Instruct-Imagen: Imagen Generation with Multi-modal Instruction

Hexiang Hu, Kelvin C.K. Chan*, Yu-Chuan Su*, Wenhu Chen**
Yandong Li, Kihyuk Sohn, Yang Zhao, Xue Ben, Boqing Gong, William W. Cohen, Ming-Wei Chang, Xuhui Jia

Highlight

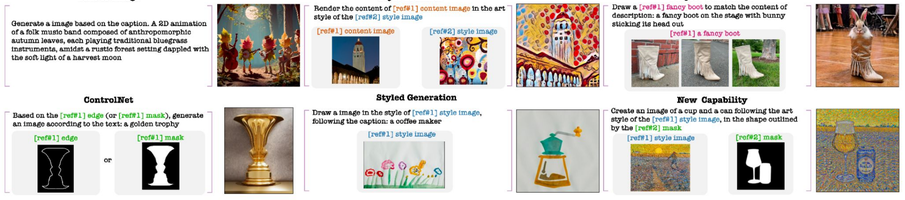
- For the first time, we show that ***multi-modal instruction tuning*** *can* make imagen models to generalize to a large variety of generation tasks.

Summary of Contributions

- We introduce ***multi-modal instruction*** for image generation, which unifies tasks that inputs different conditions (e.g. *Control2Img, Subject2Img, Style2Img*, etc.)
- We proposed a ***two-stage fine-tuning framework*** to adapt pre-trained text2image models to accept multimodal instruction.
- **Instruct-Imagen** performs in-domain tasks on par with SoTA, and shows strong ***compositional generalization*** to novel generation tasks.

Unifying Generation Tasks with *Multi-Modal Instruction*

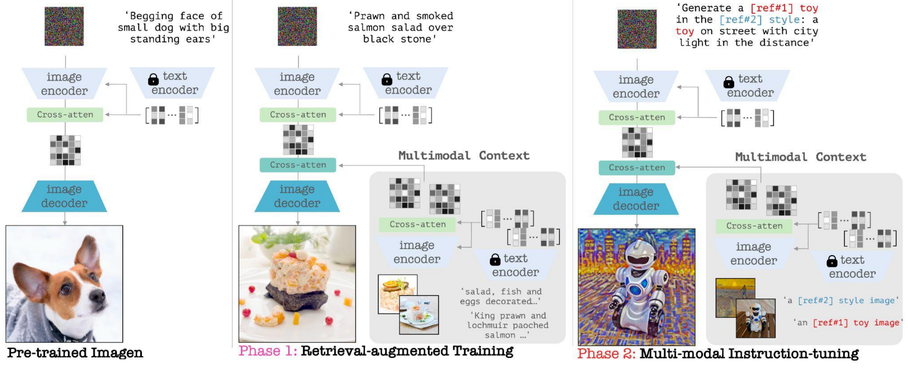
We cast existing image generation tasks in the format of multi-modal instruction



Two-Stage Multi-Modal Instruction-Tuning

S1. Retrieval-augmented continue training, **S2**. Multi-modal instruction-tuning

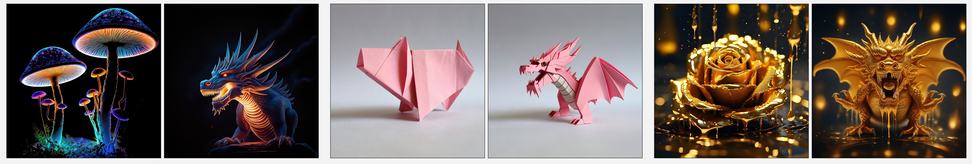
(stage 1 adapts the pre-trained model, and stage 2 fine-tunes the model with target tasks)



Instruct-Imagen Gallery

Subject-Driven Generation

Style-Driven Generation



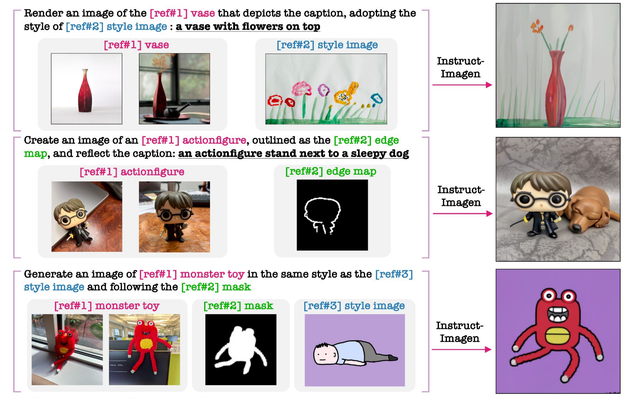
Mask-Based Editing

Mask-Free Editing

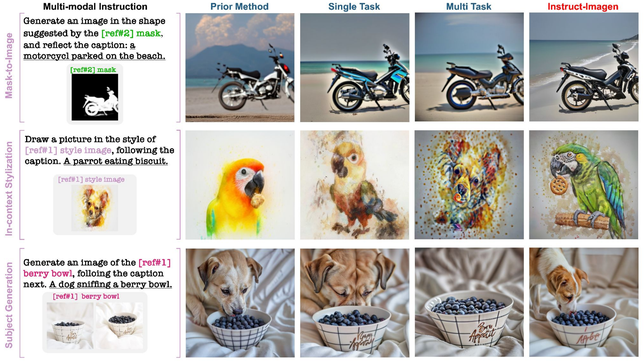
Multi-Turn Editing

Model Evaluation

Zero-Shot Compositional Generalization



In-Domain Tasks Comparison



Human Eval on In-Domain & Zero-Shot Tasks