Stylus 🖌️

Automatic Adapter Selection

Main Presenter: Michael Luo

Paper Authors: Michael Luo, Justin Wong, Brandon Trabucco, Yanping Huang, Joseph Gonzalez, Zhifeng Chen, Ruslan Salakhutdinov, Ion Stoica

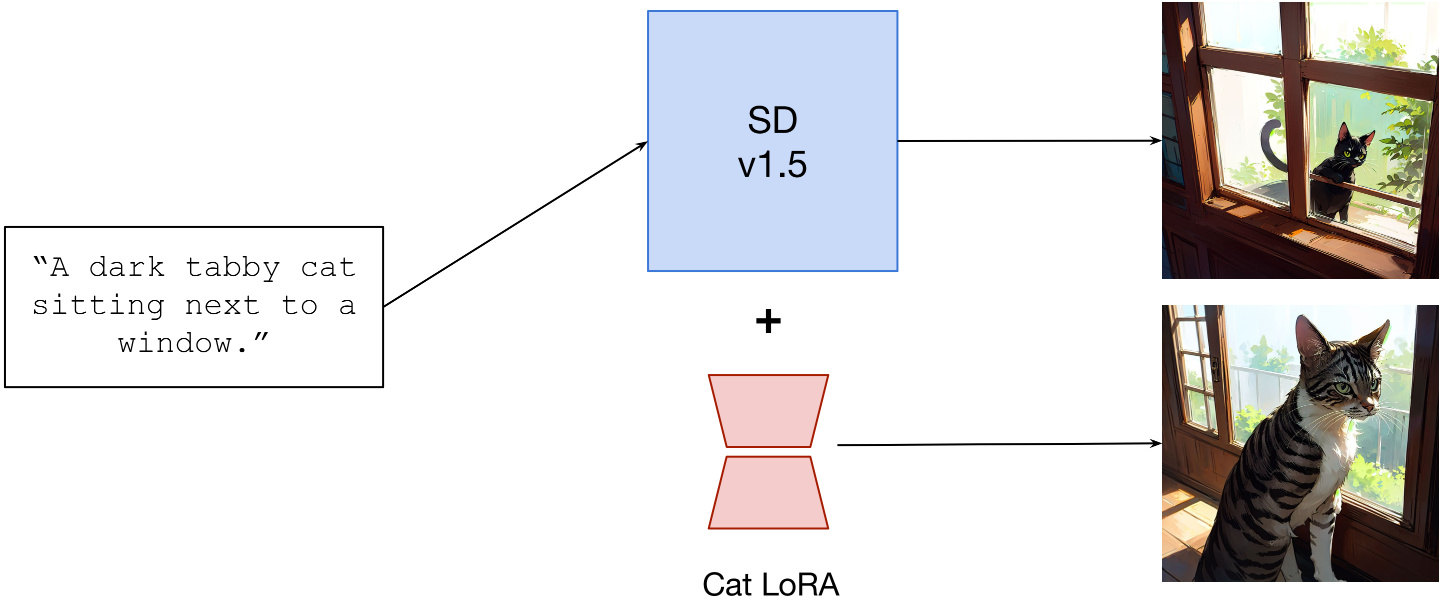
Berkeley UNIVERSITY OF CALIFORNIA

sky

BAIR BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Situation

Finetuned Adapters (i.e. LoRA, textual inversion) introduce novel concepts and styles to the base model, thereby improving image quality.



"A dark tabby cat sitting next to a window."

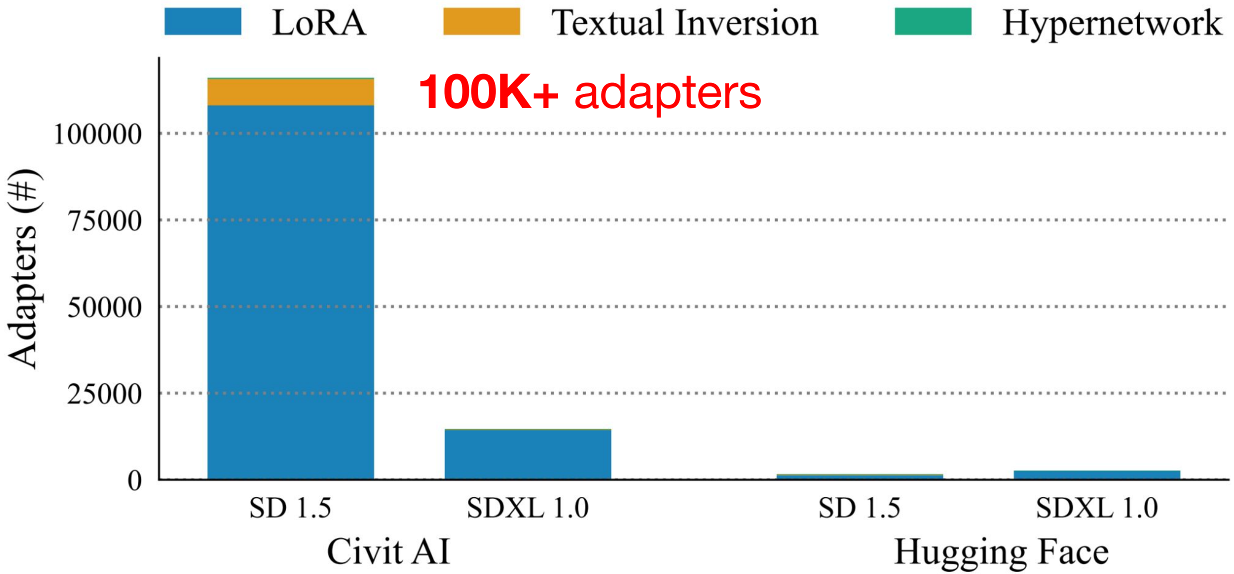
SD v1.5

+

Cat LoRA

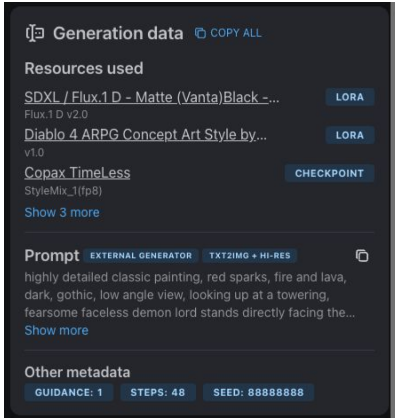
Situation

Open-source contributors have created over **100K+ adapters**!



An Emergent Problem

- Base models are no longer good enough for generating images.
- Users *manually mix and match* many checkpoints and adapters to generate the images they want.



Generation data

Resources used

SDXL / Flux.1 D - Matte (Vanta)Black -... LORA
Flux.1 D v2.0

Diablo 4 ARPG Concept Art Style by... LORA
v1.0

Copax TimeLess CHECKPOINT
StyleMix_1(fp8)

Prompt EXTERNAL GENERATOR TXT2IMG + HI-RES

highly detailed classic painting, red sparks, fire and lava, dark, gothic, low angle view, looking up at a towering, fearsome faceless demon lord stands directly facing the...

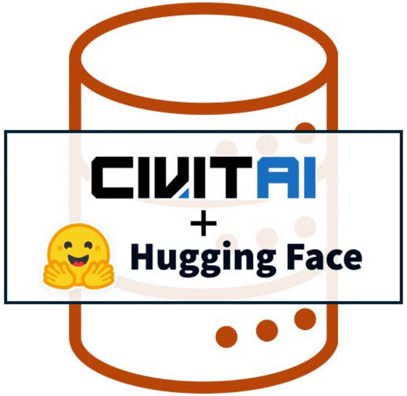
Other metadata

GUIDANCE: 1 STEPS: 48 SEED: 88888888

Image taken from Civit.ai

Stylus 🖌️

Goal: Automatically *select* and *compose* the right adapters given a user prompt.



stylus-diffusion.github.io

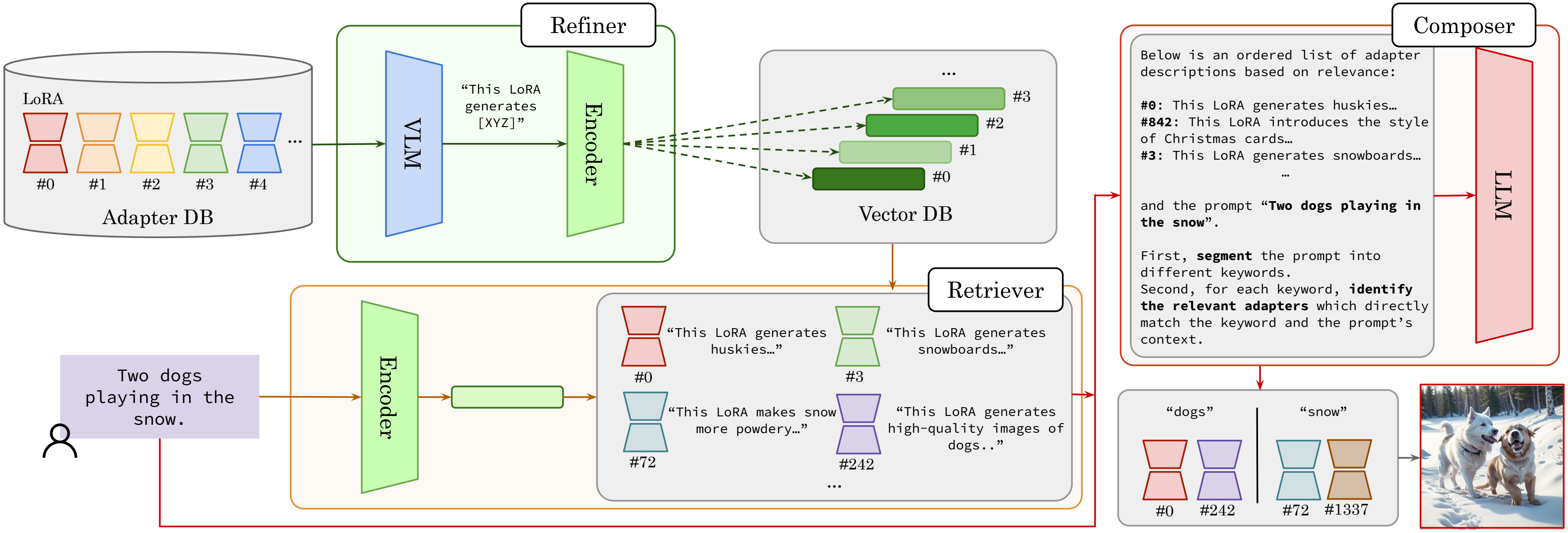
Stylus 🖌️

Goal: Automatically *select* and *compose* the right adapters given a user prompt.

Requirements:

- No training. Scales to new adapters over time.
- Performant. Must do better than existing retrieval approaches.
- Low inference overhead. Identifies the right adapters quickly.

Stylus Architecture



Refiner

Adapter DB

LoRA

#0 #1 #2 #3 #4 …

VLM

“This LoRA generates [XYZ]”

Encoder

Vector DB

…

#3

#2

#1

#0

Retriever

Two dogs playing in the snow.

Encoder

“This LoRA generates huskies…”

#0

“This LoRA generates snowboards…”

#3

“This LoRA makes snow more powdery…”

#72

“This LoRA generates high-quality images of dogs..”

#242

…

Composer

Below is an ordered list of adapter descriptions based on relevance:

#0: This LoRA generates huskies…
#842: This LoRA introduces the style of Christmas cards…
#3: This LoRA generates snowboards…
…

and the prompt “**Two dogs playing in the snow**”.

First, **segment** the prompt into different keywords.
Second, for each keyword, **identify the relevant adapters** which directly match the keyword and the prompt’s context.

LLM

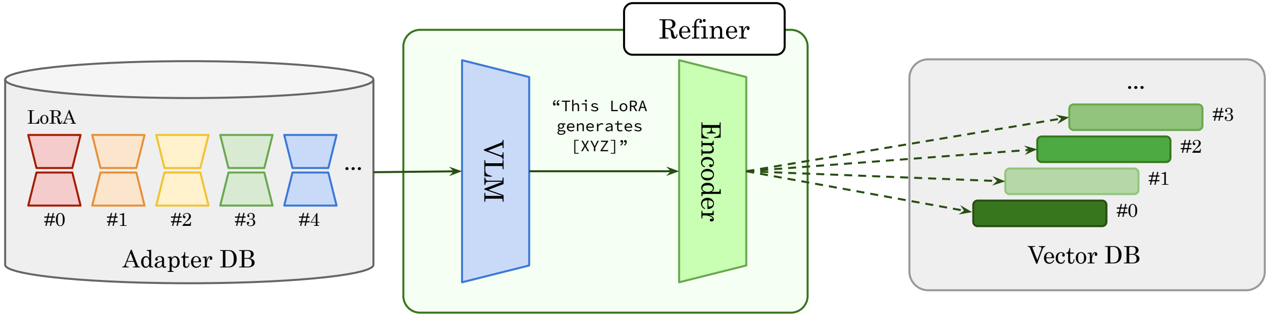
“dogs”

#0 #242

“snow”

#72 #1337

Refiner



Goal: Generate a vector embedding for each adapter.

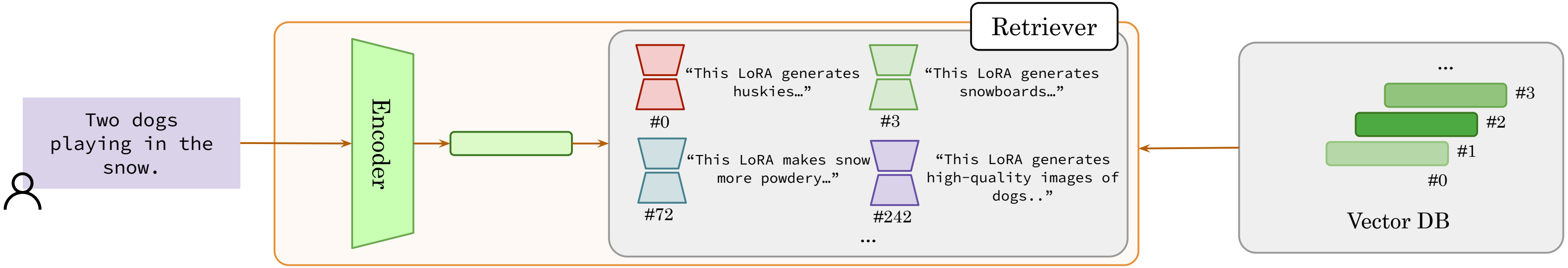
How?

- Vision Language Model (VLM) to infer adapters' descriptions
  - With adapter's model card (author description + example images)
- Embedding Model to embed adapter description

Retriever (RAG)

Goal: Retrieves the most relevant candidate adapters via similarity metric.

Problem: RAG lacks precision. Easy to add *slightly relevant* adapters.



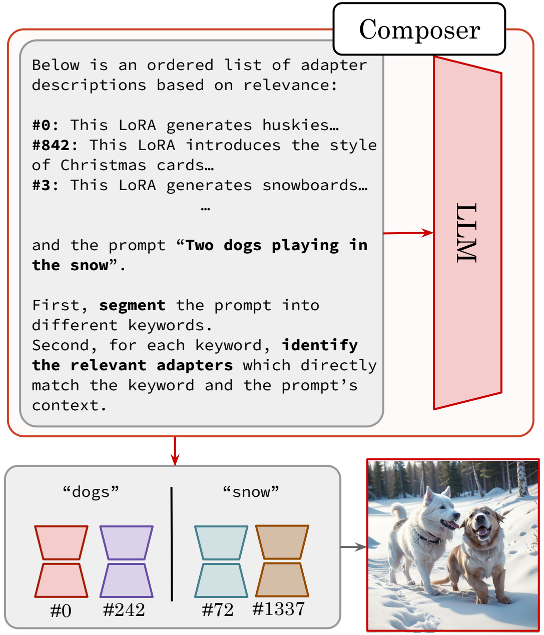
Two dogs playing in the snow.
Encoder
Retriever
"This LoRA generates huskies…"
#0
"This LoRA generates snowboards…"
#3
"This LoRA makes snow more powdery…"
#72
"This LoRA generates high-quality images of dogs.."
#242
…
…
#3
#2
#1
#0
Vector DB

Composer

Goal: Significantly improves precision for retrieving the *right* adapters.

How?

- Evaluate adapters based on *semantic relevance*.
    - Maps relevant adapters to keywords in user prompt.
        - Enforces that adapters stay relevant.
    - Prunes irrelevant adapters.
- Efficient. Only requires one LLM call.



```
Below is an ordered list of adapter
descriptions based on relevance:

#0: This LoRA generates huskies…
#842: This LoRA introduces the style
of Christmas cards…
#3: This LoRA generates snowboards…
                    …

and the prompt “Two dogs playing in
the snow”.

First, segment the prompt into
different keywords.
Second, for each keyword, identify
the relevant adapters which directly
match the keyword and the prompt’s
context.
```

Evaluation

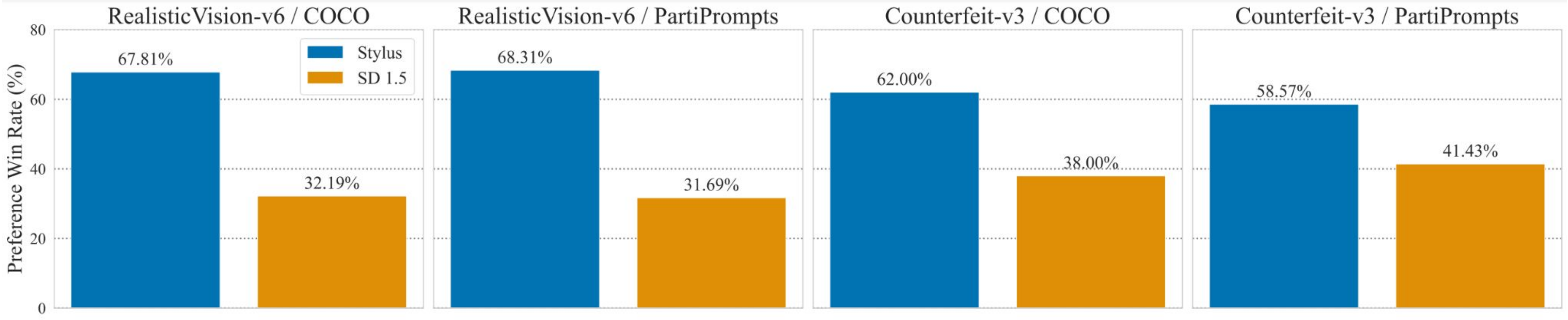
- **Prompt Datasets**
  - COCO 2014
  - PartiPrompts

- **Base Models**
  - RealisticVision (SD v1.5)
  - CounterFeit (SD v1.5)

- **Adapter Database**
  - StylusDocs v2
    - Adapters from Civit AI + Huggingface.
    - 75K generated adapter descriptions from GPT-4o.

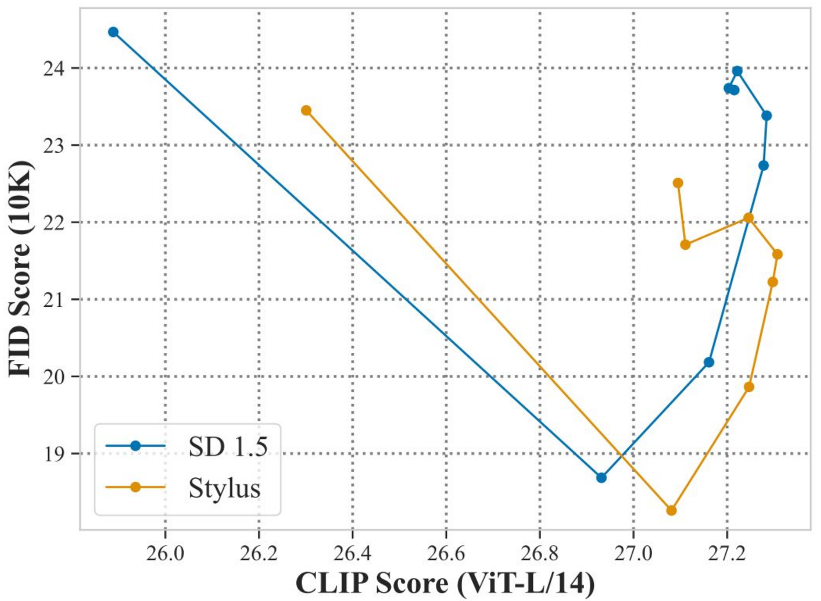
- **Evaluation**
  - Human Evaluation
  - FID/CLIP Pareto Curve
  - VLM as a judge

Human Evaluation



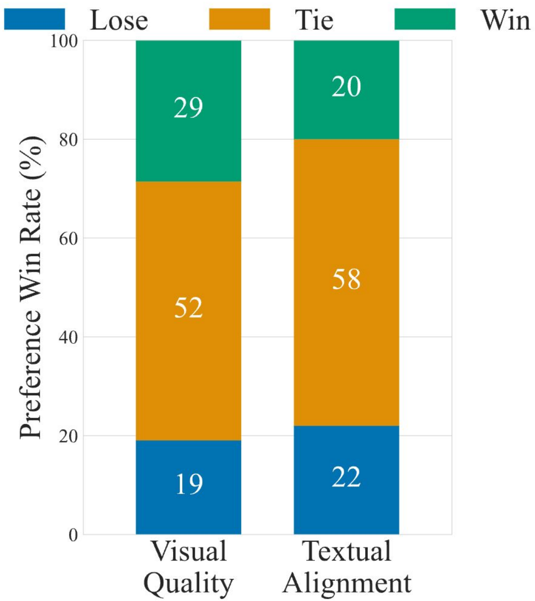
Human prefer Stylus **2x** more than the base model.

CLIP/FID Pareto Curve

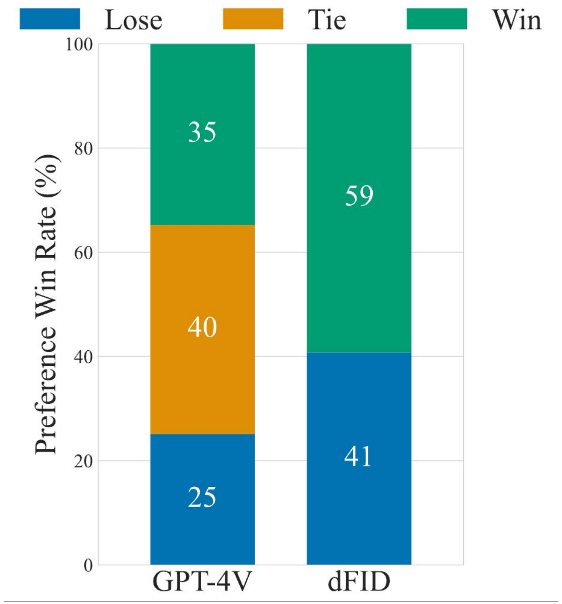


Stylus improves **pareto efficiency** for the CLIP/FID curve.

Vision Language Model (VLM) as a Judge



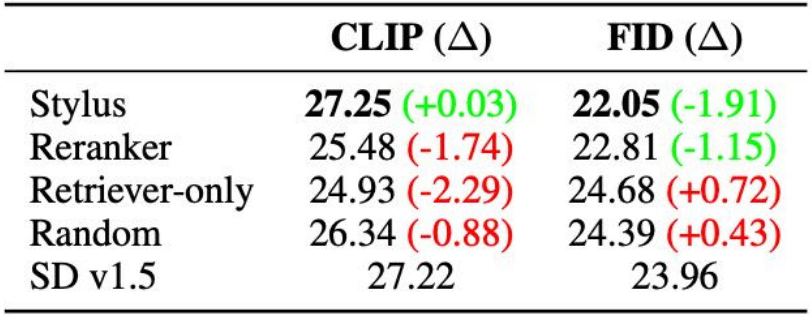
Quality



Diversity

VLM believes Stylus generates **higher quality** and **diverse** images.

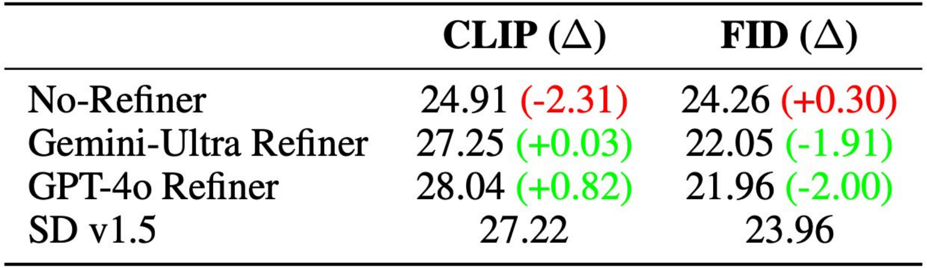
Ablations



	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

Retrieval Methods.

Stylus (with composer) is necessary. Retriever-only (RAG) hurts end2end performance.



	CLIP (Δ)	FID (Δ)
No-Refiner	24.91 (-2.31)	24.26 (+0.30)
Gemini-Ultra Refiner	27.25 (+0.03)	22.05 (-1.91)
GPT-4o Refiner	28.04 (+0.82)	21.96 (-2.00)
SD v1.5	27.22	23.96

Refiner.

Better adapter descriptions lead to much better performance.

Conclusion

Stylus automatically **selects** and **composes** adapters given a user prompt, improving *image fidelity* and *textual alignment*.

Our Contributions

- **StylusDocsv2**: a **75K** entry dataset for adapter descriptions
- Stylus's composer is the first to use an LLM to improve retrieval methods, outperforming rerankers.
- Among the first to employ VLM as a judge for evaluation.