Prompt Sliders for Fine-Grained Control, Editing and Erasing of Concepts in Diffusion Models

Deepak Sridhar

Nuno Vasconcelos

Image Synthesis: Limitations

Current Text-to-Image (T2I) models have

- limited control over fine-grained attributes
  - age, emotions etc.
- difficulty in editing complex features
  - weather, human hands and fingers etc.

InstructPix2Pix Edits

Real Image

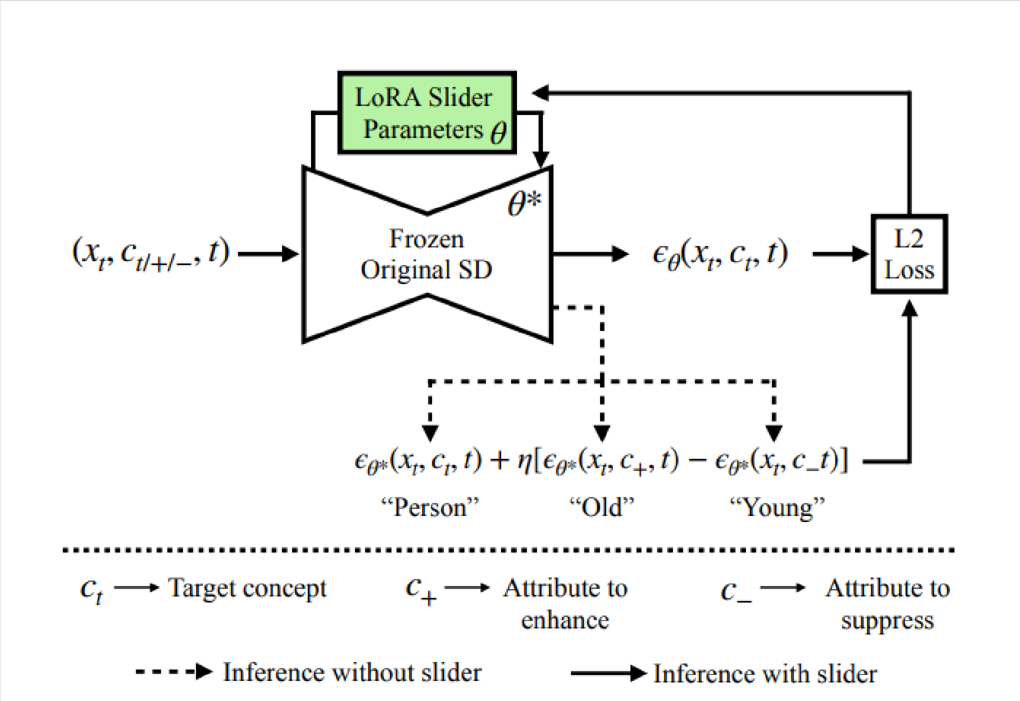
Angry

Older

Winter weather

Concept Sliders[1]: LoRA Adapters for Precise Control in Diffusion

- Introduced a method to train LoRA adapters to learn dedicated direction of a particular concept.
- This is done with a set of positive prompts and negative prompts.
- Increases the likelihood of attributes c^+ and reduces the likelihood of attribute c^- in an image x when conditioned on the target c_t.



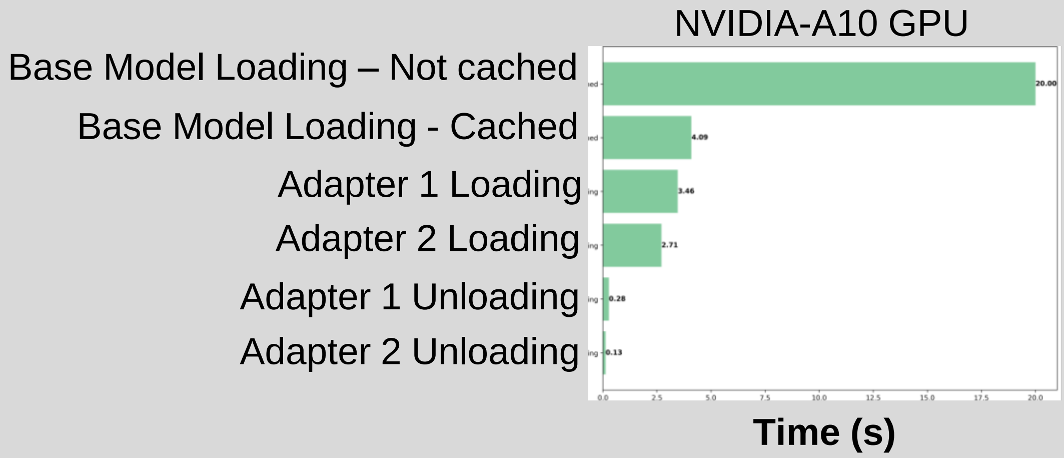
[1]*Gandikota et. al., Concept Sliders: LoRA Adaptors for Precise Control in Diffusion Models, ECCV 2024*

SVCL UCSD

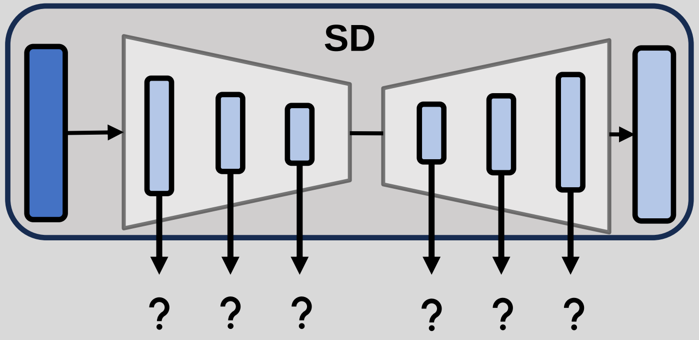
Concept Sliders: A Solution, But Is It Perfect?

- Additional parameters → **Increased model complexity** → Increased memory ↑
  - LoRA sliders require millions of parameters to train.

- Loading/unloading adapters→ **Increased inference time[2]**



- **Model-specific retraining** (SD-XL, SD v1.5) → Less flexibility
  - Requires identifying optimal adapter layers in the model



[2]https://huggingface.co/blog/lora-adapters-dynamic-loading#loading-figures

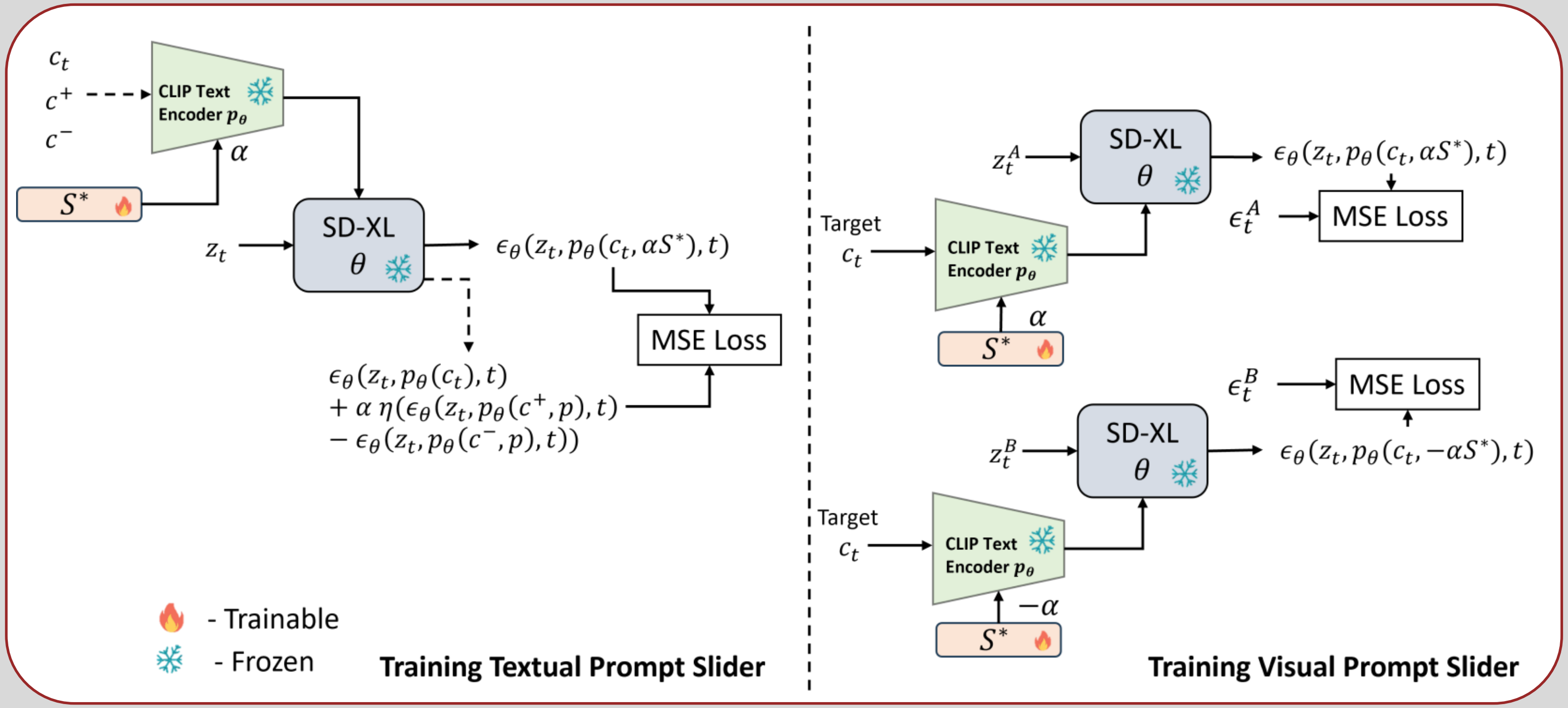
Introducing Prompt Sliders

- A textual inversion method to **learn concepts** via text embeddings.

$$S^*(\alpha) = argmin_S \, E_{\{z\sim E(x), y, \epsilon\sim N(0,1), t\}} \, | \, \epsilon_t(\alpha) - \epsilon_\theta\left(z_t, p_\theta(y, S), t\right)|_2^2$$

$$\epsilon_t(\alpha) = \epsilon_\theta(z_t, p_\theta(c_t), t) + \alpha\, \eta \sum_{\{p \in P\}} \left(\epsilon_\theta(z_t, p_\theta(c^+, p), t) - \epsilon_\theta(z_t, p_\theta(c^-, p), t)\right)$$

- Given a target concept c_t, we propose to learn the corresponding textual embedding $S^* \in R^d$ ($d = 768$ for CLIP text encoder) that encourages the distribution of c_t to exhibit more positive attributes c^+ and fewer negative attributes c^-.

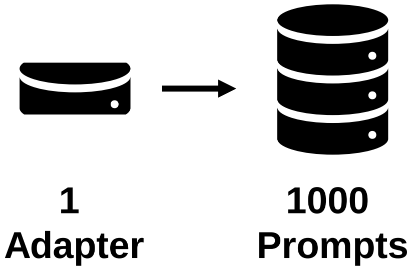


SVCL UCSD

Prompt Sliders

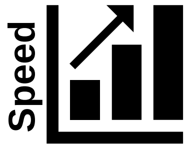
Efficient and Lightweight

1. No need for additional parameters like LoRAs.
2. Each concept requires only 3KB of storage



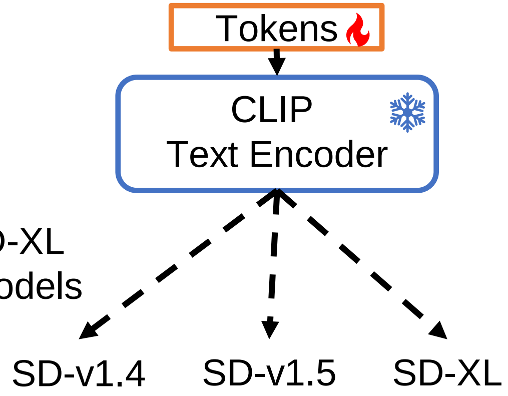
Faster inference

1. Overcomes the issue of loading/unloading adapters
2. 30% speed improvement over adapters
3. Adjust concept strength via text embedding weights



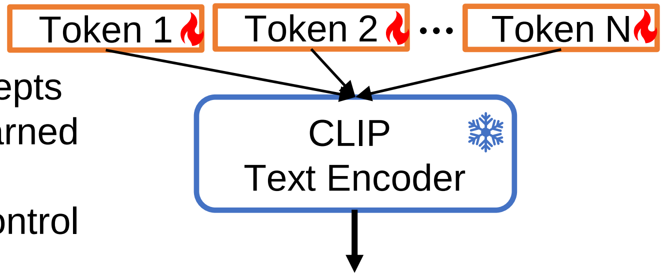
Cross-Model Flexibility

1. Generalizable across models sharing the same text encoder
2. For example, SD v1.4, v1.5, SD-XL
3. Retains performance across models



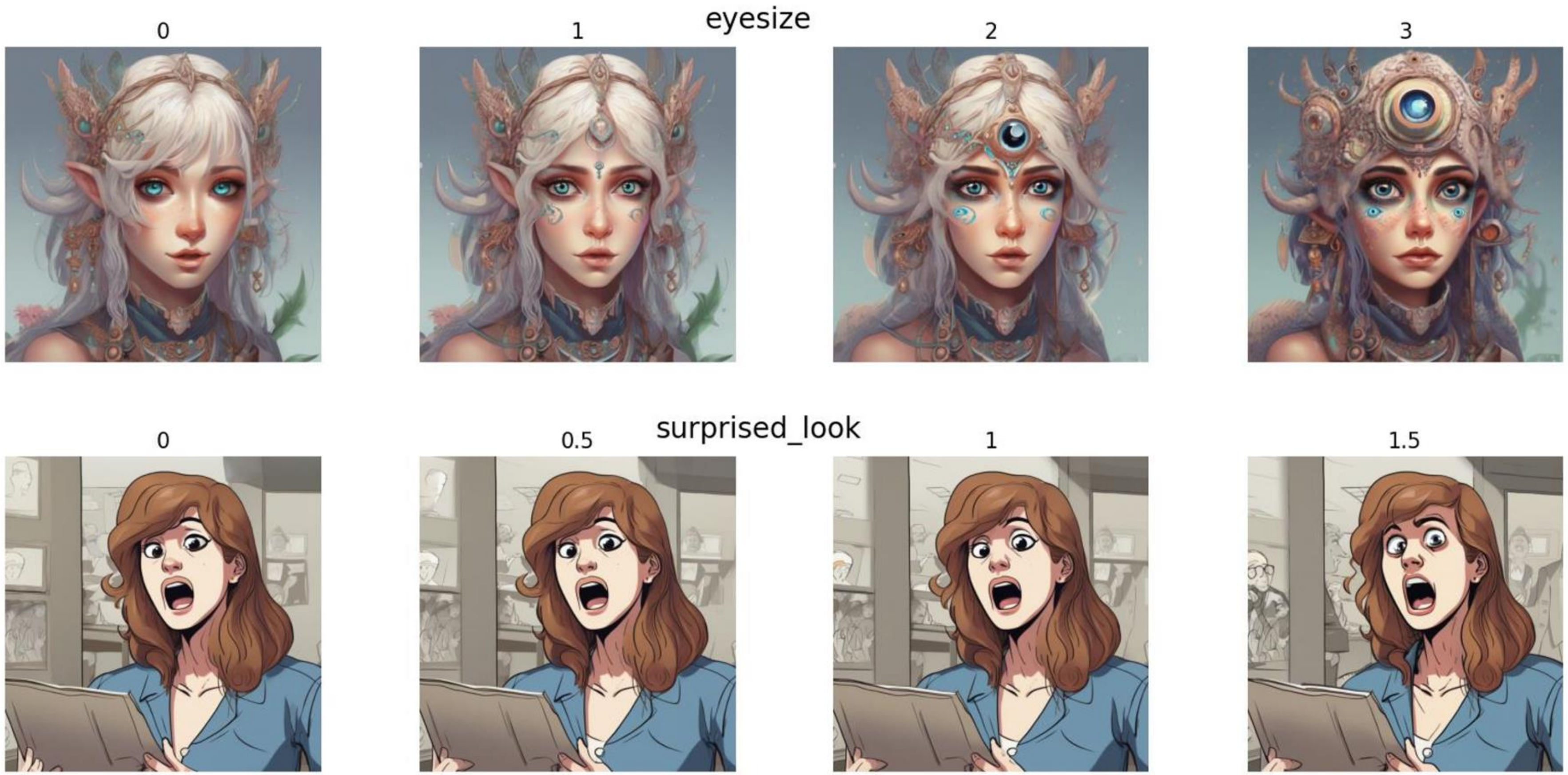
No Merging Issues

1. Combine multiple concepts easily by appending learned tokens to the prompt
2. Retains independent control

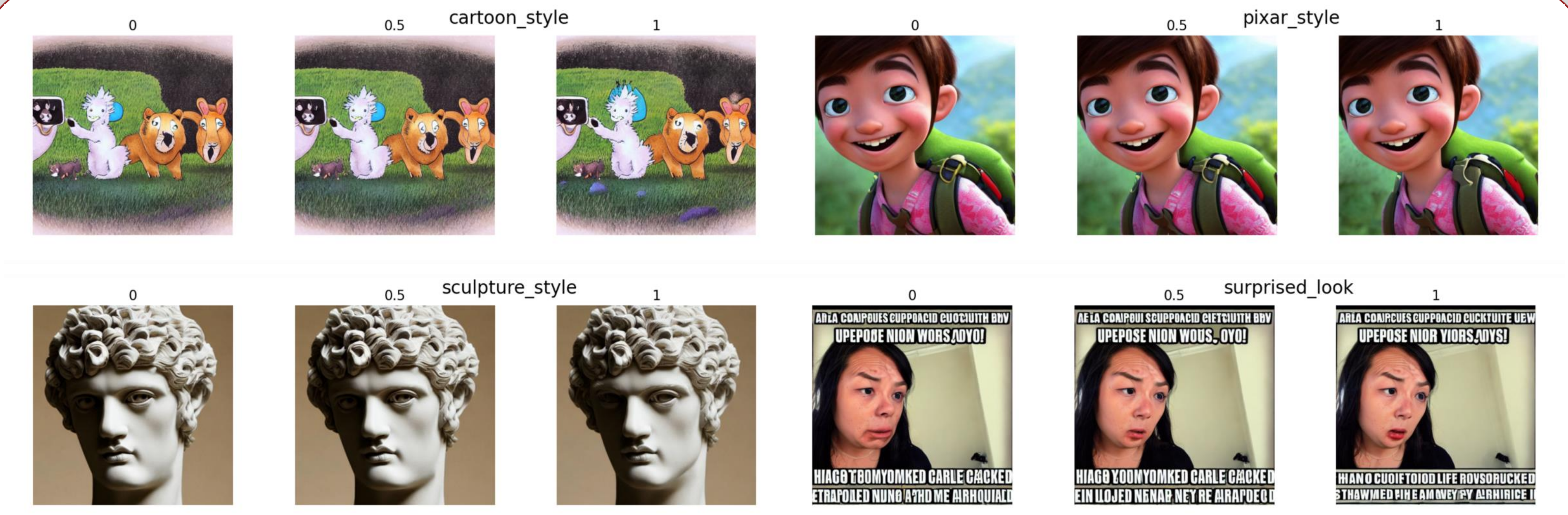


Prompt Sliders for Various Concepts

Prompt Sliders for Various Concepts



Results: Transfer to SD-v1.4 from SD-XL

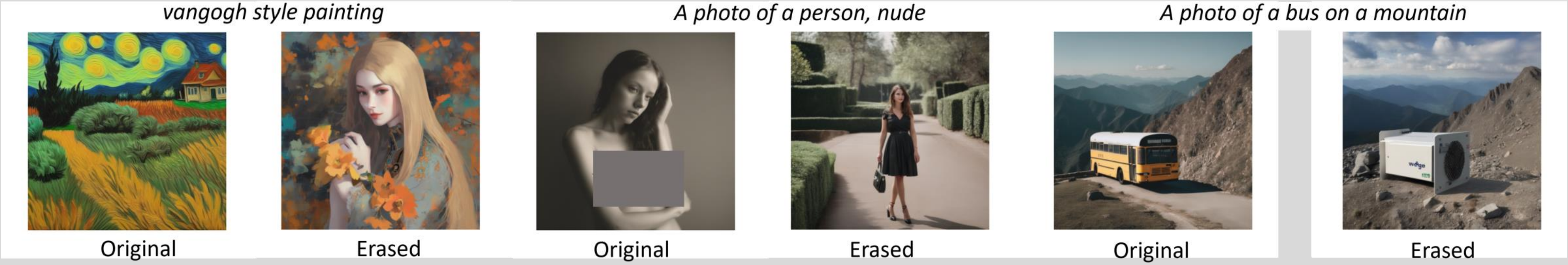


Results: Transfer to SD-v1.5 from SD-XL

Erasing Concepts with Prompt Sliders

- Using a negative α allows one to erase a concept instead of enhancing them. Formally,

$$\epsilon_t(\alpha) = \epsilon_\theta(z_t, p_\theta(c_t), t) - \alpha\,\eta \sum_{\{p \in P\}} \left(\epsilon_\theta(z_t, p_\theta(c^+, p), t) - \epsilon_\theta(z_t, p_\theta(c^-, p), t)\right)$$

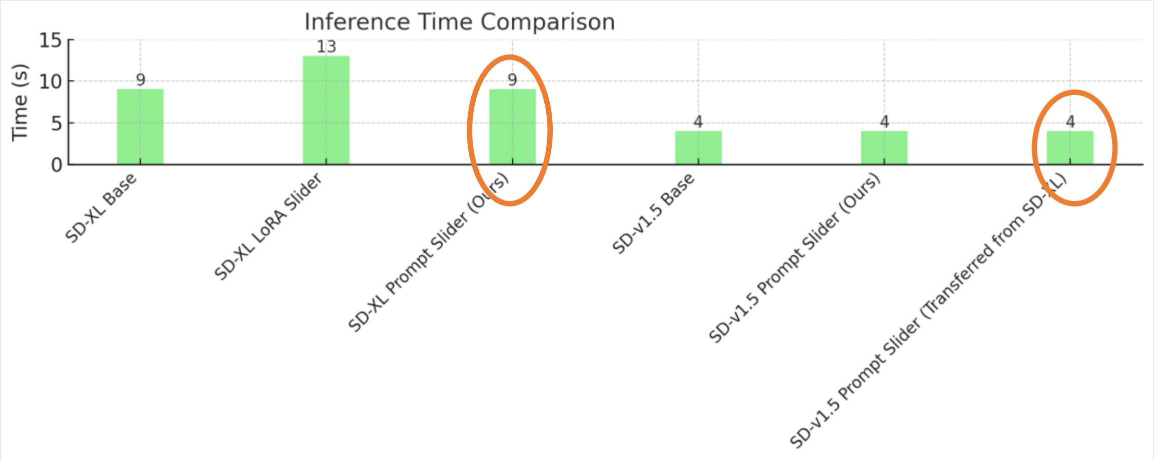


Composition of Prompt Sliders

Prompt sliders are simple to compose by just appending the learned tokens to the input prompt.

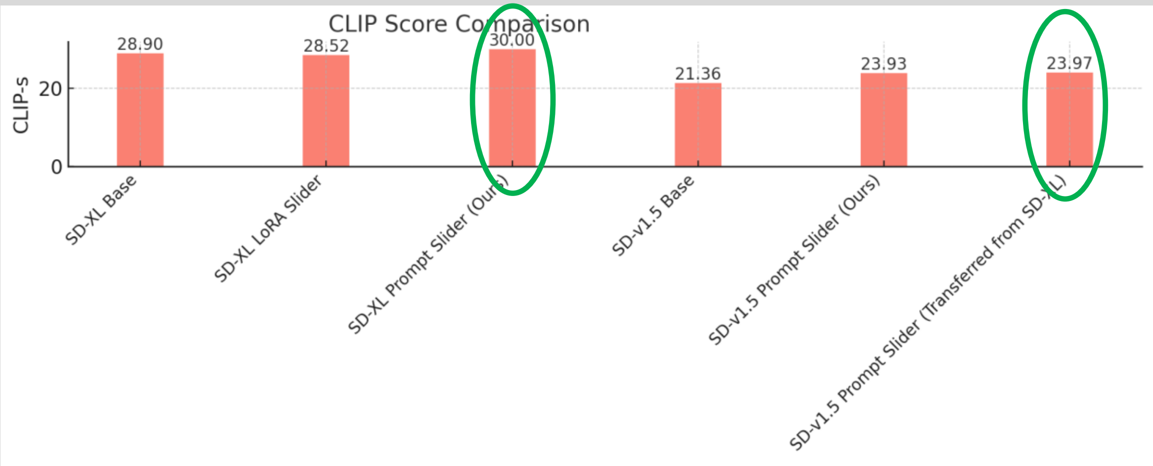
SVCL UCSD

Comparison of Inference times and Prompt Slider transfers



- Does not increase the inference time from the baseline without prompt sliders.

- Prompt Sliders enhance image-text alignment, as shown by improved CLIP scores.
- Transferring Prompt Sliders from SD-XL to SD-1.5 retains performance similar to training from scratch on SD-1.5.



Comparison with Concept Sliders

Emotion Prompt Sliders Applied on a Real Image with Inversion[3]

Real Image

[3]Brack et. al., LEDITS++: Limitless Image Editing using Text-to-Image Models, CVPR 2024

Concept Prompt Sliders Applied on a Real Image with Inversion[3]

Real Image

Manga

Borderlands

Zombie

Cyborg

Hat

Sunglasses

[3]Brack et. al., LEDITS++: Limitless Image Editing using Text-to-Image Models

Next Steps…

- Limitations
  - Image Quality deteriorates or diverges from the original image at higher guidance strength α
  - Cannot cover concepts absent in the original diffusion model without using reference images.
- Future research
  - Improve performance at higher guidance strength
  - Learn multiple concepts together with disentanglement

Thank You

Questions?

Project Page