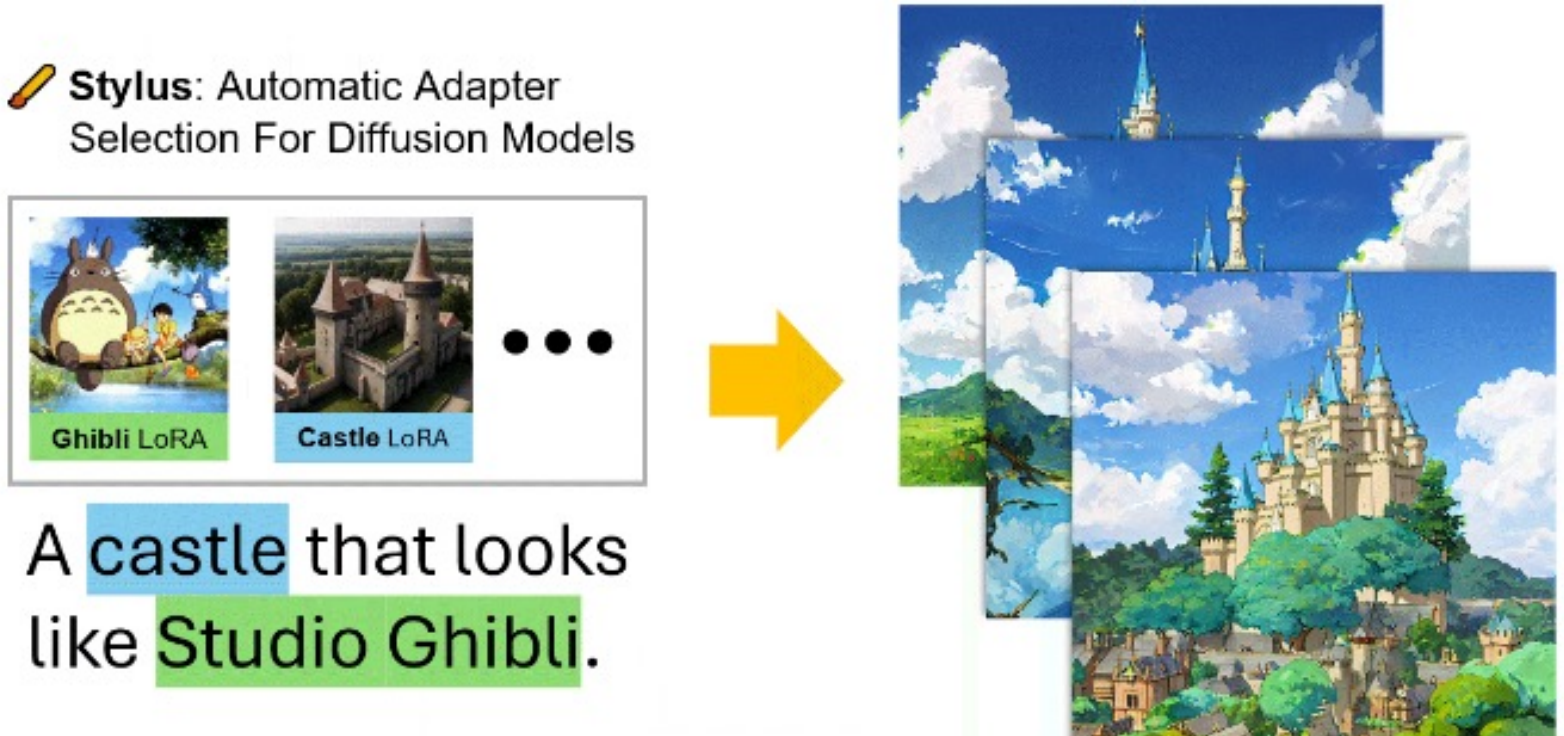
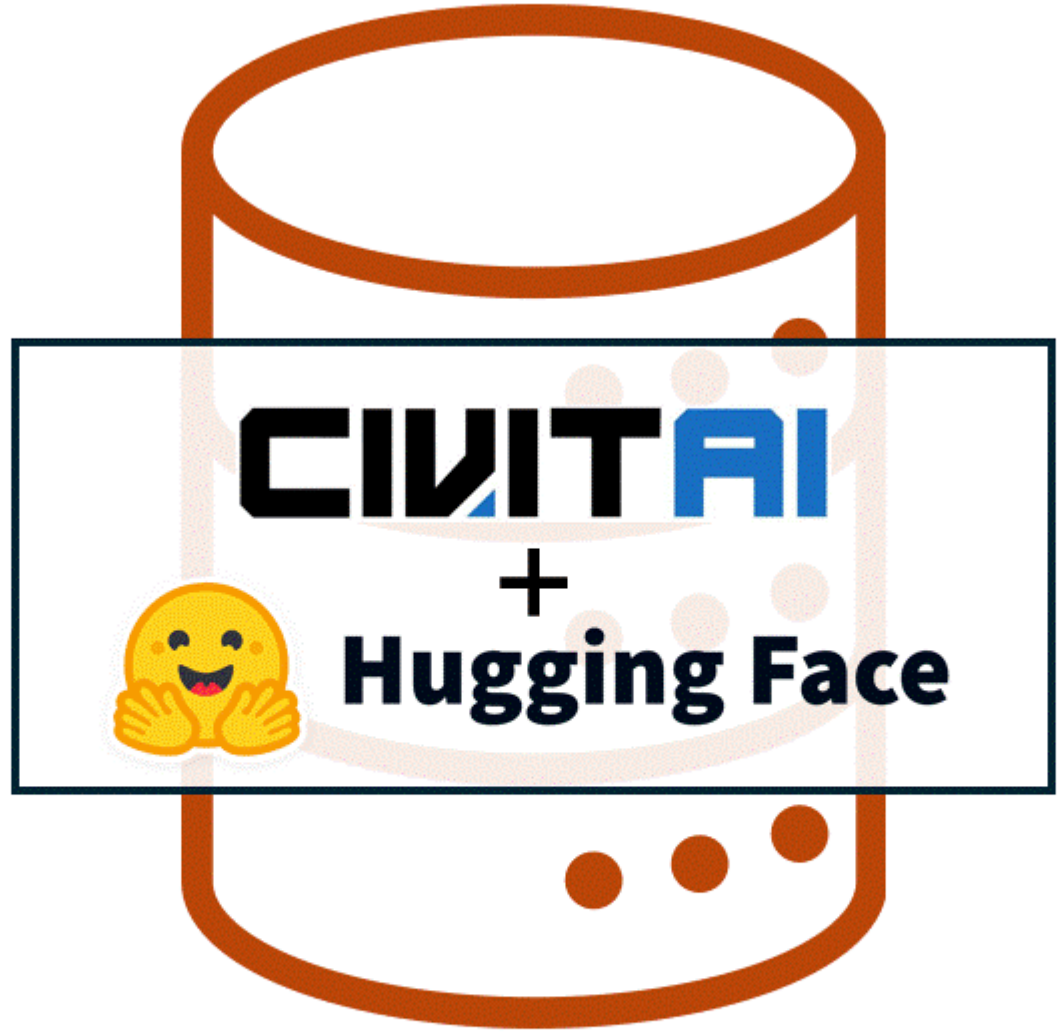
Stylus: Automatic Adapter Selection for Diffusion Models

Michael Luo [1] Justin Wong [1] Brandon Trabucco [2] Yanping Huang [3] Joseph E. Gonzalez [1] Zhifeng Chen [3] Ruslan Salakhutdinov [2] Ion Stoica [1]

[1]UC Berkeley [2]Carnegie Mellon University MLD [3]Google Deepmind

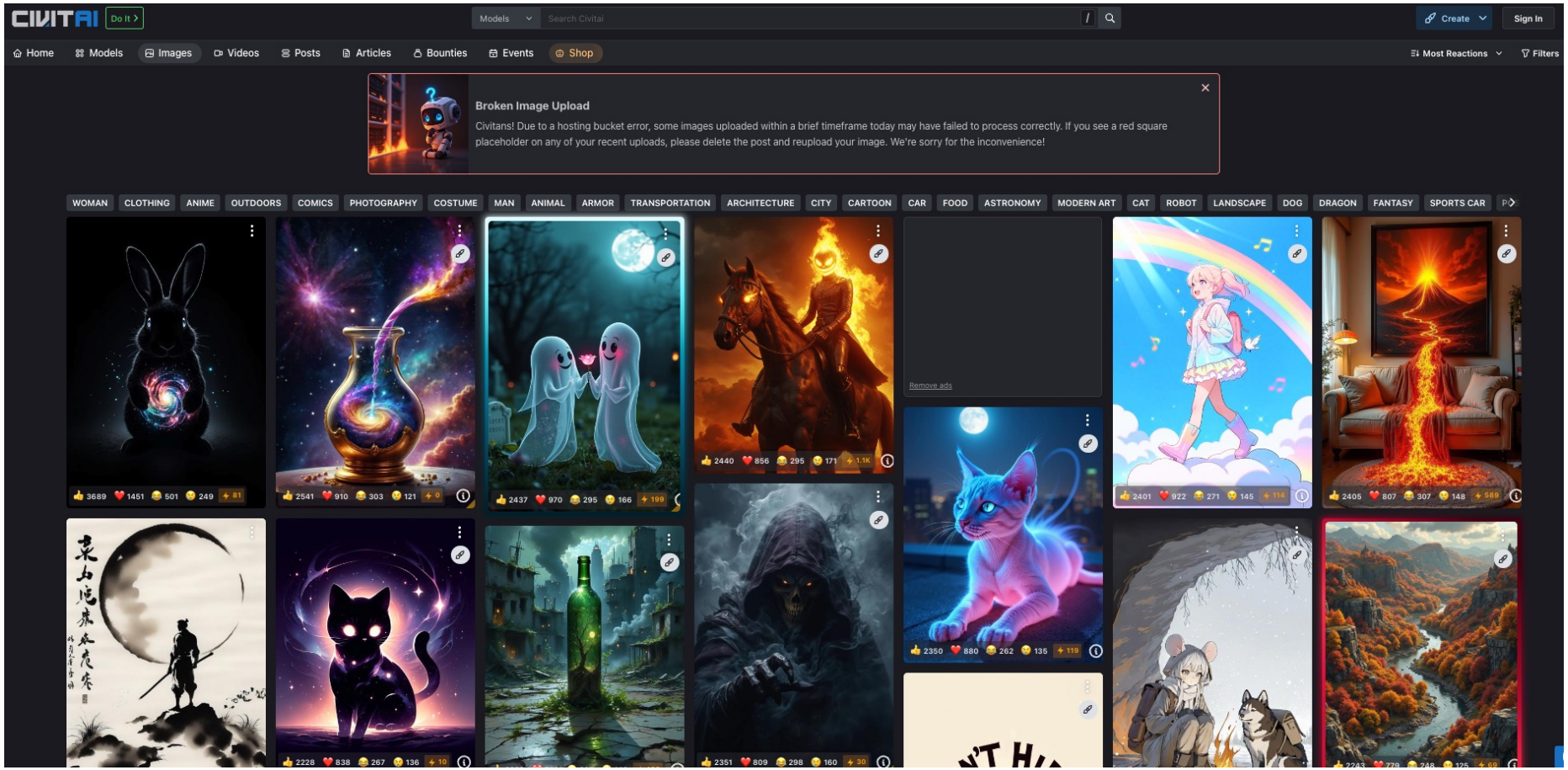




stylus-diffusion.github.io

CIVITAI

- Diffusion Model은 굉장히 다양한 이미지를 생성할 수 있다.
- CIVIT AI 커뮤니티를 Diffusion Model을 생성하는 방법을 갤러리로 저장

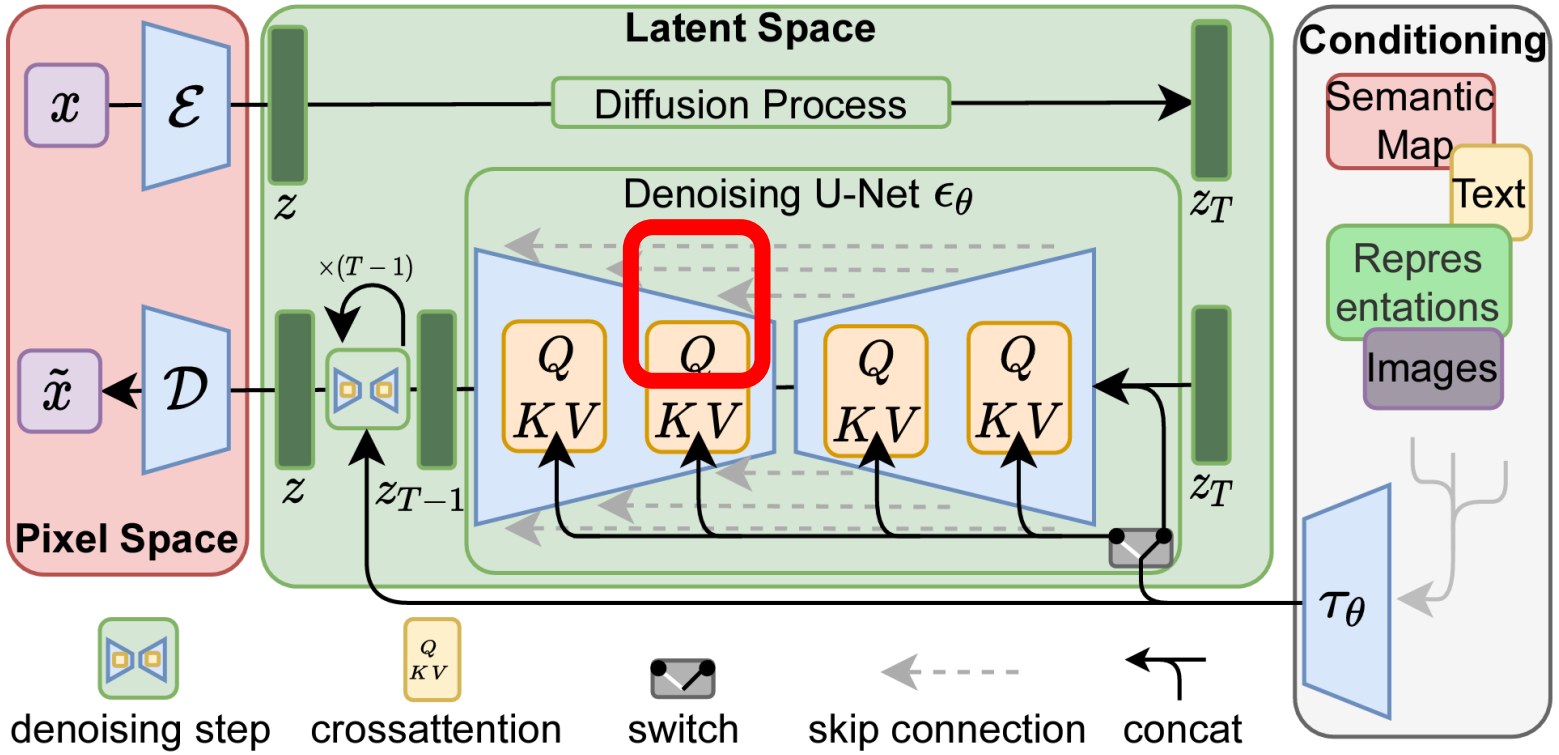


Stylus: Automatic Adapter Selection for Diffusion Models, Luo et al., 2024, NeurIPS

CIVITAI

- Diffusion Model은 굉장히 다양한 이미지를 생성할 수 있다.
- CIVIT AI 커뮤니티를 Diffusion Model을 **생성하는 방법을** 갤러리로 저장

LoRA (Low-Rank Adaptation)



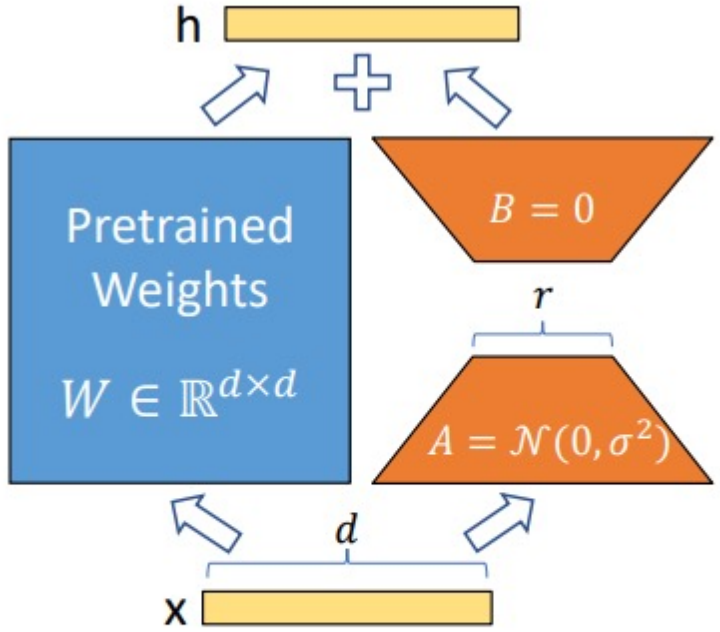


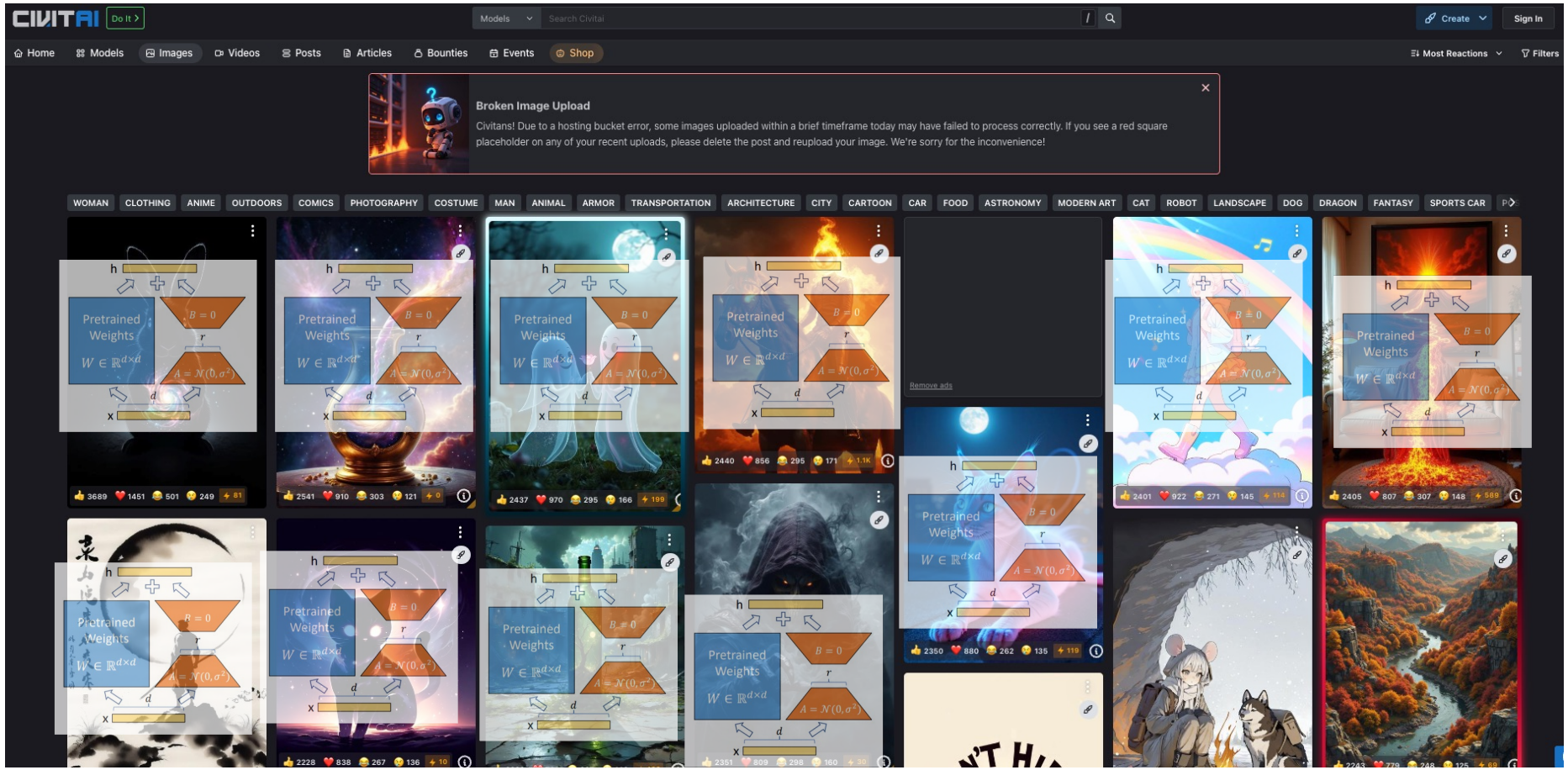
Figure 1: Our reparametrization. We only train A and B.

Stylus: Automatic Adapter Selection for Diffusion Models, Luo et al., 2024, NeurIPS

CIVITAI

- Diffusion Model은 굉장히 다양한 이미지를 생성할 수 있다.
- CIVIT AI 커뮤니티를 Diffusion Model을 **생성하는 방법을** 갤러리로 저장 **(각 이미지 마다)**

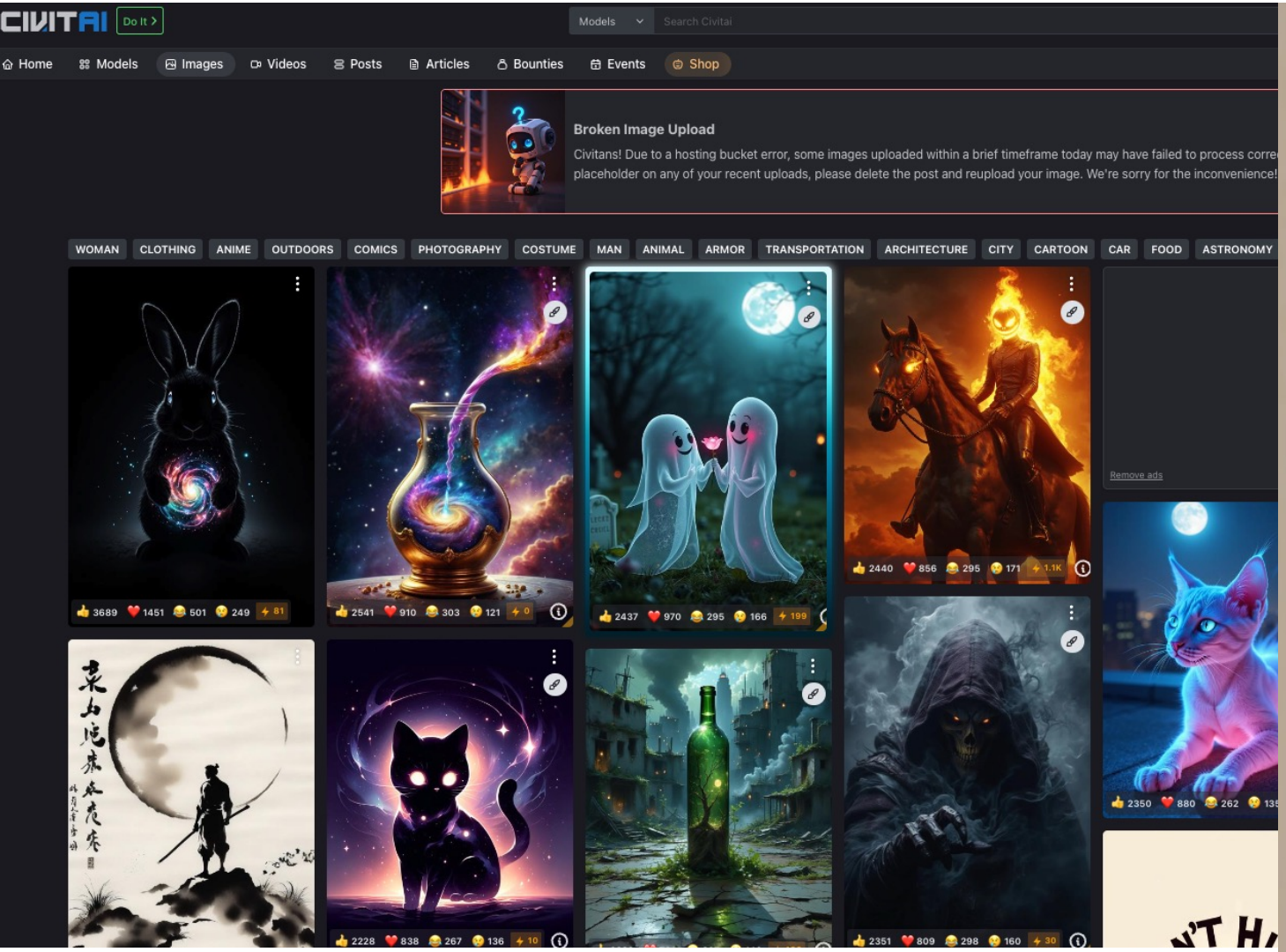
LoRA (Low-Rank Adaptation)



Stylus: Automatic Adapter Selection for Diffusion Models, Luo et al., 2024, NeurIPS

CIVITAI

- 내가 원하는 정확한 이미지를 찾기 어려움.



Stylus: Automatic Adapter Selection for Diffusion Models, Luo et al., 2024, NeurIPS

Stylus: Retrieval을 편하게!

- **입력:** 텍스트 (Two dogs playing in the snow)
- **Retriever:** VectorDB에서 LoRA 파라미터를 찾아 줌

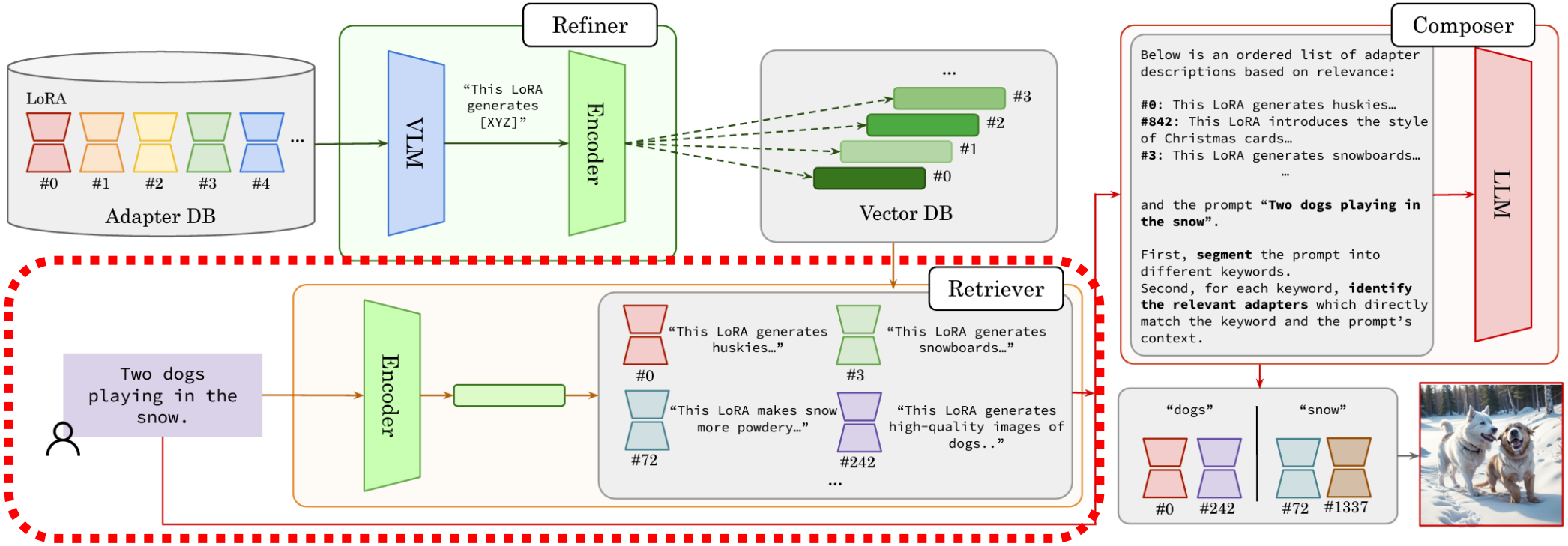


Figure 2. **Stylus algorithm.** Stylus consists of three stages. The *refiner* plugs an adapter's model card through a VLM to generate textual descriptions of an adapter's task and then through an encoder to produce the corresponding text embedding. The *retriever* fetches candidate adapters that are relevant to the entire user prompt. Finally, the *composer* prunes and jointly categorizes the remaining adapters based on the prompt's tasks, which correspond to a set of keywords.

Stylus: Retrieval을 편하게!

- **입력:** 텍스트 (Two dogs playing in the snow)
- **Composer:** LoRA들의 설명과 **입력**을 매칭해서 어떤 것을 사용할지 결정

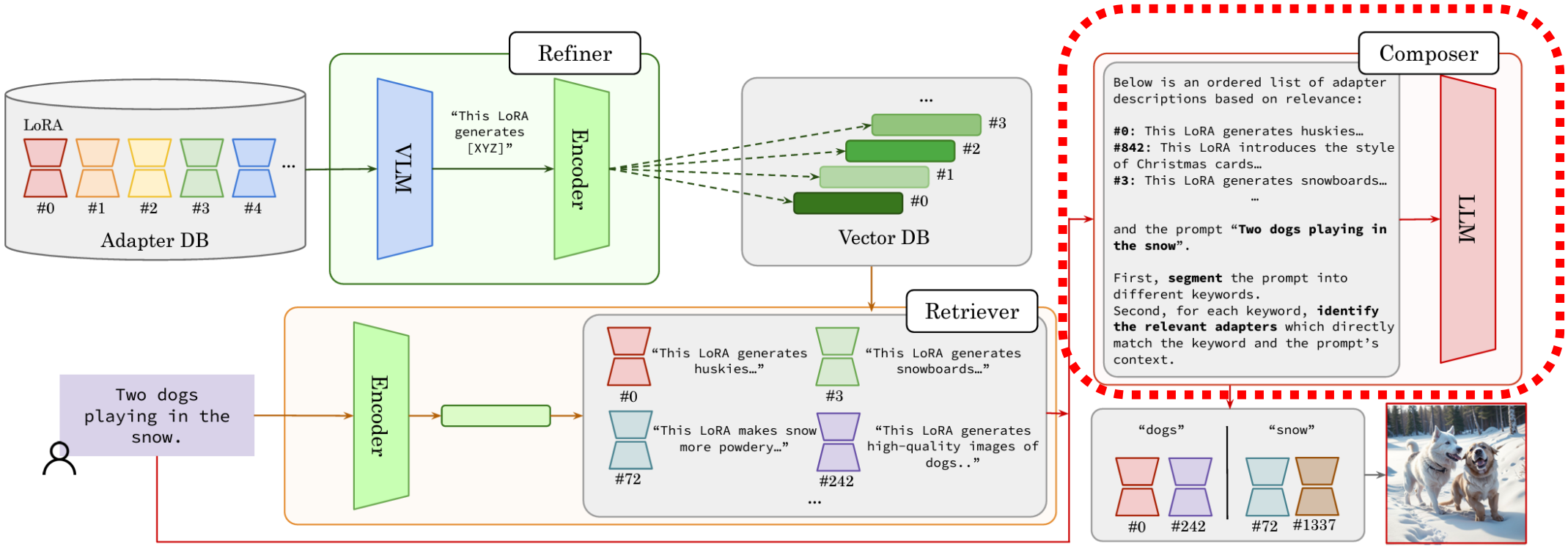


Figure 2. **Stylus algorithm.** Stylus consists of three stages. The *refiner* plugs an adapter's model card through a VLM to generate textual descriptions of an adapter's task and then through an encoder to produce the corresponding text embedding. The *retriever* fetches candidate adapters that are relevant to the entire user prompt. Finally, the *composer* prunes and jointly categorizes the remaining adapters based on the prompt's tasks, which correspond to a set of keywords.

Stylus: Retrieval을 편하게!

- **입력:** 텍스트 (Two dogs playing in the snow)
- **Composer:** LoRA들의 설명과 **입력**을 매칭해서 어떤 것을 사용할지 결정

Figure 2. **Stylus algorith** …er's model card through a VLM to generate textual descriptions of an a… …onding text embedding. The *retriever* fetches candidate adapters that are … …tly categorizes the remaining adapters based on the prompt's tasks…

Stylus: Automatic Adapter Selection for Diffusion Models, Luo et al., 2024, NeurIPS

Stylus: Retrieval을 편하게!

- **입력:** 텍스트 (Two dogs playing in the snow)
- **Retriever:** VectorDB에서 LoRA 파라미터를 찾아 줌 **(Embedding 유사도 기반)**

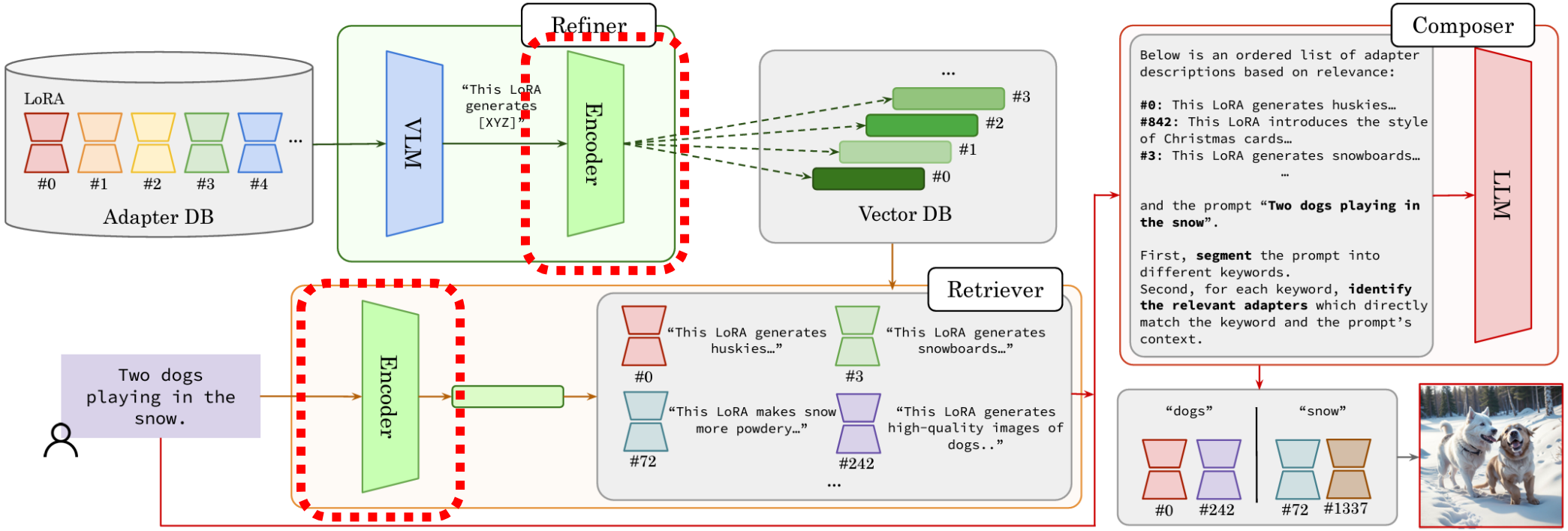
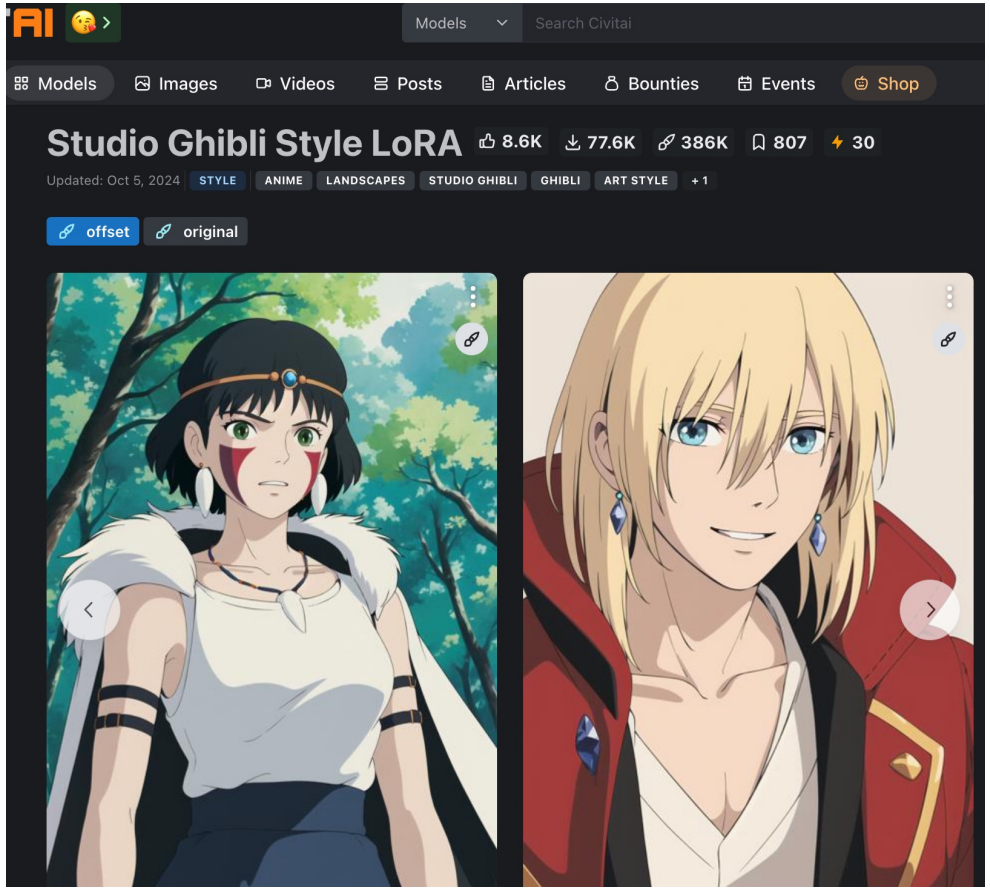
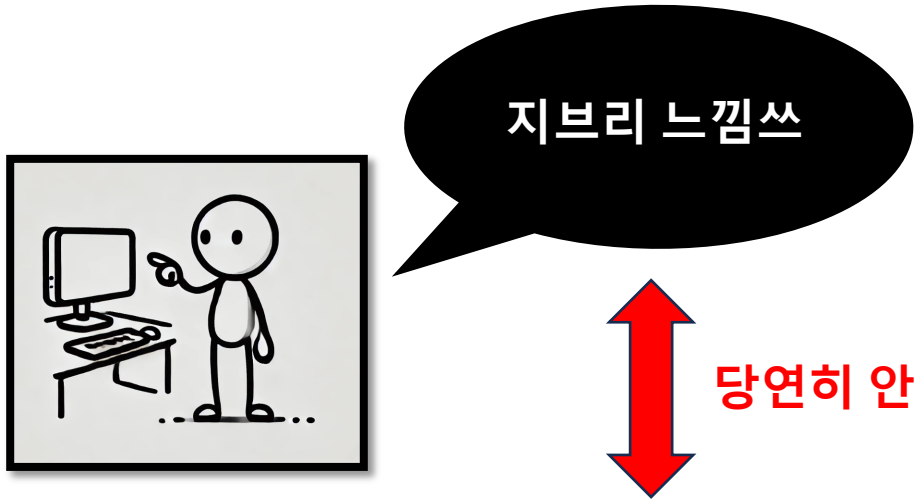


Figure 2. **Stylus algorithm.** Stylus consists of three stages. The *refiner* plugs an adapter's model card through a VLM to generate textual descriptions of an adapter's task and then through an encoder to produce the corresponding text embedding. The *retriever* fetches candidate adapters that are relevant to the entire user prompt. Finally, the *composer* prunes and jointly categorizes the remaining adapters based on the prompt's tasks, which correspond to a set of keywords.

사용자 프롬프트 vs LoRA Description





Studio Ghibli style LoRA

Making models can be expensive. Do you like what I do? Consider supporting me on Patreon or feel free to buy me a coffee ☕

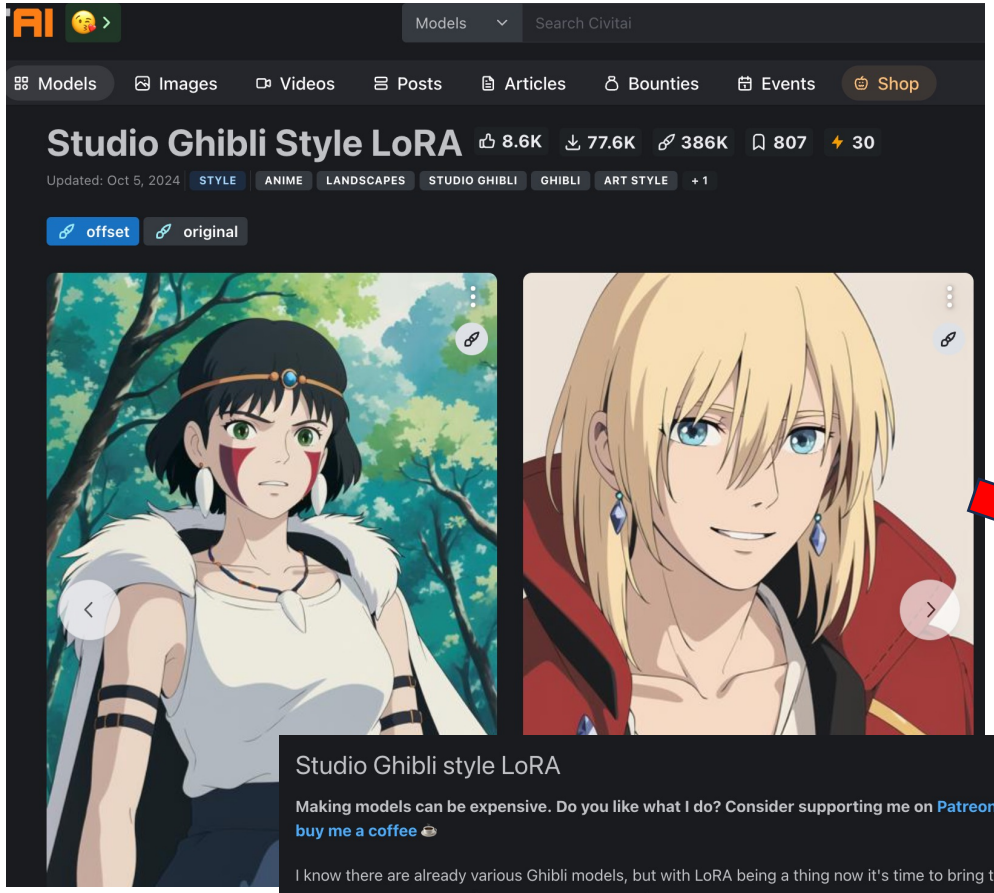
I know there are already various Ghibli models, but with LoRA being a thing now it's time to bring this style into 2023. This is good around **1 weight for the offset version and 0.65 for the old one,** on **Anything v4.**5 for a more authentic style, but it's also good on AbyssOrangeMix2. Triggers with `ghibli style` and, as you can see, it should work with any character LoRA or TI that didn't overfit on style too. Keep in mind, this is a **style** model not a "Ghibli characters" model, so movie characters in the examples are made using careful prompting, meaning it can reproduce similar characters, but it won't make them perfectly (unless you also use other TI/LoRA's).

4th pic by **Karkaz** using also his own character LoRA.

Last pic by **LottaLewds**.

Stylus: Automatic Adapter Selection for Diffusion Models, Luo et al., 2024, NeurIPS

사용자 프롬프트 vs LoRA Description



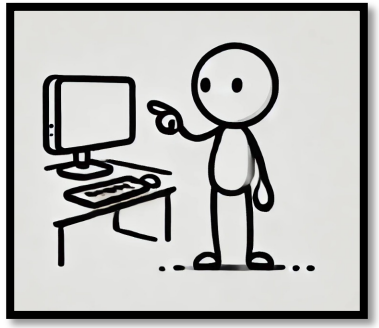
Studio Ghibli style LoRA

Making models can be expensive. Do you like what I do? Consider supporting me on Patreon or feel free to buy me a coffee

I know there are already various Ghibli models, but with LoRA being a thing now it's time to bring this style into 2023. This is good around **1 weight for the offset version and 0.65 for the old one,** on **Anything v4.**5 for a more authentic style, but it's also good on AbyssOrangeMix2. Triggers with `ghibli style` and, as you can see, it should work with any character LoRA or TI that didn't overfit on style too. Keep in mind, this is a **style** model not a "Ghibli characters" model, so movie characters in the examples are made using careful prompting, meaning it can reproduce similar characters, but it won't make them perfectly (unless you also use other TI/LoRA's).

4th pic by **Karkaz** using also his own character LoRA.

Last pic by **LottaLewds**.



지브리 느낌쓰

Ok

하울 느낌쓰

Image

Text

Vision-Language Model

Stylus: Automatic Adapter Selection for Diffusion Models, Luo et al., 2024, NeurIPS

Stylus: Retrieval을 편하게!

- **Refiner (VLM)** : CIVIT-AI **LoRA 정보 후처리**
- **Encoder (Text Embedding): 언어를 벡터로 변환**

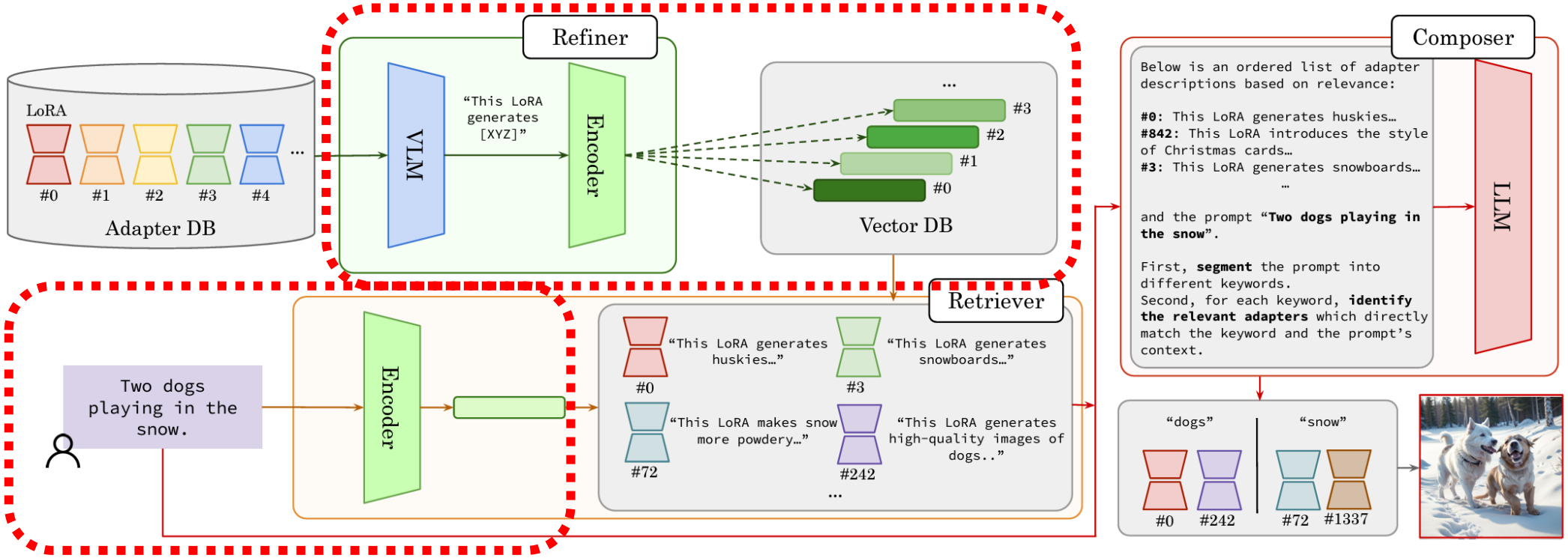


Figure 2. **Stylus algorithm.** Stylus consists of three stages. The *refiner* plugs an adapter's model card through a VLM to generate textual descriptions of an adapter's task and then through an encoder to produce the corresponding text embedding. The *retriever* fetches candidate adapters that are relevant to the entire user prompt. Finally, the *composer* prunes and jointly categorizes the remaining adapters based on the prompt's tasks, which correspond to a set of keywords.

Stable Diffusion vs Stylus **(SD + LoRA)**

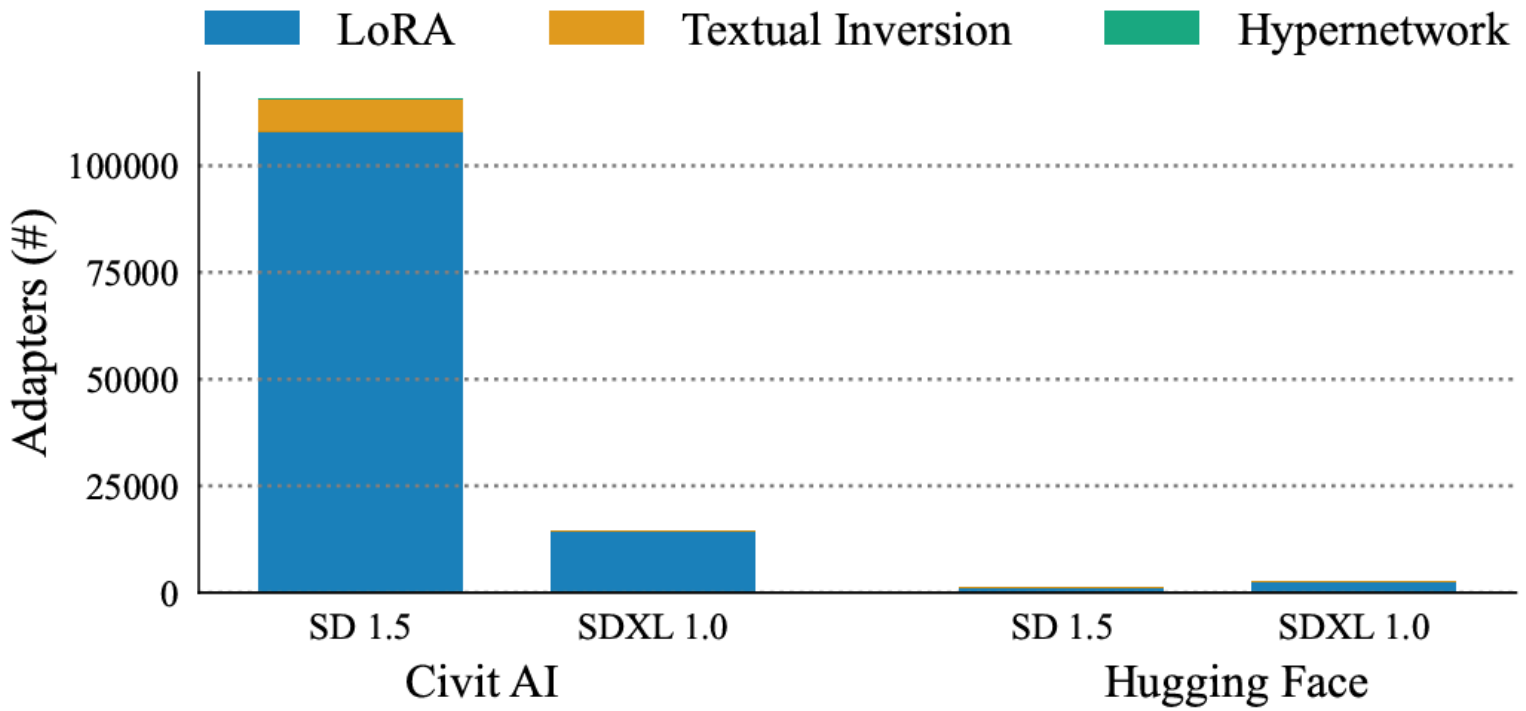


Figure 3. **Number of Adapters.** Civit AI boasts 100K+ adapters for Stable Diffusion, outpacing that of Hugging Face. Low-Rank Adaptation (LoRA) is the dominant approach for finetuning.

Stable Diffusion vs Stylus **(SD + LoRA)**



Image Prompts

Prompt 1: Photo of Dwayne Johnson, wearing military clothes and cap, dramatic lighting, <lora:TheRockV3:0.9>.
Prompt 2: Photo of Dwayne Johnson, wearing a Superman suit, high quality, <lora:TheRockV3:1>.
Prompt 3: Photo of Dwayne Johnson, wearing an Armani tuxedo, <lora:TheRockV3:0.9>

Model Card Description

- Title: Dwayne "The Rock" Johnson (LoRA)
- Tags: Celebrity, Photorealistic, Hollywood, Celeb
- Trigger Words: Th3R0ck
- Description: Had to make this one, due to Kevin Hart Lora. Recommended lora strength: 0.9. *% Author descriptions may be misleading or incomplete.*

Stable Diffusion vs Stylus **(SD + LoRA)**

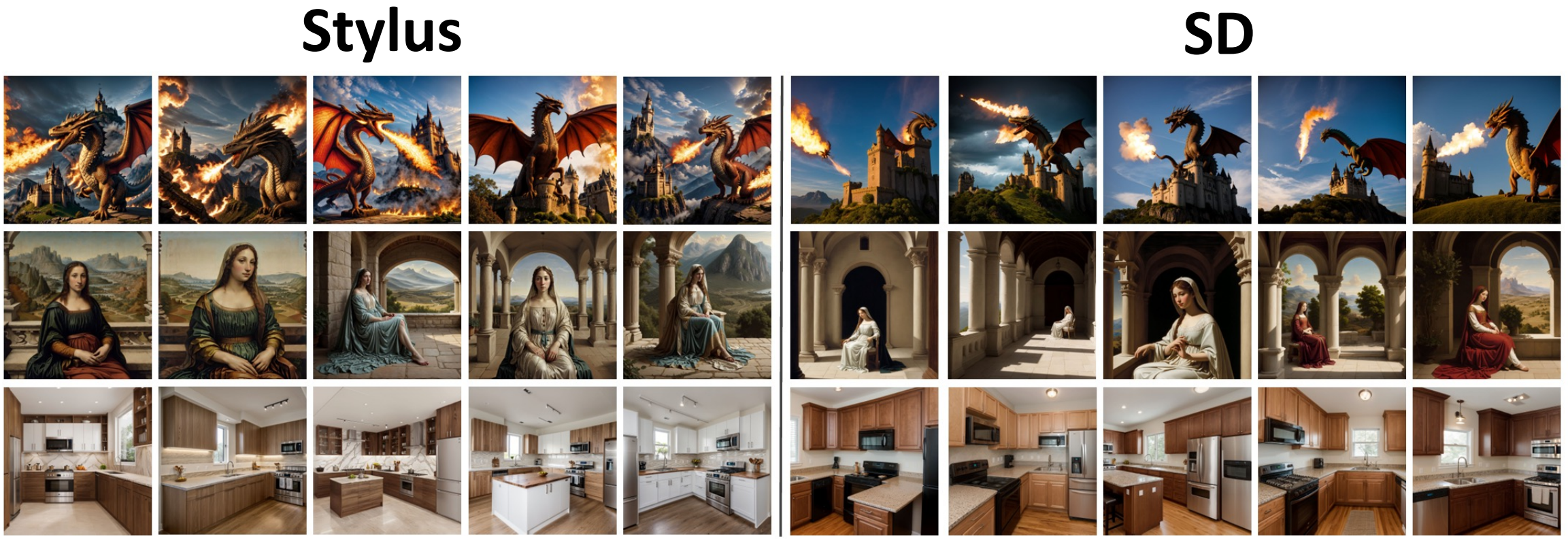


Figure 7. **Image Diversity.** Given the same prompt, our method (left) generates more diverse and comprehensive sets of images than that of existing Stable Diffusion checkpoints (right). Stylus's diversity comes from its masking scheme and the composer LLM's temperature.

Stable Diffusion vs Stylus **(SD + LoRA)**

Figure 4. **Qualitative comparison between Stylus over realistic (left) and cartoon (right) style Stable Diffusion checkpoints.** Stylus produces highly detailed images that correctly depicts keywords in the context of the prompt. For the prompt "A graffiti of a corgi on the wall", our method correctly depicts a spray-painted corgi, whereas the checkpoint generates a realistic dog.

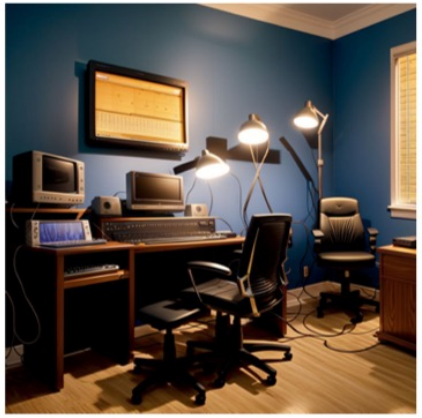
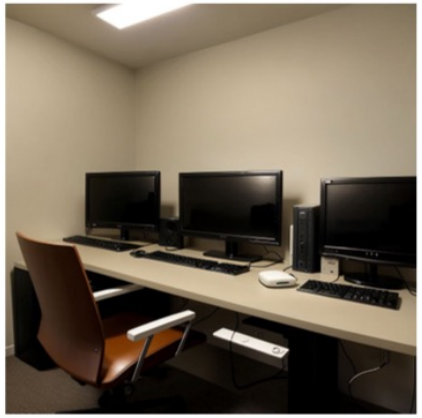
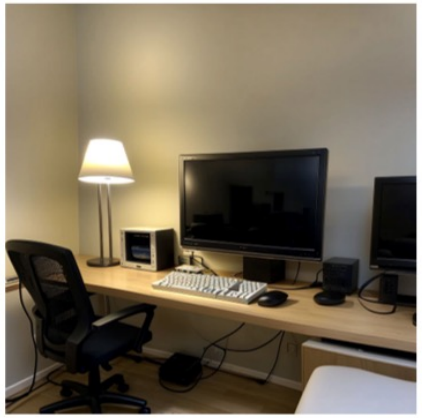
Stable Diffusion vs Stylus **(SD + LoRA)**

Stylus

SD v1.5




Wooden dish rack on a counter holding plates, saucers, a bowl, mugs and glasses.

A computer room with monitors on and a keyboard and a reading lamp.

Stable Diffusion vs Stylus **(SD + LoRA)**

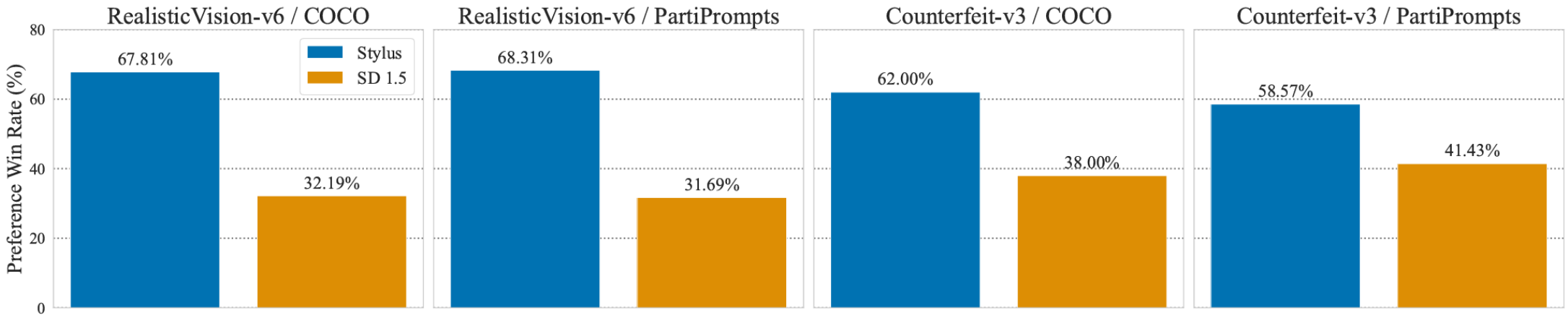
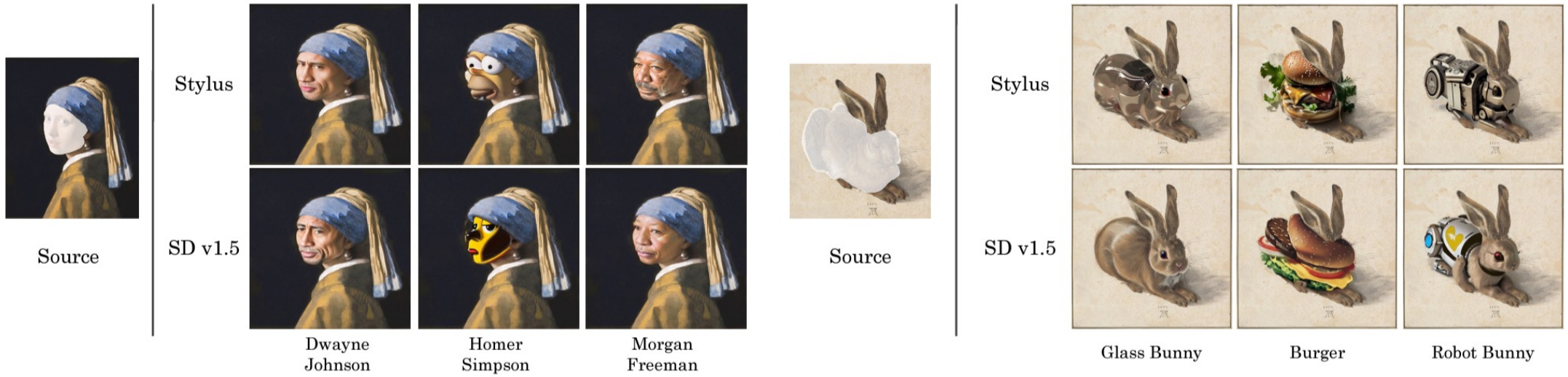


Figure 5. **Human Evaluation.** Stylus achieves a higher preference scores (2:1) over different datasets and Stable Diffusion checkpoints.

Stable Diffusion vs Stylus **(SD + LoRA)**

(a) **Image Translation.** Stylus chooses relevant adapters that better adapt new styles and elements into existing images.



(b) **Inpainting.** Stylus chooses adapters than can better introduce new characters or concepts into the inpainted mask.

Stable Diffusion vs Stylus **(SD + LoRA)**

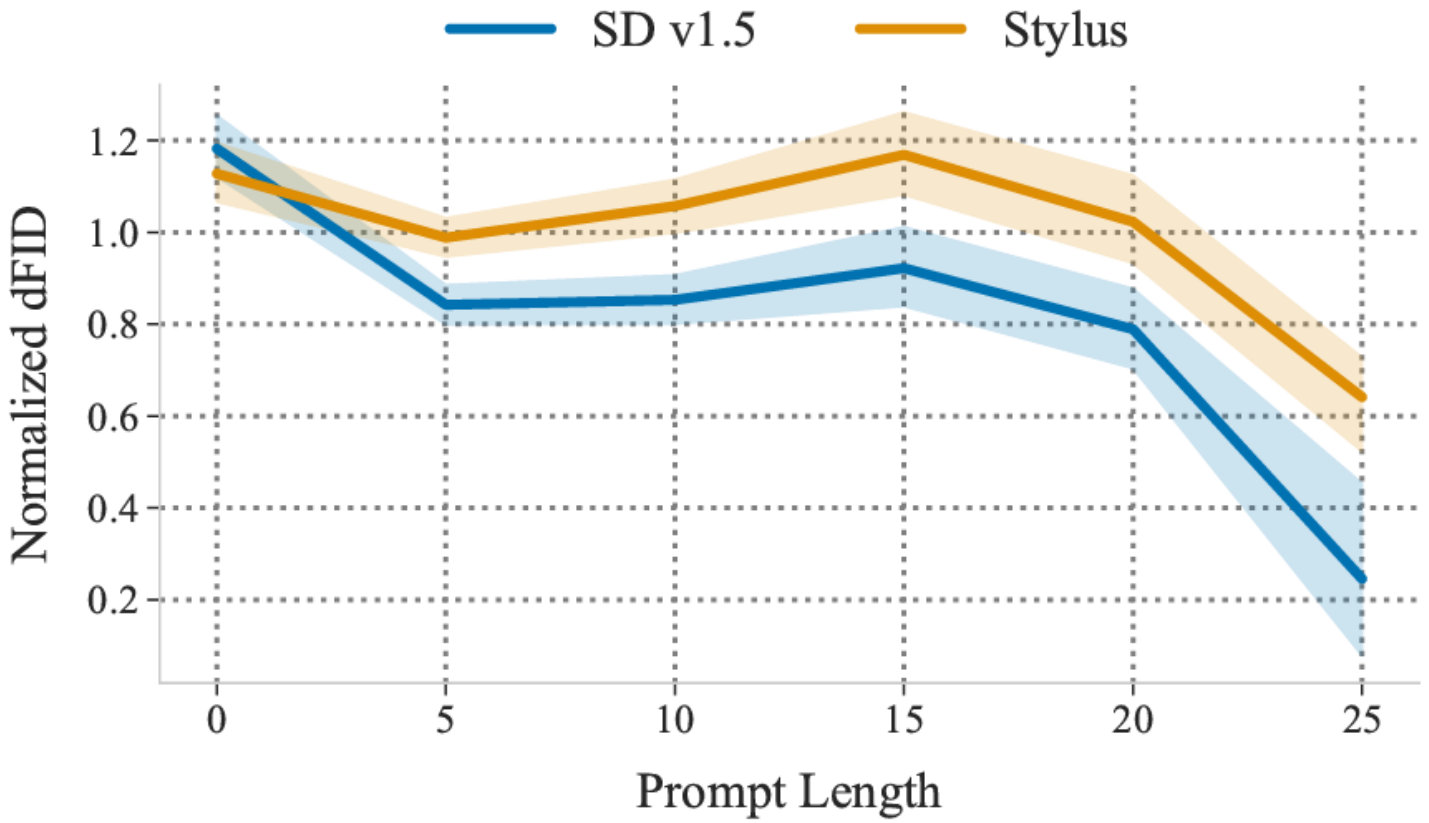


Figure 9. **Image diversity (*d*FID) across prompt length.** Stylus achieves higher diversity score than Stable Diffusion when prompt length increases.

Stable Diffusion vs Stylus **(SD + LoRA)**

	CLIP (Δ)	**FID (Δ)**
Stylus	**27.25** (+0.03)	**22.05** (-1.91)
Reranker	25.48 (-1.74)	22.81 (-1.15)
Retriever-only	24.93 (-2.29)	24.68 (+0.72)
Random	26.34 (-0.88)	24.39 (+0.43)
SD v1.5	27.22	23.96

Table 1. **Evaluation over different retrieval methods (CFG=6).** Stylus outperforms existing retrieval-based methods, attains the best FID score, and similar CLIP score to Stable Diffusion.