Towards Efficient Visual-Language Alignment of the Q-Former for Visual Reasoning Tasks

Sungkyung Kim[1*] **Adam Lee**[2*] **Junyoung Park**[3]
Andrew Chung[1,3] **Jusang Oh**[1] **Jay-Yoon Lee**[4†]
[1]Seoul National University [2]UC Berkeley [3]Weavel, Inc.
[4]Graduate School of Data Science, Seoul National University
sk0428@snu.ac.kr, alee00@berkeley.edu
{junyoung, sounho}@weavel.ai, {dhwntkd412, lee.jayyoon}@snu.ac.kr

Abstract

Recent advancements in large language models have demonstrated enhanced capabilities in visual reasoning tasks by employing additional encoders for aligning different modalities. While the Q-Former has been widely used as a general encoder for aligning several modalities including image, video, audio, and 3D with large language models, previous works on its efficient training and the analysis of its individual components have been limited. In this work, we investigate the effectiveness of parameter efficient fine-tuning (PEFT) the Q-Former using InstructBLIP with visual reasoning benchmarks ScienceQA and IconQA. We observe that applying PEFT to the Q-Former achieves comparable performance to full fine-tuning using under 2% of the trainable parameters. Additionally, we employ AdaLoRA for dynamic parameter budget reallocation to examine the relative importance of the Q-Former's sublayers with 4 different benchmarks. Our findings reveal that the self-attention layers are noticeably more important in perceptual visual-language reasoning tasks, and relative importance of FFN layers depends on the complexity of visual-language patterns involved in tasks. The code is available at https://github.com/AttentionX/InstructBLIP_PEFT.

1 Introduction

Pre-trained large language models (LLMs) can be fine-tuned with instruction tuning to align the model responses with human intentions (Taori et al., 2023; Wang et al., 2023). Recently, model alignment with instruction tuning has been extended to the image domain by using an external encoder to align visual-language modalities and enhance the model's capabilities for visual reasoning. LLaVA (Liu et al., 2023b,a) uses a projection layer to project CLIP (Radford et al., 2021) image embeddings to the text embedding space of language models. However using a projection layer to convert every CLIP embedding from an image can take up a lot of context tokens and increase inference time. BLIP-2 (Li et al., 2023b) and InstructBLIP (Dai et al., 2023) use a Q-Former for visual-language alignment that transfers visual features into a fixed number of learnable embeddings (32 in BLIP-2), which is similar to Perceiver IO (Jaegle et al., 2022) and Flamingo (Alayrac et al., 2022).

The Q-Former architecture is especially significant for its generalizability in aligning several modalities. This architecture of using cross-attention to transfer features to a small number of learnable embeddings has been used in recent studies for aligning many different modalities including image (Bai et al., 2023; Dai et al., 2023) video (Zhang et al., 2023a), and 3D (Hong et al., 2023).

However, despite the increased usage and significance of the Q-Former, prior research on its sublayers and their importance in different tasks has been limited. Elucidating the importance of each sublayer for different visual reasoning tasks can assist in designing more efficient training methods with effective parameter allocation. Moreover, although PEFT methods have been successfully applied to efficiently train language models (Hu et al., 2021; He et al., 2021; Houlsby et al., 2019; Lester et al., 2021; Li and Liang, 2021) evaluating the effectiveness of PEFT on the Q-Former and visual language models also remains under-explored. These two areas are critical for advancing the efficiency of training multimodal language models.

In this work, we evaluate the performance of PEFT on InstructBLIP with two benchmarks, ScienceQA (Lu et al., 2022) and IconQA (Lu et al., 2021), that respectively evaluate knowledge-grounded visual reasoning and perceptual visual reasoning. We apply LoRA to the Q-Former and

[*]Equal contribution.
[†]Corresponding author.

Findings of the Association for Computational Linguistics: EACL 2024, pages 15155–15165

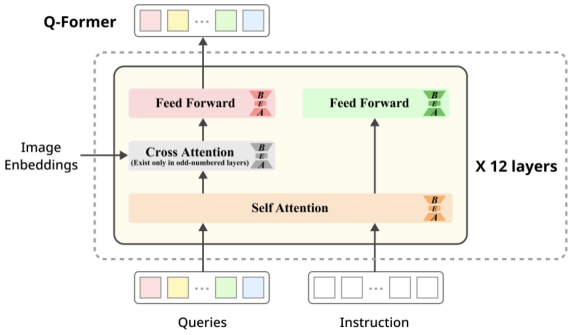


Figure 1: The detailed structure of the Q-Former with AdaLoRA weight matrices (B, E, A).

base LLMs, Flan-T5-XL (Chung et al., 2022) and Vicuna-7B (Chiang et al., 2023), and comprehensively test the performance of LoRA applied to different sublayers in the Q-Former with different ranks. We also examine the importance of each sublayer in the Q-Former for 4 different visual reasoning benchmarks using AdaLoRA (Zhang et al., 2023b), which dynamically allocates parameter budgets to improve performance. To the best of our knowledge, we are the first to inspect the effectiveness of PEFT methods on the Q-Former and analyze its sublayers for visual reasoning tasks.

Our contributions can be summarized as follows: (1) We demonstrate that applying PEFT to the Q-Former can reduce the trainable parameters to less than 2% while maintaining comparable performances. (2) We show that in contrast to full fine-tuning the Q-Former and freezing the LLM, applying PEFT to both components can achieve superior results and reduces the total trainable parameters to less than 12%. (3) We examine the significance of the different sublayers in the Q-Former using AdaLoRA, and find that the self-attention layers are relatively more important for tasks that require stronger visual-language alignment, and more intrinsic ranks on FFN layers are needed to train on complex visual-language patterns.

2 Method

In this work, we apply the PEFT method LoRA to the Q-Former and the LLM in InstructBLIP, and evaluate the performance on two visual reasoning benchmarks ScienceQA and IconQA. Additionally, we apply AdaLoRA to analyze the significance of each sublayer of the Q-Former on visual reasoning.

LoRA reduces trainable parameters by decomposing the weight update matrix $\Delta W = BA$. After fine-tuning, the weight matrix can be reparametrized by adding the weight update to the original pre-trained model weights: $W + \Delta W = W + BA$, where $W \in \mathbb{R}^{d\times k}$, $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, $r \ll \min(d, k)$. Unlike the original LoRA implementation, which confines its application to only the self-attention layers (Hu et al., 2021), we extend the use of LoRA to multiple transformer sublayers in both the Q-Former and the LLM. Specifically, we apply LoRA to the q, v layers in self-attention layers, the q, k, v, o layers in cross-attention layers, and the FFN layers in the Q-Former.

AdaLoRA decomposes the intrinsic weight update matrix $\Delta W = BEA$ with singular value decomposition (SVD). During training, the less significant singular values are adaptively pruned based on their importance scores, adjusting the rank of the weight update matrices. The importance score of the ith singular value is calculated as follows:

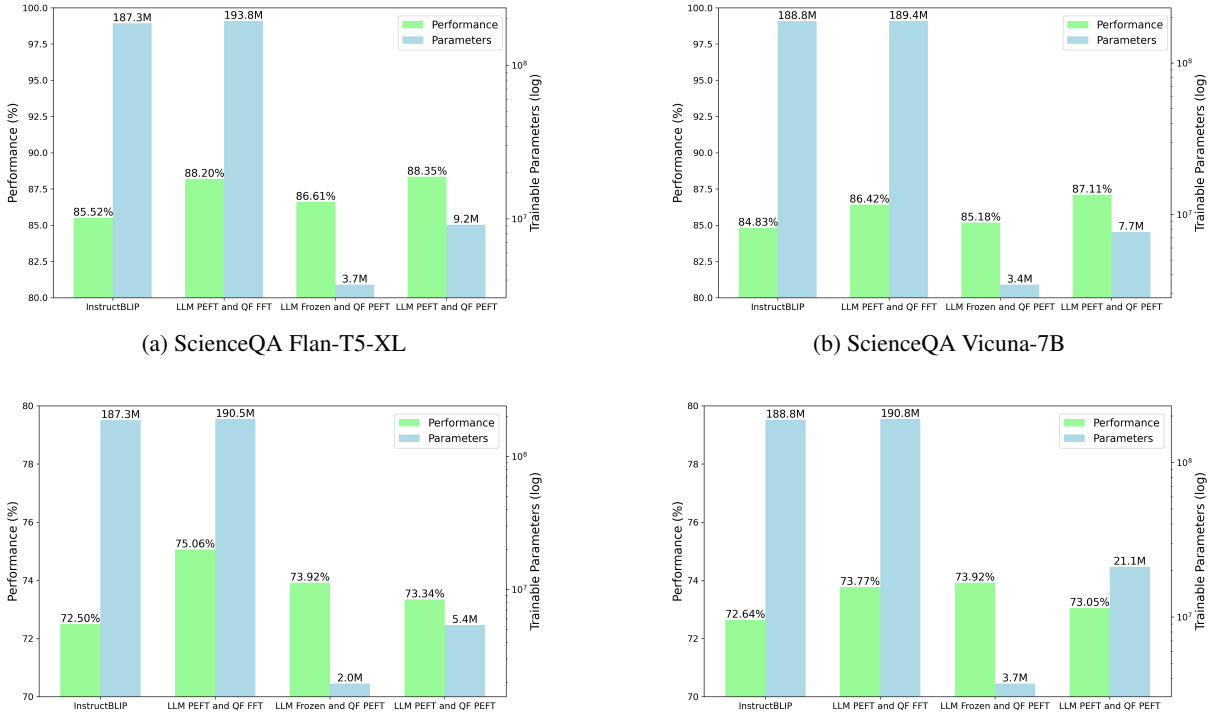
$$S_i = s(\lambda_i) + \frac{1}{d_1}\sum_{j=1}^{d_1} s(B_{ji}) + \frac{1}{d_2}\sum_{j=1}^{d_2} s(A_{ji}) \quad (1)$$

where $B \in \mathbb{R}^{d_1\times r}$, $A \in \mathbb{R}^{r\times d_2}$ and $s(\cdot)$ is a specific importance function for each entry, based on sensitivity of each weight to the training loss. As a result, applying AdaLoRA leads to appropriate rank allocation across modules for better performance. Since high-rank updates learn more complex signals, we use AdaLoRA for examining which sublayers in the Q-Former are critical for each visual reasoning task and which sublayers should be prioritized in parameter budget allocation for efficient fine-tuning. We apply AdaLoRA to the self-attention(q, v), cross-attention(q, k, v, o) layers, and FFN layers altogether for overall comparison. (Figure 1)

Base Models and Benchmarks. We employ InstructBLIP as the base model for its pioneering use of the Q-Former and its strong performance on several downstream tasks (Dai et al., 2023) including ScienceQA (IMG) (Lu et al., 2022), OCR-VQA (Mishra et al., 2019), and A-OKVQA (Schwenk et al., 2022). We use the InstructBLIP implementation of LAVIS (Li et al., 2023a) and use pre-trained Flan-T5-XL[1] and Vicuna-7B[2] HuggingFace checkpoints in our experiments.

We use two benchmarks covering Knowledge Grounded Visual Reasoning (ScienceQA) and Perceptual Visual Reasoning (IconQA) (Lu et al.,

[1] https://huggingface.co/google/flan-t5-xl
[2] https://huggingface.co/lmsys/vicuna-7b-v1.3



(a) ScienceQA Flan-T5-XL

(b) ScienceQA Vicuna-7B

(c) IconQA Flan-T5-XL

(d) IconQA Vicuna-7B

Figure 2: Comparing the performance and number of trainable parameters using Flan-T5-XL and Vicuna-7B as base models on ScienceQA and IconQA benchmarks. This compares the best performing configurations (rank value and LoRA-applied sublayers) of Q-Former full fine-tuning with LLM PEFT, Q-Former PEFT with frozen LLM, and Q-Former PEFT with LLM PEFT, against InstructBLIP (Q-Former full fine-tuning with frozen LLM). "QF" denotes Q-Former. "FFT" denotes full fine-tuning. The complete results and the training architectures are at Appendix A.

2021) tasks. These benchmarks were held-out datasets for InstructBLIP, and were not involved in training the baseline InstructBLIP model. For analyzing the Q-Former with AdaLoRA we use two additional benchmarks, Vizwiz (Gurari et al., 2018) and Flickr30k (Plummer et al., 2016).

Knowledge Grounded Visual Reasoning is a task of answering questions with a provided image related to the knowledge in diverse academic areas including physics, biology, and math. We use the ScienceQA dataset which covers a variety of science topics with corresponding extensive explanations. We only use the questions with image context (IMG). ScienceQA (IMG) has 6,218/2,097/2,017 samples for train/validation/test set.

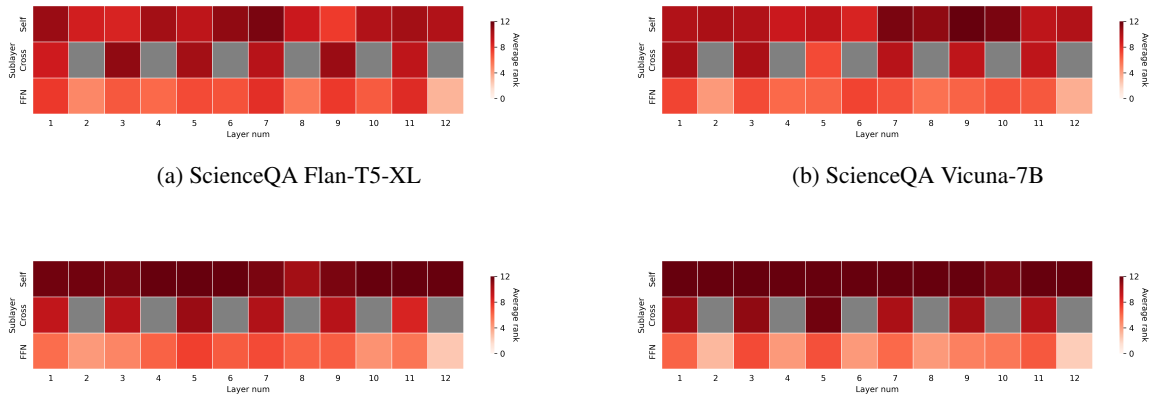
Perceptual Visual Reasoning is a task of answering questions after comprehending the abstract meanings from an image. We use IconQA (Multi-text-choice) which contains question-answer pairs for natural images that require comprehensive reasoning abilities to understand abstract diagrams. IconQA (Multi-text-choice) has 18,946/6,316/6,316 samples for train/validation/test set.

3 Experiments

3.1 PEFT Effectiveness for Visual Reasoning

We empirically analyze the effectiveness of training the Q-Former and the LLM in InstructBLIP with LoRA. (1) First, we apply LoRA to the LLM while still full fine-tuning the Q-Former. (2) Second, we apply LoRA to the Q-Former while freezing the LLM, resulting in efficient fine-tuning of the Q-Former. (3) Finally, we apply LoRA to both the Q-Former and the LLM. The performance comparison between (1), (2), (3) and the original InstructBLIP is in Figure 2. (The main results of the overall experiments are in Appendix A, and the implementation and training details can be found in Appendix B.)

We find that applying LoRA to the Q-Former yields competitive performance, matching or surpassing full-fine-tuning while using less than 2% of the original trainable parameters. Fine-tuning the base LLMs with LoRA consistently outperforms the baseline InstructBLIP model on both bench-



(a) ScienceQA Flan-T5-XL

(b) ScienceQA Vicuna-7B

(c) IconQA Flan-T5-XL

(d) IconQA Vicuna-7B

Figure 3: Heatmaps of the rank distributions of the sublayers in the Q-Former. Cross-attention layers are present in odd numbered layers only. Each value is the average of the component layers. The detailed heatmaps including additional benchmarks (Flickr30k, Vizwiz) are in Appendix E.

marks, underscoring the enhanced task-specific language capabilities by training the language model. Applying LoRA to both the Q-Former and LLM achieve superior performance on both benchmarks with fewer than 12% of the trainable parameters. Notably, we find that fine-tuning both models perform consistently higher in ScienceQA than in IconQA. This discrepancy can be attributed to ScienceQA's richer language context. Given that ScienceQA entails more language information than IconQA, training the language model appears to yield a greater boost in performance.

3.2 Analysis of Q-Former Sublayers using AdaLoRA

To investigate the significance of each sublayer in the Q-Former with 4 benchmarks (ScienceQA, IconQA, Flickr30k, Vizwiz), we use AdaLoRA to analyze the dynamically reallocated intrinsic ranks of the weight update matrices. We apply iteration-based AdaLoRA, with an initial rank of 12 and target rank of 8, to each training epoch. The final rank of each sublayer after training indicates their respective importance and the prioritization of the parameter budget. To visually represent these dynamics, we compute heatmaps (Figure 3) that average AdaLoRA ranks within the same sublayers (self-attention, cross-attention, and FFN) across each of the Q-Former's 12 layers. These heatmaps illustrate the rank distribution for each training configuration across layers and sublayers.

For IconQA, rank allocation is predominantly focused on the self-attention layers for both base LLMs, with the cross-attention layers having the second highest number of ranks. Notably FFN layers tend to have higher ranks in odd-numbered layers, where the cross-attention layers are present. For ScienceQA, the FFN layers are similarly allocated the fewest average ranks. But compared to IconQA, the average rank distribution between the three sublayers are more balanced, and self-attention layers have noticeably fewer ranks. For all configurations, FFNs in the the final layer (12) consistently have the fewest ranks.

The distribution of rank allocation can be attributed to the different types of reasoning abilities required for each task and the relationship between sublayers. IconQA is consisted of perceptual visual reasoning questions that require strong visual-language alignment. Meanwhile, ScienceQA contains questions grounded in extensive knowledge, which demands significant logical reasoning on longer texts in addition to visual-language alignment.

Given that FFN layers are adept at learning task-specific patterns (He et al., 2022), and considering the complex textual patterns inherent in ScienceQA questions, we hypothesize that these factors contributed to the observed increase in ranks and influence of FFN layers for ScienceQA relative to IconQA. Also, FFN layers in odd-numbered layers tend to have higher ranks on both tasks, as they come after the cross-attention layers, receiving image features and making them more important for learning task-specific visual patterns.

The rank allocation of self-attention layers can be attributed to the relative importance of visual-

language alignment for the task. Self-attention layers allow query embeddings to attend textual information and extract visual features that are more relevant to the text prompt. This explains the significant concentration of ranks in the self-attention layers consistently throughout all 12 layers and base language models in IconQA.

We use 2 additional benchmarks, Flickr30k and Vizwiz, for analyzing the Q-Former with AdaLoRA. Flickr30k is an image captioning task, and Vizwiz covers visual question answering. The detailed results for each benchmark is shown in Appendix E. The rank distribution is concentrated in the self-attention layers for both benchmarks, while the overall rank distribution is more even in Vizwiz than in Flickr30k. This result can be explained by the difference in complexity of the text prompts between the two benchmarks. Flickr30k's text instruction is fixed to image captioning, while Vizwiz's text instruction covers more diverse questions. Therefore, it can be explained that the resulting heatmap of Vizwiz aligns more with ScienceQA, and the result of Flickr30k aligns more with IconQA. This also indicates that result of AdaLoRA analysis in ScienceQA and IconQA generalizes well to other benchmarks.

4 Conclusion

In this work, we systematically evaluate the effectiveness of applying PEFT to the Q-Former and visual language models. Our results show that applying PEFT to the Q-Former achieves comparable performance to full fine-tuning while only utilizing less than 2% of the trainable parameters. Additionally, we employ dynamic parameter budget allocation with AdaLoRA to analyze the significance of the Q-Former's sublayers for different visual reasoning tasks. Our findings reveal that the importance of FFN layers increases when visual-language pattern becomes more complex, and importance of self-attention layers increases as significance of visual-language alignment in task increases.

Limitations

More recently, the Q-Former architecture has been used to align many different modalities beyond images and languages including 3D, depth, audio, and video. In this work, we focus on efficiently training and analyzing the Q-Former for image-text alignment and leave the study of other modalities to future works.

Ethics Statement

The datasets used in this work is publicly released by CC BY-NC-SA license, so there is no copyright issue in this paper.

Acknowledgements

This work was supported by National Research Foundation of Korea (NRF) grant (RS-2023-00280883, RS-2023-00222663), New Faculty Startup Fund from Seoul National University, National Super computing Center (KSC-2023-CRE-0176), Artificial Intelligence Industry Center Agency, Google cloud platform research credits, Seoul National University's Engineering Department, and the AI Research Group AttentionX.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P.

Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2021. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Towards a unified view of parameter-efficient transfer learning.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 2023. 3d-llm: Injecting the 3d world into large language models.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. 2022. Perceiver io: A general architecture for structured inputs outputs.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C.H. Hoi. 2023a. LAVIS: A one-stop library for language-vision intelligence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 31–41, Toronto, Canada. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 947–952.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2016. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions.

Hang Zhang, Xin Li, and Lidong Bing. 2023a. Video-llama: An instruction-tuned audio-visual language model for video understanding.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023b. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning.

A InstructBLIP PEFT Experiments

The diagram of training configurations is shown in Figure 6. The full experimental results of applying PEFT to InstructBLIP, are shown in Table 1. All the results presented in this paper are obtained after single-run experiments.

B Model Training Details for PEFT Evaluations

We conduct each experiment in Table 1 and Figure 2 using a single A100 GPU. We set the maximum epoch to 15 with early stopping of 3 patience steps. We use linear decay as a learning rate scheduler with the AdamW optimizer. For the initial learning rate, we primarily use 2e-5 for experiments which involves full fine-tuning the Q-Former, and otherwise 5e-4. For certain cases that deviate significantly from other experiments, we lower the learning rate from 2e-5 to 1e-5 and 5e-4 to 1e-4. These cases include: (1) When the model is trained on less than 8 epochs (the halfway point) by early stopping, (2) When the training is considered unstable, i.e. resulting in over 10%p lower performance than other experiment in an equivalent setup having different r value. We set the weight decay to 0.05. For batch size, we use 16 as an effective batch size across all experiments. Only difference is that (batch size, gradient accumulation iterations) were set to (8, 2) for Vicuna-7B and (16, 1) for Flan-T5-XL.

C Model Training Details for AdaLoRA Experiments

For ScienceQA and IconQA, epoch settings, effective batch sizes (16), learning rate and scheduling methods, and weight decay values are given the same as Appendix B. For Flickr30k and Vizwiz, we set the maximum epoch to 5 with early stopping of 3 patience steps. We use "linear_warmup_cosine_lr" scheduler, and set an initial learning rate of 1e-4 with batch size 8 on Vizwiz, and set an initial learning rate of 5e-5 with batch size 60 on Flickr30k.

D Instruction Templates

We provide instructions used in ScienceQA and IconQA. We use the same format from the InstructBLIP paper. We add alphabet labels for each choices and the answer. For ScienceQA, we construct the "context" section of the instruction by incorporating information from both the 'hint' and 'lecture' fields, if they are available in the dataset.

ScienceQA Context: { {hint} {lecture} } Question: { {question} } Options: { {choices} }. Answer:

IconQA <Image> Question: { {question} } Options: { {choices} }. Short answer:

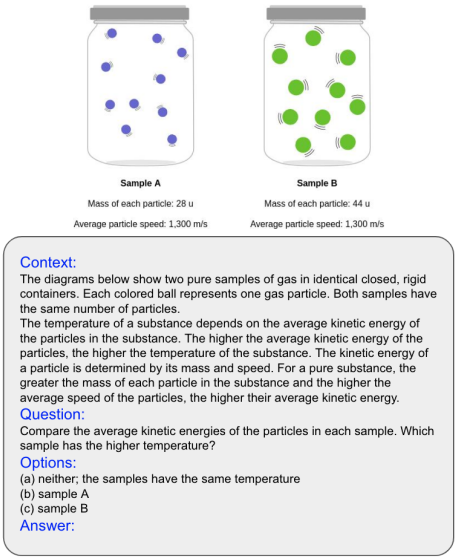


Figure 4: Example ScienceQA[3] instruction template

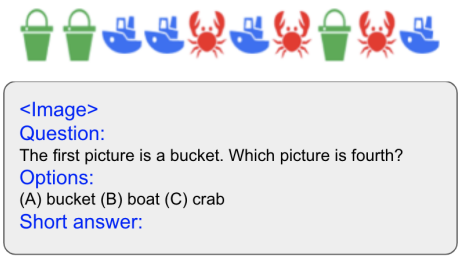


Figure 5: Example IconQA[3] instruction template

E Detailed Figures for AdaLoRA Experiments

The detailed figures of rank distribution in AdaLoRA experiments are shown in Figure 7 (Flan-T5-XL) and Figure 8 (Vicuna-7B). Also, heatmaps for Flickr30k, Vizwiz with Flan-T5-XL are shown in Figure 9 and Figure 10.

[3] https://creativecommons.org/licenses/by-nc-sa/4.0/

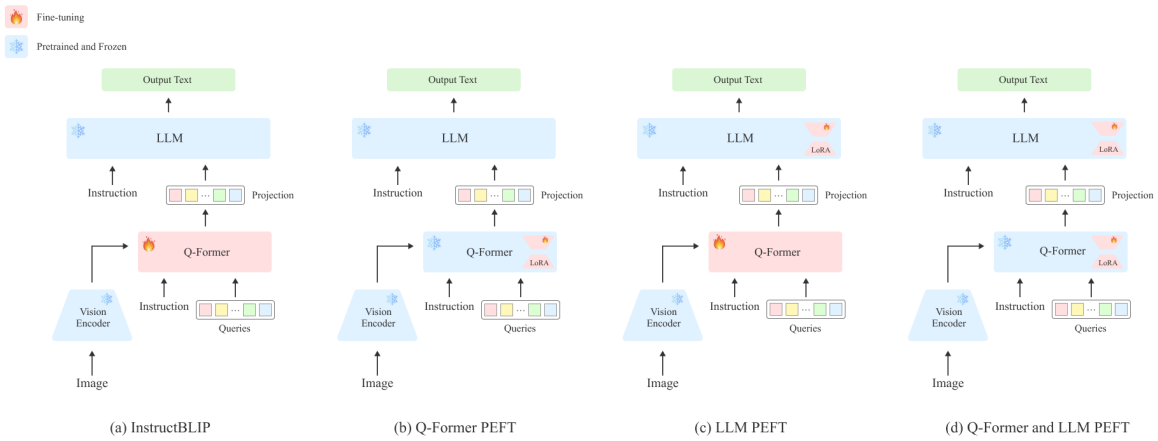
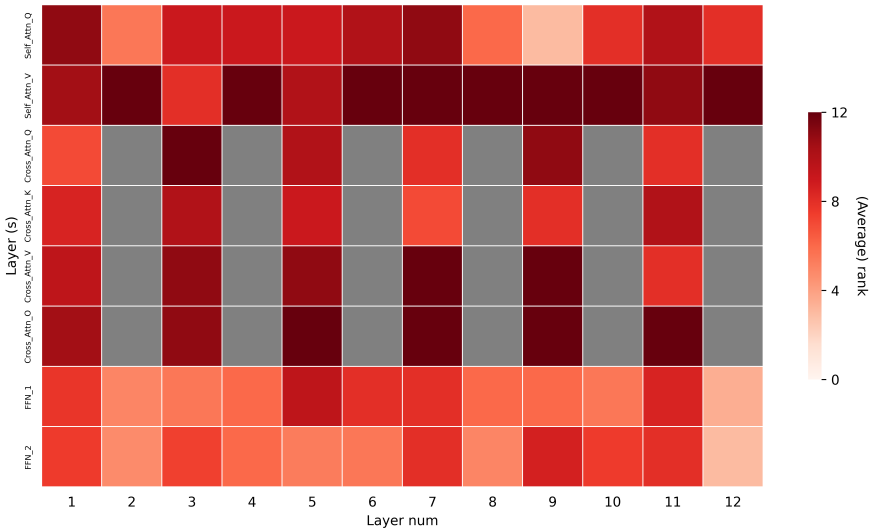


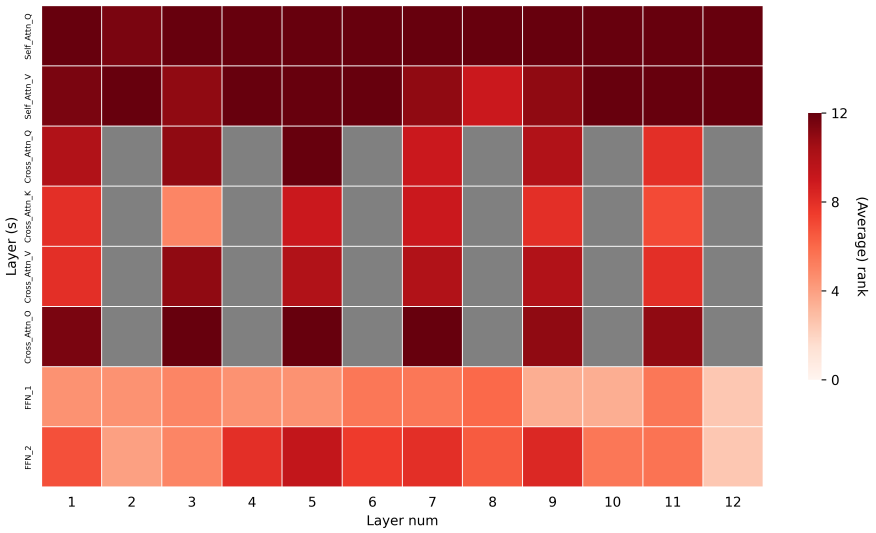
Figure 6: Applying PEFT to the Q-Former and LLM in InstructBLIP.

Method				ScienceQA				IconQA			
LLM	Q-Former	Sublayer	Base Model	r=1	r=2	r=4	r=8	r=1	r=2	r=4	r=8
LoRA	Full	ffn	Flan-T5-XL	**86.42**	86.37	85.32	86.27	73.40	**74.76**	74.00	71.52
LoRA	Full	attn	Flan-T5-XL	**87.36**	86.17	86.91	86.42	72.34	72.88	**73.45**	73.29
LoRA	Full	all	Flan-T5-XL	87.41	87.36	**88.20**	87.90	72.61	**75.06**	73.23	72.74
Freeze	LoRA	ffn	Flan-T5-XL	84.83	83.79	83.14	**85.87**	70.54	72.13	68.08	**72.40**
Freeze	LoRA	self-attn	Flan-T5-XL	**86.02**	83.74	79.57	86.02	71.82	**72.55**	72.06	71.64
Freeze	LoRA	cross-attn	Flan-T5-XL	84.13	**86.32**	84.88	85.18	72.32	72.42	72.32	**73.92**
Freeze	LoRA	all	Flan-T5-XL	85.37	86.42	83.89	**86.61**	70.19	70.50	72.82	**73.31**
LoRA	LoRA	all	Flan-T5-XL	88.00	88.10	**88.35**	88.05	71.47	**73.34**	71.41	73.18
LoRA	Full	ffn	Vicuna-7B	86.32	**86.42**	85.87	85.97	71.39	72.97	**73.02**	72.34
LoRA	Full	attn	Vicuna-7B	**86.42**	86.32	85.08	85.23	72.36	**73.16**	72.29	73.02
LoRA	Full	all	Vicuna-7B	85.03	**86.32**	85.57	85.72	**73.77**	71.71	72.93	73.15
Freeze	LoRA	ffn	Vicuna-7B	83.44	**83.74**	83.64	83.74	69.89	**72.50**	72.50	71.11
Freeze	LoRA	self-attn	Vicuna-7B	**83.19**	81.51	82.25	83.14	71.23	71.45	71.42	**71.74**
Freeze	LoRA	cross-attn	Vicuna-7B	**83.29**	83.24	83.14	82.75	71.11	72.40	71.99	**73.39**
Freeze	LoRA	all	Vicuna-7B	**85.18**	82.80	83.74	83.44	71.49	**73.92**	71.45	73.40
LoRA	LoRA	all	Vicuna-7B	85.87	**87.11**	85.08	85.62	71.72	72.01	72.61	**73.05**

Table 1: Overall performance results. "Full" indicates full fine-tuning, and the best results among 4 r values are bolded. The best results for each PEFT category, benchmark, and base language models are underlined. The underlined performances are used to compare the best performances between PEFT methods in Figure 2.

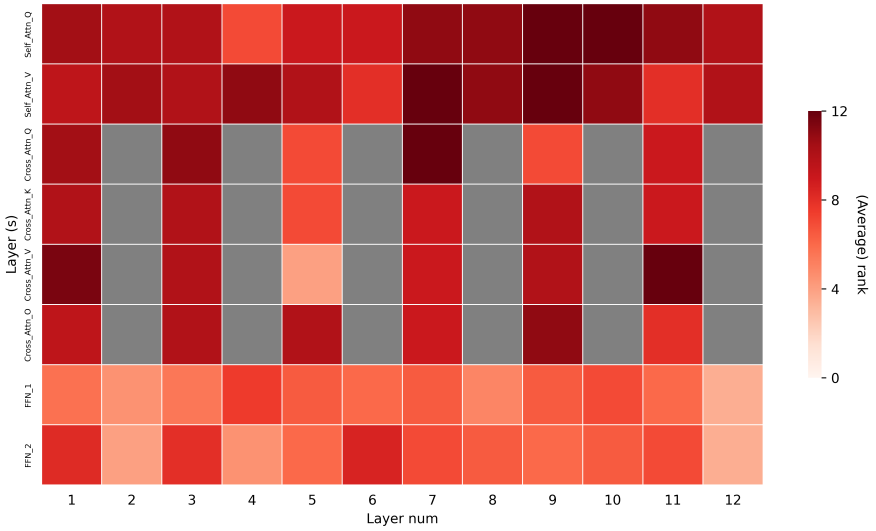


(a) ScienceQA Flan-T5-XL

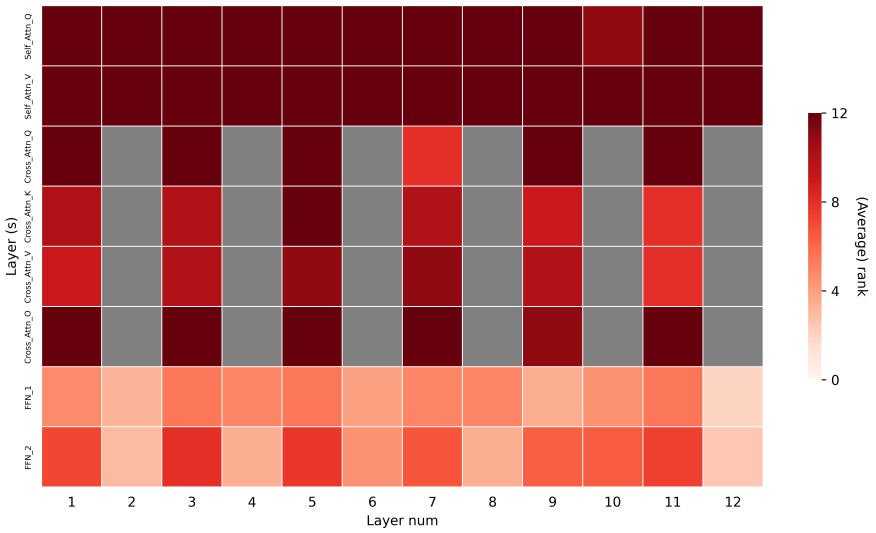


(b) IconQA Flan-T5-XL

Figure 7: Detailed heatmaps of rank distribution of modules in layers of the Q-Former. (Flan-T5-XL as a base LLM) Cross-attention layers are present in odd numbered layers only. The rank values in the feed-forward network (FFN) components are averaged across both FFN layers.

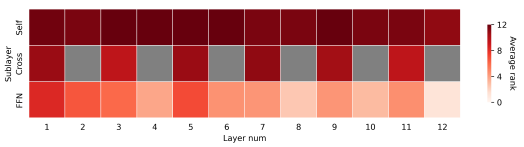


(a) ScienceQA Vicuna-7B

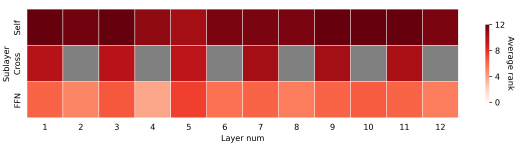


(b) IconQA Vicuna-7B

Figure 8: Detailed heatmap of rank distribution of modules in layers of the Q-Former. (Vicuna-7B as a base LLM) Cross-attention layers are present in odd numbered layers only. The rank values in the feed-forward neural network (FFN) components are averaged across both FFN layers.

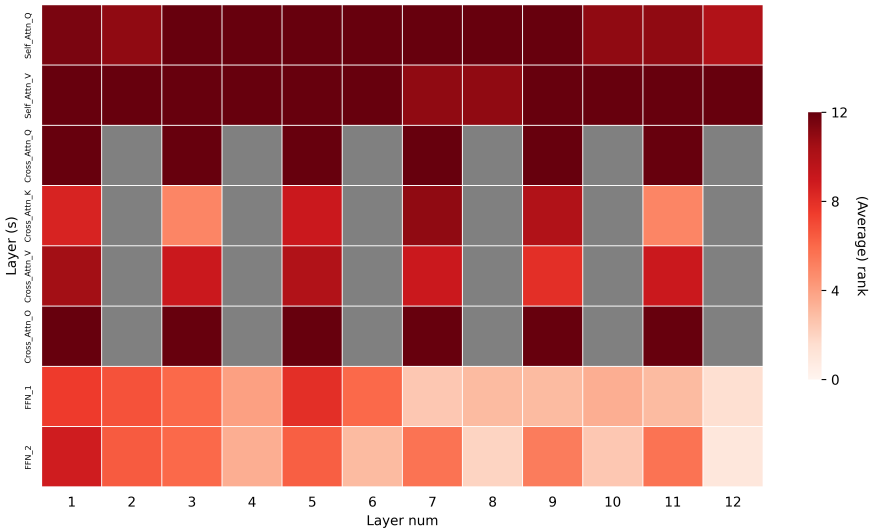


(a) Flickr30k Flan-T5-XL

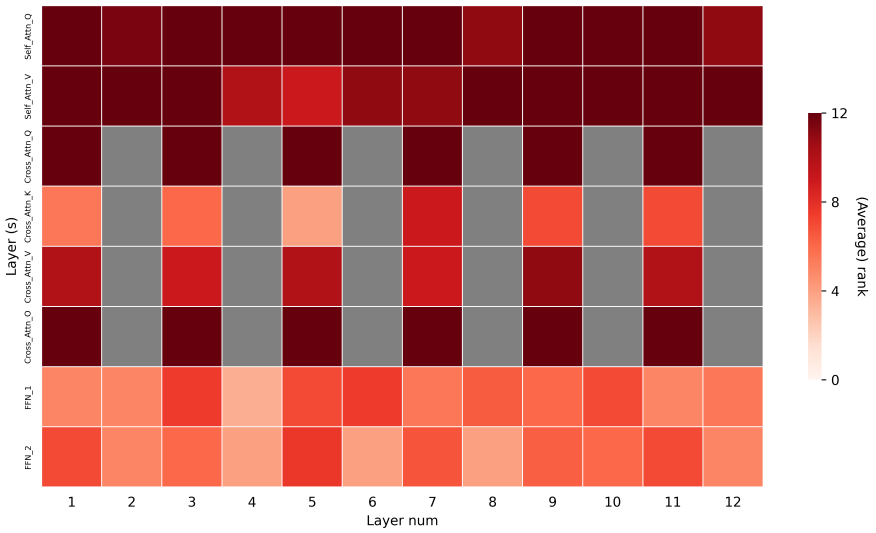


(b) Vizwiz Flan-T5-XL

Figure 9: Heatmaps of the rank distributions of the sublayers in the Q-Former for Flickr30k and Vizwiz. (Flan-T5-XL as a base LLM) Cross-attention layers are present in odd numbered layers only. Each value is the average of the component layers.



(a) Flickr30k Flan-T5-XL



(b) Vizwiz Flan-T5-XL

Figure 10: Detailed heatmaps of rank distribution of modules in layers of the Q-Former for Flickr30k and Vizwiz. (Flan-T5-XL as a base LLM) Cross-attention layers are present in odd numbered layers only. The rank values in the feed-forward network (FFN) components are averaged across both FFN layers.