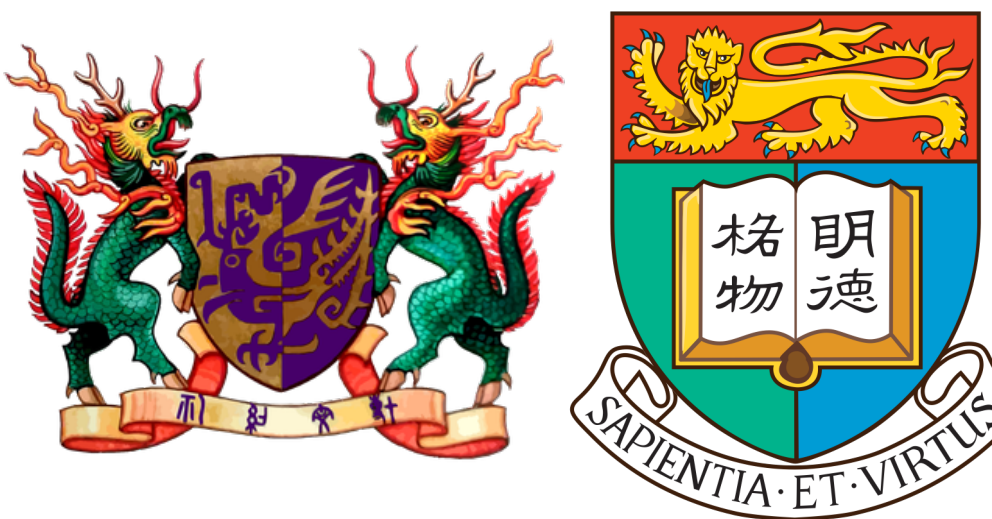




MindOmni: Unleashing Reasoning Generation in Vision Language Models with RGPO

Yicheng Xiao^{1,2}, Lin Song² * ✉, Yukang Chen³, Yingmin Luo², Yuxin Chen², Yukang Gan², Wei Huang⁴, Xiu Li¹ ✉, Xiaojuan Qi⁴ and Ying Shan²
¹Tsinghua University; ²ARC Lab, Tencent PCG; ³The Chinese University of Hong Kong; ⁴The University of Hong Kong
 *Project Lead. ✉Corresponding authors.



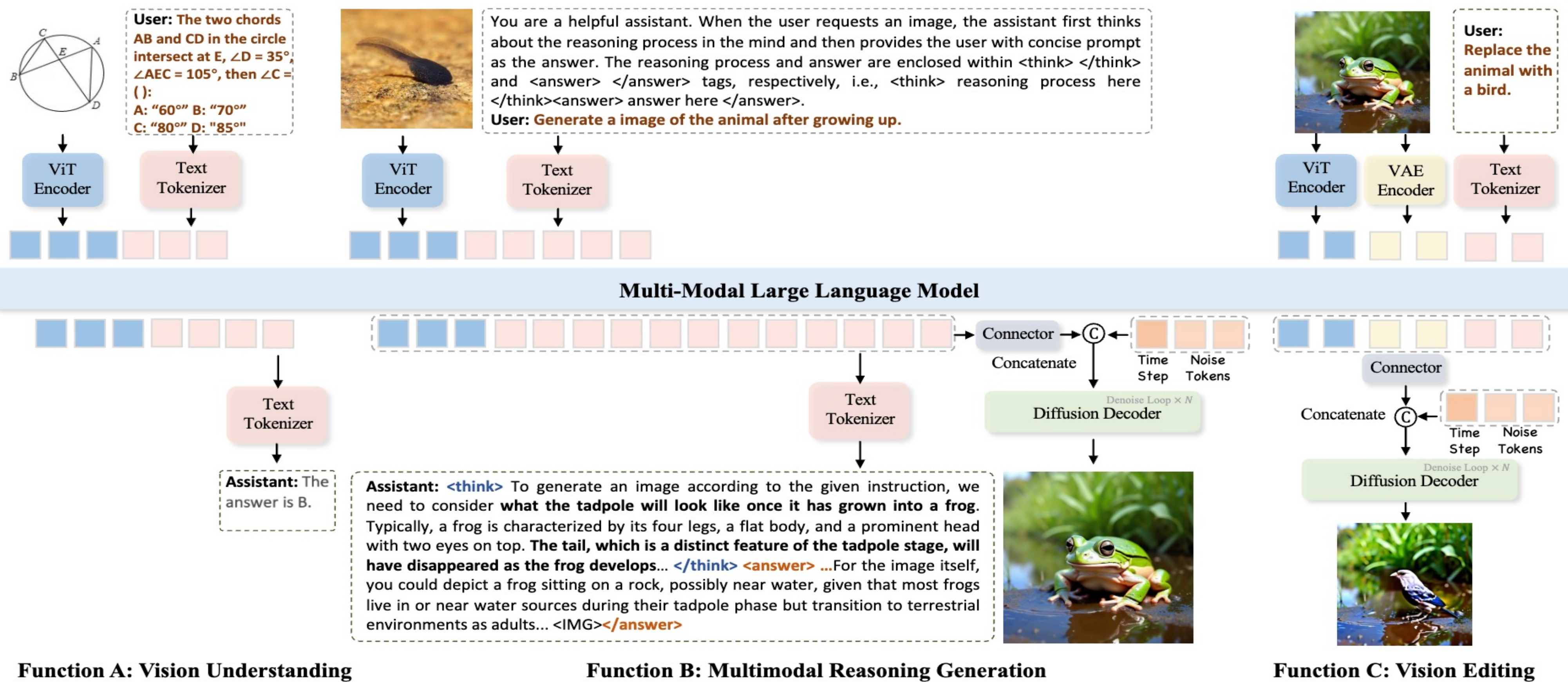
Motivation

- 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	✓	✓	✓	✓	✓	✓

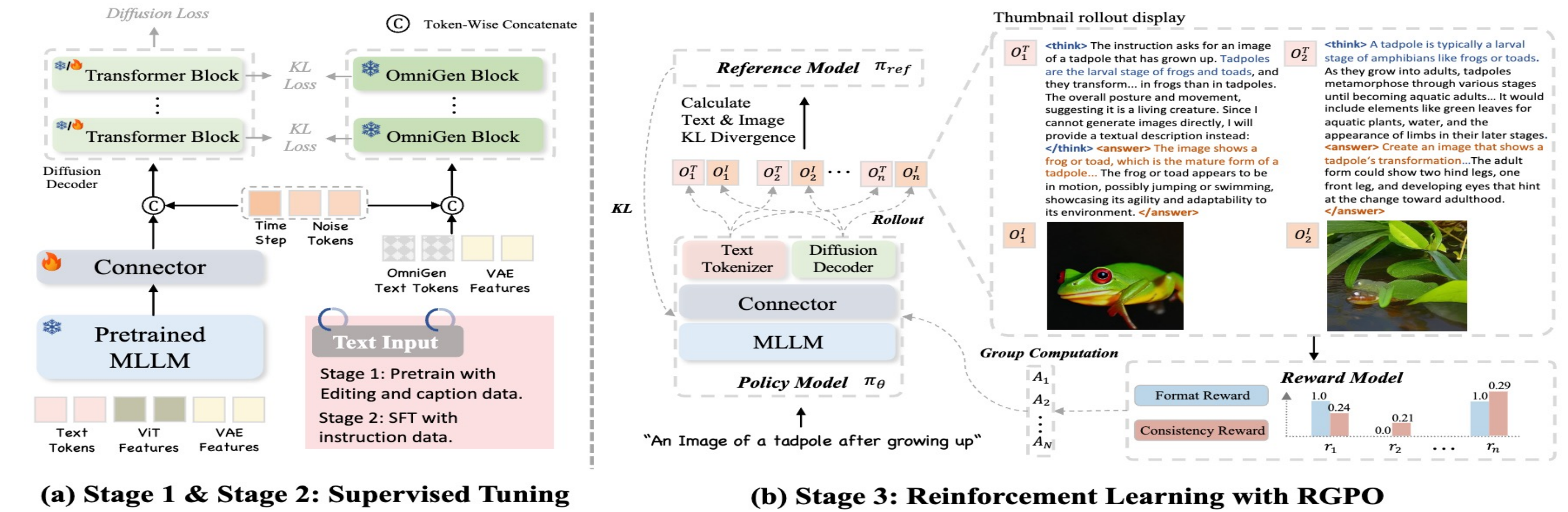
	Only Generation Model			Unified Model for Understanding and Generation						
An image of an animal with (3 + 6) lives.										
An elephant and a rabbit stand on both sides of a seesaw.										
Scene in the Sydney Opera House when New York is at noon.										
An image of China's national treasure animal.										
An image of multiple apples, the quantity of apples is the solution to the equation "x*2 + 2 = 11".										
	Sanat-1.5-1.6B	Flux.1-dev	SD3.5-large	Janus-Pro-7B	GPT-4o	GPT-4o (w/ think)	Gemini2.5-Flash	MindOmni (ours)		

Inference Pipeline of MindOmni



Training Pipeline & Results

Overview of Training Pipeline



Qualitative Results

