Language Models as Black-Box Optimizers for Vision-Language Models

Samuel Yu*, Shihong Liu*, Zhiqiu Lin*, Deepak Pathak, Deva Ramanan

CMU

Presenter: Hsuan-Yu Fan

Date: 2023/9/27

Outline

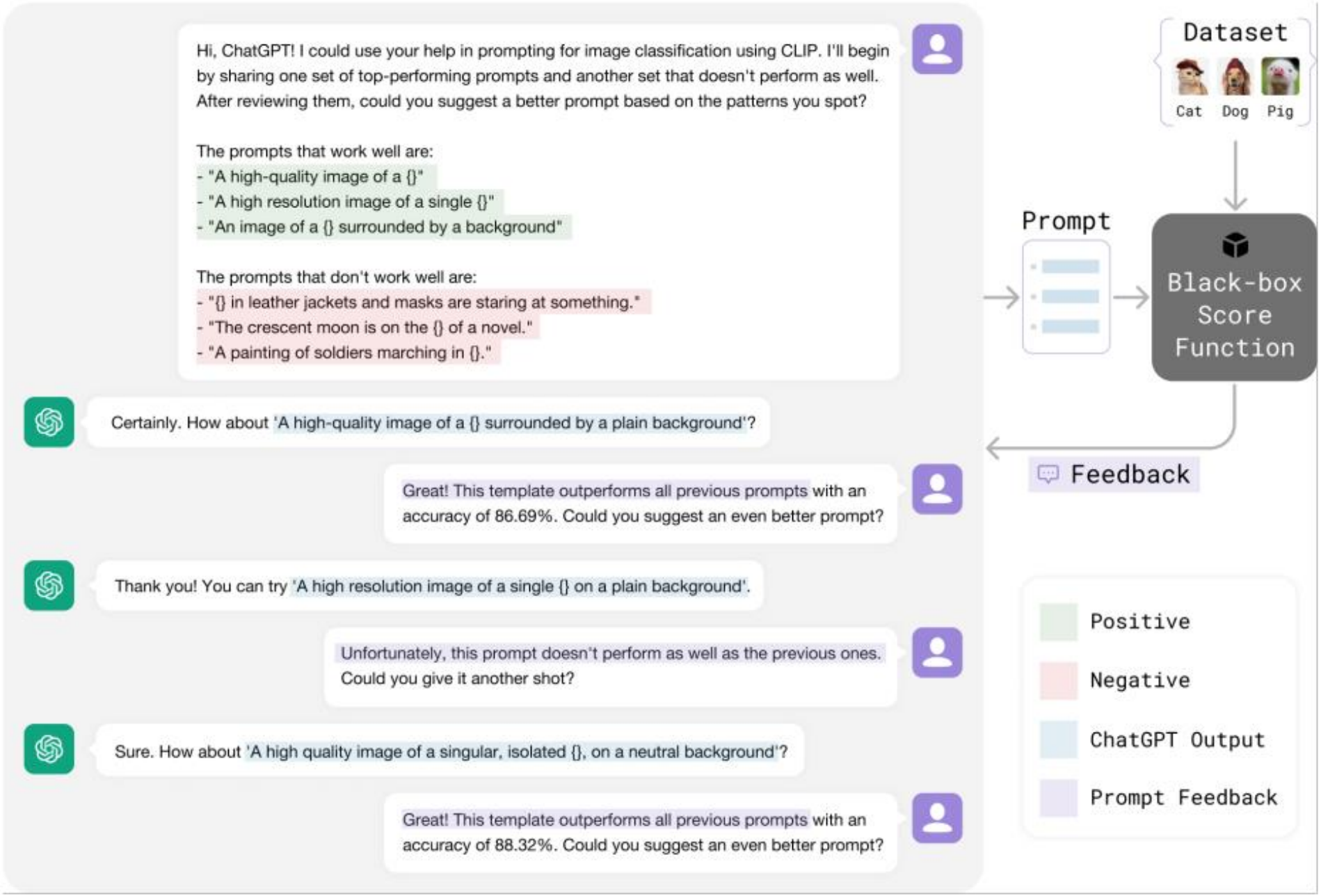
- Introduction
- Related Work
- Method
- Illustrative Task
- Additional Analysis
- Conclusion

Introduction

- Fine-tuning strategy typically relies on backpropagation, which requires transparent "white-box" access to the model weights.
- Increasing number of VLMs [1, 7, 8, 28, 29] are not releasing their weights due to privacy and legal concerns [30, 31].
- We propose employing chat-based LLMs as black-box optimizers to search for the best text prompt.
- We begin with random prompts, assess their one-shot training accuracy, and then iteratively ask ChatGPT to refine them based on the best and worst outcomes.

[1] OpenAI. Gpt-4 technical report. 2023.
[7] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems, 35:23716–23736, 2022.
[8] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. arXiv preprint arXiv:2205.14100, 2022.
[28] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917, 2022.
[29] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378, 2023.
[30] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. IEEE signal processing magazine, 37(3):50–60, 2020.
[31] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017.

Introduction



Introduction

- We find that LLMs discern between good and bad prompts and utilize the implicit "gradient" in language for more efficient searches.

Contributions:

- We introduce a novel method for black-box prompt engineering of VLMs, utilizing an LLM as an optimizer.

- We extensively explore various strategies for conversing with ChatGPT, uncovering several key factors that significantly enhance the efficiency of this tool.

- Our natural language prompts are interpretable and transfer better across CLIP architectures than previous white-box methods.

Related Work

[61] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190, 2021.
[62] Adi Haviv, Jonathan Berant, and Amir Globerson. Bertese: Learning to speak to bert. arXiv preprint arXiv:2103.05327, 2021.
[63] Weijia Shi, Xiaochuang Han, Hila Gonen, Ari Holtzman, Yulia Tsvetkov, and Luke Zettlemoyer. Toward human readable prompt tuning: Kubrick's the shining is a good movie, and a good prompt too? arXiv preprint arXiv:2212.10539, 2022.
[64] Yuan Yao, Bowen Dong, Ao Zhang, Zhengyan Zhang, Ruobing Xie, Zhiyuan Liu, Leyu Lin, Maosong Sun, and Jianyong Wang. Prompt tuning for discriminative pre-trained language models. arXiv preprint arXiv:2205.11166, 2022.
[65] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. arXiv preprint arXiv:2110.07602, 2021.
[66] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. arXiv preprint arXiv:2010.15980, 2020.
[67] Archiki Prasad, Peter Hase, Xiang Zhou, and Mohit Bansal. Grips: Gradient-free, edit-based instruction search for prompting large language models. arXiv preprint arXiv:2203.07281, 2022.

Black-box optimization of foundation models

white-box methods:

- continuous prefix-tuning [61, 62, 63, 64, 65]
- discrete token-searching [66]

black-box methods:

- heuristic-based editing [67, 68]
- continuous prefix-tuning with genetic algorithms [69, 70, 71, 72]
- discrete token searching with reinforcement learning [73, 74]

[68] Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. Reframing instructional prompts to gptk's language. arXiv preprint arXiv:2109.07830, 2021.
[69] Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. Black-box tuning for language-model-as-a-service. In International Conference on Machine Learning, pages 20841–20855. PMLR, 2022.
[70] Hanwei Xu, Yujun Chen, Yulun Du, Nan Shao, Yanggang Wang, Haiyu Li, and Zhilin Yang. Gps: Genetic prompt search for efficient few-shot learning. arXiv preprint arXiv:2210.17041, 2022.
[71] Yekun Chai, Shuohuan Wang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Clip-tuning: Towards derivative-free prompt learning with a mixture of rewards. arXiv preprint arXiv:2210.12050, 2022.
[72] Tianxiang Sun, Zhengfu He, Hong Qian, Yunhua Zhou, Xuan-Jing Huang, and Xipeng Qiu. Bbtv2: Towards a gradient-free future with large language models. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 3916–3930, 2022.
[73] Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric P. Xing, and Zhiting Hu. Rlprompt: Optimizing discrete text prompts with reinforcement learning. arXiv preprint arXiv:2205.12548, 2022.
[74] Shizhe Diao, Xuechun Li, Yong Lin, Zhichao Huang, and Tong Zhang. Black-box prompt learning for pre-trained language models. arXiv preprint arXiv:2201.08531, 2022.

Related Work

LLMs for prompt optimization

DCLIP [5]

- uses GPT3 to produce rich visual descriptions to improve zero-shot classification with CLIP [4].

APE [36]

- uses an LLM to optimize prompts for another LLM through instruction induction [78] and iterative Monte Carlo search, which involves paraphrasing the current prompt.

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PMLR, 2021.
[5] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. arXiv preprint arXiv:2210.07183, 2022.
[36] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. arXiv preprint arXiv:2211.01910, 2022.
[78] Or Honovich, Uri Shaham, Samuel R Bowman, and Omer Levy. Instruction induction: From few examples to natural language task descriptions. arXiv preprint arXiv:2205.10782, 2022.

Related Work

Few-shot adaptation of VLMs

CoOp [3, 19, 24, 83, 84, 85]

- finetune an ensemble of continuous prefix tokens using cross-entropy loss.

[3] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. International Journal of Computer Vision, 130(9):2337–2348, 2022.
[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pages 740–755. Springer, 2014.
[24] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5206– 5215, 2022.
[83] Adrian Bulat and Georgios Tzimiropoulos. Language-aware soft prompting for vision & language foundation models. arXiv preprint arXiv:2210.01115, 2022.
[84] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. arXiv preprint arXiv:2209.07511, 2022.
[85] Mohammad Mahdi Derakhshani, Enrique Sanchez, Adrian Bulat, Victor Guilherme Turrisi da Costa, Cees GM Snoek, Georgios Tzimiropoulos, and Brais Martinez. Variational prompt tuning improves generalization of vision-language models. arXiv preprint arXiv:2210.02390, 2022.

Method

General prompt engineering framework

Algorithm 1 General prompt engineering framework. This algorithm shows how humans perform prompt engineering, which motivates our method of prompt engineering using chat-based LLMs.

Require: $D_{\text{train}} = \{x, y\}_n$: training samples, $F : D \times T \rightarrow \mathbb{R}$: black-box score function

Create an initial prompt set: $\mathcal{U} \leftarrow \{p_1\}$
Evaluate the initial prompt on training set: $S \leftarrow \{F(D_{\text{train}}, p_1)\}$
while not converged **do**
    Generate a new prompt p' based on S
    Evaluate the score of the new prompt on few-shot samples: $s' = F(D_{\text{train}}, p')$
    $\mathcal{U} \leftarrow \mathcal{U} \cup \{p'\}$
    $S \leftarrow S \cup \{s'\}$
end while
return prompt with highest score $p^* \leftarrow \arg\max_{p \in \mathcal{U}} F(D_{\text{train}}, p)$

Method

Leveraging LLMs as prompt engineers (prior art)

Algorithm 1 Automatic Prompt Engineer (APE)

Require: $\mathcal{D}_{\text{train}} \leftarrow \{(Q, A)\}_n$: training examples, $f : \rho \times \mathcal{D} \mapsto \mathbb{R}$: score function

Use LLM to sample instruction proposals $\mathcal{U} \leftarrow \{\rho_1, ..., \rho_m\}$. (See Section 3.1)
while not converged **do**
    Choose a random training subset $\widetilde{\mathcal{D}}_{\text{train}} \subset \mathcal{D}_{\text{train}}$.
    for all ρ in $\mathcal{U}$ **do**
        Evaluate score on the subset $\widetilde{s} \leftarrow f(\rho, \widetilde{\mathcal{D}}_{\text{train}})$ (See Section 3.2)
    end for
    Filter the top k% of instructions with high scores $\mathcal{U}_k \subset \mathcal{U}$ using $\{\widetilde{s}_1, ..., \widetilde{s}_m\}$
    Update instructions $\mathcal{U} \leftarrow \mathcal{U}_k$ or use LLM to resample $\mathcal{U} \leftarrow \text{resample}(\mathcal{U}_k)$ (See Section 3.3)
end while

Return instruction with the highest score $\rho^\star \leftarrow \arg\max_{\rho \in \mathcal{U}_k} f(\rho, \mathcal{D}_{\text{train}})$

[36] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. arXiv preprint arXiv:2211.01910, 2022.

Method

Conversational prompting with chat-based LLMs (our approach)

Require: $D_{\text{train}} = \{x, y\}_n$: training samples, $F : D \times T \rightarrow \mathbb{R}$: black-box score function.
Require: n_{restart}: number of initial sampled prompt sets, n_{reset}: number of resets for a prompt set, n_{iter}: number of hill-climbing iterations, m: size of one initial prompt set, k: number of prompt samples send to ChatGPT.

$p^* \leftarrow \emptyset$
for $1::n_{\text{restart}}$ **do**
  Sample a new prompt set from a text corpus, $\mathcal{U}_{\text{init}} \leftarrow \{p_1, ..., p_m\}$
  for $1::n_{\text{reset}}$ **do**
    Reset to initial prompt set: $\mathcal{U} \leftarrow \mathcal{U}_{\text{init}}$
    for $1::n_{\text{iter}}$ **do**
      Sort $\mathcal{U}$ based on their scores on training samples using $\{F(D_{\text{train}}, p)\}_{p \in U}$
      $\mathcal{U}_{\text{top}} \leftarrow$ top-k prompts in $\mathcal{U}$
      $\mathcal{U}_{\text{bot}} \leftarrow$ bottom-k prompts in $\mathcal{U}$
      Generate a new prompt based on top and bottom-k prompts $p_{\text{new}} \leftarrow \text{LLM}(\mathcal{U}_{\text{top}}, \mathcal{U}_{\text{bot}})$
      $\mathcal{U} \leftarrow \mathcal{U} \cup \{p_{\text{new}}\}$
    end for
    Update best prompt: $p^* \leftarrow \arg\max_{p \in \mathcal{U} \cup \{p^*\}} F(D_{\text{train}}, p)$
  end for
end for
return prompt with highest score p^*

Illustrative Task: Few-Shot Image Classification

Experimental setup

- We apply our method to the few-shot image classification benchmark from CoOp [3]
- Covered 11 datasets, including notable ones like ImageNet [13]
- Followed the three-fold k-shot training approach from [23]
- Utilize CLIP and ChatGPT(GPT3.5) for black-box VLM and conversational prompting

[3] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. International Journal of Computer Vision, 130(9):2337–2348, 2022.
[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009.
[23] Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramana. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. arXiv preprint arXiv:2301.06267, 2023.

Illustrative Task

[25] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114, 2021.
[89] Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. To appear, 7(1):411–420, 2017.
[3] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. International Journal of Computer Vision, 130(9):2337–2348, 2022.
[23] Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramana. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. arXiv preprint arXiv:2301.06267, 2023.

Implementation details

- Sampled 1M captions from LAION-COCO [25]
- Extracted noun phrases using spaCy part-of speech tagging [89]
- Replaced one noun phrase with “{}” to create template
- Initial prompt pool: ~2M templates (2 noun phrases on average per cation)
- Algorithm parameters: 20 restarts, 50 resets, 10 iterations
- Sampled 80 prompts per restart.
- Presented top and bottom 15 prompts to ChatGPT
- Used CLIP-RN50 for experiments as per prior work [3, 23]

Illustrative Task

[3] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. International Journal of Computer Vision, 130(9):2337–2348, 2022.
[5] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. arXiv preprint arXiv:2210.07183, 2022.
[23] Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramana. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. arXiv preprint arXiv:2301.06267, 2023.
[37] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. arXiv preprint arXiv:2109.01903, 2021. https://arxiv.org/ abs/2109.01903.

Previous white-box baselines:

CoOp [3], WiSE-FT [37], Cross-Modal Adaptation [23], DCLIP [5]

Other black-box baselines:

- Including the vanilla class-agnostic templates “{classname}” and “a photo of a {classname}”

- Best Hand-Engineered templates released by OpenAI, eg., “a centered satellite photo of {classname}.” for EuroSAT [17]

- Iterative APE [36], we use 30 positive prompts for our APE implementation.

[17] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification, 2017.
[36] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. arXiv preprint arXiv:2211.01910, 2022.

Illustrative Task

Method	BB	Dataset											Avg
		Caltech	ImageNet	Aircraft	Food	Pets	Cars	SUN	UCF	DTD	EuroSAT	Flowers	
Cross-Modal [23]	✗	89.1	61.6	20.6	77.1	85.7	59.0	63.4	64.7	49.9	61.8	76.3	64.7
Wise-FT [37]	✗	85.5	58.3	18.6	71.9	81.7	55.7	56.6	59.4	44.2	52.3	65.8	59.1
CoOp [3]	✗	87.5	57.2	9.6	74.3	85.9	55.6	60.3	61.9	44.4	50.6	68.1	59.6
DCLIP [5]	✗	-	59.6	-	76.4	83.8	-	-	-	41.7	34.7	-	-
{}	✓	78.5	55.3	15.5	74.0	78.9	52.2	53.4	55.5	41.4	32.1	57.3	54.0
a photo of a {}	✓	84.5	57.9	15.9	74.0	83.2	53.9	58.0	56.9	38.8	28.6	60.2	55.6
Hand-Engineered [4]	✓	86.3	58.2	17.3	77.3	85.8	55.6	58.5	**61.5**	42.3	37.6	66.1	58.8
LAIONCOCO-1M	✓	81.4	56.2	17.4	76.5	79.6	51.3	54.9	55.8	43.1	38.6	61.3	56.0
APE	✓	89.0	59.4	17.9	77.8	85.7	55.7	60.4	58.7	43.6	46.7	66.6	60.1
Our Method	✓	**89.1**	**59.6**	**18.1**	**78.3**	**88.1**	**56.2**	**61.0**	60.2	**44.8**	**49.0**	**67.2**	**61.1**

Illustrative Task

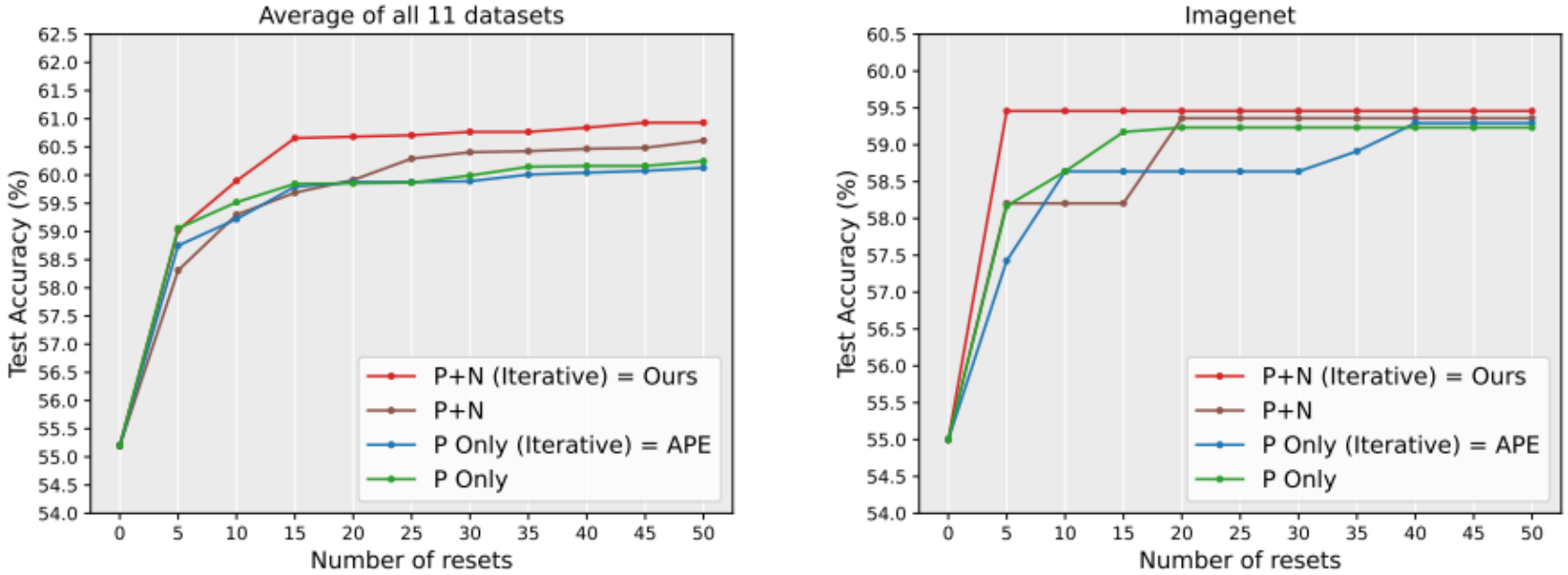
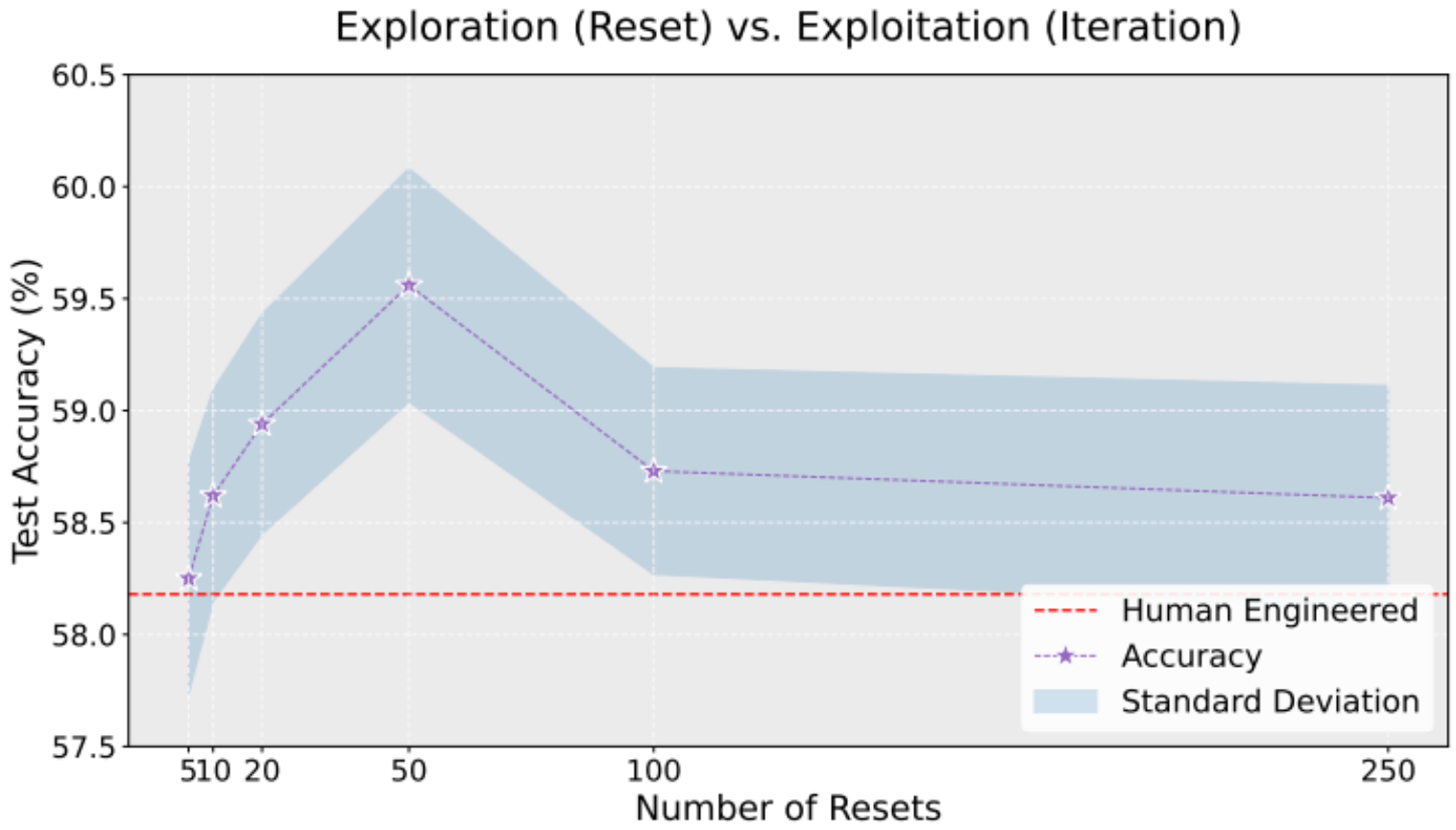


Figure 2: **Conversational feedback incorporating both positive and negative prompts leads to improved efficiency.** We ablate different configurations of using ChatGPT as black-box optimizer using conversational feedback. In particular, "P+N" denotes the use of both top-k and bottom-k prompts when conversing with ChatGPT, and "Iterative" denotes replacing the positive and negative prompts with the most recent candidates generated by ChatGPT at every iteration. For this ablation, we fix the number of restarts to 20 and iterations to 10, and ablate the test accuracy of different numbers of resets on all 11 datasets (left) and ImageNet (right). We observe that our approach (P+N Iterative) can optimize faster within a much fewer number of resets, resulting in the highest performance.

Additional Analysis

Balancing exploration and exploitation can improve the final performance

- fixed budget: 500 API calls per restart



Additional Analysis

Dataset	Example of Top Templates
Caltech [14]	`An image of a {} with a blurred background that emphasizes the subject`
DTD [90]	`The essential elements of {} are amplified with visual simplicity`
EuroSAT [17]	`A top-down view of {} arranged in a pattern {}`
Aircraft [88]	`A clear, high-quality image of a single {} with a white background`
Food [15]	`A {} featuring diverse cuisine and ingredients`
ImageNet [13]	`An image of a {} with bright and natural lighting`
Flowers [16]	`A clear and vivid photograph of the {} in its natural setting`
Pets [91]	`A {} with distinct and recognizable characteristics`
Cars [18]	`A {} featuring a wide range of color options for easy selection`
SUN [92]	`A high-resolution photo of a {} with clear background and natural lighting`
UCF [93]	`A black and white photo of a {} in motion`

Table 2: **Example templates returned by our algorithm on each dataset.** Although we do not provide ChatGPT with any information regarding the targeted dataset, we observe that the resulting templates are remarkably similar to human-engineered templates, with many domain-specific details such as “motion” and “cuisine”, and stylistic elements such as “bright and natural lighting”.

Additional Analysis

Method	RN50	→RN101	→ViT-B/32	→ViT-B/16
`a photo of a {}`	57.9	60.6	61.9	66.6
CoOp	**63.0**	20.6	31.7	39.5
Ours	59.9	**60.7**	**62.2**	**67.0**

Table 3: Black-box transferring prompts from ResNet-50 to other CLIP architectures on 16-shot ImageNet

Conclusion

- We introduce a method using LLMs to engineer prompts for black-box VLMs.
- Our approach integrates a conversational feedback loop with chat-based LLMs.
- In one-shot image classification, we outperform many black-box techniques and compete with white-box methods.
- Our method produces prompts that are more universally applicable across different black-box VLMs.