# From Holistic to Localized: Local Enhanced Adapters for Efficient Visual Instruction Fine-Tuning – Appendix

Pengkun Jiao[1], Bin Zhu[3], Jingjing Chen[2]*, Chong-Wah Ngo[3], and Yu-Gang Jiang[2]
[1]College of Computer Science and Artificial Intelligence, Fudan University, China
[2]Institute of Trustworthy Embodied AI, Fudan University, China
[3]Singapore Management University
pkjiao23@m.fudan.edu.cn
{binzhu,cwngo}@smu.edu.sg, {jingjingchen,ygj}@fudan.edu.cn

## A. Proof of Proposition 1, Corollary 1, and Corollary 2

### Proof of Proposition 1

*Statement:*

For a set of LoRAs $\{B_k A_k\}_{k=1}^{K}$, each with rank 1, where $K$ is the total number of LoRAs, the representation space of a single LoRA with rank $K$ contains, and is at least as expressive as, the union of the representation spaces of the individual rank-1 LoRAs.

*Proof:*

Each rank-1 LoRA $B_k A_k$ can be expressed as the outer product of two vectors (since rank 1 implies that both $A_k$ and $B_k$ are vectors). The combined effect of these LoRAs is given by the sum:

$$M = \sum_{k=1}^{K} B_k A_k.$$

The rank of $M$ satisfies:

$$\text{rank}(M) \leq \sum_{k=1}^{K} \text{rank}(B_k A_k) = K,$$

since the sum of $K$ rank-1 matrices has rank at most $K$.

Now, consider a single LoRA $BA$ with $\text{rank}(BA) = K$. This single LoRA can represent any linear combination of up to $K$ linearly independent rank-1 matrices. Therefore, it can represent any matrix $M$ that is the sum of $K$ rank-1 matrices.

Thus, the representation space of the single LoRA $BA$ with rank $K$ includes all possible combinations that can be formed by the individual rank-1 LoRAs $\{B_k A_k\}$.

■

*Corresponding author

### Proof of Corollary 1

*Statement:*

For a set of LoRAs $\{B_k A_k\}_{k=1}^{K}$, the combined representation capability does not exceed that of a single LoRA with rank $r = \sum_{k=1}^{K} \text{rank}(B_k A_k)$.

*Proof:*

For each LoRA matrix $B_k A_k$, let its rank be $r_k = \text{rank}(B_k A_k)$. By the rank decomposition theorem, each $B_k A_k$ can be expressed as a sum of $r_k$ rank-1 matrices. That is,

$$B_k A_k = \sum_{i=1}^{r_k} B_{k,i} A_{k,i},$$

where $B_{k,i} A_{k,i}$ are rank-1 matrices.

Consider the sum of all the LoRA matrices:

$$M = \sum_{k=1}^{K} B_k A_k = \sum_{k=1}^{K} \sum_{i=1}^{r_k} B_{k,i} A_{k,i}.$$

The total number of rank-1 matrices in this sum is $r = \sum_{k=1}^{K} r_k$.

By **Proposition 1**, the set of rank-1 matrices $\{B_{k,i} A_{k,i}\}$ can be represented within the space of a single LoRA matrix $BA$ of rank $r$. Therefore, there exist matrices $B$ and $A$ such that

$$M = BA,$$

with $\text{rank}(BA) = r$.

Since the combined effect of the original LoRA matrices $\{B_k A_k\}$ can be represented by a single LoRA matrix of rank $r$, it follows that their combined representation capability does not exceed that of a single LoRA of rank $r$.

■

**Proof of Corollary 2**

*Statement:*

For a LoRA $BA$ with rank $r$, the LoRA $B(A \odot \sigma(T))$ can be decomposed into any set of LoRA groups $\{B_k A_k\}_{k=1}^{K}$, provided that the constraint $r \geq \sum_{k=1}^{K} \text{rank}(B_k A_k)$ holds. Here, $T$ is a matrix with the same shape as $A$, and $\sigma$ is a non-linear activation function.

*Proof:*

Define $M = A \odot \sigma(T)$, where $A \in \mathbb{R}^{m \times n}$ and $\sigma(T)$ is defined as:

$$\sigma(T) = [[t_1]_m, [t_2]_m, \ldots, [t_n]_m],$$

where $[t_i]_m$ is a column vector of size $m$, with all entries equal to $t_i$. The rank of $M$ depends on the structure of both $\sigma(T)$ and $A$:

If $t_i = 0$ for certain indices, the corresponding columns of $M$ are zero, reducing the rank of $M$. Consequently, the rank of $M$, denoted $r'$, satisfies:

$$r' = \text{rank}(M) \leq \min(\text{rank}(A), \text{rank}(\sigma(T))).$$

From **Corollary 1**, a LoRA group with a total rank $r$ is at most as expressive as a single LoRA $BA$ with rank $r$. Similarly, a LoRA group with a total rank $r'$ is at most as expressive as the configuration $B(A \odot \sigma(T))$, where $\text{rank}(M) = r'$.

Now consider decomposing $B(A \odot \sigma(T))$ into $K$ LoRA groups:

$$B(A \odot \sigma(T)) \approx \sum_{k=1}^{K} B_k A_k,$$

where $\text{rank}(B_k A_k) = r_k$. By the additive nature of rank, the total rank satisfies:

$$\sum_{k=1}^{K} r_k \leq r' \leq r.$$

Thus, if the total rank satisfies $r \geq \sum_{k=1}^{K} \text{rank}(B_k A_k)$, the configuration $B(A \odot \sigma(T))$ can be decomposed into any LoRA group where the total rank does not exceed $r$.

■

## B. Implementation Details on Downstream Tasks

**Base model setting** `LLaVA-1.5-7B` integrates a `CLIP ViT-L` [7] with an image resolution of 336px and a patch size of 14, a two-layer MLP projector to map visual features into tokens, and `Vicuna v1.5` [2] as the language model. In our setup, we omit any special tokens, such as image indicator tokens, to enclose visual tokens.

**Optimization setting** For optimization, we use the Adam optimizer across all trainable parameters. During the pre-training stage, we set a global batch size of 256 and a learning rate of 0.001 with no weight decay. In the fine-tuning stage, we use a global batch size of 128, setting the learning rate to 0.0002 for the LMM adapter and 0.00002 for the vision feature projector, respectively.

**Implementation Details of Deformable Attention** For the input to deformable attention, we use the anchor feature (extracted from the second-to-last feature map of the `CLIP ViT-L` image encoder) as the query embeddings. The key and value embeddings are sampled from points in multi-level reference feature maps, specifically obtained from the 2nd-to-last, 8th-to-last, 14th-to-last, and 20th-to-last layers of the `CLIP ViT-L` encoder. The number of sampled reference points matches the number of image patches (e.g., $(336/14)^2 = 24^2$ for `LLaVA-1.5-7B`) and is initially distributed uniformly across the feature map, following the methodology described in Deformable DETR [10].

## C. Experimental setup on MLLM Benchmarks.

**Benchmarks and metircs** For the general benchmark evaluation, following LLaVA [4], we use a 558K subset of the `LAION-CC-SBU` dataset [8] and the `llava-v1.5-mix665k` dataset [4] for instruction fine-tuning. Four widely used MLLM benchmarks are used for evaluation, including MMBench [6], SEED-Bench [3], and LLaVA-Bench In-the-Wild ( $\text{LLAVA}^{\text{W}}$ ) [4]. The first three benchmarks assess various MLLM capabilities, such as perceptual understanding and visual reasoning, through binary yes/no questions (MME) or multiple-choice questions (MMBench and SEED-Bench). We use the image-only subset of SEED-Bench ($\text{SEED}^{\text{I}}$), a popular choice for many image-based MLLMs [1, 9].

**Training Data Configuration for the Two Training Stages.** For the UniFood dataset, we employ descriptive tasks—specifically, ingredient recognition and recipe generation—in the first training stage to train the vision projector and VCE module. In the second stage, the vision projector, VCE module, and LLM adaptor are trained using all available task data. For Flickr30k and ScienceQA, the full set of multimodal training data is used in both stages. For general MLLM benchmarks, please refer to the training data configuration of LLaVA 1.5 7B-LoRA [5] for details on each stage.

**Implementation Details.** Our experiments are conducted on a setup with 4 $\times$ A100 GPUs (80GB). The selected layers for LLM adapter injection match those used in `LLaVA-1.5-7B-LoRA` [4], with the adapter rank $r$ set to 128 and scaling factor $\alpha$ set to 256.

## D. More Qualitative Results

We present additional qualitative results, encompassing both general visual tasks and downstream tasks. For the downstream tasks on the UniFood dataset, since no official fine-tuned version is available, we fine-tune `LLava-1.5-7B` using vanilla LoRA on UniFood. The comparison between the baseline method and our proposed approach demonstrates that our method generates more comprehensive and accurate answers, as shown in Figures 1 and 2. Furthermore, it produces answers with significantly greater knowledge consistency on downstream tasks compared to the vanilla LoRA fine-tuning method, as illustrated in Figures 3 and 4.

## E. Performance with Varying Layer Feature Maps as Anchor Features

We assess performance by selecting different layers as anchor features to project vision features onto the UniFood dataset within the VCE module. The results, presented in Table 1, indicate that the best performance is achieved when the anchor layer aligns with the vision feature projection layer in the pretrained model. In contrast, using other layers may degrade performance, likely due to misalignment between their feature maps and the pretrained knowledge embedded in the model.

| anchor layer | IoU(↑) | F1(↑) | BLEU(↑) | Rouge-L(↑) | pMAE(↓) |
|---|---|---|---|---|---|
| -2 | **24.5** | **33.5** | **14.7** | **42.2** | **49.1** |
| -8 | 20.25 | 29.99 | 12.67 | 38.86 | 53.03 |
| -14 | 16.75 | 25.82 | 10.1 | 34.88 | 58.23 |
| -20 | 13.68 | 21.54 | 7.15 | 30.27 | 65.55 |

Table 1. Performance vs. **different anchor layers** on UniFood.

## F. Detailed Results on Memory vs. Performance of LLM Adapter

This section presents detailed results on the performance and memory usage of the LoRA adapter under varying rank settings on the UniFood dataset. As listed in Table 2, we present the results of the vanilla LoRA and Dual-LoRA across three configurations of total rank values (32, 64, and 128). The results indicate that Dual-LoRA consistently outperforms the vanilla LoRA counterparts across all rank settings (32, 64, and 128) with only a slightly higher number of parameters. Notably, Dual-LoRA achieves superior performance even with fewer trainable parameters than LoRA with higher ranks, such as Var2-Dual-LoRA vs. Var3-LoRA and Var5-LoRA, as well as Var4-Dual-LoRA vs. Var5-LoRA.

| | Rank | LMM Adapter | LMM Adapter Memory Size | *Ingredient Recognition* | | Recipe Generation | |
|---|---|---|---|---|---|---|---|
| | | | | IoU | F1 | SacreBLEU | Rouge-L |
| var1 | 32 | LoRA | 32M | 19.2 | 29.7 | 10.4 | 41.1 |
| var2 | | Dual-LoRA | 48M | **22.7** +18.1% | **33.6** +13.3% | **12.9** +24.1% | **41.3** +0.3% |
| var3 | 64 | LoRA | 64M | 20.6 | 31.3 | 12.7 | 41.3 |
| var4 | | Dual-LoRA | 96M | **23.0** +11.7% | **33.7** +7.7% | **14.4** +13.1% | **41.7** +0.9% |
| var5 | 128 | LoRA | 128M | 21.6 | 32.5 | 13.0 | 41.3 |
| var6 | | Dual-LoRA | 192M | **23.6** +8.9% | **34.4** +5.9% | **14.8** +14.3% | **42.9** +3.6% |

Table 2. Performance and memory usage of the LoRA and Dual-LoRA adapters under different rank settings on the UniFood dataset. Only the parameters of LoRA and Dual-LoRA are set to be trainable.

**User:** Can you give a short introduction to this painting?

**LLaVA-1.5-7B:**

The image is a painting of a night scene featuring a large, dark blue sky with a few stars visible. The painting also includes a moon, which is located towards the top right corner of the scene. The sky is filled with swirling clouds, creating a dynamic and dramatic atmosphere.

In the foreground, there is a small town with a church, a windmill, and a few houses. The town is situated below the night sky, creating a sense of depth and perspective in the painting. The combination of the night sky, the town, and the church creates a captivating and serene scene.

**Dual-LoRA + VCE:**

The image is a captivating representation of the night sky, **painted in the style of the famous artist Vincent Van Gogh**. The canvas is filled with a vibrant display of stars, painted in various shades of blue and white, creating a sense of depth and dimension. **The artist has skillfully used the technique of impasto, where paint is applied in thick layers, to add texture and a sense of movement to the scene.**

In the foreground, a small village is depicted, with houses and trees painted in shades of green and brown. The contrast between the dark blue night sky and the lighter colors of the village creates a striking visual effect. The painting is a beautiful representation of the artist's unique style and his ability to capture the beauty of the night sky.

Figure 1. **Qualitative Results.** Our proposed Dual-LoRA + VCE method accurately identifies the painting style, specifically Vincent Van Gogh's style, and provides more detailed insights into the artwork's analysis.


**User:** Can you explain this meme?

**LLaVA-1.5-7B:**

The meme consists of **three pictures**, each showing a different landscape. The first picture displays a snowy mountain range, likely in Iceland, with a person skiing in the foreground. The second picture features a lush green valley, possibly in Greenland, with a person standing in the middle of the scene. **The third picture showcases a cloudy sky, which could be either Iceland or Greenland, with a person standing on a snowy hill.**

The meme is accompanied by a caption that reads, "This is why I have trust issues." The caption seems to be a play on words, as the pictures depict different environments, **but the person in each picture is wearing the same backpack**. The caption might be suggesting that the person in the pictures is untrustworthy or unreliable, as they are seen in various locations with different landscapes.

**Dual-LoRA + VCE:**

The image is a split-screen collage, showcasing **two distinct landscapes**. On the top, **there's a breathtaking view of Iceland**, a country known for its stunning natural beauty. The image captures a vast expanse of green and blue hues, with mountains and valleys stretching into the distance. The sky above is a clear blue, adding to the serene atmosphere.

On the bottom, the **image transports us to Greenland**, another land of natural beauty. The landscape here is a mix of snow-covered mountains and icy plains, with people scattered across the scene, possibly exploring or enjoying the view. The sky above is a darker shade of blue, suggesting a colder climate.

The collage is framed by a white border, and the words "This is why I have trust issues" are written in the top left corner, possibly hinting at the challenges faced by those who venture into such remote and harsh environments. The image, in its entirety, provides a glimpse into the awe-inspiring landscapes of Iceland and Greenland.

Figure 2. **Qualitative Results.** Our proposed Dual-LoRA + VCE method provides more accurate answers, correctly describing both the views of Iceland and Greenland depicted in the image. However, neither the baseline method nor our approach fully captures the humor of the image, which lies in the fact that the land names do not align with their literal meanings or actual landscapes, thereby leading "trust issues".


Figure 3. **Qualitative Results.** The Dual-LoRA + VCE model provides a more accurate and consistent response, seamlessly aligning the ingredients and cooking method with the dish's presentation and its likely preparation process. In contrast, the LLaVA-1.5-7b-LoRA model, though detailed, introduces unnecessary elements and inconsistencies, rendering it less reliable for precise food analysis. This underscores the superior effectiveness of the Dual-LoRA + VCE model in resolving data conflicts and ensuring contextual accuracy.

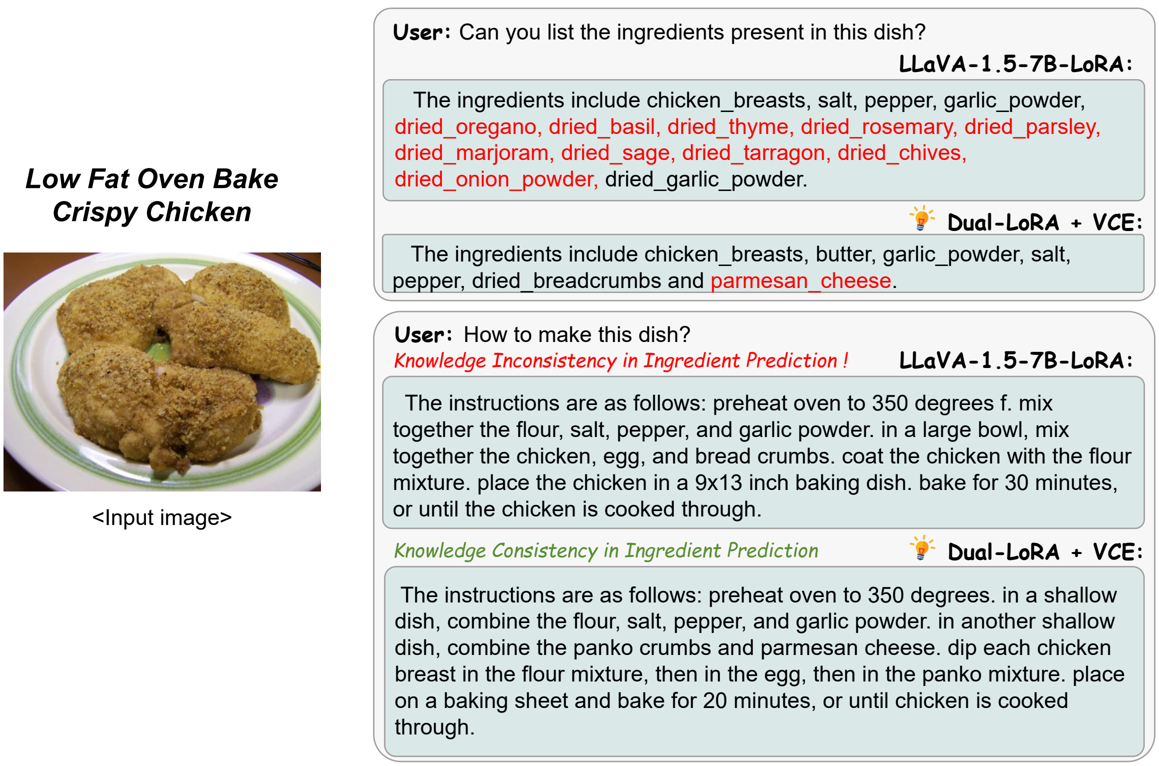


Figure 4. **Qualitative Results.** This comparison highlights the strengths of the Dual-LoRA + VCE model in generating consistent, contextually accurate responses to both ingredient prediction and recipe instructions. In contrast, the LLaVA-1.5-7b-LoRA model struggles with knowledge consistency, providing a lengthy and potentially over-generalized ingredient list and instructions that do not fully align with the dish.

## References


[1] Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13817–13827, 2024. 2

[2] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023. 2

[3] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024. 2

[4] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 2

[5] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 2

[6] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233. Springer, 2025. 2

[7] 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. 2

[8] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2

[9] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*, 2024. 2

[10] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2