HO CHI MINH UNIVERSITY OF TECHNOLOGY
FACULTY OF COMPUTER SCIENCE AND ENGINEERING

CAPSTONE PROJECT PRESENTATION

IMPROVING TEXT-TO-IMAGE MODELS WITH ARTIFICIAL INTELLIGENCE FEEDBACK

Student: Vo Manh Quyen (2014318)
Supervisors: Nguyen Quang Duc, Nguyen Duc Dung

QR

Introduction

We propose an innovative approach to improve text-to-image generation using multimodal large language models (MLLMs). In stage 1, In stage 1, we enhance user input prompts and implement a robust evaluation framework to ensure high-quality, aligned image generation with MLLMs. In stage 2, we fine-tune the model with active learning, using simulated user feedback to guide the process. The results show improvements in generating images that better align with user preferences, highlighting the potential of our approach to advance text-to-image models.

Problems & Motivation

Text-to-image models struggle with simplistic prompts, hindering their ability to understand user requests comprehensively.

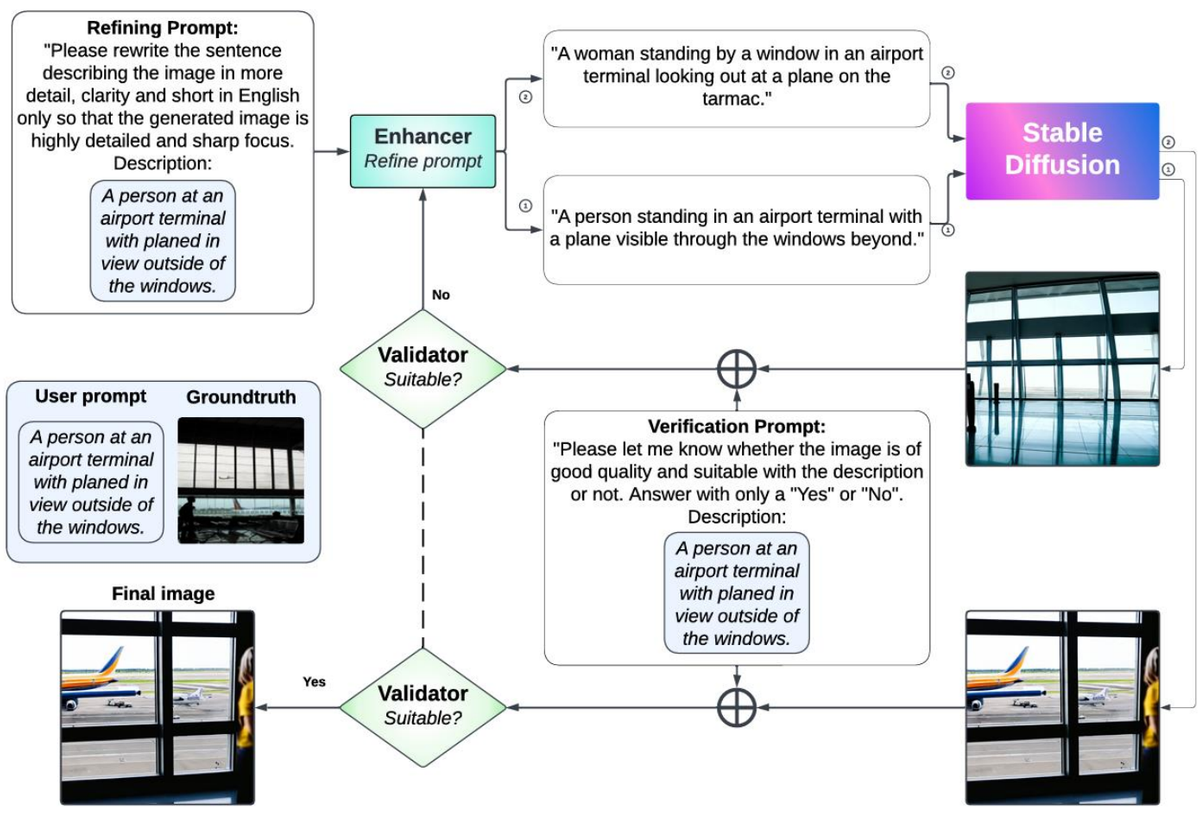
"A yellow dog" → Stable Diffusion 1.5

Motivated by the advancements in MLLMs, we aim to enhance prompt understanding and improve model performance. MLLMs offer improved text interpretation, enabling the generation of more detailed prompts and better aligning images with user preferences.

Method

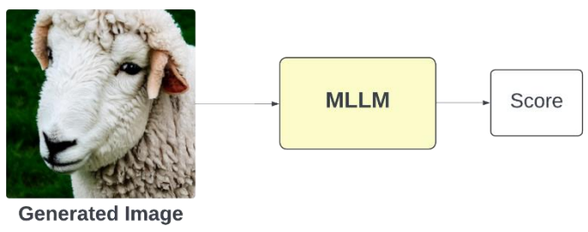
Stage 1: Improving text-to-image generation with MLLMs

With enhancer and validator as MLLMs:



Stage 2: Fine-tuning text-to-image models with active learning

Use MLLM to score generated image:



These scores serve as rewards, guiding the fine-tuning process through DDPO
Consider the objective function:

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z}),\mathbf{x}_0\sim p_\theta(\mathbf{x}_0|\mathbf{z})}\left[r(\mathbf{x}_0,\mathbf{z})\right]$$

Running gradient descent = policy gradient:

$$\nabla_\theta \mathcal{L}(\theta) = \mathbb{E}\left[\sum_{t=0}^{T} \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{z})}{p_{\theta_{\text{old}}}(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{z})} \nabla_\theta \log p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{z}) r(\mathbf{x}_0,\mathbf{z})\right]$$

Markov Decision Process formulation:

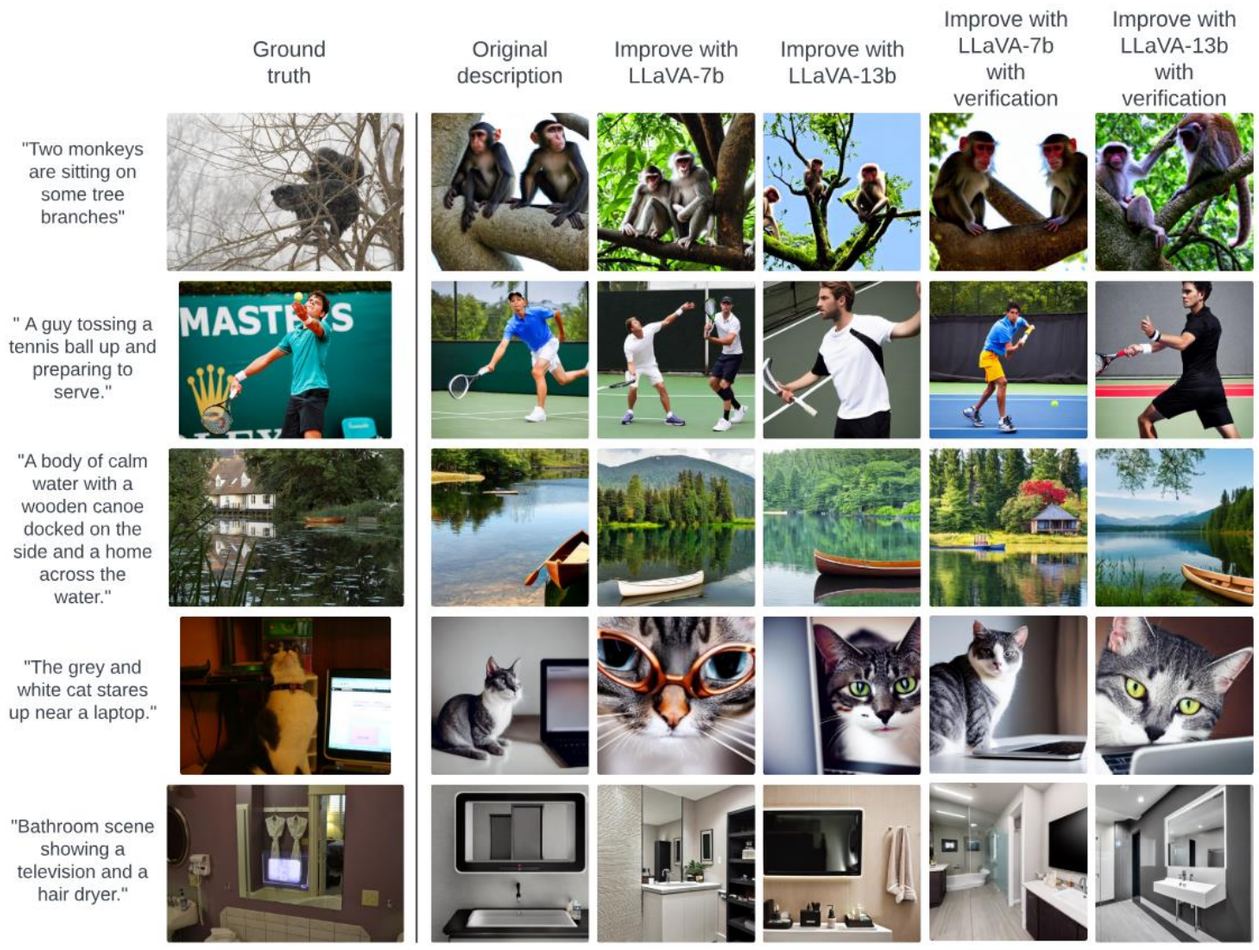
$$\mathbf{s}_t = (\mathbf{z}, \mathbf{x}_{T-t}), \mathbf{a}_t = \mathbf{x}_{T-t-1},$$
$$P_0(\mathbf{s}_0) = (p(\mathbf{z}), \mathcal{N}(\mathbf{0},\mathbf{I})), P(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t) = (\delta_{\mathbf{z}}, \delta_{\mathbf{a}_t}),$$
$$R(\mathbf{s}_t,\mathbf{a}_t) = \begin{cases} r(\mathbf{s}_{t+1}) = r(\mathbf{x}_0,\mathbf{z}) & \text{if } t = T-1, \\ 0 & \text{otherwise.} \end{cases}$$
$$\pi_\theta(\mathbf{a}_t \mid \mathbf{s}_t) = p_\theta(\mathbf{x}_{T-t-1} \mid \mathbf{x}_{T-t}, \mathbf{z})$$

Results

Stage 1: Improving text-to-image generation with MLLMs

Evaluation results on the MS-COCO dataset

Prompt Enhancer	Validator	FID↓	IS↑	Iteration
None	None	24.95 ± 0.51	33.70 ± 0.47	1
LLaVA1.5-7b	None	30.00 ± 0.60	31.64 ± 0.47	1
LLaVA1.5-13b	None	25.59 ± 0.51	33.46 ± 0.48	1
LLaVA1.6-7b	None	28.85 ± 0.55	35.72 ± 0.53	1
LLaVA1.6-13b	None	29.56 ± 0.56	35.22 ± 0.52	1
GPT-4	None	29.78 ± 0.55	35.63 ± 0.52	1
CogVLM	None	29.71 ± 0.58	34.42 ± 0.53	1
LLaVA1.5-7b	LLaVA1.5-7b	26.10 ± 0.51	33.94 ± 0.52	1.23 ± 0.57
LLaVA1.5-13b	LLaVA1.5-13b	$\mathbf{24.37 \pm 0.46}$	34.61 ± 0.50	2.39 ± 1.64
LLaVA1.6-7b	LLaVA1.6-7b	28.07 ± 0.53	35.80 ± 0.51	1.01 ± 0.09
LLaVA1.6-13b	LLaVA1.6-13b	28.50 ± 0.54	35.78 ± 0.50	1.27 ± 0.64
GPT-4	GPT-4	29.78 ± 0.57	$\mathbf{35.83 \pm 0.52}$	2.32 ± 1.69
CogVLM	CogVLM	29.80 ± 0.58	34.53 ± 0.51	1.10 ± 0.53

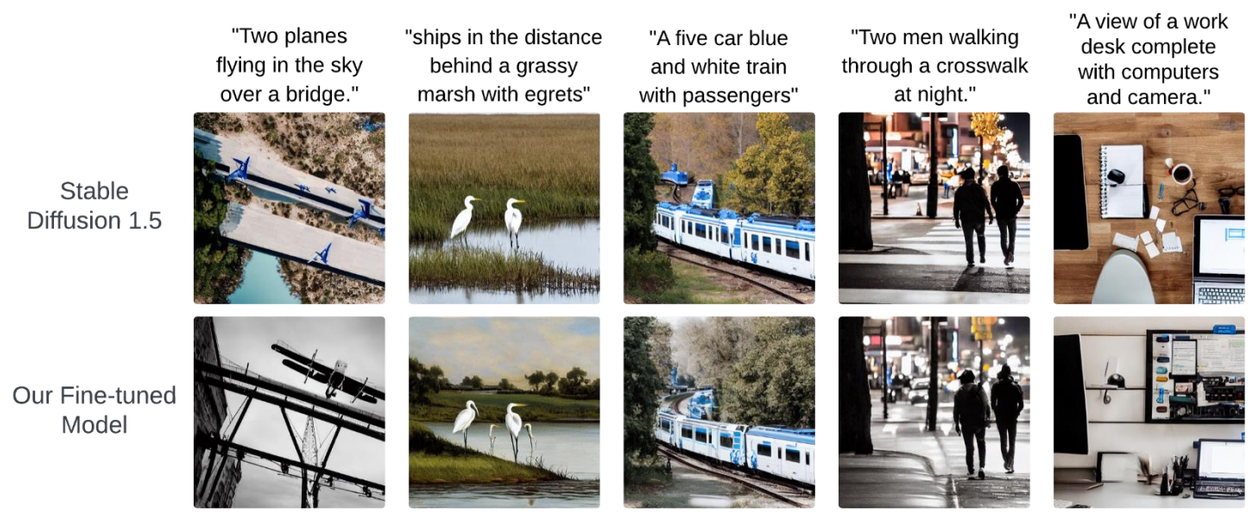


Stage 2: Fine-tuning text-to-image models with active learning

In this experiment, we assume the user prefers images that have a dark, subdued color palette with a blurred appearance, creating a somber and introspective atmosphere.

Evaluation results on the MS-COCO dataset

Model	FID ↓	Inception Score ↑	Contrast ↓	Brightness ↓
Stable Diffusion 1.5	30.94	$\mathbf{32.53 \pm 1.49}$	73.83	123.87
Our Fine-tuned Model	**29.28**	32.19 ± 1.39	**71.60**	**123.78**



Conclusion

Both of our approaches have showcased improvements in text-to-image generation. In stage 1, using MLLMs to enhance prompts, coupled with an evaluation process, has helped improve the quality and alignment of the generated image with the original prompt. In stage 2, we fine-tuned text-to-image models using active learning, focusing on aesthetic preferences. This approach enhanced visual quality while preserving performance. Our work demonstrates the potential of MLLMs in enhancing text-to-image generation to achieve higher fidelity and user satisfaction.