Mixture-of-Prompt-Experts for Multi-modal Semantic Understanding

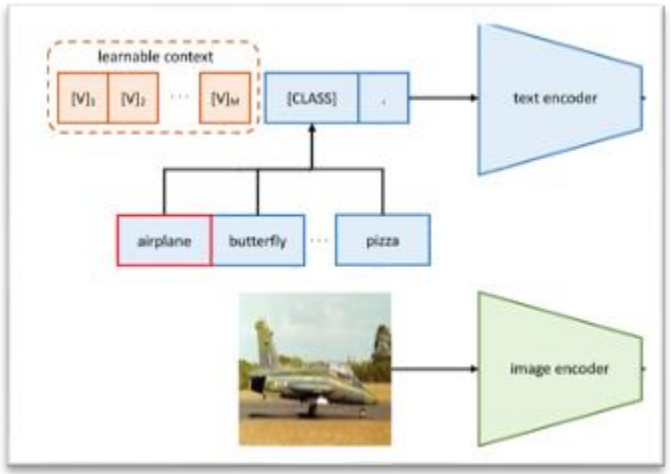
Zichen Wu, Hsiu-Yuan Huang, Fanyi Qu and Yunfang Wu

Peking University

北京大学 PEKING UNIVERSITY

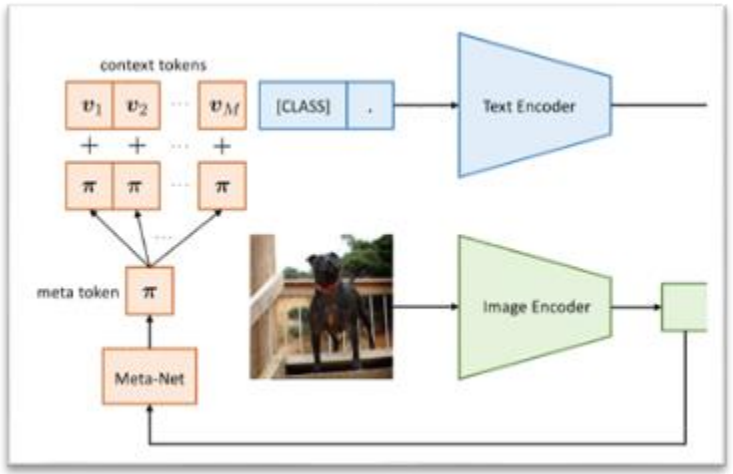
Motivation

- **Prompt tuning** in few-shot settings, especially **continuous prompt tuning**, is favored since it is data-efficient, preserves generalization and computationally feasible.
- In multimodal field, continuous prompt methods in **CLIP**-series model is largely explored



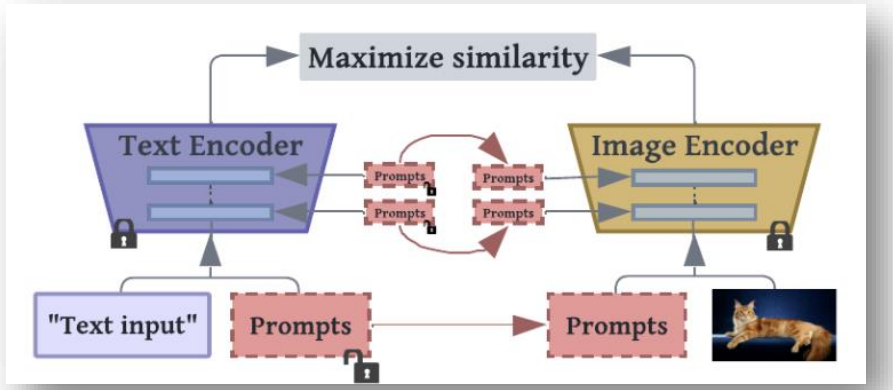
CoOP (Zhou et al. 2022)

Single modal prompt,
w/o modality interaction



CoCoOP (Zhou et al. 2022)

Single modal prompt,
w/ modality interaction

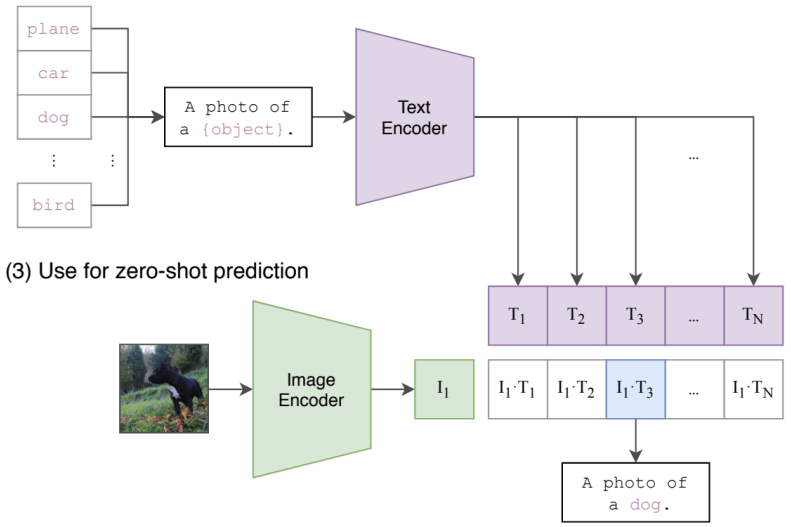


MaPLE (Khattak et al. 2023)

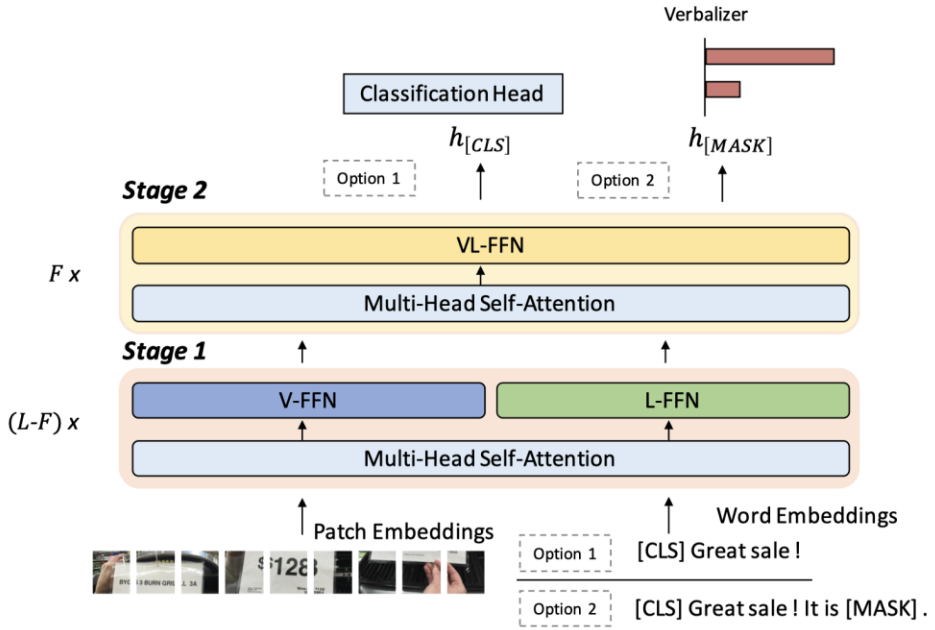
Multi modal prompt,
w/ modality interaction

Motivation

- We aim to extend the multimodal prompting research on two aspects
  - **Backbone Architecture**: Dual Encoder -> Unified Encoder
  - **Task**: Shallow multimodal task (Image Classificaion) -> Deeper multimodal semantic understanding



CLIP, dual encoder



VLMo, Unified Encoder

Motivation

- In tasks involving image-text semantic understanding, three situations may be encountered:
  - The label information is contained **within the text**.
  - The label information is contained **within the image**.
  - The label information emerges from **the interaction between the text and the image**.

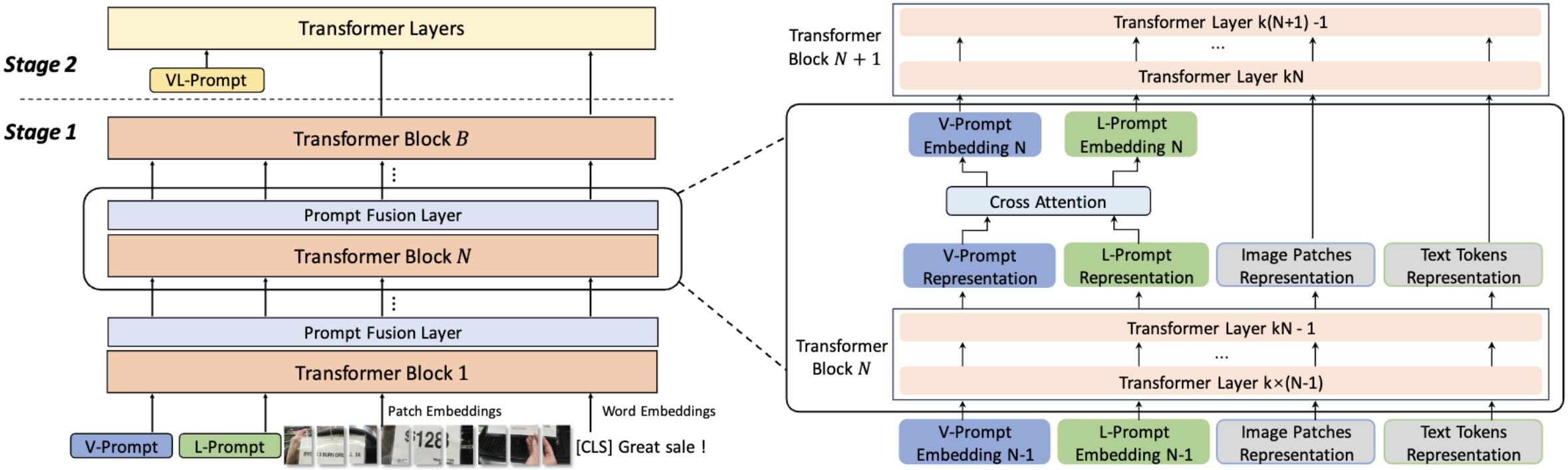
- We design three **prompt experts**, to deal with the three situations
  - **Image and text prompt expert**, extract semantic features within a single modality
  - **Unified prompt expert**, capturing inter-modality Information

- To enable a smooth transition between the single-modal and unified prompts, a **block-aware cross-attention** between prompt experts is introduced.

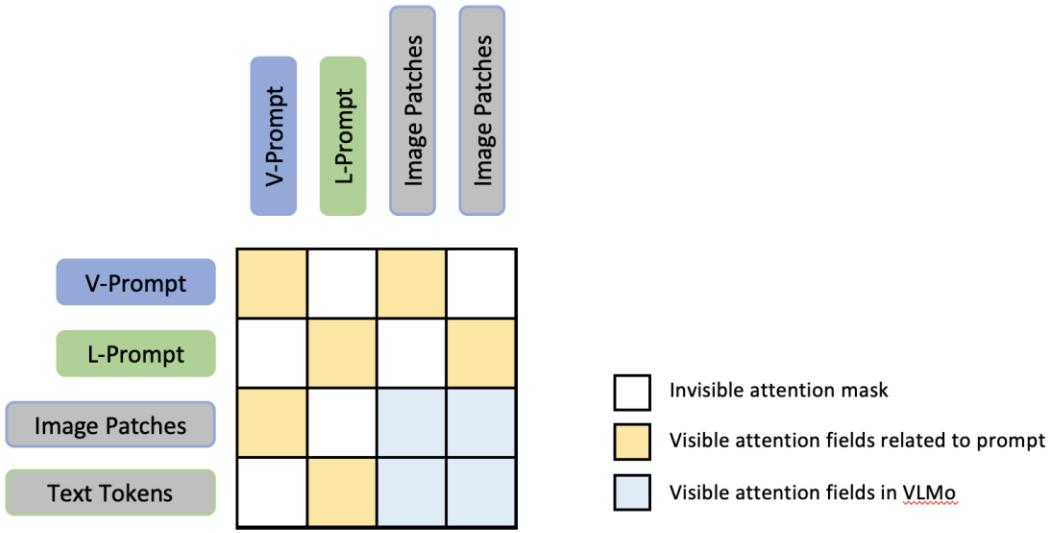
Contribution

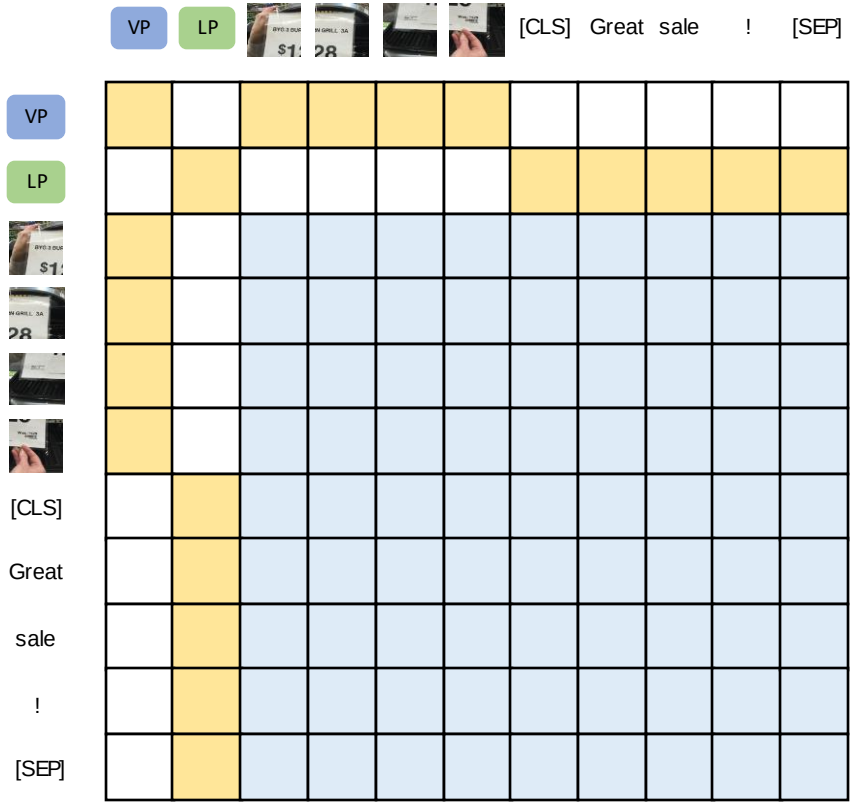
- Propose a novel mixture-of-prompt-experts, filling the blank in multi-modal continuous prompting on unified VLMs.
- Present a block-aware prompt fusion mechanism, activates deep interactions and balances the twin objectives of single-model specialization and multi-modal fusion.
- We conduct experiments on the few-shot multi-modal sarcasm detection and sentiment analysis, outperforming the previous state-of-the-art methods.

Model



Prompt Attention





Example

Experiments

- Task: Multi-modal Sarcasm Detection (Primary), Multi-modal Sentiment Analysis
- Backbone model: VLMo
- Data:
  - For MSD, we sampled 32 shots from MSDT (Cai et al., 2019) to form training/validation set.

	Full Split		Few-shot Split		Avg. length
	#Sarcasm	#Nonsarcasm	#Sarcasm	#Nonsarcasm	
Train	8642	11174	16	16	21.85
Dev	959	1451	16	16	21.79
Test	959	1450	959	1450	22.22

Task	MSD	MSA
Image		
Text	Great Sale!	Wish you a good Valentines Day.

  - For MSA, we use the sampled data from MVSA-S as Yu et al. 2022.
  - Metric: Accuracy, F1 (Macro-F1 for MSA)

Comparing Methods (Part)

- Multimodal Pre-trained Methods: CLIP, VLMo, InstructBLIP(VLLM)
- Multimodal Prompting Methods (on CLIP): CoOP, MaPLE, CMPA
- Multi-modal Sarcasm Detection methods: HFM, ResBERT, HKE
- Prompting Methods implemented on VLMo: Soft Prompt, P-Tuning, P-Tuning v2

	MSD	**MSA**
Manual Prompt Soft Prompt P-Tuning	The image-text pair is [MASK]. <text> [V1] [V2] ... [Vn] <text> [V1] [V2] ... [Vn] The image-text pair is [MASK]. <text>	Sentiment of the text: [MASK]. <text> [V1] [V2] ... [Vn] <text> [V1] [V2] ... [Vn] Sentiment of the text: [MASK]. <text>
Instruction Prompt	1.Text:<text> Answer the question: Is this image-text pair sarcastic or nonsarcastic? Answer: 2.Text:<text> Based on the image and text, answer the question: Is this image-text pair sarcastic or nonsarcastic? Answer: 3.Text: <text> Combining the text, is this sarcastic or nonsarcastic? Answer:	1.Text:<text> Answer the question: Which sentiment does this image-text pair contain, negative, neutral or positive? Answer: 2.Text:<text> Based on the image and text, answer the question: Which sentiment does this image-text pair contain, negative, neutral or positive? Answer: 3.Text:<text> Combining the text, which sentiment does this contain, negative, neutral or positive? Answer:

Results

- On sampled MSDT

Methods	w/ CH	w/ LMH	Accuracy	Precision	Recall	F1
HFM (Cai et al., 2019)	✓		58.88 (2.93)	48.93 (2.00)	65.83 (12.28)	55.73 (4.51)
resBERT (Pan et al., 2020)	✓		58.85 (2.04)	48.95 (1.62)	70.87 (4.61)	57.82 (0.45)
HKE (Liu et al., 2022a)	✓		60.76 (4.00)	52.56 (3.41)	74.43 (14.13)	61.02 (3.80)
VLMo (Bao et al., 2022)	✓		60.18 (3.62)	50.97 (4.26)	60.31 (5.28)	53.96 (3.65)
CLIP (Radford et al., 2021)	✓		61.31 (1.28)	51.24 (1.30)	68.33 (8.26)	58.26 (2.22)
InstructBLIP (zero-shot)		✓	57.59 (2.87)	47.22 (3.12)	50.33 (13.25)	47.89 (5.49)
InstructBLIP (Dai et al., 2023)		✓	60.56 (4.02)	50.57 (3.29)	**78.13** (5.94)	61.05 (2.11)
CoOp (Zhou et al., 2022b)	✓		63.00 (5.99)	53.06 (6.42)	67.71 (4.69)	59.41 (5.23)
CMPA (Liu et al., 2023b)	✓		59.94 (2.34)	49.75 (2.35)	63.43 (4.08)	55.75 (2.86)
MaPLE (Khattak et al., 2023)	✓		61.28 (2.43)	50.87 (1.95)	78.00 (6.42)	61.26 (2.80)
VLMo + Manual Prompt		✓	59.85 (2.77)	50.12 (2.64)	73.20 (8.25)	59.10 (1.15)
+ Soft Prompt	✓		62.74 (0.65)	53.26 (1.56)	57.84 (11.09)	54.74 (4.71)
+ P-Tuning (Liu et al., 2021b)		✓	60.34 (1.61)	50.11 (1.31)	75.46 (4.03)	60.21 (2.06)
+ P-Tuning v2 (Liu et al., 2021a)	✓		61.78 (3.02)	52.23 (2.94)	63.99 (15.12)	56.36 (5.98)
VLMo + MoPE-BAF	✓		64.06 (0.71)	53.69 (4.87)	71.60 (2.78)	61.32 (2.57)
+ MoPE-BAF + MP		✓	**65.32** (3.30)	**55.92** (3.91)	68.64 (7.52)	**61.32** (1.15)

Table 4: Experimental results on the MSDT dataset with 32 training samples, where the standard deviations are shown in parentheses. CH refers to the classification head and LMH refers to the language modeling head. MP means Manual Prompt.

Results

- Ablation Study

Methods	A	P	R	F1
VLMo	60.18	50.97	60.31	53.96
+ MoPE	61.73	51.35	**75.13**	60.94
+ MoPE + BAF	**64.06**	**53.69**	71.60	**61.32**
VLMo + MP	59.85	50.12	**73.20**	59.10
+ MoPE	64.06	53.98	70.73	60.93
+ MoPE + BAF	**65.32**	**55.92**	68.64	**61.32**

Table 6: Ablative analysis of the MoPE and BAF modules.

- Model Analysis
  - Prompt length
  - Block number
  - Training shots

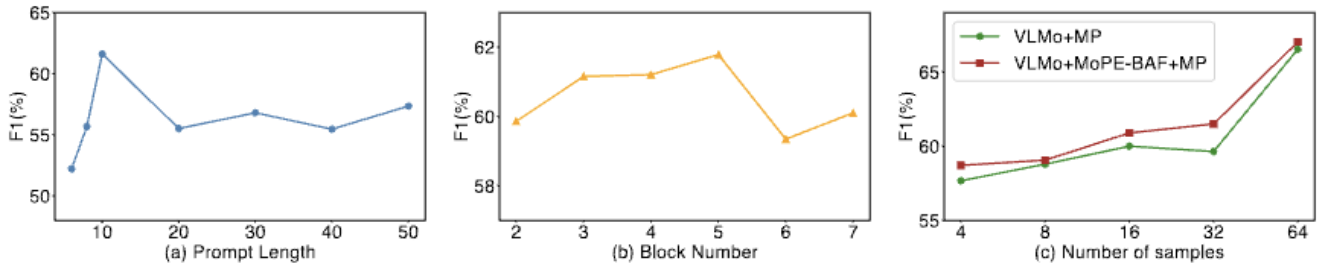


Figure 4: (a) F1 performance training MoPE with different prompt lengths. (b) F1 scores training MoPE-BAF with different block numbers. (c) Comparison between VLMo and VLMo + MoPE-BAF under different training shots.

Results

- On sampled MVSA-S

Methods	Acc	Mac-F1	Wtd-F1
MVAN	42.77	36.75	44.14
MGNNS	34.4	32.05	36.9
UP-MPF	58.21	51.08	58.49
CLIP	49.51	45.67	51.63
CoOp	51.47	40.58	48.52
MaPLe	50.49	43.06	51.74
CMPA	56.74	42.75	53.86
InstructBLIP	59.80	48.59	59.73
VLMo + MP (zero-shot)	59.07	38.53	51.94
VLMo + MP	60.79	52.62	61.27
VLMo + P-Tuning	61.03	51.28	60.75
VLMo + MoPE-BAF + MP	**63.48**	**52.92**	**62.40**

Conclusion

- MoPE-BAF, a new multi-modal soft prompt framework catering to unified VLMs for few-shot multi-modal tasks.
  - two prompt experts to serve the text and image modality separately with a better specialization ability
  - activate the interactions of prompt experts by inserting cross-modal prompt attention between adjacent Transformer blocks
- Future Work
  - incorporate task-related external knowledge into prompt design
  - MoPE has high sensitivity, how to control it?