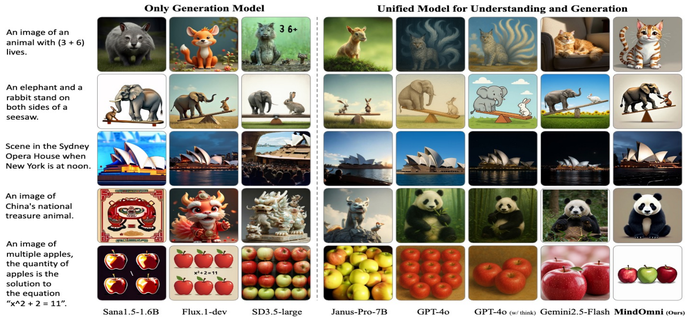
MindOmni: Unleashing Reasoning Generation in Vision Language Models with RGPO

Yicheng Xiao[1,2], Lin Song[2] *, Yukang Chen[3], Yingmin Luo[2], Yuxin Chen[2], Yukang Gan[2], Wei Huang[4], Xiu Li[1], Xiaojuan Qi[4] and Ying Shan[2]
[1]Tsinghua University; [2]ARC Lab, Tencent PCG; [3]The Chinese University of Hong Kong; [4]The University of Hong Kong
*Project Lead. Corresponding authors.

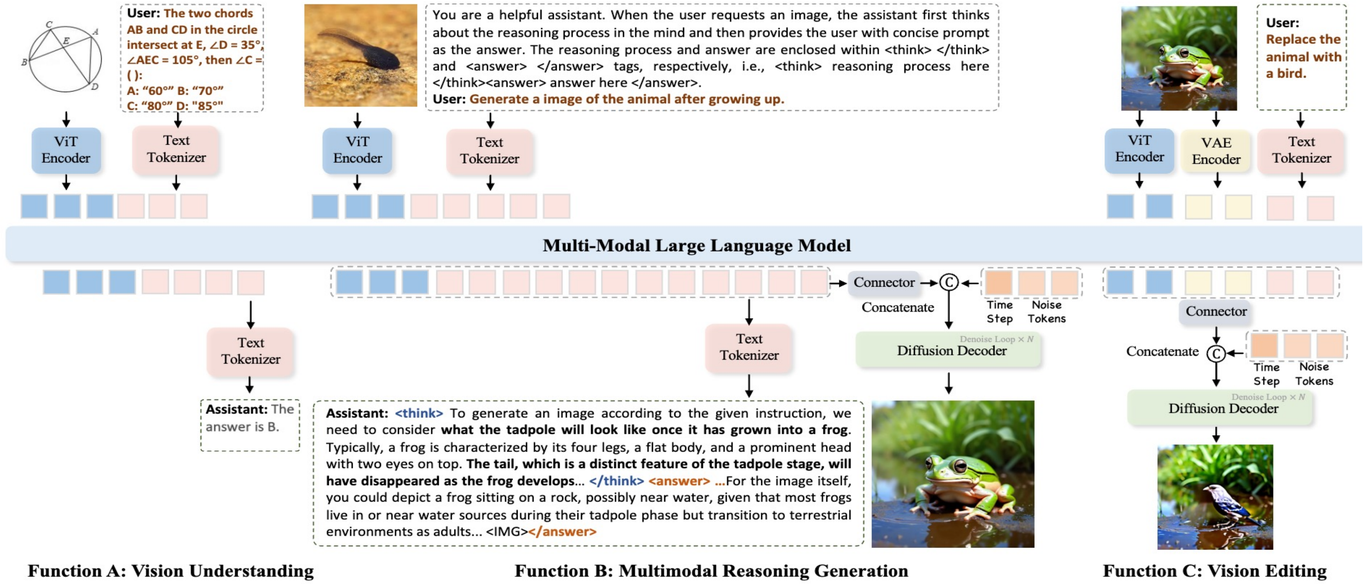
Motivation

- Recent text-to-image systems face limitations in **handling multimodal inputs** like images or audio and explicit inferences for **complex reasoning tasks**.
- The observation about the success of Deepseek-R1 prompts a fresh perspective: Can **reinforcement learning** be leveraged to **unleash the reasoning generation capabilities** of **unified Vision Language Model**

Method	X2Image Generation	Unified Und. & Gen.	End-to-End Pipeline	Fine-grained Image Editing	Explicit CoT	RL Augmented
Janus-Pro [5]	✗	✔	✔	✗	✗	✗
MetaMorph [41]	✗	✔	✔	✗	✗	✗
GoT [9]	✔	✗	✗	✔	✔	✗
MindOmni	✔	✔	✔	✔	✔	✔

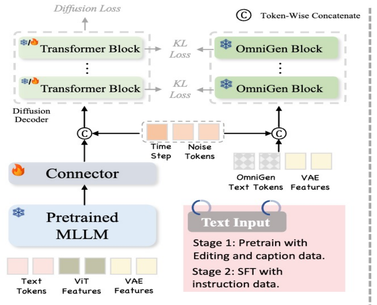


Inference Pipeline of MindOmni

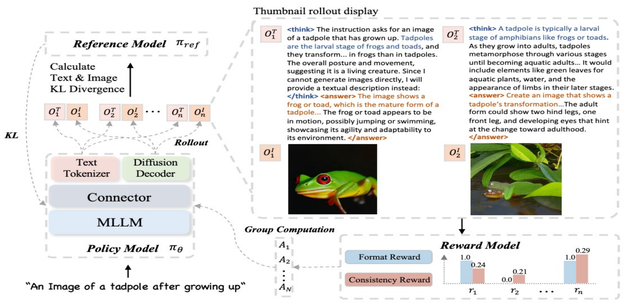


Training Pipeline & Results

Overview of Training Pipeline

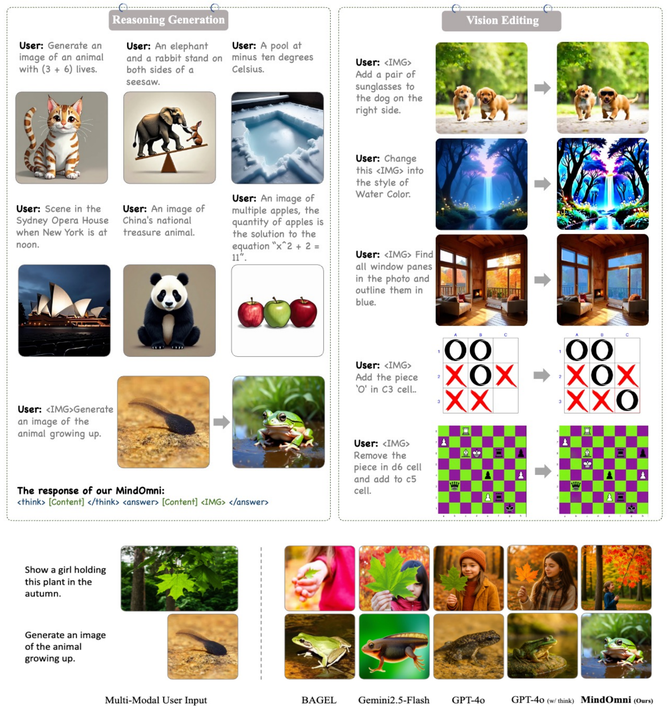


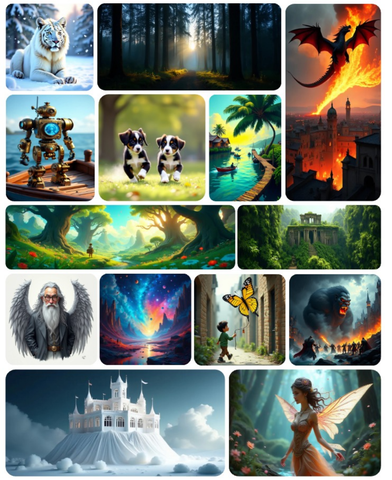
(a) Stage 1 & Stage 2: Supervised Tuning



(b) Stage 3: Reinforcement Learning with RGPO

Qualitative Results





Code

Paper