A Good Prompt Is Worth Millions of Parameters: Low-resource Prompt-based Learning for Vision-Language Models

Anonymous ACL submission

Abstract

Large pre-trained vision-language (VL) models can learn a new task with a handful of examples and generalize to a new task without fine-tuning. However, these VL models are hard to deploy for real-world applications due to their impractically huge sizes and slow inference speed. To solve this limitation, we study prompt-based low-resource learning of VL tasks with our proposed method, FEWVLM, relatively smaller than recent few-shot learners. For FEWVLM, we pre-train a sequence-to-sequence transformer model with prefix language modeling (PrefixLM) and masked language modeling (MaskedLM). Furthermore, we analyze the effect of diverse prompts for few-shot tasks. Experimental results on VQA show that FEWVLM with prompt-based learning outperforms Frozen (Tsimpoukelli et al., 2021) which is 31× larger than FEWVLM by 18.2% point and achieves comparable results to a 246× larger model, PICa (Yang et al., 2021). In our analysis, we observe that (1) prompts significantly affect zero-shot performance but marginally affect few-shot performance, (2) models with noisy prompts learn as quickly as hand-crafted prompts given larger training data, and (3) MaskedLM helps VQA tasks while PrefixLM boosts captioning performance.

1 Introduction

Fine-tuning large pre-trained language models (PLMs) have led to strong results in various domains including vision-language tasks (Devlin et al., 2018; Raffel et al., 2019; Brown et al., 2020; Radford et al., 2021). Such large PLMs can learn a new task with a few examples or generalize to a new task without fine-tuning on any training examples, i.e., few-shot and zero-shot learning (Brown et al., 2020; Radford et al., 2021; Tsimpoukelli et al., 2021). Few-shot learning overcomes the challenges of data-hungry supervised learning, where collecting human-labeled data is costly and slow. However, recent few-shot models such as GPT3 (Brown et al., 2020), Frozen (Tsimpoukelli et al., 2021), and PICa (Yang et al., 2021) are too large to deploy in small or moderate computing machines due to their gigantic model sizes

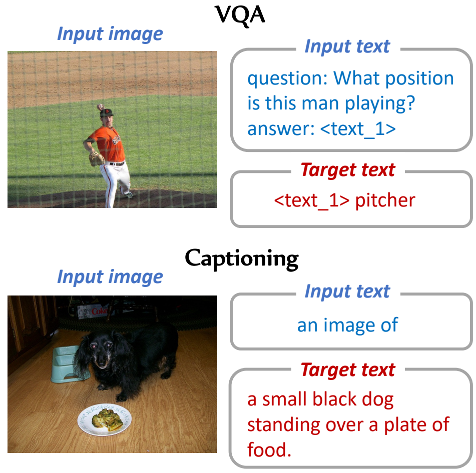

Figure 1: **Examples of VQA and Captioning tasks.** In our setup, we convert the tasks into generative tasks in which models need to generate target text given input text and an image.

In this paper, we study low-resource learning of VL tasks with our proposed method, FEWVLM, a moderate-sized vision-language model, in which we fine-tune the model with no or a handful of training examples. For FEWVLM, we pre-train a sequence-to-sequence transformer model (Cho et al., 2021; Raffel et al., 2019) with prefix language modeling (PrefixLM) and masked language modeling (MaskedLM). This setup is more practical in that training and inference can be run economically using standard computing hardware and it is expensive to obtain a large number of quality training examples in the real world. In such a few-shot setting, task-specific prompts or task descriptions are important and have shown effectiveness in few-shot NLP tasks (Gao et al., 2020; Radford et al., 2021; Schick and Schütze, 2020a,b; Brown et al., 2020).

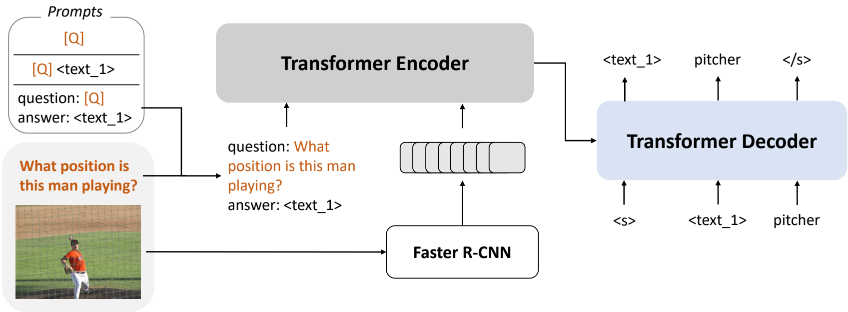

Figure 2: **Illustration of FewVLM.** This shows inference of FewVLM with prompt-based learning. Given a prompt template, we convert the question text into input text. The prompt helps the model generate correct answers.

To extend the success to VL tasks, we aim to answer the following questions for prompt-based low-resource VL learning. Q1) How does prompt design affect zero/few-shot learning on new tasks? Q2) Does prompt design still matter given larger training? Q3) How do different pre-training objectives affect zero/few-shot learning? To answer these questions, we explore various prompt formats including hand-crafted and noisy prompts on zero/few-shot VL learning datasets. In addition, we study pre-training objectives on few-shot tasks inspired by Raffel et al. (2019): prefix language modeling (PrefixLM) inspired by Raffel et al. (2019) and masked language modeling (MaskedLM). To this end, we investigate the model's performance on few-shot VL tasks including visual question answering (Goyal et al., 2017; Marino et al., 2019; Hudson and Manning, 2019), captioning (Agrawal et al., 2019; Young et al., 2014) (Fig. 1), and miniImageNet (Vinyals et al., 2016).

In our empirical analysis, our FewVLM with prompt-based learning outperforms Frozen (Tsimpoukelli et al., 2021) which is 31× larger than FewVLM by 18.2% point on zero-shot VQAv2 and achieves comparable results to a 246× larger model, PICa (Yang et al., 2021). Furthermore, we observe that (1) prompts significantly affect zero-shot performance but marginally affect few-shot performance on new tasks (§6.2 and §6.3), (2) models with noisy prompts learn as quickly as hand-crafted prompts given larger training data (§6.5), and (3) MaskedLM helps few-shot VQA tasks while PrefixLM boosts captioning performance (§6.6).

2 Related Work

Vision-language few-shot learning. Recently, several few-shot learners on vision-language tasks were proposed including GPT (Radford et al., 2019; Brown et al., 2020), Frozen (Tsimpoukelli et al., 2021), PICa (Yang et al., 2021), and SimVLM (Wang et al., 2021). Frozen (Tsimpoukelli et al., 2021) is a large language model based on GPT-2 (Radford et al., 2019), and is transformed into a multimodal few-shot learner by extending the soft prompting to incorporate a set of images and text. Their approach shows the few-shot capability on visual question answering and image classification tasks. Similarly, PICa (Yang et al., 2021) uses GPT-3 (Brown et al., 2020) to solve VQA tasks in a few-shot manner by providing a few in-context VQA examples. It converts images into textual descriptions so that GPT-3 can understand the images. SimVLM (Wang et al., 2021) is trained with prefix language modeling on weakly-supervised datasets. It demonstrates its effectiveness on a zero-shot captioning task. While these models achieve improvement on few-shot tasks, they are impractical to use in real-world applications due to their model sizes.

Language model prompting. Providing prompts or task descriptions play an vital role in improving pre-trained language models in many tasks (Gao et al., 2020; Radford et al., 2021; Schick and Schütze, 2020a,b; Brown et al., 2020). Among them, GPT models (Radford et al., 2019; Brown et al., 2020) achieved great success in prompting or task demonstrations in NLP tasks. In light of this direction, prompt-based approaches improve small pre-trained models in few-shot text classification tasks (Gao et al., 2020; Schick and Schütze, 2020a,b). CLIP (Radford et al., 2021) also explores prompt templates for image classification which affect zero-shot performance. We follow these core ideas so we aim to improve zero-shot and few-shot performance using prompts in vision-language tasks.

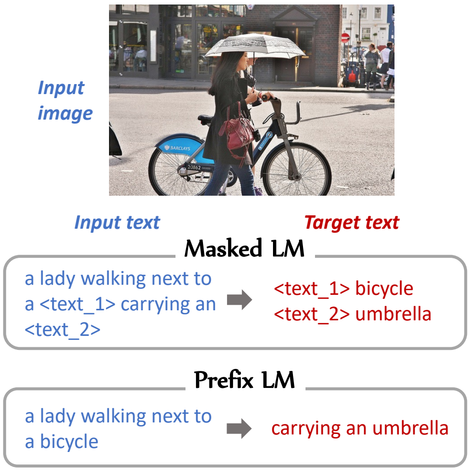

Figure 3: **Pre-training objectives.** We pre-train FEWVLM with masked language modeling (MaskedLM) and prefix language modeling (PrefixLM).

3 Analysis Setup

In this work, we study the zero-shot and few-shot performance of vision-language models $\mathcal{L}$. We introduce our analysis setup: problem formulation, analysis questions, downstream tasks and datasets, evaluation metrics, and baselines.

3.1 Problem Formulation

For zero-shot tasks, a pre-trained VL model $\mathcal{L}$ have no access to training set $\mathcal{D}_{train}$ and development set $\mathcal{D}_{dev}$, and directly makes inference on the test instances $\mathcal{D}_{test}$. For few-shot tasks, we compose a dev set $\mathcal{D}_{dev}$ from training data and ensure that $|\mathcal{D}_{train}| = |\mathcal{D}_{dev}|$ following Perez et al. (2021); Gao et al. (2020) to tune the hyper-parameters and select the model. We limit the sizes of training and development sets to meet the goal of learning from limited data. The size of $\mathcal{D}_{train}$ and $\mathcal{D}_{dev}$ are small — i.e., we set the size of both to 16 in our study.

3.2 Analysis Questions

We aim to answer the following questions in this study through experiments on multiple VL datasets.

Q1) How does prompt design affect zero/few-shot learning on new tasks? Providing a pre-trained language model with task-specific prompts or significantly improves zero-shot and few-shot performance on NLP domains (Gao et al., 2020; Schick and Schütze, 2020a,b; Brown et al., 2020). For this question, we test several ad-hoc prompts on vision-language tasks and analyze how large zero-shot and few-shot performance is affected by different prompts, hand-crafted and noisy prompts, in Sec. 6.5.

Q2) Does prompt design still matter given larger training data? As we will see in our experiments, prompts affect the zero/few-shot performance. However, prompts may have different effects when models are given different sizes of training data. To answer this question, we train models with different sizes of training data and various prompts, and compare the performance between different prompts.

Q3) How do different pre-training objectives affect zero/few-shot learning? We study two different pre-training objectives on few-shot performance: prefix language modeling (PrefixLM) inspired by Raffel et al. (2019) and masked language modeling (MaskedLM). In this setup, we pre-train our model with different objectives and test the model on zero-shot and few-shot tasks in Sec. 6.6.

3.3 Downstream Tasks and Datasets

In this work, we mainly focus on three tasks: visual question answering, captioning, and categorical learning. The visual question answering task requires models to answer a question to a given context image. We convert the visual question answering task into a generation task so that the model can generate answers in the zero-shot setting. The captioning task requires a model to generate descriptions for a given context image. The categorical learning requires a model to choose the correct category or class. We evaluate our model in an open-ended fashion to quantify fast learning of categories, in which it must generate correct labels unlike other classification methods.

We include VQAv2 (Goyal et al., 2017), OK-VQA (Marino et al., 2019), and GQA (Hudson and Manning, 2019) for visual question answering tasks, and NoCaps (Agrawal et al., 2019), and Flickr30k (Young et al., 2014) for image captioning.[1] We use Karpathy split (Karpathy and Fei-Fei, 2015) for Flickr30k, which re-splits train and val images into 29,000 / 1,014 / 1,000 for train / validation / test. For categorical learning, we include miniImageNet (Vinyals et al., 2016), a meta learning dataset. Following (Tsimpoukelli et al., 2021), we use only meta test data to evaluate FEWVLM in a few-shot manner and test on 5-way k-shot setup, where 5 classes and k examples *per class* are given.[2]

[1] We include COCO captioning results on Sec. B of Appendix.

[2] For VQA and captioning, we include k samples in total, not per class.

Table 1: **Hand-crafted prompts.** We study hand-crafted prompts on zero-shot and few-shot tasks. `[Q]` and `[A]` refer to question text and answer text, respectively. `<text_1>` is a sentinel token. We append image features to input text. Target prompts are "`[A]`" and "`<text_1> [A]`" in VQA. We use caption text as a target prompt in captioning.

Task	ID	Input prompt	Example
VQA	P1	[Q] <text_1>	**input:** What position is this man playing? <text_1> **output:** <text_1> pitcher
	P2	*question:* [Q] *answer:*	**input:** question: What position is this man playing? answer: **output:** <text_1> pitcher
	P3	*question:* [Q] *answer:* <text_1>	**input:** question: What position is this man playing? answer: <text_1> **output:** <text_1> pitcher
Captioning	Q1	*a picture of*	**input:** a picture of **output:** a small black dog standing over a plate of food.
	Q2	*a photo of*	**input:** a photo of **output:** a small black dog standing over a plate of food.
	Q3	*an image of*	**input:** an image of **output:** a small black dog standing over a plate of food.

3.4 Evaluation Metrics

To evaluate few-shot performance, we randomly sample 5 different training and dev splits and measure average performance on the 5 splits. We fine-tune the vision-language models with 200 epochs for the few-shot setup and choose the best checkpoint on the dev set. For NoCaps task, it does not have training data. Thus we use the training data from COCO captioning in the experiments following Wang et al. (2021). We evaluate on the VQAv2 validation set, GQA test-dev, OK-VQA test set, test set of Karpathy split for Flickr30k captioning, and NoCaps validation set. We adopt accuracy for VQA datasets and miniImageNet, and CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016) as evaluation metrics for captioning.

3.5 Baselines

We evaluate strong zero/few-shot vision-language learners for comparison: Frozen (Tsimpoukelli et al., 2021), PICa (Yang et al., 2021) for VQA datasets and SimVLM (Wang et al., 2021) for captioning datasets. We include Unified VLP (Zhou et al., 2020) for few-shot VQAv2 and Flickr30k. Also, we compare them with fully fine-tuned models $\mathcal{L}_{full}$ as upper bounds of few-shot models for each task; these models are fine-tuned on the entire datasets while few-shot models can access a small amount of data. For fully fine-tuned models $\mathcal{L}_{full}$, we borrow numbers from Uniter$_{large}$ (Chen et al., 2019) for VQAv2, Oscar (Li et al., 2020b) for GQA, SimVLM (Wang et al., 2021) and VinVL (Zhang et al., 2021) for NoCaps CIDER and SPICE respectively, and Unified VLP (Zhou et al., 2020) for Flickr30k captioning. We include VL-T5$_{\text{no-vqa}}$ as a baseline which is pre-trained without visual question answering datasets (Cho et al., 2021). For miniImageNet, we include Frozen and AFHN (Li et al., 2020a). Frozen is designed for few-shot learning while AFHN is for meta learning, which is smaller and faster.

4 Method

Before diving into the analysis, we introduce our model, FewVLM, to do zero/few-shot learning on VL tasks and answer the analysis questions we raised. We introduce FewVLM architecture and pre-training objectives.

4.1 Encoder-Decoder Vision-language Model

We adopt an encoder-decoder architecture (Cho et al., 2021; Vaswani et al., 2017), to encode visual and text inputs and generate target text. We represent an input image with 36 object regions from a Faster R-CNN (Ren et al., 2015) trained on Visual Genome (Krishna et al., 2017). The sets of region representations are fed into the encoder by appending them to the text Cho et al. (2021). We train the model parameters θ by minimizing the negative log-likelihood of target text y tokens given input text x and image v:

$$L_\theta = -\sum_{i=1}^{|y|} \log P_\theta(y_i | y_{<i}, x, v). \tag{1}$$

The model is not task-specific, so it is a good option for zero/few-shot settings.

4.2 Pre-training Objectives

We pre-train the models with both prefix language modeling (PrefixLM) and masked language modeling (MaskedLM). Fig. 3 illustrates the PrefixLM and MaskedLM.

Prefix language modeling. We include prefix language modeling (PrefixLM) following Raffel et al. (2019). Given an image and a span of text, this objective randomly splits the text into two separate components; the former component with the given image is used as inputs to the encoder and the latter component is used as target text to be generated by the decoder.

Masked language modeling. We follow Cho et al. (2021) to do masked language modeling. This objective is to replace random spans with numbered sentinel tokens, e.g., <text_1>, and then the masked text is fed into the encoder. Then the decoder generates the masked spans as target text. We randomly mask 15% of input text tokens and replace them with sentinel tokens.

Pre-training data. To pre-train FewVLM, we collect image-caption data from MS COCO (Lin et al., 2014; Chen et al., 2015) and Visual Genome (VG) (Krishna et al., 2017).

5 Low-resource Adaptation

In downstream tasks, we train our model with few-shot examples. Fig. 2 shows an illustration of FewVLM in inference time. Given a prompt template $\mathcal{P}$, we first get input text and target text using the template $x, y = \mathcal{P}(\text{input}, \text{label})$. Then we train model parameters by minimizing the negative log-likelihood in Eq. (1). In inference, we use the same prompt and the model generates the label text. Here we obtain the final label by removing the target prompt template.

5.1 Prompt Design

Prompts affect the performance of the vision-language model (Cho et al., 2021); we study the effect of different prompts on the zero-shot and few-shot performance on downstream tasks. Tables 1 and 11 show prompts we used in our experiments.

5.1.1 Visual Question Answering.

The visual question answering tasks (VQA, OK-VQA, and GQA) require models to answer a question to a given context image. Recent approaches (Chen et al., 2019; Tan and Bansal, 2019; Su et al., 2019; Li et al., 2019, 2020b) tackle visual question answering tasks as multi-label classification over a predefined set of answer candidates. Instead, we approach the visual question answering tasks as a generation task so that the model can produce the answers without introducing any task-specific heads. In this setup, prompts act as constraints to guide the models to generate proper formats of answers; models might generate a sentence for VQA, which is not the correct format, without prompts.

Therefore, we study several prompts for input and output as shown in Tables 1 and 11; we explore hand-crafted prompts (Table 1) and noisy prompts for ablation study (Table 11).

Hand-crafted prompts. For input prompts, we explore three different templates: "*question:* `[Q]` *answer:*" and with the <text_1> sentinel token at the end. Similarly to masked language modeling, we expect models to generate words thanks to the sentinel token. For target prompts, we explore two different templates: "`[A]`" (an answer) and "<text_1> [A]" (an answer with a sentinel token). Here, we aim to mimic MaskedLM's target text format, so the similar format helps the model quickly adapt to the new task. We call each prompt ID as in Table 1.

Noisy prompts. To understand the effect of noisy prompts in zero/few-shot learning, we include irrelevant prompts, noisy tokens, and random sentences as in Table 11. Irrelevant prompts are random questions or instructions that mislead models to answer wrong questions or follow irrelevant instructions. Noisy tokens are randomly selected from T5's vocabulary, so we test how robust our model is to random tokens. Finally, random sentences are captions from MS COCO and this gives false information to models.

5.1.2 Captioning.

In NoCaps and Flickr30k, we explore three hand-crafted input prompts: "*a picture of*", "*a photo of*", and "*an image of*". We study the effect of different word choices in this captioning task. While the three different words have similar meanings, they show different performance in zero-shot and few-shot tasks as we will see in our experiments.. For target prompts, we just train the model with the original caption without any additional prompts.

5.1.3 MiniImageNet

In miniImageNet, we train our model with a hand-crafted input prompt, "*This is* <text_1>," and target prompt, "<text_1> [A]." We compare our model with and without prompts in this dataset to study whether prompts are helpful in categorical learning.

Table 2: **Zero-shot VQA results.** We test models without any training examples. VL-T5$_{\text{no-vqa}}$ is pre-trained without VQA datasets. Compared to larger models, Frozen and PICa-Full, our models outperform them or show the comparable results.

Model	Model size	VQAv2	OK-VQA	GQA
Unified VLP	122M	0.0	-	-
VL-T5$_{\text{no-vqa}}$	224M	13.5	5.8	6.3
Frozen	7B	29.5	5.9	-
PICa	175B	-	**17.5**	-
FewVLM$_{base}$	224M	43.4	11.6	27.0
FewVLM$_{large}$	740M	**47.7**	*16.5*	**29.3**

Table 3: **Few-shot VQA results.** We report average performance over 5 different splits. The size of training and validation sets are 16 for our FewVLM and VL-T5$_{\text{no-vqa}}$, and Frozen and PICa use 4 and 16 in-context training examples, respectively. For the fair comparison to Frozen, we include FewVLM$^{*}_{base}$ with 4 training and validation examples.

Model	Model size	VQAv2	OK-VQA	GQA
Unified VLP	122M	24.3	-	-
VL-T5$_{\text{no-vqa}}$	224M	31.8	12.7	19.6
Frozen	7B	38.2	12.6	-
PICa	175B	**54.3**	**43.3**	-
FewVLM$^{*}_{base}$	224M	45.1	14.5	26.9
FewVLM$_{base}$	224M	48.2	15.0	32.2
FewVLM$_{large}$	740M	*51.1*	*23.1*	**35.7**
Fine-tuned $\mathcal{L}_{full}$	-	72.6	-	61.5

Table 4: **Zero-shot captioning results.** We use the CIDEr and SPICE metrics for evaluation.

Model	Model size	NoCaps		Flickr30k	
		CIDEr	SPICE	CIDEr	SPICE
Unified VLP	122M	-	-	24.9	7.2
VL-T5$_{\text{no-vqa}}$	224M	4.4	5.3	2.6	2.0
SimVLM$_{huge}$	-	**101.4**	-	-	-
FewVLM$_{base}$	224M	42.2	8.5	31.0	10.0
FewVLM$_{large}$	740M	*47.7*	**9.1**	**36.5**	**10.7**

Table 5: **Few-shot captioning results.** We report average performance over 5 different splits. We use the CIDEr and SPICE metrics for evaluation.

Model	Model size	NoCaps		Flickr30k	
		CIDEr	SPICE	CIDEr	SPICE
Unified VLP	122M	-	-	28.8	9.4
VL-T5$_{\text{no-vqa}}$	224M	22.0	6.8	12.8	8.3
FewVLM$_{base}$	224M	48.6	10.0	32.6	12.8
FewVLM$_{large}$	740M	**53.1**	**10.4**	**37.0**	**13.5**
Fine-tuned $\mathcal{L}_{full}$	-	112.2	13.1	67.4	17.0

6 Results and Discussion

In this section, we first discuss our main results on zero-shot and few-shot tasks and then answer the questions we raised: does prompt design matter in zero/few-shot learning?

6.1 Experiment Details

For pre-training, we set batch size 1,280 and 800 for FewVLM$_{base}$ and FewVLM$_{large}$, respectively and pre-train them with 30 epochs. We use learning rate 1e-4 with 5% linear warmup. For few-shot learning, we train models with 200 epochs, learning rate 5e-5 and 5% linear warmup and choose the best checkpoint on the dev set. For FewVLM, we use "question: [Q] answer <text_1>" (P3) as an input prompt and "<text_1> [A]" as a target prompt for visual question answering, and "an image of" (Q3) as an input prompt for captioning, which show the best performance. We will study the effect of different prompts in Sec. 6.5. The sizes of of $\mathcal{D}_{train}$ and $\mathcal{D}_{dev}$ are 16 on VQA and captioning tasks. For miniImageNet, we use 'This is <text_1>,' and "<text_1> [A]" as input and target prompts. In this data, we test with {1, 3, 5}-shots per class.

6.2 Performance on Zero-shot Learning

We evaluate the existing models in a zero-shot manner, in which models do not have access to any training data. Tables 2 and 4 show the results on VQA and captioning datasets, respectively. First, FewVLM with the hand-crafted prompt (P3) achieves better performance than other baselines on VQA datasets. In particular, our FewVLM$_{base}$ significantly outperforms Frozen which is about 31× larger than ours. Also, PICa based on GPT3 (Brown et al., 2020) shows the best performance on OK-VQA. It is noticeable that our FewVLM$_{large}$, the 246× smaller model, achieves the comparable result to PICa. Compared to VL-T5$_{\text{no-vqa}}$ which is the same architecture as ours, FewVLM$_{base}$ improves VQAv2 performance by about 30% point. As we will see in the later section, our pre-training objectives and the prompts boost the VQA performance. On NoCaps, SimVLM$_{huge}$ shows the best performance. Our FewVLM$_{base}$ significantly improves the performance compared to VL-T5$_{\text{no-vqa}}$. As we will see in the later section, our pre-training objectives and the prompts boost the VQA and captioning performance.

Table 6: **5-way miniImageNet results.** We evaluate FewVLM in a generative manner. The shot represents the number of training examples per class.

Model	Model size	1 shot	3 shots	5 shots
Frozen	7B	14.5	34.7	33.8
FewVLM$_{base}$ (no prompt)	224M	48.0	75.0	82.6
FewVLM$_{base}$	224M	57.0	78.0	84.2
FewVLM$_{large}$	740M	57.1	78.3	84.4
AFHN	-	62.3	-	78.1

Table 7: **Zero-shot results of hand-crafted prompts.** We test different input prompts in zero-shot predictions. We use a CIDEr metric for Flickr30k. Note that zero-shot setting does not require target prompts.

	no prompt	P1	P2	P3
VQAv2	3.7	9.9	19.0	43.4
	no prompt	**Q1**	**Q2**	**Q3**
Flickr30k	9.6	15.2	25.6	31.0

6.3 Performance on Few-shot Learning

Tables 3 and 5 show the few-shot performance on VQA and captioning datasets. Sizes of training and validation sets are 16 for FewVLM, VL-T5$_{\text{no-vqa}}$, and Unified VLP; and Frozen and PICa use 4 and 16 in-context demonstration examples, respectively.

On VQAv2 and OK-VQA, PICa shows the best performance while our FewVLM$_{large}$ achieves the comparable result on VQAv2. OK-VQA requires external knowledge to answer unlike other VQA datasets, so larger models and large pre-training data (prior knowledge) are necessary to improve. Interestingly, FewVLM$^{*}_{base}$, which is trained with 4 training examples, outperforms Frozen. On captioning data, FewVLM$_{base}$ notably outperforms VL-T5$_{\text{no-vqa}}$ by 31.1% point on NoCaps CIDEr.

Unified VLP slightly underperforms FewVLM on Flickr30k captioning task. We conjecture that their architecture is based on a encoder-decoder transfomer and it is pre-trained with a captioning task (Zhou et al., 2020).

6.4 MiniImageNet

Table 6 shows results on miniImageNet, where models must choose the correct class for each image. We train and evaluate FewVLM in an generative manner; the model must generate correct label text to get the credit. FewVLM significantly outperforms Frozen in all shots. Note that we train FewVLM with a few training samples while Frozen uses them as in-context demonstration. Interestingly, FewVLM with a hand-crafted prompt improves performance a lot on the 1-shot case, while it marginally improves on the 5-shot case.

6.5 Study of Prompt Design

Here we examine the effect of different prompts on FewVLM$_{base}$ in Table 7 and Figs. 6, 5, and 4. We test the model on VQAv2 and Flickr30k datasets.

6.5.1 Zero-shot Predictions

Table 7 shows the zero-shot performance on VQAv2 and Flickr30k. We observe that zero-shot results are remarkably affected by input prompts on both datasets. For input prompts, <text_1> in P1 and P3 helps the zero-shot predictions significantly compared to "no prompt" and P2. We conjecture that <text_1> guides the model to predict masked spans similarly to MaskedLM, so it improves the performance.

On Flickr30k, we examine different word choices of prompts: "a picture of" (Q1), "a photo of" (Q2), and "an image of" (Q3). For instance, using "an image of" outperforms using no prompt by 21.4 point. It is noticeable that different word choices significantly affect the zero-shot results.

6.5.2 Few-shot Predictions

We study various input prompts including irrelevant prompts, noisy tokens, and random sentences on VQAv2 (Fig. 4). First, noisy prompts and no prompt achieve near 0 accuracy on the zero-shot setting. In few-shot predictions, FewVLM with noisy prompts learns as quickly as hand-crafted prompts given larger data. For example, our model with noisy prompts achieves comparable results to the best hand-crafted prompt. Among all different types of noisy prompts, random sentences deteriorate performance the most. This is because the random sentences come from captions in MS COCO, so the model might choose the answer from wrong captions not from images. Interestingly, no prompt outperforms the other noisy prompts and even shows similar to or better than the hand-crafted prompt with larger training data. We also observe a similar phenomenon on Flickr30k; no prompt performs similar to hand-crafted prompts in Fig. 5.

In addition, we explore two different target prompts, "<text_1 [A]" and "[A]." We try to mimic the MaskedLM's target text format, so we add "<text_1" to target prompt on VQA. This might help the model's fast adaptation to a new

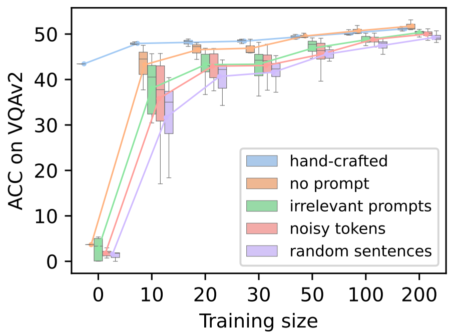

Figure 4: **VQAv2 results on noisy prompts.** We investigate different prompts on various training sizes. FEWVLM is trained with our best hand-crafted prompt (P3), irrelevant prompts, noisy tokens and random sentences. We list the prompt templates in Table 11 of appendix. We use "`<text_1> [A]`" as our target prompt.

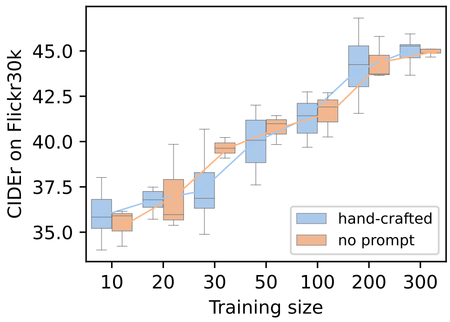

Figure 5: **Flickr30k results on hand-crafted prompts.** We investigate different hand-crafted prompts (Q1, Q2, and Q3) on various training sizes.

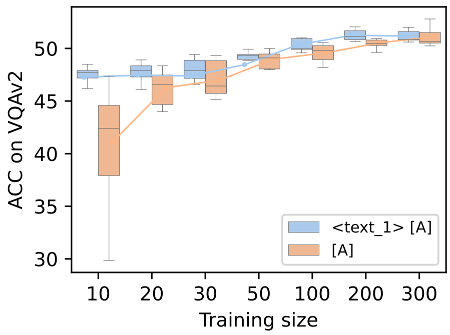

Figure 6: **VQAv2 results on different target prompts.** We investigate different target prompts with hand-crafted input prompts on various training sizes.

task since they share the same target prompt. In Fig. 6, we notice an interesting phenomenon; the target prompt "`[A]`" shows a larger variance than the other suggesting that introducing "`<text_1`" helps the model quickly adapt to a new task. However, both prompts show similar results given larger training data, e.g., 300.

Table 8: **Results on different pre-training objectives.** We test our pre-training objectives and how it affects zero-shot and few-shot performance. We train FEWVLM$_{base}$ with 16 training and validation examples.

Objective	VQAv2	GQA	Flickr30k CIDEr
Zero-shot			
MaskedLM	42.4	25.1	4.6
PrefixLM	11.9	6.7	26.8
MaskedLM + PrefixLM	**43.4**	**27.0**	**31.0**
Few-shot			
MaskedLM	46.0	31.4	18.5
PrefixLM	40.8	27.6	31.8
MaskedLM + PrefixLM	**48.2**	**32.2**	**32.6**

6.6 Pre-training Objectives

We investigate how pre-training objectives affect different tasks. We pre-train FEWVLM with different pre-training objectives: masked language modeling (MaskedLM) and prefix language modeling (PrefixLM).

In Table 8, we observe that MaskedLM helps VQA tasks while PrefixLM helps captioning tasks in zero-shot and few-shot settings. We conjecture that MaskedLM is to predict spans, which is analogous to predict correct answers to questions, and PrefixLM is to generate the rest of the given prefix, which is similar to captioning tasks. In other words, if the pre-training task is similar to the downstream tasks, then it will help performance further. When pre-training with both objectives, they create a synergetic effect and thus improve cross-task generalization.

7 Conclusion

In this work, we present FEWVLM, a few-shot prompt-based learner on vision-language tasks. On diverse datasets, FEWVLM outperforms baselines and shows comparable results to PICa which is 246× larger than ours. We observe that prompts are vital in zero-shot and few-shot tasks and each pre-training objective helps different few-shot tasks. Also, we find out that models with larger training data are not significantly affected by noisy prompts. Future work includes exploring automatic prompt generation and diverse formats of few-shot tasks such as multiple-choice VQA. Finding optimal prompts require exhaustive engineering to achieve the best performance and leads to impressive results. We leave the exploration of these directions to future investigations.

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8948–8957.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer.

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

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

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Uniter: Learning universal image-text representations.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. *arXiv preprint arXiv:2102.02779*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

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

Kai Li, Yulun Zhang, Kunpeng Li, and Yun Fu. 2020a. Adversarial feature hallucination networks for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13470–13479.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3195–3204.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. *arXiv preprint arXiv:2105.11447*.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99.

Timo Schick and Hinrich Schütze. 2020a. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.

Timo Schick and Hinrich Schütze. 2020b. It's not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29:3630–3638.

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

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2021. An empirical study of gpt-3 for few-shot knowledge-based vqa. *arXiv preprint arXiv:2109.05014*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049.

Table 9: **Model architectures.**

Hyperparameter	**FewVLM$_{base}$**	**FewVLM$_{large}$**
# Layers	12+12	24+24
Hidden dimension	768	1,024
FF hidden size	3,072	4,096
# Attention head	12	16
Attention head size	64	64

Table 10: **COCO captioning results.** We use the CIDEr and SPICE metrics for evaluation.

Model	**Model size**	**Zero-shot**		**Few-shot**	
		CIDEr	**SPICE**	**CIDEr**	**SPICE**
VL-T5$_{\text{no-vqa}}$	224M	4.9	2.0	43.0	10.8
SimVLM$_{huge}$	-	102.3	22.1	-	-
FewVLM$_{base}$	224M	84.5	8.0	98.7	18.9
FewVLM$_{large}$	740M	92.1	17.3	100.4	19.1
Unified VLP (fully supervised)	122M	-	-	117.7	21.3

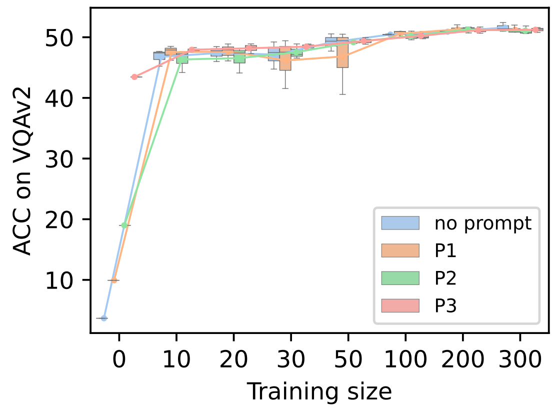

Figure 7: **VQAv2 results on hand-crafted prompts and the target prompt "<text_1> [A]".**

A Model Architectures

Table 9 shows model parameters in our model, FewVLM. FewVLM$_{base}$ and FewVLM$_{large}$ is based on VL-T5 (Cho et al., 2021) and T5 (Raffel et al., 2019), respectively.

B COCO Captioning

We evaluate our model with COCO captioning data. We use Karpathy split (Karpathy and Fei-Fei, 2015) for MS COCO captioning, which re-splits train and val images into 113,287 / 5000 / 5000 for train / validation / test. Table 10 shows the results on COCO.

C Prompt Study

Tables 7, 8, and 9 show the results of each prompt on VQAv2 and Flickr30k with various training sizes.

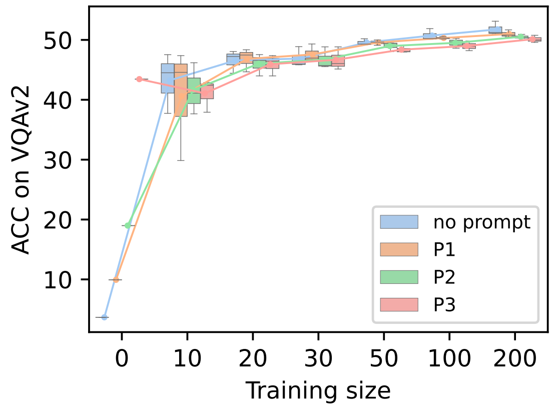

Figure 8: **VQAv2 results on hand-crafted prompts and the target prompt "[A]"**

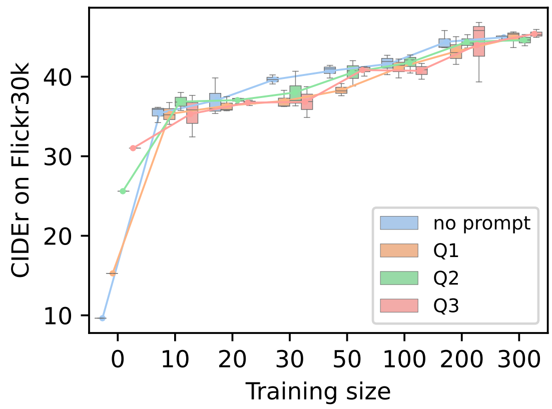

Figure 9: **Flickr30k results on hand-crafted prompts.**

D Effect of pre-training Data

We pre-train our model with different datasets: MS COCO and Visual Genome (VG), and Conceptual Captions (CC). We investigate which pre-training dataset helps the downstream tasks in a few-shot manner. In Table 12, we observe that MS COCO and VG datasets are more helpful to the downstream tasks than CC.

Table 11: **Prompt templates.** We test different input prompts on VQAv2. `[Q]` refers to input question text. We use <text_1> [A] as target text. We append image features to input text.

Input prompt template	**Category**
Fill in the blank in the below sentence: [Q]	irrelevant prompts
Question: [Q] True or False?	irrelevant prompts
[Q] What color is the floor?	irrelevant prompts
Paraphrase this into a different question? [Q]	irrelevant prompts
[Q] How many are they?	irrelevant prompts
nezg publice passed Dream [Q]	noisy tokens
benefic video starting garbagetap Talent summary [Q]	noisy tokens
gestion Bun dates youngest batteriesfeder organisationoyez [Q]	noisy tokens
[Q] chefernției geekutilisées plantingasta Pest principiiMF saddle véritable	noisy tokens
[Q] composant emergency laissé Klägereiniger swipe concentrateOSS/18 rewardprepaid	noisy tokens
[Q] A black dog is sitting on a couch.	random sentences
[Q] A man working at a kitchen counter in a room illuminated by sunlight.	random sentences
A brown purse is sitting on a green bench. [Q]	random sentences
A television that is sitting next to signs. [Q]	random sentences
[Q] A woman is wearing white pants.	random sentences

Table 12: **Results on different pre-training datasets.** We examine different pre-training datasets on each downstream tasks.

Dataset	**VQAv2**	**GQA**	**Flickr30k**
MS COCO, VG	48.2	32.2	32.6
Conceptual Captions	36.7	25.9	22.3