Optimizing Stable Diffusion Prompts in Text Space

Bachelor's Seminar

Moritz Böhme
Supervised by Niklas Deckers

Institute of Computer Science

27. June 2024

Context: Text to Image Generation

- Users generate an image from a prompt using latent diffusion

Context: Text to Image Generation

- Users generate an image from a prompt using latent diffusion
- Example: "`realistic spaceship rocket design.`"

Context: Text to Image Generation

- Users generate an image from a prompt using latent diffusion
- Example: “`realistic spaceship rocket design.`”

Most Common Scenario: Improving Aesthetics

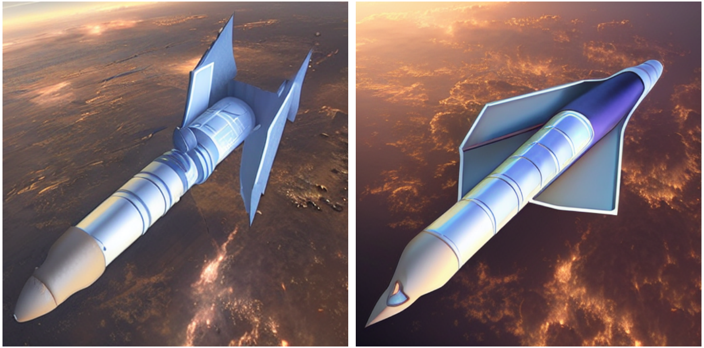
Problem: Descriptive Prompt $\nrightarrow$ Good Aesthetics

- Example: "`realistic spaceship rocket design.`" produces a matching, but unappealing image
- Prompt language distinct from user's language $\rightarrow$ Iterative trial and error (prompt engineering)

Common Solution: Prompt Modifiers

- Add suffixes (prompt modifiers): "`hd`", "`high quality`", etc.
  $\rightarrow$ "`realistic spaceship rocket design.` **`high quality`**"

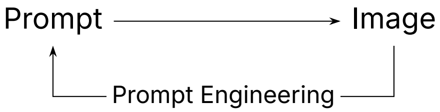
Common Solution: Prompt Modifiers



- Add suffixes (prompt modifiers): "`hd`", "`high quality`", etc.
  → "`realistic spaceship rocket design.` **`high quality`**"
- Result still not ideal

Common Solution: Prompt Modifiers

- Iterate with more or other suffixes



Manual Prompt Engineering: Flaws

- Highly arbitrary

Manual Prompt Engineering: Flaws

- Highly arbitrary
- Does not generalize

Manual Prompt Engineering: Flaws

- Highly arbitrary
- Does not generalize
- Inaccessible to non-experts

Manual Prompt Engineering: Potential Solutions

- User study to find good suffixes

Manual Prompt Engineering: Potential Solutions

- User study to find good suffixes
  - Highly user dependent

Manual Prompt Engineering: Potential Solutions

- User study to find good suffixes
  - Highly user dependent
  - Does not generalize over prompts

Manual Prompt Engineering: Potential Solutions

- User study to find good suffixes
  - Highly user dependent
  - Does not generalize over prompts

- Use classifier pretrained on user preferences to improve generated images as in Deckers et al. [1]

Modifying Prompt Embeddings [1]

`“realistic spaceship rocket design.”`
Before (left) and after (right) optimization.
Reproduced from Deckers et al. [1]

Proposed Solution

- Problem with method of Deckers et al. [1]: Does not yield an improved prompt

Proposed Solution

- Problem with method of Deckers et al. [1]: Does not yield an improved prompt
- Our proposed solution yields a prompt

Proposed Solution

- Problem with method of Deckers et al. [1]: Does not yield an improved prompt
- Our proposed solution yields a prompt
  $\rightarrow$ Users can interpret and edit prompt

Proposed Solution

- Problem with method of Deckers et al. [1]: Does not yield an improved prompt
- Our proposed solution yields a prompt
  $\rightarrow$ Users can interpret and edit prompt
  $\rightarrow$ Allows reuse for different prompts

Proposed Solution

“realistic spaceship rocket design.”

Before (left) and after (right) optimization

Proposed Solution

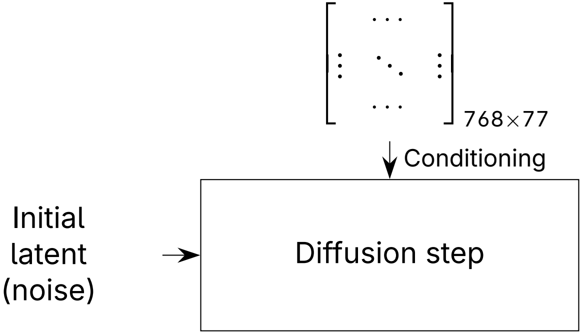
“realistic spaceship rocket design.
sts crispy affirting fanny dechomo earn ”
Before (left) and after (right) optimization

Latent Diffusion

$$\begin{bmatrix} & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & \end{bmatrix}_{768 \times 77}$$

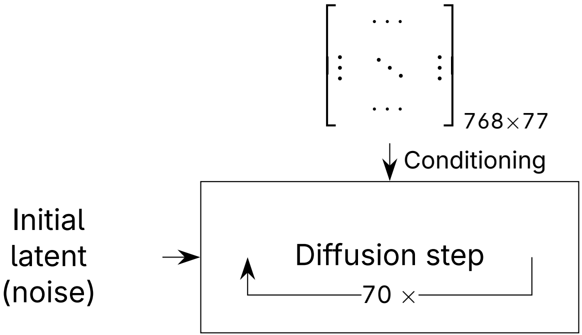
- CLIP converts prompt to embedding

Latent Diffusion



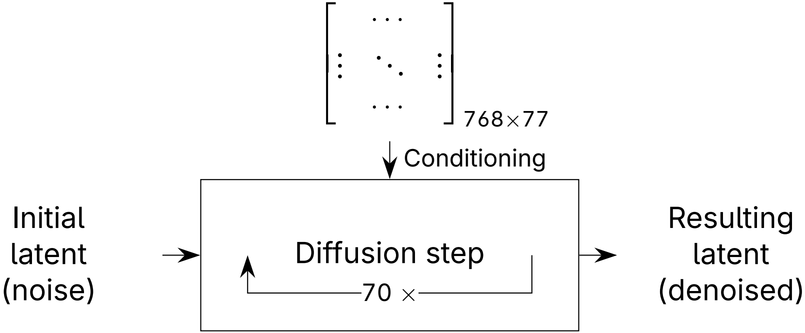
- CLIP converts prompt to embedding
- Diffusion model generates latent representation of an image using the prompt embedding as condition

Latent Diffusion



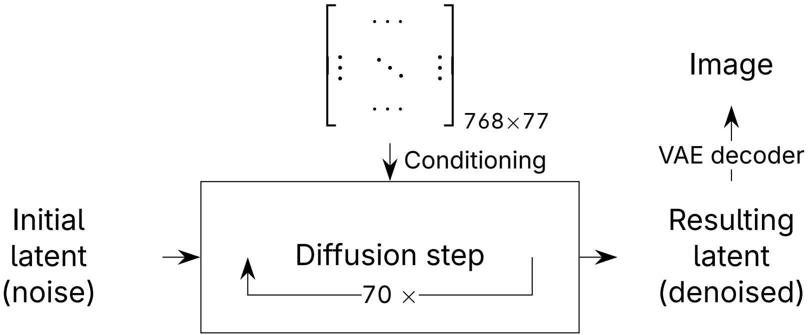
- CLIP converts prompt to embedding
- Diffusion model generates latent representation of an image using the prompt embedding as condition

Latent Diffusion



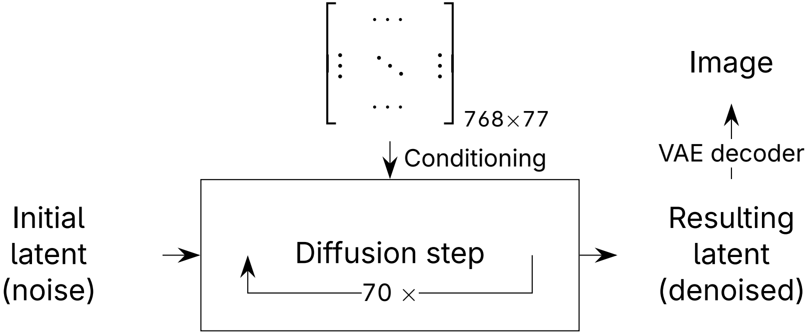
- CLIP converts prompt to embedding
- Diffusion model generates latent representation of an image using the prompt embedding as condition

Latent Diffusion



- CLIP converts prompt to embedding
- Diffusion model generates latent representation of an image using the prompt embedding as condition

Latent Diffusion



- CLIP converts prompt to embedding
- Diffusion model generates latent representation of an image using the prompt embedding as condition
- Diffusion was trained to have resulting image match the description

Generating Prompt Embedding Using CLIP

“prompt”

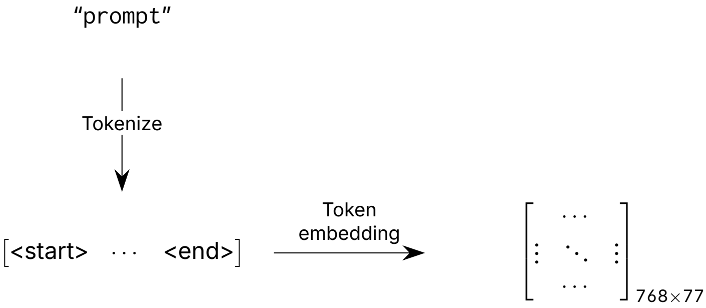
Generating Prompt Embedding Using CLIP

"prompt"

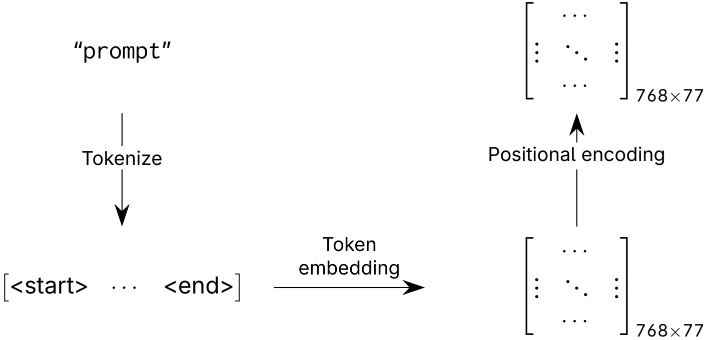
$\downarrow$ Tokenize

[<start> ··· <end>]

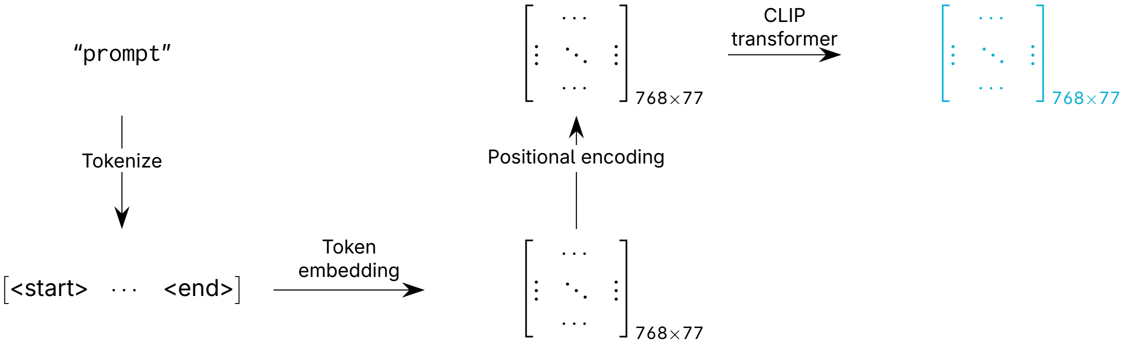
Generating Prompt Embedding Using CLIP



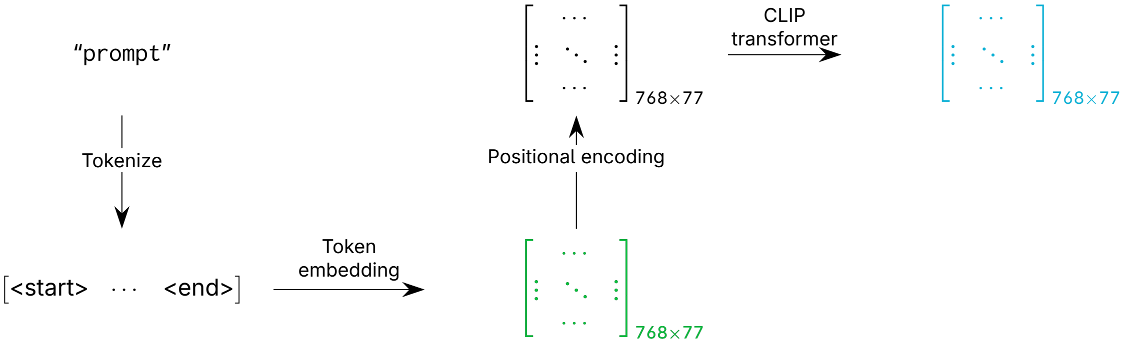
Generating Prompt Embedding Using CLIP



Generating Prompt Embedding Using CLIP



Generating Prompt Embedding Using CLIP

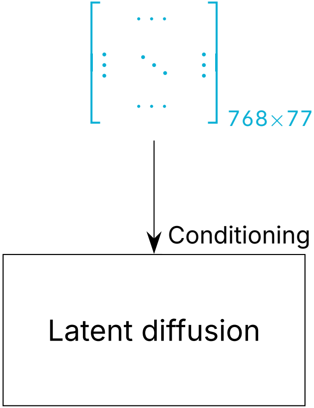


This embedding needs to be trained to control the prompt tokens

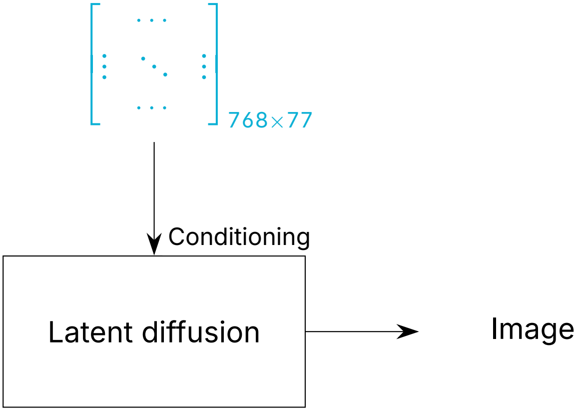
Computing Aesthetic Score Based on Schuhmann [4]

$$\begin{bmatrix} & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & \end{bmatrix}_{768\times 77}$$

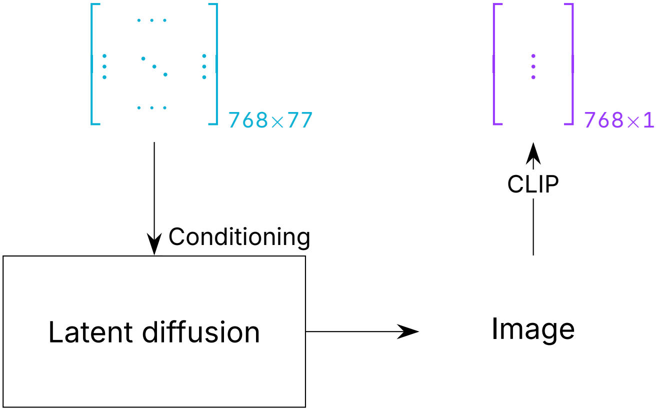
Computing Aesthetic Score Based on Schuhmann [4]



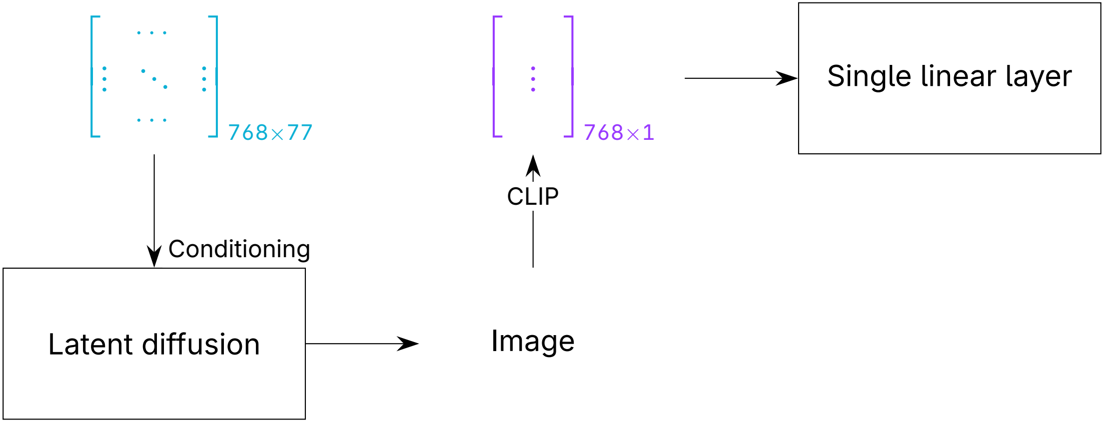
Computing Aesthetic Score Based on Schuhmann [4]



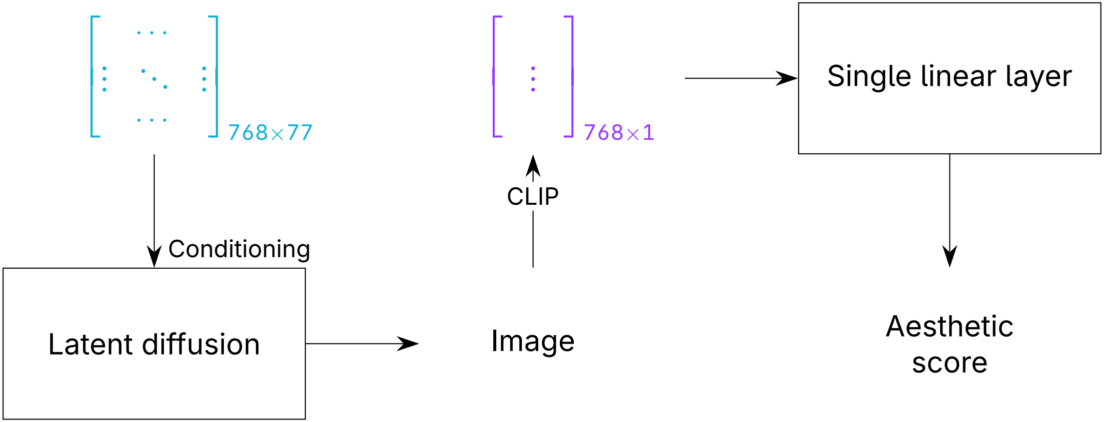
Computing Aesthetic Score Based on Schuhmann [4]



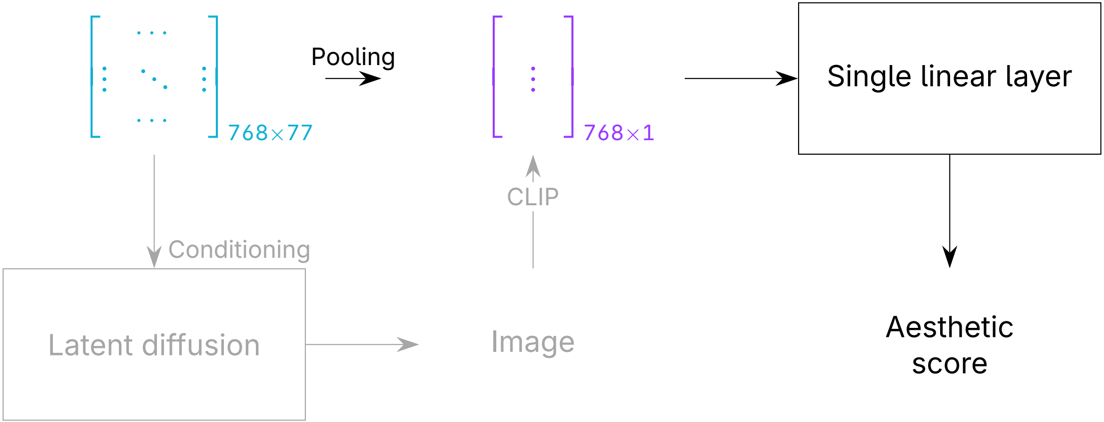
Computing Aesthetic Score Based on Schuhmann [4]



Computing Aesthetic Score Based on Schuhmann [4]



Computing Aesthetic Score Based on Schuhmann [4]



- Potential shortcut because CLIP space is the same for images and text

Constructing a Pipeline

“prompt”

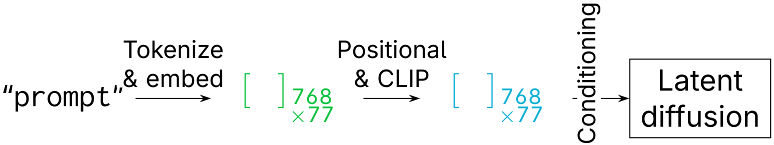
Constructing a Pipeline

“prompt” $\xrightarrow{\text{Tokenize \& embed}}$ $[\quad]_{768 \times 77}$

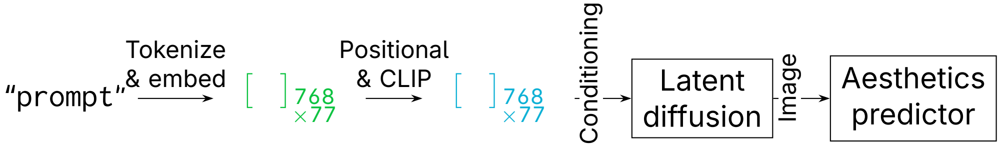
Constructing a Pipeline

$$\texttt{“prompt”} \xrightarrow{\text{Tokenize \& embed}} [\quad]_{768 \times 77} \xrightarrow{\text{Positional \& CLIP}} [\quad]_{768 \times 77}$$

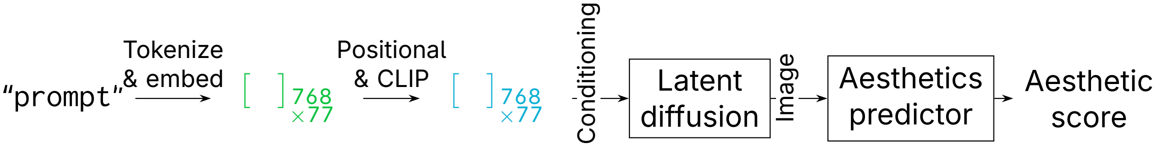
Constructing a Pipeline



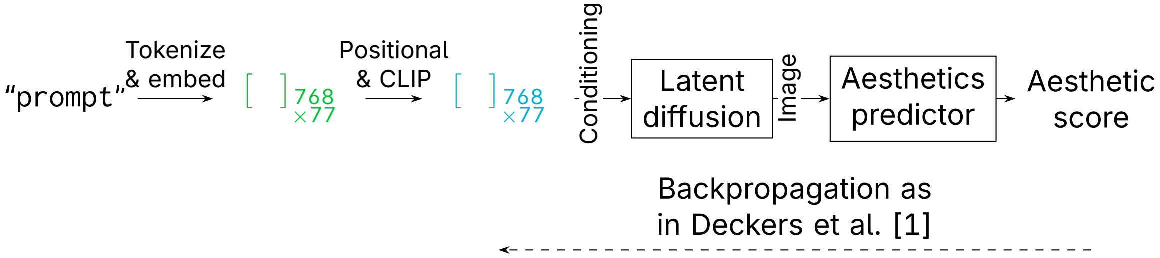
Constructing a Pipeline



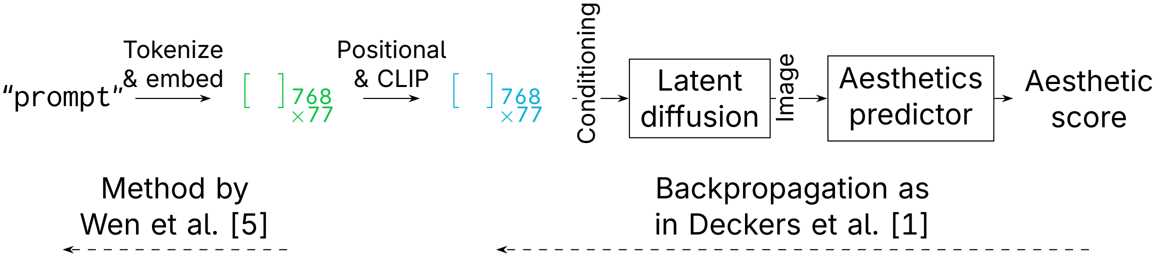
Constructing a Pipeline



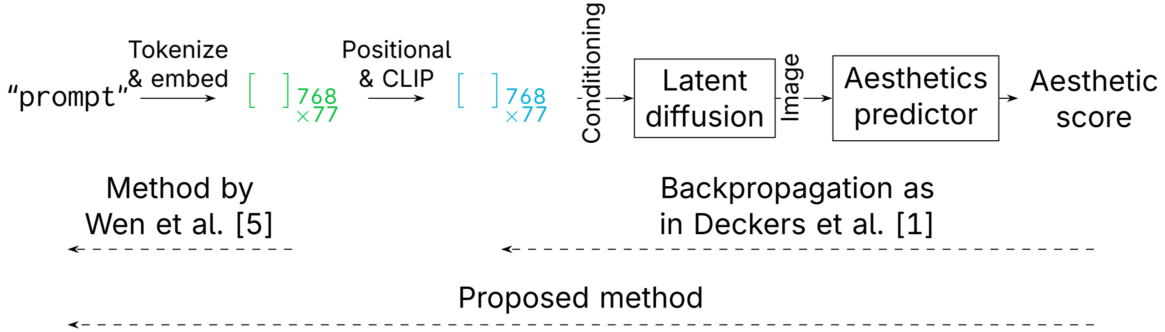
Constructing a Pipeline



Constructing a Pipeline



Constructing a Pipeline



Projection to Find Token Representation for Given Prompt Embedding

Given embeddings

$$\begin{bmatrix} \quad \end{bmatrix}_{768 \times 77}$$

$$\begin{bmatrix} \quad \end{bmatrix}_{768 \times 77}$$

$$\begin{bmatrix} \quad \end{bmatrix}_{768 \times 77}$$

Projection to Find Token Representation for Given Prompt Embedding

Given embeddings

$$[\quad]_{768\times 77} \xleftrightarrow{\text{equals}} [\quad]_{768\times 77} \longleftrightarrow \texttt{"some prompt"}$$

$$[\quad]_{768\times 77}$$

$$[\quad]_{768\times 77}$$

Projection to Find Token Representation for Given Prompt Embedding

Given embeddings

$[\quad]_{768\times77} \xleftrightarrow{\text{equals}} [\quad]_{768\times77} \longleftrightarrow$ `“some prompt”`

$[\quad]_{768\times77} \xrightarrow{\text{gets projected to}} [\quad]_{768\times77} \longleftrightarrow$ `“some other prompt”`

$[\quad]_{768\times77}$

Projection to Find Token Representation for Given Prompt Embedding

Given embeddings

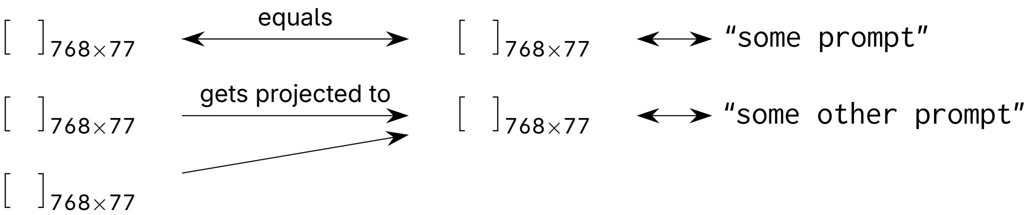
$[\quad]_{768\times77} \xleftrightarrow{\text{equals}} [\quad]_{768\times77} \longleftrightarrow$ “some prompt”

$[\quad]_{768\times77} \xrightarrow{\text{gets projected to}} [\quad]_{768\times77} \longleftrightarrow$ “some other prompt”

$[\quad]_{768\times77} \longrightarrow$

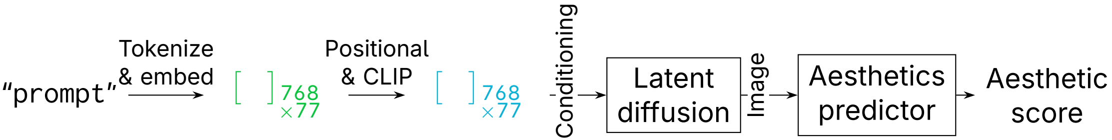
Projection to Find Token Representation for Given Prompt Embedding

Given embeddings

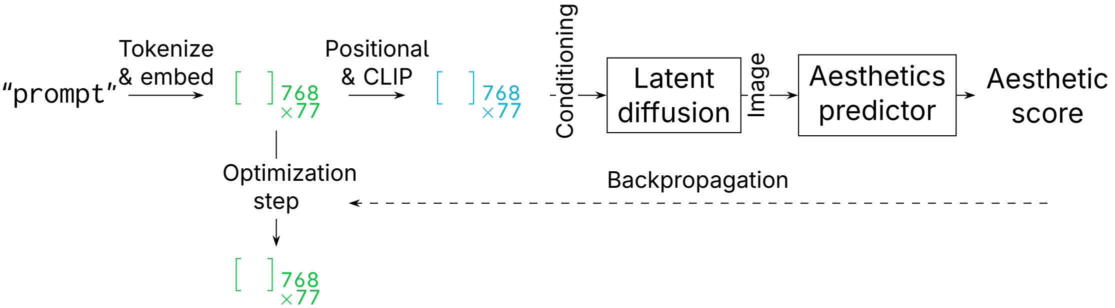


- Wen et al. [5] proposed projection into discrete token space

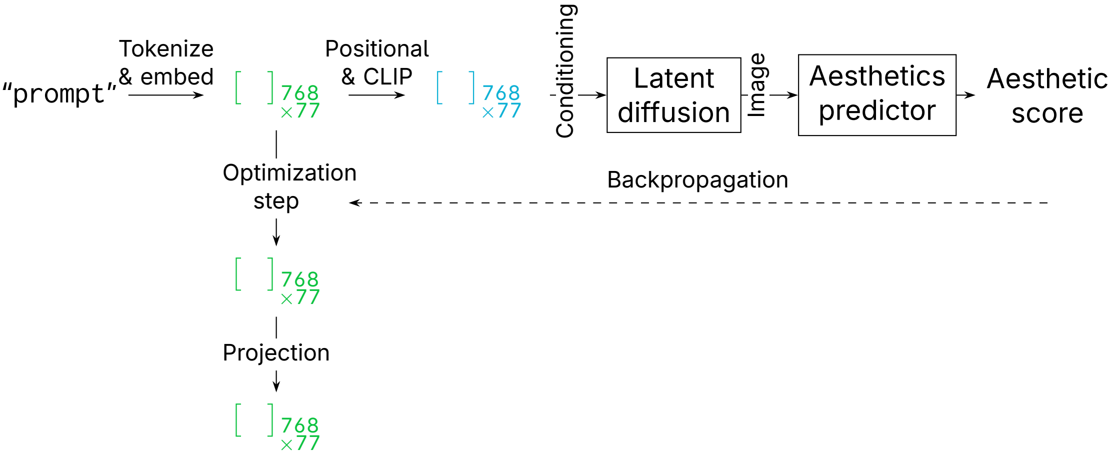
Full Proposed Method



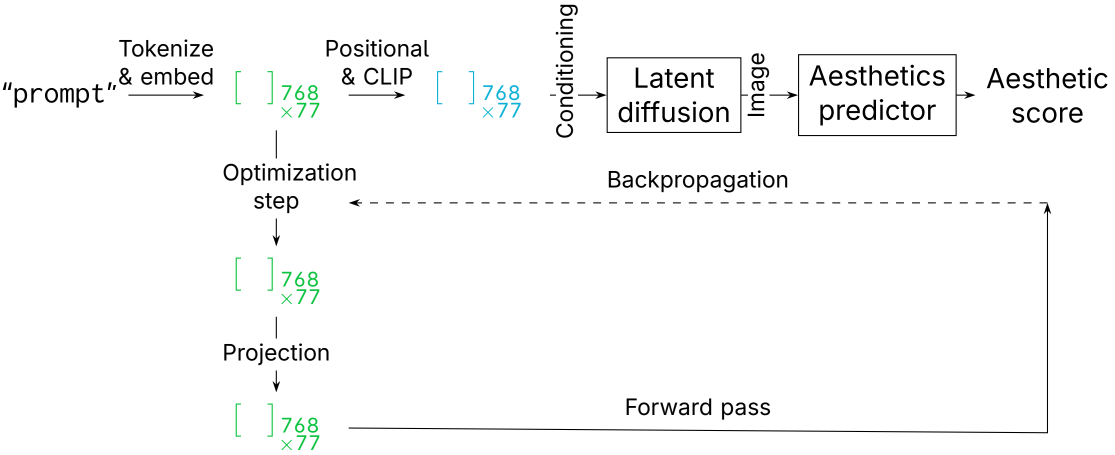
Full Proposed Method



Full Proposed Method



Full Proposed Method



Restricting Manipulation to Suffixes

- Not all 77 token embeddings are altered

Restricting Manipulation to Suffixes

- Not all 77 token embeddings are altered
- We want to change suffix tokens only

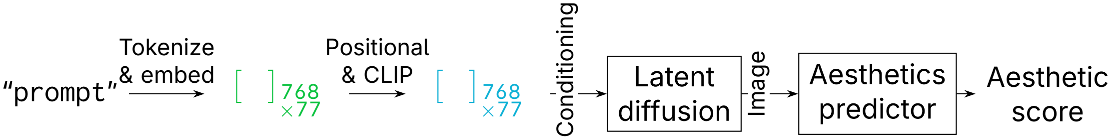
Restricting Manipulation to Suffixes

- Not all 77 token embeddings are altered
- We want to change suffix tokens only
- Prevents alteration of displayed objects

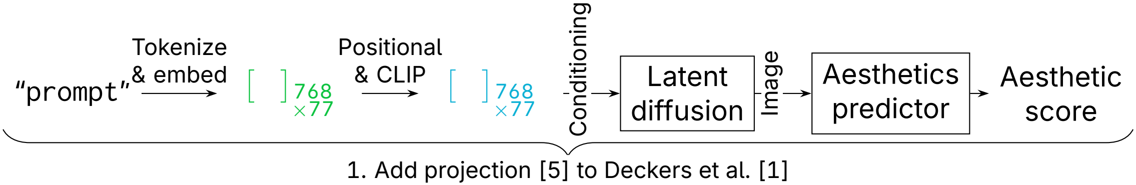
Restricting Manipulation to Suffixes

- Not all 77 token embeddings are altered
- We want to change suffix tokens only
- Prevents alteration of displayed objects
- This resembles prompt modifiers in prompt engineering

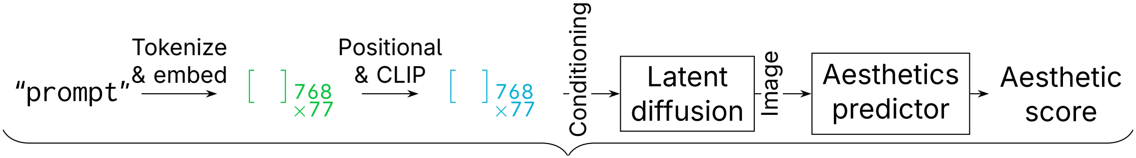
Experiments: Overview



Experiments: Overview



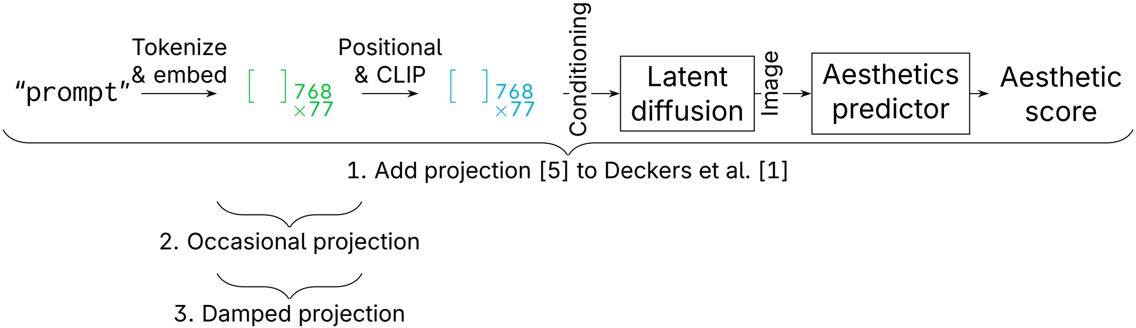
Experiments: Overview



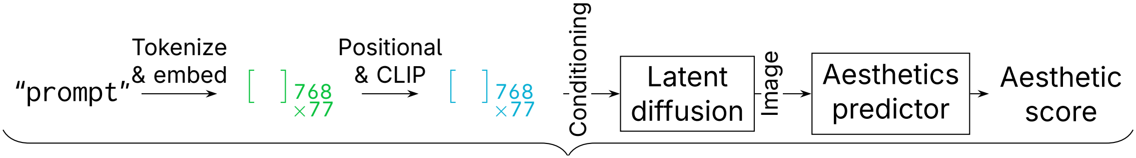
1. Add projection [5] to Deckers et al. [1]

2. Occasional projection

Experiments: Overview



Experiments: Overview



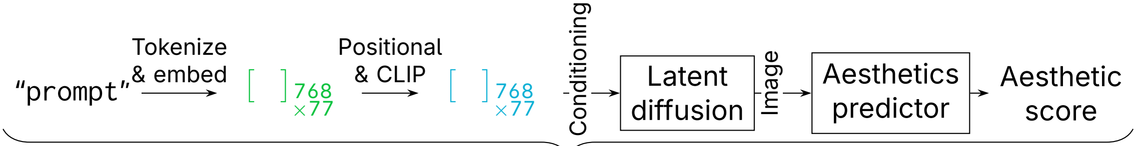
1. Add projection [5] to Deckers et al. [1]

2. Occasional projection

3. Damped projection

4. Skip image generation

Experiments: Overview



1. Add projection [5] to Deckers et al. [1]

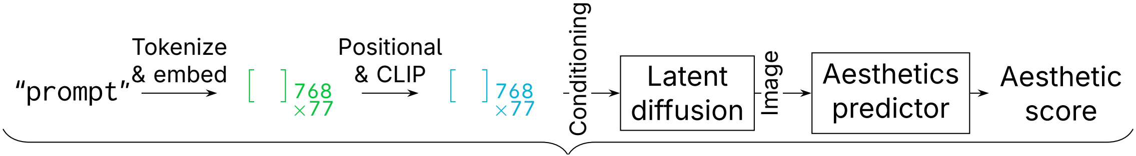
2. Occasional projection

3. Damped projection

4. Skip image generation

5. Generalization over different initial latents

Experiments: Overview



1. Add projection [5] to Deckers et al. [1]

2. Occasional projection

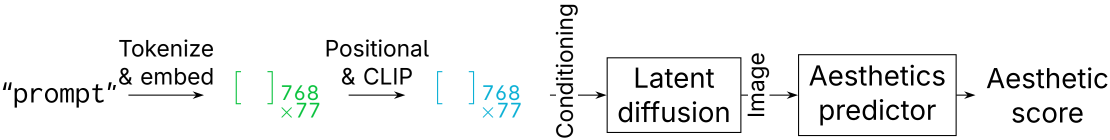
3. Damped projection

4. Skip image generation

5. Generalization over different initial latents

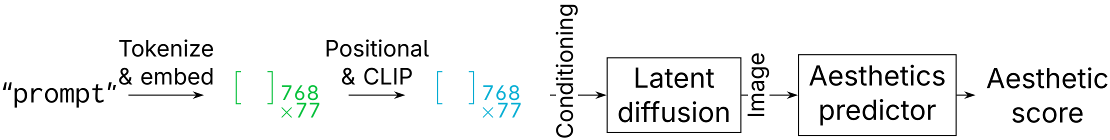
6. Generalization over different prompts

1. Add Projection [5] to Deckers et al. [1]



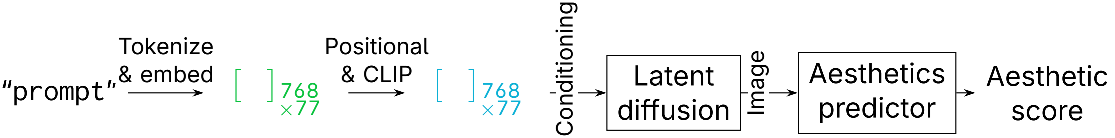
– Alter prompt embedding to improve aesthetic score [1]

1. Add Projection [5] to Deckers et al. [1]



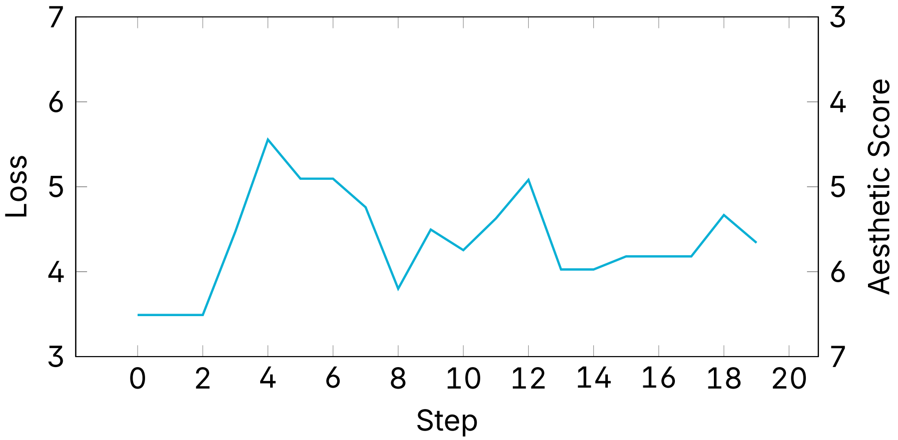
- Alter prompt embedding to improve aesthetic score [1]
- Add projection [5]

1. Add Projection [5] to Deckers et al. [1]



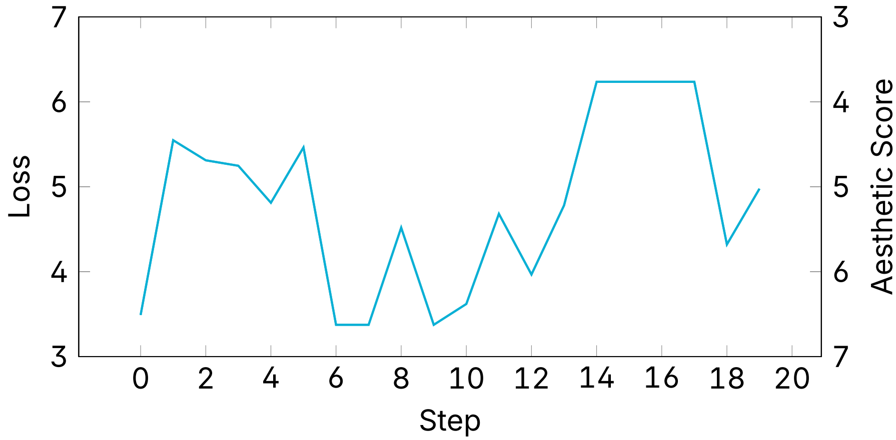
- Alter prompt embedding to improve aesthetic score [1]
- Add projection [5]
  → Results not as good as in Deckers et al. [1]

1. Add Projection [5] to Deckers et al. [1]



Hyperparameter configuration I

1. Add Projection [5] to Deckers et al. [1]

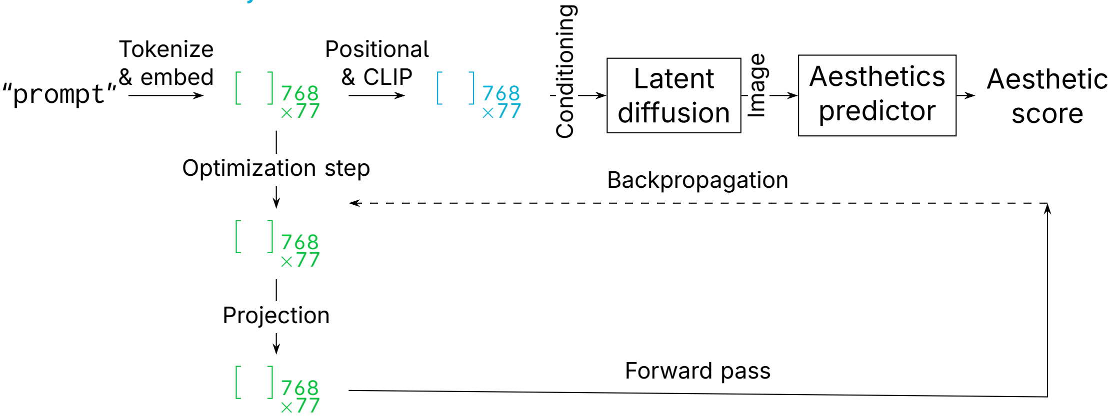


Hyperparameter configuration II

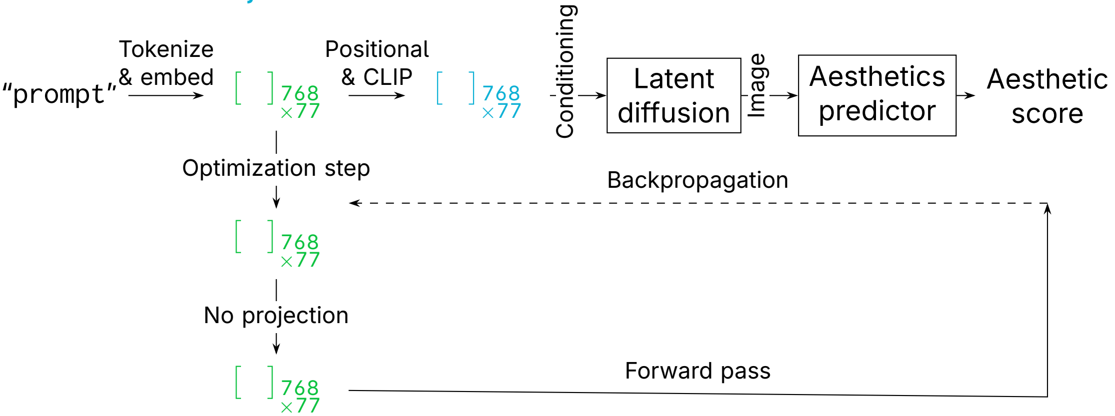
1. Add Projection [5] to Deckers et al. [1]


“realistic spaceship rocket design.
tha utilize tongue pathic oughton richegregearn ”
Before (left) and after (right) optimization

2. Occasional Projection

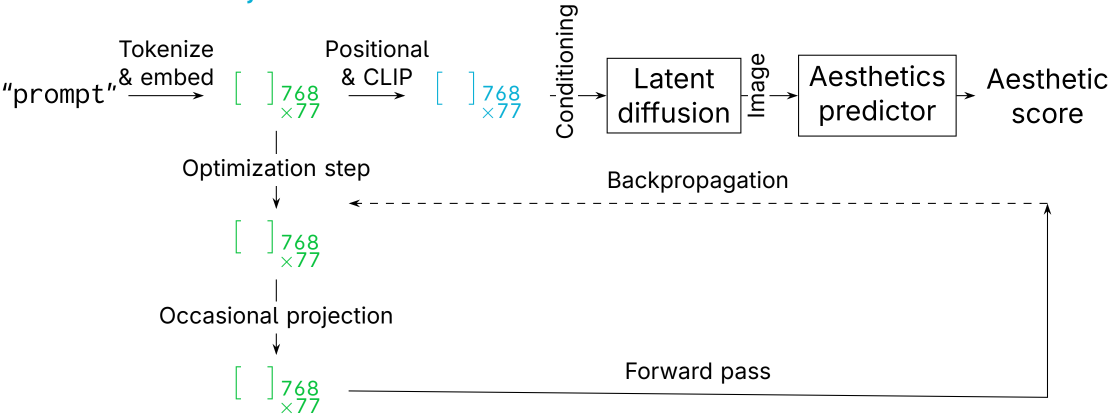


2. Occasional Projection



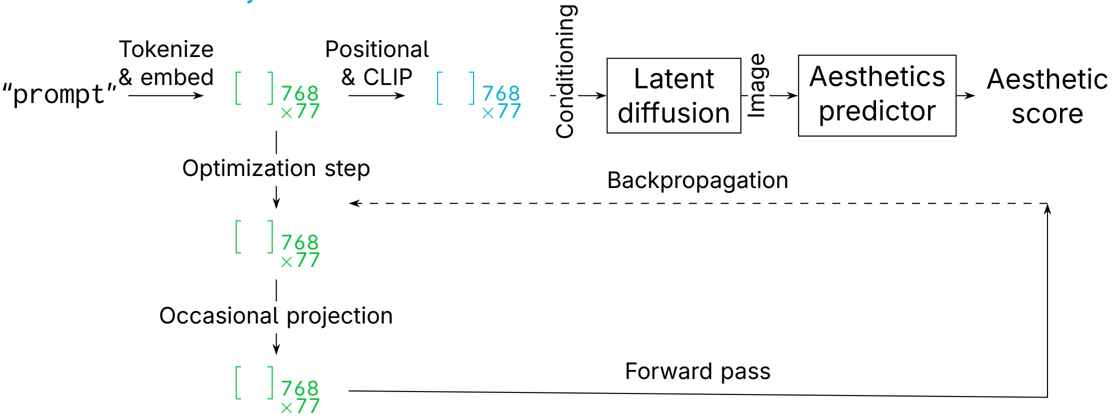
– Do not project before each step (only after final iteration)

2. Occasional Projection



- Do not project before each step (only after final iteration)
- Idea: Compromise between projection and no projection

2. Occasional Projection



- Do not project before each step (only after final iteration)
- Idea: Compromise between projection and no projection
- Add hyperparameter to decide whether to project

2. Occasional Projection

“realistic spaceship rocket design.”

Before (left) and after (right) optimization

- Results are sometimes better, but inconclusive

2. Occasional Projection


"realistic spaceship rocket design. asco indicates brides exited behavihaeasked earn "

Before (left) and after (right) optimization

- Results are sometimes better, but inconclusive

3. Damped Projection (Upcoming)

- Do not project to real embeddings

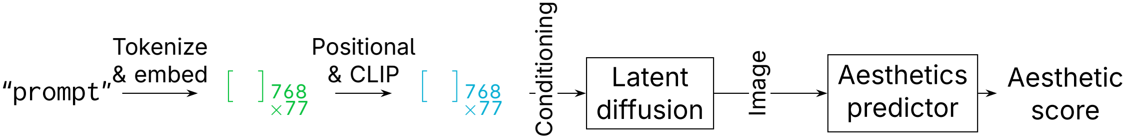
3. Damped Projection (Upcoming)

- Do not project to real embeddings
- Only nudge embeddings in direction of their discrete counterparts

3. Damped Projection (Upcoming)

- Do not project to real embeddings
- Only nudge embeddings in direction of their discrete counterparts
- Should also make exploration of embedding space easier

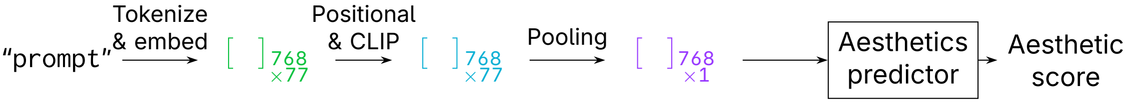
4. Skip Image Generation



4. Skip Image Generation

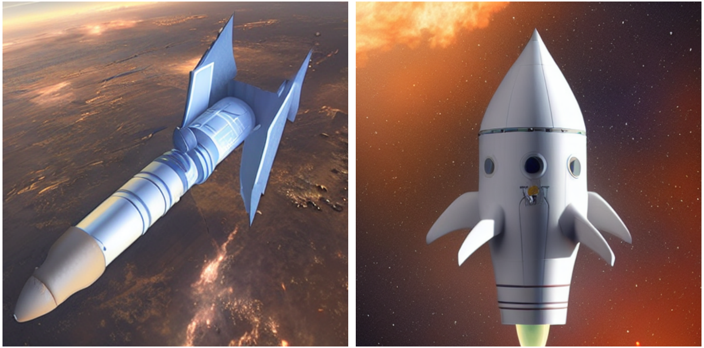
– Skip latent diffusion model

4. Skip Image Generation



- Skip latent diffusion model
- Use aesthetics predictor directly on CLIP embedding

4. Skip Image Generation



“realistic spaceship rocket design.
sts crispy affirting fanny dechomo earn ”
Before (left) and after (right) optimization

4. Skip Image Generation

“realistic spaceship rocket design. minion dumb chalksignalling pooja touches ”
Before (left) and after (right) optimization

5. & 6. Generalization Experiments (Upcoming)

- Initial latents significantly influence final image

5. & 6. Generalization Experiments (Upcoming)

- Initial latents significantly influence final image
  - Train using batches of different initial latents simultaneously

5. & 6. Generalization Experiments (Upcoming)

- Initial latents significantly influence final image
  - Train using batches of different initial latents simultaneously
    $\rightarrow$ Generalize over different initial latents

5. & 6. Generalization Experiments (Upcoming)

- Initial latents significantly influence final image
  - Train using batches of different initial latents simultaneously
    $\rightarrow$ Generalize over different initial latents
- Use same suffix for different prompts during training

5. & 6. Generalization Experiments (Upcoming)

- Initial latents significantly influence final image
  - Train using batches of different initial latents simultaneously
    $\rightarrow$ Generalize over different initial latents
- Use same suffix for different prompts during training
  - Such a suffix may not be optimal for all prompts

5. & 6. Generalization Experiments (Upcoming)

- Initial latents significantly influence final image
  - Train using batches of different initial latents simultaneously
    $\rightarrow$ Generalize over different initial latents
- Use same suffix for different prompts during training
  - Such a suffix may not be optimal for all prompts
  - Find optimal suffix for groups/clusters of prompts

5. & 6. Generalization Experiments (Upcoming)

- Initial latents significantly influence final image
  - Train using batches of different initial latents simultaneously
    $\rightarrow$ Generalize over different initial latents
- Use same suffix for different prompts during training
  - Such a suffix may not be optimal for all prompts
  - Find optimal suffix for groups/clusters of prompts
    $\rightarrow$ Generalize over different prompts

Use Different Optimization Target Metrics

- More granular aesthetics: image composition, contrast, . . .

Use Different Optimization Target Metrics

- More granular aesthetics: image composition, contrast, . . .
- Style score: comic, pixel art, film, painting, . . .

Use Different Optimization Target Metrics

- More granular aesthetics: image composition, contrast, ...
- Style score: comic, pixel art, film, painting, ...
- Artist classifier: da Vinci, van Gogh, ...

Use Different Optimization Target Metrics

- More granular aesthetics: image composition, contrast, ...
- Style score: comic, pixel art, film, painting, ...
- Artist classifier: da Vinci, van Gogh, ...
- Different classes: cat, dog, fish, ...

Use Different Optimization Target Metrics

- More granular aesthetics: image composition, contrast, ...
- Style score: comic, pixel art, film, painting, ...
- Artist classifier: da Vinci, van Gogh, ...
- Different classes: cat, dog, fish, ...
- Safety classifier: SFW, privacy, gender bias, ...

Exploration of Embedding Dimensions

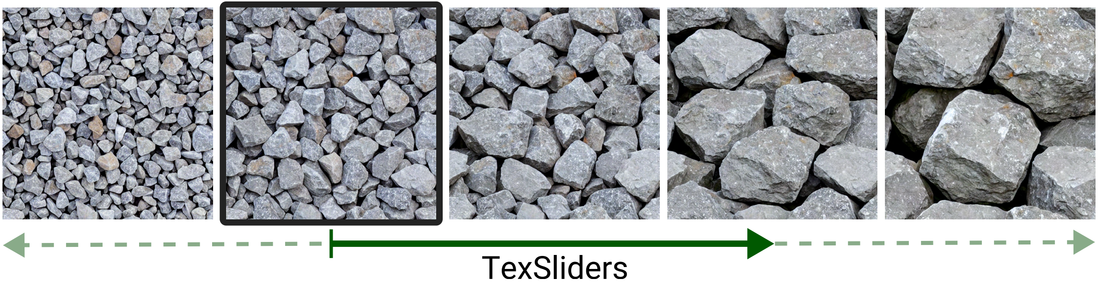


Figure 1: Reproduced from Guerrero-Viu et al. [2]

- Let users control embedding dimensions akin to Guerrero-Viu et al. [2]

Exploration of Embedding Dimensions

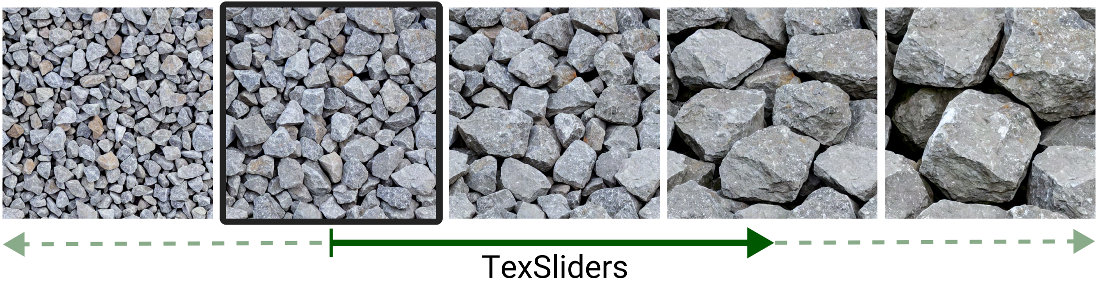


Figure 1: Reproduced from Guerrero-Viu et al. [2]

- Let users control embedding dimensions akin to Guerrero-Viu et al. [2]
  - Manipulation of texture generation

Exploration of Embedding Dimensions

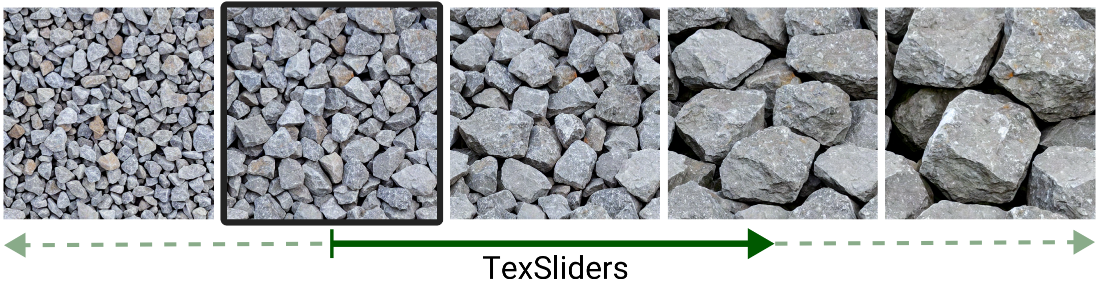


Figure 1: Reproduced from Guerrero-Viu et al. [2]

- Let users control embedding dimensions akin to Guerrero-Viu et al. [2]
  - Manipulation of texture generation
  - Example: given a stone texture, change stone size using slider

Exploration of Embedding Dimensions

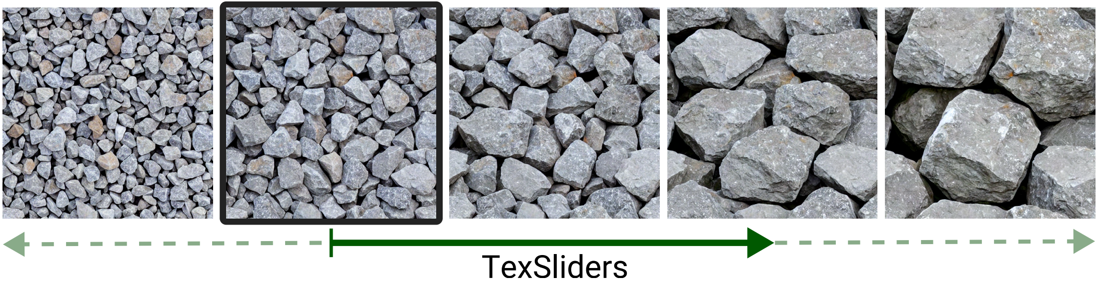


Figure 1: Reproduced from Guerrero-Viu et al. [2]

- Let users control embedding dimensions akin to Guerrero-Viu et al. [2]
  - Manipulation of texture generation
  - Example: given a stone texture, change stone size using slider
- Changes should be resembled in text

Exploration of Embedding Dimensions

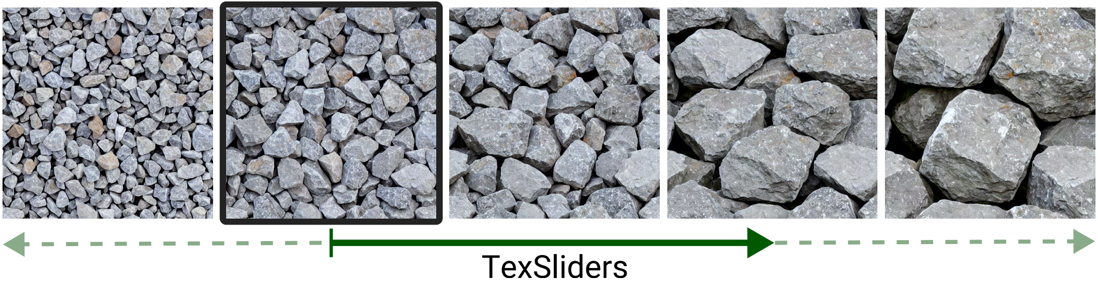


Figure 1: Reproduced from Guerrero-Viu et al. [2]

- Let users control embedding dimensions akin to Guerrero-Viu et al. [2]
  - Manipulation of texture generation
  - Example: given a stone texture, change stone size using slider
- Changes should be resembled in text

  → Useful adapter for the Infinite Index Explorer

Exploration of Affix Types

- Explore differences between prefix, infix and suffix

Exploration of Affix Types

- Explore differences between prefix, infix and suffix
- For infix: choice of insertion in prompt

Exploration of Affix Types

- Explore differences between prefix, infix and suffix
- For infix: choice of insertion in prompt
- Use replacements (beyond affixes)

Exploration of Affix Types

- Explore differences between prefix, infix and suffix
- For infix: choice of insertion in prompt
- Use replacements (beyond affixes)
- Might introduce alteration of displayed objects

Conclusion

- Assisted prompt engineering required

Conclusion

- Assisted prompt engineering required
- Our proposed solution: Generate prompt suffix to optimize aesthetic score

Conclusion

- Assisted prompt engineering required
- Our proposed solution: Generate prompt suffix to optimize aesthetic score
- Results show improvement of aesthetic score

Conclusion

- Assisted prompt engineering required
- Our proposed solution: Generate prompt suffix to optimize aesthetic score
- Results show improvement of aesthetic score
- Current suffixes lack interpretabilty

Conclusion

- Assisted prompt engineering required
- Our proposed solution: Generate prompt suffix to optimize aesthetic score
- Results show improvement of aesthetic score
- Current suffixes lack interpretabilty
- Improvements to generalization might increase interpretabilty

Conclusion

- Assisted prompt engineering required
- Our proposed solution: Generate prompt suffix to optimize aesthetic score
- Results show improvement of aesthetic score
- Current suffixes lack interpretabilty
- Improvements to generalization might increase interpretabilty
- Final goal: Find list of suffixes which work for every prompt and image seed

Conclusion

- Assisted prompt engineering required
- Our proposed solution: Generate prompt suffix to optimize aesthetic score
- Results show improvement of aesthetic score
- Current suffixes lack interpretabilty
- Improvements to generalization might increase interpretabilty
- Final goal: Find list of suffixes which work for every prompt and image seed
- Skipping latent diffusion is feasible, further investigation required

Conclusion

- Assisted prompt engineering required
- Our proposed solution: Generate prompt suffix to optimize aesthetic score
- Results show improvement of aesthetic score
- Current suffixes lack interpretabilty
- Improvements to generalization might increase interpretabilty
- Final goal: Find list of suffixes which work for every prompt and image seed
- Skipping latent diffusion is feasible, further investigation required

Thank you!

References I

[1] Niklas Deckers, Julia Peters, and Martin Potthast. Manipulating embeddings of stable diffusion prompts. *CoRR*, abs/2308.12059, 2023. doi: 10.48550/ARXIV.2308.12059. URL `https://doi.org/10.48550/arXiv.2308.12059`.

[2] Julia Guerrero-Viu, Milos Hasan, Arthur Roullier, Midhun Harikumar, Yiwei Hu, Paul Guerrero, Diego Gutierrez, Belen Masia, and Valentin Deschaintre. Texsliders: Diffusion-based texture editing in clip space. *arXiv preprint arXiv:2405.00672*, 2024. URL `https://arxiv.org/abs/2405.00672`.

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01042. URL `https://doi.org/10.1109/CVPR52688.2022.01042`.

References II

[4] Christoph Schuhmann. Laion-aesthetics, 8 2022. URL https://laion.ai/blog/laion-aesthetics/.

[5] Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 – 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a00548031e4647b13042c97c922fadf1-Abstract-Conference.html.

Introduction

Related Work

Method

Experiments

Future Work

Conclusion

References

Appendix

Generalization Over Different Initial Latents (Upcoming)

- Initial latents significantly influence final image

Generalization Over Different Initial Latents (Upcoming)

- Initial latents significantly influence final image
- Users do not typically choose seed or initial latent

Generalization Over Different Initial Latents (Upcoming)

- Initial latents significantly influence final image
- Users do not typically choose seed or initial latent
- Latent specific suffix has limited use

Generalization Over Different Initial Latents (Upcoming)

- Initial latents significantly influence final image
- Users do not typically choose seed or initial latent
- Latent specific suffix has limited use
  $\rightarrow$ Train using batches of different initial latents simultaneously

Test Generalization of Suffixes Over Multiple Prompts (Upcoming)

- Generate optimal suffix on one prompt

Test Generalization of Suffixes Over Multiple Prompts (Upcoming)

- Generate optimal suffix on one prompt
- Test effect over 10 other prompts

Test Generalization of Suffixes Over Multiple Prompts (Upcoming)

- Generate optimal suffix on one prompt
- Test effect over 10 other prompts
  - Repeat for each of the 10 other prompts

Train Prompt Independent Suffix (Upcoming)

- If suffix does not generalize in previous experiment:
  use same suffix for different prompts during training

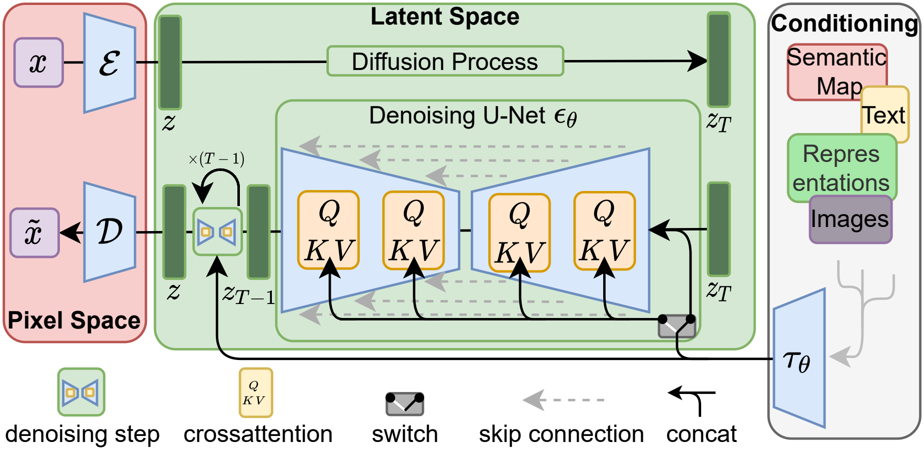
Train Prompt Independent Suffix (Upcoming)

- If suffix does not generalize in previous experiment:
  use same suffix for different prompts during training
- Such a suffix may not be optimal

Train Prompt Independent Suffix (Upcoming)

- If suffix does not generalize in previous experiment: use same suffix for different prompts during training
- Such a suffix may not be optimal
  - Find optimal suffix for groups/clusters of prompts

Latent Diffusion [3]



Rerpoduced from Rombach et al. [3]