NEURAL INFORMATION PROCESSING SYSTEMS

南京大學 NANJING UNIVERSITY

AP-Adapter: Improving Generalization of Automatic Prompts on Unseen Text-to-Image Diffusion Models

Yuchen Fu, Zhiwei Jiang, Yuliang Liu, Cong Wang,
Zexuan Deng, Zhaoling Chen, Qing Gu

State Key Laboratory for Novel Software Technology, Nanjing University, China

Contents

CONTENTS

01. **Motivation**

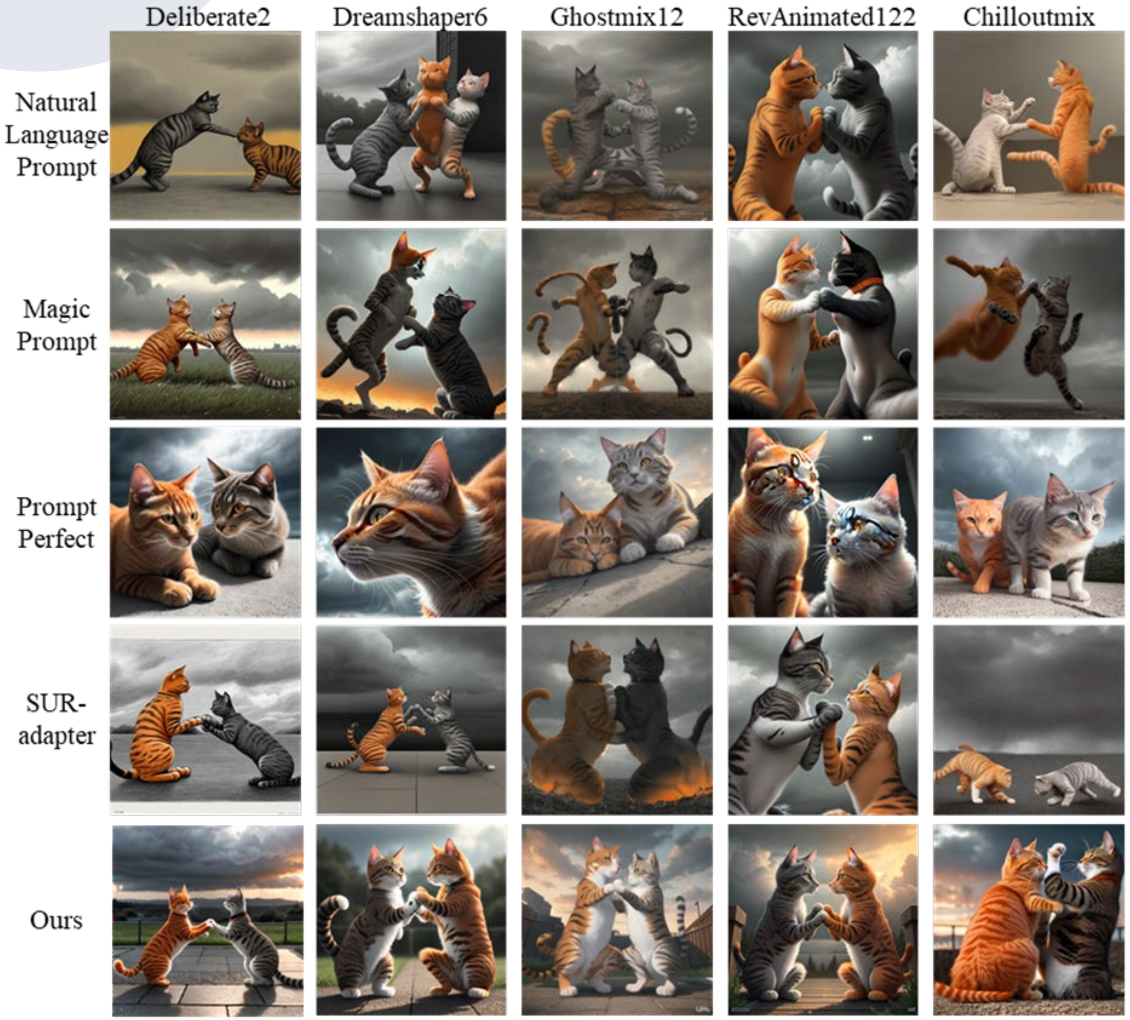
02. **Framework**

03. **Dataset**

04. **Experiments**

01 Motivation

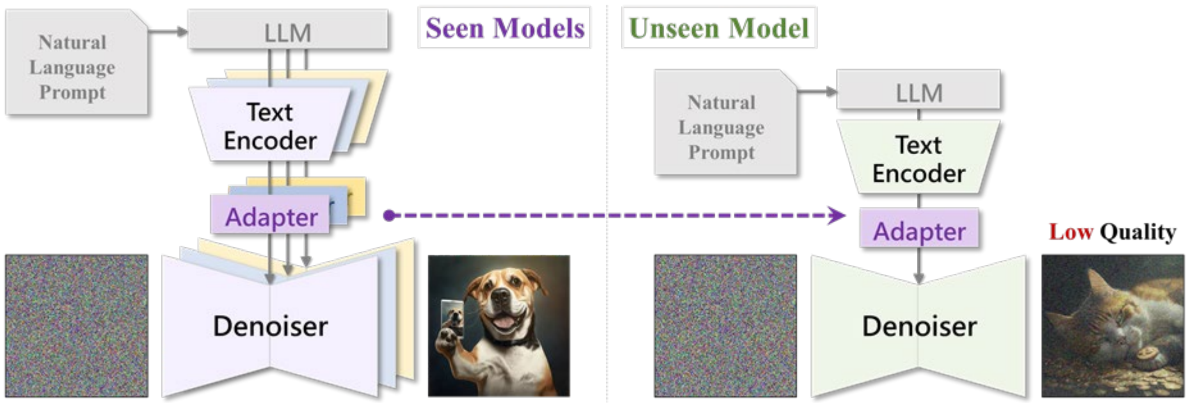
01 Motivation



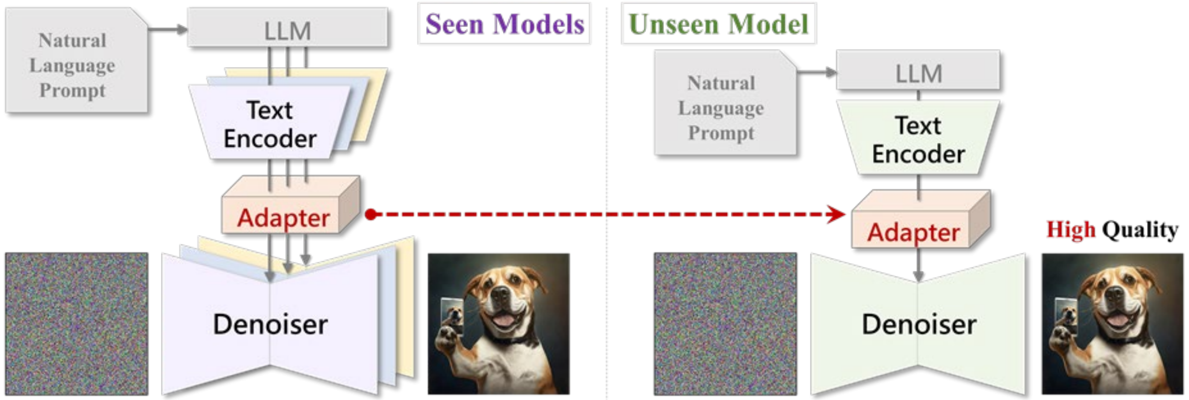
Two cats fighting against each other, with one cat being orange and the other being grey. The scene is set against a backdrop of a cloudy sky, giving the image a sense of depth and atmosphere. The style of the image is a detailed, realistic drawing.

- We explore model-generalized automatic prompt optimization (MGAPO), targeting the effectiveness of automatic prompts on unseen models.
- We propose AP-Adapter, the first method to address MGAPO.
- We collect and annotate a multi-modal, multi-domain dataset for training and evaluation.

01 Motivation



(i) Model-Specific Adapter for Prompt Representation



(ii) Model-Generalized Adapter for Prompt Representation

Model-Specific

Model-specific adapter[1] is trained specifically for one diffusion model ("one for one")

Model-Generalized

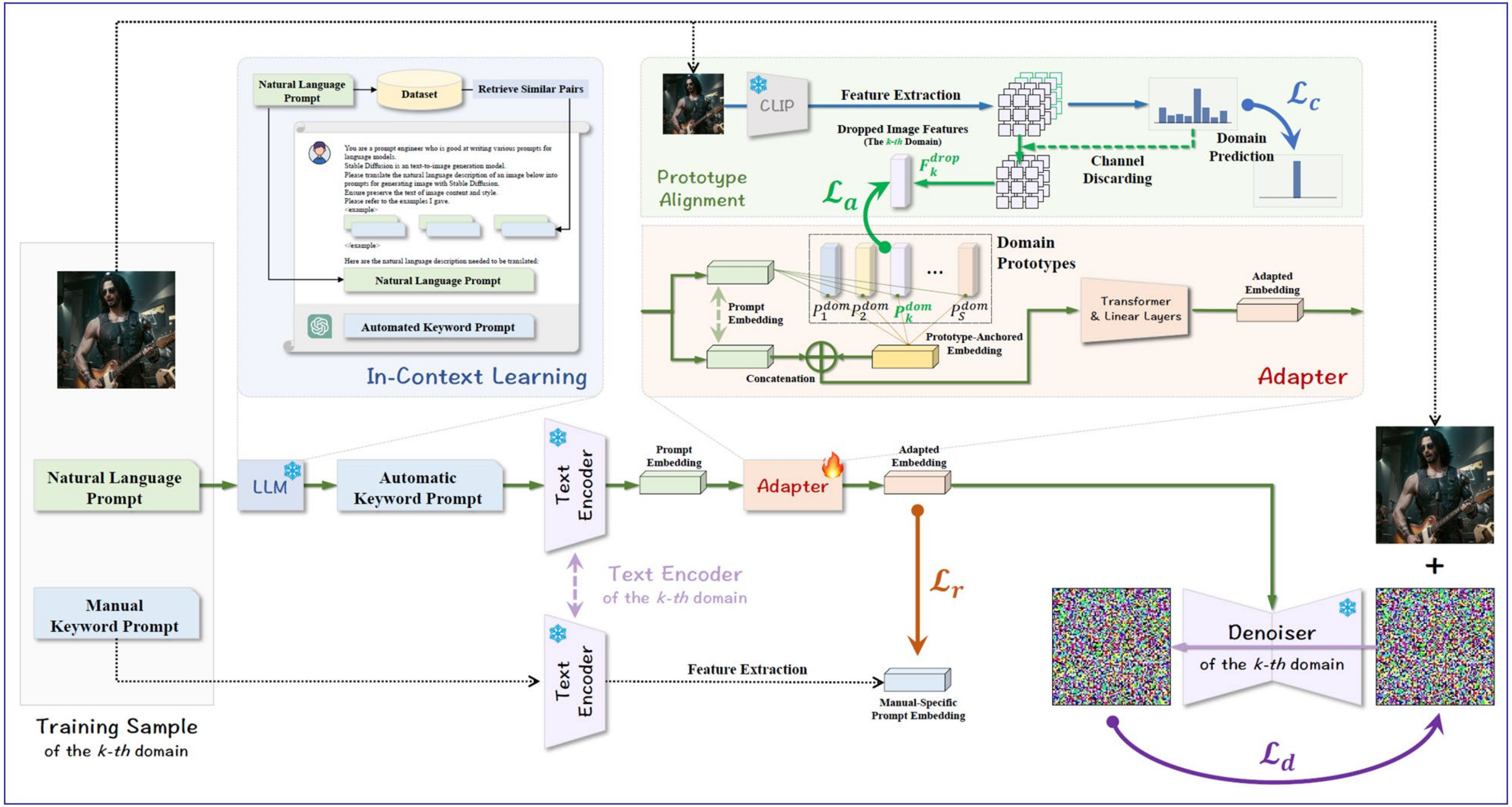
Model-generalized adapter is trained for all diffusion models ("one for all")

[1] Zhong S, Huang Z, Wen W, et al. Sur-adapter: Enhancing text-to-image pre-trained diffusion models with large language models[C]//Proceedings of the 31st ACM International Conference on Multimedia. 2023: 567-578.

02 Framework

02

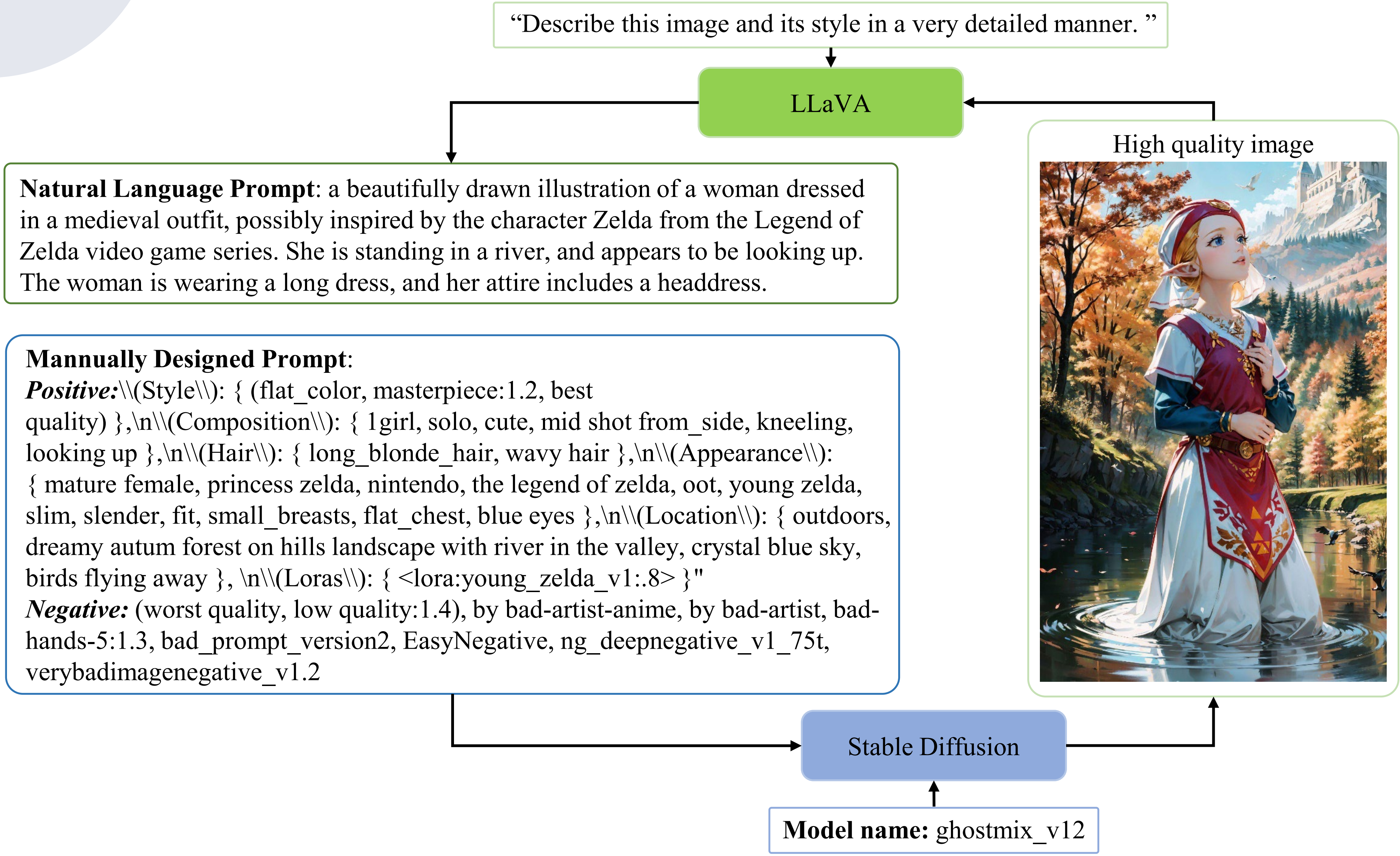
Framework



You are a prompt engineer who is good at writing various prompts for language models.
Stable Diffusion is an text-to-image generation model.
Please translate the natural language description of an image below into prompts for generating image with Stable Diffusion.
Ensure preserve the text of image content and style.
Please refer to the examples I gave.
<example>
</example>
Here are the natural language description needed to be translated:

03 Dataset Creation

03 Data Creation



“Describe this image and its style in a very detailed manner. ”

LLaVA

High quality image

Natural Language Prompt: a beautifully drawn illustration of a woman dressed in a medieval outfit, possibly inspired by the character Zelda from the Legend of Zelda video game series. She is standing in a river, and appears to be looking up. The woman is wearing a long dress, and her attire includes a headdress.

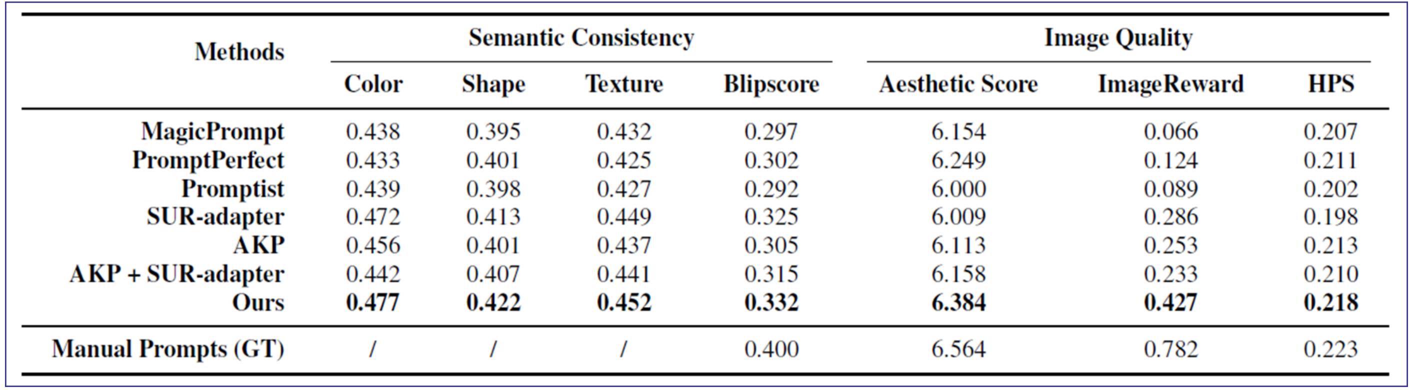
Mannually Designed Prompt:
Positive:\\(Style\\): { (flat_color, masterpiece:1.2, best quality) },\n\\(Composition\\): { 1girl, solo, cute, mid shot from_side, kneeling, looking up },\n\\(Hair\\): { long_blonde_hair, wavy hair },\n\\(Appearance\\): { mature female, princess zelda, nintendo, the legend of zelda, oot, young zelda, slim, slender, fit, small_breasts, flat_chest, blue eyes },\n\\(Location\\): { outdoors, dreamy autum forest on hills landscape with river in the valley, crystal blue sky, birds flying away }, \n\\(Loras\\): { <lora:young_zelda_v1:.8> }"
Negative: (worst quality, low quality:1.4), by bad-artist-anime, by bad-artist, bad-hands-5:1.3, bad_prompt_version2, EasyNegative, ng_deepnegative_v1_75t, verybadimagenegative_v1.2

Stable Diffusion

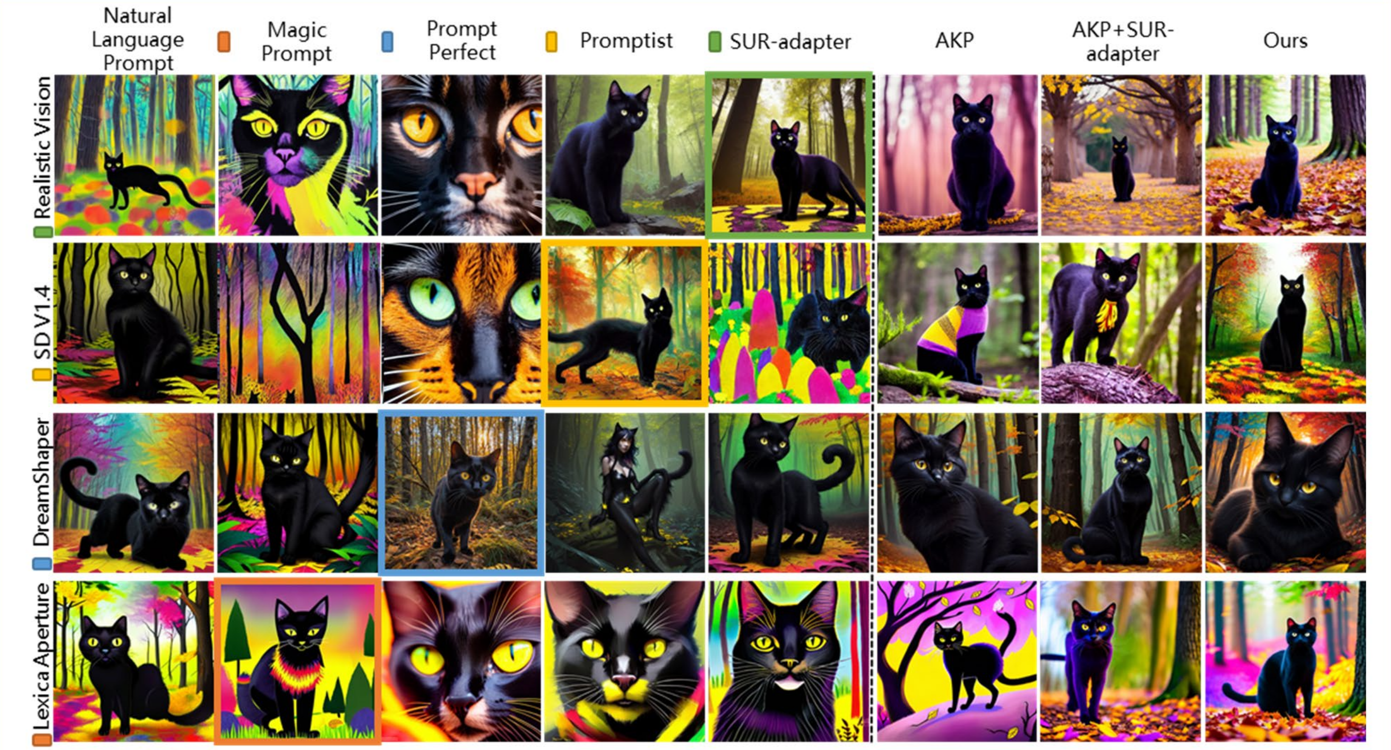
Model name: ghostmix_v12

04 Experiments

04 Experiments



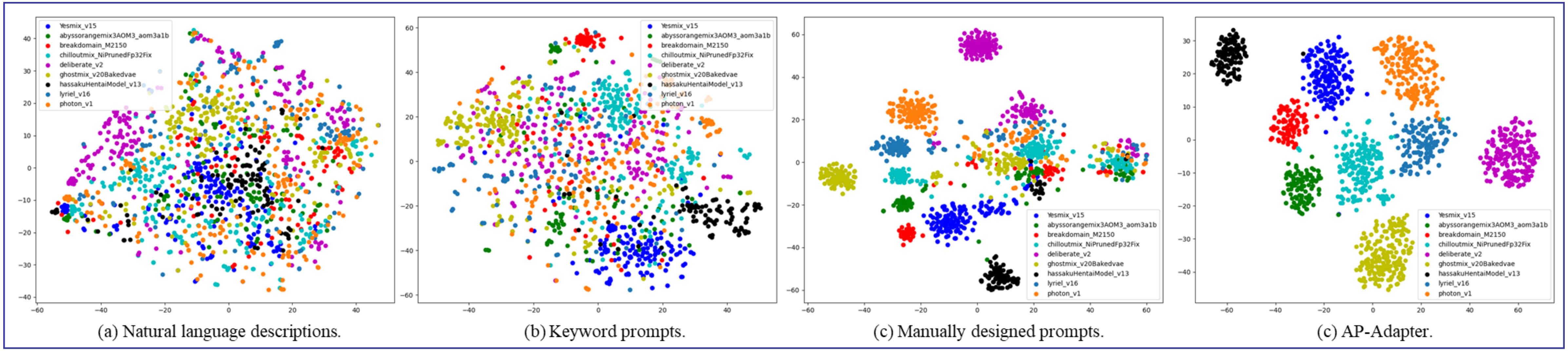
Methods	Semantic Consistency				Image Quality		
	Color	Shape	Texture	Blipscore	Aesthetic Score	ImageReward	HPS
MagicPrompt	0.438	0.395	0.432	0.297	6.154	0.066	0.207
PromptPerfect	0.433	0.401	0.425	0.302	6.249	0.124	0.211
Promptist	0.439	0.398	0.427	0.292	6.000	0.089	0.202
SUR-adapter	0.472	0.413	0.449	0.325	6.009	0.286	0.198
AKP	0.456	0.401	0.437	0.305	6.113	0.253	0.213
AKP + SUR-adapter	0.442	0.407	0.441	0.315	6.158	0.233	0.210
Ours	**0.477**	**0.422**	**0.452**	**0.332**	**6.384**	**0.427**	**0.218**
Manual Prompts (GT)	/	/	/	0.400	6.564	0.782	0.223



Natural Language Prompt: a black cat with yellow eyes, standing in a vibrant and colorful forest. The background is filled with a variety of colors, creating a visually striking and artistic atmosphere.

04 Experiments

Visualization of Conditioned Features



(a) Natural language descriptions.

(b) Keyword prompts.

(c) Manually designed prompts.

(c) AP-Adapter.

Visualization of text-conditioned domain distinctiveness.

a. Natural language description of the image.

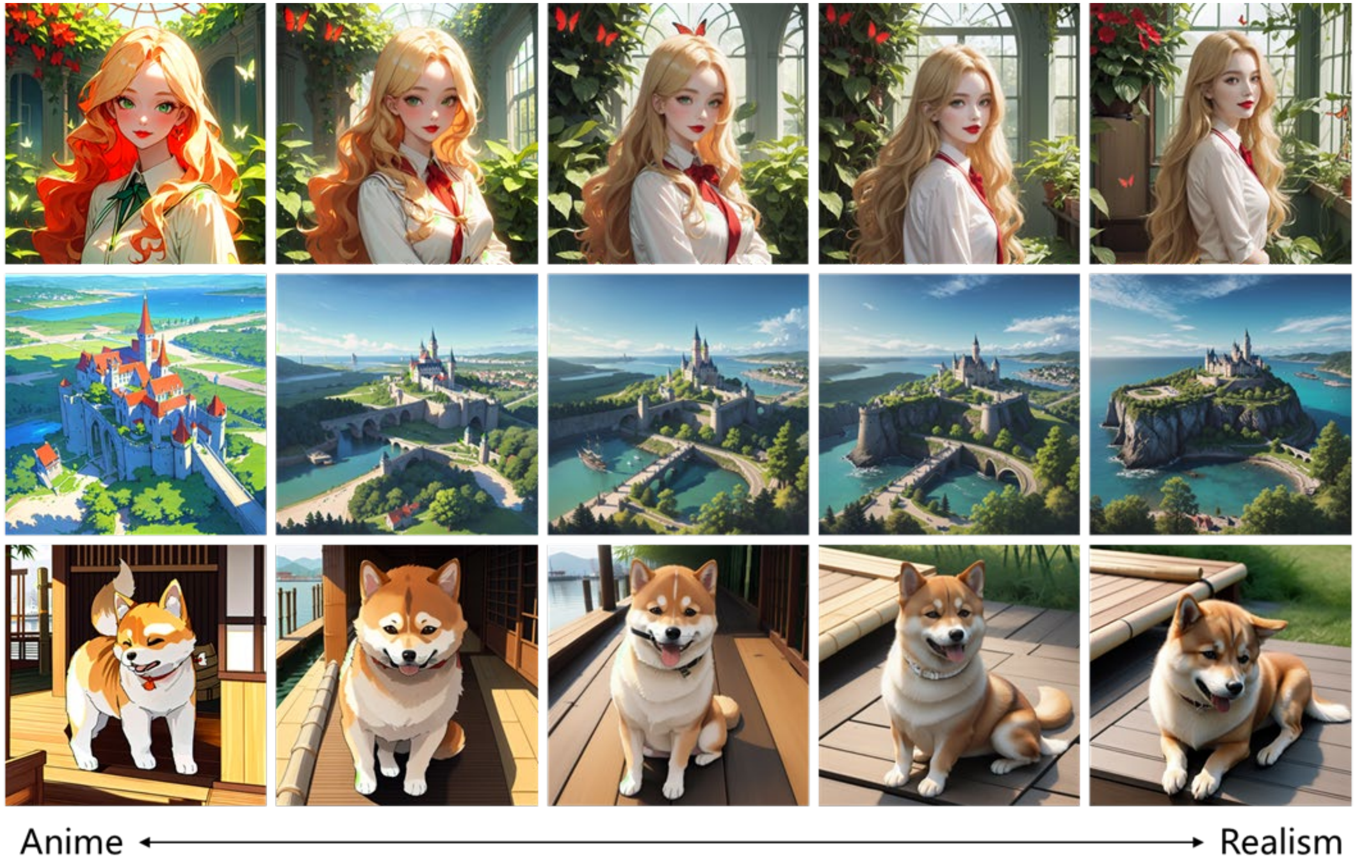
b. Keyword prompts output by the first-stage large language model.

c. Features output by the second-stage AP-adapter.

d. Manually designed prompts.

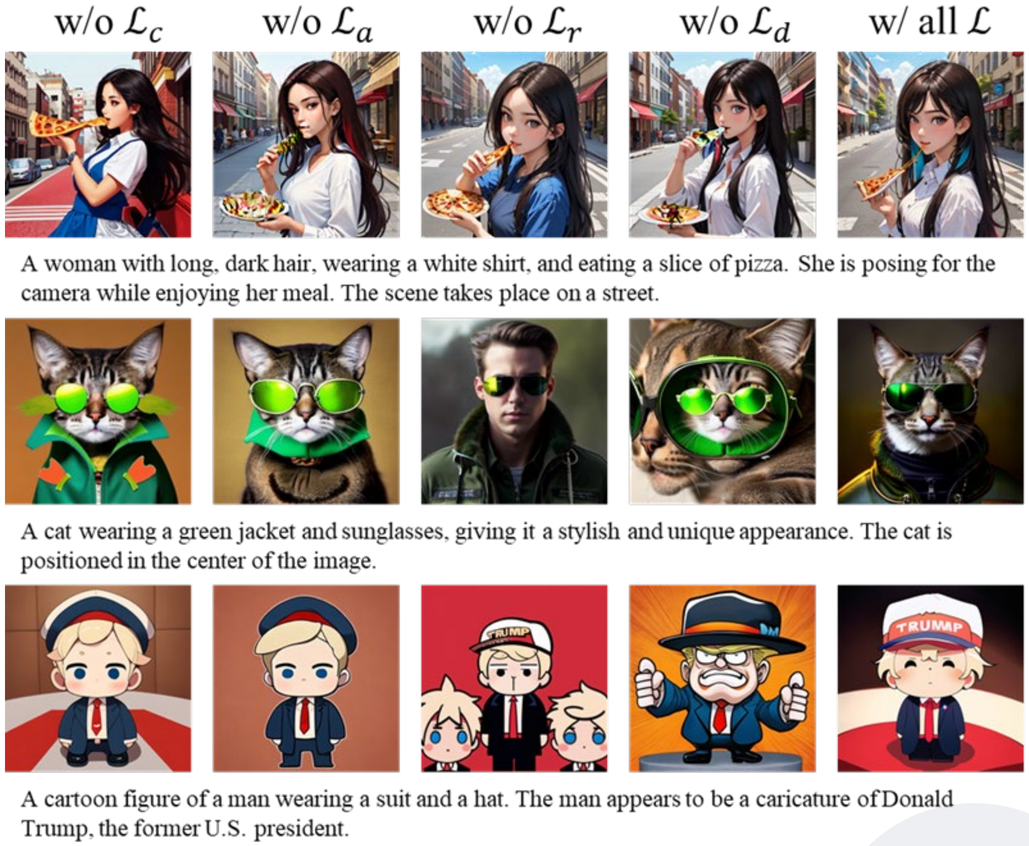
04 Experiments

Effect of Domain Prototypes



Linear combinations of domain prototypes from anime style to realism style. The blending ratio changes from left to right.

Ablation of Losses



A woman with long, dark hair, wearing a white shirt, and eating a slice of pizza. She is posing for the camera while enjoying her meal. The scene takes place on a street.

A cat wearing a green jacket and sunglasses, giving it a stylish and unique appearance. The cat is positioned in the center of the image.

A cartoon figure of a man wearing a suit and a hat. The man appears to be a caricature of Donald Trump, the former U.S. president.

Ablation Study of Loss Functions.

Thank You！

2024.11.06