MODALITY-SPECIALIZED SYNERGIZERS FOR INTERLEAVED VISION-LANGUAGE GENERALISTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advancements in Vision-Language Models (VLMs) have led to the emergence of Vision-Language Generalists (VLGs) capable of understanding and generating both text and images. However, seamlessly generating an arbitrary sequence of text and images remains a challenging task for the current VLGs. One primary limitation lies in applying a unified architecture and the same set of parameters to simultaneously model discrete text tokens and continuous image features. Recent works attempt to tackle this fundamental problem by introducing modality-aware expert models. However, they employ identical architectures to process both text and images, disregarding the intrinsic inductive biases in these two modalities. In this work, we introduce MODALITY-SPECIALIZED SYNERGIZERS (MOSS), a novel design that efficiently optimizes existing unified architectures of VLGs with modality-specialized adaptation layers, i.e., a Convolutional LoRA for modeling the local priors of image patches and a Linear LoRA for processing sequential text. This design enables more effective modeling of modality-specific features while maintaining the strong cross-modal integration gained from pretraining. In addition, to improve the instruction-following capability on interleaved text-and-image generation, we introduce LEAFINSTRUCT, the first open-sourced interleaved instruction tuning dataset comprising 184,982 high-quality instances on more than 10 diverse domains. Extensive experiments show that VLGs integrated with MOSS achieve state-of-the-art performance, significantly surpassing baseline VLGs in complex interleaved generation tasks. Furthermore, our method exhibits strong generalizability on different VLGs.

1 INTRODUCTION

As multimodal learning research advances, there is a growing trend of building Vision-Language Generalists (VLGs) (Sun et al., 2023b; 2024; Koh et al., 2023; Aghajanyan et al., 2022; Li et al., 2023b; Dong et al., 2024; Team, 2024) that can comprehend and generate *interleaved* text and images, where multiple text segments and images are presented in arbitrary sequences. Compared with previous Vision-Language Models (VLMs) (Alayrac et al., 2022; Li et al., 2023c; Liu et al., 2023c) that can only generate text and diffusion models (Ramesh et al., 2021; Rombach et al., 2022) that can only produce images, such VLGs enable a wider array of applications that require the simultaneous generation of both images and text, such as script generation (Qi et al., 2024), visual storytelling (Huang et al., 2016), and many others.

Despite these recent advancements, one notable issue of existing VLGs is that they often fail to produce coherent and high-quality interleaved text and images. As shown in the top example in Figure 1, current state-of-the-art VLG, e.g., Emu2 (Sun et al., 2024), still suffers from poor text and image quality, including heavy repetition in text and unnatural distortions in the image. We attribute this issue to a fundamental challenge: *existing VLGs use the same architecture (i.e., transformer backbone) with the same set of parameters to process both text and images, which may not be sufficient to model the distinct inductive biases in each modality given their intrinsic discrepancy.* For example, text follows a linear, left-to-right sequence, whereas images are inherently two-dimensional, composed of local priors in adjacent patches. Previous studies show that the transformer architecture (Vaswani et al., 2017) predominantly employed in current LLMs and VLGs excels at sequence modeling. But compared with convolutional architectures, transformer is less effective at modeling local priors of adjacent image patches, which is crucial for various vision tasks (Zhong et al., 2024;

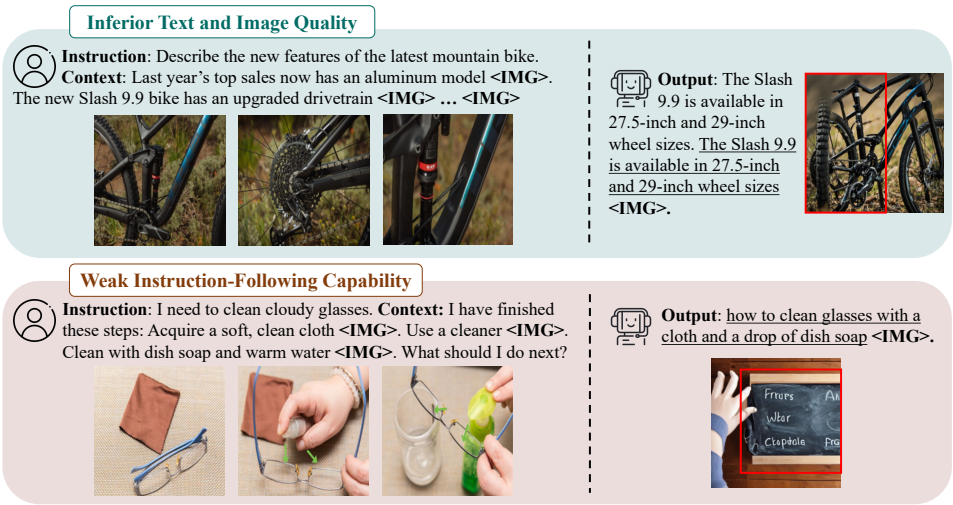


Figure 1: Failure cases of existing VLGs (Emu2 at the top and GILL at the bottom). The output text with inferior quality is highlighted with underline. The regions that impede output images' quality are highlighted with red bounding boxes.

Chen et al., 2023d). Thus, applying a unified architecture with the same set of parameters can result in poorer performance in mixed-modal generation, such as producing images with local inconsistency and distortion among adjacent patches. Recently, several works (Akbari et al., 2023; Ye et al., 2024) have proposed modality-aware expert models as an attempt to tackle this problem. However, they still apply the same architecture to process both text and images, ignoring the inherent discrepancies between these two modalities. These challenges underscore the need for careful architectural design and specialized allocation of model parameters tailored to each modality.

Another critical challenge is existing VLGs (Sun et al., 2023b; 2024) often fail to adhere to human instructions to perform interleaved generation tasks. In the bottom example in Figure 1, GILL (Koh et al., 2023) failed to accurately follow the instruction and the context to complete the next step for *"clear cloudy glasses"*. Instead, GILL produces unhelpful text that is repetitive with the input and an image irrelevant to the task. While existing VLGs are often pretrained on interleaved documents (Zhu et al., 2023c), they are only instruction-tuned for single-modality generation, e.g., either text or image generation, leading to a weak instruction-following capability of interleaved generation. Moreover, there is a lack of large-scale instruction-following data specifically designed for interleaved generation, making the interleaved instruction tuning not scalable and less feasible.

To address these fundamental challenges, we first propose MODALITY-SPECIALIZED SYNERGIZERS (MOSS), a novel framework that introduces modality-specialized parameters to seamlessly handle the inductive biases of different modalities within the unified architectures of VLGs. As lightweight adaptation layers, our proposed MOSS is generic and can be integrated into most, if not all, existing VLGs without requiring expensive pre-training. Specifically, for images, we introduce **Convolutional** Low-Rank Adaptation (Convolutional LoRA) layers to better model the local prior of image patches. For text, we employ a separate set of **Linear** Low-Rank Adaptation (LoRA) layers, acknowledging the distinct sequential modeling process of text compared to images. During finetuning, both modality-specialized architectures are zero-initialized and progressively fine-tuned to learn their modality-specific features, while the VLG's parameters remain frozen, maintaining strong cross-modal integration gained from massive pre-training. Our design allows each modality to have a better representation of modality-specific features with its own specialized parameters and optimal architectural design. Additionally, to improve VLG's instruction-following capabilities under diverse interleaved generation scenarios, we introduce LEAFINSTRUCT, the first open-sourced high-quality interleaved instruction tuning data with 184,982 instances spanning more than 10 domains. To obtain high-quality instruction data at scale, we develop a rigorous automatic pipeline.

To validate the effectiveness and generalizability of our method and dataset, we adopt our method on two different VLG backbones with discrete and continuous image token spaces, and conduct extensive experiments on multiple datasets. The results demonstrate that the VLGs instruction-tuned with our method achieves state-of-the-art performance across most evaluation aspects. Particularly, our

method can produce interleaved content with better quality, including text quality, image coherence, text-image consistency, and helpfulness. In summary, our contributions are threefold. **First**, we introduce MoSS, a novel design that enhances VLGs to generate interleaved content with modality-specialized parameters and adaptation architectures. To the best of our knowledge, we are the first to apply different adaptation architectures within an autoregressive generative model to improve interleaved generation. **Second**, to fill the blank in existing resources and improve the instruction-following capability of VLGs, we introduce the first open-sourced large-scale instruction-tuning dataset across diverse domains. **Third**, by instruction-tuning existing VLGs with a small number of parameters, we achieve significant performance improvement on most aspects of evaluation benchmarks, outperforming existing open-source baselines by 34.7% on InterleavedBench. We also demonstrate that our approach can effectively generalize to different VLG backbones.

2 Related Work

Interleaved Vision-Language Models There are two popular formulations for VLGs: The first leverages VQGAN (Esser et al., 2021) to quantize an image into a long sequence of discrete tokens and add the vocabulary in VQGAN's codebook into the vocabulary of LLMs (Aghajanyan et al., 2022; Yu et al., 2023; Yasunaga et al., 2023; Team, 2024; Jin et al., 2023). In this way, the LLMs are trained with a unified autoregressive objective to predict image tokens or text tokens. The predicted image tokens are fed into a VQGAN decoder to reconstruct images. The second formulation employs the CLIP image encoder to transform images into sequences of continuous embeddings (Koh et al., 2023; Tang et al., 2023; Zhu et al., 2023b; Sun et al., 2023b; 2024; Li et al., 2024b; Wu et al., 2023; Tian et al., 2024), which are then concatenated with text embeddings in their original order. Compared to the first approach, this formulation often requires shorter sequences to represent an image and generally yields superior performance. Our proposed method requires minimal assumptions on VLG's architectures and can be applied to many of the existing transformer-based VLGs.

Visual Instruction Tuning Xu et al. (2023a) propose MultiInstruct, the first human-label visual instruction tuning dataset to improve the generalizability of VLMs. LLaVA (Liu et al., 2023c) leverages GPT-4 to convert image captions from existing annotations into three tasks, including visual dialogues, visual question answering, and detail captions. Following studies either utilize proprietary LLMs (Dai et al., 2023; Ye et al., 2023; Yin et al., 2023; Liu et al., 2023b; Li et al., 2023a; Lyu et al., 2023; Zhu et al., 2023a; Wang et al., 2023; Chen et al., 2023b) or human efforts (Liu et al., 2023b; Xu et al., 2024) to augment visual instruction tuning tasks. Several studies target specific aspects of VLMs' capability, such as domain and instruction bias (Avrahami et al., 2022; Liu et al., 2023a), object grounding (Chen et al., 2023a), and OCR (Zhang et al., 2023b; Hu et al., 2023). Instruction tuning has also been widely applied to other vision-language tasks, such as image editing (Brooks et al., 2023a) and interleaved text-image understanding (Jiang et al., 2024). Hu et al. (2024) finetune a model that can follow multimodal instructions to generate desired images. However, most existing instruction-tuning datasets only consider the tasks where the outputs are in a single modality, i.e., either text or image. *To facilitate the training and enhance the instruction-following capabilities for VLGs, we curated* LeafInstruct*, the first instruction-tuning dataset tailored for interleaved text-image generation across diverse domains, where the inputs and outputs can contain interleaved text and multiple images.*

Parameter-Efficient Finetuning (PEFT) PEFT methods (Hu et al., 2021; Li & Liang, 2021; Karimi Mahabadi et al., 2021; Zaken et al., 2022; Jia et al., 2022; Lian et al., 2022; Jie & Deng, 2022; Liu et al., 2022; Chen et al., 2023d; Zhong et al., 2024) aim to adapt pretrained large models to various downstream tasks and have become prevalent in instruction tuning. Typically, these methods involve freezing the pretrained large models while finetuning a minimal set of newly introduced parameters. Recent studies (Wang et al., 2022; Zadouri et al., 2023; Lin et al., 2024; Shen et al., 2024) propose to combine PEFT methods with Mixture-of-Experts to mitigate task interference and enhance performance, particularly in visual instruction tuning where models need to process inputs from two modalities. *Our proposed* MoSS *is the first PEFT method that utilizes two distinct LoRA architectures—linear and convolutional—for text and image generation within autoregressive VLGs.*

3 Background: Autoregressive Vision-Language Generalists

Existing autoregressive VLGs can be broadly classified into two categories: those that represent each image as a sequence of *discrete tokens* (Yasunaga et al., 2023; Aghajanyan et al., 2022; Team, 2024),

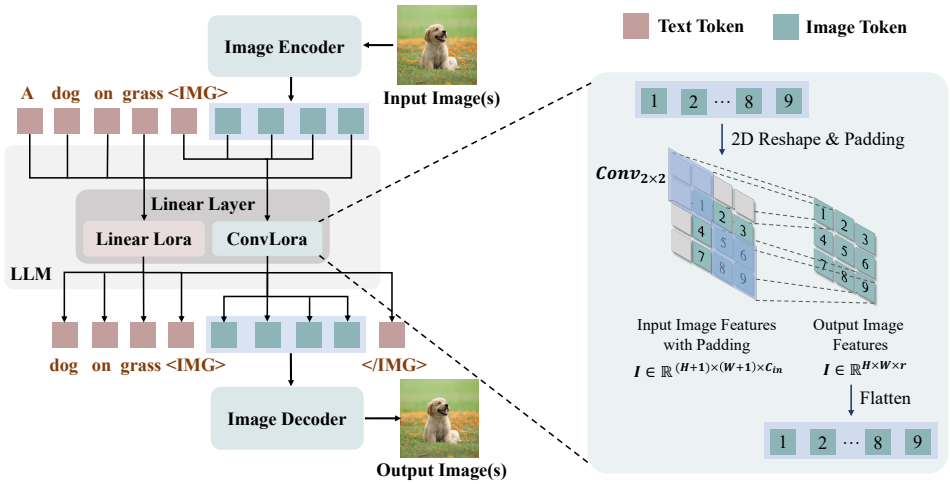


Figure 2: An autoregressive VLG with our proposed MoSS added to its linear layers. The linear LoRA on the left side is specialized to generate text tokens and the Convolutional LoRA on the right side is specialized to generate image patches. On the right handside, we show the details of convolutional operation applied to autoregressively generate image tokens. Best viewed in color.

and those that represent each image as a sequence of *continuous vectors* (Sun et al., 2023b; 2024). However, despite these differences in image representation, their underlying model architectures and formulations for vision-language generation remain largely similar. Thus, we do not differentiate them in the following formulation.

Model Architecture Autoregressive VLGs typically comprise three components: an image encoder (e.g., CLIP (Sun et al., 2023a) or VQ-VAE encoder (Gafni et al., 2022)), a decoder-only large language model (LLM), and an image decoder (e.g., a diffusion model (Podell et al., 2023) or VQ-VAE decoder (Gafni et al., 2022)). Given a sequence of interleaved text segments and images, the image encoder processes each image into a sequence of image tokens. These image tokens are then concatenated with the text tokens in their original order and input into the LLM. The LLM autoregressively predicts the next token, which could be either text or image. Finally, the image decoder takes in the predicted image tokens and reconstructs the target image.

Training Objective The training objective of VLGs can be loosely defined in the following unified autoregressive manner.

$$\arg\max_{\theta} \sum^{\mathcal{D}} \sum_{n=1}^{N} P_\theta(s_n | s_1, s_2, ..., s_{n-1}) \tag{1}$$

where θ denotes the model parameters, N denotes the input sequence length, $\mathcal{D}$ denotes the training dataset, and s_i denotes a text token or an image-patch embedding. This unified objective is optimized through two types of losses: (1) If the image is represented as discrete tokens, the CrossEntropy loss is employed to minimize the divergence between the predicted probability distribution of the image or text tokens and the ground truth distribution; (2) If the image is encoded as continuous vectors, the mean-squared-error (MSE) loss is used to minimize the difference between the predicted and actual image embeddings.

4 Modality-Specialized Synergizers (MoSS)

In this section, we first detail the two modality-specialized adaptations in MoSS: Linear LoRA for text tokens and Convolutional LoRA for image tokens. We then describe the process of synergistically integrating these adaptations into autoregressive VLGs to perform interleaved generation.

4.1 Linear Low-Rank Adaptation (LoRA)

LoRA (Hu et al., 2021) is a parameter-efficient finetuning method that freezes the pretrained model parameters and injects low-rank decomposable matrices into the layers of transformers. Formally,

given the weights in a linear layer $\mathbf{W} \in \mathbb{R}^{d_{out} \times d_{in}}$, LoRA modifies the weights by adding a decomposable weight matrix $\Delta\mathbf{W}$ to $\mathbf{W}$. Thus, for a vector $\mathbf{h} \in \mathbb{R}^{d_{in}}$, the modified linear transformation $T : \mathbb{R}^d_{in} \rightarrow \mathbb{R}^d_{out}$ becomes:

$$T(\mathbf{h}) = \mathbf{h}(\mathbf{W} + \Delta\mathbf{W})^\intercal = \mathbf{h}\mathbf{W}^\intercal + \mathbf{h}\Delta\mathbf{W}^\intercal \tag{2}$$

$\Delta\mathbf{W}$ is decomposed into two low-rank matrices, i.e., LoRA A: $\mathbf{W}_A \in \mathbb{R}^{r \times d_{in}}$ and LoRA B: $\mathbf{W}_B \in \mathbb{R}^{d_{out} \times r}$ satisfying the low-rank constraint $r \ll min(d_{out}, d_{in})$. The final expression is

$$T(\mathbf{h}) = \mathbf{h}\mathbf{W}^\intercal + \alpha\mathbf{h}\mathbf{W}_A^\intercal\mathbf{W}_B^\intercal \tag{3}$$

where $\alpha \in \mathbb{R}$ is a hyper-parameter.

4.2 Convolutional Low-Rank Adaptation (Convolutional LoRA)

We propose Convolutional LoRA, a variant of LoRA specifically designed for modeling the local structure of image hidden states during image generation, by improving the architecture proposed in Zhong et al. (2024). Detailed empirical comparison between two Convolutional LoRAs can be found in Table 10 in Appendix C.3. The previous approach first reduces the dimension of input features and then performs the convolution operation within a lower-dimension space. Since dimension reduction can cause information loss, the convolution within a reduced dimension can be less effective at modeling the local priors of image patches. On the contrary, our method performs convolution in the original input feature space and the dimension is deducted during the convolution process, which alleviates the information loss issue in the previous design.

Specifically, our approach consists of a convolutional LoRA A layer, i.e., $\text{Conv}_{k \times k}$, where the kernel size is $k \times k$, the number of input channels is c_{in}, and the number of output channels is r, as well as a LoRA B: $\mathbf{W}_B \in \mathbb{R}^{C_{out} \times r}$. Given the 2D feature $\mathbf{I} \in \mathbb{R}^{H \times W \times C_{in}}$ of an image, where H denotes the height, W denotes the width, and C_{in} denotes the number of channels of $\mathbf{I}$, the convolutional LoRA A projects down its number of channels to r and simultaneously performs convolution operation. Then the LoRA B projects its number of channels up to C_{out}. The equation 3 becomes:

$$\tilde{T}(\mathbf{I}) = \mathbf{I}\mathbf{W}^\intercal + \alpha\text{Conv}_{k \times k}(\mathbf{I})\mathbf{W}_B^\intercal \tag{4}$$

where α is a hyper-parameter.

4.3 Integrating MoSS into VLGs

As shown in Figure 2, we propose to integrate two types of adaptations into VLGs, i.e., using **Linear LoRA** for text generation and **Convolutional LoRA** for image generation. Formally, let $\mathbf{W} \in \mathbb{R}^{d_{out} \times d_{in}}$ be the weights of any linear layer in a LLM, and let $\mathbf{H} = [\mathbf{h}^t_1, ..., \mathbf{h}^t_m, \mathbf{h}^i_{m+1}, \mathbf{h}^i_{m+2}, ..., \mathbf{h}^i_{m+(H \times W)}, ..., \mathbf{h}^t_N] \in \mathbb{R}^{N \times d_{in}}$ denotes the hidden states of a sequence of interleaved text and images, where a subscript indicates position of a hidden state and the superscript indicate if a hidden state is decoded into a text token (t) or decoded into an image-patch embedding (i). We untie $\mathbf{H}$ into text hidden states $\mathbf{H}^t = [\mathbf{h}^t_1, \mathbf{h}^t_2, ..., \mathbf{h}^t_m, \mathbf{h}^t_{m+(H \times W)+1}, ..., \mathbf{h}^t_N]$ and image hidden states $\mathbf{H}^i = [[\mathbf{h}^i_{m+1}, ..., \mathbf{h}^i_{m+(H \times W)}], [\mathbf{h}^i_{n+1}, ..., \mathbf{h}^i_{n+(H \times W)}], ...]$, where $m + 1$ and $n + 1$ denote the starting positions of two subsequences of image hidden states. Each subsequence of a single image has a fixed length of $H \times W$ and we reshape the hidden states of each image in $\mathbf{H}^i$ into a 2D structure. Hence, the dimension of $\mathbf{H}^i$ becomes $B \times H \times W \times C_{in}$, where B denotes the number of images in the sequence $\mathbf{H}$. We feed $\mathbf{H}^t$ into the Equation 3 to get $\hat{\mathbf{H}}^t = T(\mathbf{H}^t)$ and $\mathbf{H}^i$ into Equation 4 to get $\hat{\mathbf{H}}^i = \tilde{T}(\mathbf{H}^i)$.

It is non-trivial to integrate convolutional operation in auto-regressive model and to the best of our knowledge, we are the first to incorporate the convolutional architecture to improve interleaved generation. The right part of Figure 2 visualizes the convolutional operation applied to a sequence of image patches. The squares on the left denote the reshaped 2-dimensional input image patches and the larger blue squares denote the 2×2 convolution kernels. The number on each square denotes the original positions of a patch in the image sequence. For demonstration purposes, we draw image patches with $H = 3$ and $W = 3$. Note that the current hidden state of an image patch can only

depend on previous hidden states since we use the autoregressive architecture. Thus, when applying the convolution operation on an image patch, the kernel only covers neighboring patches on the top and left sides of a patch. For example, the new hidden state of patch 9 is computed from patches: $5, 6, 8,$ and 9. To preserve the shape $(H \times W)$ of the input image patches, we pad the reshaped image hidden states with zero vectors on the top and left sides, as shown by the grey squares in Figure 2. Finally, we assemble $\hat{\mathbf{H}}^i$ and $\hat{\mathbf{H}}^t$ back to their original sequence to form $\hat{\mathbf{H}}$.

5 Interleaved Instruction Tuning with LeafInstruct

Existing interleaved vision-language models (Sun et al., 2023b; 2024; Dong et al., 2024) predominantly follow the training procedures that they are first pretrained on massive corpora of interleaved data such as MMC4 (Zhu et al., 2023c) and other resources and then finetuned on a mix of high-quality datasets, such as visual instruction tuning data in Liu et al. (2023c) and InstructPix2Pix (Brooks et al., 2023b). However, one significant limitation of these instruction-tuning datasets is that the outputs are typically in a single modality, e.g., either text or image, which hinders the instruction-following capability of VLGs especially in generating interleaved text and images specified by the given instructions.

5.1 Dataset: LeafInstruct

To bridge the gap between limited existing resources and the practical need for improving interleaved generation models, we curated LeafInstruct, the first comprehensive instruction tuning dataset for interleaved text-and-image generation. Each instance in our dataset consists of (1) a detailed instruction, (2) an input context with interleaved text and images, and (3) a ground-truth output also with interleaved text and images. We show an example of LeafInstruct in Figure 6, and compare it with other representative instruction-based datasets.

Dataset construction We construct a diverse instruction-tuning data collection from large-scale web resources and academic datasets, including MMDialog (Feng et al., 2023), VIST (Huang et al., 2016), WikiWeb2M (Burns et al., 2023) and YouCook2 (Zhou et al., 2018). Since the original data sources can be noisy, we meticulously devised an automatic data annotation pipeline to ensure the high quality of our curated data. We include the details on dataset construction in Appendix B. We also conducted a rigorous human assessment of our dataset (see Section 7).

Dataset Statistics After applying our rigorous data processing pipeline, we totally obtain 184,982 high-quality instances out of more than 7 million source samples. Our dataset covers a wide range of realistic instruction-tuning tasks, including multimodal document completion, multimodal dialogue, visual storytelling, multimodal script generation, and knowledge-intensive generation. In Appendix B, we show the domain distribution of LeafInstruct in Figure 7 and compare our dataset with existing datasets in Table 4. These analyses effectively demonstrate the diversity and the novelty of our dataset.

5.2 Instruction Tuning for Interleaved Generation

With our curated LeafInstruct, we enable large-scale interleaved instruction tuning so that the model can learn how to follow human instructions to generate desired interleaved text and images. To preserve the VLG's capability obtained from pre-training, we only fine-tuned the modality-specialized adaptation layers, and the remaining parameters in the VLGs are kept frozen.

Specifically, as shown on the right side of Figure 6, given the task instruction and the interleaved context as inputs, the model is trained to autoregressively generate interleaved text tokens and images with two alternative generation modes for text and images, respectively. We use a special token <IMG> to indicate where an image occurs in the interleaved sequence. The training process is as follows: **(1)** The model is set to the **text generation** mode by default. During this mode, the hidden states of newly generated tokens are always routed to linear LoRA, and only the parameters in the linear LoRA are optimized. **(2)** After the <IMG> token is generated, the model switches to **image generation** mode. The VLG takes in the updated context ended with <IMG> and is trained to generate a fixed-length $(H \times W)$ sequence of image patch embeddings autoregressively.

All the hidden states of generated image embeddings are routed to Convolutional LoRA and only the parameters in the Convolutional LoRA are fine-tuned. **(3)** When the generation of an image is finished, the model is trained to predict an end-of-image token </IMG>, and the model will resume the text generation mode. This process will be iterated until the training on a sequence is finished.

Interleaved Inference The inference procedure of our framework is largely identical to the instruction tuning, where two generation processes iterate alternatively. The only key difference is that the fine-tuned VLGs will automatically determine when to generate a text segment or an image at their own discretion. The iterative generation process terminates when the model produces the end-of-generation token </s> at the end of a response. Note that although our inference process is designed for interleaved generation, we can also handle the cases where the outputs only contain text or images, enabling a wide range of applications.

6 EXPERIMENTS

6.1 EXPERIMENT SETUP

Evaluation Benchmarks We evaluate the interleaved generation capability of our method on InterleavedBench (Liu et al., 2024). InterleavedBench is a comprehensive dataset specifically tailored for interleaved evaluation. InterleavedBench covers a diverse array of tasks, where the evaluation data are either curated by the authors (e.g., *document completion*), or re-annotated based on subsets of well-established academic evaluation benchmarks, including *visual storytelling* from VIST (Huang et al., 2016), *activity generation* from ActivityNet (Krishna et al., 2017), *script generation* from WikiHow (Yang et al., 2021), *image editing* from MagicBrush (Zhang et al., 2023a), and *multi-concept image composition* from CustomDiffusion (Kumari et al., 2023). We include more details on evaluation benchmarks in Appendix C.1.

Evaluation Metrics We adopt InterleavedEval (Liu et al., 2024), a strong reference-free evaluation metric with a high correlation with human judgments based on GPT-4o, to conduct a holistic assessment of the quality of interleaved generation. InterleavedEval prompts GPT-4o to score an interleaved output from five aspects, including Text Quality, Perceptual Quality, Image Coherence, Text-Image Coherence (TIC), and Helpfulness. For each aspect, the GPT-4o outputs a discrete score from $\{0, 1, 2, 3, 4, 5\}$, where 0 is the worst and 5 is the best. We refer to the original paper (Liu et al., 2024) for a detailed definition of each score and each evaluation aspect. We also have an additional evaluation on image editing on the *full test set* of MagicBrush using well-established metrics, including CLIPScore (Hessel et al., 2021) and DINO (Caron et al., 2021) in Table 5 in Appendix C.2.

Implementation Details To demonstrate the generalizability of our method, we adopt our MoSS to two representative autoregressive VLG backbones, i.e., Emu2 (Sun et al., 2024) and Chameleon (Team, 2024), and fine-tune them on our LEAFINSTRUCT dataset. The rank number of all the LoRA is set to 256 by default. Note that for the Chameleon model, we adopt the implementation in Chern et al. (2024) since the original model and checkpoints are not publicly available. More implementation details including hyperparameters and GPU setups can be found in Appendix A.1.

Baselines For fair comparisons, we primarily compare our methods with current state-of-the-art **open-source** VLGs, including GILL (Koh et al., 2023), MiniGPT-5 (Zheng et al., 2023a), Pretrained Emu2, and Chameleon. We also report the performance of pipelines based on **proprietary** models, including Gemini 1.5 (Reid et al., 2024)+SDXL (Podell et al., 2023) and GPT-4o (OpenAI, 2024)+DALLE 3 (Betker et al.). For these baselines, we first prompt the VLMs (e.g., GPT-4o) to generate text along with image captions, and then feed the image captions to a separate image generation model (e.g., DALLE). We report these performances only for reference purposes.

6.2 MAIN RESULTS

Quantitative Results Table 1 presents the main results of our method in comparison to the baselines. We have the following findings. **Firstly**, our approach is highly effective and efficient when it is adapted to existing VLGs. Applying our MoSS to VLGs achieved significant improvement over their original performance on all evaluation aspects. For example, compared with the original Emu2 model, Emu2+MoSS achieved a performance gain of **up to 190.2%** (on Text-Image Coherence) and **97.76%** on the average of 5 aspects, almost doubling the overall performance. **Secondly**,

Table 1: **Main results of interleaved generation on InterleavedBench**. We show the performance of pipelines based on proprietary models (Top), open-source VLGs (Middle), and the VLGs trained with our MoSS and LeafInstruct (Bottom), respectively. Note that the scale is from 0 to 5 (5 is the best). We also report the percentage of improvement in our method over the original VLG backbone in the parentheses. The best results are highlighted in **bold**.

Model	Text Quality	Perceptual Quality	Image Coherence	TIC	Helpfulness
Proprietary Models					
Gemini1.5 + SDXL	3.37	4.34	3.34	3.98	3.28
GPT-4o + DALL·E 3	3.16	4.44	3.13	4.39	3.46
Open-Source Models					
MiniGPT-5	1.31	3.44	2.06	2.66	1.76
GILL	1.44	**4.02**	2.12	2.69	1.53
Emu2	1.33	2.29	1.71	1.22	1.87
Chameleon	**3.33**	0.67	0.28	0.47	1.43
Emu2 + MoSS (Ours)	2.61 (+96.2%)	3.62 (+58.1%)	**3.41** (+99.4%)	**3.54** (+190.2%)	**2.71** (+44.9%)
Chameleon + MoSS (Ours)	2.98 (-10.5%)	2.25 (+235.8%)	1.05 (+275%)	1.7 (+261.7%)	1.82 (+27.3%)

our method beats the previous open-sourced state-of-the-art (i.e., GILL) by a large margin, i.e., **34.7%** on the average of 5 aspects. Particularly, the outputs of our method have better coherence across images (w/ 37.8% improvement in Image Coherence) and between text and images (w/ 31.6% improvement in Text-Image Coherence). Our method also exhibits better instruction-following capability and is able to generate more helpful content given the 11.5% improvement in Helpfulness. **Thirdly**, it is worth noting that the Chameleon baseline achieves good performance on Text Quality but extremely poor performance on image-related aspects. We observed that Chameleon usually generates long and comprehensive text responses with no image output, thus leading to poor performance on image-related aspects. We hypothesize the reason lies in the lack of instruction tuning on interleaved generation with both text and images. From Table 1, our approach improves the original Chameleon by a significant margin, especially on image-related aspects. This shows that our interleaved instruction tuning can effectively enhance a VLG that was previously poor at mixed-modal generation. We noticed that adding MoSS in Chameleon can cause a slight performance drop in text quality. We discuss the details of this problem in Appendix C.3, conduct an additional human evaluation of text quality, and present the results and findings in Table 11. **Fourthly**, there remains a notable gap between open-sourced VLGs and the pipeline approaches based on proprietary models, indicating building a powerful and general-purpose open-sourced VLGs is still challenging.

Per-task Performance We also show the average performance on each task on InterleavedBench in Figure 3. Specifically, our method (i.e., Emu2-MoSS) outperforms the baselines on most tasks, often by a large margin. For subject-driven generation, the slightly lower performance of our approach compared to other baselines is due to its poorer perceptual quality. We included more detailed justifications for this result in the following section. We report the per-task performance on all aspects in Figure 9 in Appendix C.5. In addition, we include more results on well-established benchmarks of image-understanding, text-to-image generation, and image-editing in Table 5, 6, 7, 8, and 9 in Appendix C.2.

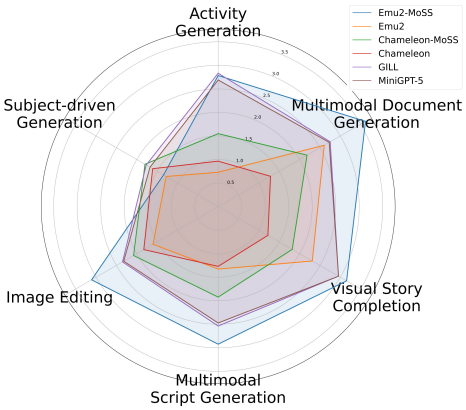


Figure 3: **Per-task performance** averaged on 5 aspects on InterleavedBench.

Qualitative Results To better interpret the results, we conducted a qualitative analysis on several open-sourced baselines and our MoSS in Figure 4. Our findings are as follows. **Firstly**, our method demonstrates better helpfulness and instruction-following capabilities. For example, in the first row in Figure 4, our method generates a more coherent visual story with more diverse content given the input. In the third row, our approach provides a more natural and reasonable next step for the user, i.e., *have a good vegan lunch*, while other baselines either jump to *dinner* (MiniGPT-5) or stick to *breakfast* (Emu2). **Secondly**, we observed that poor text quality is a common issue for many baselines. For instance, MiniGPT-5 often fails to generate explanatory text while GILL usually generates

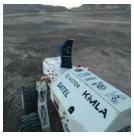
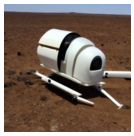
Input	MiniGPT-5	GILL	Emu2	Emu2+MoSS (Ours)
Instruction: Complete a travel story. **Context**: I visited the countryside for the weekend <IMG>.	**Output:** <IMG>	**Output:** the gardens <IMG>	**Output**: I was so excited to visit the country side for the weekend <IMG>. I was so excited to visit the country side for the weekend.	**Output**: I went to the countryside and saw many beautiful flowers <IMG>. I also saw many beautiful buildings <IMG>.
Instruction: Describe a team's success in University Rover Challenge. **Context**: During the challenge, the PCz Team placed the second <IMG>. 	**Output:** <IMG> 	**Output:** The team's rover is an autonomous vehicle that can across the surface of the moon. <IMG>	**Output**: Describe a team's success in University Rover Challenge <IMG>.	**Output**: 16 teams from 10 countries competed in the University Rover Challenge. <IMG>
Instruction: Tell me what should do next to plan vegan healthy meals. **Context**: I've done: I had oatmeal with fruit <IMG> and a tofu scramble for breakfast in the morning <IMG>.	**Output:** After you eat breakfast make sure to have enough food for dinner <IMG>.	**Output:** the tofu scramble <IMG>	**Output**: Eat vegan cereal with a non-dairy milk for a quick breakfast <IMG>.	**Output**: Have a good vegan meal for lunch <IMG>.
Instruction: Help me modify this picture <IMG> and change the bus into a train.	**Output:** <IMG>	**Output:** a bus on the road <IMG>	**Output:** <IMG>	**Output**: The train is on its track <IMG>.

Figure 4: **Qualitative results** of MoSS based on Emu2 and open-source baselines. The <IMG> tokens denote the images' positions in the interleaved sequences.

a short caption, e.g., *the gardens* in the first row and *the tofu scramble* in the third row, instead of generating useful content to solve the task. **Thirdly**, neither GILL nor MiniGPT-5 can preserve the visual appearance of the entities and scenes in the input images. Our approach, on the contrary, faithfully retains most visual characteristics, leading to significantly better Image Coherence. **Finally**, as shown in the fourth row in Figure 4, for tasks such as image editing or image composition, although MiniGPT-5 and GILL can sometimes generate images with better perceptual quality, the image contents are often irrelevant to the task, ignoring input instructions and context. In contrast, our method strives to adhere to instructions and can better condition its generation on the provided image. Due to the complexity of the task, our model may produce images with lower perceptual quality and noticeable distortions. However, when taking Helpfulness into account, the images generated by our model can be considered as the better ones compared with the baselines. We present additional qualitative results of Chameleon + MoSS in Figure 8 in Appendix C.4.

7 Discussions

Comparison between MoSS and other PEFT Methods To directly validate the performance improvement brought by our proposed MoSS, we fine-tuned Emu2 using (1) traditional linear LoRA (Hu et al., 2021) and (2) Mixture-of-Expert (MoE) LoRA (Shen et al., 2024), with the results presented in Table 2. In traditional linear LoRA, text and images share the same low-rank adaptation parameters, while in MoE-LoRA, two different sets of linear LoRA are used for images and text respectively. The routing strategy in MoE-LoRA is based on the output modality of each

Table 2: **Comparison between MoSS and existing PEFT methods**, i.e., traditional linear LoRA, and Mixture-of-Expert (MoE) LoRA. Mixture-of-Expert LoRA uses two different sets of linear LoRA for images and text, respectively. The rank number is set to 256 for all methods in this table.

Model	Text Quality	Image Quality	Image Coherence	TIC	**Helpfulness**
Emu2	1.33	2.29	1.71	1.22	1.87
+ LoRA	1.77	2.38	1.99	2.04	1.64
+ MoE-LoRA	1.98	3.28	2.66	2.62	2.01
+ MoSS (Ours)	**2.61**	**3.62**	**3.41**	**3.54**	**2.71**

hidden state, i.e., whether the hidden state is used to generate text or image. From Table 2, we effectively verify the benefits of using separate parameters for image and text. MoSS significantly outperforms the MoE-LoRA across all aspects, especially the image-related aspects such as Image Coherence and Text-Image Coherence (TIC). The conclusions from the results are two-fold. First, it shows that introducing modality-specialized architecture and parameters can effectively improve interleaved text-and-image generation. Second, it verifies that convolutional LoRA can improve image generation by better modeling the local priors of images. Additionally, we compare the computational cost of using MoSS and LoRA in Appendix C.3.

Effect of Rank Number To investigate how the number of rank r can affect the performance, we show the performance averaged on 5 aspects on InterleavedBench with the rank number equals $(32, 64, 128, 256)$ comparing LoRA, MoE-LoRA, and our MoSS in Figure 5. Our approach consistently outperforms LoRA and MoE-LoRA across all rank numbers, and as the rank number increases, the gap between MoSS and previous methods consistently grows larger. This proves the effectiveness and generalizability of MoSS across different rank sizes. Based on this experiment, we set the rank number of our approach to 256 by default. We include more results on the effect of rank in LoRA in Table 13, 14, and 15 in Appendix C.6.

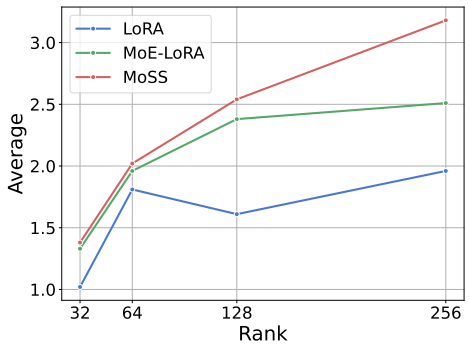


Figure 5: Performance averaged on 5 aspects with different rank numbers.

Quality Assessment of LeafInstruct To verify our LeafInstruct dataset is of high quality, we conduct a rigorous human evaluation using the multi-aspect evaluation criteria in InterleavedEval (Liu et al., 2024). Specifically, we use a scale of 0 to 3 in the evaluation, where 0 is the lowest score while 3 is the highest. We randomly sampled 200 instances from LeafInstruct and asked two human annotators with expertise in NLP and multimodal research to rate each instance from 5 aspects. We report the averaged scores from two annotators in Table 3. We show that the sampled instances consistently achieved almost full scores across all 5 aspects, which effectively demonstrated that our curated dataset is of high quality.

Table 3: **Human evalution** of randomly sampled instances from LeafInstruct. Note that the scale is from 0 to 3 (**Score 3 is the best**), which is different from the scale used in Table 1 and Table 2.

	Text Quality	Perceptual Quality	Image Coherence	TIC	Helpfulness
Score	2.89	2.96	2.77	2.87	2.71

8 Conclusion

We propose Modality-Specialized Synergizers (MoSS), a novel modality-specialized adaptation framework tailored for VLGs. MoSS dedicates a set of linear LoRA for processing text and a set of Convolutional LoRA for images, allowing each modality to have its own optimal adaptation design. Besides, we propose the first interleaved instruction tuning dataset LeafInstruct and verify the dataset quality via rigorous human evaluation. Extensive experiments on InterleavedBench showcase that our proposed method and dataset are highly effective, establishing the new state-of-the-art among open-sourced VLGs in interleaved text-and-image generation.

REFERENCES

Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, and Luke Zettlemoyer. CM3: A causal masked multimodal model of the internet. *CoRR*, abs/2201.07520, 2022. URL https://arxiv.org/abs/2201.07520.

Hassan Akbari, Dan Kondratyuk, Yin Cui, Rachel Hornung, Huisheng Wang, and Hartwig Adam. Alternating gradient descent and mixture-of-experts for integrated multimodal perception. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=uTlKUAm68H.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html.

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 18187–18197. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01767. URL https://doi.org/10.1109/CVPR52688.2022.01767.

James Betker, Gabriel Goh, Li Jing, † TimBrooks, Jianfeng Wang, Linjie Li, † LongOuyang, † JuntangZhuang, † JoyceLee, † YufeiGuo, † WesamManassra, † PrafullaDhariwal, † CaseyChu, † YunxinJiao, and Aditya Ramesh. Improving image generation with better captions. URL https://api.semanticscholar.org/CorpusID:264403242.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 18392–18402. IEEE, 2023a. doi: 10.1109/CVPR52729.2023.01764. URL https://doi.org/10.1109/CVPR52729.2023.01764.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023b.

Andrea Burns, Krishna Srinivasan, Joshua Ainslie, Geoff Brown, Bryan A. Plummer, Kate Saenko, Jianmo Ni, and Mandy Guo. Wikiweb2m: A page-level multimodal wikipedia dataset. *CoRR*, abs/2305.05432, 2023. doi: 10.48550/ARXIV.2305.05432. URL https://doi.org/10.48550/arXiv.2305.05432.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *CoRR*, abs/2306.15195, 2023a. doi: 10.48550/ARXIV.2306.15195. URL https://doi.org/10.48550/arXiv.2306.15195.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *CoRR*, abs/2311.12793, 2023b. doi: 10.48550/ARXIV.2311.12793. URL https://doi.org/10.48550/arXiv.2311.12793.

Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W. Cohen. Subject-driven text-to-image generation via apprenticeship learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023c. URL http://papers.nips.cc/paper_files/paper/2023/hash/6091bf1542b118287db4088bc16be8d9-Abstract-Conference.html.

Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023d. URL https://openreview.net/pdf?id=plKu2GByCNW.

Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. ANOLE: an open, autoregressive, native large multimodal models for interleaved image-text generation. *CoRR*, abs/2407.06135, 2024. doi: 10.48550/ARXIV.2407.06135. URL https://doi.org/10.48550/arXiv.2407.06135.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *CoRR*, abs/2305.06500, 2023. doi: 10.48550/ARXIV.2305.06500. URL https://doi.org/10.48550/arXiv.2305.06500.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi. DreamLLM: Synergistic multimodal comprehension and creation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=y01KGvd9Bw.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 12873–12883. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.01268. URL https://openaccess.thecvf.com/content/CVPR2021/html/Esser_Taming_Transformers_for_High-Resolution_Image_Synthesis_CVPR_2021_paper.html.

Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. MMDialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7348–7363, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.405. URL https://aclanthology.org/2023.acl-long.405.

Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XV*, volume 13675 of *Lecture Notes in Computer Science*, pp. 89–106. Springer, 2022. doi: 10.1007/978-3-031-19784-0_6. URL https://doi.org/10.1007/978-3-031-19784-0_6.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685, 2021. URL https://arxiv.org/abs/2106.09685.

Hexiang Hu, Kelvin C. K. Chan, Yu-Chuan Su, Wenhu Chen, Yandong Li, Kihyuk Sohn, Yang Zhao, Xue Ben, Boqing Gong, William W. Cohen, Ming-Wei Chang, and Xuhui Jia. Instruct-imagen: Image generation with multi-modal instruction. *CoRR*, abs/2401.01952, 2024. doi: 10.48550/ARXIV.2401.01952. URL https://doi.org/10.48550/arXiv.2401.01952.

Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. BLIVA: A simple multimodal LLM for better handling of text-rich visual questions. *CoRR*, abs/2308.09936, 2023. doi: 10.48550/ARXIV.2308.09936. URL https://doi.org/10.48550/arXiv.2308.09936.

Ting-Hao (Kenneth) Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross B. Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. Visual storytelling. In Kevin Knight, Ani Nenkova, and Owen Rambow (eds.), *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pp. 1233–1239. The Association for Computational Linguistics, 2016. doi: 10.18653/V1/N16-1147. URL https://doi.org/10.18653/v1/n16-1147.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge J. Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXIII*, volume 13693 of *Lecture Notes in Computer Science*, pp. 709–727. Springer, 2022. doi: 10.1007/978-3-031-19827-4_41. URL https://doi.org/10.1007/978-3-031-19827-4_41.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. MANTIS: interleaved multi-image instruction tuning. *CoRR*, abs/2405.01483, 2024. doi: 10.48550/ARXIV.2405.01483. URL https://doi.org/10.48550/arXiv.2405.01483.

Shibo Jie and Zhi-Hong Deng. Convolutional bypasses are better vision transformer adapters. *CoRR*, abs/2207.07039, 2022. doi: 10.48550/ARXIV.2207.07039. URL https://doi.org/10.48550/arXiv.2207.07039.

Yang Jin, Kun Xu, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Quzhe Huang, Bin Chen, Chenyi Lei, An Liu, Chengru Song, Xiaoqiang Lei, Di Zhang, Wenwu Ou, Kun Gai, and Yadong Mu. Unified language-vision pretraining in LLM with dynamic discrete visual tokenization. *CoRR*, abs/2309.04669, 2023. doi: 10.48550/ARXIV.2309.04669. URL https://doi.org/10.48550/arXiv.2309.04669.

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in Neural Information Processing Systems*, 34:1022–1035, 2021.

Jing Yu Koh, Daniel Fried, and Russ Salakhutdinov. Generating images with multimodal language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/43a69d143273bd8215578bde887bb552-Abstract-Conference.html.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*, 2017.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. 2023.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. MIMIC-IT: multi-modal in-context instruction tuning. *CoRR*, abs/2306.05425, 2023a. doi: 10.48550/ARXIV.2306.05425. URL https://doi.org/10.48550/arXiv.2306.05425.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *CoRR*, abs/2407.07895, 2024a. doi: 10.48550/ARXIV.2407.07895. URL https://doi.org/10.48550/arXiv.2407.07895.

Huayang Li, Siheng Li, Deng Cai, Longyue Wang, Lemao Liu, Taro Watanabe, Yujiu Yang, and Shuming Shi. Textbind: Multi-turn interleaved multimodal instruction-following in the wild. *CoRR*, abs/2309.08637, 2023b. doi: 10.48550/ARXIV.2309.08637. URL https://doi.org/10.48550/arXiv.2309.08637.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. 202:19730–19742, 2023c. URL https://proceedings.mlr.press/v202/li23q.html.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, 2021.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *CoRR*, abs/2403.18814, 2024b. doi: 10.48550/ARXIV.2403.18814. URL https://doi.org/10.48550/arXiv.2403.18814.

Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/00bb4e415ef117f2dee2fc3b778d806d-Abstract-Conference.html.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *CoRR*, abs/2401.15947, 2024. doi: 10.48550/ARXIV.2401.15947. URL https://doi.org/10.48550/arXiv.2401.15947.

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*, pp. 740–755. Springer, 2014.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *CoRR*, abs/2306.14565, 2023a. doi: 10.48550/ARXIV.2306.14565. URL https://doi.org/10.48550/arXiv.2306.14565.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *CoRR*, abs/2205.05638, 2022. doi: 10.48550/ARXIV.2205.05638. URL https://doi.org/10.48550/arXiv.2205.05638.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *CoRR*, abs/2310.03744, 2023b. doi: 10.48550/ARXIV.2310.03744. URL https://doi.org/10.48550/arXiv.2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023c. URL https://doi.org/10.48550/arXiv.2304.08485.

Minqian Liu, Zhiyang Xu, Zihao Lin, Trevor Ashby, Joy Rimchala, Jiaxin Zhang, and Lifu Huang. Holistic evaluation for interleaved text-and-image generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *CoRR*, abs/2306.09093, 2023. doi: 10.48550/ARXIV.2306.09093. URL https://doi.org/10.48550/arXiv.2306.09093.

OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. Accessed: 2024-05-26.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. *CoRR*, abs/2307.01952, 2023. doi: 10.48550/ARXIV.2307.01952. URL https://doi.org/10.48550/arXiv.2307.01952.

Jingyuan Qi, Minqian Liu, Ying Shen, Zhiyang Xu, and Lifu Huang. MULTISCRIPT: multimodal script learning for supporting open domain everyday tasks. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pp. 18888–18896. AAAI Press, 2024. doi: 10.1609/AAAI.V38I17.29854. URL https://doi.org/10.1609/aaai.v38i17.29854.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, and et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530, 2024. doi: 10.48550/ARXIV.2403.05530. URL https://doi.org/10.48550/arXiv.2403.05530.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 10674–10685. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01042. URL https://doi.org/10.1109/CVPR52688.2022.01042.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*, 2023.

Ying Shen, Zhiyang Xu, Qifan Wang, Yu Cheng, Wenpeng Yin, and Lifu Huang. Multimodal instruction tuning with conditional mixture of lora. *CoRR*, abs/2402.15896, 2024. doi: 10.48550/ARXIV.2402.15896. URL https://doi.org/10.48550/arXiv.2402.15896.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: improved training techniques for CLIP at scale. *CoRR*, abs/2303.15389, 2023a. doi: 10.48550/ARXIV.2303.15389. URL https://doi.org/10.48550/arXiv.2303.15389.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *CoRR*, abs/2307.05222, 2023b. doi: 10.48550/ARXIV.2307.05222. URL https://doi.org/10.48550/arXiv.2307.05222.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14398–14409, June 2024.

Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. Codi-2: In-context, interleaved, and interactive any-to-any generation. *CoRR*, abs/2311.18775, 2023. doi: 10.48550/ARXIV.2311.18775. URL https://doi.org/10.48550/arXiv.2311.18775.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2024.

Changyao Tian, Xizhou Zhu, Yuwen Xiong, Weiyun Wang, Zhe Chen, Wenhai Wang, Yuntao Chen, Lewei Lu, Tong Lu, Jie Zhou, Hongsheng Li, Yu Qiao, and Jifeng Dai. Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer. *CoRR*, abs/2401.10208, 2024. doi: 10.48550/ARXIV.2401.10208. URL https://doi.org/10.48550/arXiv.2401.10208.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023. doi: 10.48550/ARXIV.2302.13971. URL https://doi.org/10.48550/arXiv.2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to believe: Prompting GPT-4V for better visual instruction tuning. *CoRR*, abs/2311.07574, 2023. doi: 10.48550/ARXIV.2311.07574. URL https://doi.org/10.48550/arXiv.2311.07574.

Yaqing Wang, Sahaj Agarwal, Subhabrata Mukherjee, Xiaodong Liu, Jing Gao, Ahmed Hassan Awadallah, and Jianfeng Gao. Adamix: Mixture-of-adaptations for parameter-efficient model tuning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 5744–5760. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.388. URL https://doi.org/10.18653/v1/2022.emnlp-main.388.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal LLM. *CoRR*, abs/2309.05519, 2023. doi: 10.48550/ARXIV.2309.05519. URL https://doi.org/10.48550/arXiv.2309.05519.

Zhiyang Xu, Ying Shen, and Lifu Huang. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 11445–11465. Association for Computational Linguistics, 2023a. doi: 10.18653/V1/2023.ACL-LONG.641. URL https://doi.org/10.18653/v1/2023.acl-long.641.

Zhiyang Xu, Ying Shen, and Lifu Huang. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 11445–11465. Association for Computational Linguistics, 2023b. doi: 10.18653/V1/2023.ACL-LONG.641. URL https://doi.org/10.18653/v1/2023.acl-long.641.

Zhiyang Xu, Chao Feng, Rulin Shao, Trevor Ashby, Ying Shen, Di Jin, Yu Cheng, Qifan Wang, and Lifu Huang. Vision-flan: Scaling human-labeled tasks in visual instruction tuning, 2024.

Yue Yang, Artemis Panagopoulou, Qing Lyu, Li Zhang, Mark Yatskar, and Chris Callison-Burch. Visual goal-step inference using wikiHow. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2167–2179, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.165. URL https://aclanthology.org/2021.emnlp-main.165.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-Tau Yih. Retrieval-augmented multimodal language modeling. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 39755–39769. PMLR, 2023. URL https://proceedings.mlr.press/v202/yasunaga23a.html.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality. *CoRR*, abs/2304.14178, 2023. doi: 10.48550/ARXIV.2304.14178. URL https://doi.org/10.48550/arXiv.2304.14178.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13040–13051, June 2024.

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, Jing Shao, and Wanli Ouyang. LAMM: language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *CoRR*, abs/2306.06687, 2023. doi: 10.48550/ARXIV.2306.06687. URL https://doi.org/10.48550/arXiv.2306.06687.

Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, Candace Ross, Adam Polyak, Russell Howes, Vasu Sharma, Puxin Xu, Hovhannes Tamoyan, Oron Ashual, Uriel Singer, Shang-Wen Li, Susan Zhang, Richard James, Gargi Ghosh, Yaniv Taigman, Maryam Fazel-Zarandi, Asli Celikyilmaz, Luke Zettlemoyer, and Armen Aghajanyan. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *CoRR*, abs/2309.02591, 2023. doi: 10.48550/ARXIV.2309.02591. URL https://doi.org/10.48550/arXiv.2309.02591.

Ted Zadouri, Ahmet Üstün, Arash Ahmadian, Beyza Ermis, Acyr Locatelli, and Sara Hooker. Pushing mixture of experts to the limit: Extremely parameter efficient moe for instruction tuning. *CoRR*, abs/2309.05444, 2023. doi: 10.48550/ARXIV.2309.05444. URL https://doi.org/10.48550/arXiv.2309.05444.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1–9, 2022.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023a. URL http://papers.nips.cc/paper_files/paper/2023/hash/64008fa30cba9b4d1ab1bd3bd3d57d61-Abstract-Datasets_and_Benchmarks.html.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *CoRR*, abs/2306.17107, 2023b. doi: 10.48550/ARXIV.2306.17107. URL https://doi.org/10.48550/arXiv.2306.17107.

Kaizhi Zheng, Xuehai He, and Xin Eric Wang. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *CoRR*, abs/2310.02239, 2023a. doi: 10.48550/ARXIV.2310.02239. URL https://doi.org/10.48550/arXiv.2310.02239.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023b.

Zihan Zhong, Zhiqiang Tang, Tong He, Haoyang Fang, and Chun Yuan. Convolution meets lora: Parameter efficient finetuning for segment anything model. *CoRR*, abs/2401.17868, 2024. doi: 10.48550/ARXIV.2401.17868. URL https://doi.org/10.48550/arXiv.2401.17868.

Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In Sheila A. McIlraith and Kilian Q. Weinberger (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 7590–7598. AAAI Press, 2018. doi: 10.1609/AAAI.V32I1.12342. URL https://doi.org/10.1609/aaai.v32i1.12342.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592, 2023a. doi: 10.48550/ARXIV.2304.10592. URL https://doi.org/10.48550/arXiv.2304.10592.

Jinguo Zhu, Xiaohan Ding, Yixiao Ge, Yuying Ge, Sijie Zhao, Hengshuang Zhao, Xiaohua Wang, and Ying Shan. VL-GPT: A generative pre-trained transformer for vision and language understanding and generation. *CoRR*, abs/2312.09251, 2023b. doi: 10.48550/ARXIV.2312.09251. URL https://doi.org/10.48550/arXiv.2312.09251.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023c. URL https://openreview.net/forum?id=tOd8rSjcWz.

A MORE DETAILS OF MOSS

A.1 IMPLEMENTATION DETAILS

We leverage the Emu2 model (Sun et al., 2024), consisting of the EVA-02-CLIP-E-plus (Sun et al., 2023a) as the image encoder, the LLaMA-33B (Touvron et al., 2023), and the SDXL (Podell et al., 2023) as the image decoder, as our base model. The EVA-02-CLIP-E-plus and the LLaMA-33B is connected by a linear project-up layer and the LLaMA-33B and the SDXL is connected by a linear project-down layer. All the variants of LoRA in Section 7, including our MOSS are trained with LEAFINSTRUCT for one epoch on $8 \times$ A100 GPUs with learning rate $2e^{-5}$, batch size 1 per GPU, and a gradient accumulation step of 16. All the LoRA have a rank of 256, dropout rate of 0.05, and the LoRA α in Section 4 is set to 2×128. The kernel size of MOSS is 2×2, the stride is set to 1. During training, all parameters of the Emu2 model are kept frozen and only the LoRA parameters are updated.

B MORE DETAILS OF LEAFINSTRUCT

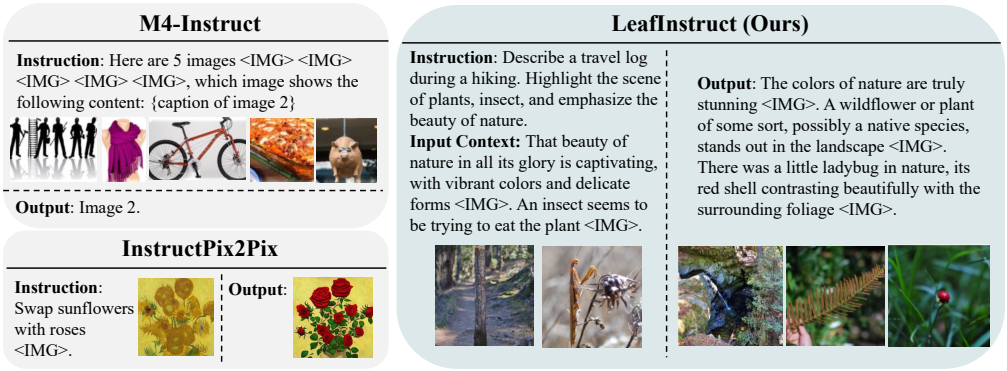


Figure 6: Comparison between existing benchmarks and our LEAFINSTRUCT. In existing datasets such as InstructPix2Pix (Brooks et al., 2023b) and Mantis-Instruct (Li et al., 2024a), the outputs are in single modality, either text or image. On the contrary, the inputs and outputs of our LEAFINSTRUCT cover multiple modalities.

Table 4: Comparison between our LEAFINSTRUCT and existing instruction tuning datasets.

Dataset Name	Input Text	Input Images	Output Text	Output Images	Publicly Available
LLaVA (Liu et al., 2023c)	Yes	Single	Single	No	Yes
MultiInstruct (Xu et al., 2023b)	Yes	Single	Single	No	Yes
Vision-Flan (Xu et al., 2024)	Yes	Single	Single	No	Yes
InstructPix2Pix (Brooks et al., 2023a)	Yes	Single	No	Single	Yes
MagicBrush (Zhang et al., 2023a)	Yes	Single	No	Single	Yes
SuTI (Chen et al., 2023c)	Yes	Multiple	No	Single	No
Instruct-Imagen (Hu et al., 2024)	Yes	Multiple	No	Single	No
Mantis-Instruct (Jiang et al., 2024)	Yes	Multiple	Yes	No	Yes
LEAFINSTRUCT (Ours)	Yes	Multiple	Yes	Multiple	Yes

B.1 MORE DETAILS IN DATASET CONSTRUCTION

We elaborate on the details of our dataset construction pipeline as follows. **Firstly**, we filter the samples based on the text length, number of images, and the coherence between text and images (measured by CLIPScore (Hessel et al., 2021)). We only keep the instances with 3 to 6 images in total. We also discard the instances with more than 12 sentences to ensure a balanced ratio between the number of textual sentences and images. **Secondly**, we leverage a state-of-the-art open-sourced LLM (i.e., `Llama-8B-Instruct`) as a text filter to discard the instances with poor text quality. **Thirdly**, we remove the instances with duplicate or perceptually highly similar images to ensure

the diversity of the images. **Finally**, we also apply Llama3 to annotate the task instruction for each instance based on the text content and rewrite the text if it's too verbose to prevent the context length from being too long.

Details of Text Quality Filter We use `Llama-8B-Instruct` model to rate the text quality of an instance with the following prompt: *"Imagine you are an expert data annotator. You are given a text material and you need to evaluate its quality in terms of whether it is coherent, fluent, easy to understand, and helpful to humans. Please be critical and rate the quality as good only when the text quality is good in all four aspects. Output 1 if you think the material is good after you consider all four aspects. Output 0 if you think the material is not good enough. Here is the text material to be evaluated: {TEXT} Only output 0 or 1 and do not output anything else. Your evaluation is:"* We discard the instances if the output from Llama is 0.

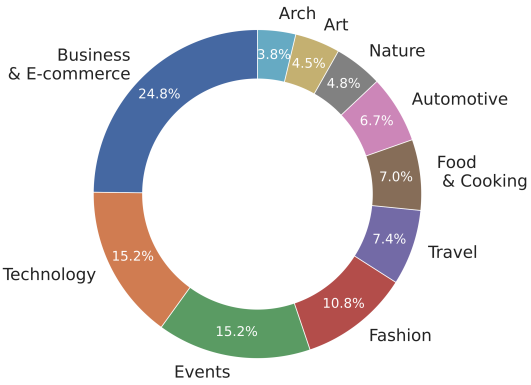


Figure 7: Domain distribution in LeafInstruct.

Details of Image Filter We empirically found that if the images are too identical in the training instances, the trained models tend to find a shortcut to simply copy the image during generation. To this end, we design a filter to discard the instances with duplicate images to improve data quality. Specifically, we leverage the LPIPS score (Zhang et al., 2018) that measures the perceptual similarity between the images. Specifically, for each instance, we enumerate each pair of images and compute their LPIPS score. If there is one pair with a score higher than 0.6, we discard the instance. We determine the threshold of 0.6 by empirical trial.

Details of Instruction Annotation We also adopt `Llama-8B-Instruct` to annotate the task instruction for each instance. We devise instructions to prompt the Llama3 model to rewrite the original text material in the pretraining dataset MMC4 into instruction-tuning instances. The input context length is 2048 and the output context length is 1024. We set the temperature as 1 to encourage the diversity of instructions. We use the following prompt: *"Imagine you are an expert instruction annotator. You are given a material. You need to read its content and output a brief task instruction with one sentence such that another person can recover the given the material given the instruction. The instruction you predict should be specifically tailored for creative interleaved content generation that consists of both text and images. Now you need to annotate a concise, accurate instruction for the following instance. Please only predict the instruction and do not output anything else. Please design the instruction for the multi-modal generation task interleaved with both text and images. Text: {TEXT} Instruction:"*.

C More Experiment Results

C.1 More Details on Evaluation Benchmarks

InterleavedBench has two splits: a context-based split in which the input of each instance is equipped with interleaved text and images; and a context-free split with text-only inputs. The context-based split contains 465 instances and the text-only split contains 350 instances. We only use the context-based split as the testing set since we mainly focus on tasks with interleaved inputs and outputs.

C.2 More Results on Established Benchmarks

Interleaved Generation and Image Editing Although InterleavedBench is a new evaluation benchmark, it also consists of testing instances from 3 well-established benchmarks including (1) visual story completion from VIST, (2) MagicBrush, and (3) multi-concept image composition from CustomDiffusion. Below, we directly report the performance of MoSS and baselines on these three

well-established benchmarks in Table 6, 7, 8, respectively. We also report the performance of MagicBrush using established metrics in Table 5.

Table 5: **Results of image editing on the full test set of MagicBrush**. We show the performance of open-source VLGs (Top), and the VLGs trained with our proposed method (Bottom), respectively.

Model	CLIP-I	DINO	CLIP-T	AVG
MiniGPT-5	72.04	41.66	25.45	46.38
GILL	72.95	43.32	24.81	47.03
SEED-X-Edit (Ge et al., 2024)	85.56	68.74	**27.28**	60.53
Emu2 + MoSS (Ours)	**85.88**	**74.92**	25.98	**62.26**

Multimodal Understanding and Text-to-Image Generation To show that our MoSS framework can also excel on tasks requiring single modality outputs i.e., the output only contains text or an image, we evaluate its performance on widely adopted image understanding benchmarks including MMBench, MME, MMMU, Pope, and MM-Vet, and text-to-image generation benchmarks including MSCOCO 30K (Lin et al., 2014), and GenEval (Ghosh et al., 2024). For MSCOCO-30K, following the previous evaluation protocol (Sun et al., 2024), we randomly sample 30,000 captions from the validation set of MSCOCO and generate 30,000 images. We report the FID between the 30,000 generated images and real images from the validation set of MSCOCO (Note for FID, the lower the better). For other benchmarks, we adopt their official implementation of the evaluation. Since LeafInstruct mainly targets tasks with interleaved outputs, we augmented it with 500,000 instances from Vision-Flan (Xu et al., 2024), a popular visual-instruction tuning dataset targeting image understanding, and 500,000 instances from LAION-COCO[1], a standard training dataset for text-to-image generation. We finetune Emu2 with LoRA, MoE-LoRA, and MoSS on the mixed dataset. We report their performance in Table 9.

Table 6: Performance of our MoSS, traditional linear LoRA, and Mixture-of-Expert (MoE) LoRA using Emu2 as the backbone model on the VIST subset in InterleavedBench.

PEFT	Text Quality	Image Quality	Image Coherence	TIC	**Helpfulness**
LoRA	0.3	0.52	0.46	0.68	0.43
MoE-LoRA	**1.76**	2.19	2.52	2.91	1.92
MoSS	1.73	**2.87**	**2.54**	**3.30**	**2.26**

Table 7: Performance of our MoSS, traditional linear LoRA, and Mixture-of-Expert (MoE) LoRA using Emu2 as the backbone model on the MagicBrush subset in InterleavedBench.

PEFT	Text Quality	Image Quality	Image Coherence	TIC	**Helpfulness**
LoRA	N/A	2.33	1.48	N/A	0.93
MoE-LoRA	N/A	2.85	1.54	N/A	1.08
MoSS	N/A	**3.43**	**2.03**	N/A	**1.27**

Table 8: Performance of our MoSS, traditional linear LoRA, and Mixture-of-Expert (MoE) LoRA using Emu2 as the backbone model on the CustomDiffusion subset in InterleavedBench.

PEFT	Text Quality	Image Quality	Image Coherence	TIC	**Helpfulness**
LoRA	N/A	2.95	1.73	N/A	1.56
MoE-LoRA	N/A	**3.36**	1.67	N/A	1.39
MoSS	N/A	3.24	**2.04**	N/A	**1.83**

[1] https://huggingface.co/datasets/laion/laion-coco

Table 9: Results on widely adopted multimodal understanding and text-to-image generation benchmarks. Note that the FID metric on MSCOCO is the lower the better.

Model	MMBench	MME	MMMU	Pope	MM-Vet	MSCOCO-30K FID (↓)	GenEval
Chameleon	32.7	604.5	**38.8**	59.8	9.7	26.7	**39.0**
Emu2+LoRA	54.1	1148.0	33.7	87.3	31.3	23.4	26.8
Emu2+MoE-LoRA	54.6	1170.3	34.1	88.1	31.9	22.7	28.1
Emu2+MoSS(Ours)	**56.0**	**1278.4**	35.8	**87.6**	**34.1**	**18.2**	28.9

From Table 9, our MoSS outperforms previous LoRA and MoE-LoRA on most of the multimodal understanding benchmarks by a notable margin, which demonstrates that MoSS can be well generalized to diverse multimodal comprehension tasks. For text-to-image generation, our MoSS achieves better performance on both benchmarks, showing the effectiveness and generalizability of our approach. Notably, our MoSS achieves significantly better FID on MSCOCO-30K, which validates that our ConvLoRA can effectively improve the quality of generated images.

C.3 Additional Analysis

Comparison between previous and our ConvLoRA To show the benefits of our modified ConvLoRA architecture compared to the ConvLoRA proposed in Zhong et al. (2024) denoted as SAM-ConvLoRA, we replace the ConvLoRA in MoSS with SAM-ConvLoRA. Specifically, we set the rank of project-down and project-up matrices in SAM-ConvLoRA to 256 which is the same number of ranks in our proposed MoSS-ConvLoRA, and adopt the multi-scale convolution kernels to the size of 2x2 and 4x4. As shown in Table 10, our MoSS-ConvLoRA consistently outperforms the previous SAM-ConvLoRA on all other evaluation aspects, which demonstrates the superiority of our proposed ConvLoRA architecture. Particularly, our MoSS-ConvLoRA achieves notably better visual qualities, including perceptual quality and image coherence, thanks to our novel design that our convolution operation is applied to the full-rank original image features instead of low-rank image features as in SAM-ConvLoRA.

Table 10: Comparison of two types of ConvLoRA.

Model	Text Quality	Image Quality	Image Coherence	TIC	**Helpfulness**
MoSS w/ SAM-ConvLoRA	2.50	3.33	3.17	3.50	2.41
MoSS w/ MoSS-ConvLoRA (Ours)	**2.61**	**3.62**	**3.41**	**3.54**	**2.71**

Human Evaluation on Text Quality of Chameleon and Chameleon + MoSS We noticed that adding MoSS in Chameleon can cause a slight performance drop in text quality. This is because the original Chameleon usually generates long and verbose text responses but with no image output. On the contrary, as the text responses in our LeafInstruct dataset are more concise to allow for including more images, after interleaved instruction tuning, our model learns to generate more concise text responses. Specifically, the average generated word length of the original Chameleon is 653, whereas that of Chameleon-MoSS is 166. The verbose responses from the original Chameleon are preferred by the LLM judge due to their verbosity bias (Zheng et al., 2023b; Saito et al., 2023), leading to a slight drop in the text quality of Chameleon-MoSS in LLM-based evaluation.

To better support this analysis, we further conduct a human evaluation of the text quality of the two models by randomly sampling 100 instances from InterleavedBench. We ask a human annotator to select the preferred text responses given the system outputs from two models. We report the Win-Tie-Loss results in Table 11. Win means our Chameleon-MoSS is better than the original Chameleon, Tie means the quality of two responses is equally good, and Loss means the original Chameleon is better.

From Table 11, the text quality of our Chameleon-MoSS is actually better than the original Chameleon. One issue we frequently observed in the original Chameleon is the text responses are overly verbose and sometimes even severely repetitive. In our evaluation protocol of text quality adopted from InterleavedEval (Liu et al., 2024), such verbosity and repetitiveness are not penal-

Table 11: Human evaluation results on text quality of the original Chameleon and our Chameleon-MoSS. "Win" indicates our Chameleon-MoSS's responses are preferred by humans.

Wins	Ties	Losses
28	54	18

ized, making the automatic evaluation results heavily biased towards the longer responses from the original Chameleon.

Performance of full-finetuning We compare the performance of full-parameter fine-tuning using LEAFINSTRUCT in Table 12. The first row represents the results of fully fine-tuning Emu2 on our proposed LEAFINSTRUCT dataset. The second row shows the results of parameter-efficient fine-tuning Emu2 using our proposed MOSS framework. As observed, while full fine-tuning allows Emu2 to achieve better performance on text generation, the model demonstrates inferior performance on image generation due to its lack of inductive bias. In contrast, tuning with MOSS, which incorporates ConvLoRA, significantly improves image generation performance, even though the number of trained parameters in full fine-tuning is substantially larger than that of MOSS. These results clearly highlight the advantages of integrating ConvLoRA into the transformer architecture for processing visual information.

Table 12: Comparison between full finetuning and parameter-efficient tuning with MOSS based on Emu2.

Model	Text Quality	Perceptual Quality	Image Coherence	TIC	**Helpfulness**
Full Finetuning	3.20	3.21	2.98	3.60	3.23
MOSS (Ours)	2.61	3.62	3.41	3.54	2.71

Computation cost of ConvLoRA We compared the computational cost of using linear LoRA and our ConvLoRA, respectively. We compute the inference time for generating 1,000 images for each model. The total inference times for linear LoRA and ConvLoRA are 4,380 seconds and 5,910 seconds, respectively. The difference between the two models is around 1.5 seconds per image, indicating the computational cost increased by ConvLoRA is not significant.

C.4 ADDITIONAL QUALITATIVE RESULTS OF CHAMELEON

We show the additional qualitative comparison between the original Chameleon and our Chameleon+MOSS in Figure 8. From the results, most of the examples of the original Chameleon only contain text without any image, which indicates the original Chameleon often fails to follow the instructions to generate desired mixed-modal content. Our Chameleon+MOSS, on the contrary, achieves much better performance as it is able to generate high-quality and coherent text and images. These results further validate the effectiveness of our approach and its strong generalizability and robustness when applying it to different VLGs, including Emu2 based on continuous image tokens and Chameleon based on discrete image tokens.

C.5 PER-TASK PERFORMANCE ON EACH EVALUATION ASPECT

We report the performance on all aspects for each task in InterleavedBench in Figure 9.

C.6 MORE RESULTS ON THE EFFECT OF RANKS

We conducted experiments to show the performance of two LoRA baselines and Lateralization LoRA on InterleavedBench with different ranks (64, 128, 256) in Table 13, 14, 15, respectively.

Input	Chameleon	Chameleon+MoSS (Ours)
Instruction: Complete a travel story. **Context**: I visited the countryside for the weekend <IMG>.	**Output:** I've always wanted to visit the coast and see the ocean. I've heard great things about the coast of Maine, so I would like to visit there. I would like to go hiking, see some lighthouses, and enjoy the fresh seafood…	**Output:** There are many popular activities to do in the countryside, including hiking, biking, camping, and visiting local attractions like museums and castles., including cycling the ridgeway Trail <IMG>, camping in the yorkshire dales <IMG>, and taking a boat tour of the Lake District <IMG>.
Instruction: Describe a team's success in University Rover Challenge. **Context**: During the challenge, the PCz Team placed the second <IMG>.	**Output:** The event challenges teams to design, build, and operate a rover that can navigate across a challenging Martian-like terrain, while performing tasks such as reconnaissance, exploration, and scientific research….	**Output:** This competition encourages universities to design, build, and test rovers that can operate on the Martian surface, replicating real-world scenarios that engineers and scientists may encounter during future Mars missions. **<IMG>**
Instruction: Tell me what should do next to plan vegan healthy meals. **Context**: I've done: I had oatmeal with fruit <IMG> and a tofu scramble for breakfast in the morning <IMG>.	**Output:** Have a healthy snack for breakfast, such as apple or an orange, or some almonds or cashews.	**Output:** Here are a few more ideas: Smoothie bowls made with fruit, nuts, and seeds <IMG>. Avocado toast with scrambled tofu <IMG>. Overnight oats made with rolled oats, milk, and fruit Chia seed pudding with fruit and nuts <IMG>. Muffins or cakes made with whole grains and plant-based milk <IMG>.
Instruction: Help me modify this picture <IMG> and change the bus into a train.	**Outputs:** Here's an edited version of the picture with the bus replaced by a train: <IMG>	**Outputs:** Of course! I'll make sure to do my best to assist you. <IMG>

Figure 8: **Qualitative results** of MoSS based on Chameleon and the original Chameleon. The <IMG> tokens denote the images' positions in the interleaved sequences.

Table 13: Performance of our MoSS, traditional linear LoRA, and Mixture-of-Expert (MoE) LoRA using Emu2 as the backbone model with rank $r = 64$.

PEFT	Text Quality	Image Quality	Image Coherence	TIC	**Helpfulness**
LoRA	1.7	1.6	1.81	1.90	1.39
MoE-LoRA	**1.86**	2.17	2.17	**2.15**	**1.66**
MoSS	1.46	**2.34**	**2.20**	2.13	1.58

Table 14: Performance of our MoSS, traditional linear LoRA, and Mixture-of-Expert (MoE) LoRA using Emu2 as the backbone model with rank $r = 128$.

PEFT	Text Quality	Image Quality	Image Coherence	TIC	**Helpfulness**
LoRA	1.25	1.43	1.61	1.79	1.30
MoE-LoRA	1.94	2.22	2.42	2.54	1.90
MoSS	**1.95**	**2.41**	**2.64**	**2.81**	**2.05**

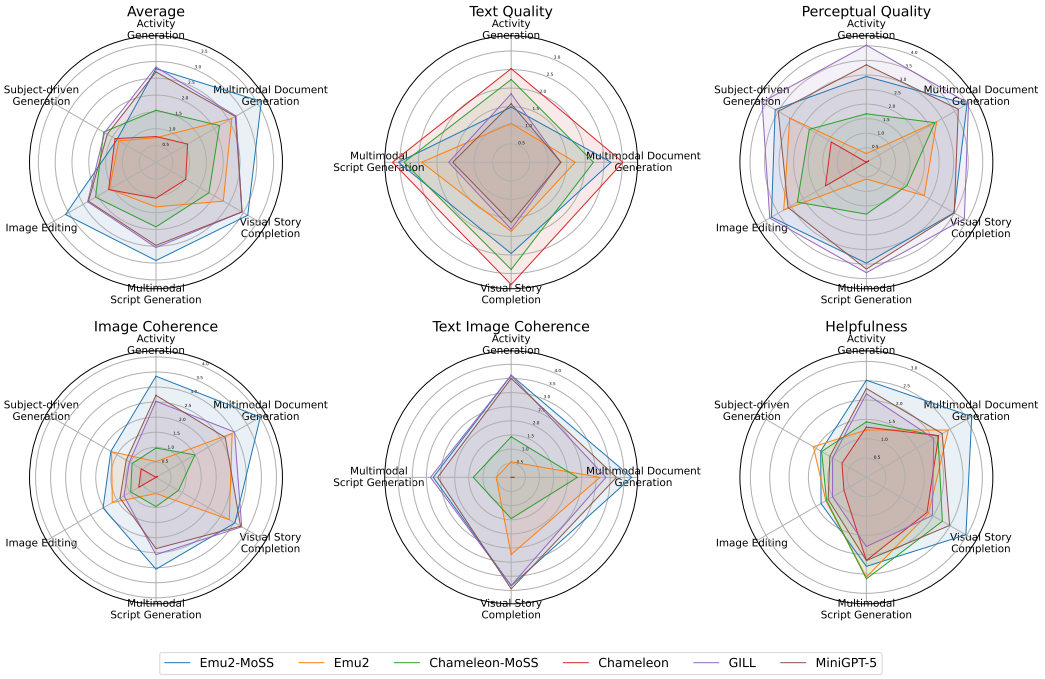


Figure 9: Per-task performance on each evaluation aspect on InterleavedBench.

Table 15: Performance of our MoSS, traditional linear LoRA, and Mixture-of-Expert (MoE) LoRA using Emu2 as the backbone model with rank $r = 256$.

PEFT	Text Quality	Image Quality	Image Coherence	TIC	**Helpfulness**
LoRA	1.77	2.38	1.99	2.04	1.64
MoE-LoRA	1.98	3.28	2.66	2.62	2.01
MoSS	**2.61**	**3.62**	**3.41**	**3.54**	**2.71**