# InstructPix2NeRF: Instructed 3D Portrait Editing from a Single Image

Jianhui Li[1], Shilong Liu[1], Zidong Liu[1], Yikai Wang[1], Kaiwen Zheng[1] , Jinhui Xu[2], Jianmin Li[1], Jun Zhu[1,2].
[1] Tsinghua University [2] Shengshu Technology

Halle B
Wed 8 May 4:45 p.m. CST — 6:45 p.m. CST

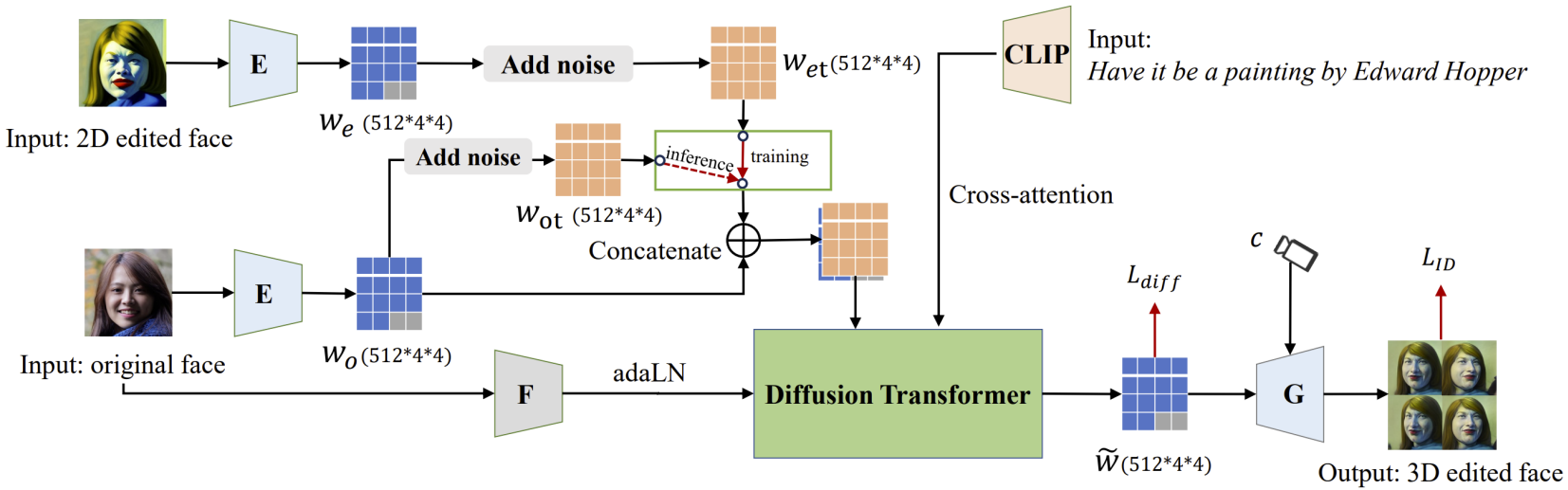
# Summary

## Our goals

- **Text-guided** 3D-aware editing from a single image
- **Multiple** natural language instructions
- **3D Consistency in identity & attributes**
- Editing interactively in several seconds

## Our pipeline

NeRF-based GAN Inversion + Contidional Latent 3D Diffusion Model



# Summary

## Results

Remove the beard

Turn the hair color to red

Make her look like a cartoon

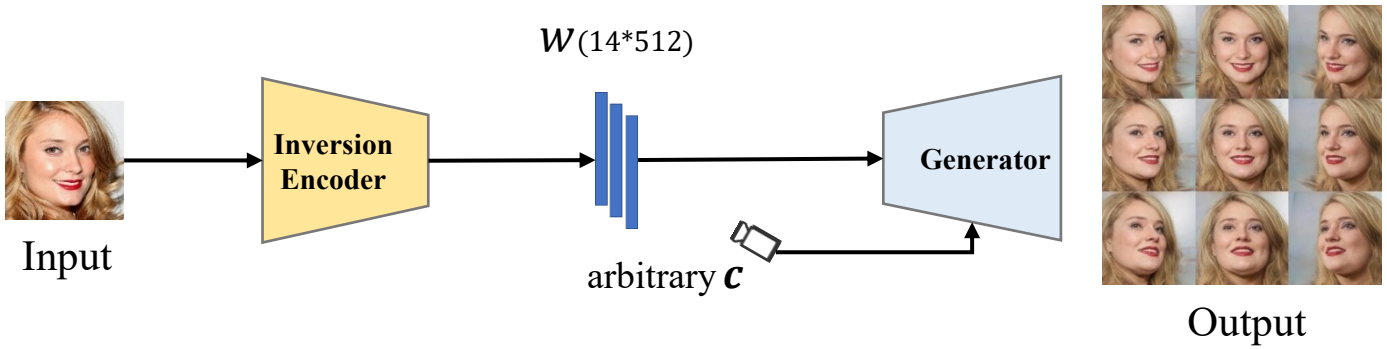
Put eyeglasses on her and turn the portrait into a bronze statue

# Related Works

## 3D GAN Inversion

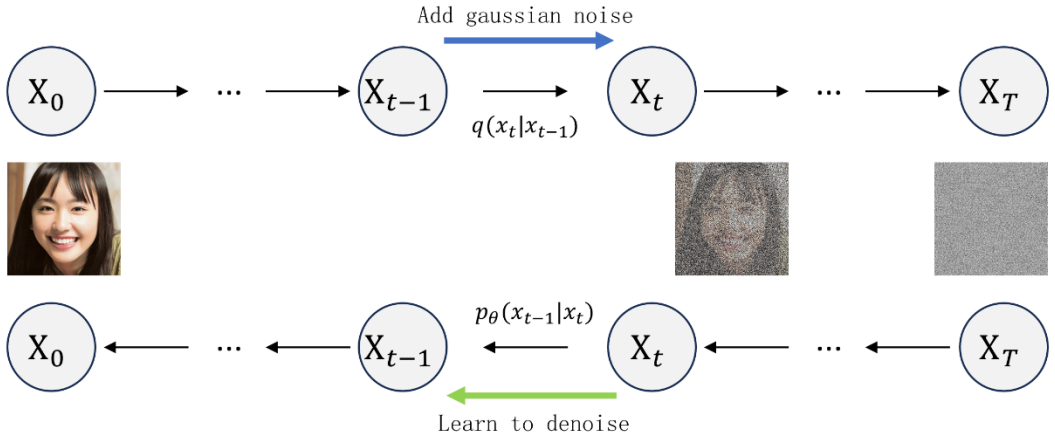
- PREIM3D

  mapping the face to $W^+$ Latent Space



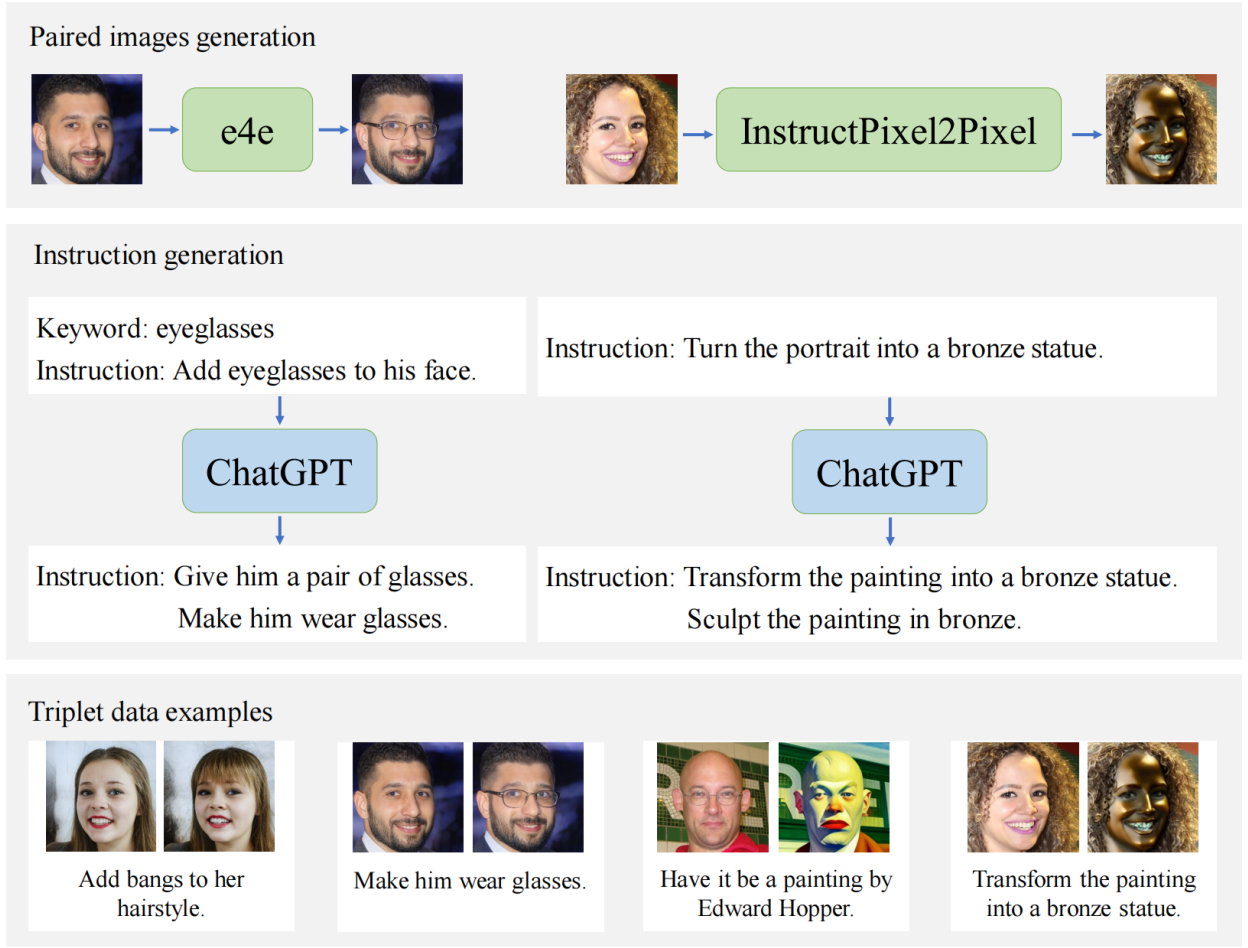

## Diffusion Model

- Adding Noise + denoising
- Conditional Diffusion



# InstructPix2NeRF

## Instruction-following Triplet Data Generation

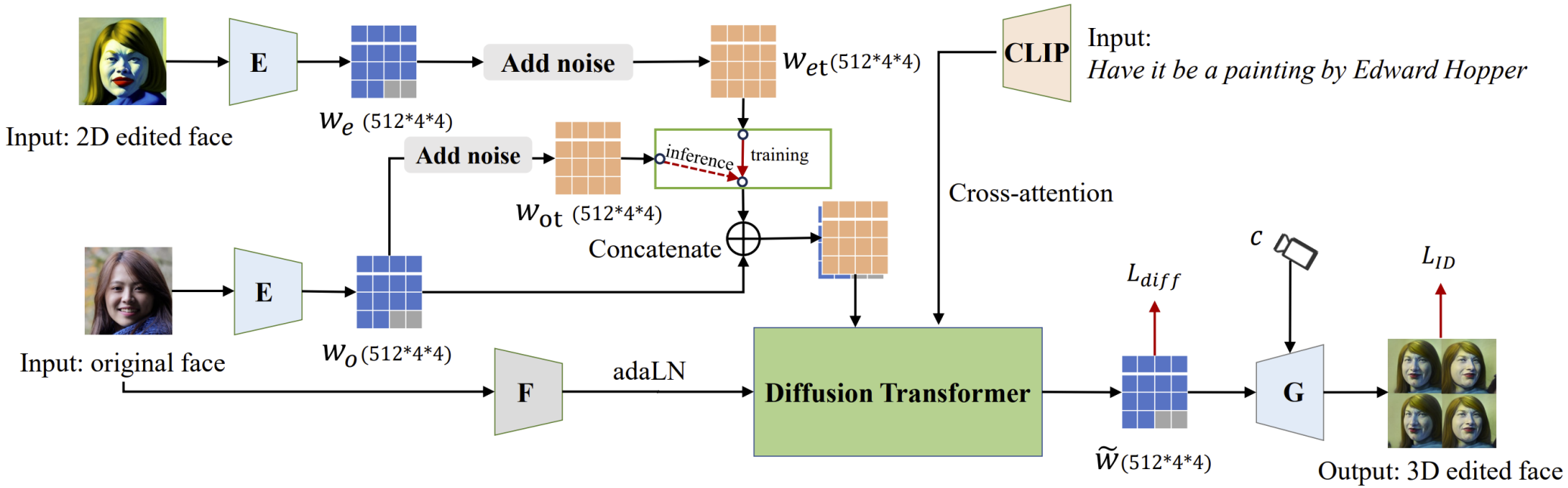


# InstructPix2NeRF

## Pipeline

Generator: $X = G(w, c)$ Encoder: $w = E(X),$ $\hat{X} = G(E(X), c).$

Conditional Diffusion: $\mathcal{L} = \mathbb{E}_{w_e, w_o, c_T, \epsilon \sim \mathcal{N}(0,1), t}[\|\epsilon - \epsilon_\theta(w_{et}, t, w_o, c_T)\|_2^2]$



# InstructPix2NeRF

## Token Position Randomization

*'turn him to a cartoon character'*

tokenization to 77 tokens: *4534 3245 4567 3123 1234 0 0 0 0 0 0 0 0 ……*

position randomization: *0 0 0 0 0 4534 3245 4567 3123 1234 0 0 0 0 0 0 0 0 ……*

## Identity Consistency Module

- Identity compensation module

  A two-layer MLP network.

$$c = t_{embedding} + ID_{embedding}$$

- Identity regularization loss

$$\mathcal{L}_{ID} = 1 - \langle F(X_e), F(G(\tilde{w}_{e0}, c_0)) \rangle$$

# InstructPix2NeRF

## Image and Text Conditioning Sampling

- Training three models

unconditional model: $c_I = \emptyset, \mathrm{c_T} = \emptyset$ with probability $p_1 = 0.05$

image-conditional model: $\mathrm{c_T} = \emptyset$ with probability $p_2 = 0.05$

Text-image-conditional model: $p_3 = 0.9$

- Predicts three score estimates

image-text conditioning: $\epsilon_\theta(w_{ot}, c_I, c_T)$

image conditioning: $\epsilon_\theta(w_{ot}, c_I, \emptyset)$

Unconditioning: $\epsilon_\theta(w_{ot}, \emptyset, \emptyset)$

$$\begin{aligned}\tilde{\epsilon}_\theta(w_{ot}, c_I, c_T) =& \epsilon_\theta(w_{ot}, \emptyset, \emptyset) \\ &+ s_I(\epsilon_\theta(w_{ot}, c_I, \emptyset) - \epsilon_\theta(w_{ot}, \emptyset, \emptyset)) \\ &+ s_T(\epsilon_\theta(w_{ot}, c_I, c_T) - \epsilon_\theta(w_{ot}, c_I, \emptyset))\end{aligned}$$

# Experiments

## Quantitative Comparison

- ID: Identity consistency

  Is it the same person?

- CLIP: directional CLIP similarity

  How well the editing result matches the editing instructions?

- AA: Attribute altering

  the change of the desired attribute.

- AD: Attribute Dependency

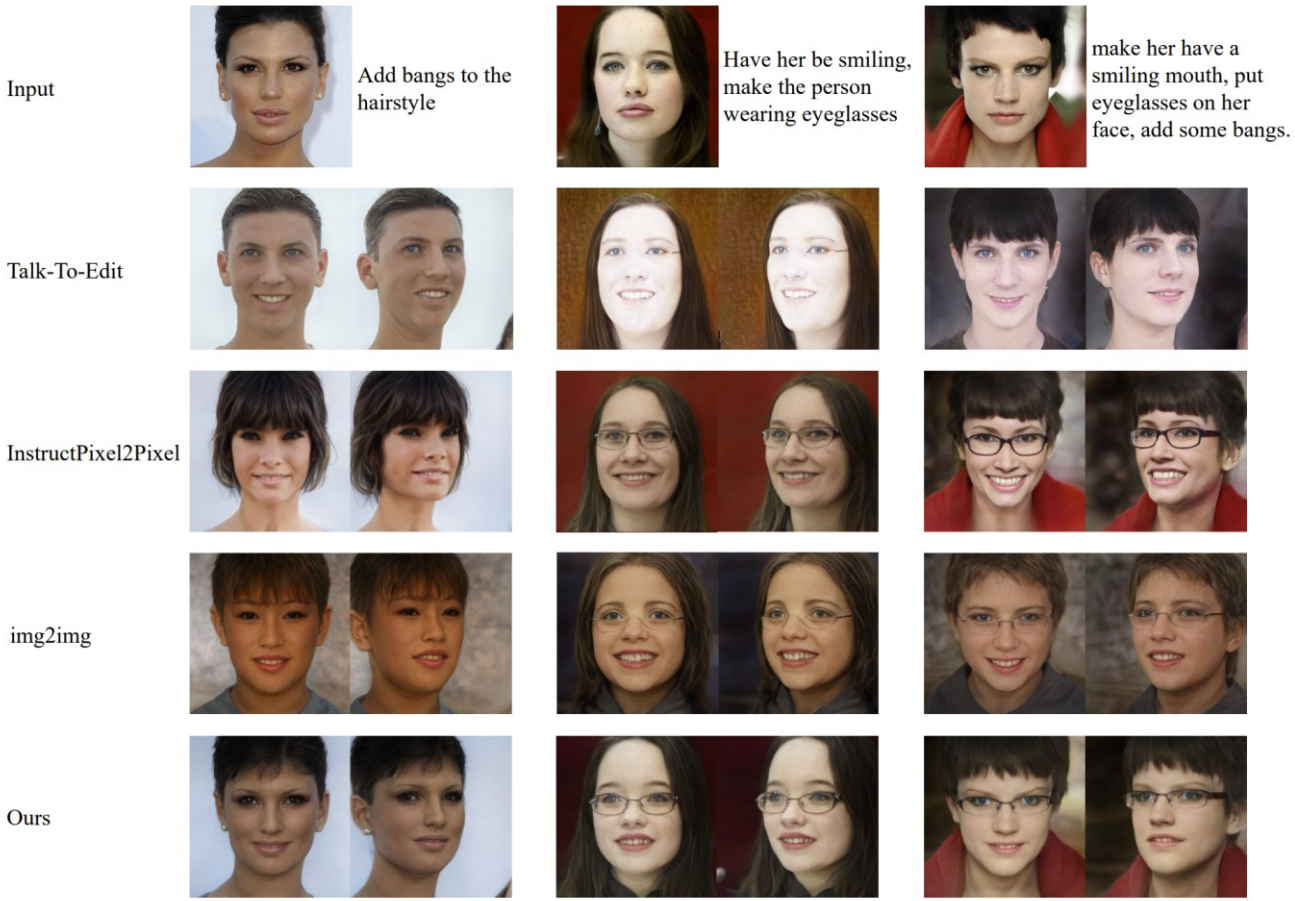
  the change in other attributes when editing certain attributes

| Attribute | Instruction example | Method | ID↑ | CLIP↑ | AA↑ | AD ↓ |
|---|---|---|---|---|---|---|
| Bangs | Let's add some bangs | Talk-To-Edit* | 0.41 | 0.07 | 0.97 | 0.56 |
| | | InstructPix2Pix | 0.40 | 0.07 | 0.80 | 0.64 |
| | | img2img | 0.40 | 0.10 | 0.99 | 0.61 |
| | | Ours | **0.56** | **0.13** | **1.05** | **0.53** |
| Eyeglasses | Make the person wearing glasses | Talk-To-Edit* | 0.46 | 0.04 | 0.69 | 0.67 |
| | | InstructPix2Pix | 0.51 | 0.17 | 3.27 | 0.65 |
| | | img2img | 0.42 | 0.17 | 3.33 | 0.79 |
| | | Ours | **0.59** | **0.20** | **3.37** | **0.64** |
| Smile | The person should smile more happily | Talk-To-Edit* | 0.43 | 0.10 | 0.54 | 0.62 |
| | | InstructPix2Pix | 0.48 | 0.16 | 1.46 | 0.66 |
| | | img2img | 0.46 | 0.15 | 1.47 | 0.76 |
| | | Ours | **0.60** | **0.18** | **1.50** | **0.61** |

| Method | ID↑ | CLIP↑ | $AA_{avg}$ ↑ | $AD_{avg}$ ↓ | $AA_{min}$ ↑ | $M_d$ ↓ | $S_d$ ↓ |
|---|---|---|---|---|---|---|---|
| Talk-To-Edit* | 0.44 | 0.05 | 0.24 | 0.58 | -0.12 | 0.125 | 0.042 |
| InstructPix2Pix | 0.46 | 0.19 | 1.48 | 0.71 | 0.31 | 0.109 | 0.039 |
| img2img | 0.37 | 0.17 | 1.39 | 0.88 | 0.12 | 0.119 | 0.039 |
| Ours(w/o TPR) | 0.50 | 0.18 | 1.50 | 0.73 | 0.08 | 0.114 | **0.038** |
| Ours | **0.55** | **0.20** | **1.53** | **0.69** | **0.52** | **0.105** | **0.038** |

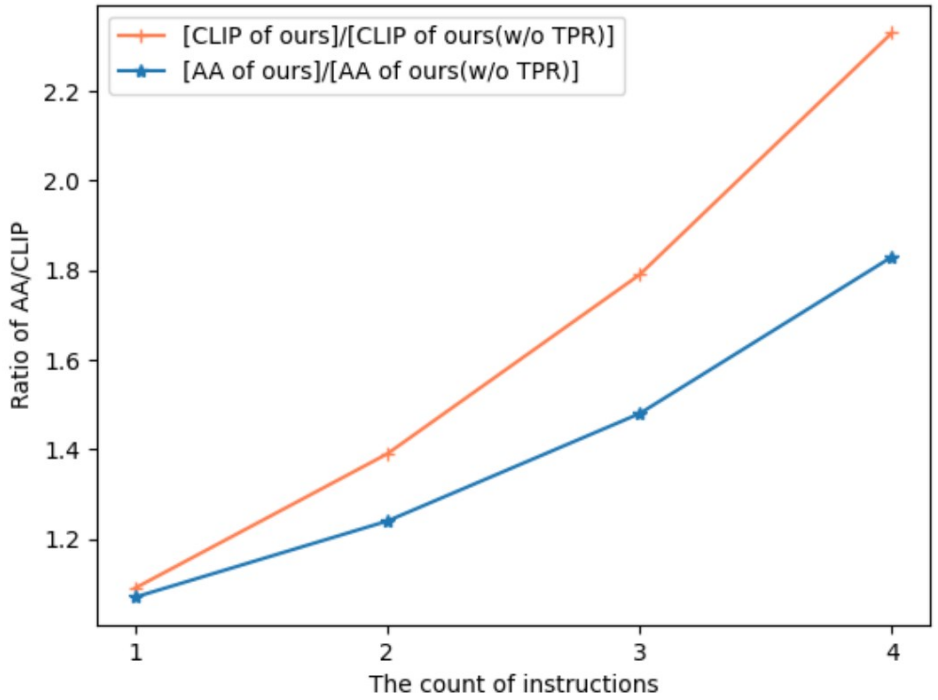
# Experiments

## Qualitative comparison



# Experiments

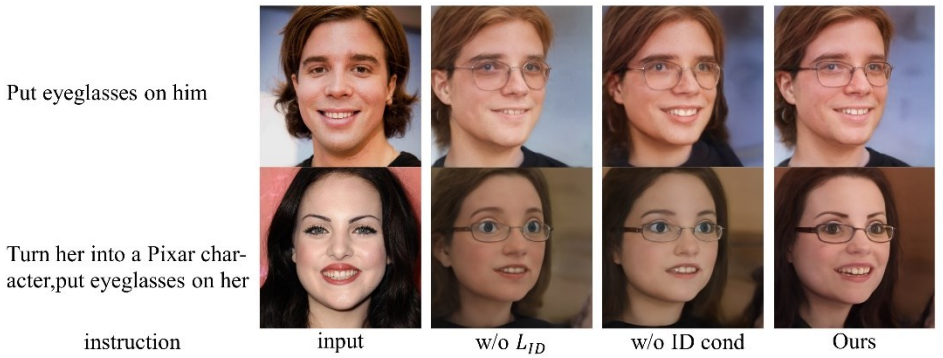
## Ablation study (TPR)

# Experiments

## Ablation study (ID module)

| Config | $\text{ID}_{bang}$ | $\text{ID}_{eyeglasses}$ | $\text{ID}_{smile}$ | $\text{ID}_{multi}$ |
|---|---|---|---|---|
| w/o $\mathcal{L}_{ID}$ | 0.47 | 0.52 | 0.55 | 0.44 |
| w/o ID cond | 0.54 | 0.55 | 0.57 | 0.50 |
| Ours | **0.56** | **0.59** | **0.60** | **0.55** |

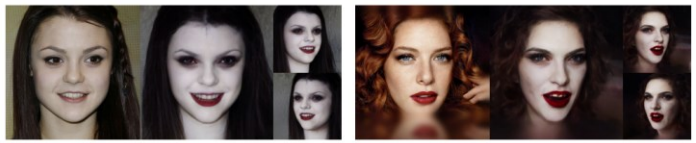



## User study

| Method | bang | eyeglasses | smile | multiple instructions |
|---|---|---|---|---|
| Talk-to-Edit | 0.742 | 0.958 | 0.817 | 0.875 |
| InstructPix2Pix | 0.833 | 0.667 | 0.725 | 0.683 |
| img2img | 0.733 | 0.758 | 0.750 | 0.783 |

# Experiments

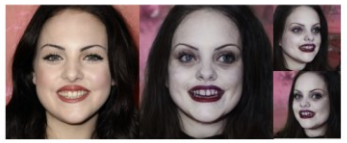
## Results



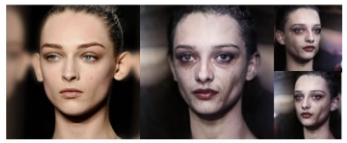
Turn her into a vampire



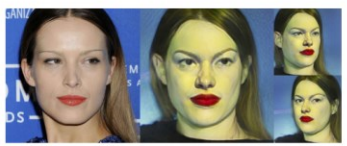
Turn her into a vampire



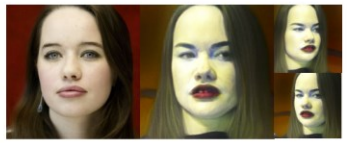
Give it a zombie makeover



Give it a zombie makeover

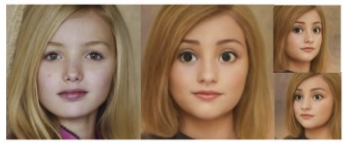


Make it look like a sketch by Edward Hopper



Make it look like a sketch by Edward Hopper

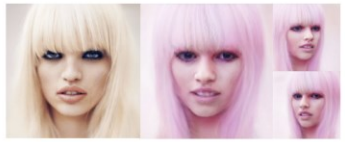
Make him a cartoon character



Make her a cartoon character

Turn the painting into a bronze statue

Turn the painting into a bronze statue



Turn the hair color to pink

Turn the hair color to pink

Turn the hair color to pink, put eyeglasses on her

Turn the hair color to pink, put eyeglasses on her

Give the portrait a comic book look

Give the portrait a comic book look

Put eyeglasses on his face, give him a goatee

Put eyeglasses on his face, give him a goatee

Remove the beard

Remove the beard

# Experiments

## Results


Make her eyes appear narrower


Make her eyes appear narrower


Put eyeglasses on her


Put eyeglasses on her


Style her hair with a pink wig


Style her hair with a pink wig


Add some beard

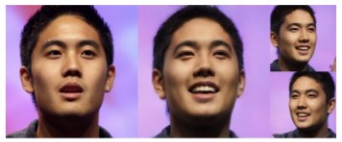

Add some beard


Shave the beard off


Shave the beard off


Give her thicker, bushier eyebrows

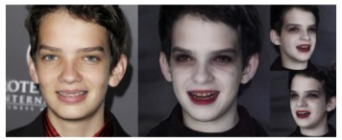

Add bangs to the hairstyle

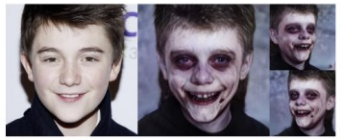

Have his wear a smile


Add a smile to her face


Turn the portrait into a bronze statue


Turn the portrait into a bronze statue


Turn him into a vampire


Give it a zombie makeover

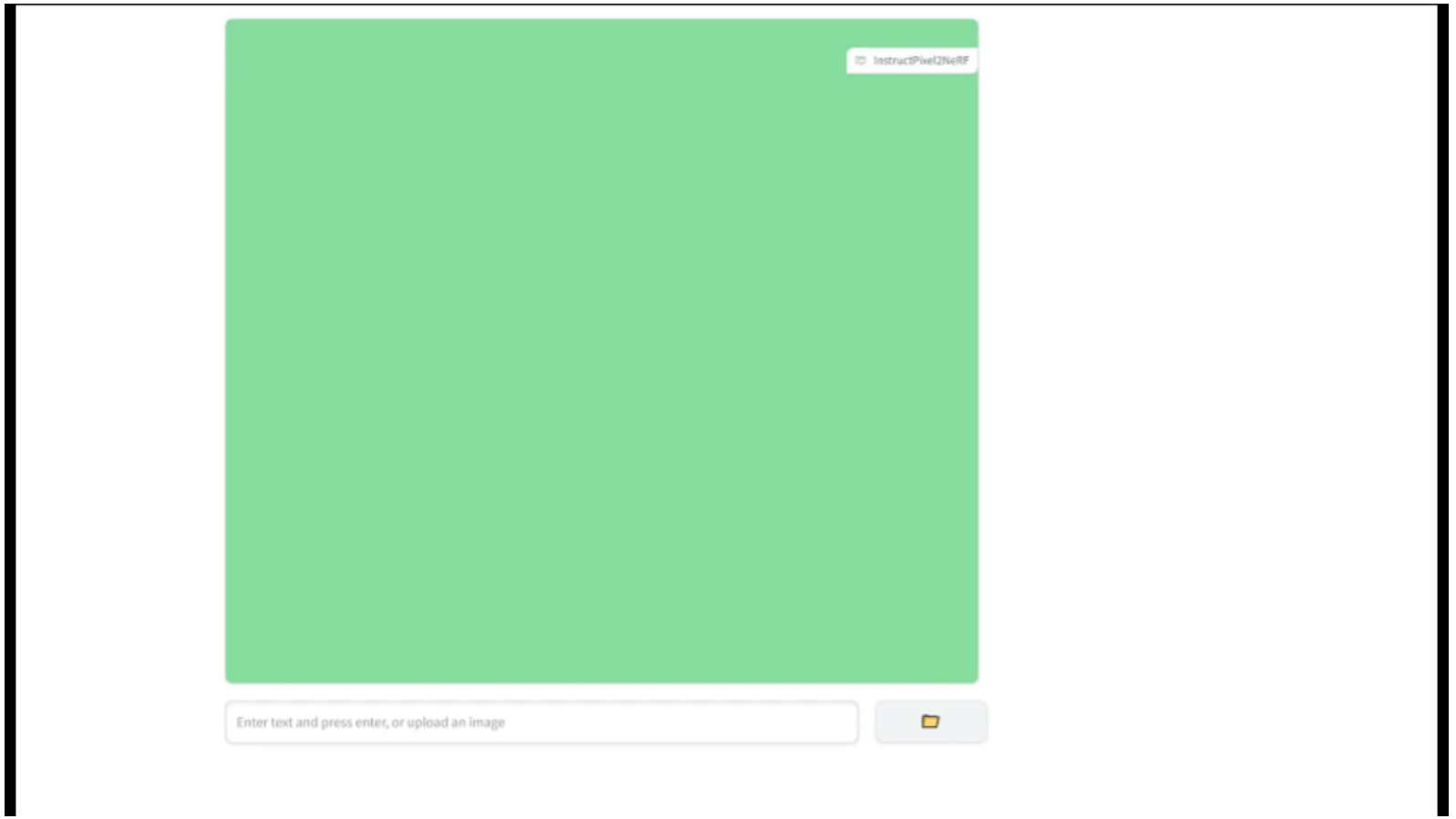

Make her look like a cartoon


Give the portrait a comic book look

# WebDemo

Tsinghua University

## Text-guided Interactive Editing



清華大學
Tsinghua University

# Thank you