

# MODALITY-SPECIALIZED SYNERGIZERS FOR INTERLEAVED VISION-LANGUAGE GENERALISTS

000  
001  
002  
003  
004  
005 **Anonymous authors**  
006 Paper under double-blind review  
007  
008  
009  
010

## ABSTRACT

011 Recent advancements in Vision-Language Models (VLMs) have led to the emer-  
012 gence of Vision-Language Generalists (VLGs) capable of understanding and gen-  
013 erating both text and images. However, seamlessly generating an arbitrary se-  
014 quence of text and images remains a challenging task for the current VLGs. One  
015 primary limitation lies in applying a unified architecture and the same set of pa-  
016 rameters to simultaneously model discrete text tokens and continuous image fea-  
017 tures. Recent works attempt to tackle this fundamental problem by introducing  
018 modality-aware expert models. However, they employ identical architectures to  
019 process both text and images, disregarding the intrinsic inductive biases in these  
020 two modalities. In this work, we introduce MODALITY-SPECIALIZED SYNER-  
021 GIZERS (MOSS), a novel design that efficiently optimizes existing unified archi-  
022 tectures of VLGs with modality-specialized adaptation layers, i.e., a Convolu-  
023 tional LoRA for modeling the local priors of image patches and a Linear LoRA  
024 for processing sequential text. This design enables more effective modeling of  
025 modality-specific features while maintaining the strong cross-modal integration  
026 gained from pretraining. In addition, to improve the instruction-following capa-  
027 bility on interleaved text-and-image generation, we introduce LEAFINSTRUCT,  
028 the first open-sourced interleaved instruction tuning dataset comprising 184,982  
029 high-quality instances on more than 10 diverse domains. Extensive experiments  
030 show that VLGs integrated with MOSS achieve state-of-the-art performance, sig-  
031 nificantly surpassing baseline VLGs in complex interleaved generation tasks. Fur-  
032 thermore, our method exhibits strong generalizability on different VLGs.  
033

## 1 INTRODUCTION

034 As multimodal learning research advances, there is a growing trend of building Vision-Language  
035 Generalists (VLGs) (Sun et al., 2023b; 2024; Koh et al., 2023; Aghajanyan et al., 2022; Li et al.,  
036 2023b; Dong et al., 2024; Team, 2024) that can comprehend and generate *interleaved* text and im-  
037 ages, where multiple text segments and images are presented in arbitrary sequences. Compared  
038 with previous Vision-Language Models (VLMs) (Alayrac et al., 2022; Li et al., 2023c; Liu et al.,  
039 2023c) that can only generate text and diffusion models (Ramesh et al., 2021; Rombach et al., 2022)  
040 that can only produce images, such VLGs enable a wider array of applications that require the si-  
041 multaneous generation of both images and text, such as script generation (Qi et al., 2024), visual  
042 storytelling (Huang et al., 2016), and many others.  
043

044 Despite these recent advancements, one notable issue of existing VLGs is that they often fail to pro-  
045 duce coherent and high-quality interleaved text and images. As shown in the top example in Figure 1,  
046 current state-of-the-art VLG, e.g., Emu2 (Sun et al., 2024), still suffers from poor text and image  
047 quality, including heavy repetition in text and unnatural distortions in the image. We attribute this  
048 issue to a fundamental challenge: *existing VLGs use the same architecture (i.e., transformer back-  
049 bone) with the same set of parameters to process both text and images, which may not be sufficient  
050 to model the distinct inductive biases in each modality given their intrinsic discrepancy.* For ex-  
051 ample, text follows a linear, left-to-right sequence, whereas images are inherently two-dimensional,  
052 composed of local priors in adjacent patches. Previous studies show that the transformer architec-  
053 ture (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;

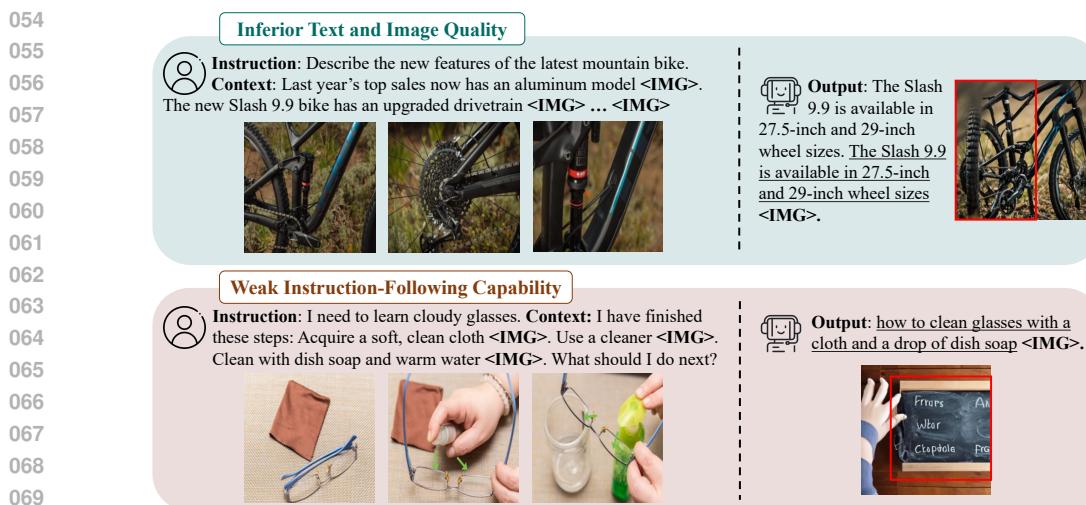


Figure 1: Failure cases of existing VLGs (Emu2 at the top and GILL at the bottom). The output text with inferior quality is highlighted with underline. The regions that impede output images’ quality are highlighted with red bounding boxes.

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),

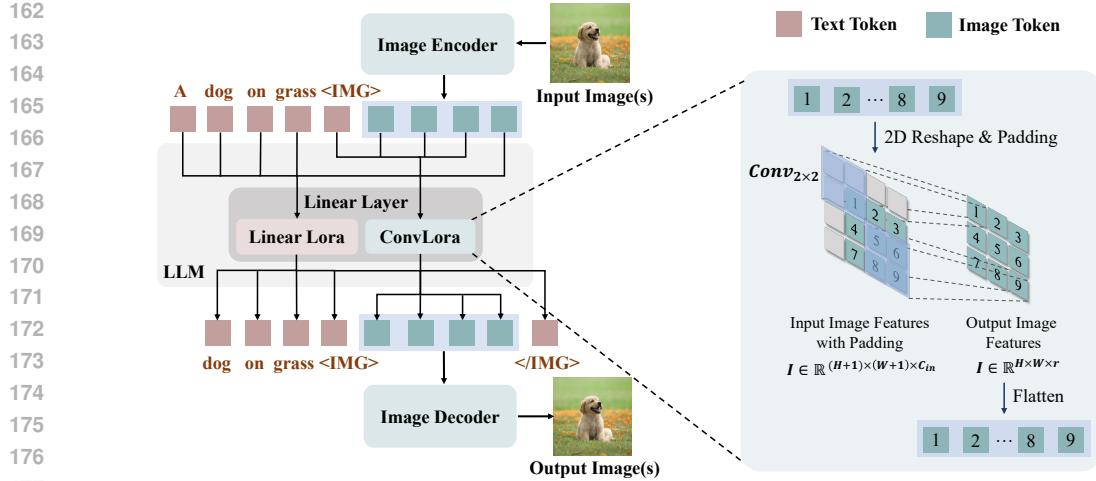


Figure 2: An autoregressive VLG with our proposed MOSS added to its linear layers. The linear LoRA on the left side is specialized to generate text tokens and the Convolutional LoRA on the right side is specialized to generate image patches. On the right handside, we show the details of convolutional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

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

where  $\theta$  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  $T : \mathbb{R}_{in}^d \rightarrow \mathbb{R}_{out}^d$  becomes:

$$T(\mathbf{h}) = \mathbf{h}(\mathbf{W} + \Delta\mathbf{W})^\top = \mathbf{h}\mathbf{W}^\top + \mathbf{h}\Delta\mathbf{W}^\top \quad (2)$$

$\Delta\mathbf{W}$  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

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

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). 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:

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

where  $\alpha$  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  $\mathbf{H} = [\mathbf{h}_1^t, \dots, \mathbf{h}_m^t, \mathbf{h}_{m+1}^i, \mathbf{h}_{m+2}^i, \dots, \mathbf{h}_{m+(H \times W)}^i, \dots, \mathbf{h}_N^t] \in \mathbb{R}^{N \times d_{in}}$  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  $\mathbf{H}^t = [\mathbf{h}_1^t, \mathbf{h}_2^t, \dots, \mathbf{h}_m^t, \mathbf{h}_{m+(H \times W)+1}^t, \dots, \mathbf{h}_N^t]$  and image hidden states  $\mathbf{H}^i = [[\mathbf{h}_{m+1}^i, \dots, \mathbf{h}_{m+(H \times W)}^i], [\mathbf{h}_{n+1}^i, \dots, \mathbf{h}_{n+(H \times W)}^i], \dots]$ , 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  $2 \times 2$  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

270 the convolution operation on an image patch, the kernel only covers neighboring patches on the top  
 271 and left sides of a patch. For example, the new hidden state of patch 9 is computed from patches:  
 272 5, 6, 8, and 9. To preserve the shape ( $H \times W$ ) of the input image patches, we pad the reshaped image  
 273 hidden states with zero vectors on the top and left sides, as shown by the grey squares in Figure 2.  
 274 Finally, we assemble  $\hat{\mathbf{H}}^i$  and  $\hat{\mathbf{H}}^t$  back to their original sequence to form  $\hat{\mathbf{H}}$ .  
 275

## 276 5 INTERLEAVED INSTRUCTION TUNING WITH LEAFINSTRUCT

277 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 Instruct-Pix2Pix (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.

### 287 5.1 DATASET: LEAFINSTRUCT

288 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.

295 **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., 296 2016), WikiWeb2M (Burns et al., 2023) and YouCook2 (Zhou et al., 2018). Since the original 297 data sources can be noisy, we meticulously devised an automatic data annotation pipeline to ensure 298 the high quality of our curated data. We include the details on dataset construction in Table 3 in 299 Appendix B. We also conducted a rigorous human assessment of our dataset (see Section 7).

302 **Dataset Statistics** After applying our rigorous data processing pipeline, we totally obtain 184,982 303 high-quality instances out of more than 7 million source samples. Our dataset covers a wide range 304 of realistic instruction-tuning tasks, including multimodal document completion, multimodal dialogue, 305 visual storytelling, multimodal script generation, and knowledge-intensive generation. In 306 Appendix B, we show the domain distribution of LEAFINSTRUCT in Figure 7 and compare our 307 dataset with existing datasets in Table 4. These analyses effectively demonstrate the diversity and 308 the novelty of our dataset.

### 309 5.2 INSTRUCTION TUNING FOR INTERLEAVED GENERATION

311 With our curated LEAFINSTRUCT, we enable large-scale interleaved instruction tuning so that the 312 model can learn how to follow human instructions to generate desired interleaved text and images. 313 To preserve the VLG’s capability obtained from pre-training, we only fine-tuned the modality- 314 specialized adaptation layers, and the remaining parameters in the VLGs are kept frozen.

315 Specifically, as shown on the right side of Figure 6, given the task instruction and the interleaved 316 context as inputs, the model is trained to autoregressively generate interleaved text tokens and images 317 with two alternative generation modes for text and images, respectively. We use a special token 318 <IMG> to indicate where an image occurs in the interleaved sequence. The training process is 319 as follows: (1) The model is set to the **text generation** mode by default. During this mode, the 320 hidden states of newly generated tokens are always routed to linear LoRA, and only the parameters 321 in the linear LoRA are optimized. (2) After the <IMG> token is generated, the model switches 322 to **image generation** mode. The VLG takes in the updated context ended with <IMG> and is 323 trained to generate a fixed-length ( $H \times W$ ) sequence of image patch embeddings autoregressively. All 324 the hidden states of generated image embeddings are routed to Convolutional LoRA and only

324 the parameters in the Convolutional LoRA are fine-tuned. (3) When the generation of an image is  
 325 finished, the model is trained to predict an end-of-image token </IMG>, and the model will resume  
 326 the text generation mode. This process will be iterated until the training on a sequence is finished.  
 327

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

## 336 6 EXPERIMENTS

### 338 6.1 EXPERIMENT SETUP

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

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

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

365 **Baselines** For fair comparisons, we primarily compare our methods with current state-of-the-art  
 366 **open-source** VLGs, including GILL (Koh et al., 2023), MiniGPT-5 (Zheng et al., 2023), Pretrained  
 367 Emu2, and Chameleon. We also report the performance of pipelines based on **proprietary** mod-  
 368 els, including Gemini 1.5 (Reid et al., 2024)+SDXL (Podell et al., 2023) and GPT-4o (OpenAI,  
 369 2024)+DALLE 3 (Betker et al.). For these baselines, we first prompt the VLMs (e.g., GPT-4o)  
 370 to generate text along with image captions, and then feed the image captions to a separate image  
 371 generation model (e.g., DALLE). We report these performances only for reference purposes.

### 373 6.2 MAIN RESULTS

375 **Quantitative Results** Table 1 presents the main results of our method in comparison to the base-  
 376 lines. We have the following findings. **Firstly**, our approach is highly effective and efficient when  
 377 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

378  
 379 **Table 1: Main results of interleaved generation on InterleavedBench.** We show the performance  
 380 of pipelines based on proprietary models (Top), open-source VLGs (Middle), and the VLGs trained  
 381 with our MOSS and LEAFINSTRUCT (Bottom), respectively. Note that the scale is from 0 to 5 (5  
 382 is the best). We also report the percentage of improvement in our method over the original VLG  
 383 backbone in the parentheses. The best results are highlighted in **bold**.

Model	Text Quality	Perceptual Quality	Image Coherence	TIC	Helpfulness
<b>Proprietary Models</b>					
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
<b>Open-Source Models</b>					
MiniGPT-5	1.31	3.44	2.06	2.66	1.76
GILL	1.44	<b>4.02</b>	2.12	2.69	1.53
Emu2	1.33	2.29	1.71	1.22	1.87
Chameleon	<b>3.33</b>	0.67	0.28	0.47	1.43
Emu2 + MOSS (Ours)	2.61 (+96.2%)	3.62 (+58.1%)	<b>3.41</b> (+99.4%)	<b>3.54</b> (+190.2%)	<b>2.71</b> (+44.9%)
Chameleon + MoSS (Ours)	2.98 (-10.5%)	2.25 (+235.8%)	1.05 (+275%)	1.7 (+261.7%)	1.82 (+27.3%)

393  
 394 Emu2 model, Emu2+MoSS achieved a performance gain of **up to 190.2%** (on Text-Image Coher-  
 395 ence) and **97.76%** on the average of 5 aspects, almost doubling the overall performance. **Secondly**,  
 396 our method beats the previous open-sourced state-of-the-art (i.e., GILL) by a large margin, i.e.,  
 397 **34.7%** on the average of 5 aspects. Particularly, the outputs of our method have better coherence  
 398 across images (w/ 37.8% improvement in Image Coherence) and between text and images (w/ 31.6%  
 399 improvement in Text-Image Coherence). Our method also exhibits better instruction-following ca-  
 400 pability and is able to generate more helpful content given the 11.5% improvement in Helpfulness.  
 401 **Thirdly**, it is worth noting that the Chameleon baseline achieves good performance on Text Quality  
 402 but extremely poor performance on Perceptual Quality, Image Coherence, and Text-Image Coher-  
 403 ence. We observed that Chameleon usually generates long and comprehensive text responses with  
 404 no image output, thus leading to poor performance on image-related aspects. We hypothesize the  
 405 reason lies in the lack of instruction tuning on interleaved generation with both text and images.  
 406 From Table 1, our approach improves the original Chameleon by a significant margin, especially on  
 407 image-related aspects. This shows that our interleaved instruction tuning can effectively enhance a  
 408 VLG that was previously poor at mixed-modal generation. **Fourthly**, there remains a notable gap  
 409 between open-sourced VLGs and the pipeline approaches based on proprietary models, indicating  
 410 building a powerful and general-purpose open-sourced VLGs is still a challenging research problem.

410 **Per-task Performance** We also show the average  
 411 performance on each task on InterleavedBench in  
 412 Figure 3. Specifically, our method (i.e., Emu2-  
 413 MoSS) outperforms the baselines on most tasks, of-  
 414 ten by a large margin. For subject-driven generation,  
 415 the slightly lower performance of our approach com-  
 416 pared to other baselines is due to its poorer percep-  
 417 tual quality. We included more detailed justifications  
 418 for this result in the following section. We report the  
 419 per-task performance on all aspects in Figure 8 in  
 420 Appendix C.3. In addition, we include more results  
 421 in Table 5, 6, 7, and 8 in Appendix C.2.

422 **Qualitative Results** To better interpret the results,  
 423 we conducted a qualitative analysis on several open-  
 424 sourced baselines and our MOSS in Figure 4. Our  
 425 findings are as follows. **Firstly**, our method demon-  
 426 strates better helpfulness and instruction-following  
 427 capabilities. For example, in the first row in Fig-  
 428 ure 4, our method generates a more coherent visual story with more diverse content given the input.  
 429 In the third row, our approach provides a more natural and reasonable next step for the user, i.e., *have*  
 430 *a good vegan lunch*, while other baselines either jump to *dinner* (MiniGPT-5) or stick to *breakfast*  
 431 (Emu2). **Secondly**, we observed that poor text quality is a common issue for many baselines. For  
 432 instance, MiniGPT-5 often fails to generate explanatory text while GILL usually generates a short  
 433 caption, e.g., *the gardens* in the first row and *the tofu scramble* in the third row, instead of generat-

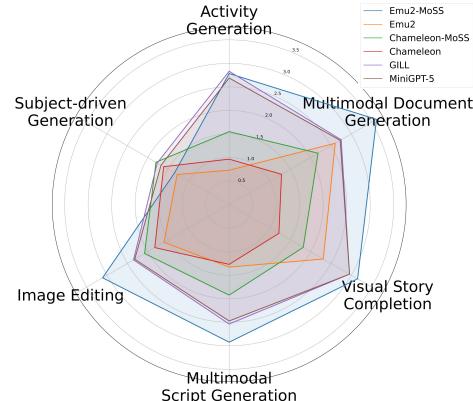


Figure 3: **Per-task performance** averaged on 5 aspects on InterleavedBench.

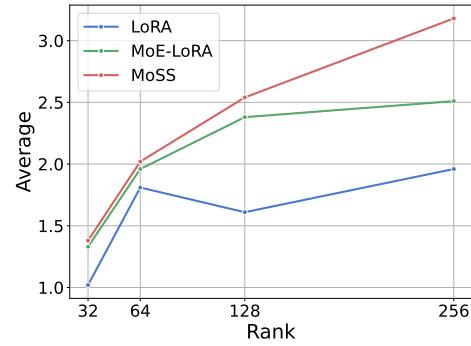
	<b>Input</b>	<b>MiniGPT-5</b>	<b>GILL</b>	<b>Emu2</b>	<b>Emu2+MoSS (Ours)</b>	
432	<b>Instruction:</b> Complete a travel story. <b>Context:</b> I visited the countryside for the weekend <IMG>.	<b>Output:</b> <IMG>	<b>Output:</b> the gardens <IMG>	<b>Output:</b> I was so excited to visit the country side for the weekend <IMG>. I was so excited to visit the country side for the weekend.	<b>Output:</b> I went to the countryside and saw many beautiful flowers <IMG>. I also saw many beautiful buildings <IMG>.	
433						
434	<b>Instruction:</b> Describe a team's success in University Rover Challenge. <b>Context:</b> During the challenge, the PCz Team placed the second <IMG>.	<b>Output:</b> <IMG>	<b>Output:</b> The team's rover is an autonomous vehicle that can across the surface of the moon. <IMG>	<b>Output:</b> Describe a team's success in University Rover Challenge <IMG>.	<b>Output:</b> 16 teams from 10 countries competed in the University Rover Challenge. <IMG>	
435						
436	<b>Instruction:</b> Tell me what should do next to plan vegan healthy meals. <b>Context:</b> I've done: I had oatmeal with fruit <IMG> and a tofu scramble for breakfast in the morning <IMG>.	<b>Output:</b> After you eat breakfast make sure to have enough food for dinner <IMG>.	<b>Output:</b> the tofu scramble <IMG>	<b>Output:</b> Eat vegan cereal with a non-dairy milk for a quick breakfast <IMG>.	<b>Output:</b> Have a good vegan meal for lunch <IMG>.	
437						
438						
439						
440						
441						
442	<b>Instruction:</b> Help me modify this picture <IMG> and change the bus into a train.	<b>Output:</b> <IMG>	<b>Output:</b> a bus on the road <IMG>	<b>Output:</b> <IMG>	<b>Output:</b> The train is on its track <IMG>.	
443						
444						
445						
446						
447						
448						
449						
450						
451						
452						
453						
454						
455						
456						
457						
458						
459						
460						
461						
462						
463						
464						
465	<b>Figure 4: Qualitative results of MOSS and open-source baselines. The &lt;IMG&gt; tokens denote the images' positions in the interleaved sequences.</b>					
466						
467	ing useful content to solve the task. <b>Thirdly</b> , 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. <b>Finally</b> , 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.					
468						
469						
470						
471						
472						
473						
474						
475						
476						
477						
478	<h2>7 DISCUSSIONS</h2>					
479						
480	<b>Comparison between MOSS and other PEFT Methods</b> 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 hidden state, i.e., whether the hidden state is used to generate text or image. From Table 2, we					
481						
482						
483						
484						
485						

486  
 487 **Table 2: Comparison between MOSS and existing PEFT methods**, i.e., traditional linear LoRA,  
 488 and Mixture-of-Expert (MoE) LoRA. Mixture-of-Expert LoRA uses two different sets of linear  
 489 LoRA for images and text, respectively. The rank number is set to 256 for all methods in this table.  
 490

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)	<b>2.61</b>	<b>3.62</b>	<b>3.41</b>	<b>3.54</b>	<b>2.71</b>

495 effectively verify the benefits of using separate parameters for image and text. MOSS significantly  
 496 outperforms the MoE-LoRA across all aspects, especially the image-related aspects such as Image  
 497 Coherence and Text-Image Coherence (TIC). The conclusions from the results are two-fold. First,  
 498 it shows that introducing modality-specialized architecture and parameters can effectively improve  
 499 interleaved text-and-image generation. Second, it verifies that convolutional LoRA can improve  
 500 image generation by better modeling the local priors of images.  
 501

502 **Effect of Rank Number** To investigate how the  
 503 number of rank  $r$  can affect the performance,  
 504 we show the performance averaged on 5 aspects  
 505 on InterleavedBench with the rank number equals  
 506 (32, 64, 128, 256) comparing LoRA, MoE-LoRA,  
 507 and our MOSS in Figure 5. Our approach con-  
 508 sistently outperforms LoRA and MoE-LoRA across all  
 509 rank numbers, and as the rank number increases, the  
 510 gap between MOSS and previous methods con-  
 511 sistently grows larger. This proves the effectiveness  
 512 and generalizability of MOSS across different rank  
 513 sizes. Based on this experiment, we set the rank  
 514 number of our approach to 256 by default. We in-  
 515 clude more results on the effect of rank in LoRA in  
 516 Table 9, 10, and 11 in Appendix C.4.



517 Figure 5: Performance averaged on 5 aspects  
 518 with different rank numbers.  
 519

520 **Quality Assessment of LEAFINSTRUCT** To verify our LEAFINSTRUCT dataset is of high quality,  
 521 we conduct a rigorous human evaluation using the multi-aspect evaluation criteria in InterleavedE-  
 522 val (Liu et al., 2024). Specifically, we use a scale of 0 to 3 in the evaluation, where 0 is the lowest  
 523 score while 3 is the highest. We randomly sampled 200 instances from LEAFINSTRUCT and asked  
 524 two human annotators with expertise in NLP and multimodal research to rate each instance from 5  
 525 aspects. We report the averaged scores from two annotators in Table 3. We show that the sampled  
 526 instances consistently achieved almost full scores across all 5 aspects, which effectively demonstrated  
 527 that our curated dataset is of high quality.  
 528

529 Table 3: **Human evalution** of randomly sampled instances from LeafInstruct. Note that the scale is  
 530 from 0 to 3 (**Score 3 is the best**), which is different from the scale used in Table 1 and Table 2.  
 531

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

## 532 8 CONCLUSION

533 We propose MODALITY-SPECIALIZED SYNERGIZERS (MOSS), a novel modality-specialized  
 534 adaptation framework tailored for VLGs. MOSS dedicates a set of linear LoRA for processing  
 535 text and a set of Convolutional LoRA for images, allowing each modality to have its own optimal  
 536 adaptation design. Besides, we propose the first interleaved instruction tuning dataset LEAFIN-  
 537 STRUCT and verify the dataset quality via rigorous human evaluation. Extensive experiments on  
 538 InterleavedBench showcase that our proposed method and dataset are highly effective, establishing  
 539 the new state-of-the-art among open-sourced VLGs in interleaved text-and-image generation.  
 540

540 REFERENCES  
541

- 542 Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal,  
543 Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, and Luke Zettlemoyer. CM3: A  
544 causal masked multimodal model of the internet. *CoRR*, abs/2201.07520, 2022. URL <https://arxiv.org/abs/2201.07520>.
- 545 Hassan Akbari, Dan Kondratyuk, Yin Cui, Rachel Hornung, Huisheng Wang, and Hartwig Adam.  
546 Alternating gradient descent and mixture-of-experts for integrated multimodal perception. In  
547 *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=uT1KUAm68H>.
- 548 Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel  
549 Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford,  
550 Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick,  
551 Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski,  
552 Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual  
553 language model for few-shot learning. 2022. URL [http://papers.nips.cc/paper\\_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html).
- 554 Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of  
555 natural images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*  
556 2022, New Orleans, LA, USA, June 18-24, 2022, pp. 18187–18197. IEEE, 2022. doi: 10.1109/  
557 CVPR52688.2022.01767. URL <https://doi.org/10.1109/CVPR52688.2022.01767>.
- 558 James Betker, Gabriel Goh, Li Jing, † TimBrooks, Jianfeng Wang, Linjie Li, † LongOuyang, †  
559 JuntangZhuang, † JoyceLee, † YufeiGuo, † WesamManassra, † PrafullaDhariwal, † CaseyChu, †  
560 YunxinJiao, and Aditya Ramesh. Improving image generation with better captions. URL <https://api.semanticscholar.org/CorpusID:264403242>.
- 561 Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image  
562 editing instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*  
563 2023, Vancouver, BC, Canada, June 17-24, 2023, pp. 18392–18402. IEEE, 2023a. doi: 10.  
564 1109/CVPR52729.2023.01764. URL <https://doi.org/10.1109/CVPR52729.2023.01764>.
- 565 Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image  
566 editing instructions. In *CVPR*, 2023b.
- 567 Andrea Burns, Krishna Srinivasan, Joshua Ainslie, Geoff Brown, Bryan A. Plummer, Kate Saenko,  
568 Jianmo Ni, and Mandy Guo. Wikiweb2m: A page-level multimodal wikipedia dataset. *CoRR*,  
569 abs/2305.05432, 2023. doi: 10.48550/ARXIV.2305.05432. URL <https://doi.org/10.48550/arXiv.2305.05432>.
- 570 Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and  
571 Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of  
572 the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.
- 573 Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing  
574 multimodal Ilm’s referential dialogue magic. *CoRR*, abs/2306.15195, 2023a. doi: 10.48550/  
575 ARXIV.2306.15195. URL <https://doi.org/10.48550/arXiv.2306.15195>.
- 576 Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and  
577 Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *CoRR*,  
578 abs/2311.12793, 2023b. doi: 10.48550/ARXIV.2311.12793. URL <https://doi.org/10.48550/arXiv.2311.12793>.
- 579 Wenhui Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and  
580 William W. Cohen. Subject-driven text-to-image generation via apprenticeship learning. In  
581 Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine  
582 (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neu-*  
583 *ral Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December  
584 10 - 16, 2023*, 2023c. URL [http://papers.nips.cc/paper\\_files/paper/2023/hash/6091bf1542b118287db4088bc16be8d9-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2023/hash/6091bf1542b118287db4088bc16be8d9-Abstract-Conference.html).

- 594 Zhe Chen, Yuchen Duan, Wenhui Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision  
 595 transformer adapter for dense predictions. In *The Eleventh International Conference on Learning  
 596 Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023d. URL  
 597 <https://openreview.net/pdf?id=plKu2GByCNW>.
- 598 Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. ANOLE: an open, autoregressive, native large  
 599 multimodal models for interleaved image-text generation. *CoRR*, abs/2407.06135, 2024. doi:  
 600 10.48550/ARXIV.2407.06135. URL <https://doi.org/10.48550/arXiv.2407.06135>.
- 602 Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang,  
 603 Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-  
 604 language models with instruction tuning. *CoRR*, abs/2305.06500, 2023. doi: 10.48550/ARXIV.  
 605 2305.06500. URL <https://doi.org/10.48550/arXiv.2305.06500>.
- 606 Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian  
 607 Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi.  
 608 DreamLLM: Synergistic multimodal comprehension and creation. In *The Twelfth International  
 609 Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=y01KGvd9Bw>.
- 610 Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-  
 611 resolution image synthesis. In *IEEE Conference on Computer Vision and Pattern  
 612 Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 12873–12883. Computer Vi-  
 613 sion 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](https://openaccess.thecvf.com/content/CVPR2021/html/Esser_Taming_Transformers_for_High-Resolution_Image_Synthesis_CVPR_2021_paper.html).
- 614 Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao,  
 615 and Qingwei Lin. MMDialog: A large-scale multi-turn dialogue dataset towards multi-modal  
 616 open-domain conversation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.),  
 617 *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume  
 618 1: Long Papers)*, pp. 7348–7363, Toronto, Canada, July 2023. Association for Computational  
 619 Linguistics. doi: 10.18653/v1/2023.acl-long.405. URL <https://aclanthology.org/2023.acl-long.405>.
- 620 Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-  
 621 a-scene: Scene-based text-to-image generation with human priors. In Shai Avidan, Gabriel J.  
 622 Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision  
 623 - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings,  
 624 Part XV*, volume 13675 of *Lecture Notes in Computer Science*, pp. 89–106. Springer, 2022. doi:  
 625 10.1007/978-3-031-19784-0\\_6. URL [https://doi.org/10.1007/978-3-031-19784-0\\_6](https://doi.org/10.1007/978-3-031-19784-0_6).
- 626 Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying  
 627 Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation.  
 628 *arXiv preprint arXiv:2404.14396*, 2024.
- 629 Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a  
 630 reference-free evaluation metric for image captioning. In *EMNLP*, 2021.
- 631 Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and  
 632 Weizhu Chen. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685,  
 633 2021. URL <https://arxiv.org/abs/2106.09685>.
- 634 Hexiang Hu, Kelvin C. K. Chan, Yu-Chuan Su, Wenhui Chen, Yandong Li, Kihyuk Sohn, Yang  
 635 Zhao, Xue Ben, Boqing Gong, William W. Cohen, Ming-Wei Chang, and Xuhui Jia. Instruct-  
 636 imagen: Image generation with multi-modal instruction. *CoRR*, abs/2401.01952, 2024. doi:  
 637 10.48550/ARXIV.2401.01952. URL <https://doi.org/10.48550/arXiv.2401.01952>.
- 638 Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. BLIVA: A simple multi-  
 639 modal LLM for better handling of text-rich visual questions. *CoRR*, abs/2308.09936, 2023. doi:  
 640 10.48550/ARXIV.2308.09936. URL <https://doi.org/10.48550/arXiv.2308.09936>.

- 648 Ting-Hao (Kenneth) Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya  
 649 Agrawal, Jacob Devlin, Ross B. Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra,  
 650 C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell.  
 651 Visual storytelling. In Kevin Knight, Ani Nenkova, and Owen Rambow (eds.), *NAACL HLT*  
 652 *2016, The 2016 Conference of the North American Chapter of the Association for Computational*  
 653 *Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pp.  
 654 1233–1239. The Association for Computational Linguistics, 2016. doi: 10.18653/V1/N16-1147.  
 655 URL <https://doi.org/10.18653/v1/n16-1147>.
- 656 Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge J. Belongie, Bharath Hariha-  
 657 ran, and Ser-Nam Lim. Visual prompt tuning. In Shai Avidan, Gabriel J. Brostow, Moustapha  
 658 Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th*  
 659 *European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXIII*, vol-  
 660 *ume 13693 of Lecture Notes in Computer Science*, pp. 709–727. Springer, 2022. doi: 10.1007/  
 661 978-3-031-19827-4\\_41. URL [https://doi.org/10.1007/978-3-031-19827-4\\_41](https://doi.org/10.1007/978-3-031-19827-4_41).
- 662 Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhui Chen. MANTIS:  
 663 interleaved multi-image instruction tuning. *CoRR*, abs/2405.01483, 2024. doi: 10.48550/ARXIV.  
 664 2405.01483. URL <https://doi.org/10.48550/arXiv.2405.01483>.
- 665 Shibo Jie and Zhi-Hong Deng. Convolutional bypasses are better vision transformer adapters.  
 666 *CoRR*, abs/2207.07039, 2022. doi: 10.48550/ARXIV.2207.07039. URL <https://doi.org/10.48550/arXiv.2207.07039>.
- 667 Yang Jin, Kun Xu, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Quzhe Huang, Bin Chen,  
 668 Chenyi Lei, An Liu, Chengru Song, Xiaoqiang Lei, Di Zhang, Wenwu Ou, Kun Gai, and  
 669 Yadong Mu. Unified language-vision pretraining in LLM with dynamic discrete visual tok-  
 670 enization. *CoRR*, abs/2309.04669, 2023. doi: 10.48550/ARXIV.2309.04669. URL <https://doi.org/10.48550/arXiv.2309.04669>.
- 671 Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank  
 672 hypercomplex adapter layers. *Advances in Neural Information Processing Systems*, 34:1022–  
 673 1035, 2021.
- 674 Jing Yu Koh, Daniel Fried, and Russ Salakhutdinov. Generating images with multimodal language  
 675 models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey  
 676 Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on*  
 677 *Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, Decem-  
 678 ber 10 - 16, 2023*, 2023. URL [http://papers.nips.cc/paper\\_files/paper/2023/hash/43a69d143273bd8215578bde887bb552-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2023/hash/43a69d143273bd8215578bde887bb552-Abstract-Conference.html).
- 679 Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning  
 680 events in videos. In *International Conference on Computer Vision (ICCV)*, 2017.
- 681 Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept  
 682 customization of text-to-image diffusion. 2023.
- 683 Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and  
 684 Ziwei Liu. MIMIC-IT: multi-modal in-context instruction tuning. *CoRR*, abs/2306.05425, 2023a.  
 685 doi: 10.48550/ARXIV.2306.05425. URL <https://doi.org/10.48550/arXiv.2306.05425>.
- 686 Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan  
 687 Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models.  
 688 *CoRR*, abs/2407.07895, 2024a. doi: 10.48550/ARXIV.2407.07895. URL <https://doi.org/10.48550/arXiv.2407.07895>.
- 689 Huayang Li, Siheng Li, Deng Cai, Longyue Wang, Lemao Liu, Taro Watanabe, Yujiu Yang, and  
 690 Shuming Shi. Textbind: Multi-turn interleaved multimodal instruction-following in the wild.  
 691 *CoRR*, abs/2309.08637, 2023b. doi: 10.48550/ARXIV.2309.08637. URL <https://doi.org/10.48550/arXiv.2309.08637>.

- 702 Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-  
 703 image pre-training with frozen image encoders and large language models. 202:19730–19742,  
 704 2023c. URL <https://proceedings.mlr.press/v202/li23q.html>.
- 705 Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In  
 706 *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the*  
 707 *11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*,  
 708 pp. 4582–4597, 2021.
- 710 Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng  
 711 Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models.  
 712 *CoRR*, abs/2403.18814, 2024b. doi: 10.48550/ARXIV.2403.18814. URL <https://doi.org/10.48550/arXiv.2403.18814>.
- 714 Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your  
 715 features: A new baseline for efficient model tuning. In Sanmi Koyejo, S. Mo-  
 716 hamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neu-*  
 717 *ral Information Processing Systems 35: Annual Conference on Neural Information Pro-*  
 718 *cessing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - Decem-*  
 719 *ber 9, 2022*, 2022. URL [http://papers.nips.cc/paper\\_files/paper/2022/hash/00bb4e415ef117f2dee2fc3b778d806d-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2022/hash/00bb4e415ef117f2dee2fc3b778d806d-Abstract-Conference.html).
- 721 Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and  
 722 Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *CoRR*, abs/2401.15947,  
 723 2024. doi: 10.48550/ARXIV.2401.15947. URL <https://doi.org/10.48550/arXiv.2401.15947>.
- 725 Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large  
 726 multi-modal model with robust instruction tuning. *CoRR*, abs/2306.14565, 2023a. doi: 10.48550/  
 727 ARXIV.2306.14565. URL <https://doi.org/10.48550/arXiv.2306.14565>.
- 729 Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and  
 730 Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context  
 731 learning. *CoRR*, abs/2205.05638, 2022. doi: 10.48550/ARXIV.2205.05638. URL <https://doi.org/10.48550/arXiv.2205.05638>.
- 733 Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction  
 734 tuning. *CoRR*, abs/2310.03744, 2023b. doi: 10.48550/ARXIV.2310.03744. URL <https://doi.org/10.48550/arXiv.2310.03744>.
- 737 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023c. URL  
 738 <https://doi.org/10.48550/arXiv.2304.08485>.
- 739 Minqian Liu, Zhiyang Xu, Zihao Lin, Trevor Ashby, Joy Rimchala, Jiaxin Zhang, and Lifu Huang.  
 740 Holistic evaluation for interleaved text-and-image generation. In *Proceedings of the 2024 Con-*  
 741 *ference on Empirical Methods in Natural Language Processing*, 2024.
- 743 Chenyang Lyu, Minghao Wu, Longyue Wang, Xiting Huang, Bingshuai Liu, Zefeng Du, Shuming  
 744 Shi, and Zhaopeng Tu. Macaw-llm: Multi-modal language modeling with image, audio, video,  
 745 and text integration. *CoRR*, abs/2306.09093, 2023. doi: 10.48550/ARXIV.2306.09093. URL  
 746 <https://doi.org/10.48550/arXiv.2306.09093>.
- 747 OpenAI. Hello gpt-4o. <https://openai.com/index/hello-gpt-4o/>, 2024. Accessed: 2024-  
 748 05-26.
- 749 Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe  
 750 Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image  
 751 synthesis. *CoRR*, abs/2307.01952, 2023. doi: 10.48550/ARXIV.2307.01952. URL <https://doi.org/10.48550/arXiv.2307.01952>.
- 754 Jingyuan Qi, Minqian Liu, Ying Shen, Zhiyang Xu, and Lifu Huang. MULTISCRIPT: multimodal  
 755 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*

- 756        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>.
- 760        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.
- 764        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>.
- 775        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>.
- 780        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>.
- 784        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>.
- 787        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>.
- 792        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.
- 796        Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. Codis-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>.
- 800        Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2024.
- 801        Changyao Tian, Xizhou Zhu, Yuwen Xiong, Weiyun Wang, Zhe Chen, Wenhui Wang, Yun-tao 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>.
- 806        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>.

- 810 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N  
 811 Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon,  
 812 U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett  
 813 (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates,  
 814 Inc., 2017. URL [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/3f5ee243547dee91fb053c1c4a845aa-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fb053c1c4a845aa-Paper.pdf).
- 815
- 816 Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to  
 817 believe: Prompting GPT-4V for better visual instruction tuning. *CoRR*, abs/2311.07574, 2023.  
 818 doi: 10.48550/ARXIV.2311.07574. URL <https://doi.org/10.48550/arXiv.2311.07574>.  
 819
- 820 Yaqing Wang, Sahaj Agarwal, Subhabrata Mukherjee, Xiaodong Liu, Jing Gao, Ahmed Hassan  
 821 Awadallah, and Jianfeng Gao. Adamix: Mixture-of-adaptations for parameter-efficient model  
 822 tuning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022*  
 823 *Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi,*  
 824 *United Arab Emirates, December 7-11, 2022*, pp. 5744–5760. Association for Computational  
 825 Linguistics, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.388. URL <https://doi.org/10.18653/v1/2022.emnlp-main.388>.  
 826
- 827 Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any  
 828 multimodal LLM. *CoRR*, abs/2309.05519, 2023. doi: 10.48550/ARXIV.2309.05519. URL  
 829 <https://doi.org/10.48550/arXiv.2309.05519>.  
 830
- 831 Zhiyang Xu, Ying Shen, and Lifu Huang. Multiinstruct: Improving multi-modal zero-shot learning  
 832 via instruction tuning. In Anna Rogers, Jordan L. Boyd-Graber, and Naoki Okazaki (eds.),  
 833 *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume*  
 834 *1: Long Papers)*, ACL 2023, Toronto, Canada, July 9-14, 2023, pp. 11445–11465. Association  
 835 for Computational Linguistics, 2023a. doi: 10.18653/V1/2023.ACL-LONG.641. URL <https://doi.org/10.18653/v1/2023.acl-long.641>.  
 836
- 837 Zhiyang Xu, Ying Shen, and Lifu Huang. Multiinstruct: Improving multi-modal zero-shot learning  
 838 via instruction tuning. In Anna Rogers, Jordan L. Boyd-Graber, and Naoki Okazaki (eds.),  
 839 *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume*  
 840 *1: Long Papers)*, ACL 2023, Toronto, Canada, July 9-14, 2023, pp. 11445–11465. Association  
 841 for Computational Linguistics, 2023b. doi: 10.18653/V1/2023.ACL-LONG.641. URL <https://doi.org/10.18653/v1/2023.acl-long.641>.  
 842
- 843 Zhiyang Xu, Chao Feng, Rulin Shao, Trevor Ashby, Ying Shen, Di Jin, Yu Cheng, Qifan Wang, and  
 844 Lifu Huang. Vision-flan: Scaling human-labeled tasks in visual instruction tuning, 2024.  
 845
- 846 Yue Yang, Artemis Panagopoulou, Qing Lyu, Li Zhang, Mark Yatskar, and Chris Callison-Burch.  
 847 Visual goal-step inference using wikiHow. In Marie-Francine Moens, Xuanjing Huang, Lucia  
 848 Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods*  
 849 *in Natural Language Processing*, pp. 2167–2179, Online and Punta Cana, Dominican Republic,  
 850 November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.  
 851 165. URL <https://aclanthology.org/2021.emnlp-main.165>.  
 852
- 853 Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang,  
 854 Mike Lewis, Luke Zettlemoyer, and Wen-Tau Yih. Retrieval-augmented multimodal language  
 855 modeling. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan  
 856 Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023,*  
 857 *23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning*  
 858 *Research*, pp. 39755–39769. PMLR, 2023. URL <https://proceedings.mlr.press/v202/yasunaga23a.html>.  
 859
- 860 Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen  
 861 Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian  
 862 Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models  
 863 with multimodality. *CoRR*, abs/2304.14178, 2023. doi: 10.48550/ARXIV.2304.14178. URL  
<https://doi.org/10.48550/arXiv.2304.14178>.

- 864 Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei  
 865 Huang. *mplug-owl2: Revolutionizing multi-modal large language model with modality collabora-*  
 866 *tion*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*  
 867 (*CVPR*), pp. 13040–13051, June 2024.
- 868 Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai,  
 869 Xiaoshui Huang, Zhiyong Wang, Jing Shao, and Wanli Ouyang. LAMM: language-assisted multi-  
 870 modal instruction-tuning dataset, framework, and benchmark. *CoRR*, abs/2306.06687, 2023. doi:  
 871 10.48550/ARXIV.2306.06687. URL <https://doi.org/10.48550/arXiv.2306.06687>.
- 872 Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun  
 873 Babu, Bin Tang, Brian Karrer, Shelly Sheynin, Candace Ross, Adam Polyak, Russell Howes,  
 874 Vasu Sharma, Puxin Xu, Hovhannes Tamoyan, Oron Ashual, Uriel Singer, Shang-Wen Li, Susan  
 875 Zhang, Richard James, Gargi Ghosh, Yaniv Taigman, Maryam Fazel-Zarandi, Asli Celikyilmaz,  
 876 Luke Zettlemoyer, and Armen Aghajanyan. Scaling autoregressive multi-modal models: Pretrain-  
 877 ing and instruction tuning. *CoRR*, abs/2309.02591, 2023. doi: 10.48550/ARXIV.2309.02591.  
 878 URL <https://doi.org/10.48550/arXiv.2309.02591>.
- 879 Ted Zadouri, Ahmet Üstün, Arash Ahmadian, Beyza Ermis, Acyr Locatelli, and Sara Hooker. Push-  
 880 ing mixture of experts to the limit: Extremely parameter efficient moe for instruction tuning.  
 881 *CoRR*, abs/2309.05444, 2023. doi: 10.48550/ARXIV.2309.05444. URL <https://doi.org/10.48550/arXiv.2309.05444>.
- 882 Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. Bitfit: Simple parameter-efficient fine-tuning  
 883 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.
- 884 Kai Zhang, Lingbo Mo, Wenhui Chen, Huan Sun, and Yu Su. Magicbrush: A man-  
 885 ually annotated dataset for instruction-guided image editing. In Alice Oh, Tristan Nau-  
 886 mann, 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](http://papers.nips.cc/paper_files/paper/2023/hash/64008fa30cba9b4d1ab1bd3bd3d57d61-Abstract-Datasets_and_Benchmarks.html).
- 887 Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable  
 888 effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer  
 889 Vision and Pattern Recognition*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.
- 890 Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong  
 891 Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *CoRR*,  
 892 abs/2306.17107, 2023b. doi: 10.48550/ARXIV.2306.17107. URL <https://doi.org/10.48550/arXiv.2306.17107>.
- 893 Kaizhi Zheng, Xuehai He, and Xin Eric Wang. Minigpt-5: Interleaved vision-and-language gener-  
 894 ation via generative vokens. *CoRR*, abs/2310.02239, 2023. doi: 10.48550/ARXIV.2310.02239.  
 895 URL <https://doi.org/10.48550/arXiv.2310.02239>.
- 896 Zihan Zhong, Zhiqiang Tang, Tong He, Haoyang Fang, and Chun Yuan. Convolution meets lora:  
 897 Parameter efficient finetuning for segment anything model. *CoRR*, abs/2401.17868, 2024. doi:  
 898 10.48550/ARXIV.2401.17868. URL <https://doi.org/10.48550/arXiv.2401.17868>.
- 899 Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from  
 900 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>.
- 910 Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing  
 911 vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592,

918 2023a. doi: 10.48550/ARXIV.2304.10592. URL <https://doi.org/10.48550/arXiv.2304.10592>.  
919  
920

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

925 Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Young-  
926 jae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open,  
927 billion-scale corpus of images interleaved with text. In *Thirty-seventh Conference on Neu-  
928 ral Information Processing Systems Datasets and Benchmarks Track*, 2023c. URL <https://openreview.net/forum?id=t0d8rSjcWz>.  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971

## 972 A MORE DETAILS OF MOSS

### 973 A.1 IMPLEMENTATION DETAILS

976 We leverage the Emu2 model (Sun et al., 2024), consisting of the EVA-02-CLIP-E-plus (Sun et al.,  
 977 2023a) as the image encoder, the LLaMA-33B (Touvron et al., 2023), and the SDXL (Podell et al.,  
 978 2023) as the image decoder, as our base model. The EVA-02-CLIP-E-plus and the LLaMA-33B is  
 979 connected by a linear project-up layer and the LLaMA-33B and the SDXL is connected by a linear  
 980 project-down layer. All the variants of LoRA in Section 7, including our MOSS are trained with  
 981 LEAFINSTRUCT for one epoch on  $8 \times$  A100 GPUs with learning rate  $2e^{-5}$ , batch size 1 per GPU,  
 982 and a gradient accumulation step of 16. All the LoRA have a rank of 256, dropout rate of 0.05, and  
 983 the LoRA  $\alpha$  in Section 4 is set to  $2 \times 128$ . The kernel size of MOSS is  $2 \times 2$ , the stride is set to 1.  
 984 During training, all parameters of the EMU2 model are kept frozen and only the LoRA parameters  
 985 are updated.

### 986 A.2 MORE DETAILS OF MOSS

## 988 B MORE DETAILS OF LEAFINSTRUCT

M4-Instruct		LeafInstruct (Ours)			
<b>Instruction:</b> Here are 5 images <IMG> <IMG> <IMG> <IMG>, which image shows the following content: {caption of image 2}		<b>Instruction:</b> Describe a travel log during a hiking. Highlight the scene of plants, insect, and emphasize the beauty of nature.			
 <b>Output:</b> Image 2.		<b>Input Context:</b> That beauty of nature in all its glory is captivating, with vibrant colors and delicate forms <IMG>. An insect seems to be trying to eat the plant <IMG>.			
<b>InstructPix2Pix</b> <table border="1" style="margin-left: auto; margin-right: auto;"> <tr> <td><b>Instruction:</b> Swap sunflowers with roses &lt;IMG&gt;.</td> <td><b>Output:</b> </td> </tr> </table>		<b>Instruction:</b> Swap sunflowers with roses <IMG>.	<b>Output:</b> 	<b>Output:</b> The colors of nature are truly stunning <IMG>. A wildflower or plant of some sort, possibly a native species, stands out in the landscape <IMG>. There was a little ladybug in nature, its red shell contrasting beautifully with the surrounding foliage <IMG>.	
<b>Instruction:</b> Swap sunflowers with roses <IMG>.	<b>Output:</b> 				
					

1003 Figure 6: Comparison between existing benchmarks and our LEAFINSTRUCT. In existing datasets  
 1004 such as InstructPix2Pix (Brooks et al., 2023b) and Mantis-Instruct (Li et al., 2024a), the outputs are  
 1005 in single modality, either text or image. On the contrary, the inputs and outputs of our LEAFIN-  
 1006 STRUCT cover multiple modalities.

1008 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

### 1019 B.1 MORE DETAILS IN DATASET CONSTRUCTION

1022 We elaborate on the details of our dataset construction pipeline as follows. **Firstly**, we filter the  
 1023 samples based on the text length, number of images, and the coherence between text and images  
 1024 (measured by CLIPScore (Hessel et al., 2021)). We only keep the instances with 3 to 6 images in  
 1025 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

1026 LLM (i.e., Llama-8B-Instruct) as a text filter to discard the instances with poor text quality.  
 1027 **Thirdly**, we remove the instances with duplicate or perceptually highly similar images to ensure  
 1028 the diversity of the images. **Finally**, we also apply Llama3 to annotate the task instruction for each  
 1029 instance based on the text content and rewrite the text if it’s too verbose to prevent the context length  
 1030 from being too long.

1031

1032 **Details of Text Quality Filter** We use  
 1033 Llama-8B-Instruct model to rate the text  
 1034 quality of an instance with the following  
 1035 prompt: “*Imagine you are an expert data an-*  
 1036 *nnotator. You are given a text material and you*  
 1037 *need to evaluate its quality in terms of whether*  
 1038 *it is coherent, fluent, easy to understand, and*  
 1039 *helpful to humans. Please be critical and rate*  
 1040 *the quality as good only when the text quality is*  
 1041 *good in all four aspects. Output 1 if you think*  
 1042 *the material is good after you consider all four*  
 1043 *aspects. Output 0 if you think the material is*  
 1044 *not good enough. Here is the text material to*  
 1045 *be evaluated: {TEXT} Only output 0 or 1 and*  
 1046 *do not output anything else. Your evaluation*  
 1047 *is:.” We discard the instances if the output from*  
 1048 *Llama is 0.*

1049

1050 **Details of Image Filter** We empirically found that if the images are too identical in the training  
 1051 instances, the trained models tend to find a shortcut to simply copy the image during generation. To  
 1052 this end, we design a filter to discard the instances with duplicate images to improve data quality.  
 1053 Specifically, we leverage the LPIPS score (Zhang et al., 2018) that measures the perceptual similarity  
 1054 between the images. Specifically, for each instance, we enumerate each pair of images and compute  
 1055 their LPIPS score. If there is one pair with a score higher than 0.6, we discard the instance. We  
 1056 determine the threshold of 0.6 by empirical trial.

1057

1058 **Details of Instruction Annotation** We also adopt Llama-8B-Instruct to annotate the task in-  
 1059 struction for each instance. We devise instructions to prompt the Llama3 model to rewrite the  
 1060 original text material in the pretraining dataset MMC4 into instruction-tuning instances. The in-  
 1061 put context length is 2048 and the output context length is 1024. We set the temperature as 1 to  
 1062 encourage the diversity of instructions. We use the following prompt: “*Imagine you are an expert*  
 1063 *instruction annotator. You are given a material. You need to read its content and output a brief*  
 1064 *task instruction with one sentence such that another person can recover the given the material given*  
 1065 *the instruction. The instruction you predict should be specifically tailored for creative interleaved*  
 1066 *content generation that consists of both text and images. Now you need to annotate a concise,*  
 1067 *accurate instruction for the following instance. Please only predict the instruction and do not output*  
 1068 *anything else. Please design the instruction for the multi-modal generation task interleaved with*  
 1069 *both text and images. Text: {TEXT} Instruction:.”*

1070

## C MORE EXPERIMENT RESULTS

1071

### C.1 MORE DETAILS ON EVALUATION BENCHMARKS

1072

1073 InterleavedBench has two splits: a context-based split in which the input of each instance is equipped  
 1074 with interleaved text and images; and a context-free split with text-only inputs. The context-based  
 1075 split contains 465 instances and the text-only split contains 350 instances. We only use the context-  
 1076 based split as the testing set since we mainly focus on tasks with interleaved inputs and outputs.

1077

### C.2 MORE RESULTS ON ESTABLISHED BENCHMARKS

1078

1079 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,

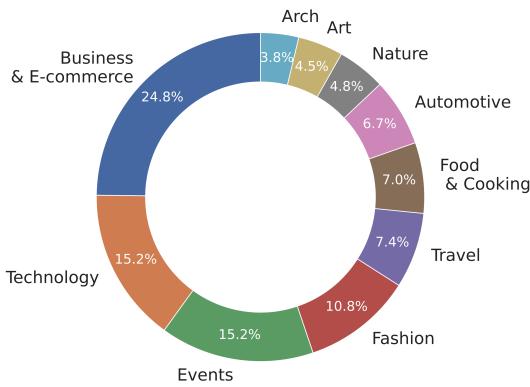


Figure 7: Domain distribution in LeafInstruct.

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	<b>27.28</b>	60.53
Emu2 + MOSS (Ours)	<b>85.88</b>	<b>74.92</b>	25.98	<b>62.26</b>

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	<b>1.76</b>	2.19	2.52	2.91	1.92
MOSS	1.73	<b>2.87</b>	<b>2.54</b>	<b>3.30</b>	<b>2.26</b>

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	<b>3.43</b>	<b>2.03</b>	N/A	<b>1.27</b>

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	<b>3.36</b>	1.67	N/A	1.39
MOSS	N/A	3.24	<b>2.04</b>	N/A	<b>1.83</b>

### C.3 PER-TASK PERFORMANCE ON EACH EVALUATION ASPECT

We report the performance on all aspects for each task in InterleavedBench in Figure 8.

### C.4 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 9, 10, 11, respectively.

