

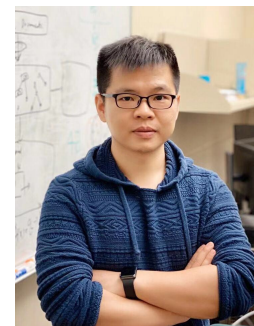
# Modality-Specialized Synergizers for Interleaved Vision-Language Generalists

---

Zhiyang Xu\*, Minqian Liu\*, Ying Shen, Joy Rimchala, Jiaxin Zhang, Qifan Wang, Yu Cheng, Lifu Huang

**Virginia Tech, Intuit AI Research, Meta AI, CUHK, UC Davis**

**ICLR 2025**



\*: Equal Contribution.

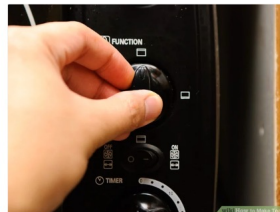
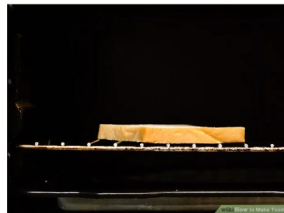
# Interleaved Text-and-Image Generation

- Interleaved generation<sup>[1,2]</sup> that requires models generating **both text and images** is an increasingly important task in multimodal learning.



**Input:** Given the task “How to make a toast in an oven” and the first two steps, predict the subsequent steps to complete the task.

**Step 1:** Put the slices of bread flat on the oven rack. **Step 2:** Turn on the broiler of the oven, or set the heat on the toaster oven.



**Output:** Here're the subsequent steps:

**Step 3:** Use tongs to flip the bread over half-way through the toasting. **Step 4:** Remove the toast from the oven



[1] Chameleon: Mixed-Modal Early-Fusion Foundation Models. Chameleon Team at Meta. 2024.

[2] Emu3: Next-Token Prediction is All You Need. Emu3 Team at BAAI. 2024.

# Challenges in Interleaved Text-and-Image Generation

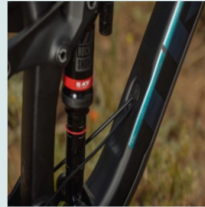
- Existing models use the same architecture with the same set of parameters to process both text and images, which may be **insufficient to model the distinct inductive biases** in each modality given their intrinsic discrepancy.

## Inferior Text and Image Quality



**Instruction:** Describe the new features of the latest mountain bike.

**Context:** Last year's top sales now has an aluminum model <IMG>. The new Slash 9.9 bike has an upgraded drivetrain <IMG> ... <IMG>



**Output:** The Slash 9.9 is available in 27.5-inch and 29-inch wheel sizes. The Slash 9.9 is available in 27.5-inch and 29-inch wheel sizes <IMG>.



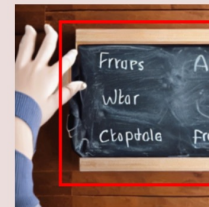
## Weak Instruction-Following Capability



**Instruction:** I need to learn cloudy glasses. **Context:** I have finished these steps: Acquire a soft, clean cloth <IMG>. Use a cleaner <IMG>. Clean with dish soap and warm water <IMG>. What should I do next?



**Output:** how to clean glasses with a cloth and a drop of dish soap <IMG>.



# Modality-Specialized Synergizers (MoSS)

- Introduce **modality-specialized modules** to model different inductive biases.

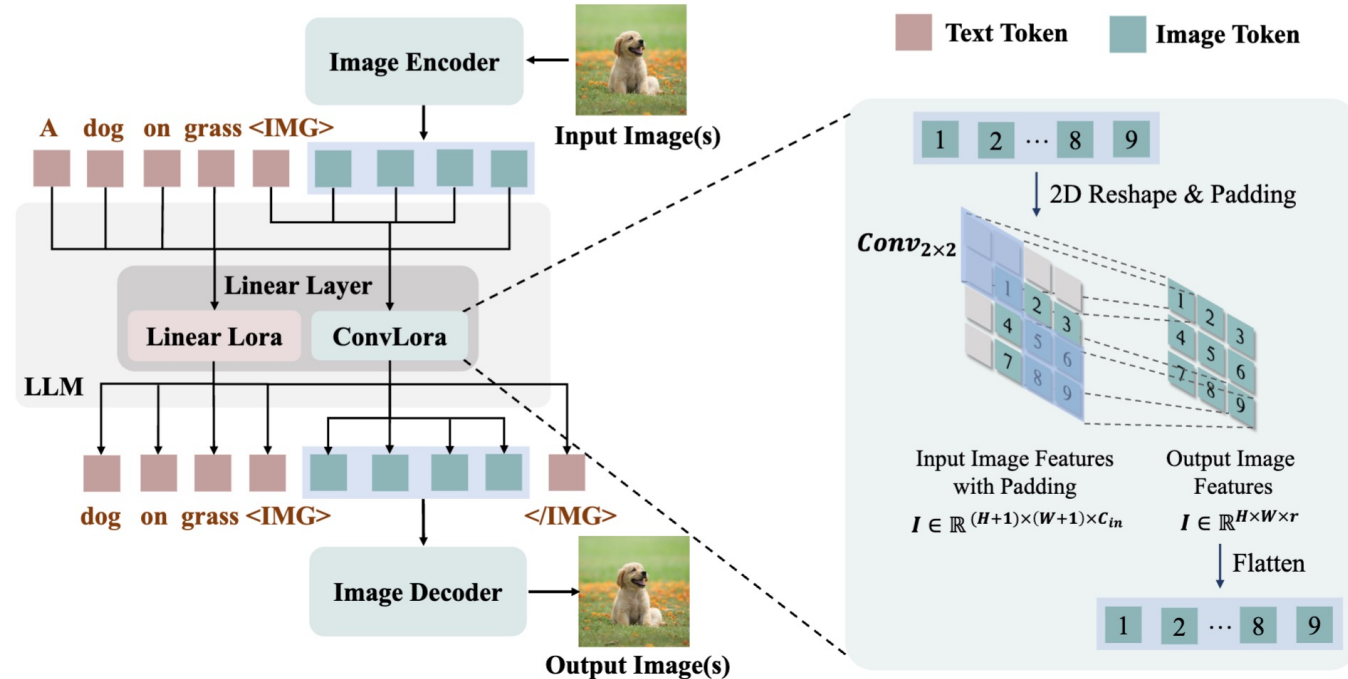


Figure 2: An autoregressive VLG with our proposed MoSS added to its linear layers. The linear LoRA on the left side is specialized to generate text tokens and the Convolutional LoRA on the right side is specialized to generate image patches. On the right handside, we show the details of convolutional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



# Interleaved Instruction Tuning with LeafInstruct

- We also LeafInstruct, first open-sourced **interleaved instruction tuning** dataset comprising 184,982 high-quality instances on more than 10 diverse domains.

## M4-Instruct

**Instruction:** Here are 5 images <IMG> <IMG> <IMG> <IMG> <IMG>, which image shows the following content: {caption of image 2}



**Output:** Image 2.

## InstructPix2Pix

**Instruction:** Swap sunflowers with roses <IMG>.



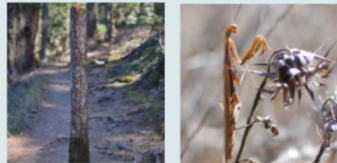
**Output:**



## LeafInstruct (Ours)

**Instruction:** Describe a travel log during a hiking. Highlight the scene of plants, insect, and emphasize the beauty of nature.

**Input Context:** 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>.



**Output:** 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>.

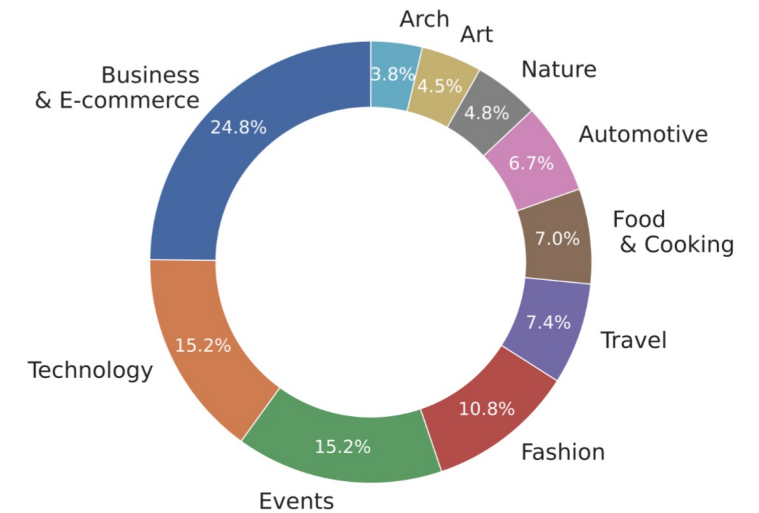


Figure 7: Domain distribution in LeafInstruct.



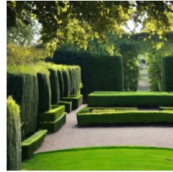



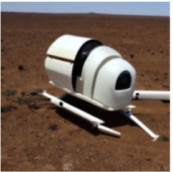



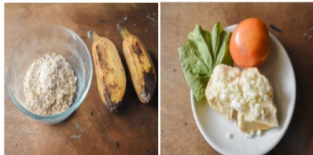




# Main Results of Interleaved Generation

- Our approach is **highly effective and efficient** when it is adapted to existing interleaved generation models.

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%)

# Qualitative Results

- Our approach demonstrates better **helpfulness** and **instruction-following** capabilities.
- Text quality is commonly poor in previous baselines.
- Baselines' generated images are not consistent with inputs.

Input	MiniGPT-5	GILL	Emu2	Emu2+MoSS (Ours)
<p><b>Instruction:</b> Complete a travel story. <b>Context:</b> I visited the countryside for the weekend &lt;IMG&gt;.</p> 	<p><b>Output:</b> &lt;IMG&gt;</p> 	<p><b>Output:</b> the gardens &lt;IMG&gt;</p> 	<p><b>Output:</b> I was so excited to visit the country side for the weekend &lt;IMG&gt;. I was so excited to visit the country side for the weekend.</p> 	<p><b>Output:</b> I went to the countryside and saw many beautiful flowers &lt;IMG&gt;. I also saw many beautiful buildings &lt;IMG&gt;.</p> 
<p><b>Instruction:</b> Describe a team's success in University Rover Challenge. <b>Context:</b> During the challenge, the PCz Team placed the second &lt;IMG&gt;.</p> 	<p><b>Output:</b> &lt;IMG&gt;</p> 	<p><b>Output:</b> The team's rover is an autonomous vehicle that can across the surface of the moon. &lt;IMG&gt;</p> 	<p><b>Output:</b> Describe a team's success in University Rover Challenge &lt;IMG&gt;.</p> 	<p><b>Output:</b> 16 teams from 10 countries competed in the University Rover Challenge. &lt;IMG&gt;</p> 
<p><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 &lt;IMG&gt; and a tofu scramble for breakfast in the morning &lt;IMG&gt;.</p> 	<p><b>Output:</b> After you eat breakfast make sure to have enough food for dinner &lt;IMG&gt;.</p> 	<p><b>Output:</b> the tofu scramble &lt;IMG&gt;</p> 	<p><b>Output:</b> Eat vegan cereal with a non-dairy milk for a quick breakfast &lt;IMG&gt;.</p> 	<p><b>Output:</b> Have a good vegan meal for lunch &lt;IMG&gt;.</p> 

# Conclusion

- We introduce MoSS, a novel framework that efficiently adapts VLGs by using **modality-specialized** architectures and parameters for each modality.
- We created LeafInstruct, the first open-sourced large-scale **interleaved instruction tuning** dataset with 184K high-quality instances spanning more than 10 domains.
- MoSS achieves state-of-the-art performance, significantly surpassing baseline VLGs in complex interleaved generation tasks.