香港大學 THE UNIVERSITY OF HONG KONG

香港中文大學 The Chinese University of Hong Kong

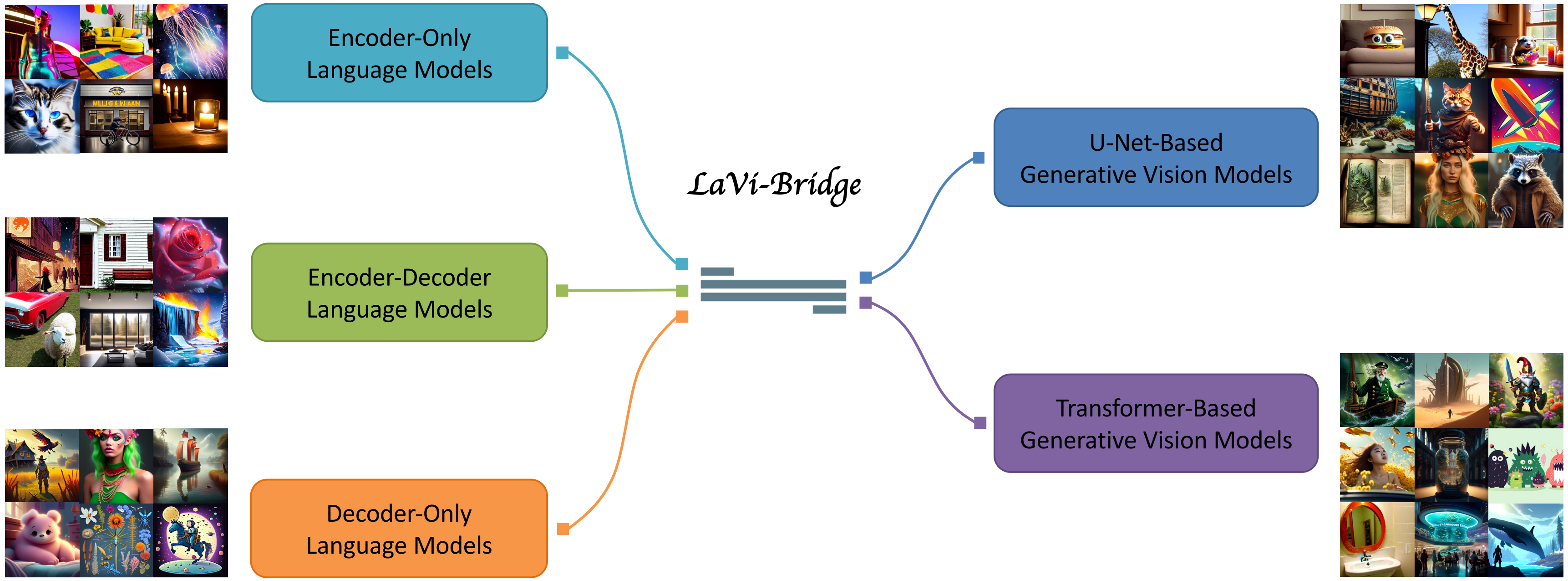
香港科技大學

Bridging Different Language Models and Generative Vision Models for Text-to-Image Generation

Shihao Zhao, Shaozhe Hao, Bojia Zi, Huaizhe Xu, Kwan-Yee K. Wong

Introduction - Overview

Motivation: As language and vision models progress, there is potential to improve diffusion models with advanced components, with a key goal being the integration of diverse language and vision models.


Encoder-Only Language Models
Encoder-Decoder Language Models
Decoder-Only Language Models
LaVi-Bridge
U-Net-Based Generative Vision Models
Transformer-Based Generative Vision Models

LaVi-Bridge enables the integration of unrelated pre-trained language and vision models for text-to-image generation. It provides a flexible, plug-and-play solution without modifying the original model weights.

Introduction - Contribution

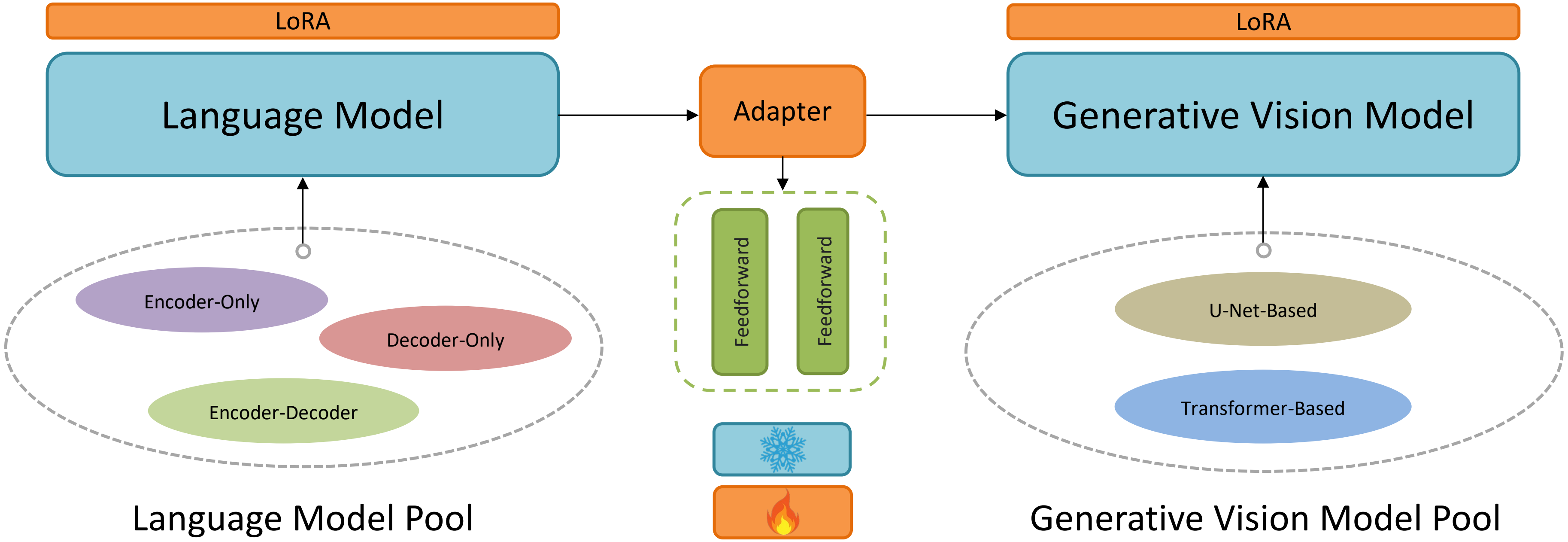
1. LaVi-Bridge can connect various pre-trained language models (**encoder-only, encoder-decoder, decoder-only**) and generative vision models (**U-Net-based and Transformer-based**).

2. LaVi-Bridge utilizes **LoRA and adapters**, eliminating the need to modify the original weights of the models. It is more flexible and requires relatively small computing resources.

3. We concluded that integrating **superior models enhances performance in their modalities**. For example, the diffusion model using Llama-2 shows excellent semantic understanding, while the one using the transformer in PixArt produces more aesthetically pleasing images.

Method - Pipeline

We select one model each from the language and vision pools, **freeze them**, and **incorporate LoRA into both**. The connection between the models is **established via an adapter**, and we only train the weights introduced by LoRA and the adapter.



Method - Language and Vision Alignment

1. **Freeze** the language model f^{θ_1} and the vision model g^{θ_2}.

2. Inject **LoRA $\Delta\boldsymbol{\theta}$** into both language and vision models: $f^{\theta_1+\Delta\theta_1}$ and $g^{\theta_2+\Delta\theta_2}$.

3. Use an **adapter $\boldsymbol{h}$** to connect the language model and vision model through cross-attention layers.

$$c = f(y),$$

$$Q = W_q(z), K = W_k(c), V = W_v(c),$$

$$CrossAttention(Q, K, V) = softmax(Q \cdot K^T) \cdot V,$$

⇒

$$c = f^{\theta_1+\Delta\theta_1}(y),$$

$$Q = W_q^{\theta_2+\Delta\theta_2}(z), K = {W_k}^{\theta_2+\Delta\theta_2}(h(c)), V = W_v^{\theta_2+\Delta\theta_2}(h(c)),$$

$$CrossAttention(Q, K, V) = softmax(Q \cdot K^T) \cdot V.$$

Experiment - Settings

Training involved 1 million text-image pairs: 600k from the COCO2017 trainset and 400k from an internal dataset. We used AdamW with a learning rate of 1×10^{-4} and trained for 50k steps.

We conducted evaluations on different combinations of language and vision models:

- Short Prompts: Subset of COCO2014. FID and aesthetic score (vision), CLIP score (text).
- Long Prompts: Use Llama-2 to expand the short prompts. Aesthetic score (vision), CLIP score (text).
- Compositional Prompts: Compbench. Metrics in Compbench (text).

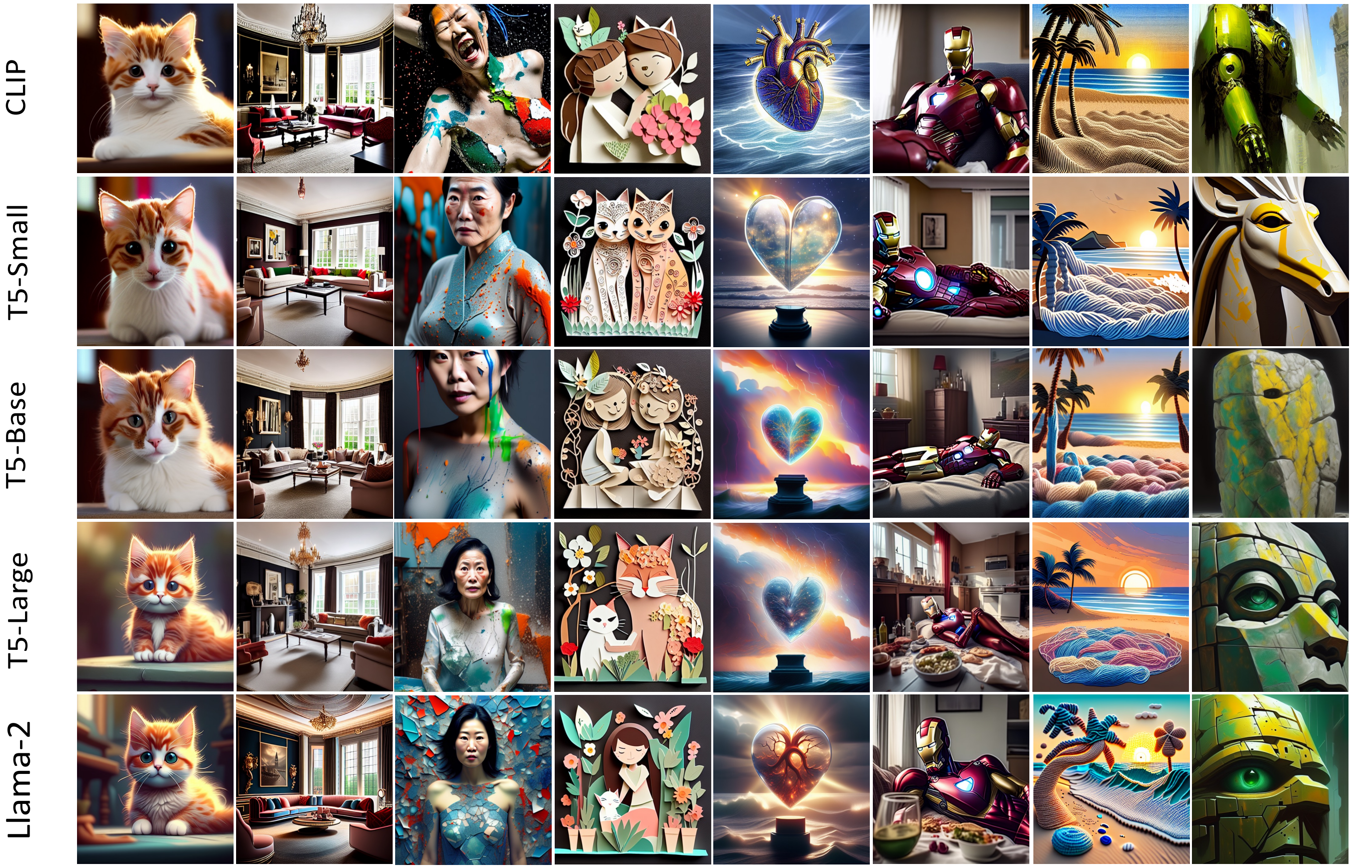
Experiment - Evaluation

A cute cat.

A middle-aged woman of Asian descent, her dark hair streaked with silver, appears fractured and splintered, intricately embedded within a sea of broken porcelain. The porcelain glistens with splatter paint patterns in a harmonious blend of glossy and matte blues, greens, oranges, and reds, capturing her dance in a surreal juxtaposition of movement and stillness. Her skin tone, a light hue like the porcelain, adds an almost mystical quality to her form.

An illustration of a human heart made of translucent glass, standing on a pedestal amidst a stormy sea. Rays of sunlight pierce the clouds, illuminating the heart, revealing a tiny universe within.

Digital illustration of a beach scene crafted from yarn. The sandy beach is depicted with beige yarn, waves are made of blue and white yarn crashing onto the shore. A yarn sun sets on the horizon, casting a warm glow. Yarn palm trees sway gently, and little yarn seashells dot the shoreline.



London luxurious interior living-room.

A paper craft art depicting a girl giving her cat a gentle hug. Both sit amidst potted plants, with the cat purring contentedly while the girl smiles. The scene is adorned with handcrafted paper flowers and leaves.

Marvel movie character, iron man, dress up to match movie character, full body photo, American apartment, lying down, life in distress, messy, lost hope, food, wine, hd, 8k, real, reality, super detail, 8k post photo manipulation, real photo.

An ancient stone Colossus with eye, Stephan Martinière, dark yellow and light emerald, color zone painting, Denis Sarazhin, dark emerald and silver, robotic expressionism, high detail.

Evaluation on Different Language Models:

Vision Model: U-Net of Stable Diffusion V1.4.

	CLIP	T5-Small	T5-Base	T5-Large	Llama-2
Short - FID	23.57	22.98	22.62	23.11	**21.80**
Short - Aesthetics	5.609	5.813	**5.888**	5.881	5.883
Short - CLIP Score	0.3102	0.3122	0.3149	0.3156	**0.3172**
Long - Aesthetics	6.003	6.206	6.284	6.305	**6.355**
Long - CLIP Score	0.3120	0.3111	0.3179	0.3193	**0.3231**
Comp - Color	0.3578	0.3368	0.3856	0.3889	**0.4859**
Comp - Shape	0.3752	0.2962	0.3266	0.3552	**0.4285**
Comp - Texture	0.4506	0.3728	0.4132	0.4524	**0.5055**
Comp - Spatial	0.1296	0.1456	0.1569	0.1582	**0.1914**
Comp - Non-Spatial	0.3009	0.2984	0.3054	0.3068	**0.3106**
Comp - Complex	0.2985	0.2728	0.3055	0.3072	**0.3094**

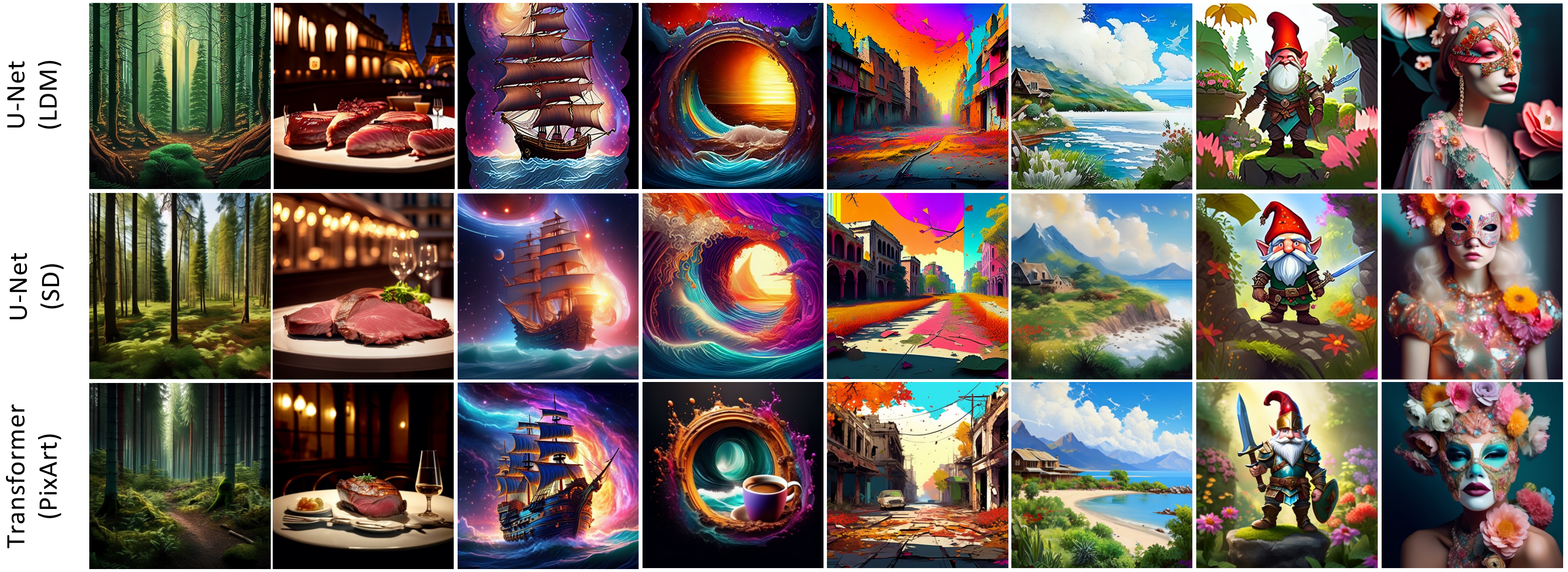
Experiment - Evaluation

Forest.

Pirate ship trapped in a cosmic maelstrom nebula, rendered in cosmic beach whirlpool engine, volumetric lighting, spectacular, ambient lights, light pollution, cinematic atmosphere, art nouveau style, illustration art artwork by SenseiJaye, intricate detail.

Abandoned city with ruined buildings, long deserted streets, cars aged by time, trees, flowers, scattered leaves, empty street, vibrant colors, lineart.

A fierce garden gnome warrior, clad in armor crafted from leaves and bark, brandishes a tiny sword and shield. He stands valiantly on a rock amidst a blooming garden, surrounded by colorful flowers and towering plants. A determined expression is painted on his face, ready to defend his garden kingdom.



Smooth meat table, restaurant, Paris, elegant, lights.

A swirling, multicolored portal emerges from the depths of an ocean of coffee, with waves of the rich liquid gently rippling outward. The portal engulfs a coffee cup, which serves as a gateway to a fantastical dimension. The surrounding digital art landscape reflects the colors of the portal, creating an alluring scene of endless possibilities.

A natural landscape painting with white clouds floating in the blue sky. There are several mountains below with some plants growing on the mountains. There is a sea below the mountains. There is a house made of stone and wood on the shore. There are many green plants next to the house.

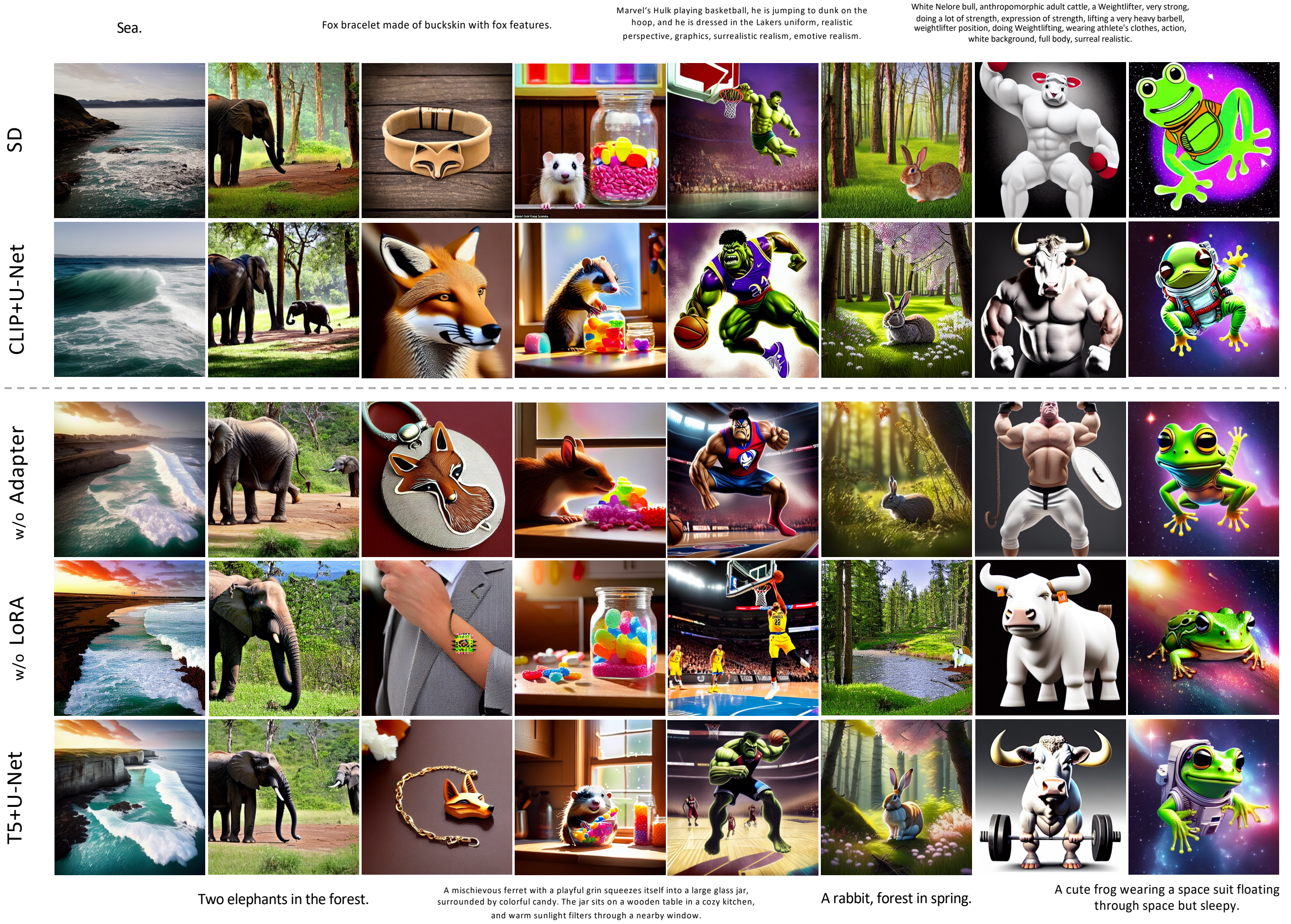
Portrait photography, a woman in a glamorous makeup, wearing a mask with tassels, in the style of midsommar by Ari Aster, made of flowers, bright pastel colors, prime lense.

	U-Net(LDM)	U-Net(SD)	Transformer(PixArt)
Short - FID	25.94	23.11	**23.02**
Short - Aesthetics	5.703	5.881	**6.145**
Short - CLIP Score	0.3126	0.3156	**0.3172**
Long - Aesthetics	6.122	6.305	**6.406**
Long - CLIP Score	0.3189	0.3193	**0.3210**
Comp - Color	**0.4099**	0.3889	0.3689
Comp - Shape	**0.3724**	0.3552	0.3316
Comp - Texture	**0.5046**	0.4524	0.4553
Comp - Spatial	0.1550	0.1582	**0.1725**
Comp - Non-Spatial	0.3004	0.3068	**0.3098**
Comp - Complex	0.3060	**0.3072**	0.3014

Evaluation on Different Vision Models:

Language Model: T5-Large.

Experiment - Ablation Study


Sea.
Fox bracelet made of buckskin with fox features.
Marvel's Hulk playing basketball, he is jumping to dunk on the hoop, and he is dressed in the Lakers uniform, realistic perspective, graphics, surrealistic realism, emotive realism.
White Nelore bull, anthropomorphic adult cattle, a Weightlifter, very strong, doing a lot of strength, expression of strength, lifting a very heavy barbell, weightlifter position, doing Weightlifting, wearing athlete's clothes, action, white background, full body, surreal realistic.
SD
CLIP+U-Net
w/o Adapter
w/o LoRA
T5+U-Net
Two elephants in the forest.
A mischievous ferret with a playful grin squeezes itself into a large glass jar, surrounded by colorful candy. The jar sits on a wooden table in a cozy kitchen, and warm sunlight filters through a nearby window.
A rabbit, forest in spring.
A cute frog wearing a space suit floating through space but sleepy.

1. **Training w/o LaVi-Bridge.**
2. **Training w/o LoRA and Adapter.**

	SD	CLIP+U-Net	w/o Adapter	w/o LoRA	T5+U-Net
Short - FID	**20.32**	23.57	23.81	**22.35**	23.11
Short - Aesthetics	**5.899**	5.609	5.807	5.829	**5.881**
Short - CLIP Score	**0.3132**	0.3102	0.3147	0.3107	**0.3156**
Long - Aesthetics	**6.120**	6.003	6.131	6.273	**6.305**
Long - CLIP Score	**0.3171**	0.3120	0.3106	0.3097	**0.3193**
Comp - Color	0.3570	**0.3578**	0.3550	0.2485	**0.3889**
Comp - Shape	0.3563	**0.3752**	0.3044	0.2944	**0.3552**
Comp - Texture	0.4028	**0.4506**	0.4001	0.3190	**0.4524**
Comp - Spatial	0.1225	**0.1296**	**0.1651**	0.0956	0.1582
Comp - Non-Spatial	**0.3104**	0.3009	0.3065	0.2998	**0.3068**
Comp - Complex	**0.3042**	0.2985	0.2878	0.2687	**0.3072**

香港大學
THE UNIVERSITY OF HONG KONG

香港中文大學
The Chinese University of Hong Kong

香港科技大學

Thank you !

https://arxiv.org/abs/2403.07860

https://github.com/ShihaoZhaoZSH/LaVi-Bridge