HybridPrompt: Bridging Language Models and Human Priors in Prompt Tuning for Visual Question Answering

Zhiyuan Ma, Zhihuan Yu, Jianjun Li*, Guohui Li

Huazhong University of Science and Technology (HUST), China
{zhiyuanma,zhihuanyu,jianjunli,guohuili}@hust.edu.cn

Abstract

Visual Question Answering (VQA) aims to answer the natural language question about a given image by understanding multimodal content. However, the answering quality of most existing visual-language pre-training (VLP) methods is still limited, mainly due to: (1) Incompatibility. Upstream pre-training tasks are generally incompatible with downstream question answering tasks, which makes the knowledge from the language model not well transferable to downstream tasks, and greatly limits their performance in few-shot scenarios; (2) Under-fitting. They generally do not integrate human priors to compensate for universal knowledge from language models, so as to fit the challenging VQA problem and generate reliable answers. To address these issues, we propose HybridPrompt, a cloze- and verify-style hybrid prompt framework with bridging language models and human priors in prompt tuning for VQA. Specifically, we first modify the input questions into the cloze-style prompts to narrow the gap between upstream pre-training tasks and downstream VQA task, which ensures that the universal knowledge in the language model can be better transferred to subsequent human prior-guided prompt tuning. Then, we imitate the cognitive process of human brain to introduce topic and sample related priors to construct a dynamically learnable prompt template for human prior-guided prompt learning. Finally, we add fixed-length learnable free-parameters to further enhance the generalizability and scalability of prompt learning in the VQA model. Experimental results verify the effectiveness of HybridPrompt, showing that it achieves competitive performance against previous methods on widely-used VQAv2 dataset and obtains new state-of-the-art results. Our code is released at: https://github.com/zhizhi111/hybrid.

Introduction

Visual Question Answering (VQA) (Yu et al. 2017, 2018) is a classic and challenging multimodal comprehension task, which aims to learn cross-modal semantic content from image-text pair to answer a given natural language question. Inspired by the success of vision-language pre-training (VLP), we have recently witnessed a boosting number of research works on VQA (Chen et al. 2020; Li et al. 2020; Xu et al. 2021; Li et al. 2021b; Radford et al. 2021; Li et al. 2019b; Lu et al. 2019; Kim, Son, and Kim 2021; Cui et al. 2021; Ma et al. 2022b; Li et al. 2021a; Jia et al. 2021; Ma et al. 2022a; Liu et al. 2022).

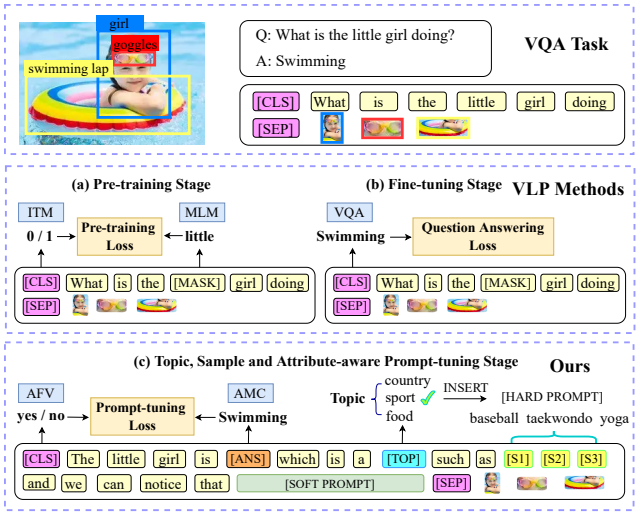


Figure 1: Example of VLP methods and our proposed HybridPrompt method for VQA task.

Though achieving remarkable progress, existing VLP-based methods still suffer from the following two limitations. (1) *Incompatibility*. **Firstly**, upstream pre-training tasks are generally incompatible with downstream question answering task. Taking Figure 1 as an example, the VLP-based methods mainly exploit the "Pre-training + Fine-tuning" paradigm to train a vision-language model, aiming at obtaining a correct answer such as *"Swimming"* for VQA. However, in the upstream pre-training task, i.e., in the pre-training stage in Figure 1 (a), previous VLP methods generally only carry out simple mask-language-modeling (MLM), image-text-matching (ITM) or some extended tasks to respectively learn masked contextual word representations (e.g., the word *"little"* has been replaced by a special token `[MASK]`) and coarse-grained image-text matching (e.g., *"What is the little girl doing"* and its corresponding description image), which are obviously different from the downstream VQA task depicted in the next fine-tuning stage in Figure 1 (b). In the VQA task, the pre-trained model needs to deeply understand the

*Corresponding author.

nature of the question, so as to obtain a reliable answer such as *"Swimming"* for accurate visual question answering. Especially in the new few-shot domains, accurate and reliable answers are more difficult to obtain due to the lack of a large amount of annotated VQA data to support model fine-tuning. (2) *Under-fitting.* **Secondly**, prior models generally do not integrate human priors (e.g., topic, sample and attribute-related priors) to compensate for universal knowledge from language models to fit the challenging VQA problem and generate reliable answers. For example, in Figure 1, when the question asks *"What is the little girl doing?"*, the VQA system is expected to firstly explore the topic of the answer such as *"country"*, *"sport"* or *"food"* as in Figure 1 (c). There is no doubt that clarity on the topic of the answer will greatly contribute to the success of VQA. Moreover, the listing of similar entities (e.g., *"baseball"*, *"taekwondo"* and *"yoga"*) in the same topic will obviously help the VQA system to learn a sample-aware answer, since similar entities tend to have similar topic-shared features, which will make the prediction of the answer (*i.e., "Swimming"*) easier. Last but not the least, prior models also fail to learn scenario-specific features. For example, we can notice that in Figure 1, the image shows that *the little girl is wearing a swimming lap*. If such an scenario- or attribute-aware representation can be captured by the VQA model, it will help reach a more accurate answer.

To address the aforementioned limitations, we propose HybridPrompt, a *cloze-* and *verify-style* hybrid prompt framework with bridging language models and human priors in prompt tuning for VQA. Specifically, to address the first limitation, we stand on the shoulder of prompt learning paradigm (Liu et al. 2021) to propose a "pre-training + prompt-tuning" based approach, which modifies the input questions into *cloze-* and *verify-style* prompts to mitigate the gap between upstream MLM & ITM tasks and downstream VQA task. Based on such a framework, we further address the second limitation by designing a topic, sample and attribute-aware hybrid prompt template to dynamically integrate human priors and perform prompt-tuning for achieving more accurate and reliable answer prediction. Experiments on widely-used VQA v2 dataset demonstrate the effectiveness of HybridPrompt, showing that it achieves competitive performance and obtains new state-of-the-art results.

Related Work

General VQA Methods (General-VQA). As a vital yet challenging multimodal task, VQA recently has drawn more and more attention. In most general VQA methods, attention mechanism and multi-modal fusion are two fundamental techniques. The attention mechanism has been widely explored in both computer vision (CV) and natural language processing (NLP), and has also been jointly applied in many VQA models such as SAN (Yang et al. 2016), UpDn (Anderson et al. 2018), BAN (Kim, Jun, and Zhang 2018), DFAF (Gao et al. 2019) and MCAN (Yu et al. 2019), which effectively build a crucial bridge for joint reasoning between multimodal features and significantly enhance the accuracy of VQA models. Moreover, to achieve further interactions between visual and textual features for answer prediction, a series of the VQA models have been proposed by employing various advanced bilinear pooling strategies, such as MLB (Kim et al. 2017), MCB (Fukui et al. 2016), MUTAN (Ben-Younes et al. 2017), BLOCK (Ben-Younes et al. 2019), and MHEF (Lao et al. 2021b). Recently, by taking advantages of graph networks (GN) in relational reasoning, some classical GN-based VQA models have been proposed (e.g., MuRel (Cadene et al. 2019), ReGAT (Li et al. 2019a), MN-GMN (Khademi 2020) and DC-GCN (Huang et al. 2020)) and shown promising results.

VLP-Based VQA Methods (VLP-VQA). Despite significant improvements, previous general VQA methods still cannot align visual and textual features well enough for joint training. To this end, a series of vision-language pre-training (VLP)-based VQA models have been proposed to mitigate the predicament. Specifically, the VLP-VQA models are firstly pretrained on large-scale paired image-text corpus to obtain unified-modality representations and then transferred to downstream task to benefit visual question answering. **ViLBERT** (Lu et al. 2019) and **LXMERT** (Tan and Bansal 2019) are two pioneering researches by adopting a dual-stream architecture to separately encode visual and textual features and then perform multimodal fusion via a unified encoder. But they are computationally expensive. In view of this, some recent works tend to use the single-stream architecture to fuse them and show promising performance. For example, **UNITER** (Chen et al. 2020) adopts Faster R-CNN (Ren et al. 2015) to extract feature sequences of Region of Interest (RoI) from images, and then concatenates them with textual sequence into a unified Transformer encoder, which is mainly supervised by MLM, MRM and ITM tasks for pre-training. Simultaneously, **OSCAR** (Li et al. 2020) introduces the object tags into image-text sequence and constructs a binary contrastive loss to learn image-text alignments. Later, **VinVL** (Zhang et al. 2021) extends the binary contrastive loss into 3-way contrastive loss to effectively transfer to VQA task for fine-tuning. Unlike them, **ViLT** (Kim, Son, and Kim 2021) is the first to use pre-trained ViT (Dosovitskiy et al. 2020) to extract visual features and directly employ linear mapping method to embed image blocks for faster VQA prediction. Likewise, subsequent **ALBEF** (Li et al. 2021a) and **UNIMO** (Li et al. 2021b) also adopt such a unified architecture and leverage cross-modal contrastive learning to align the image and text before fusing them through cross-modal attention. Recently, **E2E-VLP** (Xu et al. 2021) builds a new unified Transformer framework to jointly learn visual representations and semantic alignments between image and text for end-to-end vision-language pre-training. **ROSITA** (Cui et al. 2021) further introduces intra-modal and cross-modal knowledge graph into the model and uses a novel SKM strategy to pretrain, which effectively enhance the semantic alignments. Very recently, **TCL** (Yang et al. 2022) takes advantage of localized and structural information and proposes triple contrastive learning for VLP by leveraging both cross-modal and intra-modal contrastive objectives to provide complementary benefits in representation learning. Generally speaking, these VLP-VQA methods firstly use MLM, ITM and their extended tasks to pre-train a large-scale language model,

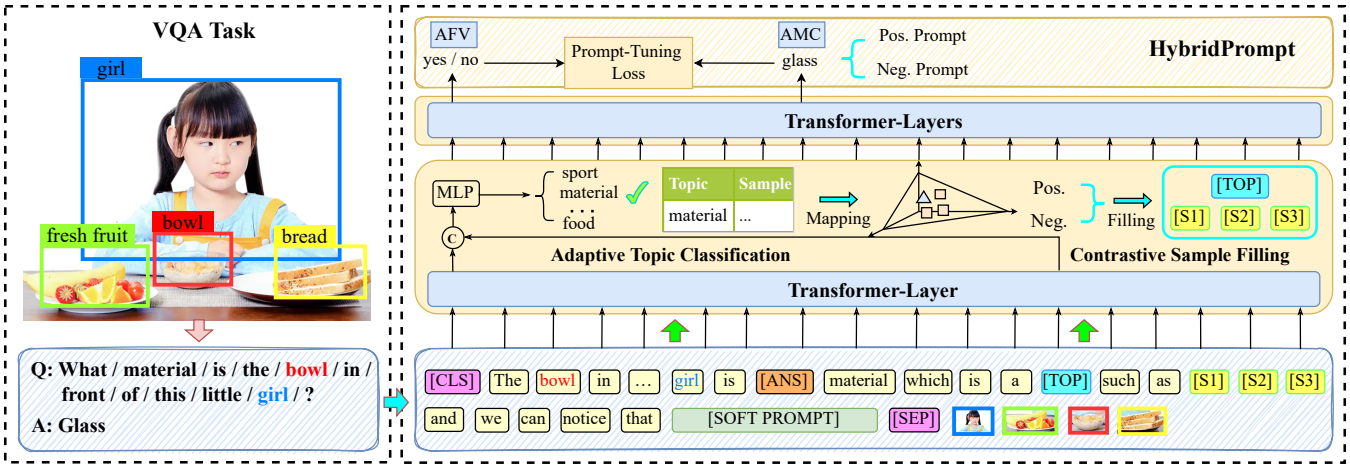


Figure 2: The framework of proposed HybridPrompt.

and then directly transfer the pre-trained language model to downstream VQA task for answer prediction but leveraging a brand new QA objective, which is completely different from the previous pre-training objectives. This is also the reason why the effect of VQA has not been significantly improved with the increasing scale of model parameters and pre-training corpus, that is, the lack of bridging language model and human prior knowledge in the fine-tuning of VQA.

Prompt Learning Techniques (PLT). Enhancing the prior capabilities of models by incorporating more knowledge into language models (LMs) via prompts has recently sparked the interest of NLP researchers (Petroni et al. 2019; Shin et al. 2020; Jiang et al. 2020; Li and Liang 2021; Zhong, Friedman, and Chen 2021; Lester, Al-Rfou, and Constant 2021; Gao, Fisch, and Chen 2021; Liu et al. 2021), which is another line of research relevant to our work. In this new dubbed "*pre-train, prompt and predict*" paradigm, instead of adapting pre-trained LMs to downstream tasks via objective engineering, downstream tasks are reorganized to look more like those solved during the original LM training with the help of a textual *prompt*. For example, when recognizing the emotion of a social media post, "I missed the bus today.", we may continue with a prompt "I felt so __". The above example is a *prefix-style* prompt, that is, "I missed the bus today. I felt so __" as in Prefix-Tuning (Li and Liang 2021) and PromptTuning (Lester, Al-Rfou, and Constant 2021). However, these approaches require manually constructing context prompts for each data sample, which is practically infeasible. Different from existing prompt learning techniques, in this work, we introduce a topic, sample and attribute-guided learnable hybrid template into VLP models to bridge the universal knowledge from LMs and specialized knowledge from human priors in prompt-tuning for reliable VQA.

Methodology

We first present the problem definition, and then briefly introduce our observations, based on which we propose our HybridPrompt framework.

Problem Definition. Following prior work (Lao et al. 2021a), we define the VQA task as predicting the most likely answer a from an answer dictionary $\mathcal{A}$, giving an image $\boldsymbol{I}$ from image set $\mathcal{I}$ and a question $\boldsymbol{Q}$ from question set $\mathcal{Q}$. A VQA dataset with N training instances is denoted as $\mathcal{D} = \{(\boldsymbol{I}_i, \boldsymbol{Q}_i), a_i\}_{i=1}^{N}$, where $\boldsymbol{I}_i \in \mathcal{I}$ and $\boldsymbol{Q}_i \in \mathcal{Q}$ are the image and question input of the i-th instance, while $a_i \in \mathcal{A}$ denotes the correct answer of the i-th instance. Considering that a question to an image may have multiple correct answers (e.g., *"carpet"* and *"rug"*) in widely-used VQA datasets (Goyal et al. 2017), the VQA model can be formally defined as learning a mapping function $f : \mathcal{Q} \times \mathcal{I} \rightarrow [0, 1]^{|\mathcal{A}|}$ from the multimodal inputs $\mathcal{Q} \times \mathcal{I}$ to the answer space $\mathcal{A}$, then producing a probability distribution over $\mathcal{A}$ and selecting the answer with the highest probability as output. The probability distribution can be formally defined as,

$$P(\mathcal{A} \mid \boldsymbol{I}_i, \boldsymbol{Q}_i) = \text{softmax}(f_\theta(\boldsymbol{I}_i, \boldsymbol{Q}_i)) \tag{1}$$

where θ denotes the trainable parameters of the VQA model. We can employ a cross-entropy loss for the answer mapping classification task. Formally,

$$\mathcal{L}_{\text{AMC}} = -\frac{1}{N} \sum_{i}^{N} \sum_{j}^{|\mathcal{A}|} a_{ij}^{*} \log \left(P(a_{ij} \mid \boldsymbol{I}_i, \boldsymbol{Q}_i) \right) \tag{2}$$

where a_{ij} denotes the j^{th} answer candidate in $\mathcal{A}$ for the i^{th} training instance, and a_{ij}^{*} is its ground truth label.

Intuitive Observations. We make some intuitive yet meaningful observations about VQA: Firstly, to answer a question, it is helpful to narrow down the search range for an answer by giving the topic category of the answer. Secondly, inspired by the success of contrastive learning, the accuracy of visual question answering can be further improved if several topic-related samples (positive samples) are given to prompt answer generation or several topic-unrelated samples (negative samples) are given to exclude topic-unrelated answers. Thirdly, when searching for answers under the same topic, using a soft-prompting strategy (Li and Liang 2021; Lester,

Al-Rfou, and Constant 2021) to train a set of attribute-related representations of answers for prompt-tuning is beneficial for improving the VQA model's ability to distinguish similar answers, which can further improve the model's performance.

Framework Overview. Based on the above observations, we propose a hybrid prompt framework HybridPrompt, which aims to introduce a dynamically trainable template for aforementioned topic, sample and attribute-aware downstream prompt-tuning on VQA task. The proposed HybridPrompt framework mainly comprises four parts, as shown in Figure 2. Specifically, a Hybrid Weak Prompt (HWP) layer (Sec.) is first used to construct a *cloze-style* weak prompt with `[ANS]`, `[TOP]` and $[S_i]$ slots for answer prediction, topic classification and sample filling respectively. This layer makes the VQA task be more close to the upstream MLM pre-training task, which significantly reduces its difficulty. These templates with unfilled slots are then fed to the next Dynamic Hard Prompt (DHP) layer (Sec.) for dynamic adaptive topic classification. By fusing hidden information from `[TOP]` and `[CLS]` slots, their topic categories can be obtained by feeding the fused representations into an MLP layer followed by a softmax function. Then, by searching a topic-sample library based on human knowledge and sampling randomly, we can obtain the topic-related samples as positive samples, whereas topic-unrelated ones as negative samples for contrastive learning. Subsequently, these hard templates filled with topic and samples are fed into the next Trainable Soft Prompt (TSP) layer (Sec.) for training and updating of soft prompts by employing multi-layer Transformers. Finally, an Answer Mapping Classification (AMC) layer (Sec.) is designed to conduct the final answer mapping classification by predicting a probability distribution over answer dictionary $\mathcal{A}$ under the obtained hybrid prompts. Note the AMC loss is a contrastive loss based on positive and negative samples, which not only guides the VQA model to give correct predictions, but also guides the model to give correct predictions only if the given prompts are correct. Moreover, a fresh answer filling verification (AFV) loss is designed to supervise the answer prediction process from a global perspective, which imitates the commonly-used ITM pretraining task and makes the VQA model benefit more from pre-training stage of the language models.

Hybrid Weak Prompt (HWP) Layer

The HWP layer is designed to modify the questions of VQA tasks into the *cloze-style* hybrid prompt templates for adapting to upstream MLM task and facilitating subsequent prompt-tuning process. Following Liu et al. (2021), we define our weak prompt operation as a prompting function $f_{\text{prompt}}(\cdot)$, which aims to map the question $\boldsymbol{Q}$ into a weak prompt template $\mathbf{T}$,

$$\mathbf{T} = f_{\text{prompt}}(\boldsymbol{Q}) \tag{3}$$

the weak prompting operation mainly comprises two steps: linguistic reorganization and weak prompt concatenation.

Linguistic Reorganization This step is used to modify each batch of questions into declarative statements, which include general questions and special questions. For general questions (i.e., the answer type is *yes* or *no*), we simply search for typical words in the sentence and reorganize the sentence by placing them after the subject. The typical words we use include *be* verbs (e.g., *is*, *am*, *are*, *was* or *were*), *auxiliary* verbs (e.g., *do*, *does*, *did*, *have*, *has* or *had*) and *modal* verbs (e.g., *may*, *can*, *need*, *might*, *must*, *dare*, *will*, *shall*, *would*, or *should*). For more complex special questions, we use the following rules to reorganize the questions into initial weak prompt template $\mathbf{T}_{\text{initial}}$:

- ***what / who / why / how***[1]***+ be + [x]:*** The initial weak prompt template is: [x] + *be* + `[ANS]`;
- ***when / where + be + [x]:*** The initial weak prompt template is: [x] + *be* + at / in the + `[ANS]`;
- ***what / whose / which* + [x_1] + *be* + [x_2]:** The initial weak prompt template is: [x_2] + *be* + `[ANS]` + [x_1];
- **interrogative + auxiliary / modal + [x]:** The initial weak prompt template in this case is similar to the previous rules, but different from them in that it needs to replace the *be* verb with the corresponding *auxiliary* or *modal* verb, based on the type of the specific interrogative pronoun.

Weak Prompt Concatenation This step is used to introduce human knowledge and concatenate it with the above initial weak prompt template to obtain the final weak prompt template $\mathbf{T}$. Specifically, $\mathbf{T}$ = `[CLS]` + $\mathbf{T}_{\text{initial}}$ + 'which is a' + `[TOP]` + 'such as' + $[S_1]$ + $[S_2]$ + $[S_3]$ + 'and we can notice that' + `[SOFT PROMPT]` + `[SEP]` + $\mathbf{v}_1 + \mathbf{v}_2 + \cdots + \mathbf{v}_m$, where `[CLS]`, `[SEP]`, `[TOP]` and $[S_i]$ respectively denote classification token, separate token, topic token and sample token, while `[SOFT PROMPT]` represents for a set of continuous randomly initialized tensors with fixed length. Moreover, $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_m$ respectively denotes visual tokens from an image detected by the Faster-RCNN (Ren et al. 2015) as in UNITER (Chen et al. 2020), and m is the number of detected visual tokens from the image.

Dynamic Hard Prompt (DHP) Layer

The DHP layer is the key to implementing topic and sample-aware prompt tuning in our work, which aims to dynamically learn the possible position of the answer in the latent topic space by exploiting a unified Transformer layer, and give out points near the answer as positive samples whereas points far away from it as negative samples for contrastive learning. Specifically, we first define the Transformer layer used for dynamic topic classification as function $f_{\text{topic}}(\cdot)$,

$$\mathbf{H} = f_{\text{topic}}(\mathbf{T}) \tag{4}$$

Note $\mathbf{H}_0$ is the hidden state of `[CLS]`, and $\mathbf{H}_{n+4}$ is the representation of `[TOP]`, where n denotes the length of $\mathbf{T}_{\text{initial}}$. The DHP layer mainly comprises two steps: adaptive topic classification and contrastive sample filling.

Adaptive Topic Classification To obtain topic-aware sentence-level representation to narrow down the search range for an answer, we fuse the hidden state of `[CLS]` and

[1]The special questions guided by *how* include *how long*, *how far*, *how much*, *how old*, *how heavy*, *how tall*, etc.

[TOP]	[S$_1$]	[S$_2$]	[S$_3$]	[TOP]	[S$_1$]	[S$_2$]	[S$_3$]	[TOP]	[S$_1$]	[S$_2$]	[S$_3$]
Judge	yes	no	unknown	Brand	nike	apple	dell	Material	plastic	wood	metal
Color	white	blue	red	Number	-10	3.14	58	Pattern	strips	solid	plaid
Type	sarcasm	humor	sorrow	Animal	cat	dog	horse	Gender	male	female	woman
Time	7.00 am	afternoon	night	Country	usa	uk	china	Weather	rainy	sunny	windy
Sport	baseball	taekwondo	yoga	Fruit	banana	orange	apple	Food	bread	milk	cake
Age	1	10	young	Name	big ben	united	jack	Reason	safety	fast	stability
Orientation	left	right	north	Location	beach	outside	street	Person	man	mom	boy
Comment	poor	clear	good	Object	umbrella	kite	frisbee	Others	unknown	unknown	unknown

Table 1: The categories of topics and their possible corresponding samples.

[TOP] by leveraging an MLP layer and a softmax function as follows,

$$P(c|\mathbf{T}) = \text{softmax}\left(\text{MLP}\left(\mathbf{H}_0, \mathbf{H}_{n+4}\right)\right) \tag{5}$$

where c denotes the latent topic categories of the answer, and $P(c|\mathbf{T})$ is a probability distribution over all categories. The categories labeled by us on VQA v2 dataset are summarized in Table 1, which includes $K = 24$ categories in total. We then employ a cross-entropy loss to supervise the adaptive topic classification process as follows,

$$\mathcal{L}_{\text{ATC}} = -\frac{1}{N}\sum_{i}^{N}\sum_{j}^{K} c_{ij}^{*}\log(P(c_{ij} \mid \mathbf{T})) \tag{6}$$

where c_{ij} denotes the j-th topic category for the i-th training instance, and c_{ij}^{*} is its ground truth label annotated by us. Finally, a topic word $\boldsymbol{w}_{top}$ can be obtained by selecting the category with the highest probability over all labeled topic categories, which will be filled at the [TOP] slot of the $\mathbf{T}$.

Contrastive Sample Filling To obtain sample-aware entity-level representation for further fine-grained locating of the answer, we design the contrastive sample filling module. Specifically, we search the topic-sample pairs in Table 1 to match the topic word $\boldsymbol{w}_{top}$ obtained in the previous classification module and obtain three topic-related items as the positive samples for filling. Note that the three samples shown in Table 1 are randomly selected from the sample library with a maximum size of 10. On the other hand, we also search for topic words that are not related to $\boldsymbol{w}_{top}$ and their corresponding three random samples as negative samples for filling, which are used for pushing away the distance between the answer and the negative samples. Finally, the template filled with a topic word and several samples can be represented by $\mathbf{T}_{\text{filled}} = \{\mathbf{T}_{\text{filled}}^{+}, \mathbf{T}_{\text{filled}}^{-}\}$.

Trainable Soft Prompt (TSP) Layer

The TSP layer is the key to implementing attribute-aware prompt for predicting a more fine-grained answer, which aims to employ the multi-layer Transformers to adaptively train a set of continuous randomly initialized tensors to fit a specific yet challenging answer and improve the personalized representation ability of the VQA model. Specifically, we adopt 12 layers of Transformer as backbone, which is used for adaptive attribute-aware training and can be defined as function $f_{\text{attri}}(\cdot)$,

$$\mathbf{H}_{\text{attri}}^{(l)} = f_{\text{attri}}^{(l)}(\mathbf{H}_{\text{attri}}^{(l-1)}) \tag{7}$$

$$\mathbf{H}_{\text{attri}}^{(0)} = \mathbf{T}_{\text{filled}} \tag{8}$$

where l denotes layer number, and $\mathbf{H}_{\text{attri}}^{(12)}$ is the output of the last layer. The TSP layer retrieves the hidden feature $\mathbf{H}_{\text{ans}}^{(12)}$ of the [ANS] state and the hidden feature $\mathbf{H}_{\text{cls}}^{(12)}$ of the [CLS] state for prediction and verification of the next AMC layer.

Answer Mapping Classification (AMC)

The AMC layer is designed to predict a final answer based on the training of a hybrid prompt template. Specifically, given hidden features $\mathbf{H}_{\text{ans}}^{(12)}$ and $\mathbf{H}_{\text{cls}}^{(12)}$, AMC is trained to map the hidden feature of the answer into an answer space $\mathcal{A}$ to select a most likely answer and verify whether the answer is the correct one under given the correct prompt (i.e., $\mathbf{T}_{\text{filled}}^{+}$).

For answer mapping classification, HybridPrompt first feeds hidden feature $\mathbf{H}_{\text{ans}}^{(12)}$ into an MLP, and then uses a softmax function to obtain the final answer prediction probability distribution $P_{\text{answer}}(\mathcal{A} \mid \mathbf{T})$. Note $\mathbf{T}$ here refers to the aforementioned hybrid prompt template from $(\boldsymbol{I}_i, \boldsymbol{Q}_i)$ of the i-th training instance. Formally,

$$P_{\text{answer}}(\mathcal{A} \mid \mathbf{T}) = \text{softmax}(\text{MLP}(\mathbf{H}_{\text{ans}}^{(12)})) \tag{9}$$

The answer mapping classification is a contrastive learning process over positive and negative prompt templates, which can be constrained by a cross-entropy loss $\mathcal{L}_{\text{AMC}}$,

$$\mathcal{L}_{\text{AMC}}^{+} = -\frac{1}{N}\sum\nolimits_{i}^{N}\sum\nolimits_{j}^{|\mathcal{A}|} a_{ij}^{*}\log\left(P(a_{ij} \mid \mathbf{T}_{\text{filled}}^{+})\right) \tag{10}$$

$$\mathcal{L}_{\text{AMC}}^{-} = \frac{1}{N}\sum\nolimits_{i}^{N}\sum\nolimits_{j}^{|\mathcal{A}|} a_{ij}^{*}\log\left(P(a_{ij} \mid \mathbf{T}_{\text{filled}}^{-})\right) \tag{11}$$

$$\mathcal{L}_{\text{AMC}} = \mathcal{L}_{\text{AMC}}^{+} + \mathcal{L}_{\text{AMC}}^{-} \tag{12}$$

where a_{ij} denotes the j-th answer candidate in $\mathcal{A}$ for the i-th training instance, and a_{ij}^{*} is its ground truth label.

For answer filling verification, HybridPrompt first feeds hidden feature $\mathbf{H}_{\text{cls}}^{(12)}$ into an MLP, and then further utilizes a sigmoid function to obtain the final answer verification probability distribution $P_{\text{verify}}(\mathcal{A} \mid \mathbf{T})$. Formally,

$$P_{\text{verify}}(\mathcal{A} \mid \mathbf{T}) = \text{sigmoid}(\text{MLP}(\mathbf{H}_{\text{cls}}^{(12)})) \tag{13}$$

The answer filling verification process can also be constrained by a cross-entropy loss. Formally,

$$\mathcal{L}_{\text{AFV}} = -\frac{1}{N}\sum\nolimits_{i}^{N}\sum\nolimits_{j}^{|\mathcal{A}|} \ell_{ij}^{*}\log\left(P(\ell_{ij} \mid \mathbf{T})\right) \tag{14}$$

Methods	test-dev (%)				test-std (%)
	All	Y/N	Num.	Other	All
UpDn (2018)	65.32	81.82	44.21	56.05	65.67
MuRel (2019)	68.03	84.77	49.84	57.85	68.41
DFAF (2019)	70.22	86.09	53.32	60.49	70.34
ReGAT (2019a)	70.27	86.08	54.42	60.33	70.58
MCAN (2019)	70.63	86.82	53.26	60.72	70.90
DC-GCN (2020)	71.21	87.32	53.75	61.45	71.54
LENA (2021)	69.39	85.87	49.97	59.52	69.70
MHEF (2021b)	69.91	86.80	45.52	59.90	69.94
HybridPrompt	**76.12**	**91.65**	**59.54**	**66.62**	**76.30**

Table 2: Comparison results with General-VQA methods. Bold indicates the winner.

where ℓ_{ij} is a binary value that denotes the verification result in the j-th answer candidate in $\mathcal{A}$ for the i-th training instance, while ℓ^*_{ij} is a soft pseudo-label computed by the formula:

$$\ell^*_{ij} = a^*_{ij} \log \left(P(a_{ij} \mid \mathbf{T}^+_{\text{filled}}) \right) \tag{15}$$

Similar to the above AMC loss, answer filling verification is also a contrastive learning process, and its loss can be computed by,

$$\mathcal{L}_{\text{AFV}} = \mathcal{L}^+_{\text{AFV}} + \mathcal{L}^-_{\text{AFV}} \tag{16}$$

In sum, our model is trained to minimize the following total objective:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{ATC}} + \mathcal{L}_{\text{AMC}} + \mathcal{L}_{\text{AFV}} \tag{17}$$

Experiments

Experimental Setup

Datasets. We evaluate our model on VQA v2 dataset (Goyal et al. 2017), which is the most commonly-used VQA benchmark dataset and is manually built on the images from MSCOCO (Lin et al. 2014). The dataset is split into training (83K images and 444K questions), validation (41K images and 214K questions), and test (81K images and 448K questions) sets.

Implementation Details. HybridPrompt adopts 12-layer Transformers as the backbone. The initial learning rate is set to $8e^{-5}$, and the weight decay is set as 0.01. The batch size is set to 128 and the number of iterative steps in training is set to 20000. The length of soft prompts and the number of negative templates for each instance are all set to 4. Note the hyperparameters are all tuned with grid-search over the validation set. We adopt the AdamW (Loshchilov and Hutter 2018) optimizer to optimize the model and all experiments are performed on 4 NVIDIA RTX3090 GPUs with PyTorch.

Baselines. For a more holistic and objective comparison, we compare HybridPrompt versus the latest "8+11" SOTA models, and classify them into two groups:

- General-VQA methods without pre-training, including **UpDn** (Anderson et al. 2018), **MuRel** (Cadene et al. 2019), **DFAF** (Gao et al. 2019), **ReGAT** (Li et al. 2019a), **MCAN** (Yu et al. 2019), **DC-GCN** (Huang et al. 2020), **LENA** (Han et al. 2021) and **MHEF** (Lao et al. 2021b);

Methods	test-dev (%)	test-std (%)
ViLBERT (Lu et al. 2019)	70.55	70.92
LXMERT (Tan and Bansal 2019)	72.42	72.54
UNITER (Chen et al. 2020)	72.70	72.91
OSCAR (Li et al. 2020)	73.16	73.44
ViLT (Kim, Son, and Kim 2021)	71.26	-
E2E-VLP (Xu et al. 2021)	73.25	73.67
UNIMO (Li et al. 2021b)	73.79	74.02
ROSITA (Cui et al. 2021)	73.91	73.97
ALBEF (Li et al. 2021a)	74.54	74.70
VinVL (Zhang et al. 2021)	75.95	76.12
TCL (Yang et al. 2022)	74.90	74.92
HybridPrompt (Ours)	**76.12**	**76.30**

Table 3: Comparison results with VLP-VQA methods. Bold indicates the winner.

- VLP-VQA methods based on pretraining-then-tuning, including **ViLBERT** (Lu et al. 2019), **LXMERT** (Tan and Bansal 2019), **UNITER** (Chen et al. 2020), **OSCAR** (Li et al. 2020), **ViLT** (Kim, Son, and Kim 2021), **E2E-VLP** (Xu et al. 2021), **UNIMO** (Li et al. 2021b), **ROSITA** (Cui et al. 2021), **ALBEF** (Li et al. 2021a), **VinVL** (Zhang et al. 2021) and **TCL** (Yang et al. 2022).

All baselines use the *base* model for a fair comparison.

Overall Performance Comparison

As shown in Table 2, HybridPrompt significantly outperforms the General-VQA methods and demonstrates excellent performance on the test-dev and test-std sets. Specifically, compared with the current best attention-based method MCAN, the multimodal fusion method MHEF and the graph-based method DC-GCN, HybridPrompt respectively exceeds 7.62%, 9.09% and 6.65% on accuracy score of test-std, indicating the strong application prospect of the *"pre-training + prompt-tuning "* paradigm in multimodal comprehension. Further, from the perspective of visual language pre-training, HybridPrompt also achieves competitive performance against existing VLP-VQA models and obtains the new state-of-the-art results on VQA v2 dataset, as shown in Table 3. Specifically, HybridPrompt outperforms the latest model TCL by 1.63% and 1.84% on test-dev and test-std, respectively. Compared with the best model VinVL, HybridPrompt also achieves a certain amount of improvement, which demonstrates its effectiveness.

Ablation Studies

We conduct ablation studies to evaluate the effectiveness of each component, as shown in Table 4. Specifically, $\#1$ denotes the complete model; $\#2$ w/o answer filling verification means we train the model without AFV loss; $\#3$ w/o trainable soft prompts means we remove the fixed-length soft prompts in our hybrid template to train the model; $\#4$ w/o contrastive sample filling means we remove all the negative templates and only adopt $\mathcal{L}^+_{\text{AMC}}$ and $\mathcal{L}^+_{\text{AFV}}$ losses; $\#5$ w/o adaptive topic classification means ATC loss is removed and `[TOP]` state is implicitly updated; $\#6$ w/o weak prompt concatenation means the topic, sample and attribute-aware

#	Model	Accuracy (%)			
		dev	Δ	std	Δ
1	**Complete model**	**76.12**	-	**76.30**	-
2	w/o answer filling verification	75.42	0.70	75.68	0.62
3	w/o trainable soft prompts	75.06	1.06	75.18	1.12
4	w/o contrastive sample filling	74.35	1.77	74.49	1.81
5	w/o adaptive topic classification	71.08	5.04	71.15	5.15
6	w/o weak prompt concatenation	68.21	7.91	68.35	7.95

Table 4: Ablation study on VQA v2 dataset.

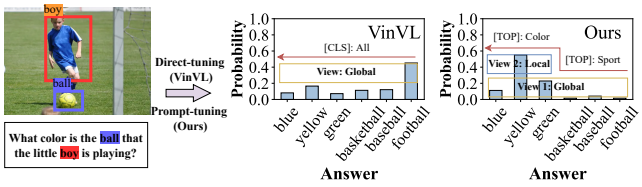


Figure 3: Case study.

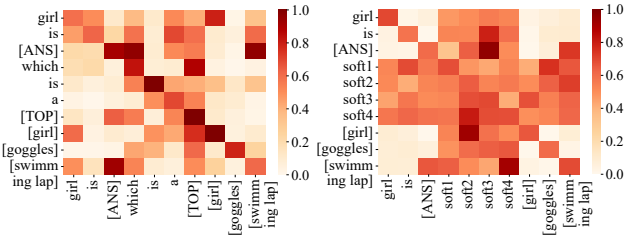
hybrid weak prompt is not added and we only adopt linguistic reorganization module to modify the inputs and predict the answer at the [ANS] position. From Table 4, we can observe that removing each component will result in a performance degradation. Particularly, w/o weak prompt adding and w/o adaptive topic classification respectively cause 7.95% and 5.15% absolute drops in accuracy for test-std, which further verifies the effectiveness of our hybrid prompting for VQA.

Further Analysis

Case Study. To verify whether our model can gradually narrow the view range over the latent answer space and perform more accurate classification, we visualize the probability distribution from our AMC layer and VinVL's classification layer. From the visualized distributions in Figure 3, we can see that compared with VinVL, HybridPrompt can more precisely predict the answer *"yellow"* by gradually narrowing the search range from *global* view to topic-related *local* view (i.e., View 1 → View 2), demonstrating its superiority in obtaining reliable answer for VQA.

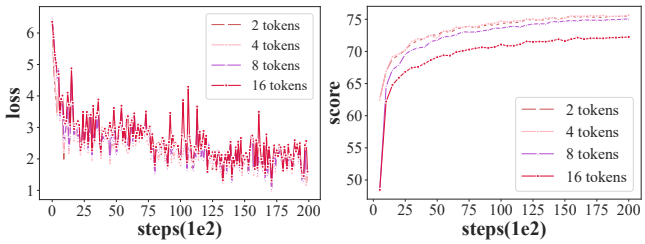
Attention Visualization. To more clearly illustrate what the hard prompts and soft prompts have learned, we visualize the attention weights from the last Transformer-layer in TSP, as shown in Figure 4. From Figure 4a, we can observe that HybridPrompt's [ANS] slot is very good at gaining knowledge from [TOP] prompt as well as cross-modal visual prompt (i.e., *"swimming lap"*), which demonstrates the effectiveness of our hard prompting. From Figure 4b, we can see that *soft prompts* receive most of the attention from other tokens, which validates its latent ability in learning personalized semantics of the whole sentence.

The Length of Soft Prompts. To further explore the influence of different lengths of the soft prompts on model training, we conduct this set of hyperparameter experiment. From Figure 5a, we can see that with the increase of training steps, all models can effectively converge. In particular, the longer



(a) Hard Prompting. (b) Soft Prompting.

Figure 4: Attention visualization.



(a) Loss decrease curve. (b) Accuracy increase curve.

Figure 5: Loss decrease and accuracy increase curves under different soft prompt lengths.

the soft-prompt length is, the more steps are needed for convergence. For example, the soft prompts with 4 tokens will converge after 10,000 steps, while the soft prompts with 16 tokens requires 17500 steps for convergence. It also suggests that there may be more implicit knowledge to be learned. From Figure 5b, we can further see that for the VQA task, it is not the longer the soft prompt length, the higher the accuracy. Especially for the soft prompts with 16 tokens, it only achieves about 70% accuracy when converges. We guess the reason might be that too-long soft prompts will interfere with the attention among other tokens, making it difficult to accurately predict answer from the [ANS] hidden state.

Conclusion

In this paper, we propose HybridPrompt, a *cloze-* and *verify-style* hybrid prompt framework with bridging language models and human priors in prompt tuning for VQA. Specifically, we first modify the input questions into the *cloze-style* prompts to mitigate the gap between upstream pre-training tasks and downstream VQA task. Then, we further propose a dynamically learnable hybrid prompt template for accurate and reliable answer prediction. Experiments on commonly-used VQA v2 dataset demonstrate the effectiveness of HybridPrompt, showing that it outperforms previous VQA methods and obtains new state-of-the-art results. For future work, we intend to explore human prior-guided prompt-tuning approaches from a visual perspective. We also plan to develop non-template prompt generation techniques to see if better performance can be achieved.

Acknowledgements

We would like to thank all anonymous reviewers for their valuable comments. The work was partially supported by the National Natural Science Foundation of China under Grant No. 62272176 and 61672252.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of CVPR*, 6077–6086.

Ben-Younes, H.; Cadene, R.; Cord, M.; and Thome, N. 2017. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of ICCV*, 2612–2620.

Ben-Younes, H.; Cadene, R.; Thome, N.; and Cord, M. 2019. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In *Proceedings of the AAAI*, volume 33, 8102–8109.

Cadene, R.; Ben-Younes, H.; Cord, M.; and Thome, N. 2019. Murel: Multimodal relational reasoning for visual question answering. In *Proceedings of CVPR*, 1989–1998.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *Proceedings of ECCV*, 104–120.

Cui, Y.; Yu, Z.; Wang, C.; Zhao, Z.; Zhang, J.; Wang, M.; and Yu, J. 2021. ROSITA: Enhancing Vision-and-Language Semantic Alignments via Cross-and Intra-modal Knowledge Integration. In *Proceedings of ACM MM*, 797–806.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of ICLR*.

Fukui, A.; Park, D. H.; Yang, D.; Rohrbach, A.; Darrell, T.; and Rohrbach, M. 2016. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In *Proceedings of EMNLP*, 457–468.

Gao, P.; Jiang, Z.; You, H.; Lu, P.; Hoi, S. C.; Wang, X.; and Li, H. 2019. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of CVPR*, 6639–6648.

Gao, T.; Fisch, A.; and Chen, D. 2021. Making Pre-trained Language Models Better Few-shot Learners. In *Proceedings of the ACL*, 3816–3830.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of CVPR*, 6904–6913.

Han, Y.; Guo, Y.; Yin, J.; Liu, M.; Hu, Y.; and Nie, L. 2021. Focal and Composed Vision-semantic Modeling for Visual Question Answering. In *Proceedings of ACM MM*, 4528–4536.

Huang, Q.; Wei, J.; Cai, Y.; Zheng, C.; Chen, J.; Leung, H.-f.; and Li, Q. 2020. Aligned dual channel graph convolutional network for visual question answering. In *Proceedings of ACL*, 7166–7176.

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

Jiang, Z.; Xu, F. F.; Araki, J.; and Neubig, G. 2020. How can we know what language models know? *In Journals of TACL*, 8: 423–438.

Khademi, M. 2020. Multimodal neural graph memory networks for visual question answering. In *Proceedings of ACL*, 7177–7188.

Kim, J.; On, K. W.; Lim, W.; Kim, J.; Ha, J.; and Zhang, B. 2017. Hadamard Product for Low-rank Bilinear Pooling. In *ICLR 2017*.

Kim, J.-H.; Jun, J.; and Zhang, B.-T. 2018. Bilinear attention networks. *In Proceedings of NeurIPS*, 31.

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of ICML*, 5583–5594.

Lao, M.; Guo, Y.; Liu, Y.; Chen, W.; Pu, N.; and Lew, M. S. 2021a. From Superficial to Deep: Language Bias driven Curriculum Learning for Visual Question Answering. In *Proceedings of ACM MM*, 3370–3379.

Lao, M.; Guo, Y.; Pu, N.; Chen, W.; Liu, Y.; and Lew, M. S. 2021b. Multi-stage hybrid embedding fusion network for visual question answering. *Neurocomputing*, 423: 541–550.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the EMNLP 2021*, 3045–3059.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021a. Align before fuse: Vision and language representation learning with momentum distillation. *In Proceedings of NeurIPS*, 34.

Li, L.; Gan, Z.; Cheng, Y.; and Liu, J. 2019a. Relation-aware graph attention network for visual question answering. In *Proceedings of ICCV*, 10313–10322.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019b. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Li, W.; Gao, C.; Niu, G.; Xiao, X.; Liu, H.; Liu, J.; Wu, H.; and Wang, H. 2021b. UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning. In *Proceedings of ACL*, 2592–2607.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of ECCV*, 121–137.

Li, X. L.; and Liang, P. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the ACL*, 4582–4597.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proceedings of ECCV*, 740–755.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2021. Pre-train, prompt, and predict: A systematic survey

of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Liu, Y.; Wei, W.; Peng, D.; and Zhu, F. 2022. Declaration-based Prompt Tuning for Visual Question Answering. *arXiv preprint arXiv:2205.02456*.

Loshchilov, I.; and Hutter, F. 2018. Decoupled Weight Decay Regularization. In *Proceedings of ICLR 2018*.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Proceedings of NeurIPS*, 13–23.

Ma, Z.; Li, J.; Li, G.; and Cheng, Y. 2022a. UniTranSeR: A Unified Transformer Semantic Representation Framework for Multimodal Task-Oriented Dialog System. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 103–114.

Ma, Z.; Li, J.; Li, G.; and Huang, K. 2022b. CMAL: A Novel Cross-Modal Associative Learning Framework for Vision-Language Pre-Training. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4515–4524.

Petroni, F.; Rocktäschel, T.; Riedel, S.; Lewis, P.; Bakhtin, A.; Wu, Y.; and Miller, A. 2019. Language Models as Knowledge Bases? In *Proceedings of EMNLP*, 2463–2473.

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

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Proceedings of NeurIPS*, 91–99.

Shin, T.; Razeghi, Y.; Logan IV, R. L.; Wallace, E.; and Singh, S. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of EMNLP*, 4222–4235.

Tan, H.; and Bansal, M. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proceedings of EMNLP*, 5099–5110.

Xu, H.; Yan, M.; Li, C.; Bi, B.; Huang, S.; Xiao, W.; and Huang, F. 2021. E2E-VLP: End-to-End Vision-Language Pre-training Enhanced by Visual Learning. In *Proceedings of ACL*, 503–513.

Yang, J.; Duan, J.; Tran, S.; Xu, Y.; Chanda, S.; Chen, L.; Zeng, B.; Chilimbi, T.; and Huang, J. 2022. Vision-Language Pre-Training with Triple Contrastive Learning. *arXiv preprint arXiv:2202.10401*.

Yang, Z.; He, X.; Gao, J.; Deng, L.; and Smola, A. 2016. Stacked attention networks for image question answering. In *Proceedings of CVPR*, 21–29.

Yu, Z.; Yu, J.; Cui, Y.; Tao, D.; and Tian, Q. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of CVPR*, 6281–6290.

Yu, Z.; Yu, J.; Fan, J.; and Tao, D. 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of ICCV*, 1821–1830.

Yu, Z.; Yu, J.; Xiang, C.; Fan, J.; and Tao, D. 2018. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE TNNLS*, 29(12): 5947–5959.

Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; and Gao, J. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of CVPR*, 5579–5588.

Zhong, Z.; Friedman, D.; and Chen, D. 2021. Factual Probing Is [MASK]: Learning vs. Learning to Recall. In *Proceedings of NAACL*, 5017–5033.