MANIPULATING EMBEDDINGS OF STABLE DIFFUSION PROMPTS

-

MASTER THESIS

JULIA PETERS

SUPERVISED BY NIKLAS DECKERS

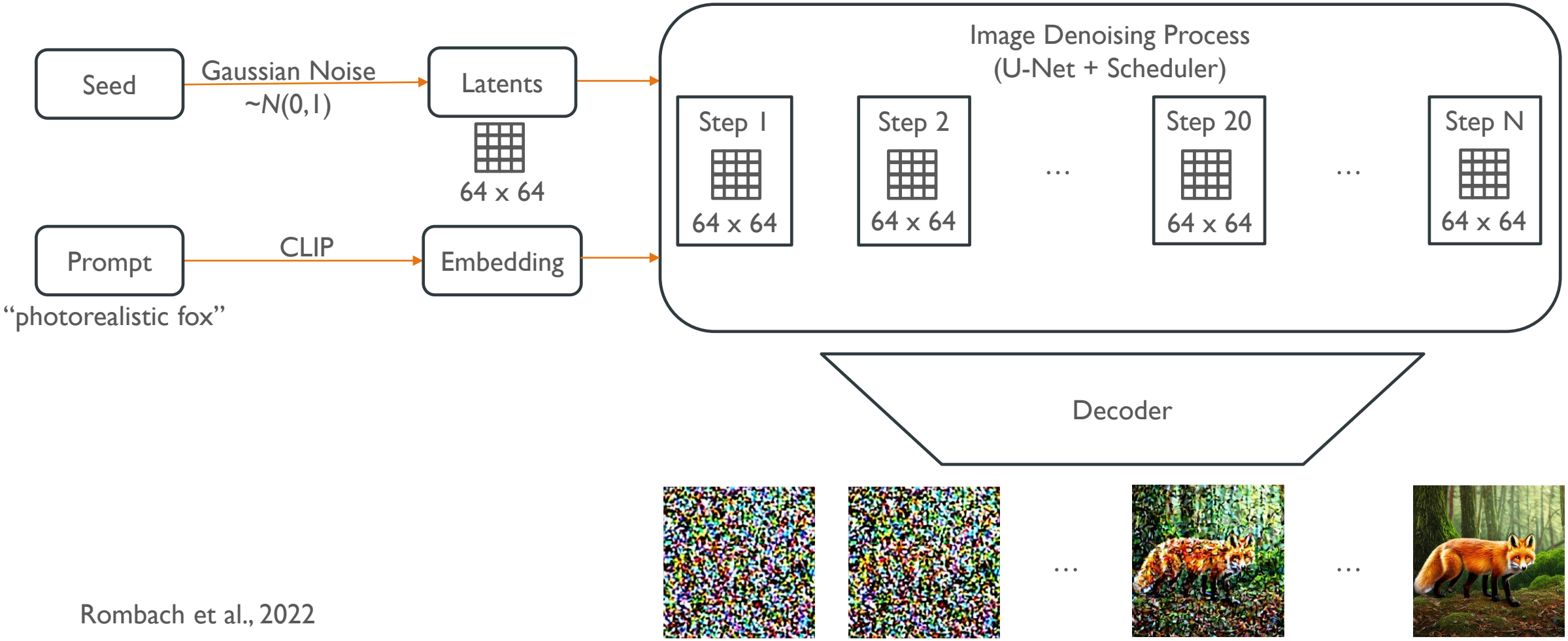
LEIPZIG, 11.05.23

OVERVIEW

- **Problem description**: Writing Stable Diffusion prompts
  - Image generation with Stable Diffusion
  - Problems with prompt engineering
- **Technical background**: Seed / embedding interpolation
- **Other approaches**: Guided image generation
  - ControlNet
  - Image generation based on brain activity
- **Approaches**: Guided image generation
  - Pipeline for image generation guidance
  - User interaction
  - Seed independent image generation
- **Conclusion / Future work**

PROBLEM DESCRIPTION: WRITING STABLE DIFFUSION PROMPTS

IMAGE GENERATION WITH STABLE DIFFUSION



Rombach et al., 2022

IMAGE GENERATION WITH STABLE DIFFUSION II

➔ Problem: Initially generated image does not entirely satisfy the expectations

➔ Approaches:

- Trying different seeds
- Prompt engineering (prompt refinement)
  - ➔ Manual trial and error

IMPLICATIONS OF PROMPT ENGINEERING

- Not a "pretty" image
- Low contrast in the background

- Not realistic enough
- Too much light

- Small prompt adjustments or using different seeds can lead to completely different images

➔ Further undesired details in regenerated images

➔ Potential outcome: Loss of interest after a few attempts

IMPLICATIONS OF PROMPT ENGINEERING II

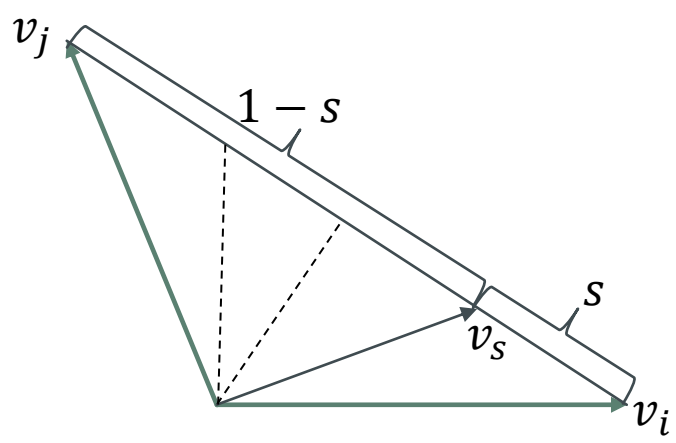
- Art of Prompt Engineering (Oppenlaender et al., 2022; Liu et al., 2022)
  - New communities to enable users to share best practices emerge
  - Definition of design guidelines for the production of better text to image outcomes

- Still lacking control to generate the desired image output

TECHNICAL BACKGROUND: SEED / CLIP EMBEDDING INTERPOLATION

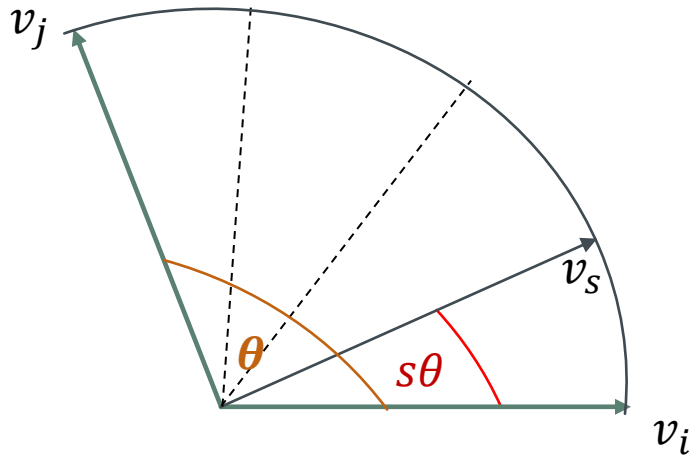
INTERPOLATION METHODS

LERP (Linear Interpolation)



- Interpolating along the line joining the tips of v_i and v_j

SLERP (Spherical Linear Interpolation)



- Rotation along the shortest arc on a unit sphere connecting two endpoints

Shoemake, 1985

CLIP EMBEDDING INTERPOLATION

- In CLIP embedding space a smooth linear interpolation (Tevet et al., 2022) and a spherical linear interpolation (Ramesh et al., 2022) can be performed

CLIP EMBEDDING INTERPOLATION

Seed: 987271

"A dream of an apple tree, stormy sky, high detail, concept art, matte painting"

"epic landscape with a lake, golden hour, misty ground, rocky ground, distant mountains, hazy, foggy, atmospheric perspective"

CLIP EMBEDDING INTERPOLATION

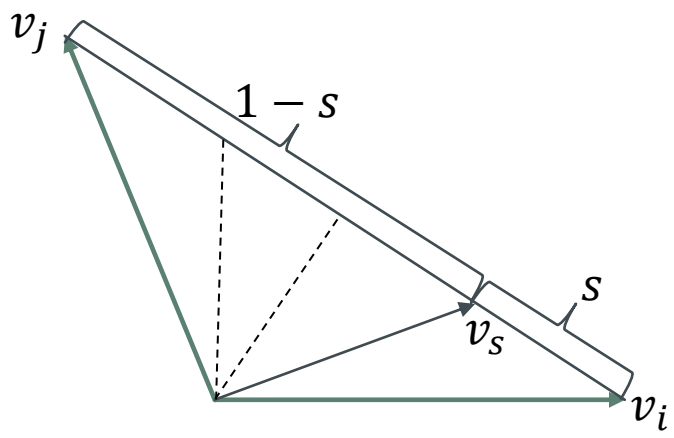
LERP

SLERP

➔ Prompt space continuous
➔ Application of gradient descent feasible

SEED INTERPOLATION

LERP (Linear Interpolation)



- **Problem**: Vector magnitude decreases in the midpoint
  - ➔ Deviating variance
  - ➔ Not appropriate for Gaussian distributed latent space
- **Result**: Blurry images close to midpoint
- **Solution**: LERP with adjusted variance

White, 2016

SEED INTERPOLATION

"a cybernetic samoyed and beagle, concept art, detailed face and body, detailed decor, fantasy, highly detailed, cinematic lighting, digital art painting, winter, nature, running"

Seed: 61582

Seed: 9168745

SEED INTERPOLATION

LERP with adjusted variance

SLERP

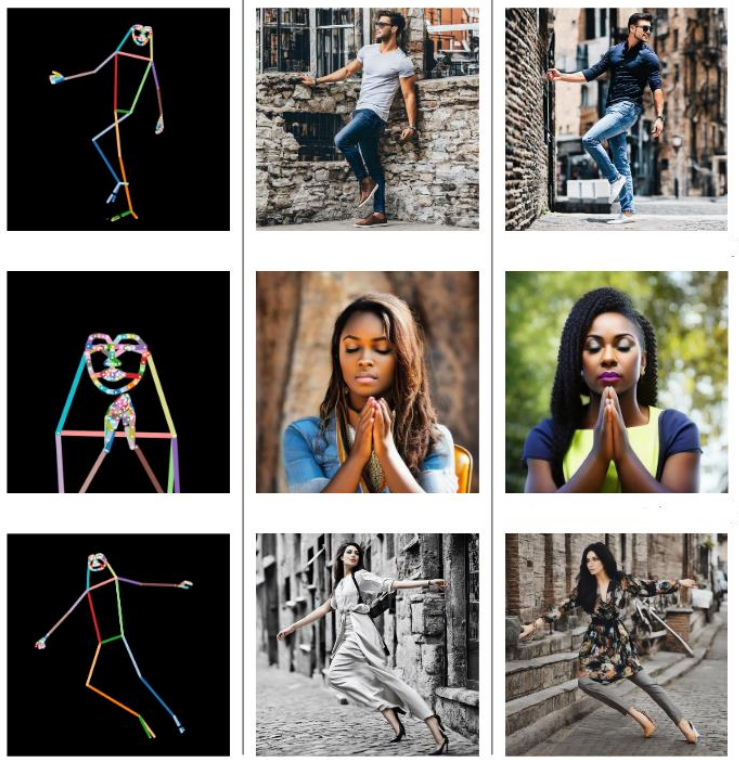
- Infinite number of images between two seeds and a predefined prompt
  - ➔ Infinite number of images for every prompt

OTHER APPROACHES: GUIDED IMAGE GENERATION

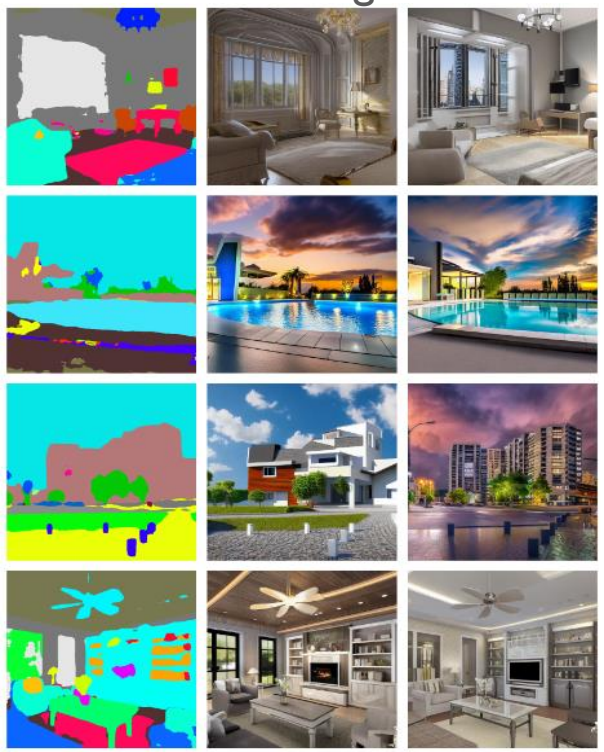
CONTROLNET

➔ Neural net structure controlling large diffusion models by supporting additional inputs (Zhang et al., 2023)

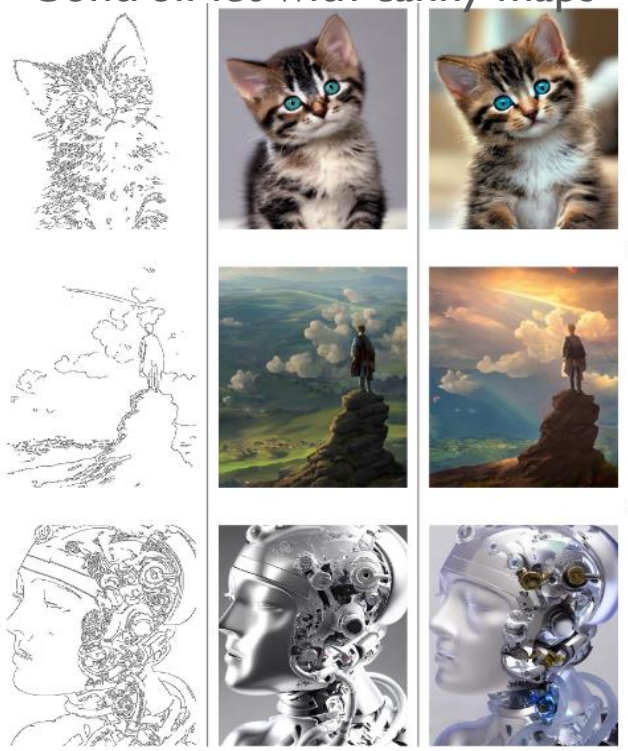
ControlNet with human pose



ControlNet with segmentation map



ControlNet with canny maps



(Images: Zhang et al., 2023)

➔ Many more additional inputs possible

IMAGE GENERATION BASED ON BRAIN ACTIVITY

(github.com/yu-takagi/StableDiffusionReconstruction)

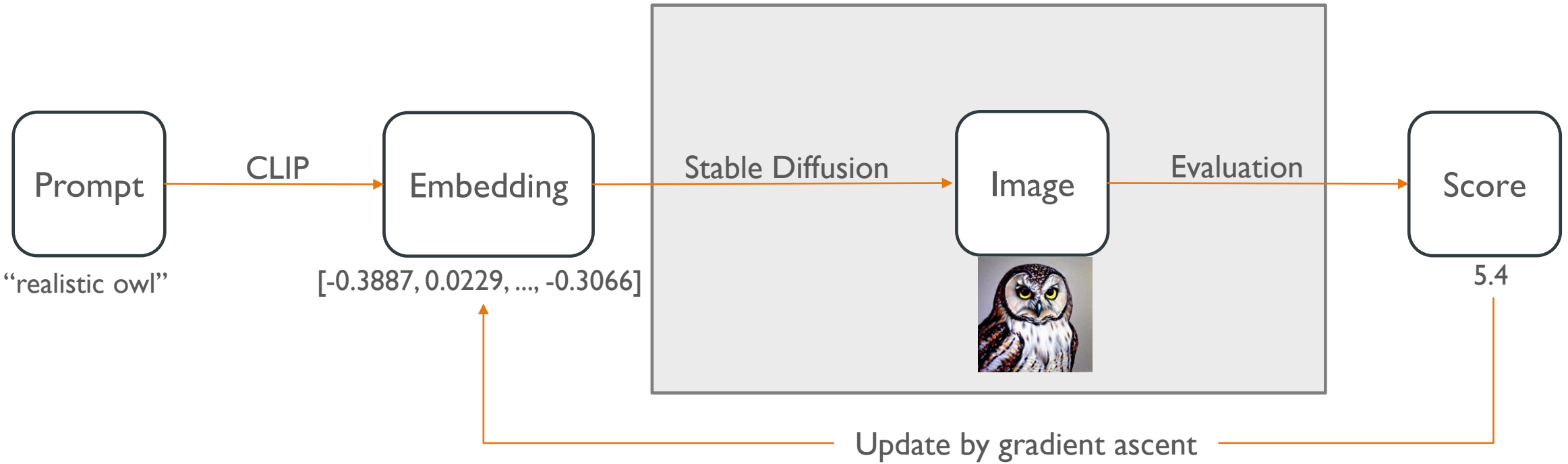
Reconstructed visual images from functional Magnetic Resonance Imaging (fMRI)

(Takagi et al., 2022)

APPROACHES: GUIDED IMAGE GENERATION

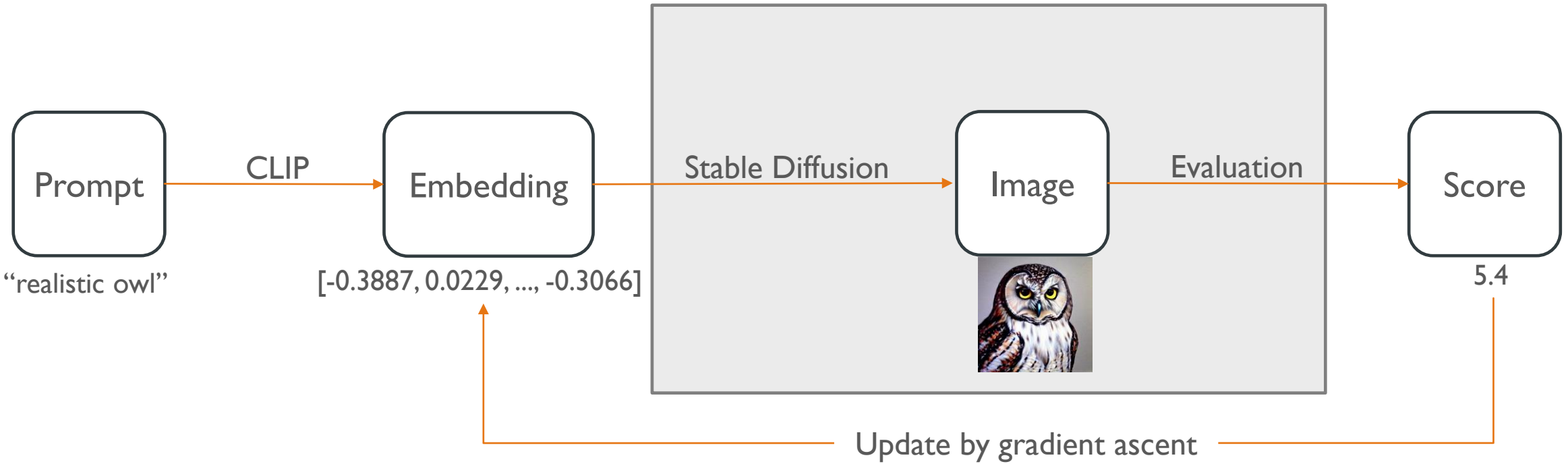
APPROACH I
METRIC BASED IMAGE GENERATION

1. PIPELINE FOR METRIC BASED IMAGE GENERATION



- Evaluation w.r.t. replaceable user defined metric
- Metric requirement: differentiability

1. PIPELINE FOR METRIC BASED IMAGE GENERATION

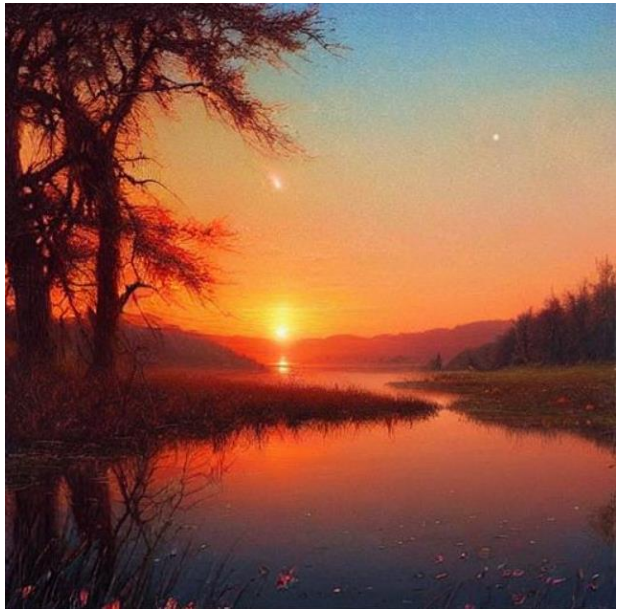


- **Advantage**: targeted prompt manipulation without prompt engineering

1. METRIC CHOICE FOR SCORE COMPUTATION

1. Simple Metric Ideas
   - Grayscale
   - Blurriness

Updated embedding, blurriness decreased (100 iterations)

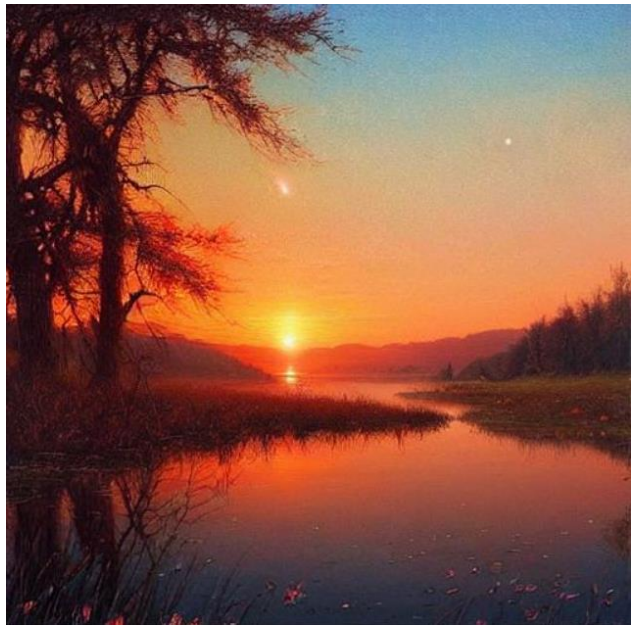


Original embedding

Updated embedding, blurriness increased (100 iterations)

1. METRIC CHOICE FOR SCORE COMPUTATION

2. LAION Aesthetic Predictor V2
   - MLP trained on 2.37B image - rating pairs ranging from 1 – 10



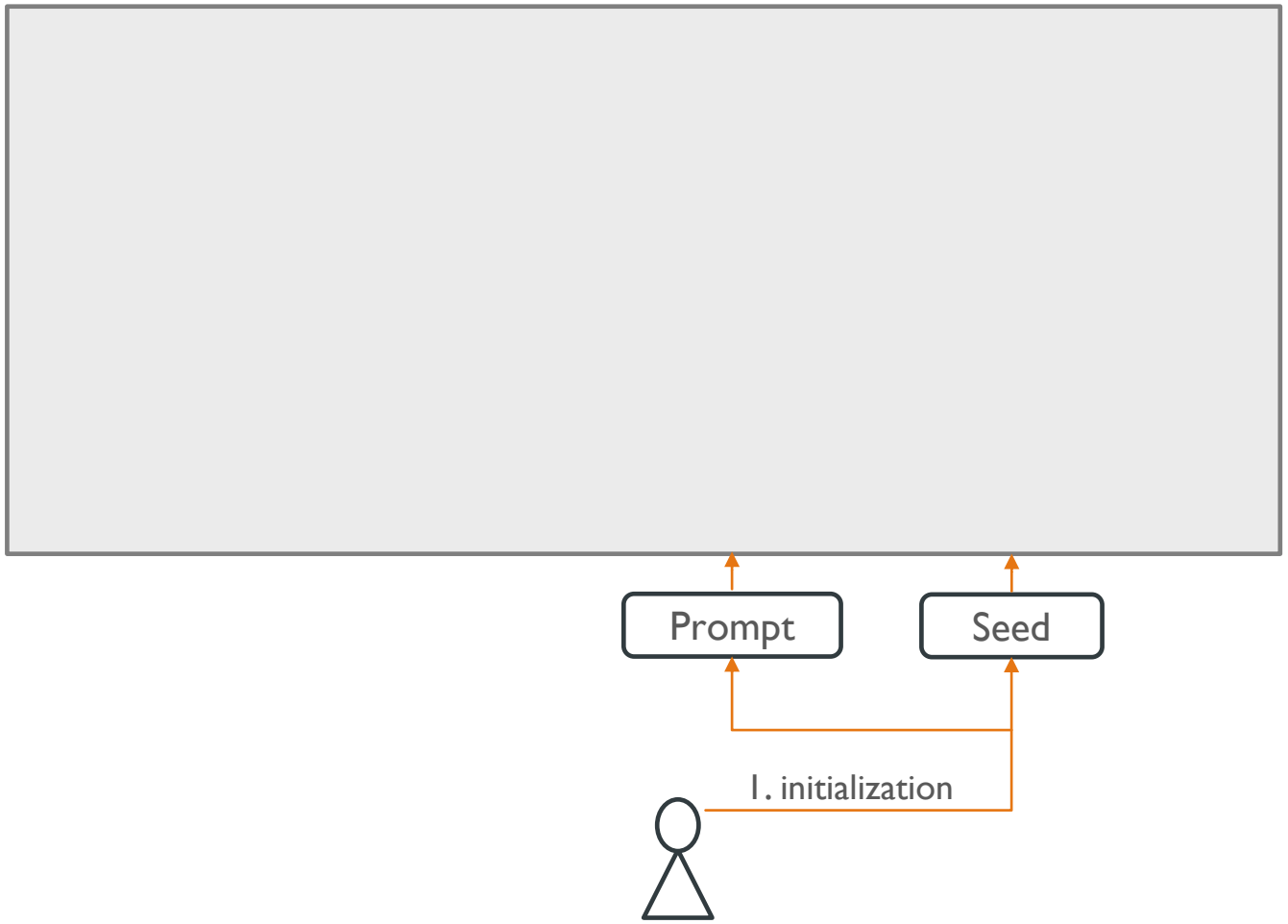
Score: 5.8

Embedding update
for
> 300 iterations

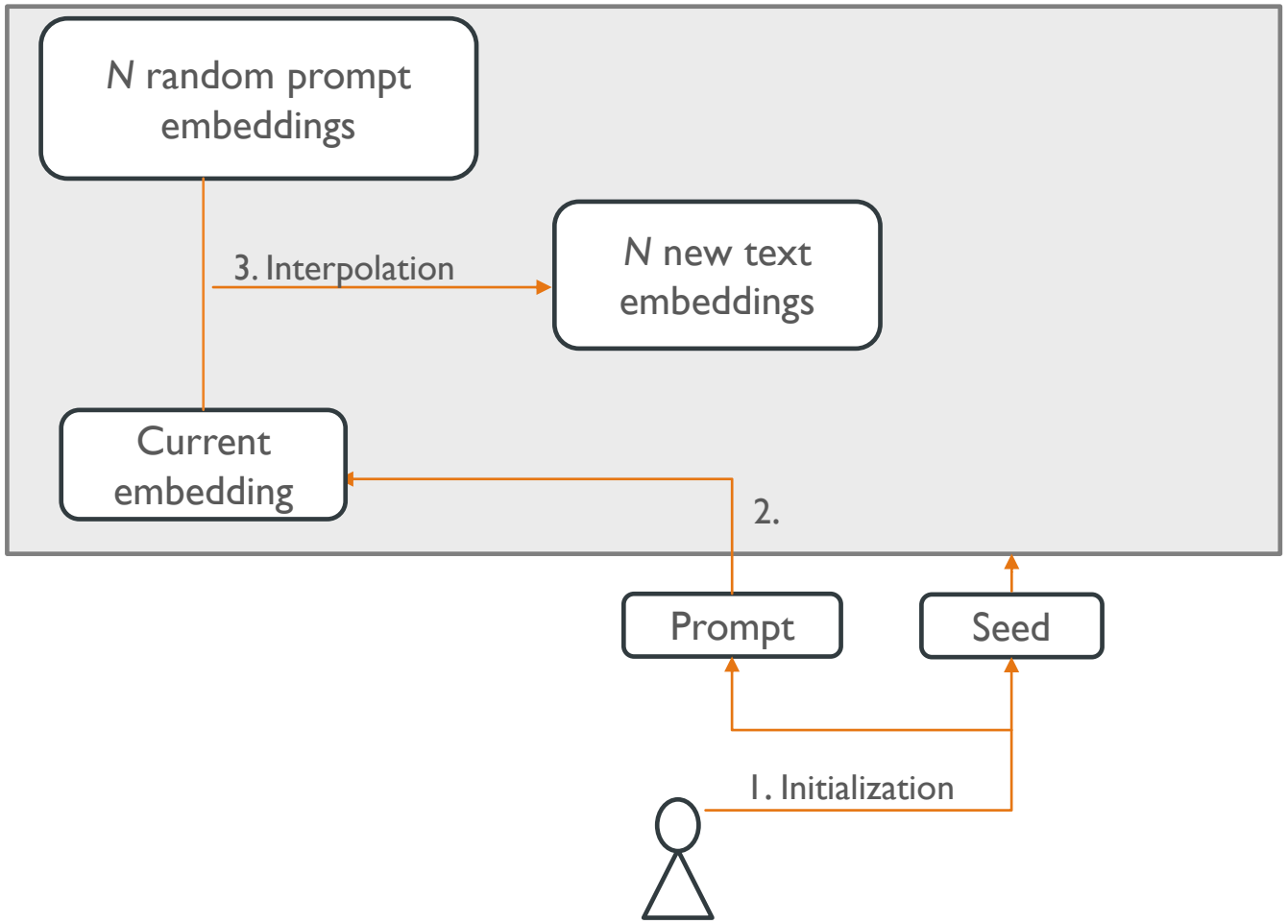
Score: 7.0

APPROACH II
USER INTERACTION

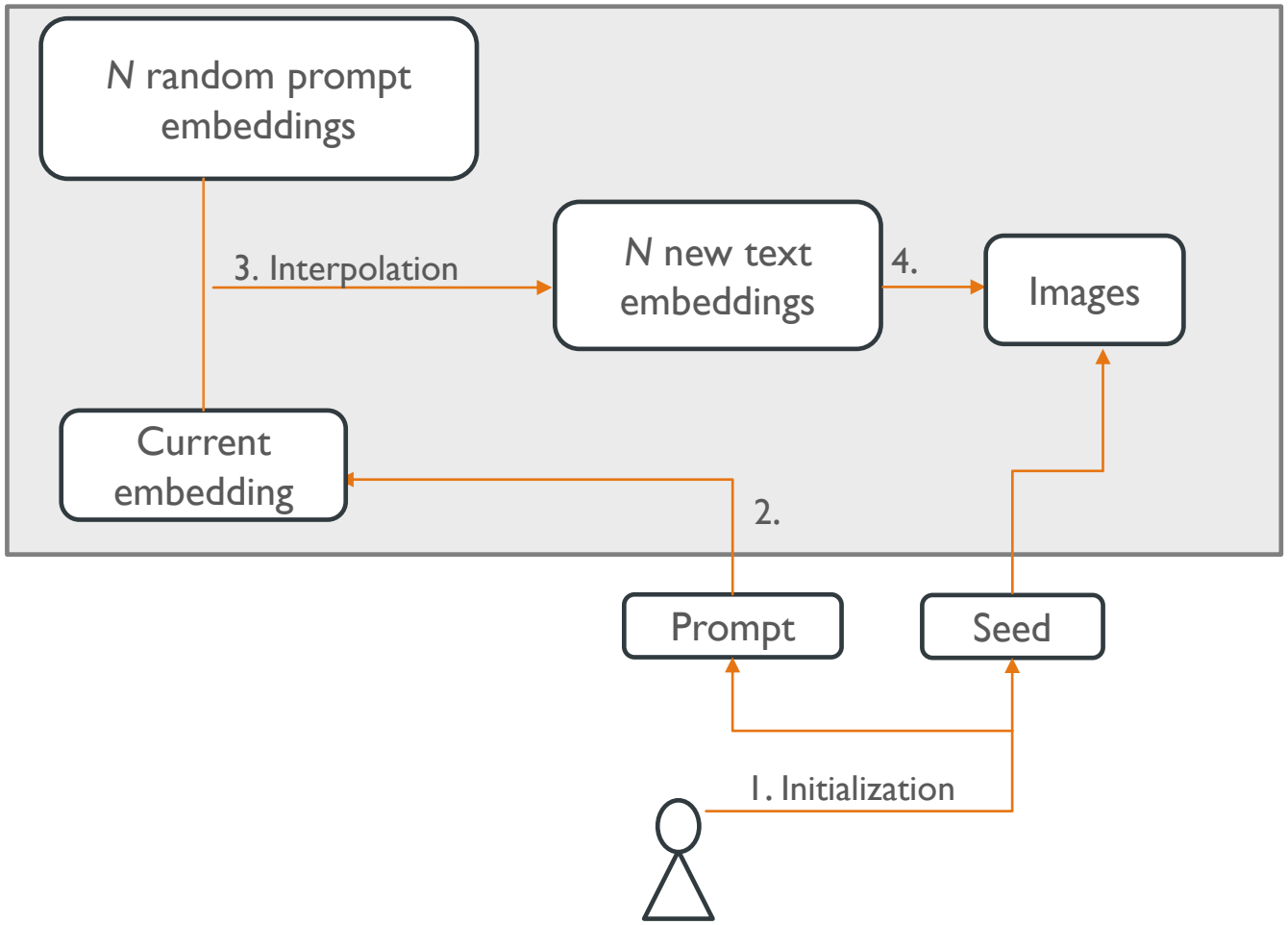
2. USER INTERACTION



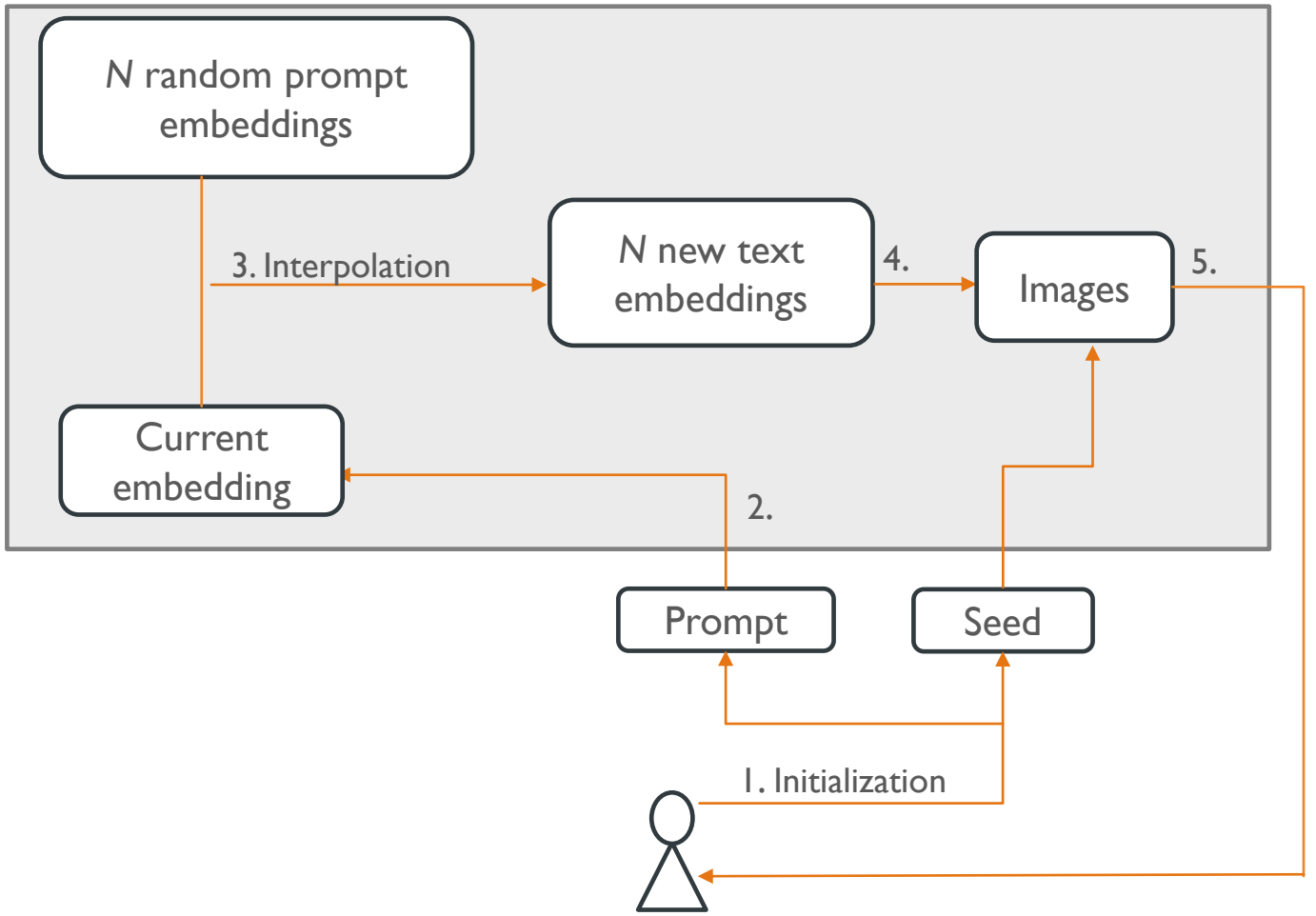
2. USER INTERACTION



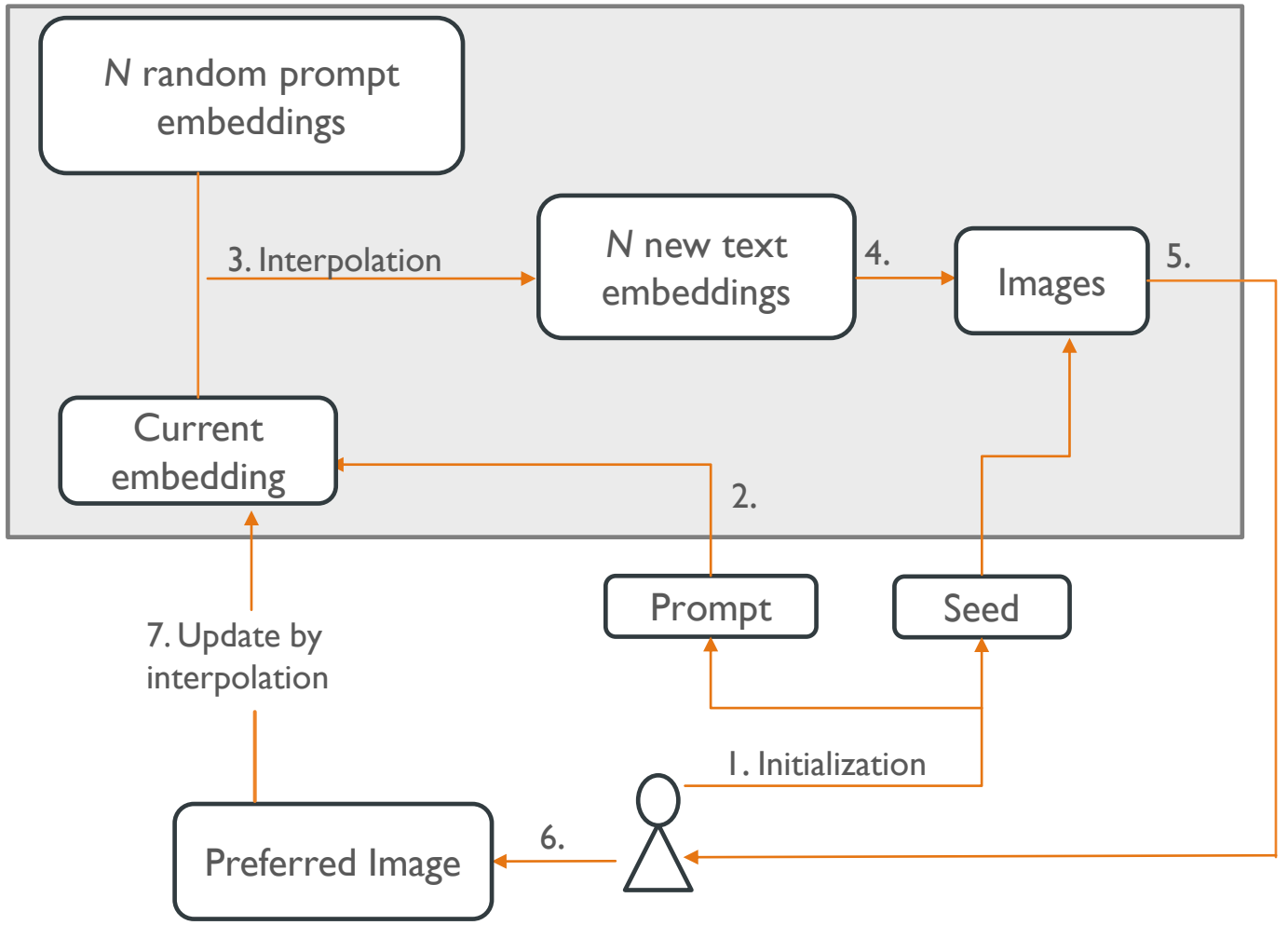
2. USER INTERACTION



2. USER INTERACTION

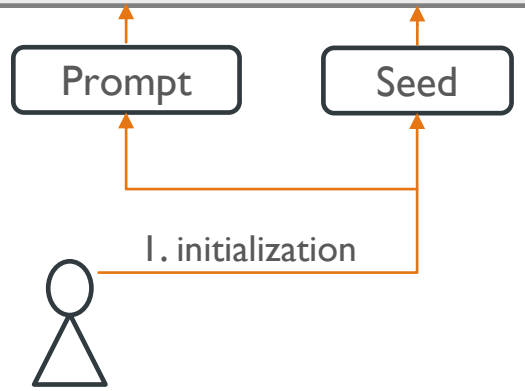


2. USER INTERACTION



2. USER INTERACTION EXAMPLE

- Initializing User Interaction Pipeline
  - Seed: 93769
  - Prompt: "flat design, astronaut flying"



2. USER INTERACTION EXAMPLE: PREFERENCE SELECTION

Current embedding
(user prompt)

Images based on new text embeddings for user selection:

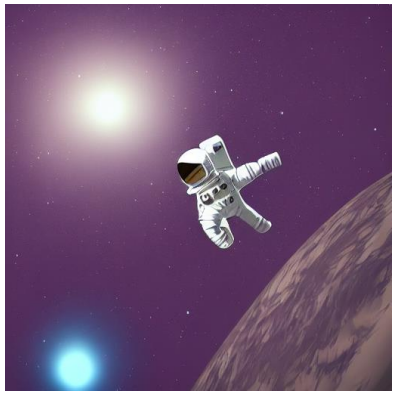


Image1

Image2

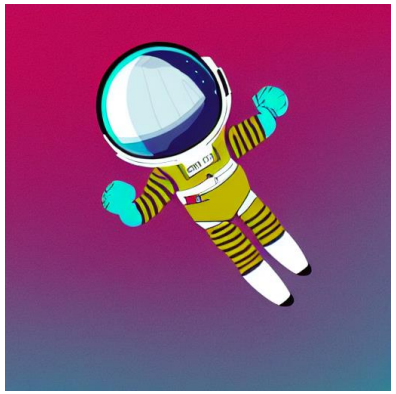


Image3

Image4

Image5

2. USER INTERACTION EXAMPLE: UPDATE USER PROMPT EMBEDDING

- User selection:
  - Image3
  - Interpolation value: 0.65

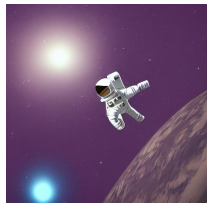


Image1

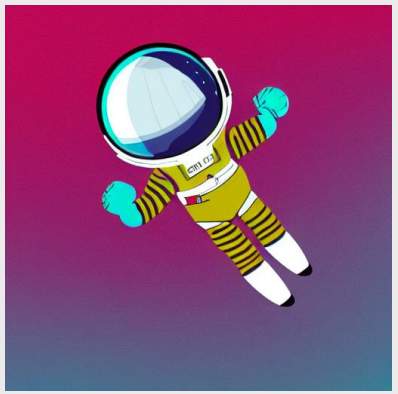

Image2


Image3


Image4


Image5

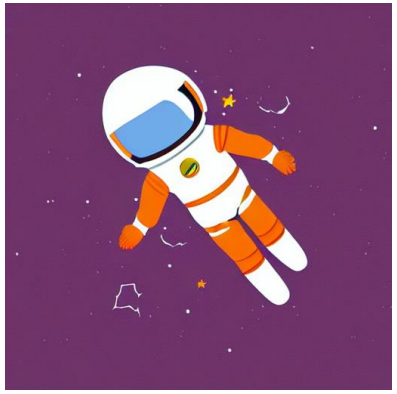

Current embedding (0.35)


Embedding based on Image3 (0.65)

Interpolation


Current embedding updated

2. USER INTERACTION EXAMPLE: PREFERENCE SELECTION



Current embedding
(user prompt updated)

Images based on new text embeddings for user selection:

Image1

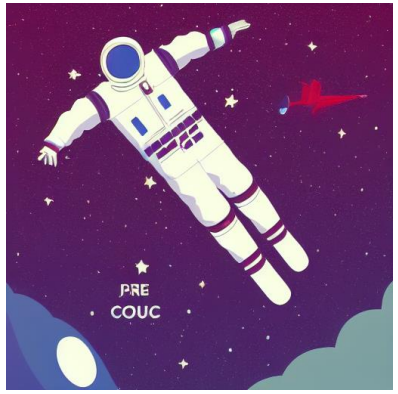


Image2

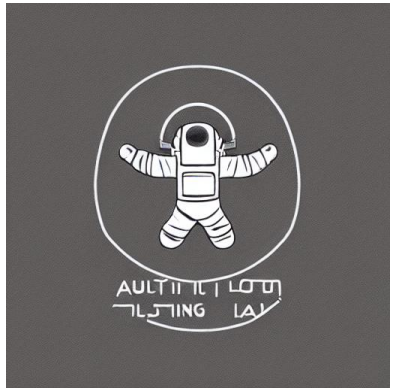


Image3

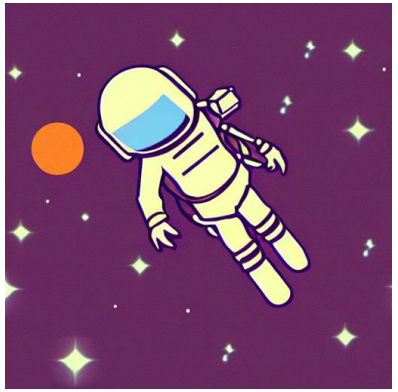


Image4

Image5

2. USER INTERACTION EXAMPLE: UPDATE USER PROMPT EMBEDDING

- User selection:
  - Image1
  - Interpolation value: 0.4


Image1

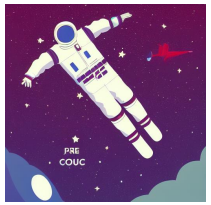


Image2

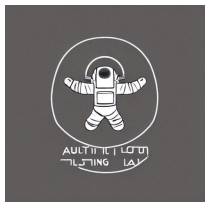


Image3

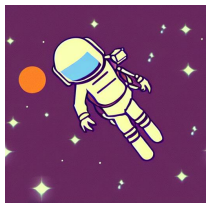


Image4

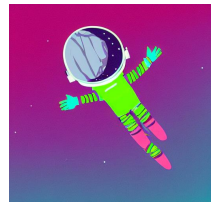
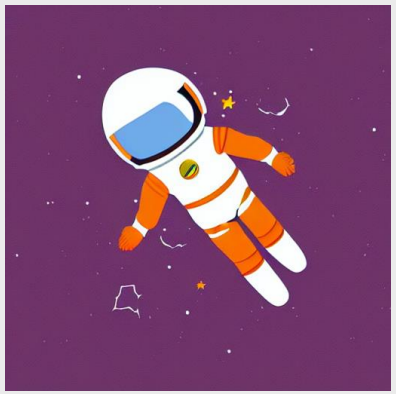
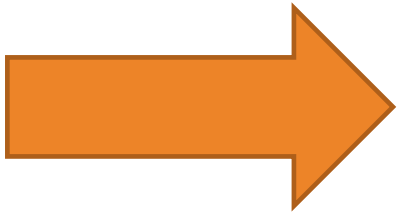


Image5


Current embedding (0.6)


Embedding based on Image1 (0.4)


Interpolation


Current embedding updated

2. USER INTERACTION EXAMPLE: RESULT

Image after ~20 iterations

➔ Modified prompt embedding without prompt engineering

APPROACH III
SEED INDEPENDENT IMAGE GENERATION

3. WIP: SEED INDEPENDENT IMAGE GENERATION

Fixed prompt: "hummingbird mascot with adorable eyes, friendly, waving to the camera"

First attempt with initial seed

Seed trial and error

Specific seed providing the preferred image

Problem: Good prompt works for specific seeds

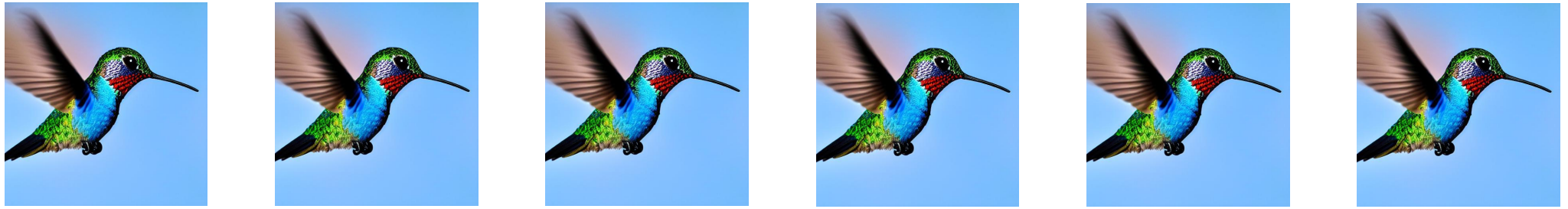
Goal: Avoiding the repeated seed adaption by obtaining the desired image independent from the seed
➔ Optimized text embedding (prompt) required providing the same image for every seed

3. WIP: SEED INDEPENDENT IMAGE GENERATION (SIMULATION)

initial prompt

optimized prompt

gradually adjusted seed



fixed random seed

➔ **Idea**:

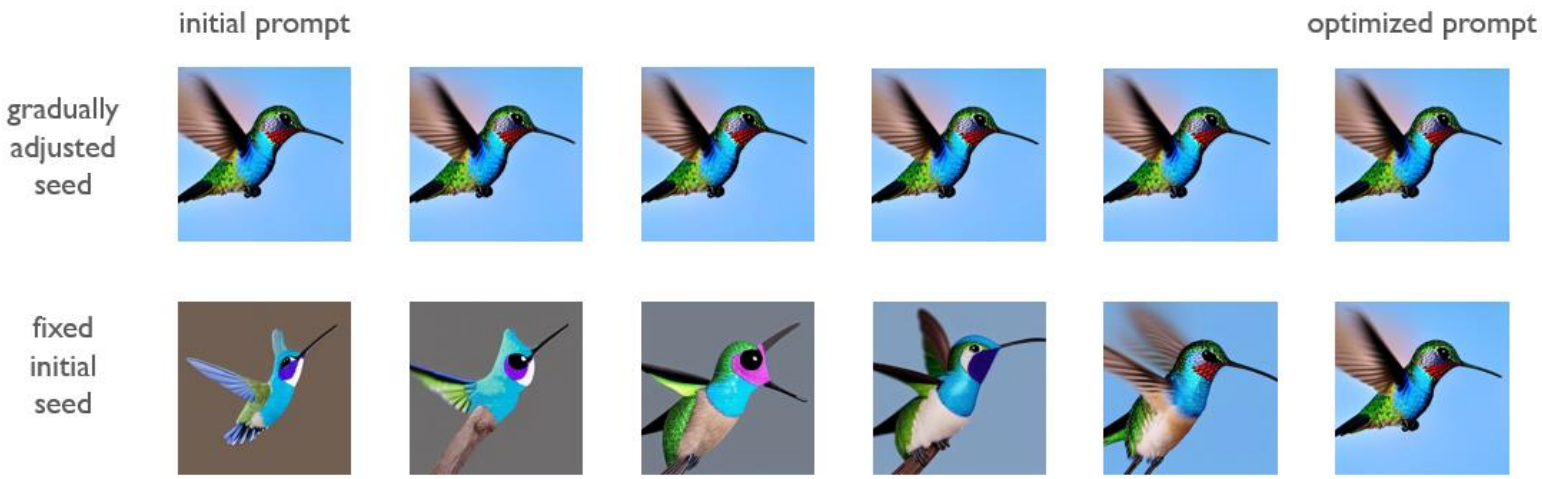
- Starting with the seed enabling the target image
- Gradually increase the distance between used seed and specific seed enabling the target image
- Maintaining the image similarity by updating the prompt embedding

3. WIP: SEED INDEPENDENT IMAGE GENERATION

➔ **Approach**:

1. In alternating steps, update:
   - Seed latent such that similarity of image and target image decreases (gradient ascent)
   - Prompt embedding such that similarity of image and target image increases (gradient descent)
2. Return optimized text embedding

➔ Same image for different seeds



3. WIP: SEED INDEPENDENT IMAGE GENERATION

➔ **Outlook**:

- Universal prompt embedding resulting in the same image for different seeds
  - ➔ Obtaining of a very precise textual description
  - ➔ Specific changes in text, like colour of an object
- Selecting preferred parts of the image to be fixed and regenerating the surrounding area

CONCLUSION

CONCLUSION

- Lack of control to create satisfactory results with text-to-image models
  - → need for systems to further enable the user to produce desired results beyond prompt engineering
- More freedom for the user by adapting flexible image generation procedures to guide the process

→ 3 Approaches

- Metric based prompt adjustment
- Prompt adjustment by user ratings
- Universal seed independent prompts

- Limitation: extensive runtime and computation due to high number of iterations of gradient ascent

FUTURE WORK:

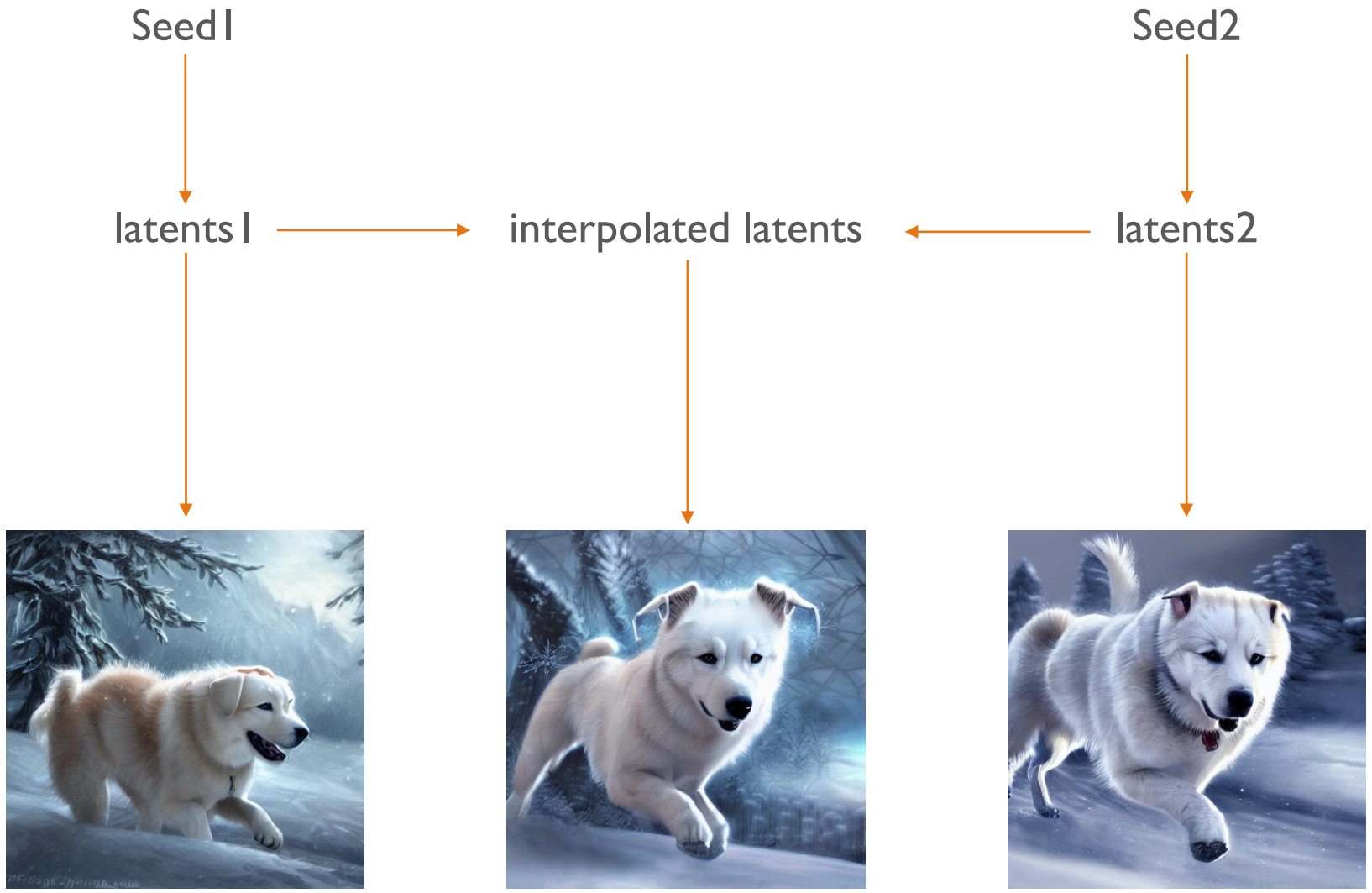
- More user experiments for improving the realization of the user intent in image generation
- Application of the experiments to other modalities:
  - Video
  - Audio

REFERENCES

- Rombach, Robin and Blattmann, Andreas and Lorenz, Dominik and Esser, Patrick and Ommer, Björn. "High-resolution image synthesis with latent diffusion models" Proceedings of the IEEE/CVF Conference on Computer Vision and Patter, pp. 10684-10695 Recognition, 2022
- Liu, Vivian and Chilton, Lydia B. "Design Guidelines for Prompt Engineering Text-to-Image Generative Models" In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 2022
- Oppenlaender, Jonas. "Prompt Engineering for Text-Based Generative Art", arXiv preprint arXiv:2204.13988, 2022
- White, Tom. "Sampling Generative Networks" arXiv preprint arXiv :1609.04468, 2016
- Zhang, Lvmin and Agrawala, Maneesh. "Adding conditional control to text-to-image diffusion models", arXiv preprint arXiv:2302.05543, 2023
- Takagi, Yu and Nishimoto, Shinji. "High-resolution image reconstruction with latent diffusion models from human brain activity", bioRxiv, 2022
- Tevet, Guy and Gordon, Brian and Hertz, Amir and Bermano, Amit H and Cohen-Or, Daniel. "Motionclip: Exposing human motion generation to clip space" Computer Vision--ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23--27, 2022, Proceedings, Part XXII, pp. 358-374, 2022
- Ramesh, Aditya and Dhariwal, Prafulla and Nichol, Alex and Chu, Casey and Chen, Mark. "Hierarchical text-conditional image generation with clip latents", arXiv preprint arXiv:2204.06125, 2022
- Shoemake, Ken. "Animating rotation with quaternion curves". In ACM Siggraph, pp245–254, 1985.

BACKUP SLIDES

SEED INTERPOLATION



LERP: VARIANCE AND VECTOR LENGTH FOR $Z \sim N(0,1)$

$$s_z = \sqrt{Var(Z)} = \sqrt{Var(sX + (s-1)Y)}$$

$$= \sqrt{s^2 Var(X) + (s-1)^2 Var(Y) + 2s(s-1)Cov(X,Y)}$$

$$= \sqrt{s^2 \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} + (s-1)^2 \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n} + 2s(s-1)\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 (y_i - \bar{y})^2}{n}}$$

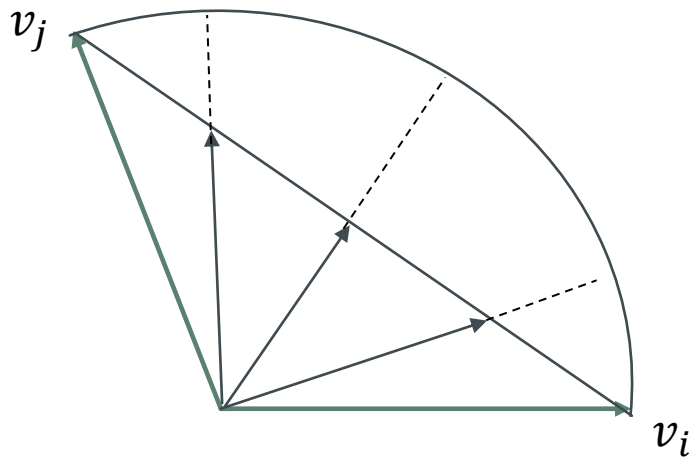
$$\overset{X,Y \sim N(0,1) \rightarrow \bar{x}=\bar{y}=0}{=} \sqrt{s^2 \frac{\sum_{i=1}^{n} x_i^2}{n} + (s-1)^2 \frac{\sum_{i=1}^{n} y_i^2}{n} + 2s(s-1)\frac{\sum_{i=1}^{n} x_i^2 y_i^2}{n}}$$

$$= \sqrt{\frac{\sum_{i=1}^{n}(sx_i + (s-1)y_i)^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} {z_i}^2}{n}} = \frac{|\vec{z}|}{\sqrt{n}}$$

➔ smaller vector length $|\vec{z}|$ ⇔ less variance

VARIANCE ADJUSTED LERP VS SLERP

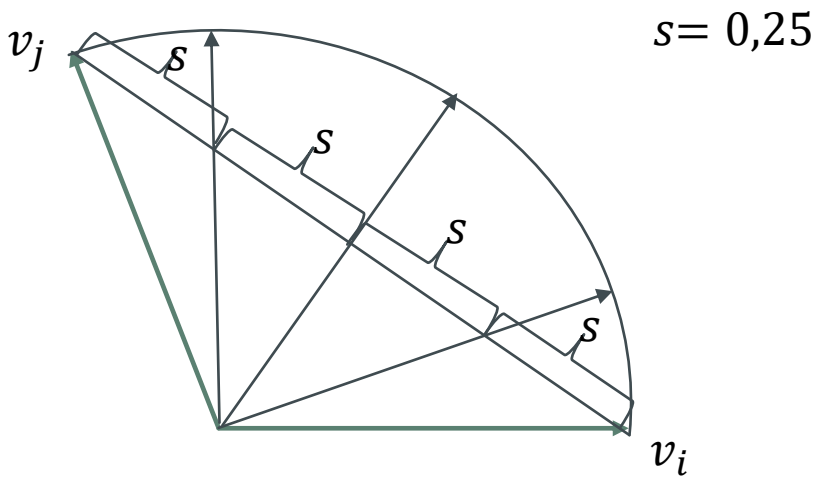
Variance adjusted LERP (Linear Interpolation)



- Interpolating along the line joining the tips of v_i and v_j
- Increasing length of interpolated vectors by aligning variances

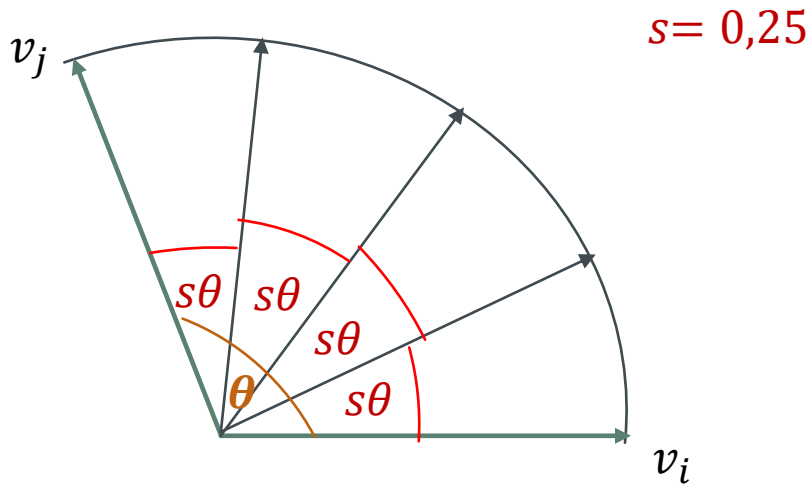
VARIANCE ADJUSTED LERP VS SLERP

Variance adjusted LERP (Linear Interpolation)



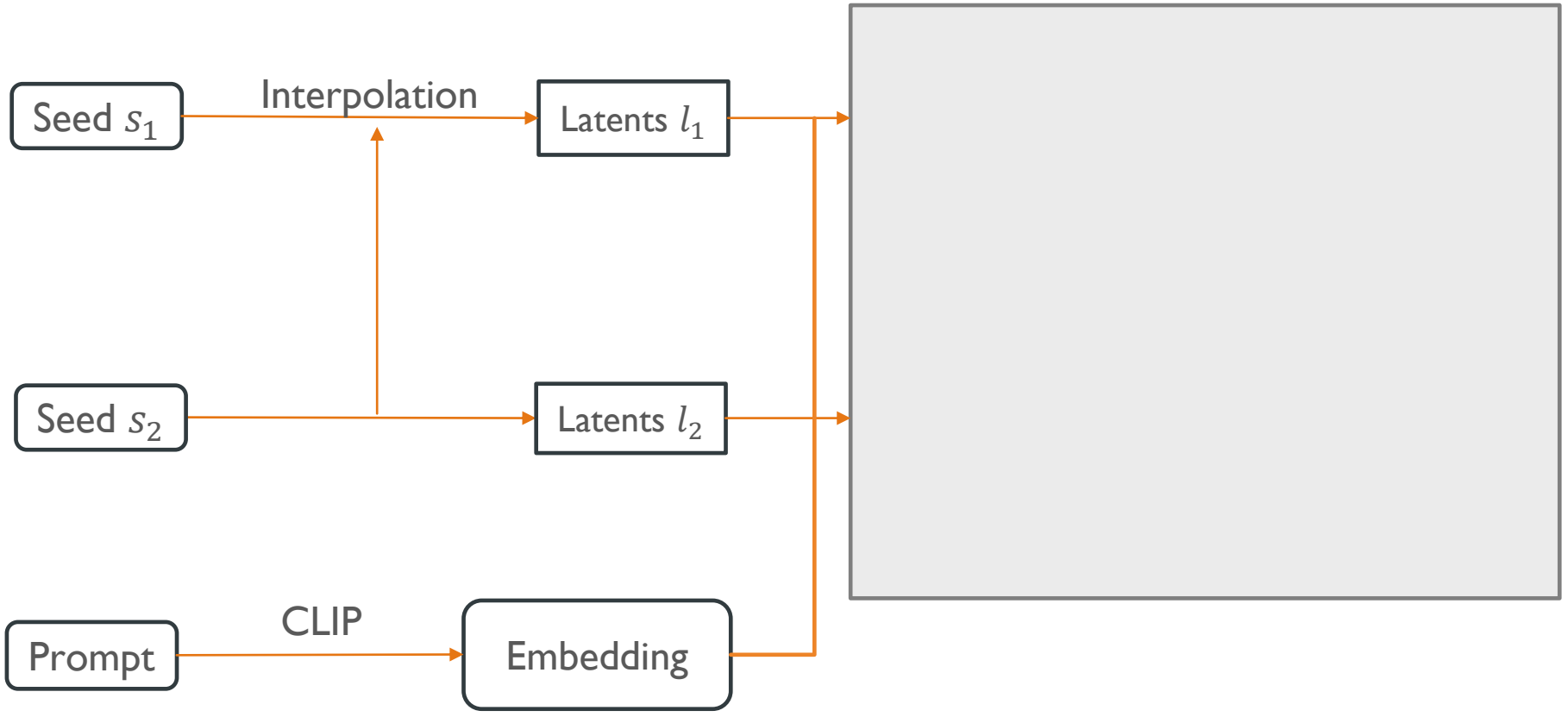
- Vector is moving faster, when interpolating closer to the midpoint (larger distance must be covered)

SLERP (Spherical Linear Interpolation)

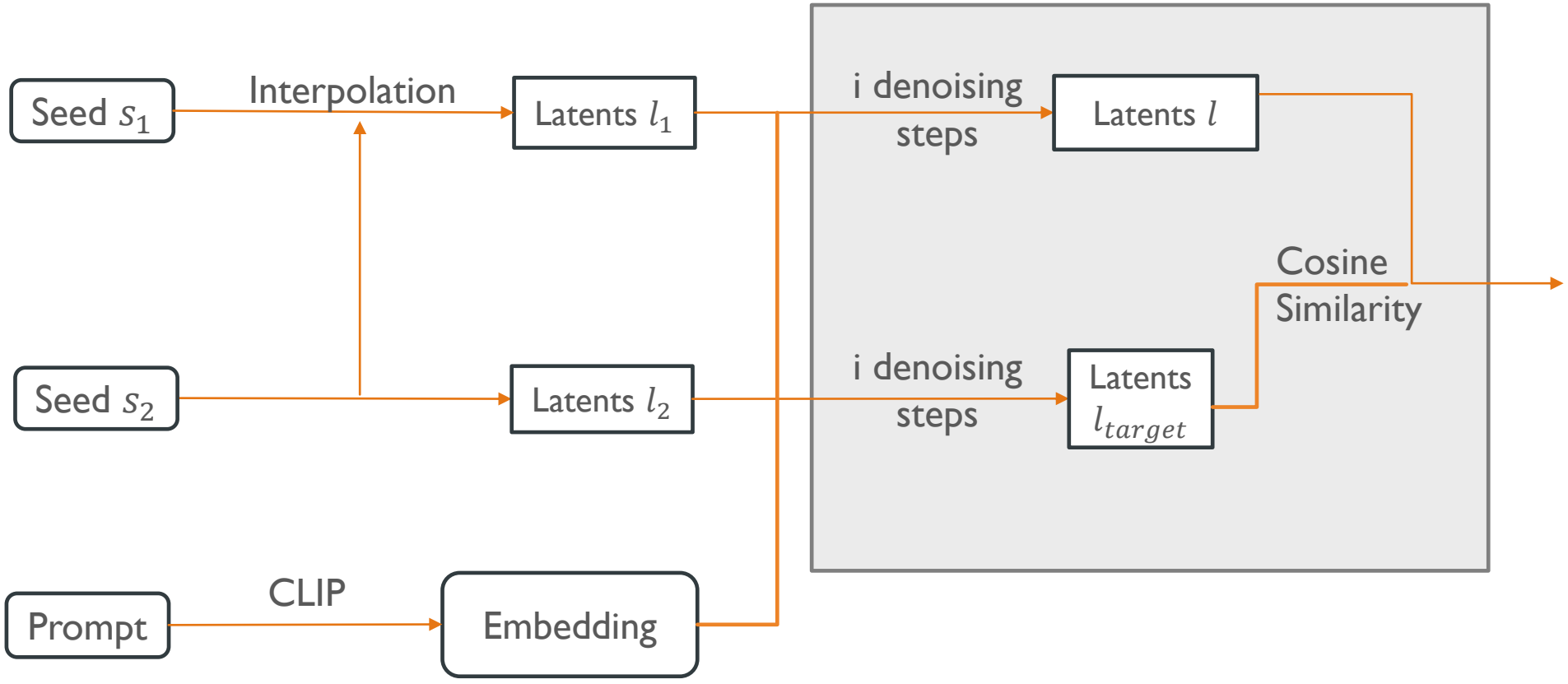


- Moving at constant velocity (→ potentially smoother interpolation)

3. WIP: SEED INDEPENDENT IMAGE GENERATION



3. WIP: SEED INDEPENDENT IMAGE GENERATION



3. WIP: SEED INDEPENDENT IMAGE GENERATION