TUE-AM-181

Uncovering the Disentanglement Capability in Text-to-Image Diffusion Models

Qiucheng Wu[1], Yujian Liu[1], Handong Zhao[2], Ajinkya Kale[2], Trung Bui[2], Tong Yu[2], Zhe Lin[2], Yang Zhang[3], Shiyu Chang[1]

[1]UC Santa Barbara, [2]Adobe Research, [3]MIT-IMB Watson AI Lab

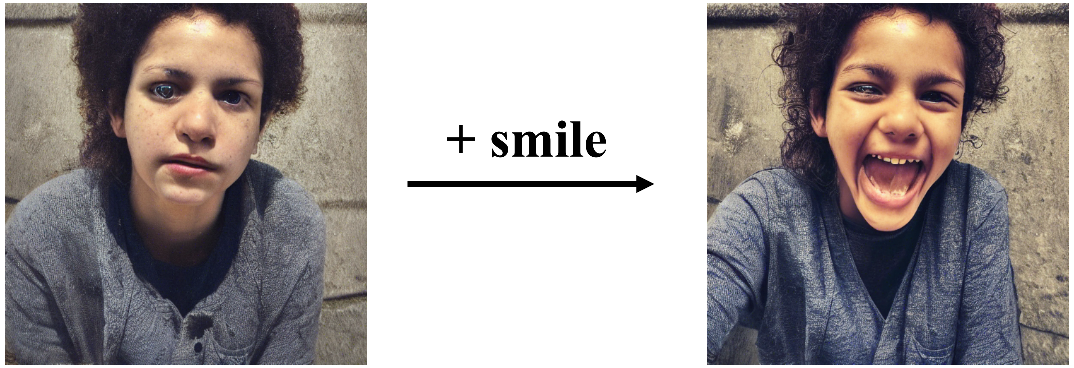
Presenter: Qiucheng Wu

Project

Code

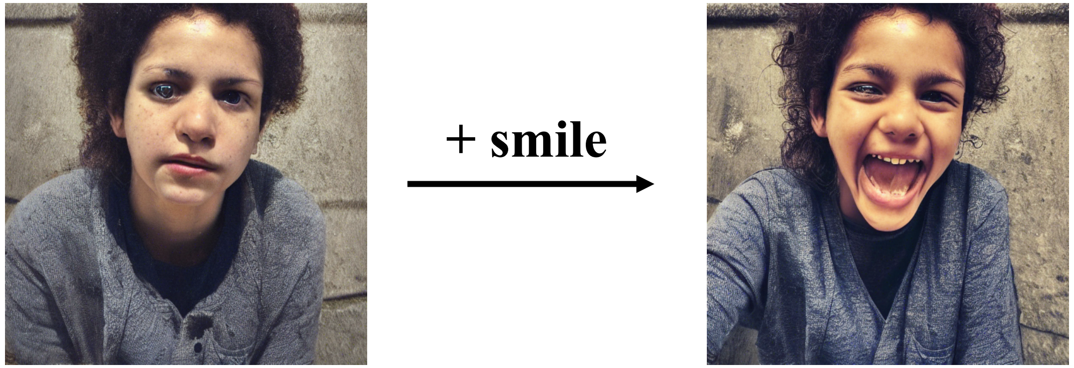
Overview

- **Disentanglement** is a desired property in generative models**.**
  - E.g., a **disentangled** model can generate a person with **expression changed** but **identity preserved**.



Overview

- **Disentanglement** is a desired property in generative models**.**
  - E.g., a **disentangled** model can generate a person with **expression changed** but **identity preserved**.

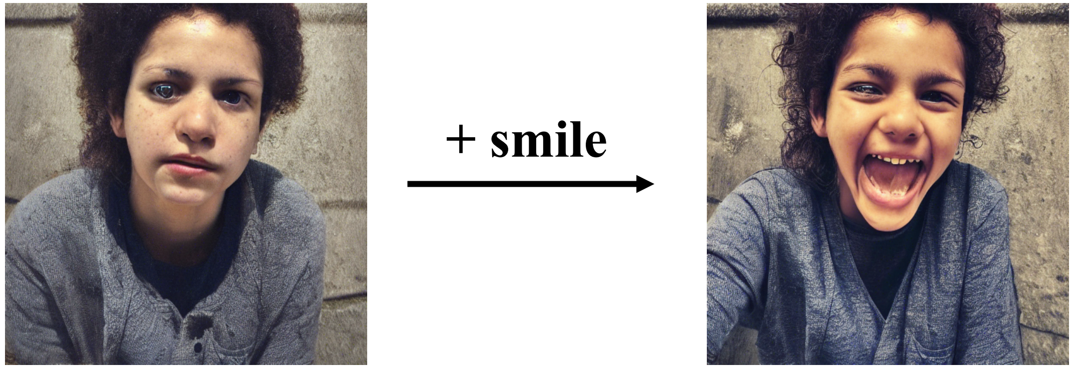


- Many generative models (e.g., GANs) inherently have this disentanglement property.
- Our research question:

*Does a pre-trained **text-to-image model** have the **disentanglement capability?***

Overview

- **Disentanglement** is a desired property in generative models**.**
    - E.g., a **disentangled** model can generate a person with **expression changed** but **identity preserved**.



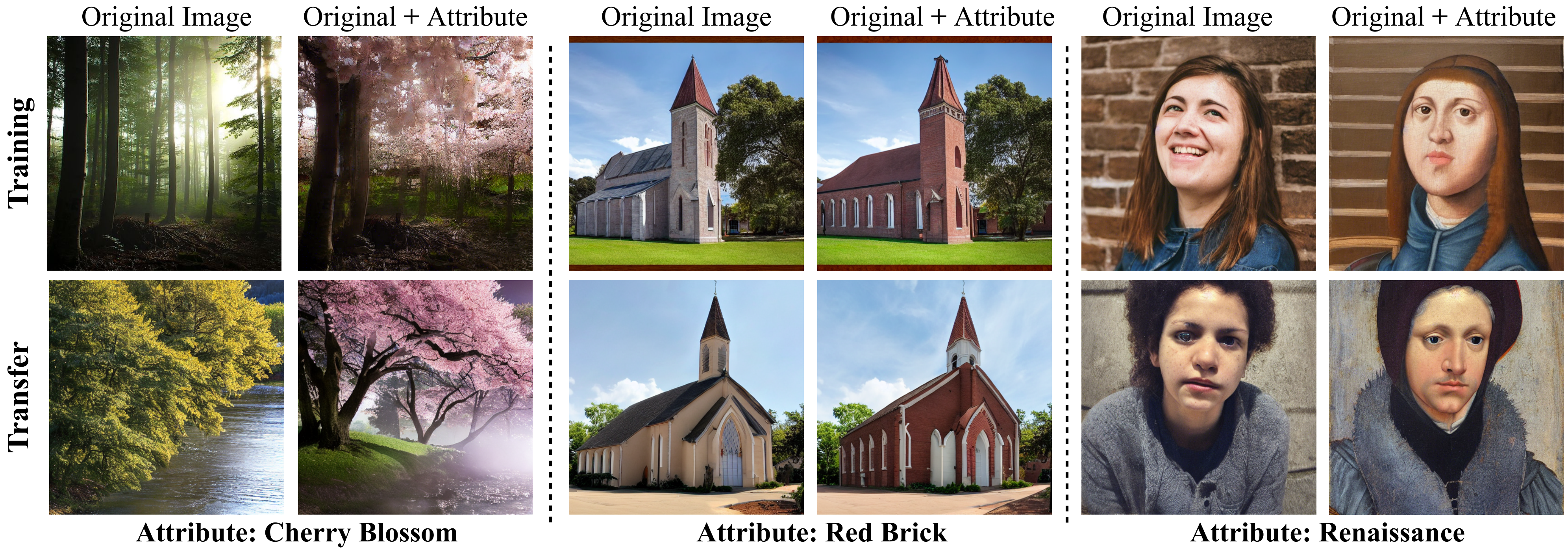
- Many generative models (e.g., GANs) inherently have this property.
- Our research question:

*Does a pre-trained **text-to-image model** have the **disentanglement capability?***

Yes!

Overview

- In this work, we discover the **disentanglement capability** in **text-to-image diffusion** model.
- Our finding leads to a simple disentangle editing framework.
- The framework can effectively edit a wide range of attributes without changing the contents.



Disentanglement in Diffusion: Preliminary Exploration

- We find the stable diffusion model inherently enables disentanglement.
- Goal: Generate an image of the **same person** with only **facial expression** changed.

Disentanglement in Diffusion: Preliminary Exploration

- We find the stable diffusion model inherently enables disentanglement.
- Goal: Generate an image of the **same person** with only **facial expression** changed.
- Consider two text input embeddings:

  $\boldsymbol{c}^{(\mathbf{0})}$(style-neutral): "A photo of person"

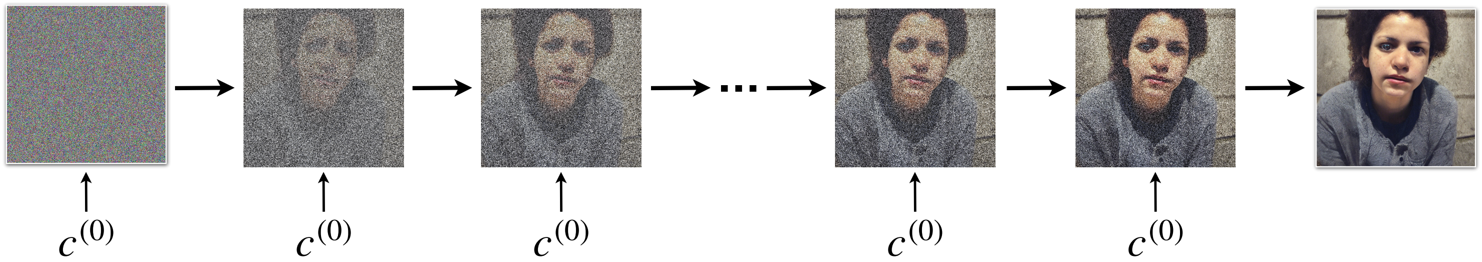
  $\boldsymbol{c}^{(\mathbf{1})}$(with style): "A photo of person with smile"

Disentanglement in Diffusion: Preliminary Exploration

$\boldsymbol{c}^{(\mathbf{0})}$(style-neutral): "A photo of person"

$\boldsymbol{c}^{(\mathbf{1})}$(with style): "A photo of person with smile"

- **Original: Directly feed $\boldsymbol{c}^{(\mathbf{0})}$.**



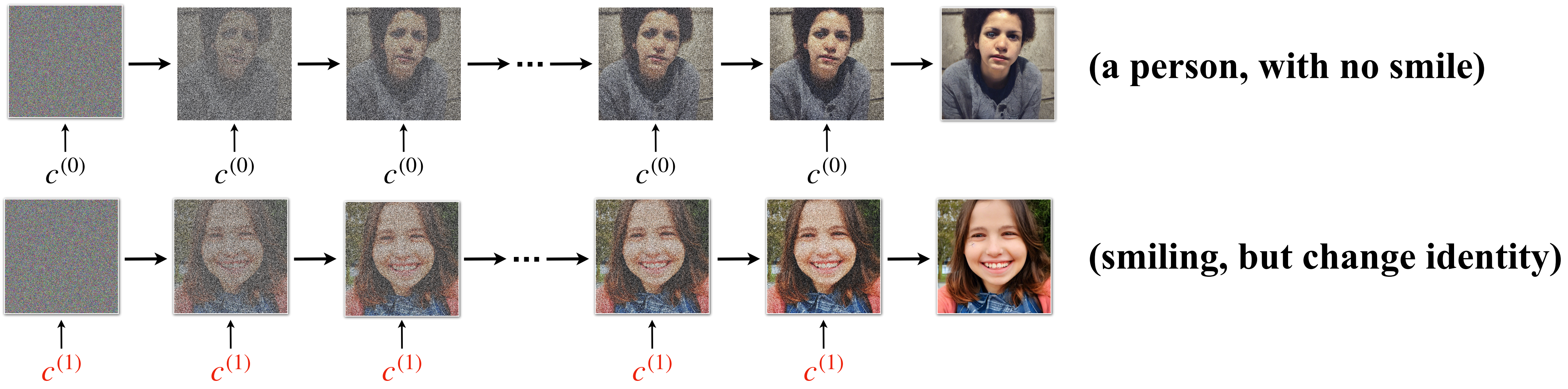
(a person, with no smile)

Disentanglement in Diffusion: Preliminary Exploration

$\boldsymbol{c}^{(0)}$(style-neutral): "A photo of person"

$\boldsymbol{c}^{(1)}$(with style): "A photo of person with smile"

- **Case 1: Full replacement**

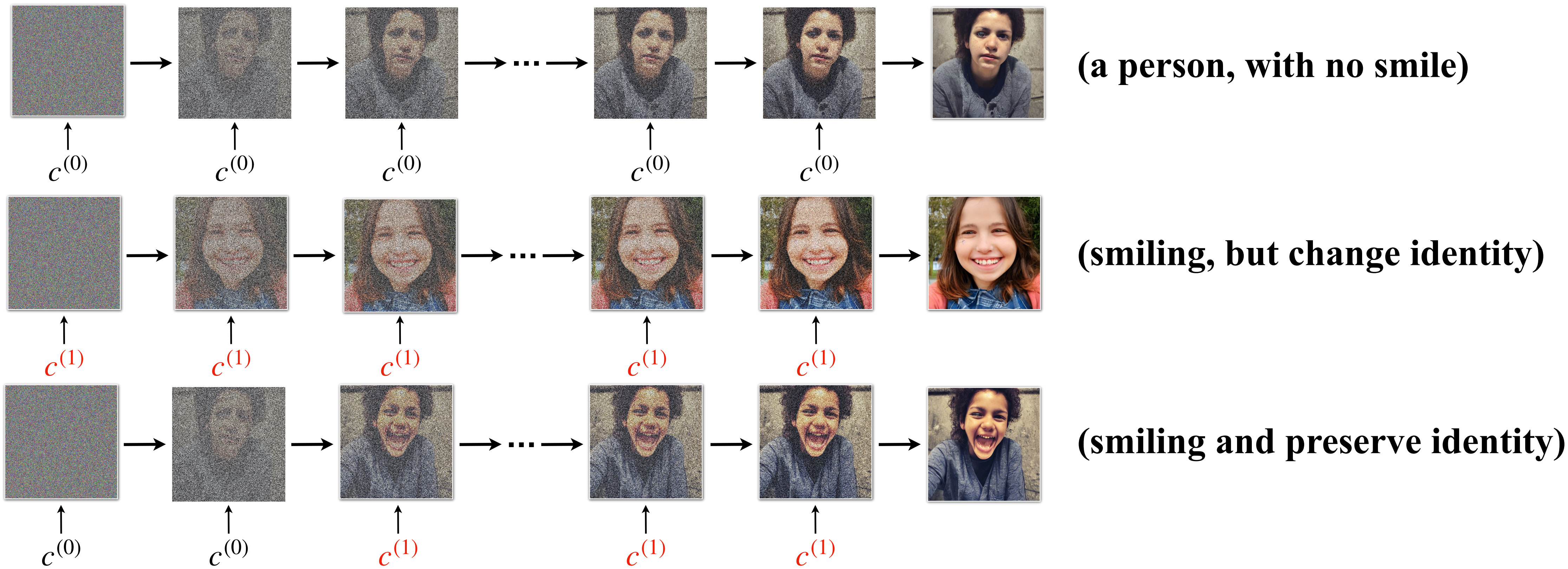


Disentanglement in Diffusion: Preliminary Exploration

$\boldsymbol{c}^{(0)}$(style-neutral): "A photo of person"

$\boldsymbol{c}^{(1)}$(with style): "A photo of person with smile"

- **Case 2: Partial replacement**

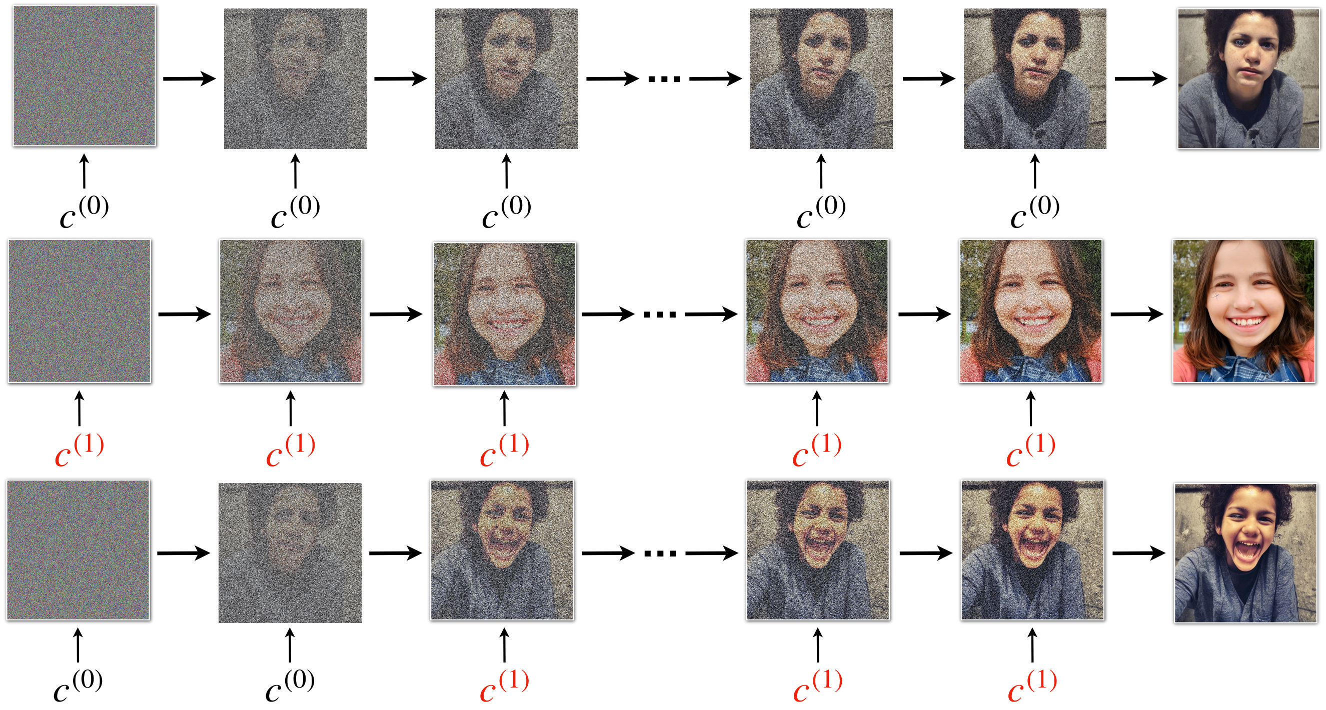


Disentanglement in Diffusion: Preliminary Experiment

- Goal: Generate an image of the **same person** with only **facial expression** changed.
- Consider two text input embeddings:

  $\boldsymbol{c}^{(\mathbf{0})}$: "A photo of person"

  $\boldsymbol{c}^{(\mathbf{1})}$: "A photo of person with smile"



- **Conclusion:**
  - The stable diffusion model inherently enables disentanglement.
  - The disentanglement can be triggered by **partially replacing the text embeddings.**

Optimizing for Disentanglement

- Our method optimizes a soft combination of two text embeddings:
  - $\boldsymbol{c}^{(0)}$: *"A castle"*
  - $\boldsymbol{c}^{(1)}$: *"A children drawing of castle"*

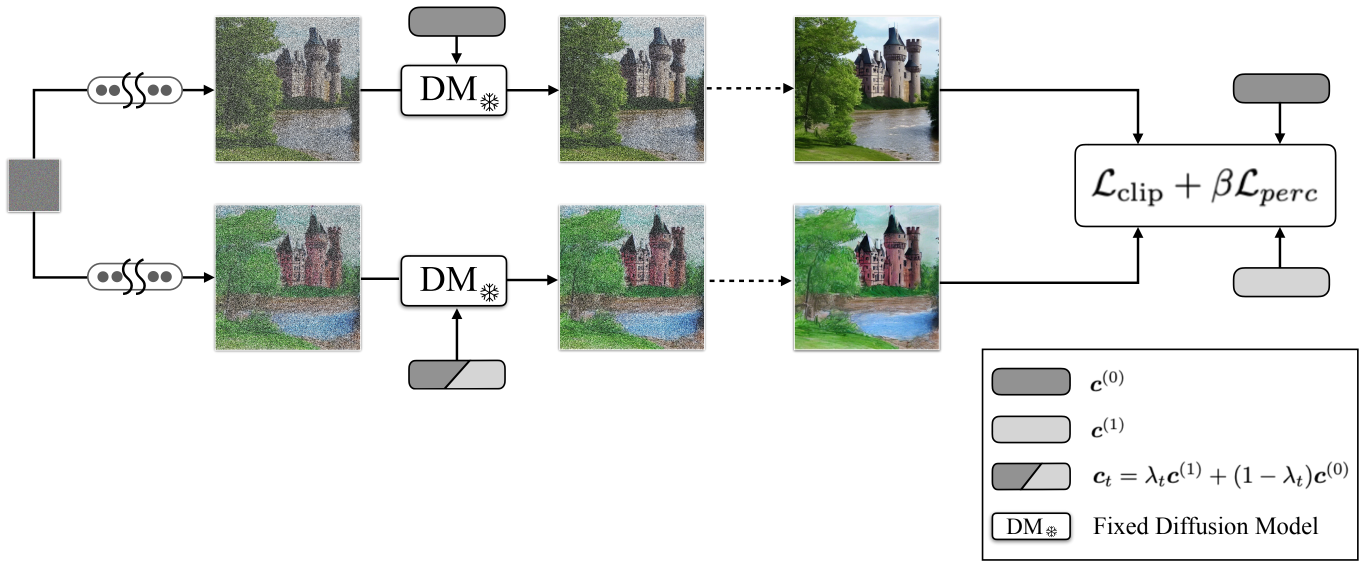
$$\boldsymbol{c_t} = \boldsymbol{\lambda_t} \boldsymbol{c}^{(1)} + (\mathbf{1} - \boldsymbol{\lambda_t}) \boldsymbol{c}^{(0)}$$

Optimizing for Disentanglement

- Our method optimizes a soft combination of two text embeddings:
  - $\boldsymbol{c}^{(0)}$: *"A castle"*
  - $\boldsymbol{c}^{(1)}$: *"A children drawing of castle"*

$$\boldsymbol{c_t} = \boldsymbol{\lambda_t c^{(1)}} + (\boldsymbol{1} - \boldsymbol{\lambda_t})\boldsymbol{c^{(0)}}$$

- The stable diffusion conditions on $\boldsymbol{c_t}$ to synthesize image with modified style (children drawing).
- $\boldsymbol{\lambda_t}$ Optimization:
  - CLIP loss to control style
  - Perceptual loss to preserve content

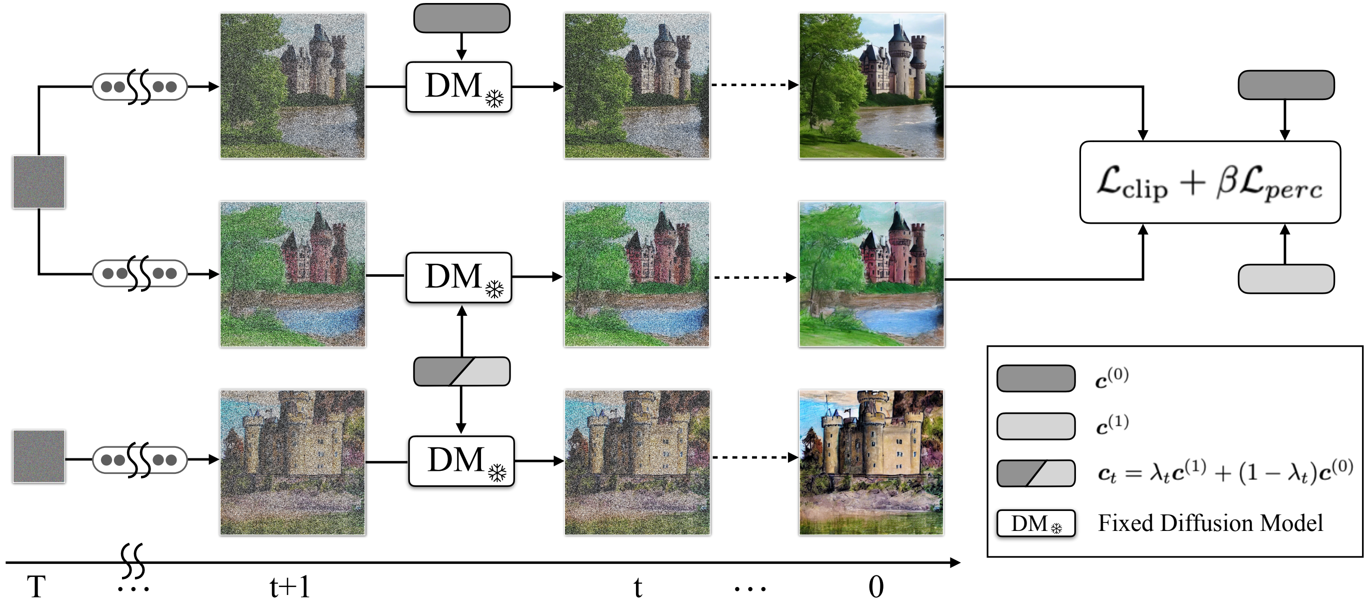


Optimizing for Disentanglement

- Our method optimizes a soft combination of two text embeddings:
  - $\boldsymbol{c}^{(0)}$: *"A castle"*
  - $\boldsymbol{c}^{(1)}$: *"A children drawing of castle"*

$$\boldsymbol{c_t} = \boldsymbol{\lambda_t c^{(1)} + (1 - \lambda_t) c^{(0)}}$$

- The stable diffusion conditions on $\boldsymbol{c_t}$ to synthesize image with modified style (children drawing).
- $\boldsymbol{\lambda_t}$ can be transferred to novel images and lead to similar editing effects.

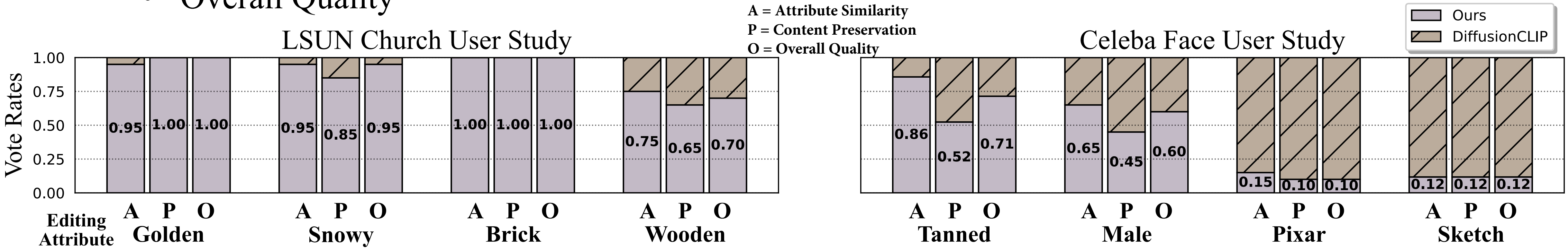


Experiment: Disentanglement Capability

- Our method is able to disentangle a wide range of attributes.
  - Global attributes: scenery styles, architecture materials, etc.
  - Local attributes: facial expressions, etc.

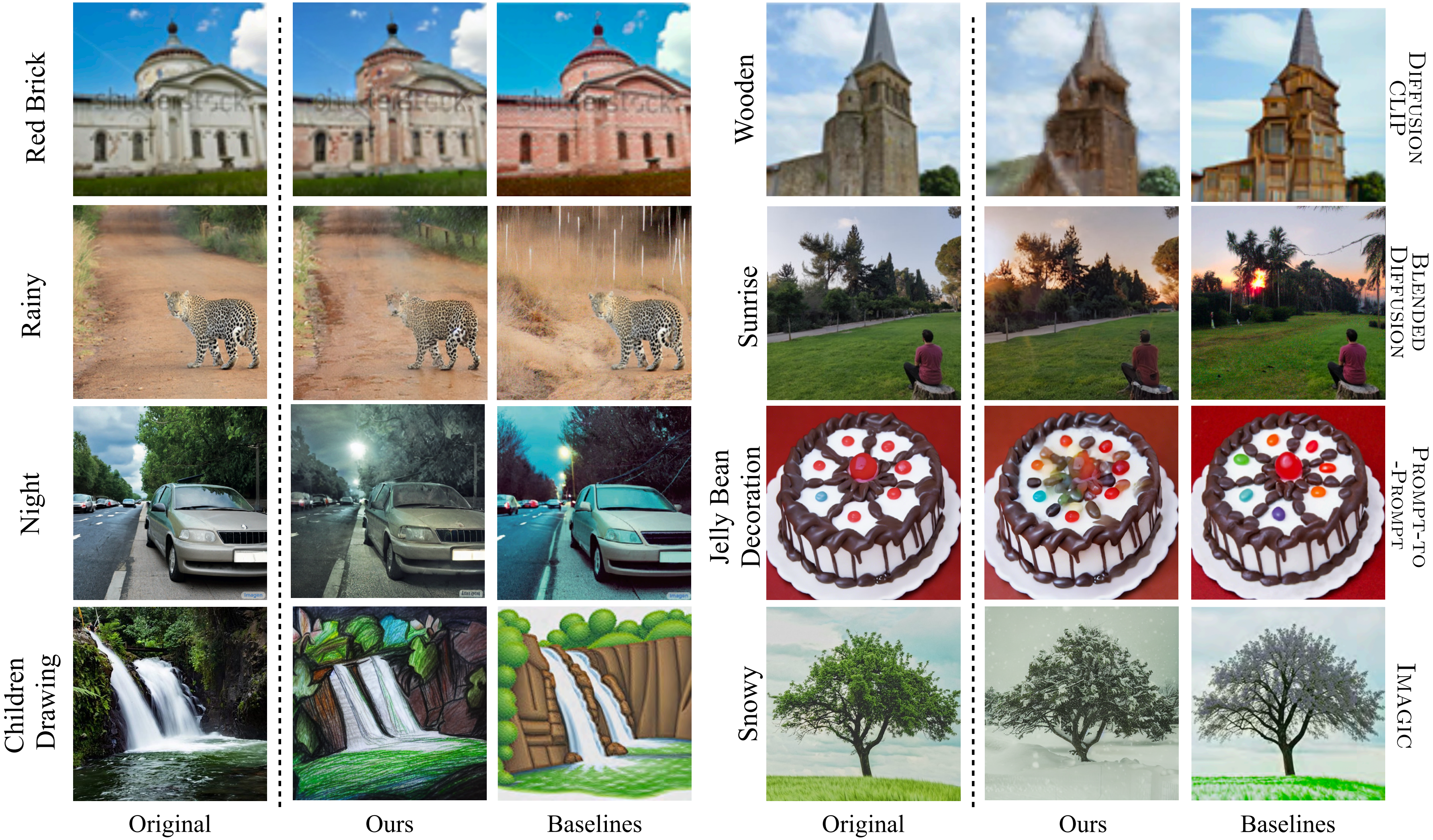
Experiment: Image Editing

- Based on the subjective study, our method shows advantages in image editing.
  - Datasets: LSUN Church (Scene), Celeba Face (Person)
  - Baseline: DiffusionCLIP
  - Our method outperforms DiffusionCLIP in 6 out of 8 attributes with following metrics:
    - Attribute Similarity
    - Content Preservation
    - Overall Quality



Experiment: Image Editing

- Our method shows competitive editing performance compared with strong baselines.



Limitations

		Scenes	Person
✔	**Global**	**Styles** (children drawing, cyberpunk, anime), **Building appearance** (wooden, red brick), **Weather & time** (sunset, night, snowy)	**Styles** (renaissance, Egyptian mural, sketch, Pixar) **Appearance** (young, tanned, male)
	Local	Cherry blossom, rainbow, foothills	**Expressions** (smiling, crying, angry)
✘	**Small edits**	Cake toppings, remove people on the street	Hats, hair colors, earrings

- We explore a wide range of attributes and find **small edits** are hard to be disentangled.
- Diffusion model has weaker control over these fine-grained details.

A photo of person, wearing hat

A cake, jelly beans decorations

Thank you!

Project

Code