate object positions. We show that CPT can be easily extended to achieve promising zero/few-shot performance on other VL tasks, such as visual relation detection, visual commonsense reasoning and visual question answering.

Our contributions are summarized as threefold: (1) We present a novel color-based prompt tuning framework for VLP models, which reformulates visual grounding into a fill-in-the-blank problem using color-based co-referential markers. (2) We present a principled approach to search for high-quality cross-modal prompt configurations. (3) We conduct comprehensive experiments which demonstrate the effectiveness of the proposed model.

2. Preliminary

In this work, we adopt VinVL (Zhang et al., 2021) as the backbone, which is a representative VLP model that achieves strong performance on various tasks. We briefly introduce the pre-training and vanilla fine-tuning procedure of the model.

Vision-language Pre-training. Given an image–text pair (I, t), a set of objects $\{v_1, v_2, \ldots, v_n\}$ is first detected from the image via object detectors. Then image and text are transformed into a sequence of tokens $\{\texttt{[IMG]}, v_1, v_2, \ldots, v_n, \texttt{[CLS]}, w_1, w_2, \ldots, w_m, \texttt{[SEP]}\}$, where $\{w_1, w_2, \ldots, w_m\}$ are text tokens of t, and [IMG], [CLS] and [SEP] are special tokens. The input representations are fed into Transformers (Vaswani et al., 2017) to produce the hidden representations $\{\mathbf{h}_{\texttt{[IMG]}}, \mathbf{h}_v^1, \mathbf{h}_v^2, \ldots, \mathbf{h}_v^n, \mathbf{h}_{\texttt{[CLS]}}, \mathbf{h}_w^1, \mathbf{h}_w^2, \ldots, \mathbf{h}_w^m, \mathbf{h}_{\texttt{[SEP]}}\}$.

To mine self-supervised signals, there are two widely adopted pre-training tasks, including masked language modeling (MLM) and image–text matching (ITM). The MLM pre-training task randomly replaces some text tokens with a special [MASK] token, and recovers the masked token from $\mathbf{h}_{\texttt{[MASK]}}$ using an MLM head. The ITM pre-training task discriminates whether a given image and text pair matches based on $\mathbf{h}_{\texttt{[CLS]}}$ using an ITM head.

Vanilla Fine-tuning. Here we take visual grounding as an example to illustrate the fine-tuning procedure. Given an image I and a query text q, visual grounding aims to locate the corresponding region in I. A common practice for the task is to first detect a set of region proposals via object detectors, and then classify or rank the proposals to select the target region (Lu et al., 2019; Chen et al., 2020). Specifically, the image and text inputs are fed into the pre-trained Transformers, and then the hidden representations of region proposals are optimized via classification or ranking loss, where new task-specific parameters are introduced. As a result, fine-tuned VLP models need large amounts of labeled data to stimulate the visual grounding capability.

3. Cross-modal prompt tuning (CPT)

To establish fine-grained connections between image regions and text in a data-efficient way, a good cross-modal prompt tuning framework should take full advantage of co-referential signals from both modalities, and prevent the gap between pre-training and tuning. To this end, CPT reformulates visual grounding into a fill-in-the-blank problem, as shown in Fig. 1. Specifically, CPT consists of two components: (1) a *visual sub-prompt* that uniquely marks the image regions with colored blocks, and (2) a *textual sub-prompt* that puts the query text into a color-based template. Equipped with CPT, it is straightforward for VLP models to ground the query text by filling in the mask with the color text of the target image region, where the objective form is identical to pre-training.

Visual Sub-prompt. Given an image and its region proposals $\mathcal{R} = \{v_1, v_2, \ldots, v_n\}$, visual sub-prompt aims to uniquely mark the image regions with natural visual makers. Interestingly, we note that colored bounding boxes are widely used to mark objects in images *for visualization* in the literature. Inspired by this, we bridge the image regions and query text through a set of colors $\mathcal{C}$, where each color $c_i = (c_v^i, c_w^i) \in \mathcal{C}$ is defined by its visual appearance c_v^i (e.g., RGB (255, 0, 0)) and color text c_w^i (e.g., *red*). Then we mark each region proposal v_i in the image with a unique color c_v^i, resulting in a set of colored image proposals $\Psi(\mathcal{R}; \mathcal{C})$, where $\Psi(\cdot)$ denotes visual sub-prompt.

In principle, there are multiple plausible choices to mark the regions with colors, including bounding boxes, solid blocks, or solid segmentation masks. In our experiments, we find that coloring the object with solid blocks and segmentation masks yields better results than bounding boxes, since solid colors are more obvious prompting signals for VLP models. Note that the addition of visual sub-prompt to the raw image does not change the architecture or parameters of VLP models.

Textual Sub-prompt. Given the image regions marked by visual sub-prompt, textual sub-prompt aims to prompt VLP models to resolve

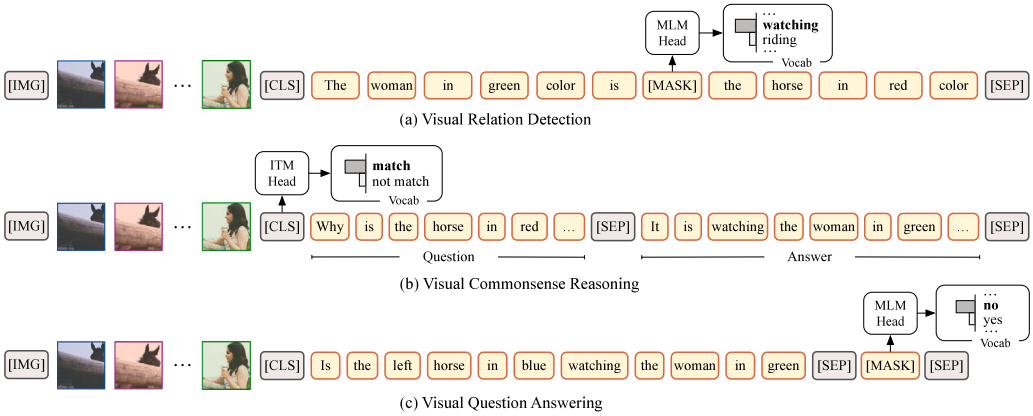


Fig. 2. CPT framework for other vision-language tasks with reused pre-trained heads. ITM: image–text matching.

Algorithm 1 Cross-modal Prompt Tuning

Require: $\mathcal{C}$: Set of colors
1: Visual sub-prompt: Mark the image regions $\mathcal{R}$ with color set $\mathcal{C}$ as $\Psi(\mathcal{R};\mathcal{C})$
2: Textual sub-prompt: Put the query into template as $\mathcal{T}(q)$
3: Infer the prediction $P(v = v_i|\mathcal{R}, q)$ from the visual and textual sub-prompts as Eq (1)
4: **if** zero-shot **then**
5: Take $P(v = v_i|\mathcal{R}, q)$ as the final prediction
6: **else**
7: Supervise the model using labeled data points $\mathcal{D}$: $\mathcal{L} = -\sum_{(\mathcal{R},q,v^\star)\in\mathcal{D}} \log P(v^\star|\mathcal{R}, q)$

the query text. Specifically, the query text q (e.g., "*the horse watched by the woman*") is transformed into a fill-in-the-blank query using a template $\mathcal{T}(\cdot)$ as follows:

$\mathcal{T}(q) =$ `[CLS]` q `is in [MASK] color [SEP]`

In this way, VLP models are prompted to decide the color of which region is more appropriate to fill in the mask (e.g., *red* or *blue*) as follows:

$$P(v = v_i|\mathcal{R}, q) = P(\texttt{[MASK]} = c_w^i|\Psi(\mathcal{R};\mathcal{C}), \mathcal{T}(q)) = \frac{\exp(\mathbf{h}_{\texttt{[MASK]}}^\top \mathbf{c}_w^i)}{\sum_{c_j\in\mathcal{C}} \exp(\mathbf{h}_{\texttt{[MASK]}}^\top \mathbf{c}_w^j)}, \quad (1)$$

where v is the target region, $\mathbf{c}_w^i$ is the embedding of c_w^i in the pre-trained MLM head. Note that the procedure does not introduce any new parameters, and also mitigates the gap between pre-training and tuning, and therefore improves the data efficiency of tuning VLP models.

Equipped with CPT, VLP models can readily perform zero-shot visual grounding without any labeled data, since the cross-modal representations of colors and their combination with other concepts (e.g., objects, attributes and relations) have been well learned by VLP models during pre-training. When a few labeled instances $\mathcal{D}$ are available, VLP models can be further tuned by CPT using the objective: $\mathcal{L} = -\sum_{(\mathcal{R},q,v^\star)\in\mathcal{D}} \log P(v^\star|\mathcal{R}, q)$. We provide the pseudo-code of the above procedure in Algorithm 1.

Cross-Modal Prompt Search. Previous works in textual prompt tuning show that prompt configurations (e.g., textual templates) can have a significant influence on the performance (Jiang et al., 2020). It is therefore desirable to search for cross-modal prompt configurations (i.e., color set $\mathcal{C}$) in a principled approach. Intuitively, $\mathcal{C}$ should consist of colors to which VLP models are the most sensitive. To obtain a color $c_i = (c_v^i, c_w^i)$, a naive approach is to adopt the most frequent color text in the pre-training text as c_w^i, and its standard RGB as c_v^i (e.g., $c_i = ((255, 0, 0), \textit{red})$). However, this solution is sub-optimal, since it determines the color text without considering its visual appearance, and the visual appearance of a color in real-world images often differs from its standard RGB.

To address the challenge, we present a principled algorithm that probes the strongly activated cross-modal signals in VLP models for prompt construction. Specifically, we first identify a candidate set of color texts $\hat{\mathcal{C}}_w$ and visual appearances $\hat{\mathcal{C}}_v$. For each visual appearance candidate $\hat{c_v} \in \hat{\mathcal{C}}_v$, we feed into VLP models a pseudo-data instance consisting of a pure colored block of $\hat{c_v}$ and a text: "`[CLS]` a photo in `[MASK]` color `[SEP]`". Then we compute the decoding score $s(\hat{c_v}, \hat{c_w})$ for each color text candidate $\hat{c_w} \in \hat{\mathcal{C}}_w$ as in Eq. (1), where a larger decoding score indicates higher correlation between $\hat{c_v}$ and $\hat{c_w}$. Finally, the color set is obtained from the visual appearances and text pairs that achieve the highest correlation scores. We refer readers to the appendix for the pseudo-code. In practice, to make the raw content of the colored image regions available to VLP models, a transparency hyperparameter $\alpha \in (0, 1)$ is applied to color visual appearances.

4. CPT for other VL tasks

By providing natural cross-modal co-referential signals, CPT can also enable zero/few-shot capabilities of VLP models on other challenging VL tasks that require fine-grained object-centric reasoning, such as visual relation detection (VRD), visual commonsense reasoning (VCR), and visual question answering (VQA), as shown in Fig. 2. CPT can be useful in explicitly informing object positions in grounded task inputs in a prompting paradigm. Here we briefly introduce the overall framework for each task, and refer readers to the appendix for more model details.

Visual Relation Detection. Given a pair of objects in an image, VRD aims to identify their semantic relations (Krishna et al., 2017). To explicitly indicate the object positions without changing model architectures, we first mark the image regions with visual sub-prompt, and put the object pair in the query template as: "`[CLS]` The s_w in c_w^i color is `[MASK]` the o_w in c_w^j color `[SEP]`", where s_w is the subject text, o_w is the object text, and c_w^i and c_w^j are the corresponding color texts. Then VLP models are prompted to recover the relation texts from masked tokens in the template. Besides better data efficiency, another advantage of CPT is that the semantic labels can be produced from open vocabularies, instead of fixed label sets.

Visual Commonsense Reasoning. Given a question, VCR aims to select the answer (and rationale) sentences through commonsense reasoning (Zellers et al., 2019). The task refers visual objects in text using object labels and image regions. To provide grounding clues for objects in text, we mark image regions with visual sub-prompt. Then we concatenate the question and answer, and decorate the resultant

Table 1
Accuracies of grounding referring expressions. Ext.: extra data augmentation or heuristic rules. FT: vanilla fine-tuning, FT-ATT: attention weights of fine-tuned VLP models, Blk: colored block, Seg: colored segmentation mask, Aug: extra data augmentation. We report mean (and standard deviation) performance over 5 random splits.

Shot	Model	Ext.	RefCOCO			RefCOCO+			RefCOCOg	
			val	testA	testB	val	testA	testB	val	test
0	Random		15.9 (0.2)	19.4 (0.6)	13.4 (0.4)	16.1 (0.1)	13.3 (0.6)	20.0 (0.2)	18.8 (0.4)	19.2 (0.3)
	FT-ATT (Cao et al., 2020)		26.9	26.1	30.6	27.1	26.7	30.8	36.6	36.3
	VC[b] (Zhang et al., 2018)	✓	–	33.3	30.1	–	34.6	31.6	–	–
	ARN[b] (Liu et al., 2019a)	✓	34.3	36.4	33.1	34.5	36.0	33.8	–	–
	KPRN[b] (Liu et al., 2019b)	✓	35.0	34.7	37.0	36.0	35.2	37.0	–	–
	DTWREG[b] (Sun et al., 2021)	✓	39.2	41.1	37.7	39.2	40.1	38.1	–	–
	GPV (Gupta et al., 2022)	✓	44.6	41.4	47.1	42.9	39.1	47.6	52.1	52.3
	ReCLIP[b] (Subramanian et al., 2022)	✓	45.8	46.1	47.1	47.9	50.1	45.1	59.3	59.0
	Pseudo-Q[b] (Jiang et al., 2022)	✓	56.0	58.3	54.1	38.9	45.1	32.1	46.3	47.4
	CPT-Blk (ours)		26.9	27.5	27.4	25.4	25.0	27.0	32.1	32.3
	CPT-Seg (ours)		32.2	36.1	30.3	31.9	35.2	28.8	36.7	36.5
	CPT-Aug (ours)	✓	**69.8**	**75.1**	**62.6**	**57.7**	**65.0**	**48.2**	**63.9**	**63.3**
1	FT (Zhang et al., 2021)		16.5 (4.9)	12.0 (6.6)	23.5 (5.7)	22.2 (7.6)	20.6 (9.3)	25.7 (5.2)	26.9 (8.4)	26.9 (8.1)
	FT-ATT (Cao et al., 2020)		26.9 (0.6)	26.3 (0.7)	30.9 (0.7)	26.7 (0.3)	26.1 (0.7)	30.6 (0.1)	36.1 (0.2)	36.6 (0.2)
	GPV (Gupta et al., 2022)	✓	48.7 (3.7)	47.6 (5.8)	49.0 (1.7)	47.1 (2.6)	45.0 (3.7)	**49.3** (1.1)	54.6 (2.4)	55.0 (2.1)
	CPT-Blk (ours)		34.1 (1.3)	37.7 (1.7)	32.2 (1.5)	35.9 (4.1)	40.4 (5.4)	32.2 (2.6)	39.7 (3.4)	39.9 (3.0)
	CPT-Seg (ours)		37.2 (0.9)	41.5 (1.5)	33.2 (1.7)	37.9 (4.0)	42.3 (5.9)	33.9 (2.4)	43.1 (2.9)	43.4 (3.1)
	CPT-Aug (ours)	✓	**70.2** (0.6)	**75.5** (0.7)	**63.7** (0.7)	**57.6** (0.4)	**65.2** (0.2)	48.5 (0.6)	**63.7** (0.4)	**63.9** (0.4)
16	FT (Zhang et al., 2021)		39.8 (4.2)	45.5 (5.0)	34.9 (3.0)	41.8 (3.0)	47.3 (3.1)	36.2 (2.3)	47.5 (4.1)	47.8 (4.7)
	FT-ATT (Cao et al., 2020)		29.8 (0.5)	31.4 (1.1)	32.1 (0.1)	30.3 (1.3)	32.2 (1.9)	32.2 (0.6)	37.7 (0.8)	38.1 (0.6)
	GPV (Gupta et al., 2022)	✓	56.5 (0.5)	58.7 (1.5)	53.0 (0.6)	57.1 (1.4)	59.8 (1.7)	**53.4** (1.1)	60.2 (0.5)	60.4 (0.5)
	CPT-Blk (ours)		44.8 (3.3)	51.4 (4.1)	38.2 (2.3)	41.5 (1.3)	48.2 (2.1)	34.7 (0.9)	47.8 (2.1)	48.2 (2.8)
	CPT-Seg (ours)		45.3 (1.8)	53.3 (3.0)	37.5 (1.3)	44.8 (0.9)	52.5 (1.2)	36.6 (1.2)	51.0 (2.6)	51.4 (2.8)
	CPT-Aug (ours)	✓	**71.2** (0.7)	**76.8** (1.0)	**64.9** (0.5)	**58.2** (0.3)	**65.7** (0.4)	49.0 (0.3)	**64.1** (0.6)	**64.0** (0.6)
Oracle (full)[a] (Zhang et al., 2021)		81.8	87.2	74.3	74.5	80.8	64.3	74.6	75.7	

[a] Fine-tuned on the full training set.
[b] Task-specific models.

text with textual sub-prompt, where each object in text is referred by the corresponding color (e.g., *in red*), as shown in Fig. 2.

Most VLP models include an image–text matching (ITM) pre-training task, where an ITM head is used to discriminate whether an image matches a given text (see Section 2). Intuitively, a question concatenated with the correct answer can better describe the image than the question concatenated with a wrong answer, and therefore should be assigned with a higher ITM score. Therefore, ITM head can be used in the same form as pre-training to judge whether the image and question–answer text match (i.e., correctly answer) for zero/few-shot VCR. We find that this ITM-based approach yields better performance in dealing with sentence-level VCR answers than MLM head in experiments. When a few training instances are available, we can further optimize the VLP model to classify the image–text pair into the ITM vocabulary {*matched*, *not matched*} in the identical objective as pre-training.

Visual Question Answering. We further investigate whether CPT can benefit tasks that do not require explicit object modeling. We are interested in the question: given a grounded question, where the object positions are provided, can CPT stimulate pre-trained capabilities of VLP models for zero/few-shot VQA? To obtain high-quality object grounding results for CPT, we use a strong visual grounding model pre-trained on a large collection of labeled datasets (Yao et al., 2022). Then we mark the image regions and object text using visual and textual sub-prompts. Finally, we concatenate the question with "`[SEP] [MASK]`", and reuse the MLM head to predict answers from the mask.

5. Experiments

We empirically evaluate CPT on different VL tasks in zero/few-shot scenarios. In our experiments, we use the same color configurations for different tasks. We refer readers to the appendix for implementation and dataset details.

5.1. Visual grounding experiments

Datasets. We adopt three widely used visual grounding datasets, including RefCOCO (Yu et al., 2016), RefCOCO+ (Yu et al., 2016) and RefCOCOg (Mao et al., 2016). To better approximate the few-shot scenario where only a few labeled instances are available, following Gao et al. (2021a), we use a few-shot validation set (consisting of 16 instances) for all experiments.

Evaluation Metrics. Following Lu et al. (2019), we adopt the accuracy of grounding results as the evaluation metrics. An expression is considered correctly grounded if the IoU of the top predicted region and the ground truth is greater than 0.5. Moreover, since model training on limited data can suffer from instability, following Dodge et al. (2020) and Gao et al. (2021a), we report average results over 5 random training set splits, as well as the standard deviation. For fair comparisons, the training and validation sets are identical for our baselines and CPT.

Baselines. We compare our model with a series of strong baselines. (1) Vanilla fine-tuning (FT) for VinVL (Zhang et al., 2021). This model adopts the same backbone as CPT, and serves as the most direct baseline. (2) Attention weights of fine-tuned VinVL model (FT-ATT). Previous works show that the attention weights of VLP models are strong grounding indicators (Cao et al., 2020; Chen et al., 2020). Following these works, we score each image region using the average text-to-image attention weights from all text tokens in the query across all attention heads.[2] (3) Task-specific visual grounding models. We compare with state-of-the-art models tailored for zero-shot visual grounding (Zhang et al., 2018; Liu et al., 2019a,b; Sun et al., 2021; Subramanian et al., 2022; Jiang et al., 2022). These works typically utilize extra data augmentation, such as image-level ground-truth referring expressions (Zhang et al., 2018; Liu et al., 2019a,b; Sun

[2] We also experiment with maximum attention score from text tokens or attention heads, or image-to-text attentions, and adopt the best practice.

Table 2
Results of VRD on visual genome test set.

Shot	Model	VRD			
		R@50	R@100	mR@50	mR@100
0	Random	1.5 (0.0)	1.8 (0.1)	1.2 (0.1)	1.6 (0.1)
	GPV (Gupta et al., 2022)	2.2	2.9	1.4	2.1
	CPT	**29.3**	**30.5**	**13.0**	**14.5**
1	FT (Zhang et al., 2021)	4.1 (0.1)	4.7 (0.0)	6.7 (0.3)	7.6 (0.4)
	GPV (Gupta et al., 2022)	11.6 (1.2)	16.2 (2.9)	7.1 (1.1)	10.5 (1.7)
	CPT	**18.0** (2.8)	**20.0** (3.0)	**23.9** (0.3)	**26.3** (0.3)
4	FT (Zhang et al., 2021)	7.3 (1.5)	7.9 (1.7)	11.8 (1.0)	13.2 (0.9)
	GPV (Gupta et al., 2022)	12.1 (2.3)	16.7 (3.3)	10.3 (0.8)	15.6 (2.1)
	CPT	**17.7** (0.6)	**19.3** (0.6)	**28.5** (1.5)	**32.1** (1.0)
16	FT (Zhang et al., 2021)	10.4 (0.7)	11.2 (0.8)	19.7 (0.1)	21.7 (0.1)
	GPV (Gupta et al., 2022)	13.7 (1.0)	19.2 (1.4)	15.3 (1.0)	24.1 (1.3)
	CPT	**18.4** (1.0)	**20.0** (1.1)	**32.5** (0.5)	**36.1** (0.6)
32	FT (Zhang et al., 2021)	11.7 (0.2)	12.4 (0.3)	22.0 (0.1)	24.1 (0.0)
	GPV (Gupta et al., 2022)	16.7 (0.6)	**23.0** (0.8)	16.7 (1.6)	24.8 (2.0)
	CPT	**20.8** (0.1)	22.3 (0.1)	**34.0** (0.1)	**37.7** (0.3)
Oracle (full) (Zhang et al., 2021)		65.1	67.4	20.6	22.5

et al., 2021), generated pseudo-queries (Jiang et al., 2022), or heuristic rules for resolving referring expressions (Subramanian et al., 2022). For example, Pseudo-Q (Jiang et al., 2022) generates pseudo-queries using templates according to the detected nouns, attributes and spatial relations. (4) General prompt-based models. GPV (Gupta et al., 2022) is a strong general purpose model that can output object positions and text tokens to address various tasks in a prompting fashion. GPV is pre-trained with augmented pseudo-queries which are generated in a similar way to Pseudo-Q (Jiang et al., 2022). We evaluate three variants of our model: CPT-Blk uses colored blocks as visual sub-prompt, and CPT-Seg leverages colored segmentation masks. CPT-Aug further pre-trains CPT-Seg using pseudo-queries from Pseudo-Q (Jiang et al., 2022) with Eq. (1).

Results. From Table 1 we observe that: (1) CPT consistently outperforms the fine-tuning baseline by a large margin across different datasets and shot settings. For example, using colored blocks as visual sub-prompts, CPT-Blk achieves 17.3 absolute accuracy points improvement on average with one shot in RefCOCO evaluation. This indicates that CPT can effectively improve sample efficiency in tuning VLP models. (2) Coloring objects with segmentation masks (CPT-Seg) achieves even better results than blocks. The reason is that solid colors that fit the outlines of objects are more common in real-world images, making CPT-Seg more natural visual sub-prompts. (3) CPT achieves substantially more stable performance than fine-tuning. For example, CPT-Seg achieves 76.2% reduction of standard deviation with one shot in RefCOCO. This shows that a coherent tuning approach from pre-training can lead to substantially more stable few-shot adaptation. (4) With simple data augmentation, CPT-Aug achieves state-of-the-art performance on zero/few-shot visual grounding, outperforming strong prompt-tuned and task-specific models. Notably, CPT-Aug achieves 75.1 zero-shot grounding accuracy on RefCOCO testA set, outperforming the previous state-of-the-art by 16.8 accuracy points. The reason is that CPT more naturally connects visual and text signals with color-based prompts, and therefore maximally stimulates the visual grounding capabilities of VLP models.

5.2. Experiments on other VL tasks

We further evaluate CPT on visual relation detection, visual commonsense reasoning and question answering.

Visual Relation Detection. We adopt the widely used Visual Genome dataset (Krishna et al., 2017), which contains 50 visual relation types. During training, K labeled instances are provided for each relation. Since Visual Genome does not provide segmentation masks, we use colored blocks in visual sub-prompt. For GPV, following Subramanian et al. (2022) and Yao et al. (2021), the query subject and object are cropped in the image to indicate their positions, and the model is prompted with "what is the relation between s_w and o_w" to decode the relation. FT-ATT cannot handle VRD which requires producing relation labels. Following Chen et al. (2019), we use recall@N (R@N) and mean recall@N (mR@N) over different relations as evaluation metrics.

From Table 2 we observe that: (1) CPT outperforms baselines in different shot settings and metrics. For example, using one shot, CPT outperforms fine-tuning by 15.3 points on R@100 and 18.7 points on mR@100, showing reasonable performance on both common relations and long-tail relations. (2) We note while the macro performance (mR@N) of CPT monotonically increases as the shot number grows, the micro results (R@N) drop first in 1- and 4-shot settings. This is due to the distribution gap between the balanced training set (i.e., K shot for each relation) and long-tail test set. Since the relations in pre-training data also follow a long-tail distribution, CPT can achieve a high starting point for micro performance.

Visual Commonsense Reasoning and Visual Question Answering. We adopt the popular VCR dataset (Zellers et al., 2019) for visual commonsense reasoning, and GQA dataset (Hudson and Manning, 2019) for visual question answering. We report the accuracy of selecting answers and rationales for VCR, and accuracy of answers for GQA. More shots are provided due to the difficulty of the tasks. For visual sub-prompts we adopt segmentation masks provided by the VCR dataset, and bounding boxes from Yao et al. (2022) on GQA. In experiments, we find that while providing reasoning clues, colors in prompt templates can sometimes disturb image understanding. To address the issue, we simply use the weighted average score given by CPT equipped with and without colors. GPV is prompted with the question concatenated with each candidate answer sentence, and decodes *yes/no* for answer selection in VCR. Since the answers of GQA is typically short, GPV directly decodes the answers based on the prompting question.

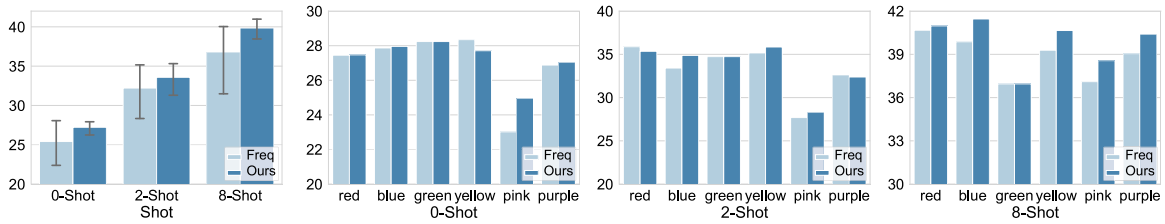
From Table 3 we can see that, CPT can significantly improve the data efficiency of VLP models for visual commonsense reasoning and visual question answering. Notably, the zero-shot performance of CPT even surpasses vanilla fine-tuning trained with 128 shots on both datasets. This shows that CPT can effectively prompt VLP models to handle both sentence-level and token-level answers. In comparison, it can be challenging for GPV to deal with sentence-level answers in question answering. In addition, ensembling color-free prompt templates helps alleviate the color disturbance problem. Removing the color-free prompt templates leads to 2.3 and 1.8 points degradation in 16-shot setting on VCR (Q → AR) and GQA respectively.

Table 3
Results on VCR validation set and GQA test-dev set.

Shot	Model	VCR			GQA
		Q → A	QA → R	Q → AR	test-dev
0	Random	25.0	25.0	6.3	0.1
	GPV (Gupta et al., 2022)	29.1	26.8	8.1	34.6
	CPT	**43.8**	**39.0**	**17.8**	**36.0**
4	FT (Zhang et al., 2021)	31.5 (5.6)	30.2 (6.0)	10.4 (3.0)	12.2 (4.4)
	GPV (Gupta et al., 2022)	29.9 (0.3)	28.3 (0.3)	8.8 (0.1)	34.6 (0.0)
	CPT	**44.4** (0.3)	**41.4** (1.0)	**18.1** (1.6)	**36.7** (5.9)
16	FT (Zhang et al., 2021)	32.1 (7.9)	35.7 (1.7)	12.8 (4.0)	17.5 (2.7)
	GPV (Gupta et al., 2022)	29.6 (0.5)	28.0 (0.4)	8.7 (0.2)	34.6 (0.0)
	CPT	**45.3** (1.6)	**41.2** (1.8)	**19.4** (1.3)	**43.6** (3.4)
64	FT (Zhang et al., 2021)	41.1 (2.6)	38.8 (2.0)	14.6 (2.0)	22.1 (1.2)
	GPV (Gupta et al., 2022)	30.3 (0.7)	28.4 (0.4)	8.8 (0.3)	34.6 (0.0)
	CPT	**45.7** (0.8)	**42.5** (0.7)	**19.2** (1.4)	**50.9** (1.1)
128	FT (Zhang et al., 2021)	43.0 (2.5)	39.7 (4.8)	14.6 (1.5)	23.7 (0.7)
	GPV (Gupta et al., 2022)	30.7 (0.3)	28.6 (0.4)	9.2 (0.2)	34.8 (0.4)
	CPT	**45.7** (0.9)	**44.5** (0.9)	**20.1** (0.6)	**51.0** (0.7)
Oracle (full) (Zhang et al., 2021)		63.9	68.3	48.3	65.1

Table 4
Top 6 colors from the frequency-based baseline and our cross-modal prompt search method.

Model	Color #1	Color #2	Color #3	Color #4	Color #5	Color #6
Freq	(255,0,0), red	(0,0,0), black	(0,0,255), blue	(0,255,0), green	(255,255,0), yellow	(165,42,42), brown
Ours	(240,0,30), red	(155,50,210), purple	(255,255,25), yellow	(0,10,255), blue	(255,170,230), pink	(0,255,0), green



(a) Overall results. (b) Zoom-in study of individual colors in different shots.

Fig. 3. Results of utilizing different colors for visual grounding, including (a) an overall evaluation of top 6 colors from different models, and (b) a zoom-in study of individual colors.

5.3. Influence of prompt configurations

We investigate the influence of colors, the key ingredients in the visual grounding of CPT. Specifically, we compare colors obtained from the frequency-based baseline (Freq), which uses the most frequent color names in text and their standard RGB value (see Section 3), and our Cross-modal Prompt Search (CPS) method in two dimensions, including an overall evaluation of top N colors and a zoom-in study of individual colors. The analysis is conducted based on CPT-Blk on the RefCOCO validation set.

Overall Evaluation of Top N Colors. We first show the top 6 colors recommended by each approach in Table 4. To evaluate the overall performance of the top colors from different models, we evaluate CPT equipped with each recommended color and report the mean accuracy and standard deviation over different colors. From the results in Fig. 3(a), we observe that the top colors produced by CPS achieve both higher mean accuracy and lower standard deviation than the baseline method in different shot settings. The reason is that CPS probes sensitive colors in VLP models for prompt construction, and therefore is able to effectively select the colors for better visual grounding.

Zoom-In Study of Individual Colors. To investigate the fine-grained influence of specific colors in CPT's visual grounding, we further perform a zoom-in study of individual colors. To align the colors for comparison, we merge the top 6 colors from the baseline and CPS, and remove the colors that are not included in the models' complete color sets (e.g., *black* $\notin C$ for CPS). We report the accuracies in Fig. 3(b), from which we observe that: (1) The performance of different colors varies greatly in prompting VLP models in the same shot settings, and the optimal colors are different in different shot settings. The results indicate the large influence of cross-modal prompt configurations, consistent with the findings from recent studies in textual prompt tuning (Jiang et al., 2020; Gao et al., 2021a). (2) Colors produced by CPS achieve comparable or superior performance compared with the baseline in individual colors. The results show that given the color texts, CPS can properly adjust the color visual appearance to improve the visual grounding performance. (3) We note that in some cases, colors produced by CPS slightly underperform the baseline. We hypothesize the reason is that CPS uses a single textual template to compute the decoding scores for color adjustment, which can be biased. The problem can potentially be addressed by ensembling templates as in Qin and Eisner (2021), which we leave for future work.

Performance on Color-involved Instances. Despite the effectiveness, adding colors in prompts might also disturb the understanding of raw images and text. We empirically assess the performance of CPT on color-involved instances. For the 2262 referring expressions containing color texts on RefCOCO+ testA set, CPT can achieve a reasonable 42.3% grounding accuracy with one shot, as compared with 28.9% of fine-tuning. The reason is that establishing strong cross-modal connections is more important in zero/few-shot scenarios, and a capable model can easily learn to distinguish the colors of raw objects and artificial markers.

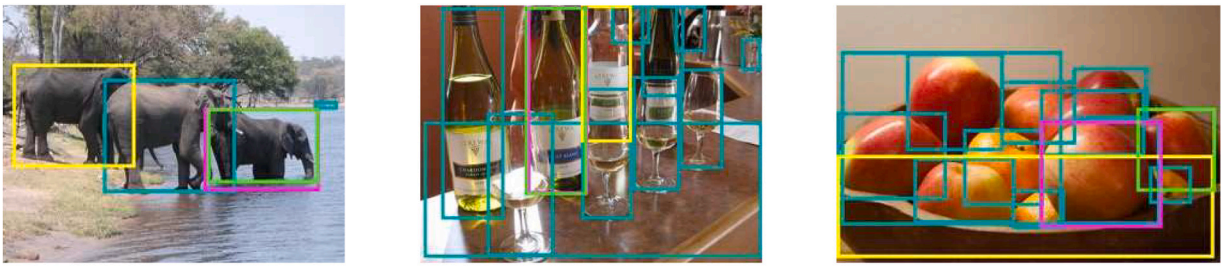


Fig. 4. Case study. The bounding boxes given by image region proposals (olive), ground-truth annotation (pink), CPT (light green), and fine-tuning baseline (yellow) are highlighted accordingly.

5.4. Case study

To provide a more intuitive understanding of CPT, we conduct a case study in few-shot setting. From Fig. 4 we can observe that: (1) CPT enables VLP models to distinguish targets distracted by the same type of objects using only a few instances, while fine-tuning struggles to succeed (left two figures). (2) CPT can be distracted by hard candidates (e.g., objects of the same type that require complex reasoning), but typically produces reasonable predictions. In the right figure, CPT predicts a nearby *apple* while fine-tuning predicts a *bowl*. The reason is that CPT reuses the pre-trained head of VLP models, which helps prevent outrageous results that typically happen in few-shot fine-tuning.

6. Related work

Prompt tuning for pre-trained models is a rapidly emerging field (Petroni et al., 2019; Brown et al., 2020; Schick and Schütze, 2021; Lester et al., 2021). Existing works for VLP models can be roughly divided into two categories, including natural language prompts and embedding-based prompts. We refer readers to the appendix for more related works.

Natural Language Prompts. To avoid the gap between pre-training and tuning, VLP models can be prompted with natural language templates, and produce the answer by filling in the mask (Radford et al., 2021; Liu et al., 2022; Zeng, 2022; Li et al., 2022a) or casual language generation (Wang et al., 2021b; Tsimpoukelli et al., 2021; Gupta et al., 2022; Kamath et al., 2022; Alayrac et al., 2022). To deal with object positions as task inputs/outputs, existing natural language prompt-based models either utilize extensive human annotations (Yang et al., 2022a; Wang et al., 2022; Yao et al., 2022; Li et al., 2022b; Kamath et al., 2021) or generate heuristic pseudo-data with external tools (Gupta et al., 2022; Kamath et al., 2022; Cho et al., 2021) to learn specialized position embeddings. Some works also generate text prompts that are related to the image content to enhance image representations (Rao et al., 2021; Wang et al., 2021a; Lin et al., 2022). Natural language prompts can improve the data efficiency of VLP models, enabling strong zero/few-shot capabilities. Moreover, different tasks can also be handled in a unified language modeling framework.

Embedding-based Prompts. To facilitate automatic prompt template search, some works learn prompts as new parameters. The parameters can be static pseudo text representations (Zhou et al., 2021; Ju et al., 2021; Sun et al., 2022; Zhu et al., 2022), disturbance vectors on images (Bahng et al., 2022; Liang et al., 2022), dynamic conditional representations (Zhou et al., 2022; Han et al., 2022), or lightweight additional modules (Gao et al., 2021b; Zhang and Ré, 2022; Jia et al., 2022; Lüddecke and Ecker, 2022; Yang et al., 2022b). When only new parameters are tuned, embedding-based prompts can be parameter-efficient. However, it can be difficult for embedding-based prompts to perform zero-shot tasks.

Most existing prompt tuning methods cannot establish fine-grained connections between text and image regions in zero/few-shot scenarios. In comparison, CPT prompts VLP models with natural co-referential markers in both image and text, which enables zero/few-shot fine-grained capabilities in locating and indicating objects for various VL tasks.

Vision-language Pre-training Models. Existing VLP models can be roughly divided into three categories: (1) *Masked language modeling* based VLP models are mainly pre-trained to recover the masked tokens (Lu et al., 2019; Su et al., 2019; Tan and Bansal, 2019; Li et al., 2020; Yu et al., 2021); (2) *Auto-regressive language modeling* based VLP models model image and text tokens with Transformer decoders auto-regressively (Ramesh et al., 2021; Wang et al., 2021b; Alayrac et al., 2022); (3) *Contrastive learning* based VLP models are pre-trained to holistically match image–text pairs (Radford et al., 2021; Li et al., 2021). In this work, we focus on prompting masked language modeling based VLP models due to their prevalence and superior performance, while applying CPT to other VLP models should also be applicable.

Visual Grounding. Most existing works on visual grounding learn to classify or rank image region candidates based on the expressions in a fully supervised fashion (Mao et al., 2016; Zhang et al., 2018; Lu et al., 2019; Chen et al., 2020), requiring large amounts of costly human-annotated data. To alleviate the reliance on human annotation, some works have investigated zero/few-shot grounding of new object types (Sadhu et al., 2019; Blukis et al., 2020), whereas amounts of training data are still needed for existing object types. Jiang et al. (2022) generate pseudo referring queries for data augmentation. Subramanian et al. (2022) enhance spatial reasoning capability of CLIP via heuristic rules. To refer to objects in images, Zellers et al. (2021) and Hessel et al. (2022) highlight objects with colored blocks, but require labeled data to learn the correlation between colors and object texts. Rohrbach et al. (2016) and Chen et al. (2018) explore weakly supervised approaches but are limited in dataset-specific in-domain training. In comparison, CPT prompts general VLP models for zero/few-shot visual grounding in a fill-in-the-blank fashion independent of specific object types, and can also be extended to indicate object positions for other VL tasks.

7. Conclusion and future work

In this work, we present a novel color-based prompt tuning framework for VLP models. To facilitate prompt construction, we present a principled approach to search for cross-modal prompt configurations. Comprehensive experimental results demonstrate the effectiveness of CPT on zero/few-shot VL tasks. In principle, color is one of the prominent attributes that can serve as natural prompts. In future, we will explore prompting VLP models with other cross-modal attributes, such as shape, size and status for more expressive, efficient and robust prompt tuning.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (No. 62236004).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.aiopen.2024.01.004.

References

Alayrac, Jean-Baptiste, Donahue, Jeff, Luc, Pauline, Miech, Antoine, Barr, Iain, Hasson, Yana, Lenc, Karel, Mensch, Arthur, Millican, Katie, Reynolds, Malcolm, et al., 2022. Flamingo: a visual language model for few-shot learning. arXiv preprint arXiv:2204.14198.

Anderson, Peter, He, Xiaodong, Buehler, Chris, Teney, Damien, Johnson, Mark, Gould, Stephen, Zhang, Lei, 2018a. Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of CVPR. pp. 6077–6086.

Anderson, Peter, Wu, Qi, Teney, Damien, Bruce, Jake, Johnson, Mark, Sünderhauf, Niko, Reid, Ian, Gould, Stephen, van den Hengel, Anton, 2018b. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: Proceedings of CVPR.

Antol, Stanislaw, Agrawal, Aishwarya, Lu, Jiasen, Mitchell, Margaret, Batra, Dhruv, Zitnick, C Lawrence, Parikh, Devi, 2015. VQA: Visual question answering. In: Proceedings of ICCV. pp. 2425–2433.

Bahng, Hyojin, Jahanian, Ali, Sankaranarayanan, Swami, Isola, Phillip, 2022. Visual prompting: Modifying pixel space to adapt pre-trained models. arXiv preprint arXiv:2203.17274.

Blukis, Valts, Knepper, Ross A., Artzi, Yoav, 2020. Few-shot object grounding and mapping for natural language robot instruction following. arXiv preprint arXiv:2011.07384.

Brown, Tom B, Mann, Benjamin, Ryder, Nick, Subbiah, Melanie, Kaplan, Jared, Dhariwal, Prafulla, Neelakantan, Arvind, Shyam, Pranav, Sastry, Girish, Askell, Amanda, et al., 2020. Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Cao, Jize, Gan, Zhe, Cheng, Yu, Yu, Licheng, Chen, Yen-Chun, Liu, Jingjing, 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In: Proceedings of ECCV. Springer, pp. 565–580.

Chen, Kan, Gao, Jiyang, Nevatia, Ram, 2018. Knowledge aided consistency for weakly supervised phrase grounding. In: Proceedings of CVPR. pp. 4042–4050.

Chen, Yen-Chun, Li, Linjie, Yu, Licheng, El Kholy, Ahmed, Ahmed, Faisal, Gan, Zhe, Cheng, Yu, Liu, Jingjing, 2020. UNITER: Universal image-text representation learning. In: Proceedings of ECCV. Springer, pp. 104–120.

Chen, Tianshui, Yu, Weihao, Chen, Riquan, Lin, Liang, 2019. Knowledge-embedded routing network for scene graph generation. In: Proceedings of CVPR. pp. 6163–6171.

Cho, Jaemin, Lei, Jie, Tan, Hao, Bansal, Mohit, 2021. Unifying vision-and-language tasks via text generation. In: Proceedings of ICML. In: PMLR, pp. 1931–1942.

Das, Abhishek, Kottur, Satwik, Gupta, Khushi, Singh, Avi, Yadav, Deshraj, Moura, José MF, Parikh, Devi, Batra, Dhruv, 2017. Visual dialog. In: Proceedings of CVPR. pp. 326–335.

Dodge, Jesse, Ilharco, Gabriel, Schwartz, Roy, Farhadi, Ali, Hajishirzi, Hannaneh, Smith, Noah, 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. arXiv preprint arXiv:2002.06305.

Gao, Tianyu, Fisch, Adam, Chen, Danqi, 2021a. Making pre-trained language models better few-shot learners. In: Proceedings of ACL.

Gao, Peng, Geng, Shijie, Zhang, Renrui, Ma, Teli, Fang, Rongyao, Zhang, Yongfeng, Li, Hongsheng, Qiao, Yu, 2021b. CLIP-adapter: Better vision-language models with feature adapters. arXiv preprint arXiv:2110.04544.

Gupta, Tanmay, Kamath, Amita, Kembhavi, Aniruddha, Hoiem, Derek, 2022. Towards general purpose vision systems: An end-to-end task-agnostic vision-language architecture. In: Proceedings of CVPR. pp. 16399–16409.

Han, Guangxing, Ma, Jiawei, Huang, Shiyuan, Chen, Long, Chellappa, Rama, Chang, Shih-Fu, 2022. Multimodal few-shot object detection with meta-learning based cross-modal prompting. arXiv preprint arXiv:2204.07841.

Hessel, Jack, Hwang, Jena D, Park, Jae Sung, Zellers, Rowan, Bhagavatula, Chandra, Rohrbach, Anna, Saenko, Kate, Choi, Yejin, 2022. The abduction of sherlock holmes: A dataset for visual abductive reasoning. arXiv preprint arXiv:2202.04800.

Hudson, Drew A., Manning, Christopher D., 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of CVPR. pp. 6700–6709.

Jia, Menglin, Tang, Luming, Chen, Bor-Chun, Cardie, Claire, Belongie, Serge, Hariharan, Bharath, Lim, Ser-Nam, 2022. Visual prompt tuning. arXiv preprint arXiv:2203.12119.

Jiang, Haojun, Lin, Yuanze, Han, Dongchen, Song, Shiji, Huang, Gao, 2022. Pseudo-Q: Generating pseudo language queries for visual grounding. In: Proceedings of CVPR. pp. 15513–15523.

Jiang, Zhengbao, Xu, Frank F., Araki, Jun, Neubig, Graham, 2020. How can we know what language models know? TACL 8, 423–438.

Ju, Chen, Han, Tengda, Zheng, Kunhao, Zhang, Ya, Xie, Weidi, 2021. Prompting visual-language models for efficient video understanding. arXiv preprint arXiv:2112.04478.

Kamath, Amita, Clark, Christopher, Gupta, Tanmay, Kolve, Eric, Hoiem, Derek, Kembhavi, Aniruddha, 2022. Webly supervised concept expansion for general purpose vision models. arXiv preprint arXiv:2202.02317.

Kamath, Aishwarya, Singh, Mannat, LeCun, Yann, Misra, Ishan, Synnaeve, Gabriel, Carion, Nicolas, 2021. MDETR: Modulated detection for end-to-end multi-modal understanding. arXiv preprint arXiv:2104.12763.

Krishna, Ranjay, Zhu, Yuke, Groth, Oliver, Johnson, Justin, Hata, Kenji, Kravitz, Joshua, Chen, Stephanie, Kalantidis, Yannis, Li, Li-Jia, Shamma, David A, et al., 2017. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. IJCV 123 (1), 32–73.

Lester, Brian, Al-Rfou, Rami, Constant, Noah, 2021. The power of scale for parameter-efficient prompt tuning. In: Proceedings of EMNLP. pp. 3045–3059.

Li, Wei, Gao, Can, Niu, Guocheng, Xiao, Xinyan, Liu, Hao, Liu, Jiachen, Wu, Hua, Wang, Haifeng, 2021. UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. In: Proceedings of ACL. Association for Computational Linguistics, pp. 2592–2607.

Li, Bin, Weng, Yixuan, Sun, Bin, Li, Shutao, 2022a. Towards visual-prompt temporal answering grounding in medical instructional video. arXiv preprint arXiv:2203.06667.

Li, Xiujun, Yin, Xi, Li, Chunyuan, Zhang, Pengchuan, Hu, Xiaowei, Zhang, Lei, Wang, Lijuan, Hu, Houdong, Dong, Li, Wei, Furu, et al., 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In: Proceedings of ECCV. Springer, pp. 121–137.

Li, Liunian Harold, Zhang, Pengchuan, Zhang, Haotian, Yang, Jianwei, Li, Chunyuan, Zhong, Yiwu, Wang, Lijuan, Yuan, Lu, Zhang, Lei, Hwang, Jenq-Neng, et al., 2022b. Grounded language-image pre-training. In: Proceedings of CVPR. pp. 10965–10975.

Liang, Sheng, Zhao, Mengjie, Schütze, Hinrich, 2022. Modular and parameter-efficient multimodal fusion with prompting. In: Findings of ACL. pp. 2976–2985.

Lin, Bingqian, Zhu, Yi, Chen, Zicong, Liang, Xiwen, Liu, Jianzhuang, Liang, Xiaodan, 2022. ADAPT: Vision-language navigation with modality-aligned action prompts. In: Proceedings of CVPR. pp. 15396–15406.

Liu, Xuejing, Li, Liang, Wang, Shuhui, Zha, Zheng-Jun, Meng, Dechao, Huang, Qingming, 2019a. Adaptive reconstruction network for weakly supervised referring expression grounding. In: Proceedings of ICCV. pp. 2611–2620.

Liu, Xuejing, Li, Liang, Wang, Shuhui, Zha, Zheng-Jun, Su, Li, Huang, Qingming, 2019b. Knowledge-guided pairwise reconstruction network for weakly supervised referring expression grounding. In: Proceedings of ACM MM. pp. 539–547.

Liu, Yuhang, Wei, Wei, Peng, Daowan, Zhu, Feida, 2022. Declaration-based prompt tuning for visual question answering. arXiv e-prints.

Liu, Pengfei, Yuan, Weizhe, Fu, Jinlan, Jiang, Zhengbao, Hayashi, Hiroaki, Neubig, Graham, 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. arXiv preprint arXiv:2107.13586.

Lu, Jiasen, Batra, Dhruv, Parikh, Devi, Lee, Stefan, 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Proc. NeurIPS 32, 13–23.

Lüddecke, Timo, Ecker, Alexander, 2022. Image segmentation using text and image prompts. In: Proceedings of CVPR. pp. 7086–7096.

Mao, Junhua, Huang, Jonathan, Toshev, Alexander, Camburu, Oana, Yuille, Alan L, Murphy, Kevin, 2016. Generation and comprehension of unambiguous object descriptions. In: Proceedings of CVPR. pp. 11–20.

Petroni, Fabio, Rocktäschel, Tim, Riedel, Sebastian, Lewis, Patrick, Bakhtin, Anton, Wu, Yuxiang, Miller, Alexander, 2019. Language models as knowledge bases? In: Proceedings of EMNLP-IJCNLP. pp. 2463–2473.

Qin, Guanghui, Eisner, Jason, 2021. Learning how to ask: Querying LMs with mixtures of soft prompts. In: Proceedings of NAACL. pp. 5203–5212.

Radford, Alec, Kim, Jong Wook, Hallacy, Chris, Ramesh, Aditya, Goh, Gabriel, Agarwal, Sandhini, Sastry, Girish, Askell, Amanda, Mishkin, Pamela, Clark, Jack, et al., 2021. Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020.

Ramesh, Aditya, Pavlov, Mikhail, Goh, Gabriel, Gray, Scott, Voss, Chelsea, Radford, Alec, Chen, Mark, Sutskever, Ilya, 2021. Zero-shot text-to-image generation. arXiv preprint arXiv:2102.12092.

Rao, Yongming, Zhao, Wenliang, Chen, Guangyi, Tang, Yansong, Zhu, Zheng, Huang, Guan, Zhou, Jie, Lu, Jiwen, 2021. DenseCLIP: Language-guided dense prediction with context-aware prompting. arXiv preprint arXiv:2112.01518.

Rohrbach, Anna, Rohrbach, Marcus, Hu, Ronghang, Darrell, Trevor, Schiele, Bernt, 2016. Grounding of textual phrases in images by reconstruction. In: Proceedings of ECCV. Springer, pp. 817–834.

Sadhu, Arka, Chen, Kan, Nevatia, Ram, 2019. Zero-shot grounding of objects from natural language queries. In: Proceedings of ICCV. pp. 4694–4703.

Schick, Timo, Schütze, Hinrich, 2021. It's not just size that matters: Small language models are also few-shot learners. In: Proceedings of NAACL. pp. 2339–2352.

Su, Weijie, Zhu, Xizhou, Cao, Yue, Li, Bin, Lu, Lewei, Wei, Furu, Dai, Jifeng, 2019. VL-BERT: Pre-training of generic visual-linguistic representations. In: Proceedings of ICLR.

Subramanian, Sanjay, Merrill, William, Darrell, Trevor, Gardner, Matt, Singh, Sameer, Rohrbach, Anna, 2022. ReCLIP: A strong zero-shot baseline for referring expression comprehension. In: Proceedings of ACL. pp. 5198–5215.

Sun, Ximeng, Hu, Ping, Saenko, Kate, 2022. DualCoOp: Fast adaptation to multi-label recognition with limited annotations. arXiv preprint arXiv:2206.09541.

Sun, Mingjie, Xiao, Jimin, Lim, Eng Gee, Liu, Si, Goulermas, John Y, 2021. Discriminative triad matching and reconstruction for weakly referring expression grounding. TPAMI 43 (11), 4189–4195.

Tan, Hao, Bansal, Mohit, 2019. LXMERT: Learning cross-modality encoder representations from transformers. In: Proceedings of EMNLP-IJCNLP. pp. 5100–5111.

Tellex, Stefanie, Kollar, Thomas, Dickerson, Steven, Walter, Matthew, Banerjee, Ashis, Teller, Seth, Roy, Nicholas, 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In: Proceedings of AAAI. volume 25.

Tsimpoukelli, Maria, Menick, Jacob, Cabi, Serkan, Eslami, SM, Vinyals, Oriol, Hill, Felix, 2021. Multimodal few-shot learning with frozen language models. arXiv preprint arXiv:2106.13884.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, Kaiser, Łukasz, Polosukhin, Illia, 2017. Attention is all you need. In: Proceedings of NeurIPS. pp. 5998–6008.

Wang, Mengmeng, Xing, Jiazheng, Liu, Yong, 2021a. ActionCLIP: A new paradigm for video action recognition. arXiv preprint arXiv:2109.08472.

Wang, Peng, Yang, An, Men, Rui, Lin, Junyang, Bai, Shuai, Li, Zhikang, Ma, Jianxin, Zhou, Chang, Zhou, Jingren, Yang, Hongxia, 2022. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: Proceedings of ICML.

Wang, Zirui, Yu, Jiahui, Yu, Adams Wei, Dai, Zihang, Tsvetkov, Yulia, Cao, Yuan, 2021b. SimVLM: Simple visual language model pretraining with weak supervision. arXiv preprint arXiv:2108.10904.

Yang, Zhengyuan, Gan, Zhe, Wang, Jianfeng, Hu, Xiaowei, Ahmed, Faisal, Liu, Zicheng, Lu, Yumao, Wang, Lijuan, 2022a. UniTAB: Unifying text and box outputs for grounded vision-language modeling. In: Proceedings of ECCV.

Yang, Hao, Lin, Junyang, Yang, An, Wang, Peng, Zhou, Chang, Yang, Hongxia, 2022b. Prompt tuning for generative multimodal pretrained models. arXiv preprint arXiv:2208.02532.

Yao, Yuan, Chen, Qianyu, Zhang, Ao, Ji, Wei, Liu, Zhiyuan, Chua, Tat-Seng, Sun, Maosong, 2022. PEVL: Position-enhanced pre-training and prompt tuning for vision-language models. arXiv preprint arXiv:2205.11169.

Yao, Yuan, Zhang, Ao, Han, Xu, Li, Mengdi, Weber, Cornelius, Liu, Zhiyuan, Wermter, Stefan, Sun, Maosong, 2021. Visual distant supervision for scene graph generation. In: Proceedings of ICCV. pp. 15816–15826.

Yu, Licheng, Poirson, Patrick, Yang, Shan, Berg, Alexander C, Berg, Tamara L, 2016. Modeling context in referring expressions. In: Proceedings of ECCV. Springer, pp. 69–85.

Yu, Fei, Tang, Jiji, Yin, Weichong, Sun, Yu, Tian, Hao, Wu, Hua, Wang, Haifeng, 2021. ERNIE-ViL: Knowledge enhanced vision-language representations through scene graphs. In: Proceedings of AAAI. volume 35, pp. 3208–3216.

Zellers, Rowan, Bisk, Yonatan, Farhadi, Ali, Choi, Yejin, 2019. From recognition to cognition: Visual commonsense reasoning. In: Proceedings of CVPR. pp. 6720–6731.

Zellers, Rowan, Lu, Ximing, Hessel, Jack, Yu, Youngjae, Park, Jae Sung, Cao, Jize, Farhadi, Ali, Choi, Yejin, 2021. MERLOT: Multimodal neural script knowledge models. In: Proceedings of NeurIPS. volume 34.

Zeng, Yawen, 2022. Point prompt tuning for temporally language grounding. In: Proceedings of SIGIR. pp. 2003–2007.

Zhang, Pengchuan, Li, Xiujun, Hu, Xiaowei, Yang, Jianwei, Zhang, Lei, Wang, Lijuan, Choi, Yejin, Gao, Jianfeng, 2021. VinVL: Revisiting visual representations in vision-language models. In: Proceedings of CVPR. pp. 5579–5588.

Zhang, Hanwang, Niu, Yulei, Chang, Shih-Fu, 2018. Grounding referring expressions in images by variational context. In: Proceedings of CVPR. pp. 4158–4166.

Zhang, Michael, Ré, Christopher, 2022. Contrastive adapters for foundation model group robustness. arXiv preprint arXiv:2207.07180.

Zhou, Kaiyang, Yang, Jingkang, Loy, Chen Change, Liu, Ziwei, 2021. Learning to prompt for vision-language models. arXiv preprint arXiv:2109.01134.

Zhou, Kaiyang, Yang, Jingkang, Loy, Chen Change, Liu, Ziwei, 2022. Conditional prompt learning for vision-language models. In: Proceedings of CVPR. pp. 16816–16825.

Zhu, Beier, Niu, Yulei, Han, Yucheng, Wu, Yue, Zhang, Hanwang, 2022. Prompt-aligned gradient for prompt tuning. arXiv preprint arXiv:2205.14865.