UNIVERSITÄT
LEIPZIG

Manipulating Embeddings of Stable Diffusion Prompts to Control Image Compositionality

Bachelor Thesis

Dinara Imambayeva

Supervised by Niklas Deckers

Leipzig, 18.01.2024

OVERVIEW

- Motivation
  - Problem Description
- Theoretical Background
- Related Works
- Approach
- Experimental Setup
- Results
- Further Approaches
- Summary
- Outlook

MOTIVATION

Image generation with Stable Diffusion

- User expectation: the generated image corresponds to the prompt

`A picture of a white cat and a black cat` → Stable Diffusion →

MOTIVATION

Image generation with Stable Diffusion

- Reality:

A picture of a white cat and a black cat → Stable Diffusion →

failed composition of multiple objects

failed attribute binding

missing objects

MOTIVATION: PROBLEM DESCRIPTION

- Often failed image compositionality:
  - Color leakage
  - Incorrect number of objects
  - Missing objects
  - Failed attribute binding
  - Failed composition of multiple objects

- Problem: generated images do not satisfy the user

MOTIVATION: PROBLEM DESCRIPTION

- Possible approaches:
  - Trying different seeds

seed = 154

seed = 510321

seed = 99

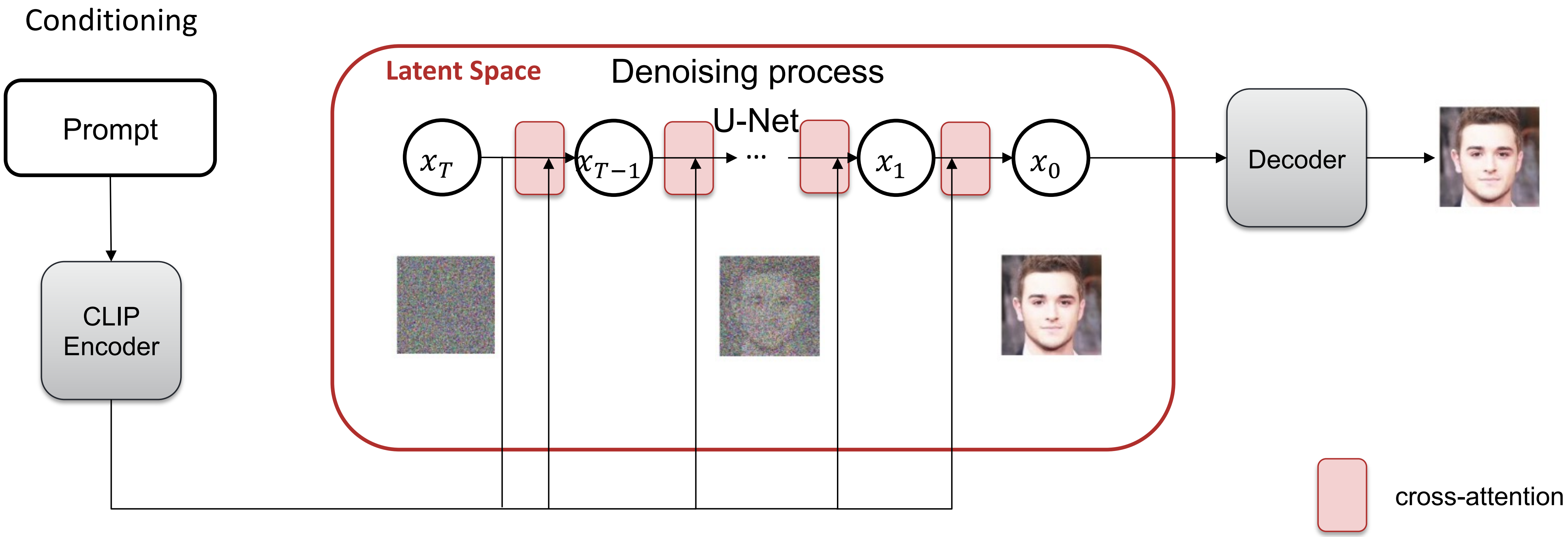
MOTIVATION: PROBLEM DESCRIPTION

- Possible approaches:
  - Prompt engineering

- Requires sometimes many attempts
- Still not desired output
- No user control, user frustration

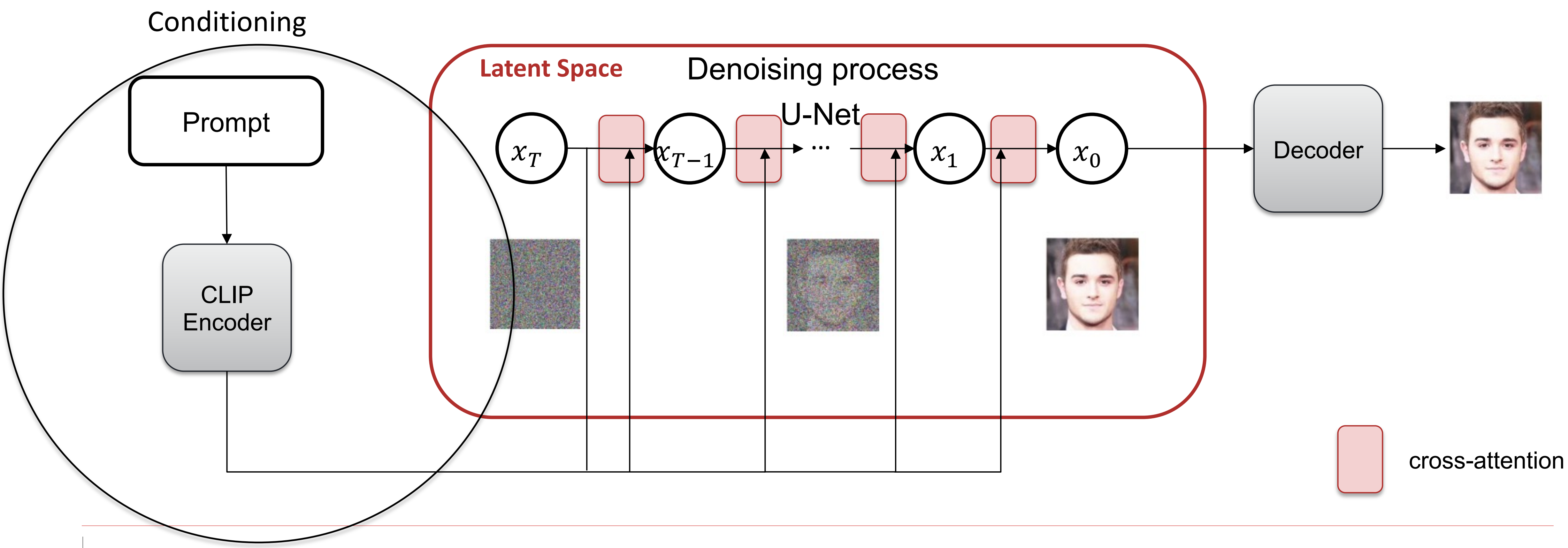
THEORETICAL BACKGROUND

- Image generation with Stable Diffusion



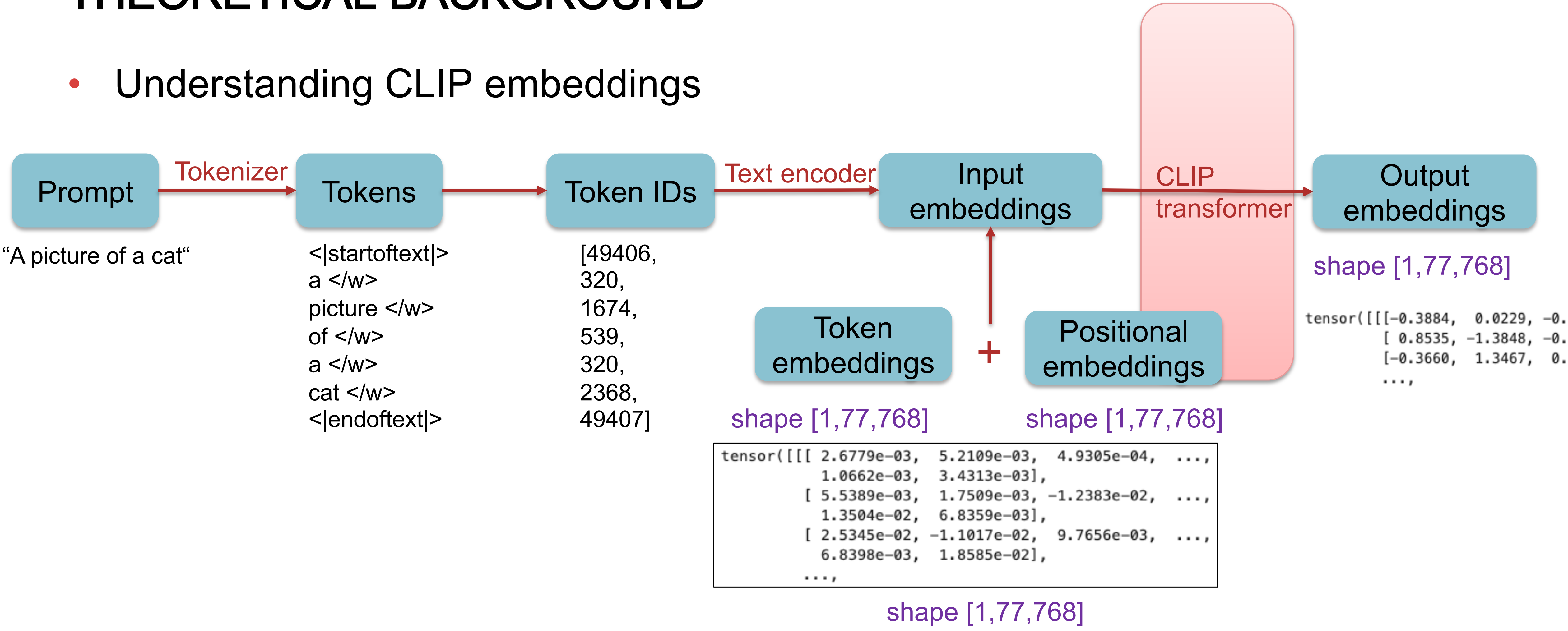
THEORETICAL BACKGROUND

- Image generation with Stable Diffusion



THEORETICAL BACKGROUND

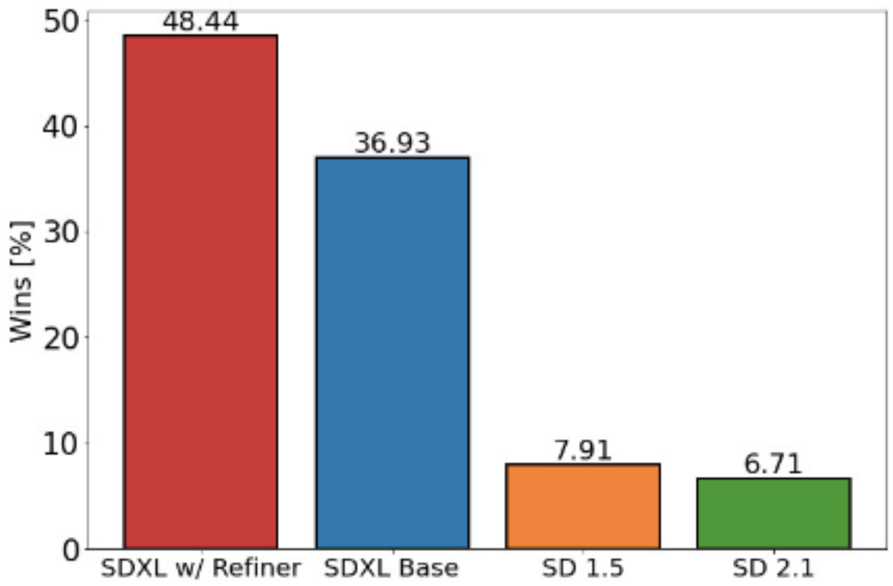
- Understanding CLIP embeddings



THEORETICAL BACKGROUND

Stable Diffusion XL

- Released in July 2023
- High-resolution image synthesis
- Higher fidelity
- Modified architecture:
    - Larger U-Net
    - 2 CLIP text encoders
- Additional refinement model



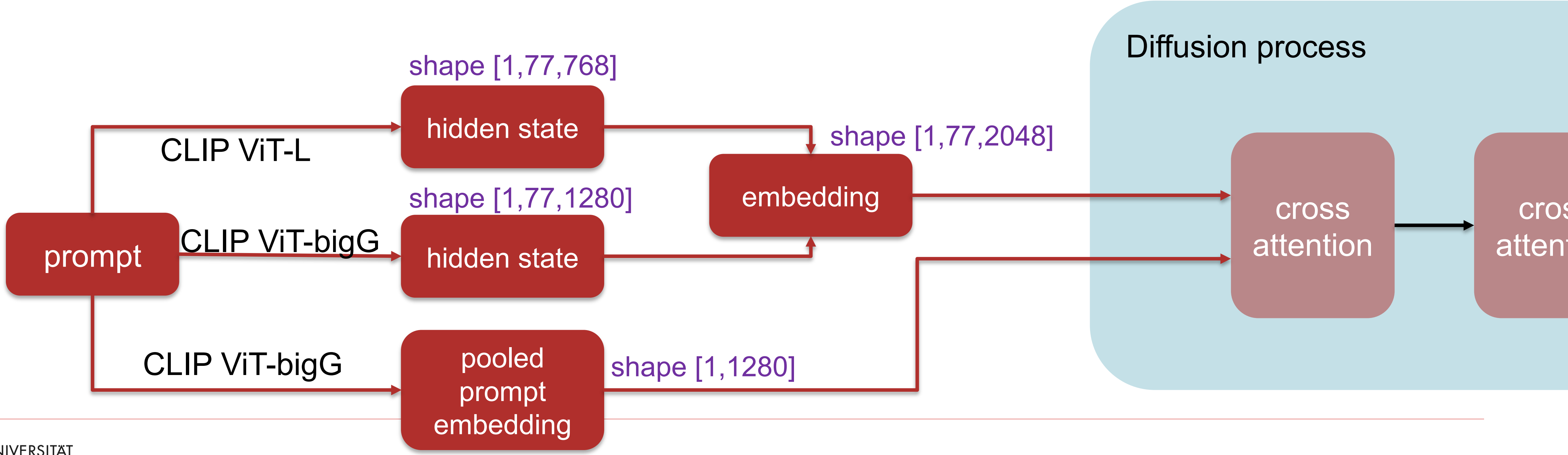
Comparison of user preferencies

Source: Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., ... & Rombach, R. (2023). Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

THEORETICAL BACKGROUND

Stable Diffusion XL

- 2 CLIP text encoders for high-resolution image synthesis
  - CLIP ViT-L
  - CLIP ViT-bigG



THEORETICAL BACKGROUND

Stable Diffusion XL

A picture of a white cat and a black cat

Stable Diffusion XL

THEORETICAL BACKGROUND

Stable Diffusion XL

A picture of a white cat and a black cat

Stable Diffusion XL

Still failed compositionality!

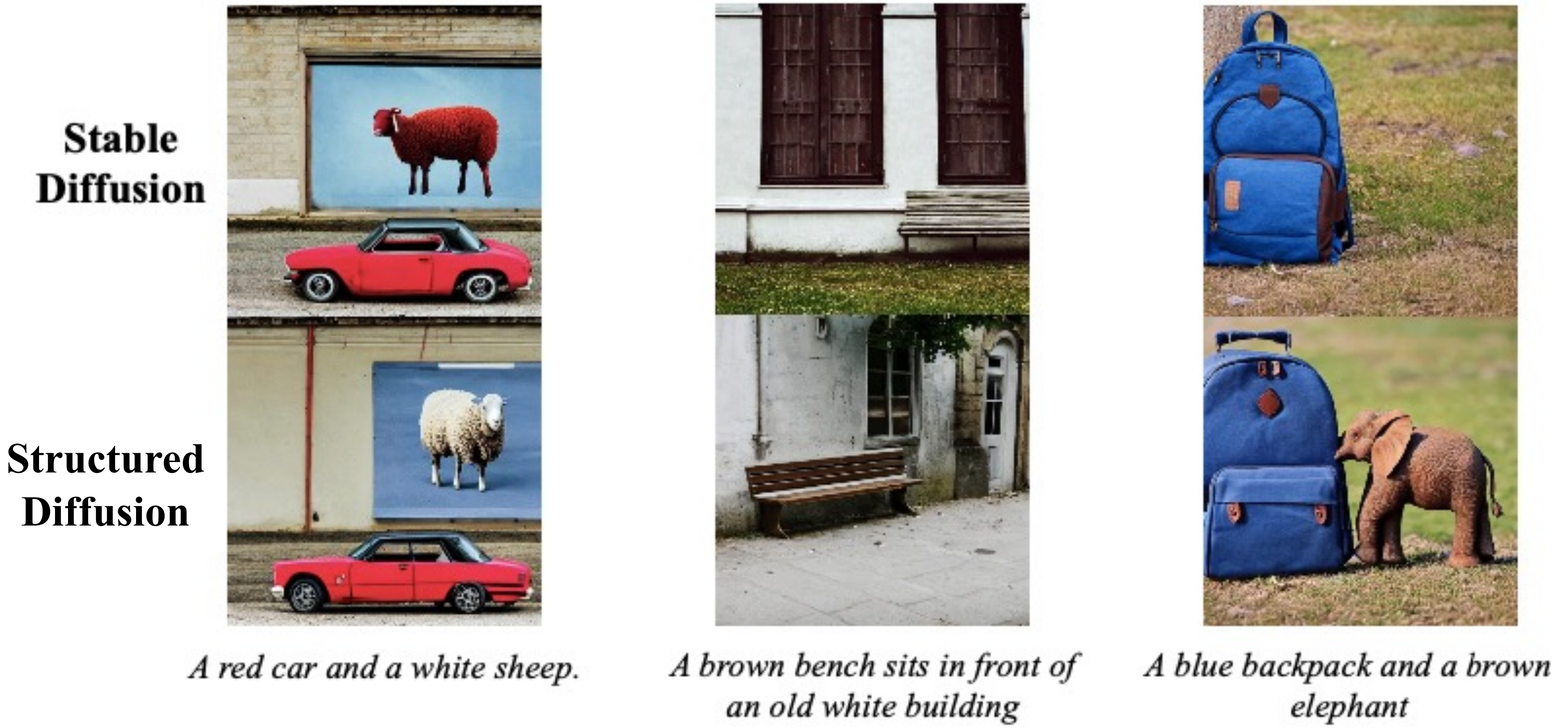
RELATED WORKS: COMPOSABLE DIFFUSION

- Compositional image generation
- Each concept is generated separately
- 2 compositional operators: Conjunction and Negation

Source:Liu, N., Li, S., Du, Y., Torralba, A., & Tenenbaum, J. B. (2022, October). Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*(pp. 423-439). Cham: Springer Nature Switzerland.

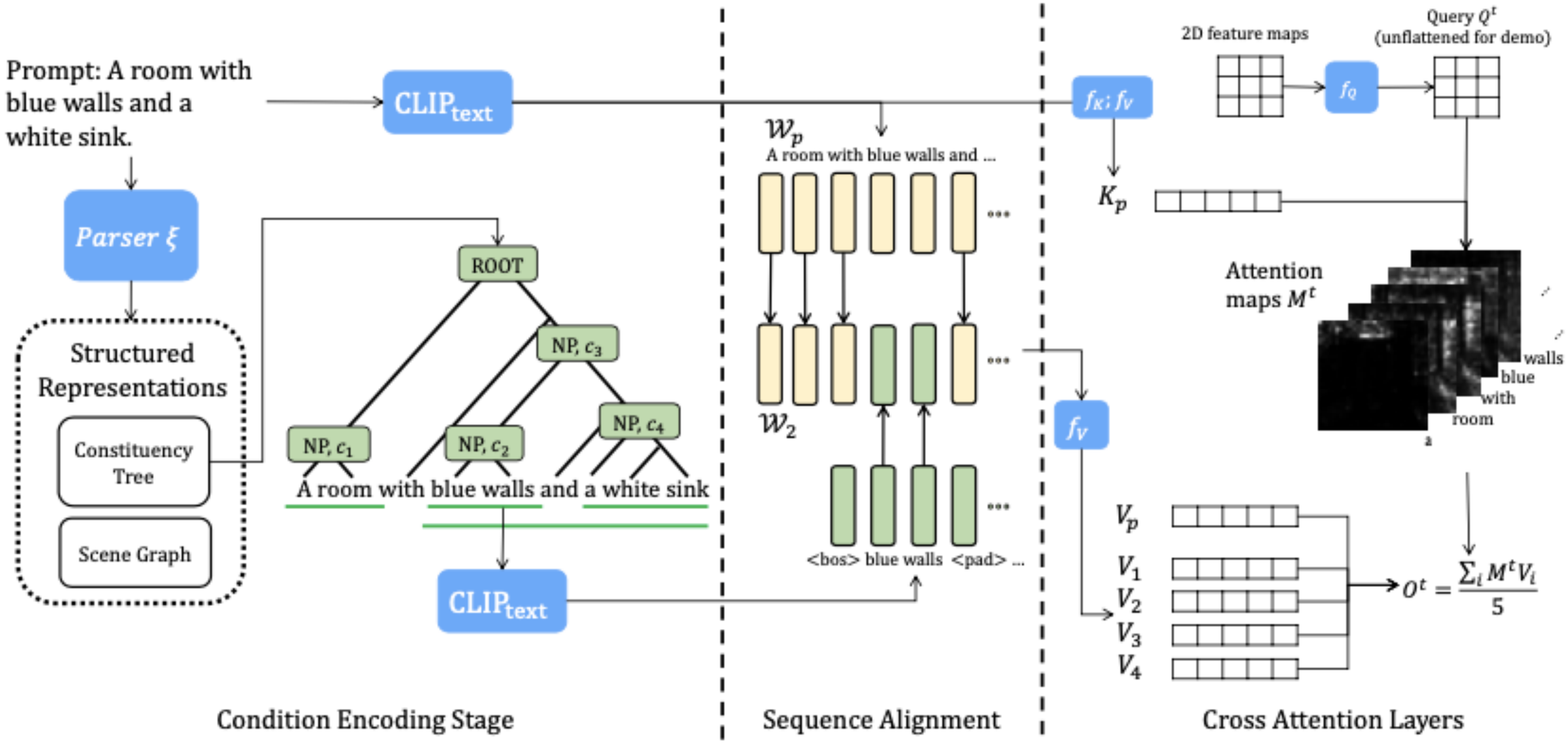
RELATED WORKS: STRUCTURED DIFFUSION

- Training-free guidance for compositional text-to-image synthesis
- **Idea**: separate encoding of noun phrases combined with manipulation in cross-attention layers
- 5-8% advantage compared to Stable Diffusion



Source: Feng, W., He, X., Fu, T. J., Jampani, V., Akula, A., Narayana, P., ... & Wang, W. Y. (2022). Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*.

RELATED WORKS: STRUCTURED DIFFUSION



Source: Feng, W., He, X., Fu, T. J., Jampani, V., Akula, A., Narayana, P., ... & Wang, W. Y. (2022). Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*.

RELATED WORKS: POOLING

- Prompt manipulation by pooling the noun phrases
- Idea: additional attribute binding through pooling

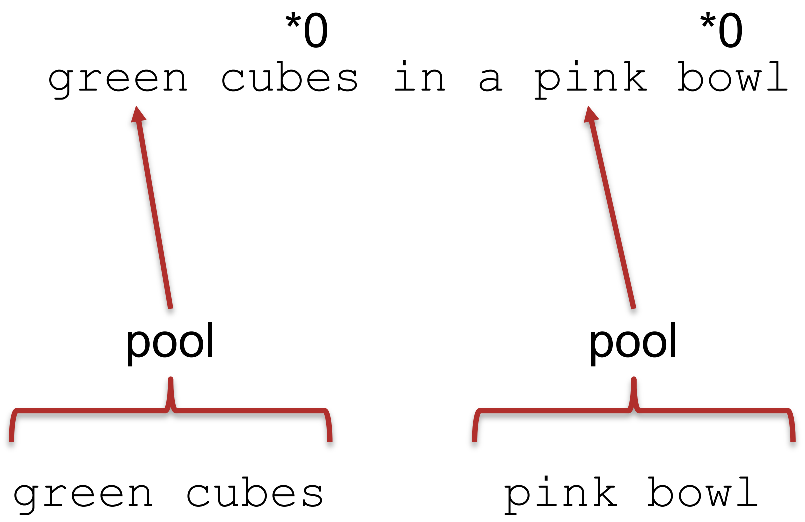
`green cubes in a pink bowl`
without NP pooling

`green cubes in a pink bowl`
with NP pooling

Source: https://colab.research.google.com/drive/1izMKdvBMfThVSRp8Tg4Fc1eiGVP0NKZP?usp=sharing#scrollTo=wa3iaNL3Mq28

RELATED WORKS: POOLING

- Common technique in NLP
- Conversion of the NP embedding vectors into a single vector
- Pooling:
  - Maximum pooling
  - Mean square pooling
  - Mean pooling

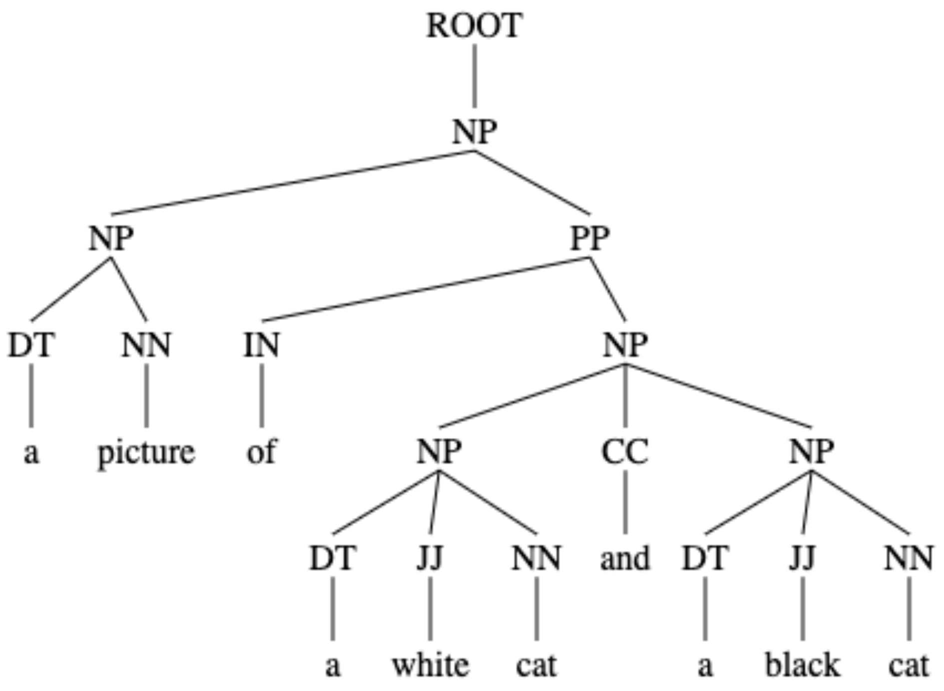


$\boxed{[a,b].\mathrm{mean}/\sqrt{2}}$

a – vector for `green`
b – vector for `cubes`

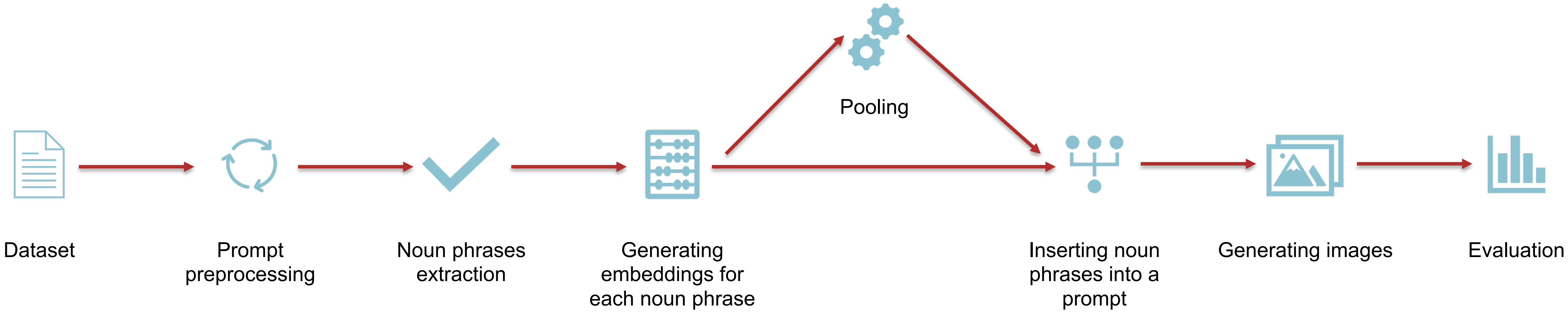
APPROACH

- Combining Structured Diffusion approach with pooling
- Extract noun phrases from the constituency tree for each prompt
- Embedding manipulation:
  - Separate embedding of NPs
  - Additional pooling of NPs



- Adjusted approach for SDXL:
  - Generating embeddings with both CLIP encoders
  - No manipulation in pooled prompt embeddings

EXPERIMENTAL SETUP



EXPERIMENTAL SETUP

Not every prompt can demonstrate the problem

Datasets

- **CC-500** (Concept Conjuction dataset)
- Prompt format: “a [colorA] [objectA] and a [colorB] [objectB]”

- **ABC-6K** (Attribute Binding Contrast dataset)
- 3200 prompt pairs
- „a kitchen with white appliances and brown cupboards“
- „a kitchen with brown appliances and white cupboards“

EXPERIMENTAL SETUP

Technical setup

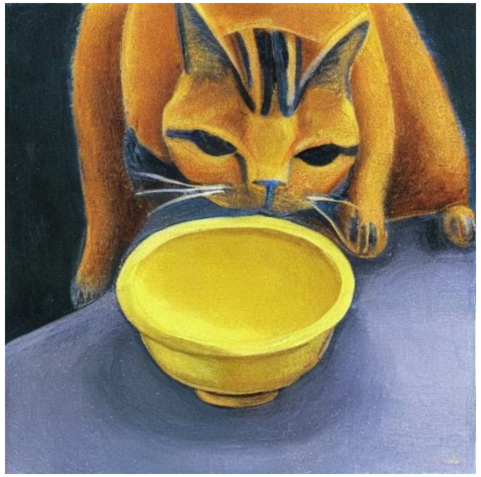
- Stable Diffusion Model: "runwayml/stable-diffusion-v1-5“
- Stable Diffusion XL Base 0.9 (released July 2023)

- Fixed seed: 0

RESULTS: CC500

SD 1.5 | Embedding NPs | Embedding NPs + Pooling

a yellow bowl and a blue cat



a red bowl and a blue cup



RESULTS: CC500

`a yellow car and a red cat`

SD XL

Embedding NPs

Embedding NPs + Pooling

RESULTS: CC500

`a pink cow and a brown sheep`

SD XL

Embedding NPs

Embedding NPs + Pooling

RESULTS

Evaluation criteria for manual evaluation on CC-500

- Adopted from Structured Diffusion

Categories:

1. Zero or one object is depicted
2. Two objects are depicted
3. Two object are depicted with correct colors
4. Image looks natural

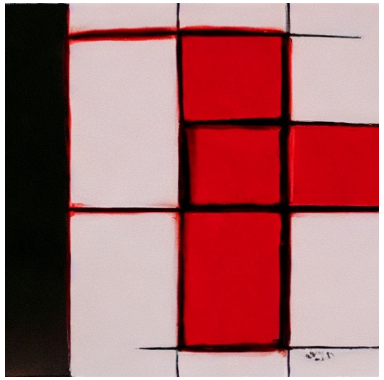
RESULTS

SD 1.5	Zero/One object	Two objects	Two objects with correct color	Natural looking image
SD	82%	12%	6%	48%
NP embedding	82%	10%	8%	40%
NP embedding + pooling	94%	6%	0%	34%

SD XL	Zero/One object	Two objects	Two objects with correct color	Natural looking image
SD	40%	42%	18%	44%
NP embedding	42%	44%	14%	30%
NP embedding + pooling	40%	46%	14%	66%

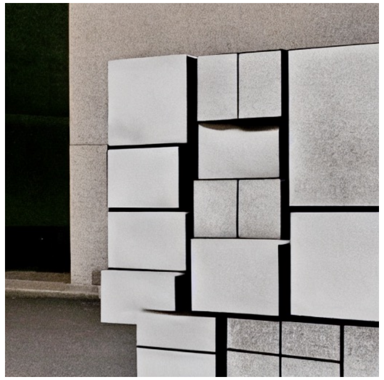
FURTHER APPROACHES: SECOND APPROACH

Idea: Can we subtract/add the color to improve attribute binding?



red cube

-



cube

+

car

=

Generated by Stable Diffusion 1.5, seed = 0

FURTHER APPROACHES: SECOND APPROACH

Experiment on SDXL

`A white cat with black ears and markings`

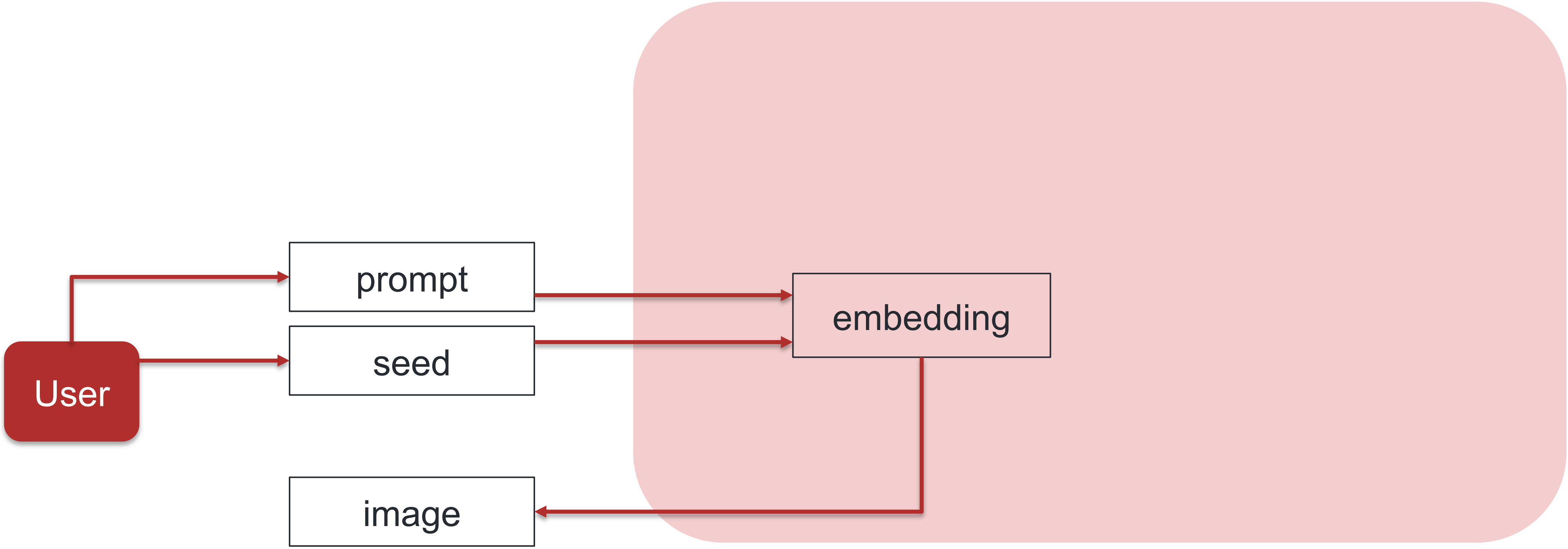
No modification

NP embedding + pooling

Adding `black` vector to the according NP

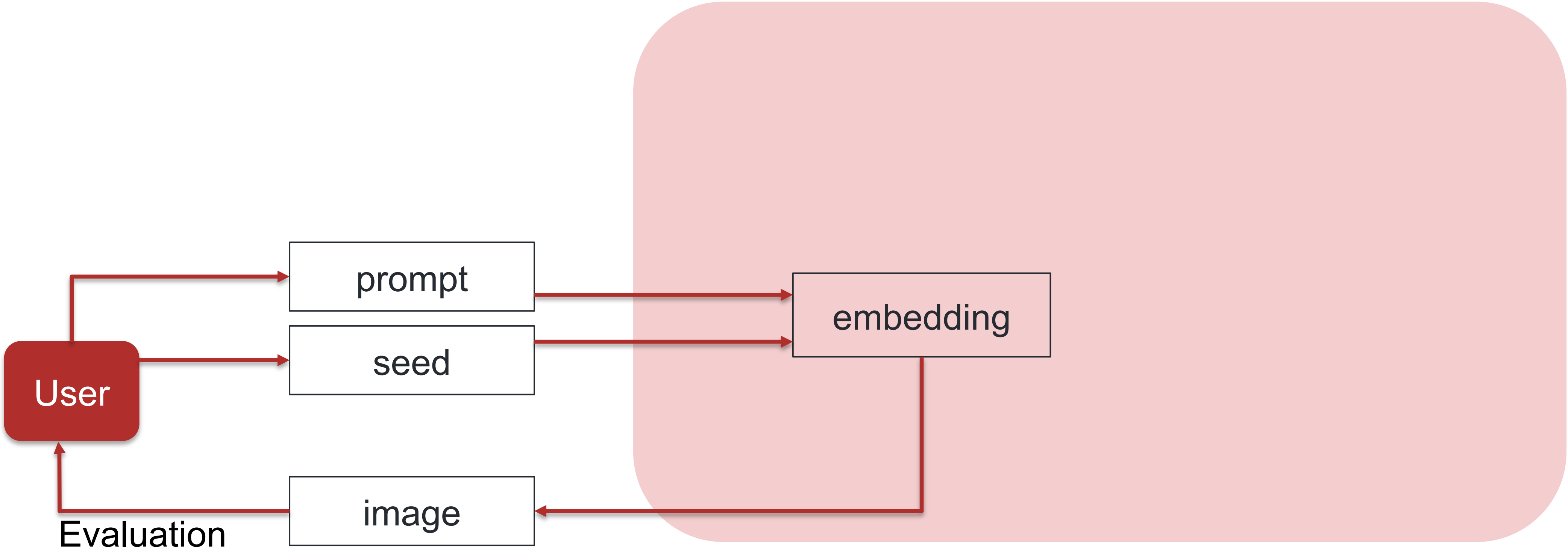
FURTHER APPROACHES: SECOND APPROACH

Idea: User interaction interface for image adjustment



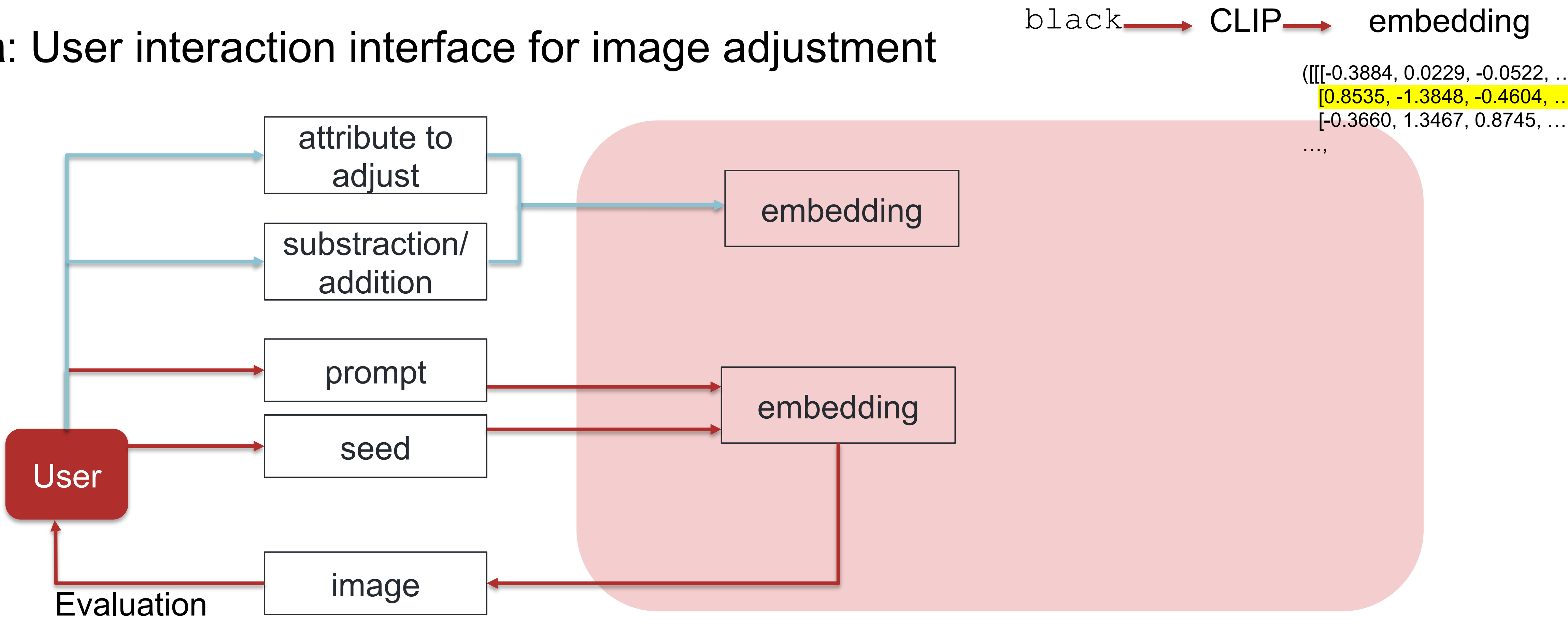
FURTHER APPROACHES: SECOND APPROACH

Idea: User interaction interface for image adjustment

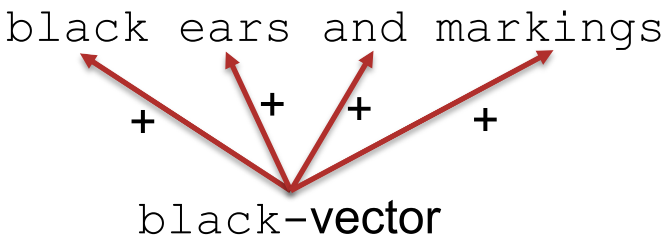


FURTHER APPROACHES: SECOND APPROACH

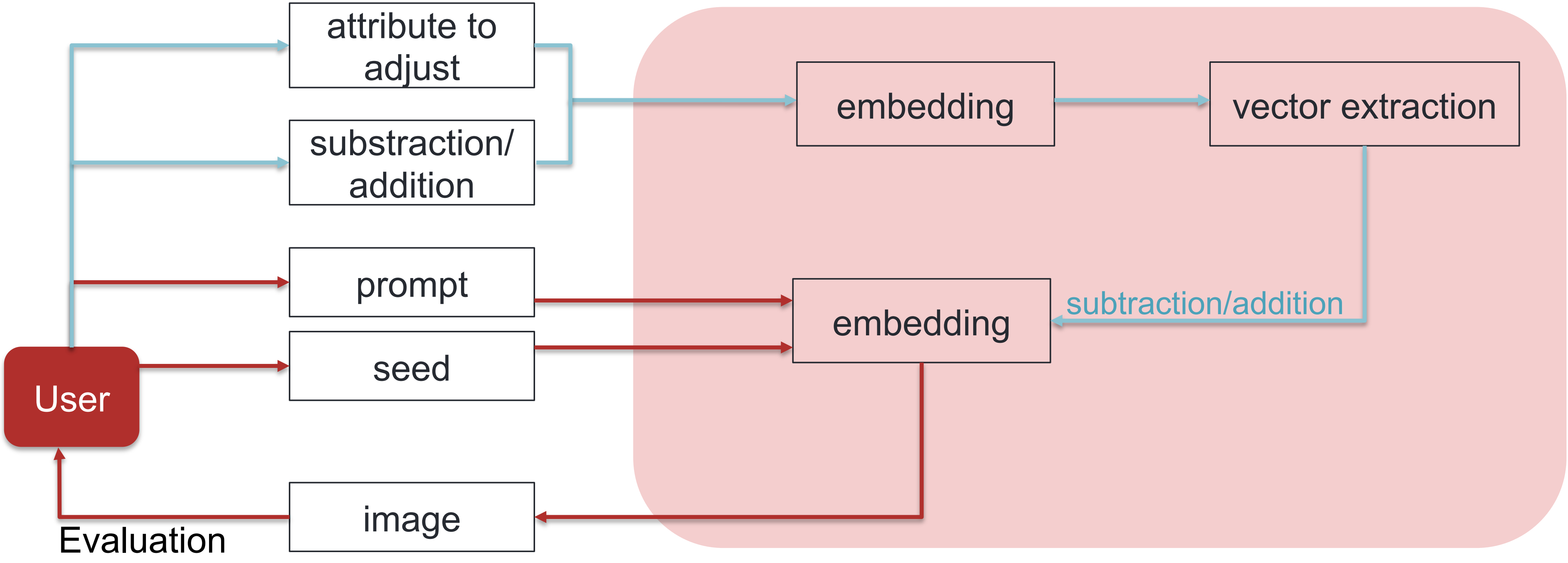
Idea: User interaction interface for image adjustment



FURTHER APPROACHES: SECOND APPROACH

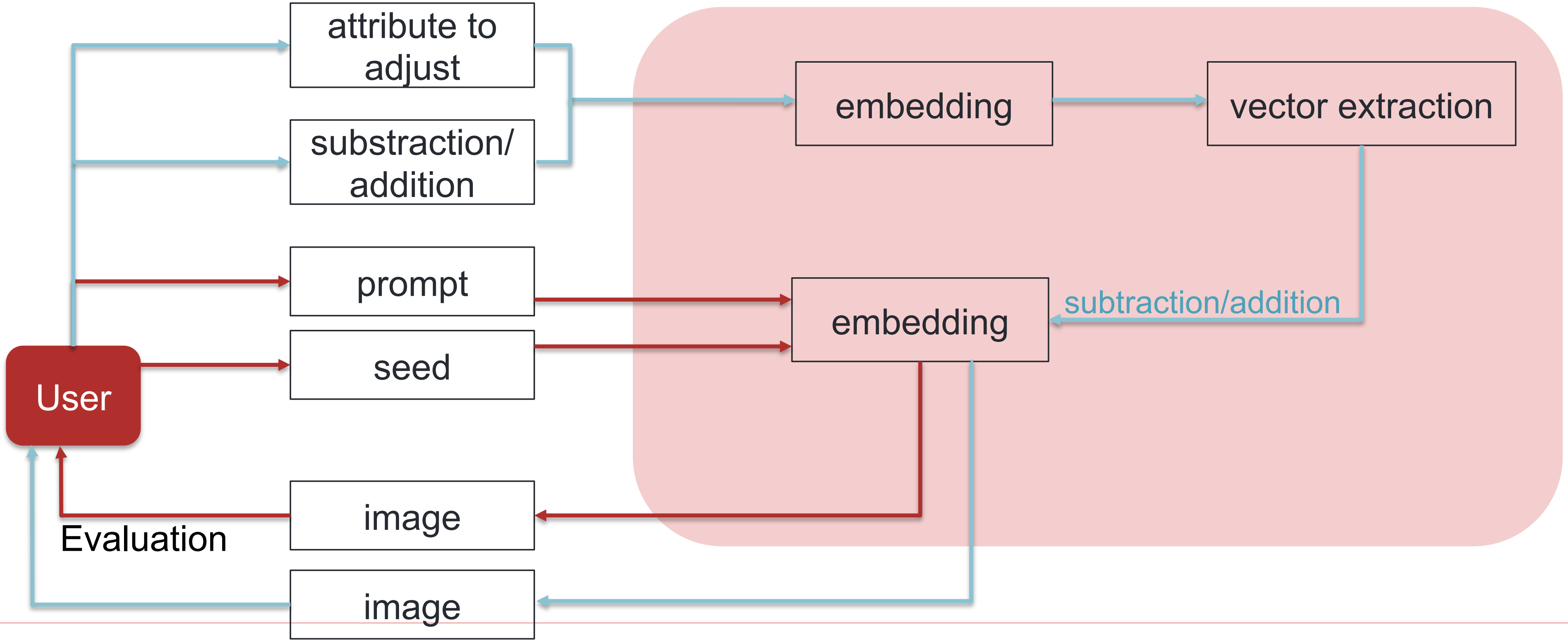


Idea: User interaction interface for image adjustment



FURTHER APPROACHES: SECOND APPROACH

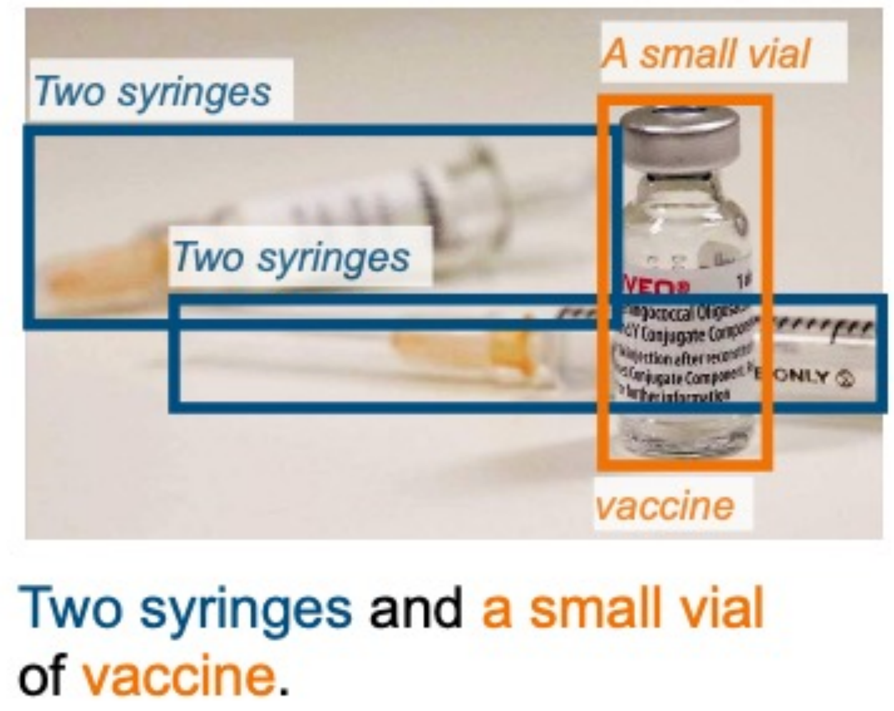
Idea: User interaction interface for image adjustment



FURTHER APPROACHES: EVALUATION

Automatic evaluation with GLIP

- Framework for object detection and phrase grounding
- Phrase grounding – identifying the correspondence between phrases in a prompt and objects in an image
- Object detection as a grounding task



Source: Li, L. H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., ... & Gao, J. (2022). Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10965-10975).

SUMMARY

- Failed compositionality problem limits user control over output
- Improving image compositionality approaches needed
- My approach:
    - Separate embedding of NPs
    - Additional pooling of NPs

- Limitation: architecture-dependent approach

OUTLOOK

Current Objectives:

- Further image generation and evaluation of results (CC-500 and ABC-6K)
- Implementation of automatic evaluation with GLIP

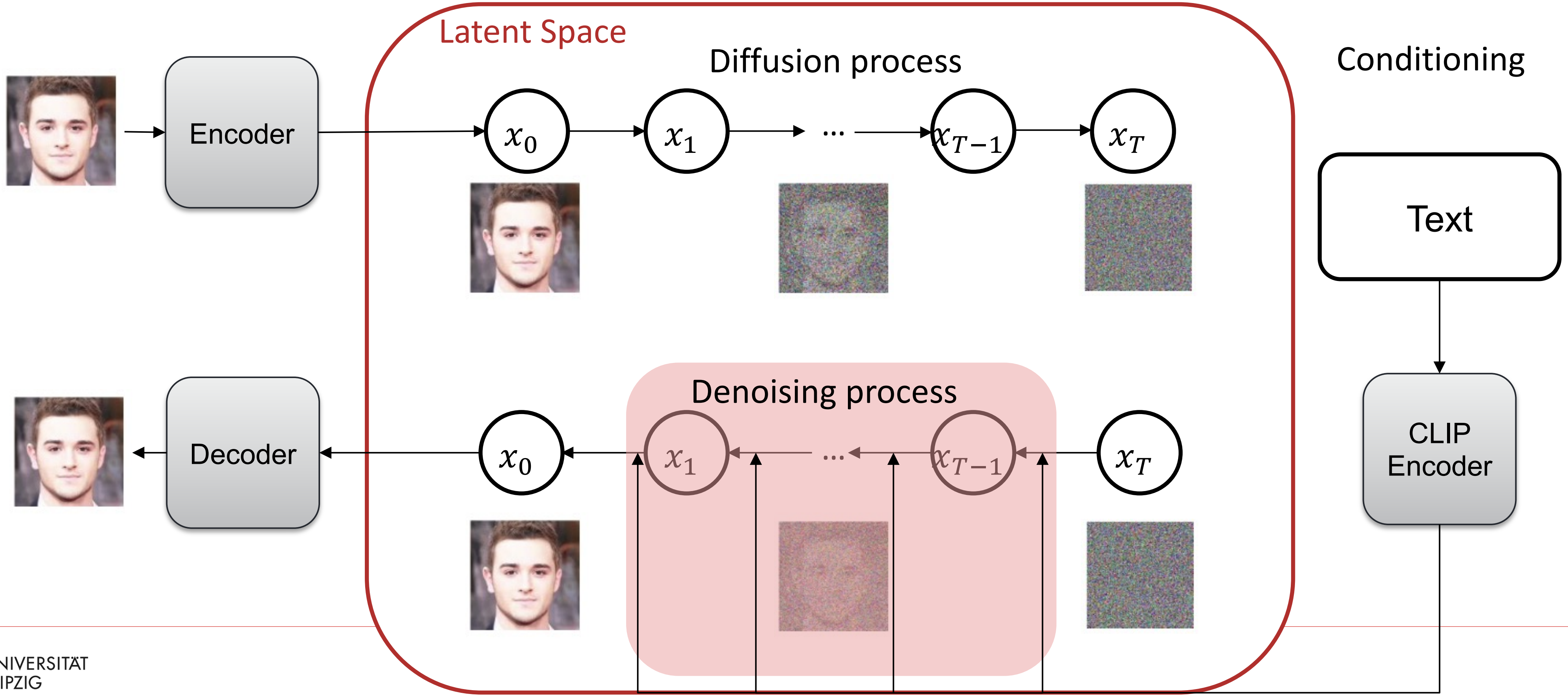
Future Objectives (beyond thesis):

- Refinement of the second approach
- Implementation of a user interface

BACKUP SLIDES

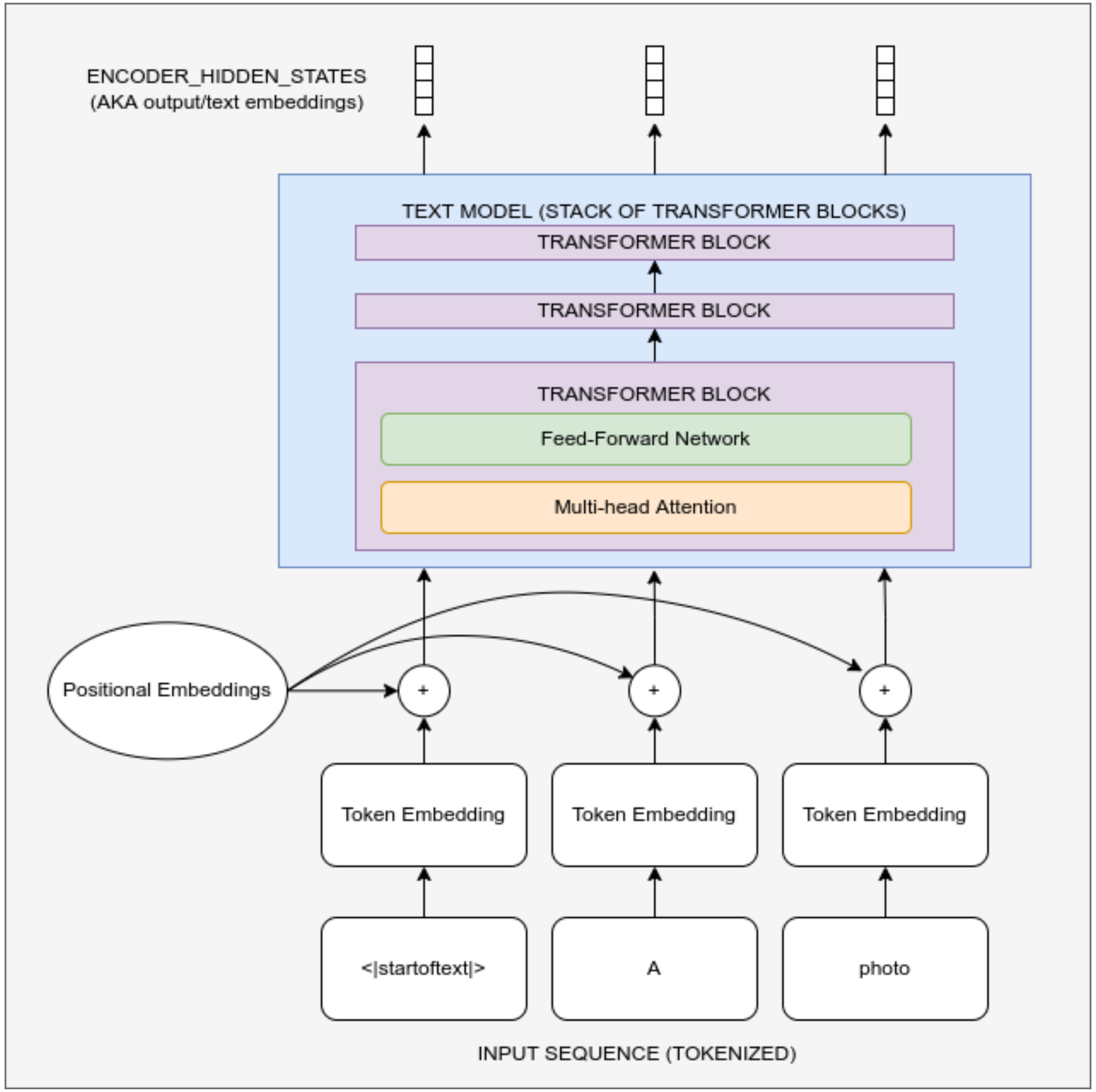
TECHNICAL BACKGROUND

– Image generation with Stable Diffusion



CLIP

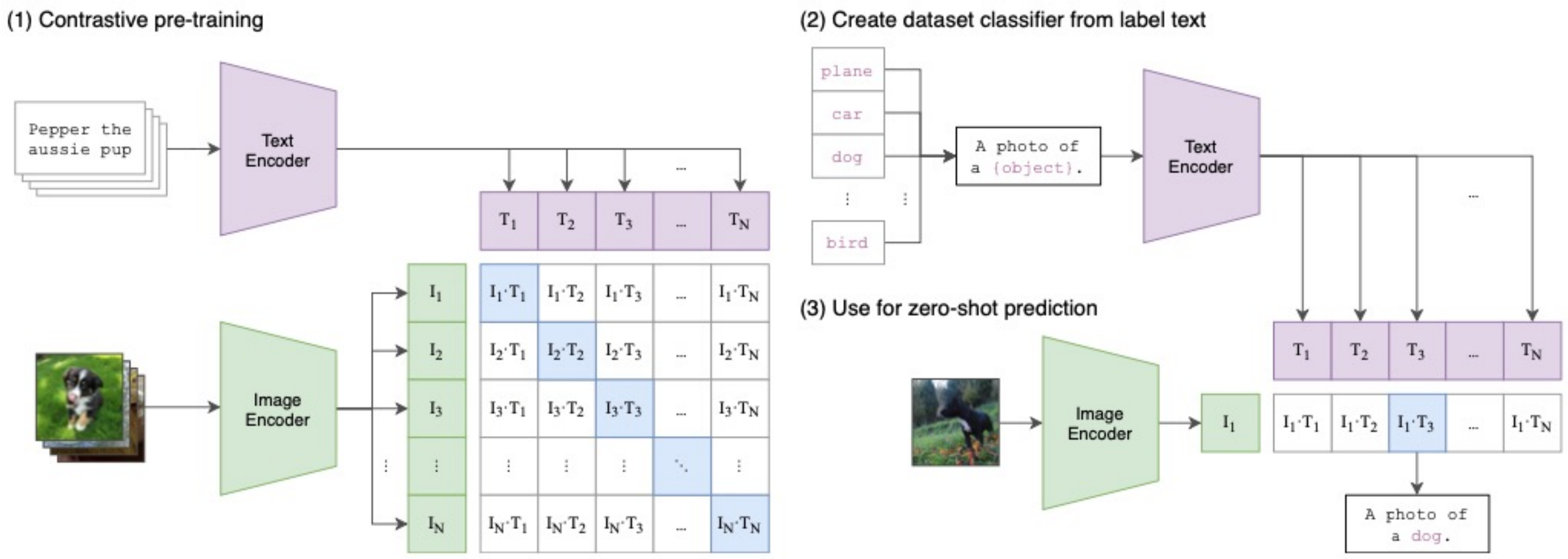
- Prompt-embedding pipeline
- CLIP encoder has transformer architecture:
    - 12 layers
    - 8 attention-heads



Source: https://colab.research.google.com/github/fastai/diffusion-nbs/blob/master/Stable%20Diffusion%20Deep%20Dive.ipynb#scrollTo=297340b2-08ef-4fa4-a19a-0beffff7701b

CLIP

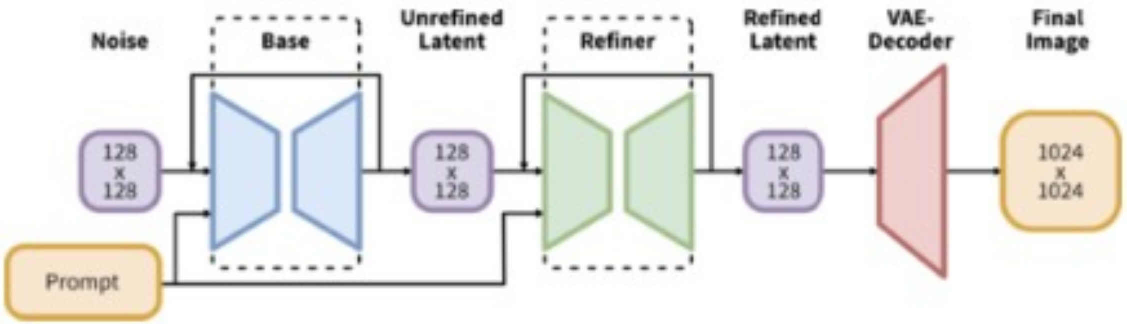
- CLIP text encoder = transformer
- Training in multimodal space



Source: Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR.

STABLE DIFFUSION XL

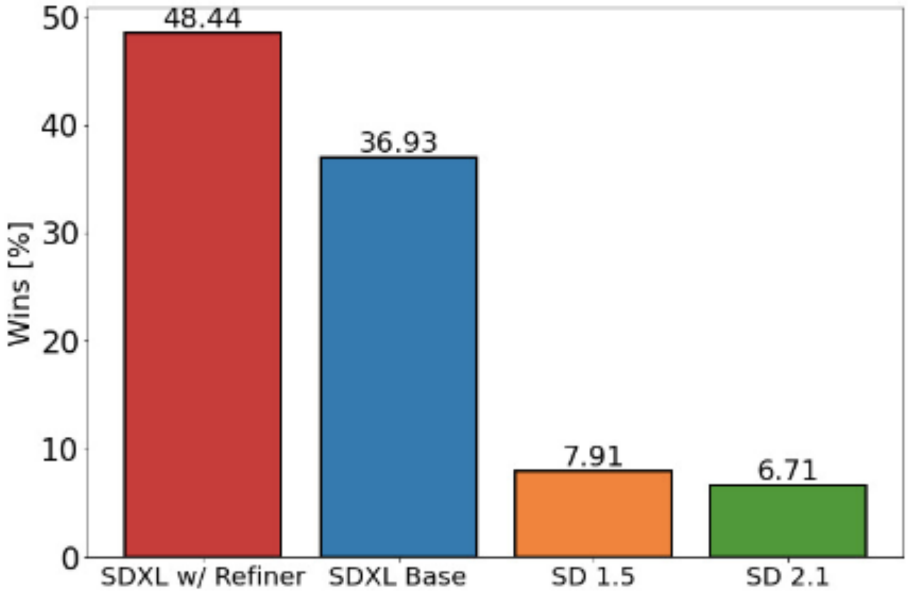
Two-stage-pipeline



Comparison between SD 1.5 and SDXL

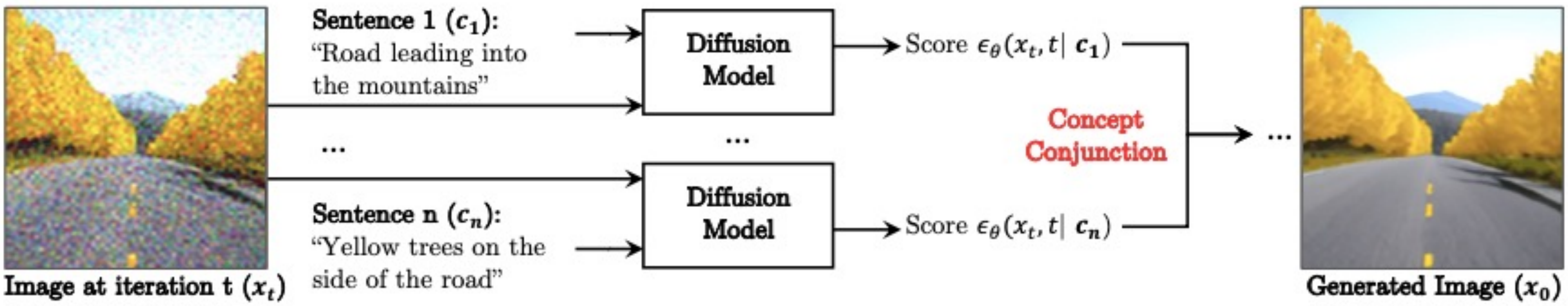
Model	*SDXL*	SD 1.4/1.5
# of UNet params	2.6B	860M
Transformer blocks	[0, 2, 10]	[1, 1, 1, 1]
Channel mult.	[1, 2, 4]	[1, 2, 4, 4]
Text encoder	CLIP ViT-L & OpenCLIP ViT-bigG	CLIP ViT-L
Context dim.	2048	768
Pooled text emb.	OpenCLIP ViT-bigG	N/A

Comparison of user preferences



Source: Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., ... & Rombach, R. (2023). Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

COMPOSABLE DIFFUSION



STRUCTURED DIFFUSION: CROSS-ATTENTION

Algorithm 1 StructureDiffusion Guidance.

Require:

Input: Prompt $\mathcal{P}$, Parser ξ, decoder ψ, trained diffusion model ϕ.

Output: Generated image x.

```
Retrieve concept set C = [c_1, ..., c_k] by traversing ξ(P);
W_p ← CLIP_text(P),  W_i ← CLIP_text(c_i);                          i = 1, ..., k
for t = T, T − 1, ..., 1 do
    for each cross attention layer in φ do
        Obtain previous layer's output X^t.
        Q^t ← f_Q(X^t),  K_p ← f_K(W_p),  V_i ← f_V(W̄_i);          i = p, 1, ..., k
        Obtain attention maps M^t from Q^t, K_p;                     {Eq. 1}
        Obtain O^t from M^t, {V_i}, and feed to following layers;    {Eq. 4}
    end for
end for
Feed z^0 to decoder ψ(·) to generate x.
```

Retrieve concept set $\mathcal{C} = [c_1, \ldots, c_k]$ by traversing $\xi(\mathcal{P})$;

$\mathcal{W}_\text{p} \leftarrow \text{CLIP}_\text{text}(\mathcal{P}),\ \ \mathcal{W}_i \leftarrow \text{CLIP}_\text{text}(c_i);$ $\quad i = 1, \ldots, k$

for $t = T, T-1, \ldots, 1$ **do**

  for each cross attention layer in ϕ **do**

    Obtain previous layer's output $\mathcal{X}^t$.

    $Q^t \leftarrow f_Q(\mathcal{X}^t),\ \ K_\text{p} \leftarrow f_K(\mathcal{W}_\text{p}),\ \ V_i \leftarrow f_V(\overline{\mathcal{W}}_i);$ $\quad i = \text{p}, 1, \ldots, k$

    Obtain attention maps M^t from Q^t, K_p; $\quad$ {Eq. $\boxed{1}$}

    Obtain O^t from $M^t, \{V_i\}$, and feed to following layers; $\quad$ {Eq. $\boxed{4}$}

  end for

end for

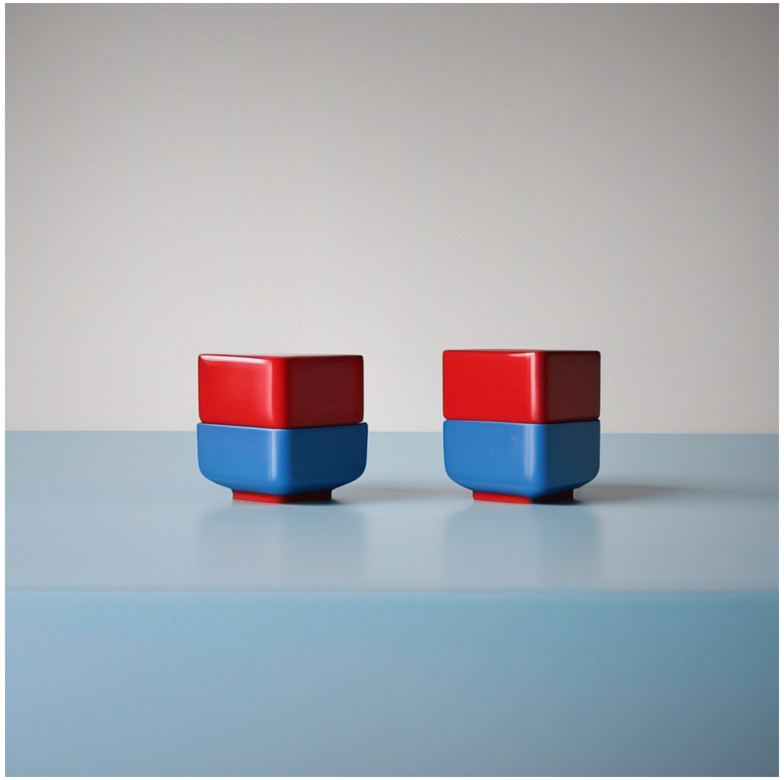
Feed z^0 to decoder $\psi(\cdot)$ to generate x.

STABLE DIFFUSION XL

Why not modifying projected pooled embedding?

→ No visual improvement

`Red cube in a blue bowl`



Not manipulated

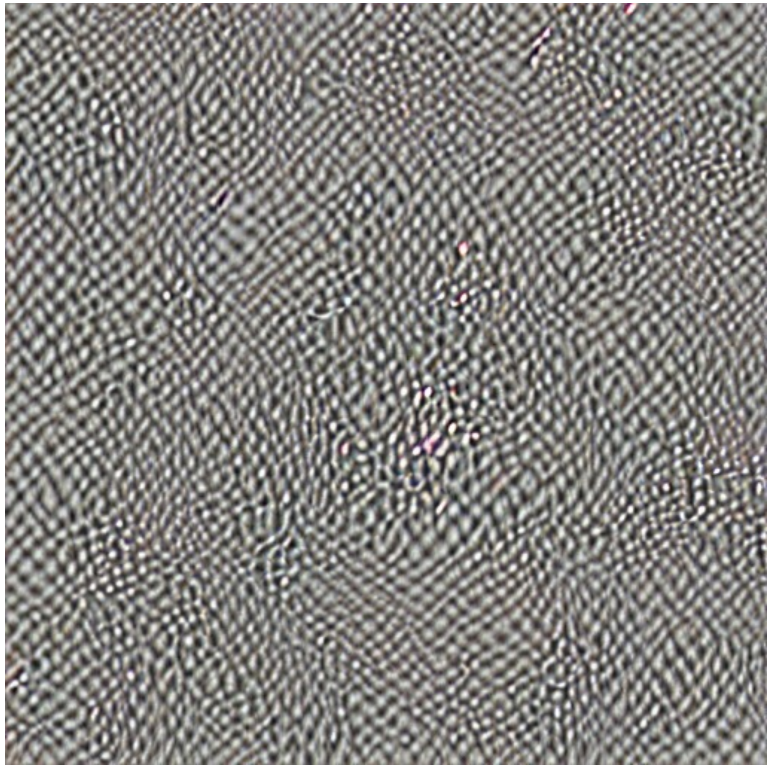
Manipulating both hidden states

Manipulating both hidden states + projected pooled layer

STABLE DIFFUSION

Why not manipulating starttoken?

- `Red cube in a blue bowl`
- Impact of the start token



Pooling all tokens

Pooling all tokens without start token