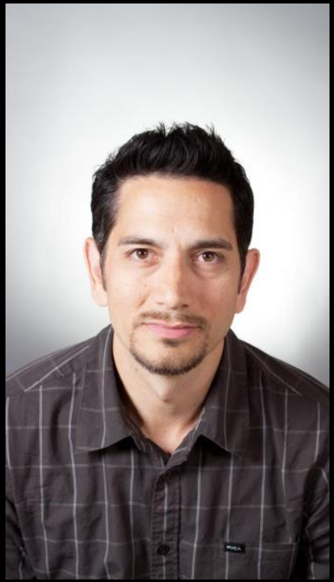
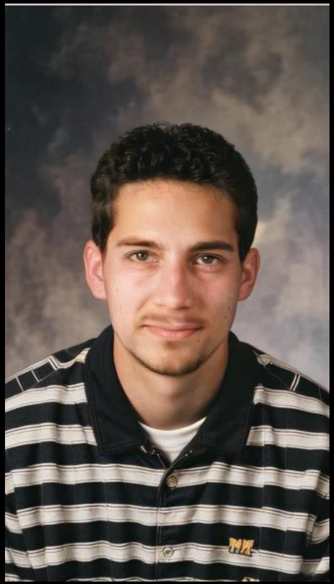
InstantPortrait: One-Step Portrait Editing via Diffusion Multi-Objective Distillation

Zhixin Lai [1*] Keqiang Sun [2] FuYun Wang [2] Dhritiman Sagar [1] Erli Ding [1*]

[1] Snap Inc [2] The Chinese University of Hong Kong
[*] Equal contribution. Work done while at Snap Inc



Input



Output

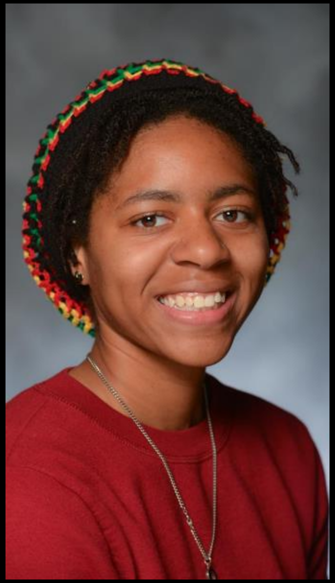
Editing Prompt:

[illegible]

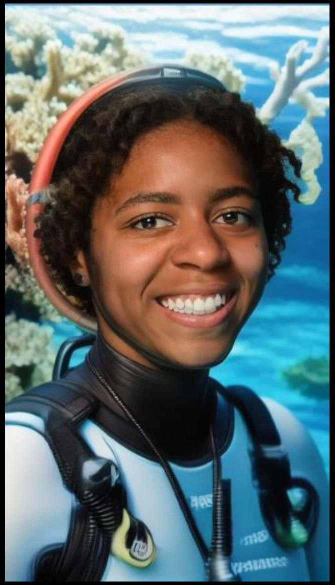
[illegible]

[illegible]

[illegible] texture, uneven tones, muted tones, low detail



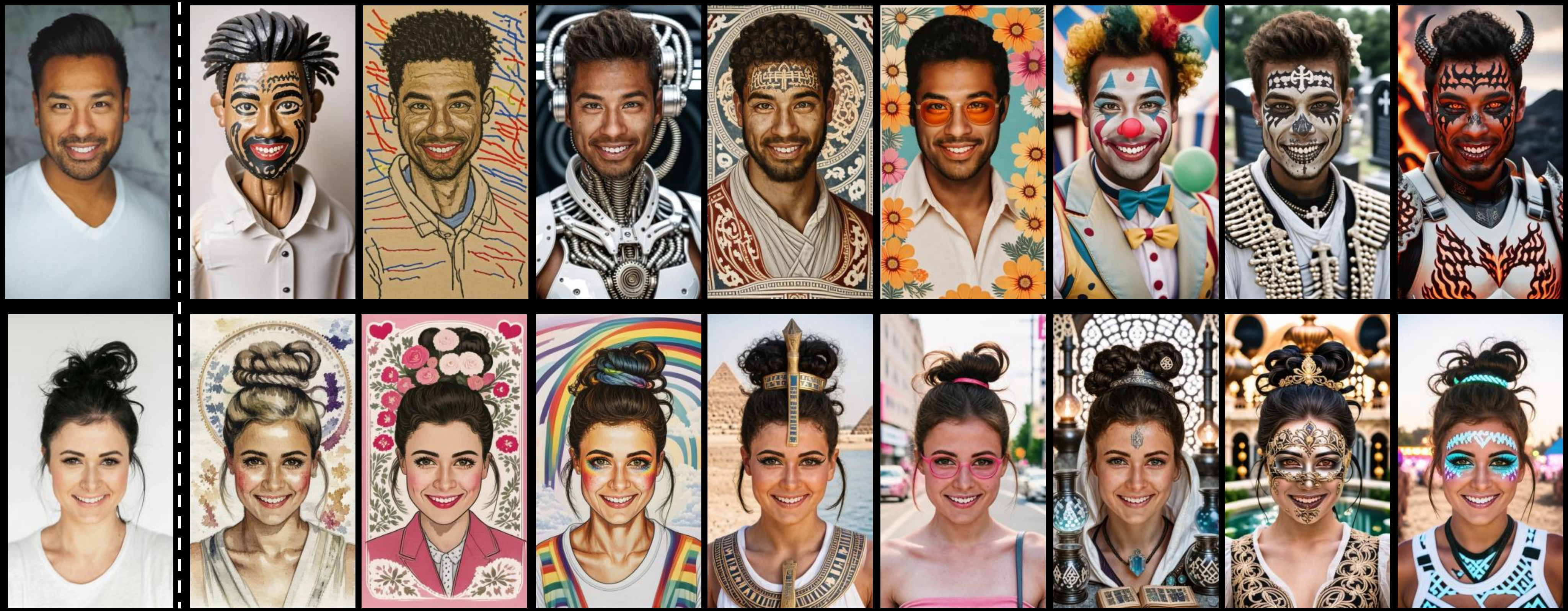
Input



Output

Editing Prompt: [illegible]

Output of Instant-Portrait Trained with Diffusion Multi-Objective Distillation



Input

Output images conditioned by the input images and editing instructions

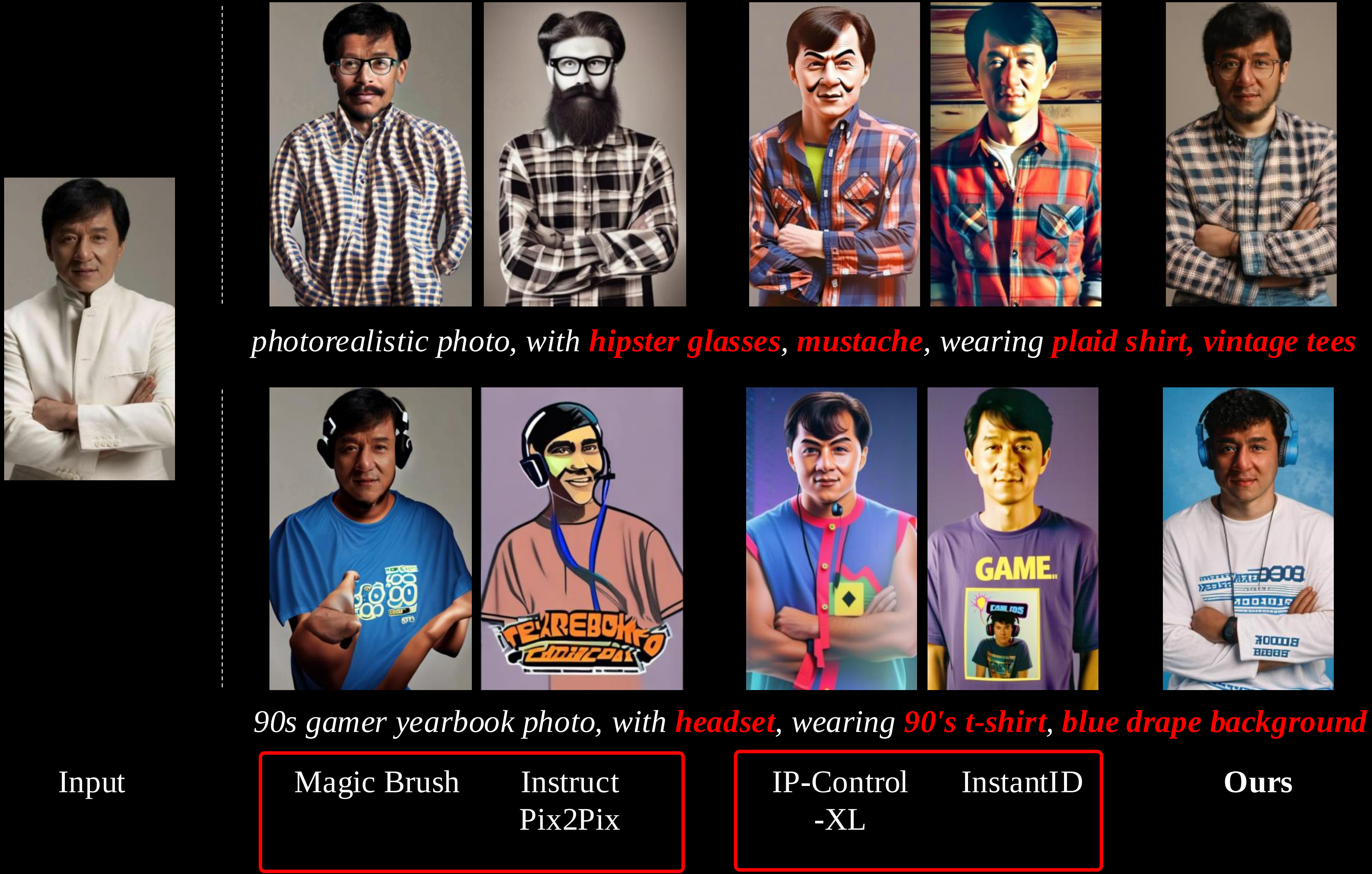
Motivation

Challenges

Identity Preservation

Fidelity to Instruction

Fast model inference



photorealistic photo, with ***hipster glasses****,* ***mustache****, wearing* ***plaid shirt, vintage tees***

90s gamer yearbook photo, with ***headset****, wearing* ***90's t-shirt****,* ***blue drape background***

Input

Magic Brush

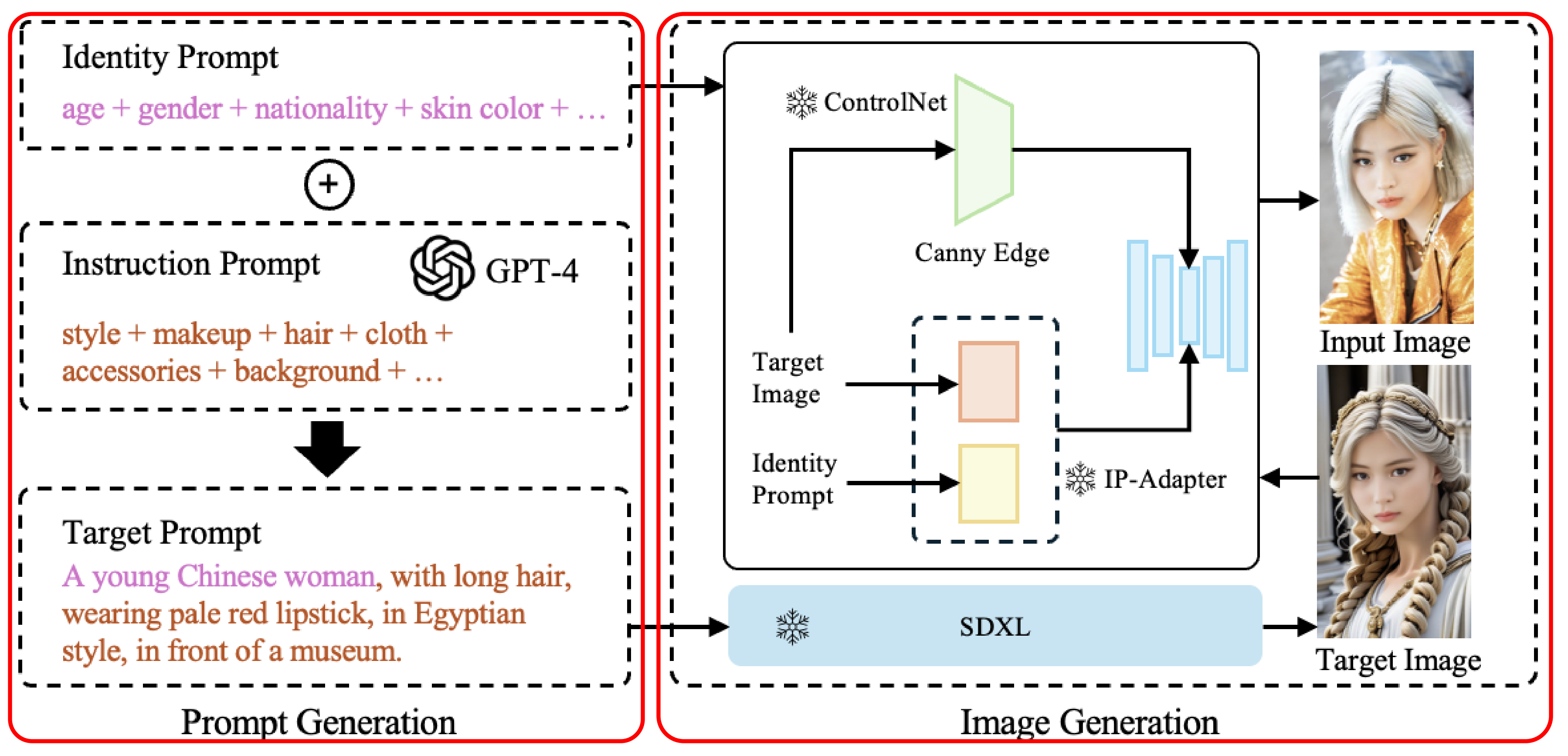
Instruct Pix2Pix

IP-Control -XL

InstantID

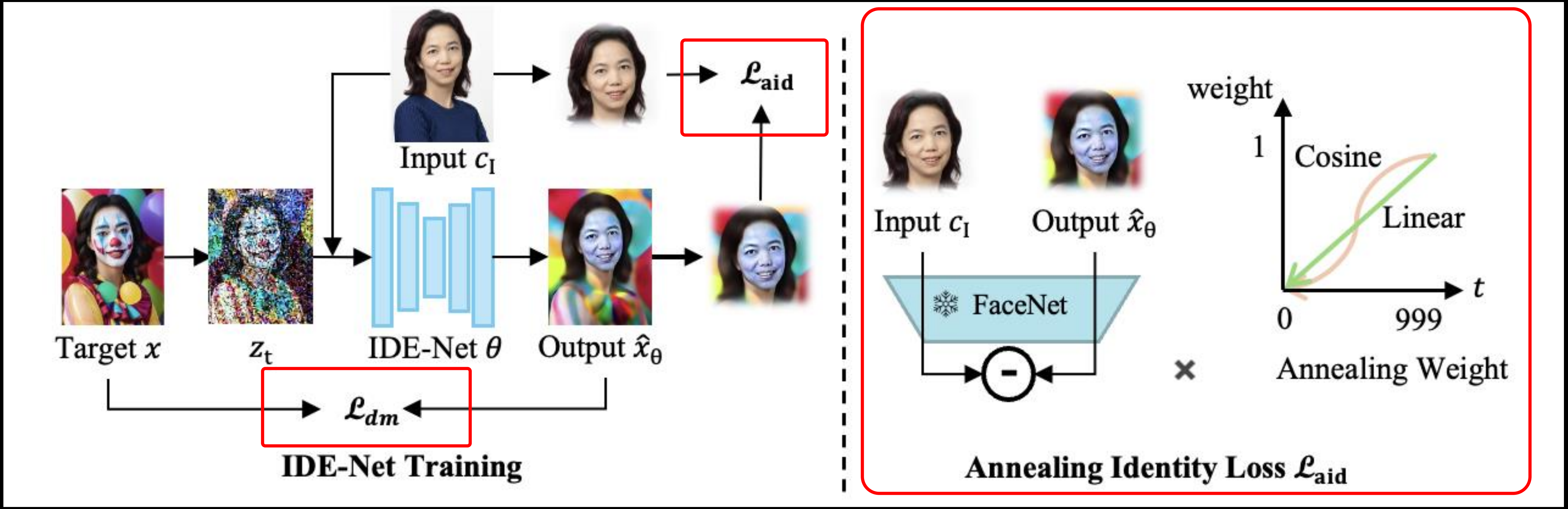
Ours

Dataset

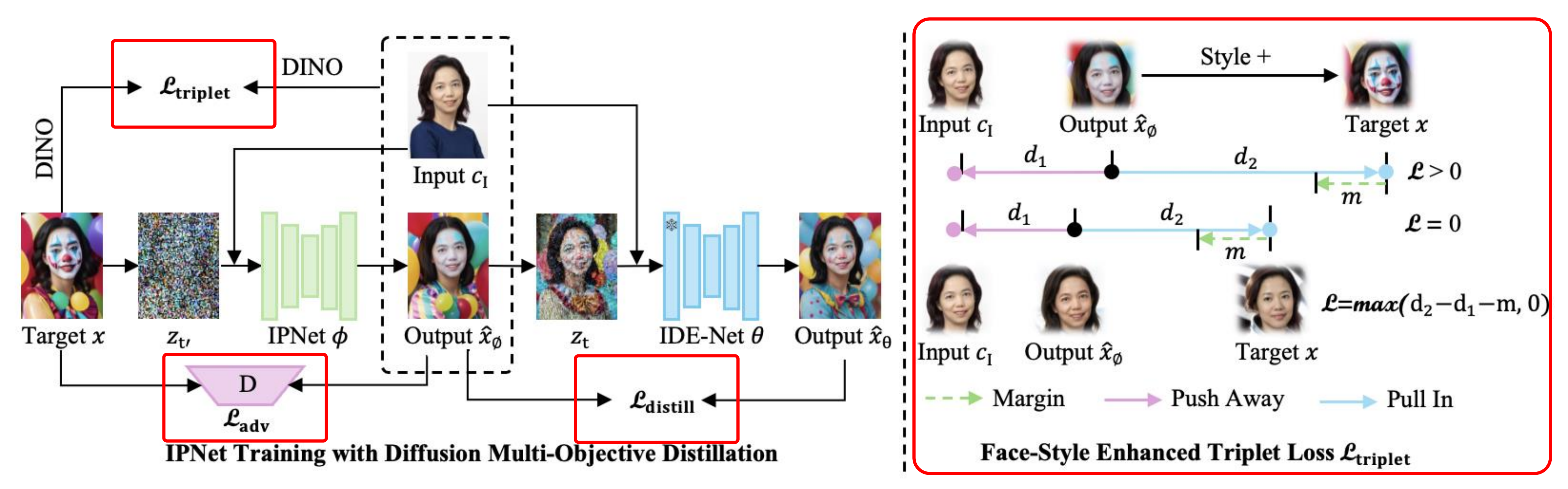

Identity Prompt
age + gender + nationality + skin color + …
Instruction Prompt
GPT-4
style + makeup + hair + cloth +
accessories + background + …
Target Prompt
A young Chinese woman, with long hair,
wearing pale red lipstick, in Egyptian
style, in front of a museum.
Prompt Generation
ControlNet
Canny Edge
Target Image
Identity Prompt
IP-Adapter
SDXL
Input Image
Target Image
Image Generation

Method

IDE-Net Training

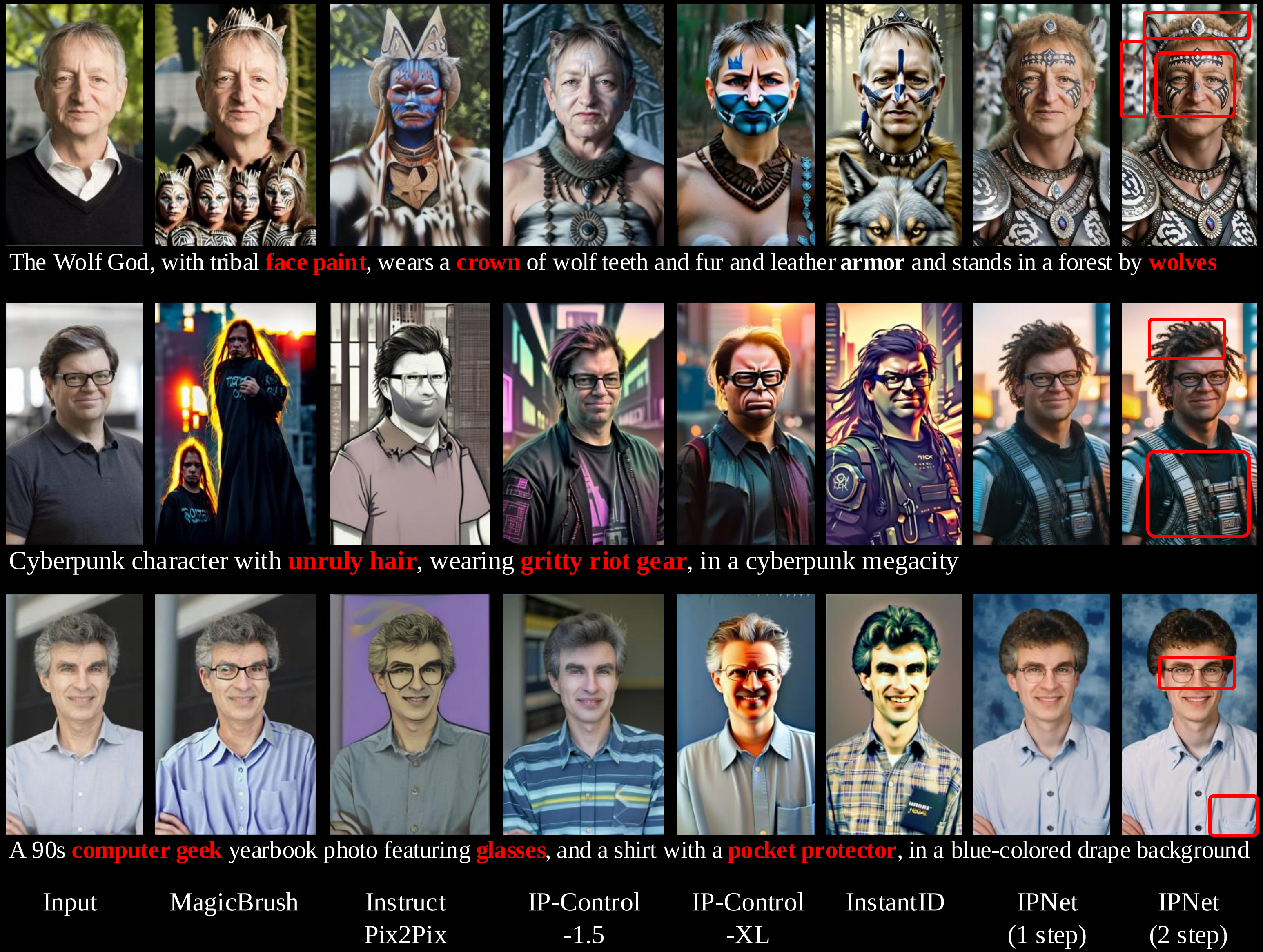


IPNet Training



Experiments

vs SOTA

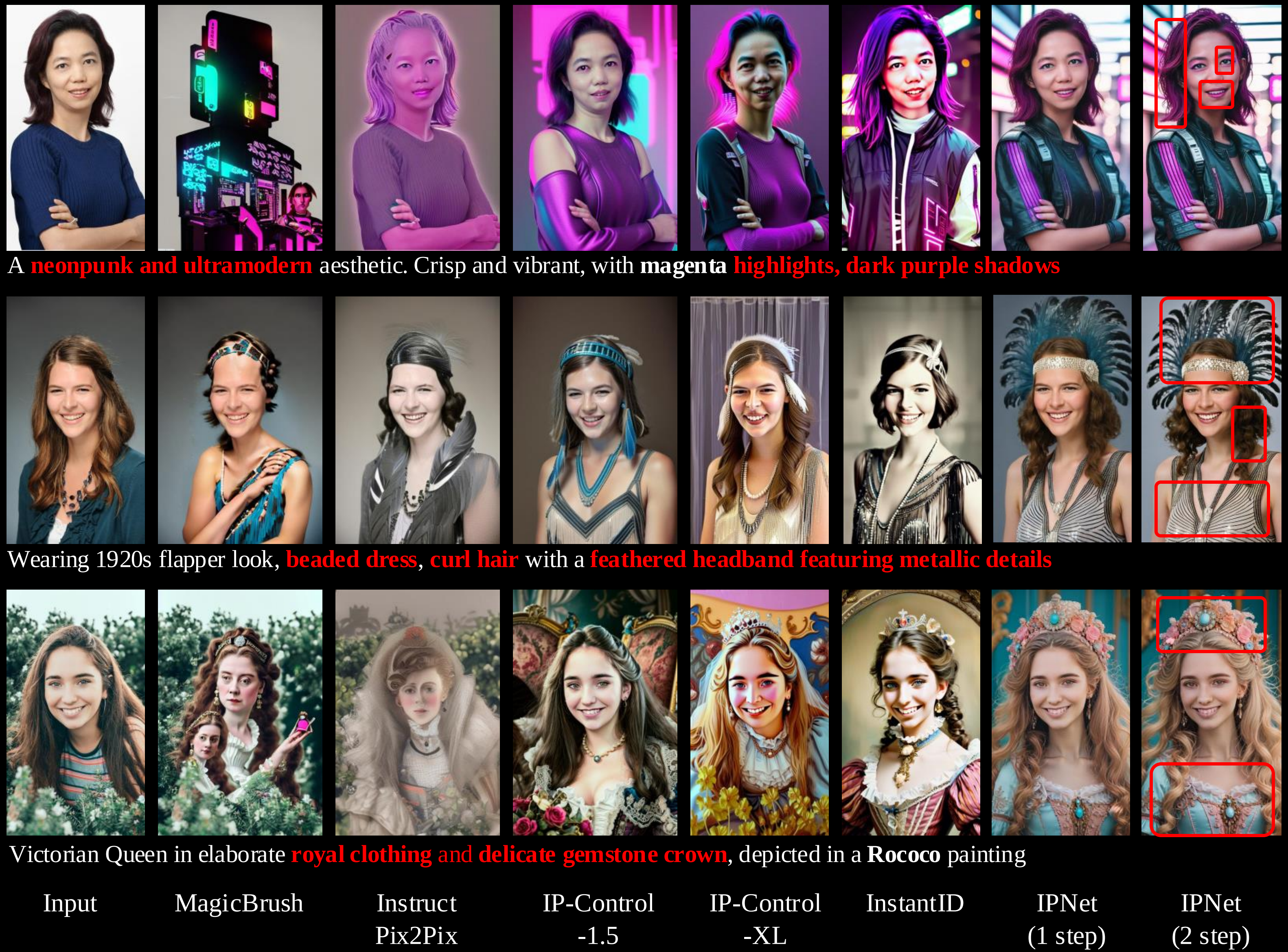


The Wolf God, with tribal **face paint**, wears a **crown** of wolf teeth and fur and leather **armor** and stands in a forest by **wolves**

Cyberpunk character with **unruly hair**, wearing **gritty riot gear**, in a cyberpunk megacity

A 90s **computer geek** yearbook photo featuring **glasses**, and a shirt with a **pocket protector**, in a blue-colored drape background

vs SOTA

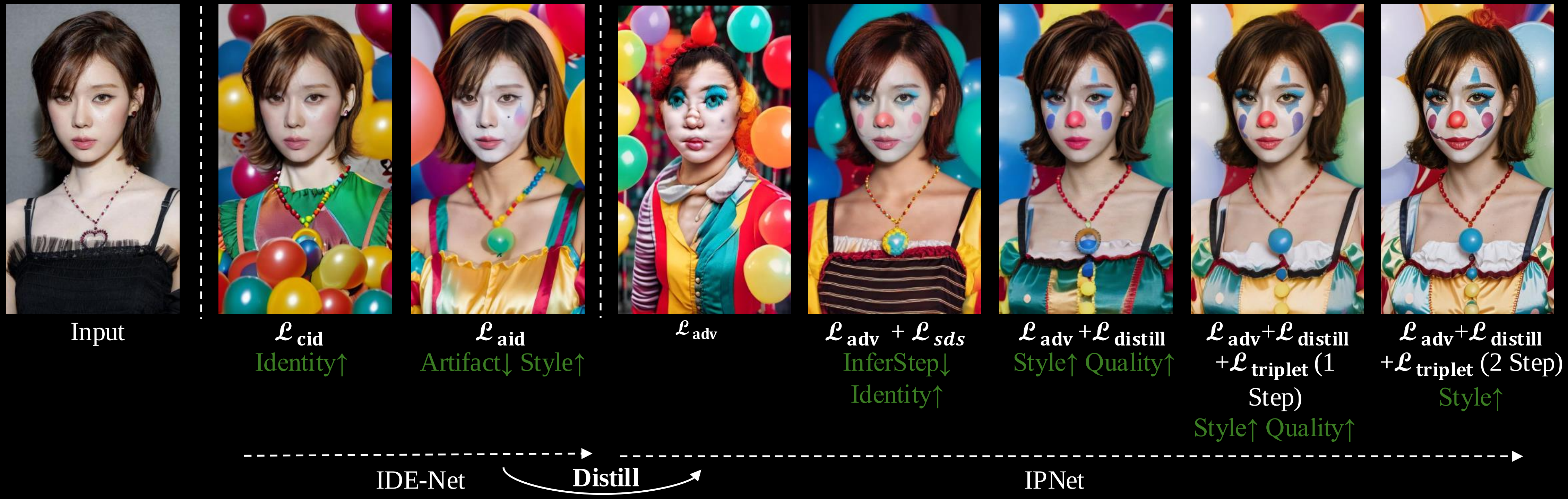


A **neonpunk and ultramodern** aesthetic. Crisp and vibrant, with **magenta** **highlights, dark purple shadows**

Wearing 1920s flapper look, **beaded dress**, **curl hair** with a **feathered headband featuring metallic details**

Victorian Queen in elaborate **royal clothing** and **delicate gemstone crown**, depicted in a **Rococo** painting

Ablation: progressive improvement over model training and distillation



Input

$\mathcal{L}_{\mathbf{cid}}$

Identity↑

$\mathcal{L}_{\mathbf{aid}}$

Artifact↓ Style↑

$\mathcal{L}_{\mathbf{adv}}$

$\mathcal{L}_{\mathbf{adv}} + \mathcal{L}_{sds}$

InferStep↓ Identity↑

$\mathcal{L}_{\mathbf{adv}}+\mathcal{L}_{\mathbf{distill}}$

Style↑ Quality↑

$\mathcal{L}_{\mathbf{adv}}+\mathcal{L}_{\mathbf{distill}}+\mathcal{L}_{\mathbf{triplet}}$ (1 Step)

Style↑ Quality↑

$\mathcal{L}_{\mathbf{adv}}+\mathcal{L}_{\mathbf{distill}}+\mathcal{L}_{\mathbf{triplet}}$ (2 Step)

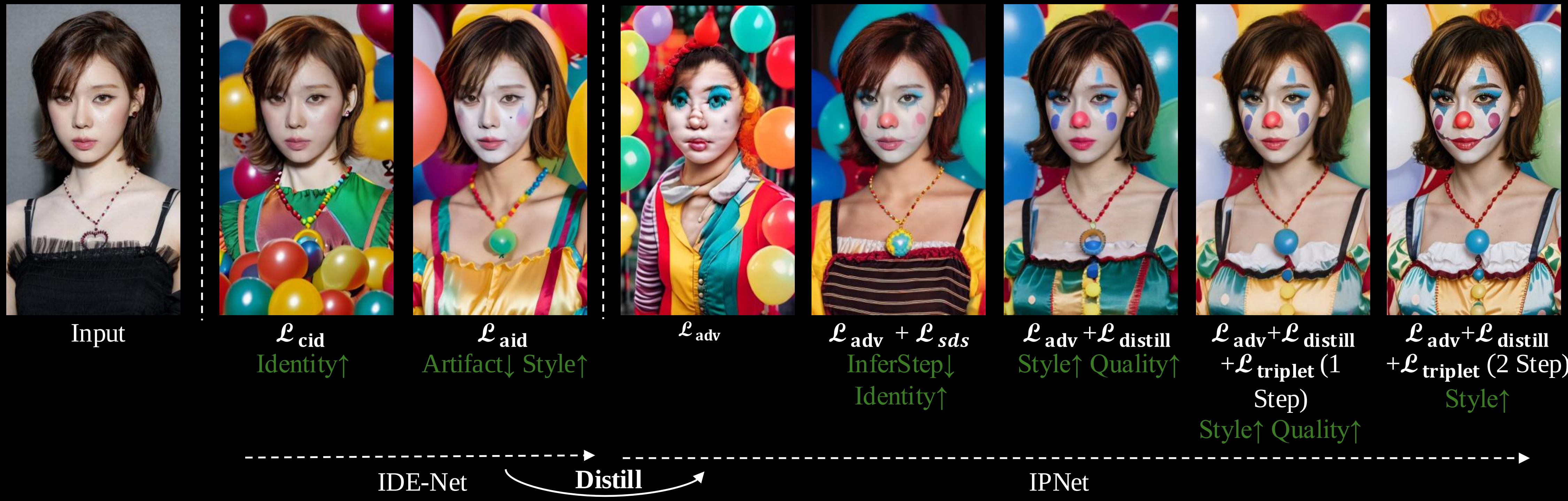
Style↑

IDE-Net

Distill

IPNet

$\mathcal{L}_{\text{aid}}$ Ablation



Input

$\mathcal{L}_{\text{cid}}$
Identity↑

$\mathcal{L}_{\text{aid}}$
Artifact↓ Style↑

$\mathcal{L}_{\text{adv}}$

$\mathcal{L}_{\text{adv}} + \mathcal{L}_{sds}$
InferStep↓
Identity↑

$\mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{distill}}$
Style↑ Quality↑

$\mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{distill}} + \mathcal{L}_{\text{triplet}}$ (1 Step)
Style↑ Quality↑

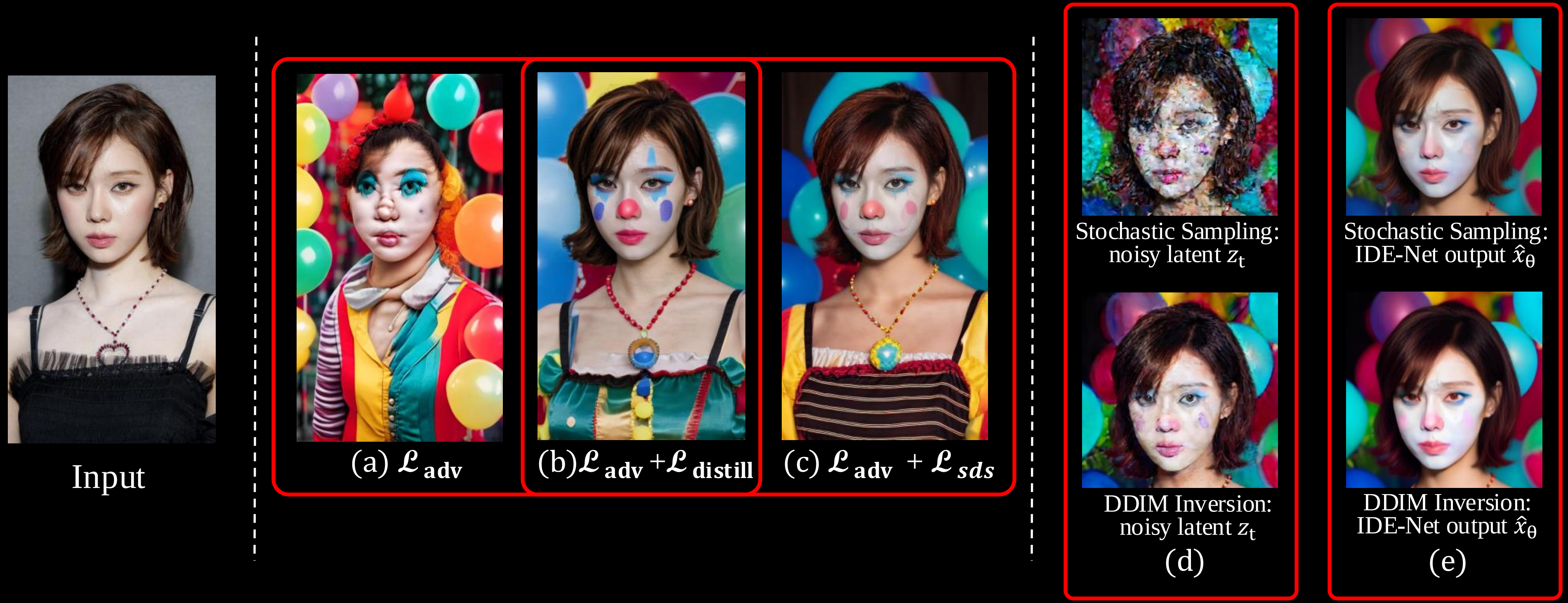
$\mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{distill}} + \mathcal{L}_{\text{triplet}}$ (2 Step)
Style↑

IDE-Net

Distill

IPNet

$\mathcal{L}_{\text{distill}}$ Ablation: DDIM vs Stochastic Sampling



Input

(a) $\mathcal{L}_{\text{adv}}$

(b)$\mathcal{L}_{\text{adv}}+\mathcal{L}_{\text{distill}}$

(c) $\mathcal{L}_{\text{adv}} + \mathcal{L}_{sds}$

Stochastic Sampling: noisy latent z_t

DDIM Inversion: noisy latent z_t

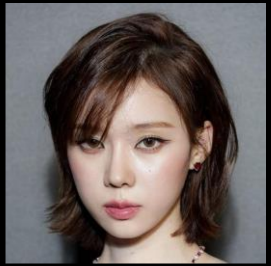
(d)

Stochastic Sampling: IDE-Net output $\hat{x}_\theta$

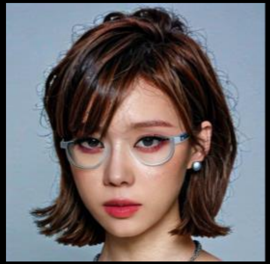
DDIM Inversion: IDE-Net output $\hat{x}_\theta$

(e)

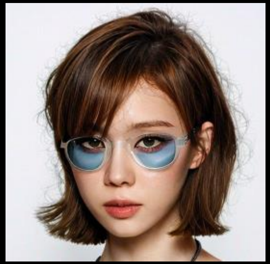
$\mathcal{L}_{\textbf{triplet}}$ Ablation



Input

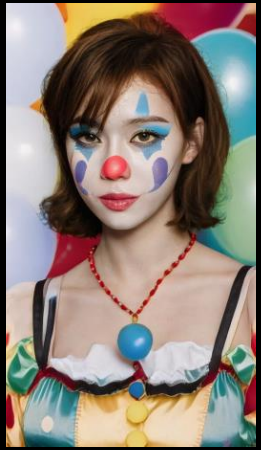


w/o $\mathcal{L}_{\textbf{triplet}}$

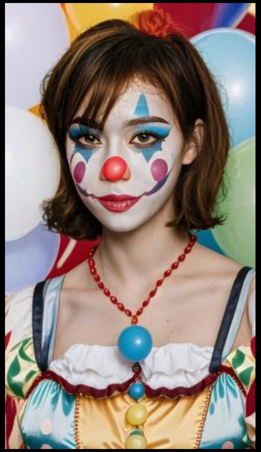


with $\mathcal{L}_{\textbf{triplet}}$

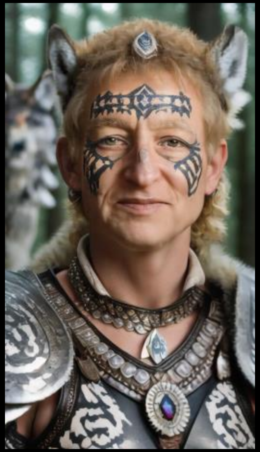
Ablation: Style Boost via Iterative Inference



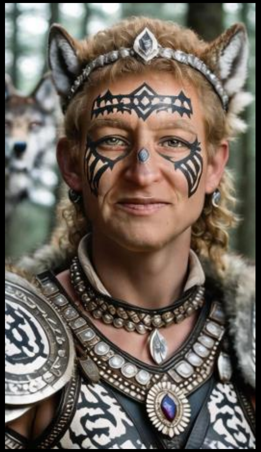
1 Step



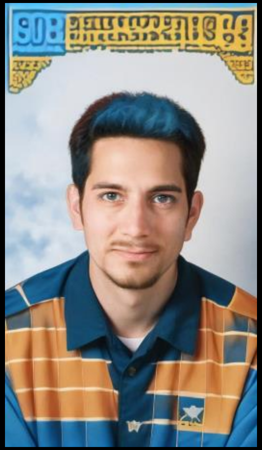
2 Step



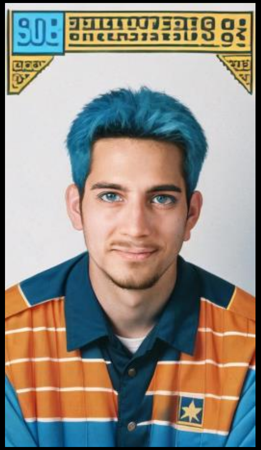
1 Step



2 Step

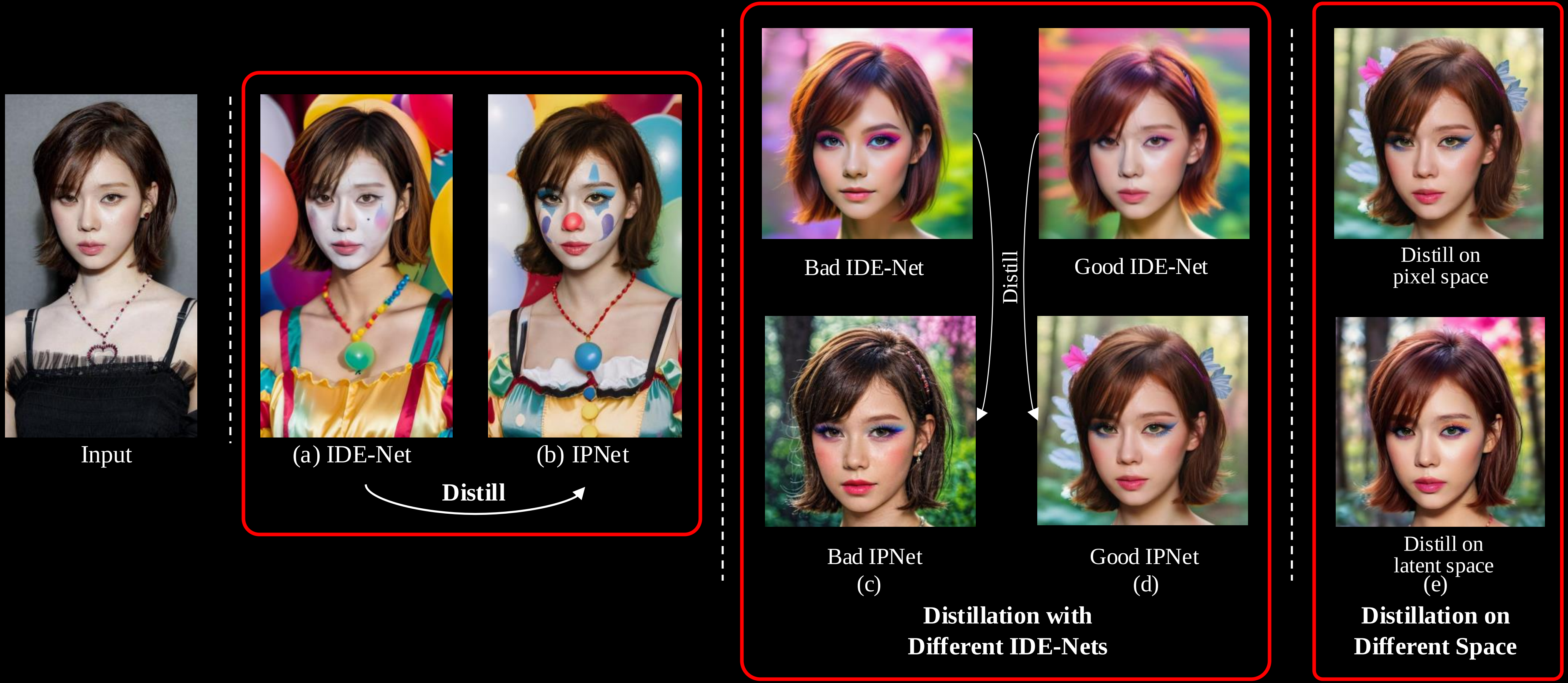


1 Step



2 Step

Ablation: Diffusion Multi-Object Distillation



Input

(a) IDE-Net

(b) IPNet

Distill

Bad IDE-Net

Good IDE-Net

Distill

Bad IPNet
(c)

Good IPNet
(d)

Distillation with Different IDE-Nets

Distill on pixel space

Distill on latent space
(e)

Distillation on Different Space

Thank you!