# BooW-VTON: Boosting In-the-Wild Virtual Try-On via Mask-Free Pseudo Data Training

**arXiv, 2024**

5 citations

Tianjin University, Alibaba Group

# INTRODUCTION

## Virtual Try-On

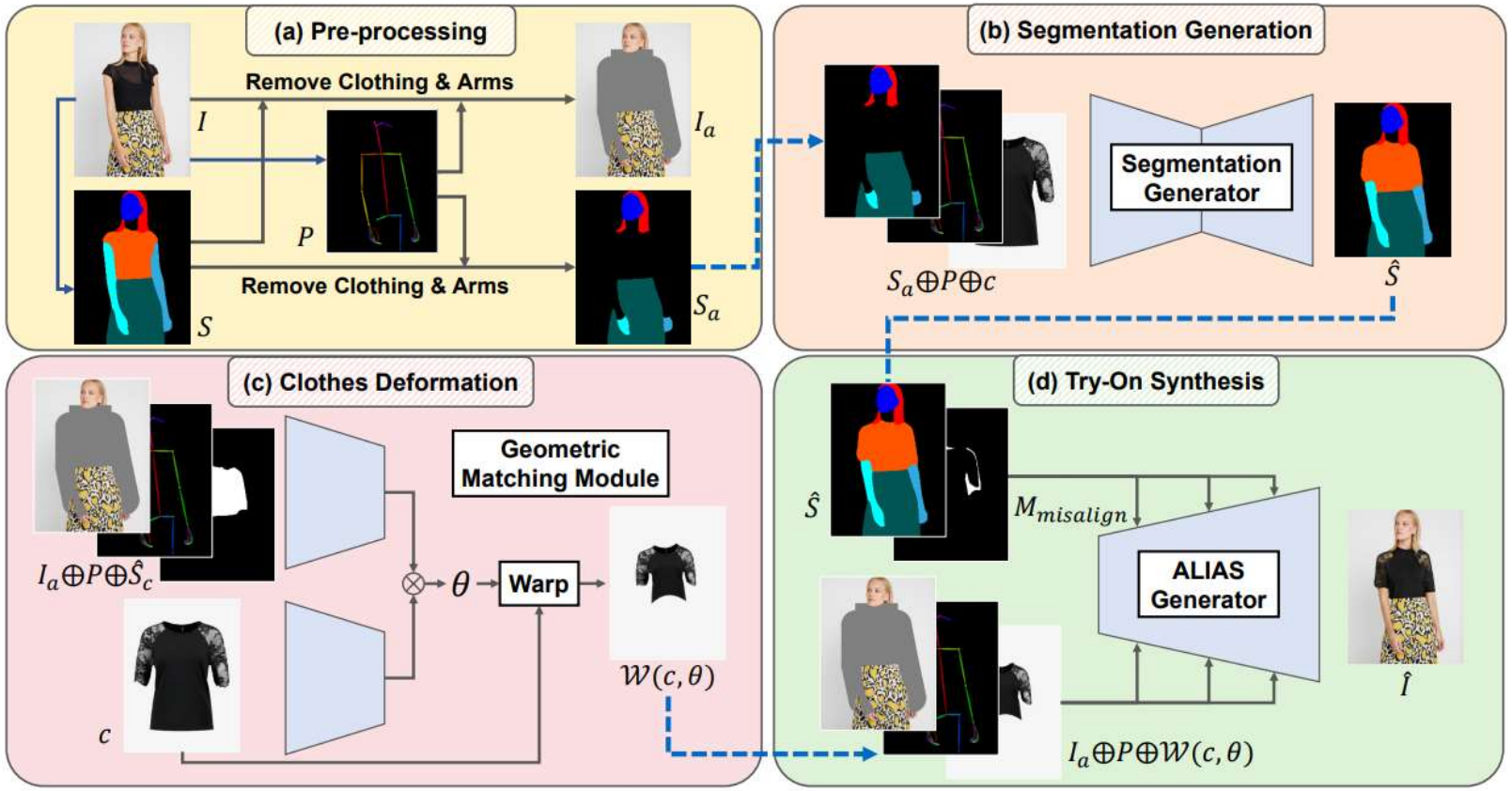
- Aims to generate realistic try-on images of a specific person while preserving the original pose and body feature in source images
- Issue
  - Render garments onto correct human body parts while preserving non try-on contents
  - E.g., body structure, external objects

# INTRODUCTION

## Previous works

- Extract human pose parser to maintain a structure of output person
- Utilize an independent image warping modules to deform garment images

VITON-HD:

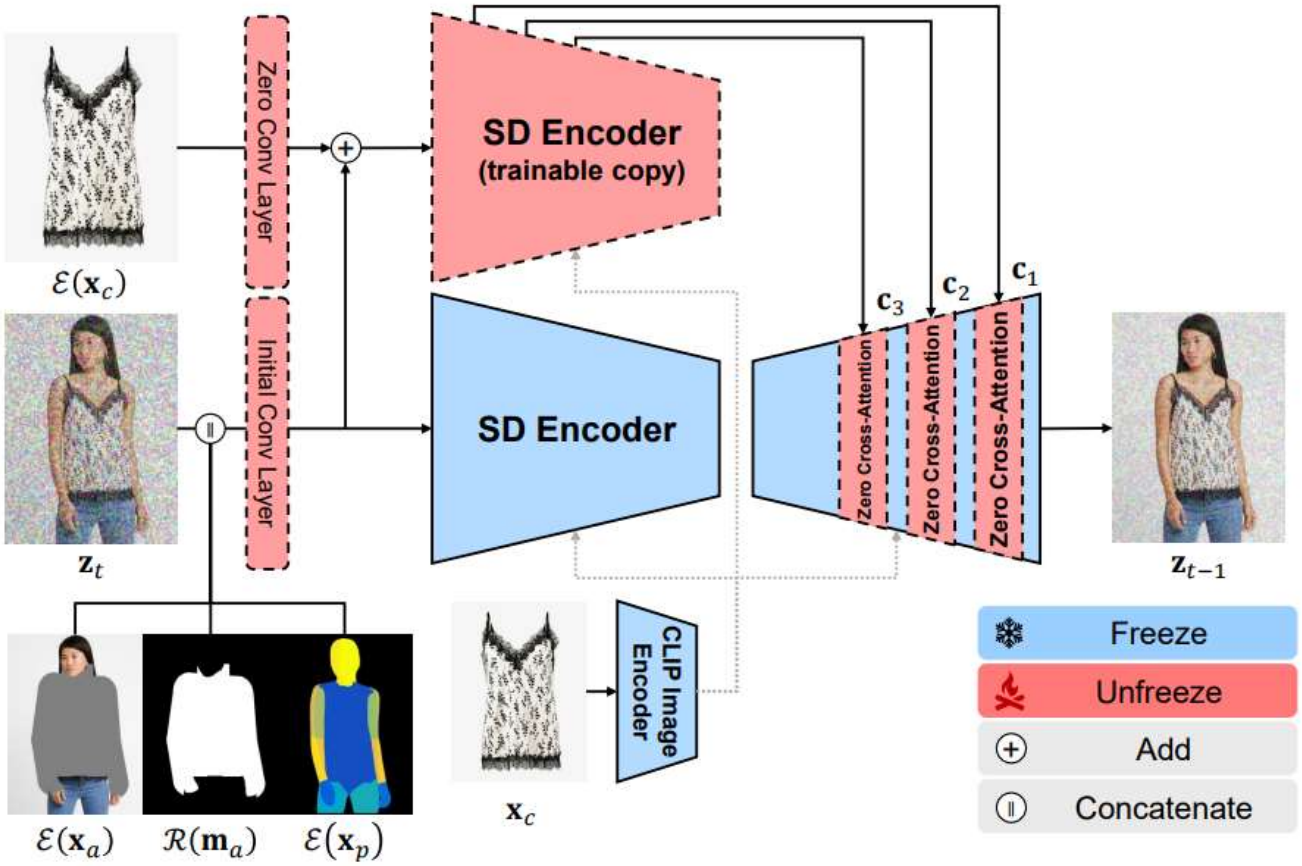


# INTRODUCTION

## Current works

- Encode garment images and human pose using pre-trained encoders
- Utilize the powerful diffusion models

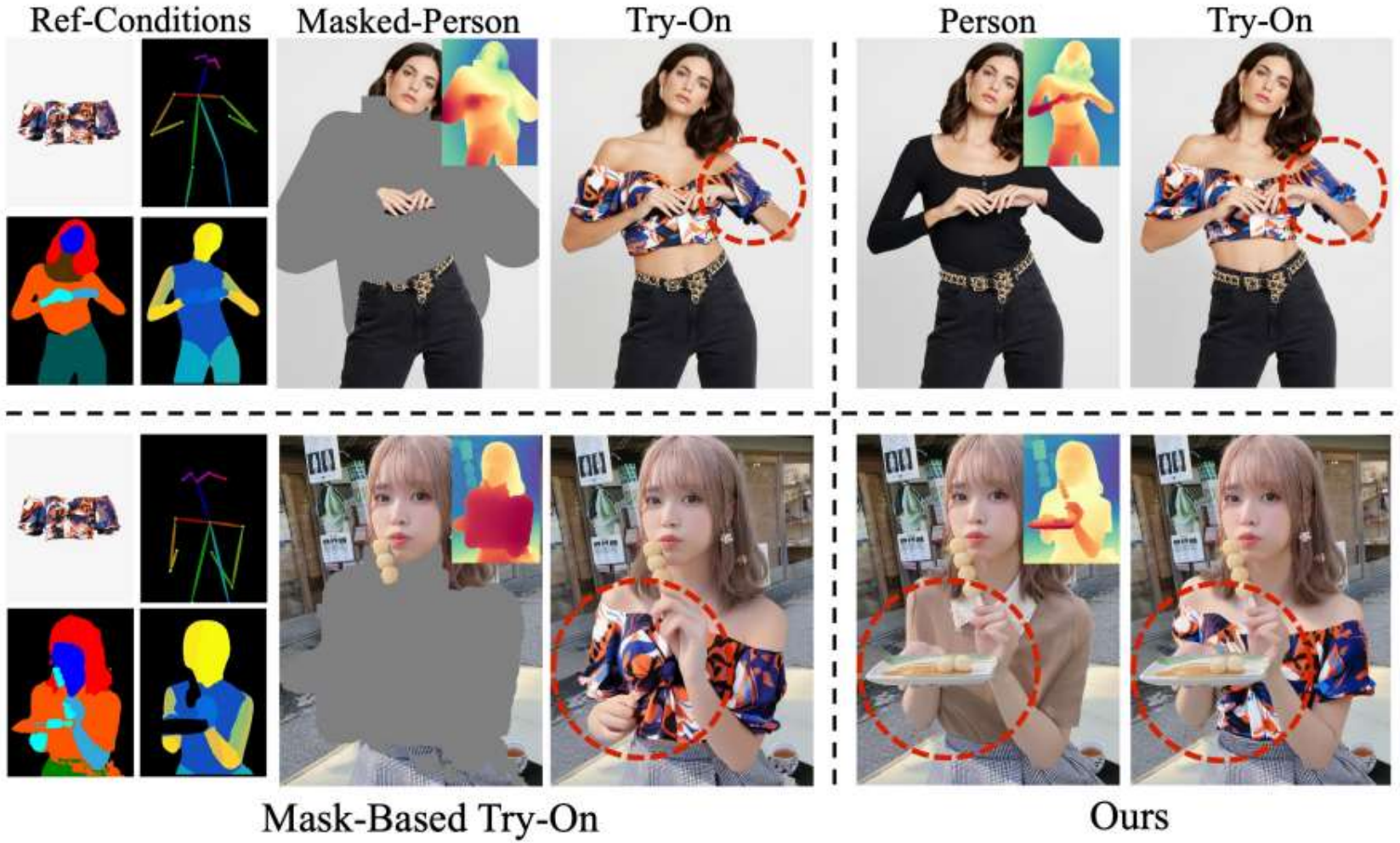
StableVITON:



# INTRODUCTION

## Problem

- Existing methods require masking the person image → significant loss of spatial information

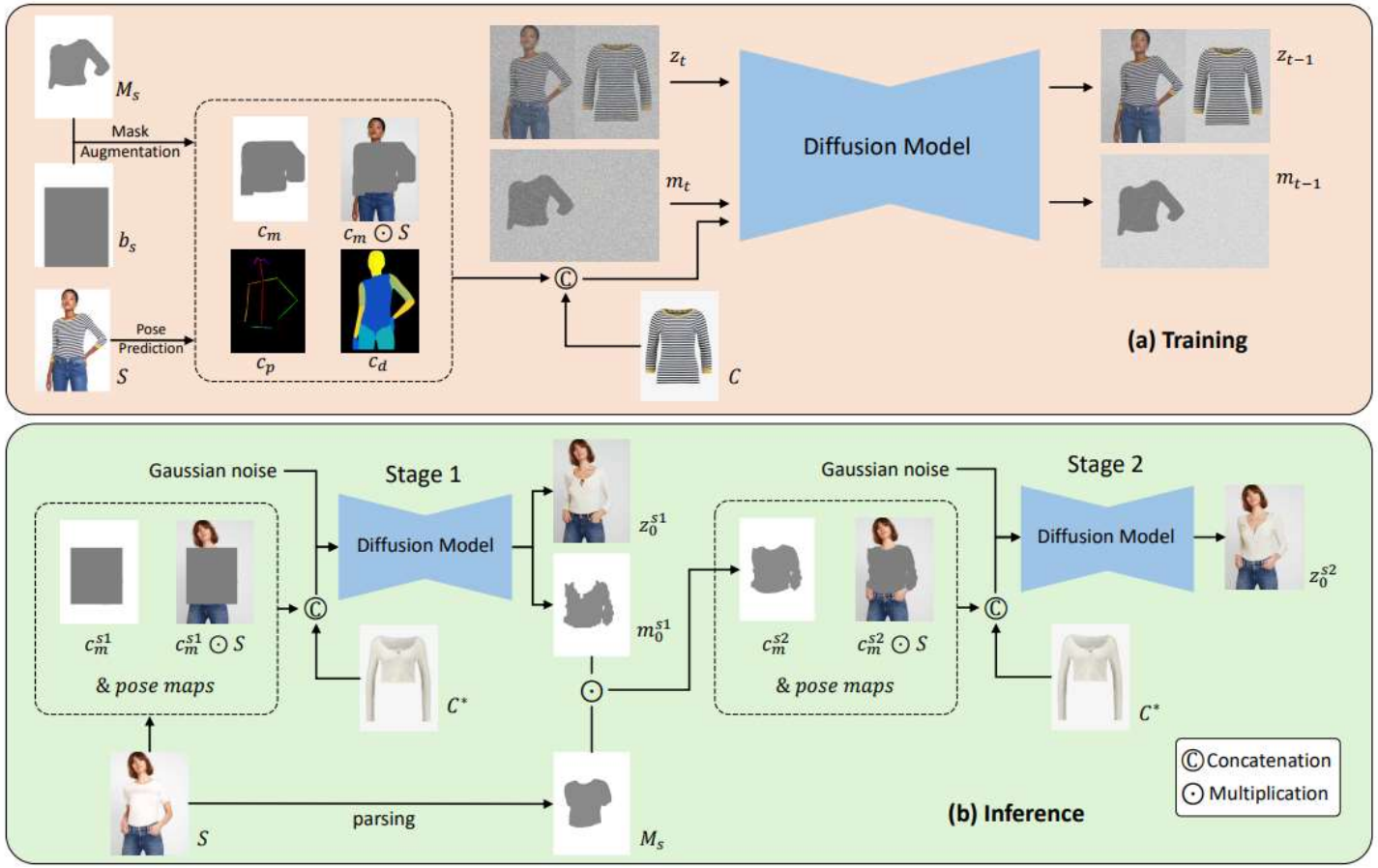


# INTRODUCTION

## Current works 2

- **TPD** mitigates information loss by reducing the mask area through precise mask
- Fails to fully resolve the issues caused by mask

TPD:



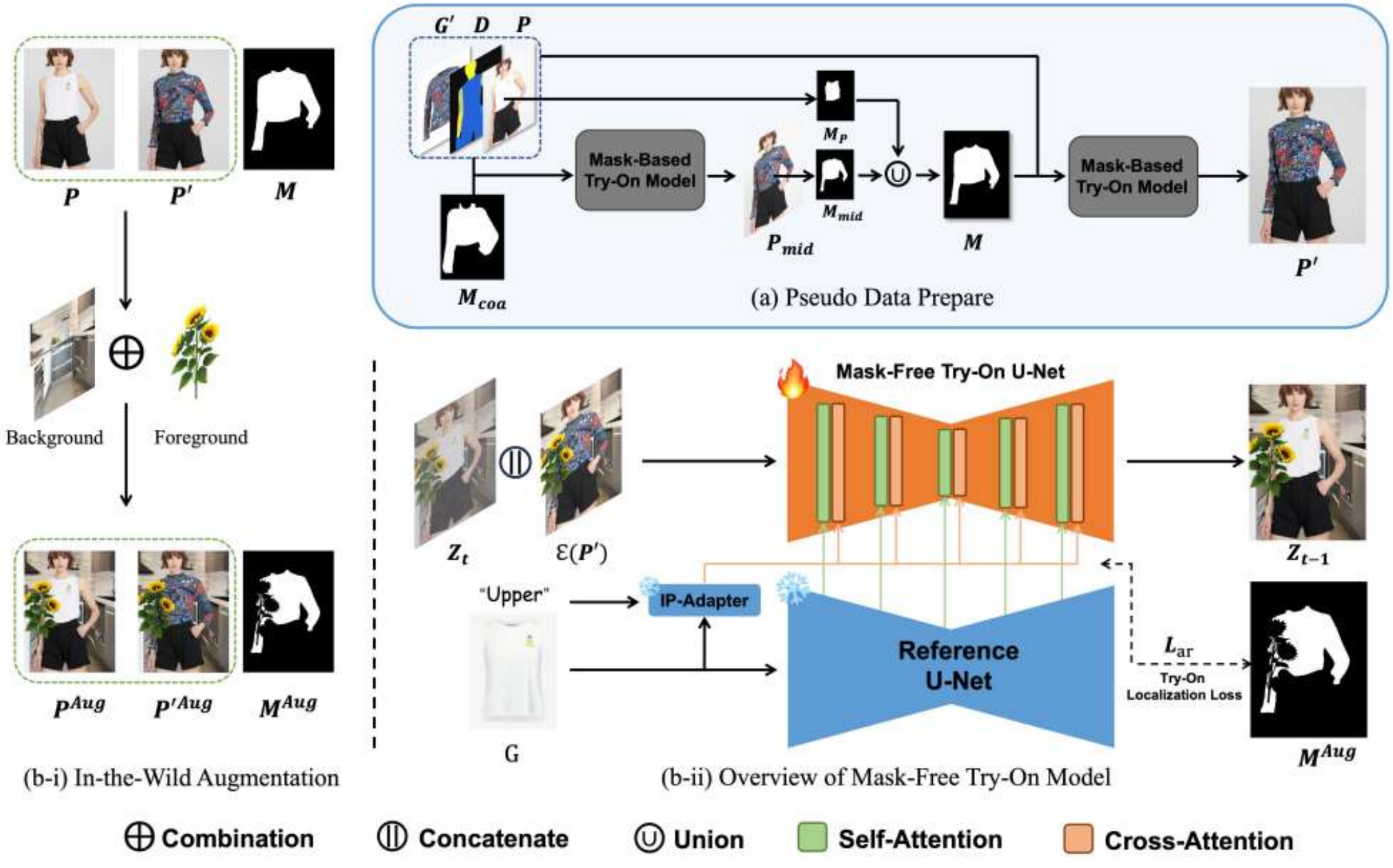
# INTRODUCTION

## Contribution

- Propose BooW-VTON, a mask-free in-the-wild virtual try-on diffusion model without any additional parser
  - In the-wild data augmentation
  - Try-on localization loss

# METHOD

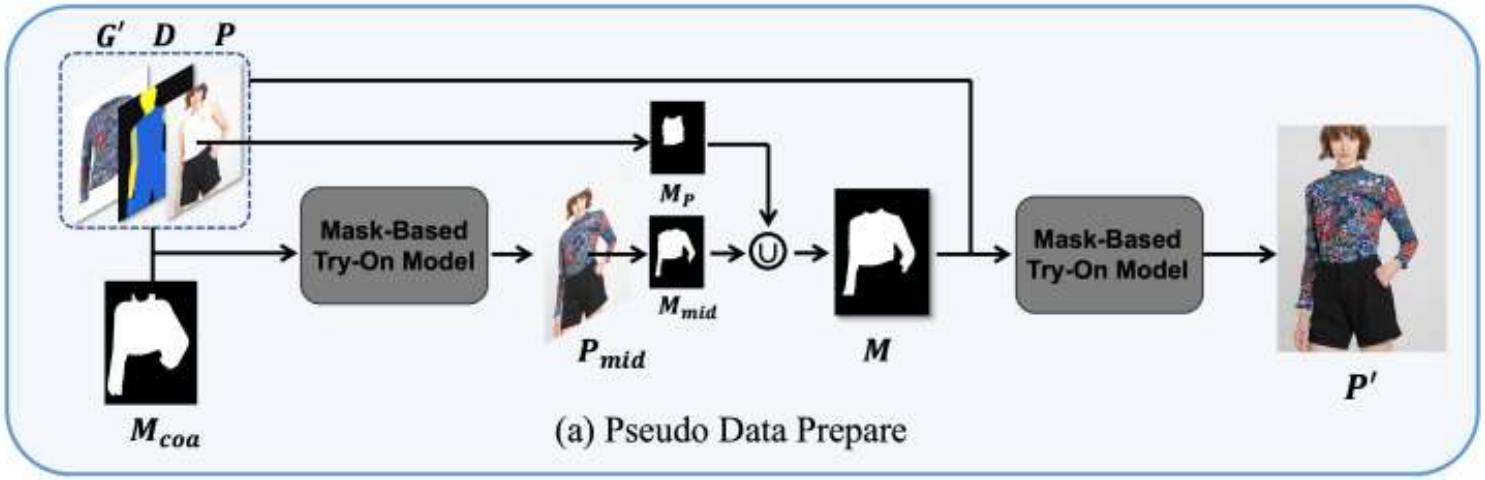
## Overview



# METHOD

## (a) Pseudo Data Prepare

- Paired setting: $P\&G \rightarrow P$
- Unpaired setting: $P\&G' \rightarrow P'$




$P$: person image
$P'$: result of person $P$ wearing garment $G'$
$G$: garment worn by the $P$
$G'$: another garment image different from $G$

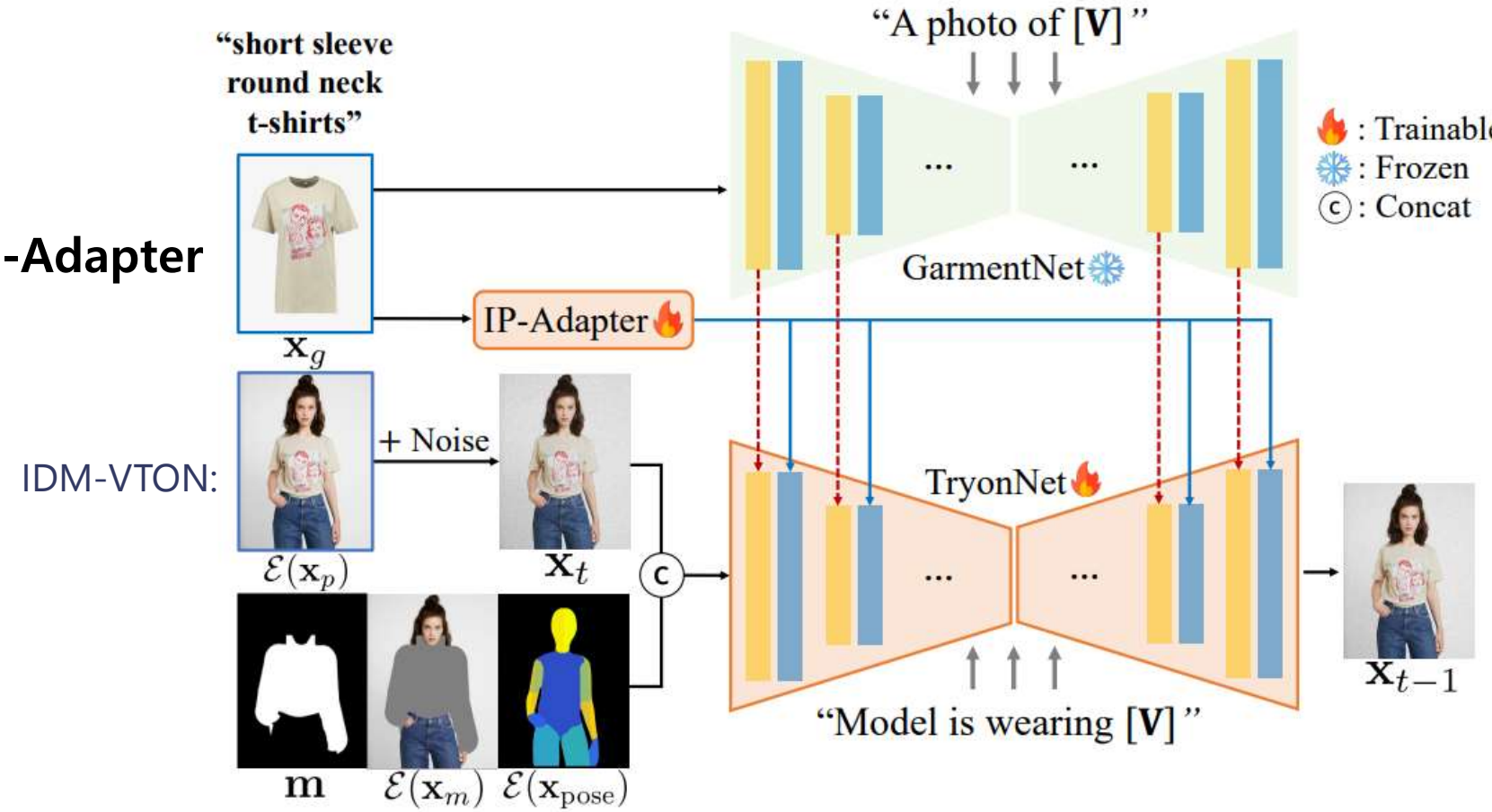
- The results $P_{mid}$ contain imperfections caused by the mask
- Adopt a two-stage inference approach
- For next step: $\{P, P', M\}$

# METHOD

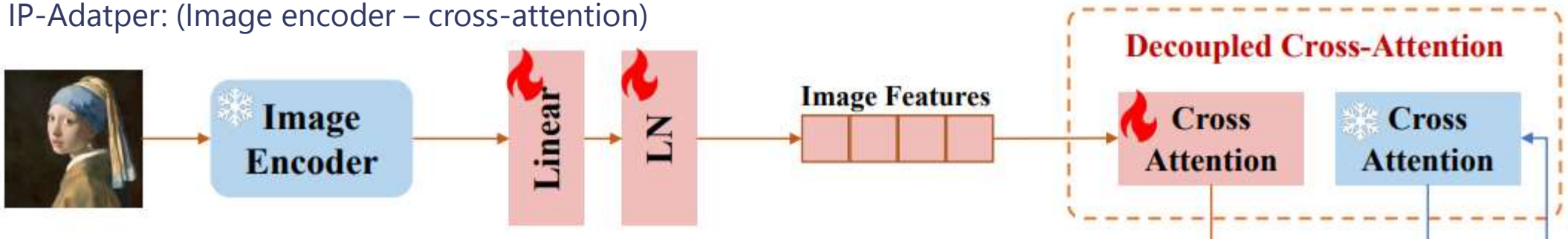
## (a) Pseudo Data Prepare

- Model: **IDM-VTON** with **IP-Adapter**

IDM-VTON:




IP-Adatper: (Image encoder – cross-attention)



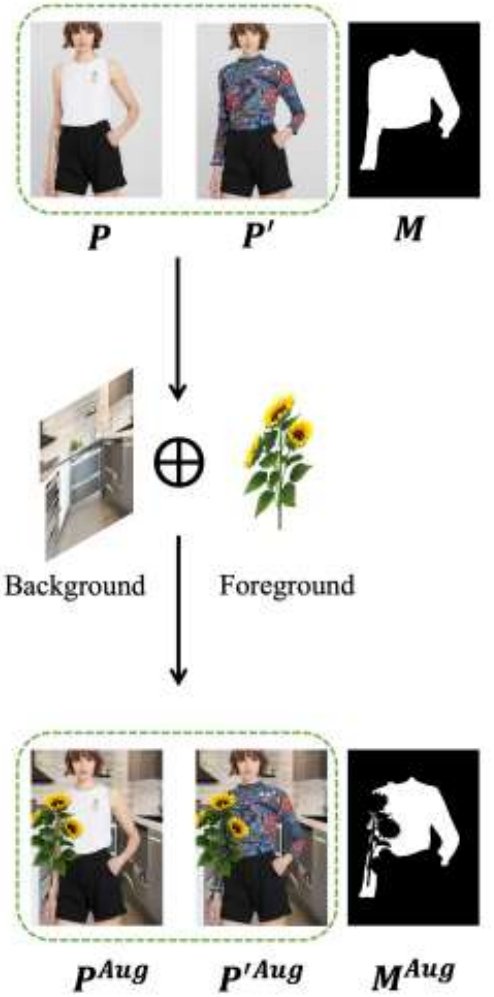
# METHOD

## (b-i) In-the-Wild Augmentation

- $\{P, P', M\}$: simple try-on samples
- Background: T2I model **SDXL**


(a) Generation of Background

{ } in a china garden
{ } in a snowy winter landscape with pine trees
{ } in a bustling urban street scene



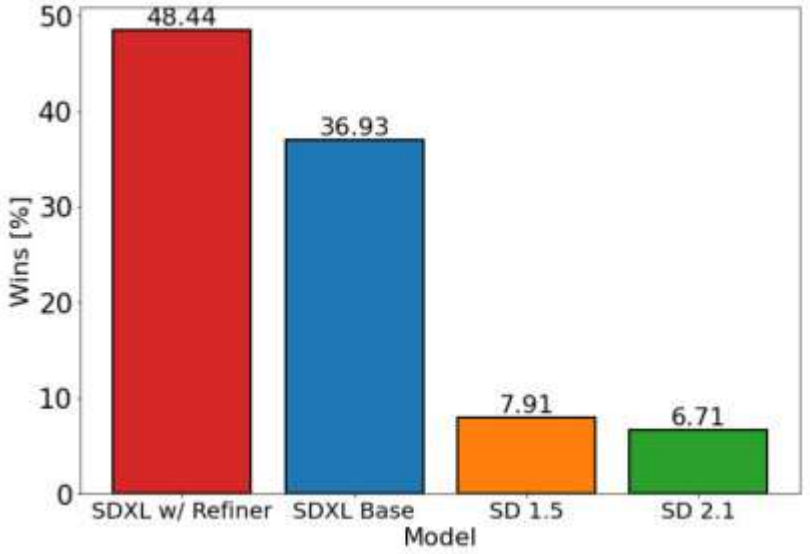

(b-i) In-the-Wild Augmentation

$P$: person image
$P'$: result of person $P$ wearing garment $G'$

# METHOD

## (b-i) In-the-Wild Augmentation

- **SDXL** vs **SD**
  - X3 U-Net backbone
  - Two text encoders
  - Refinement model



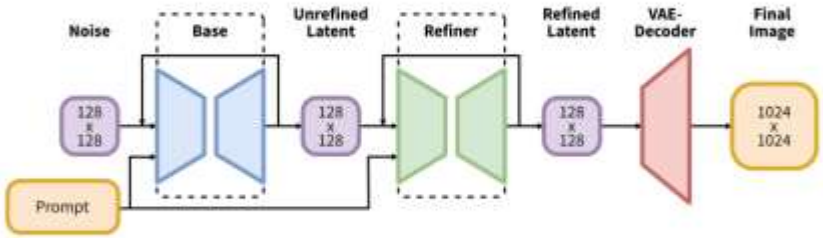



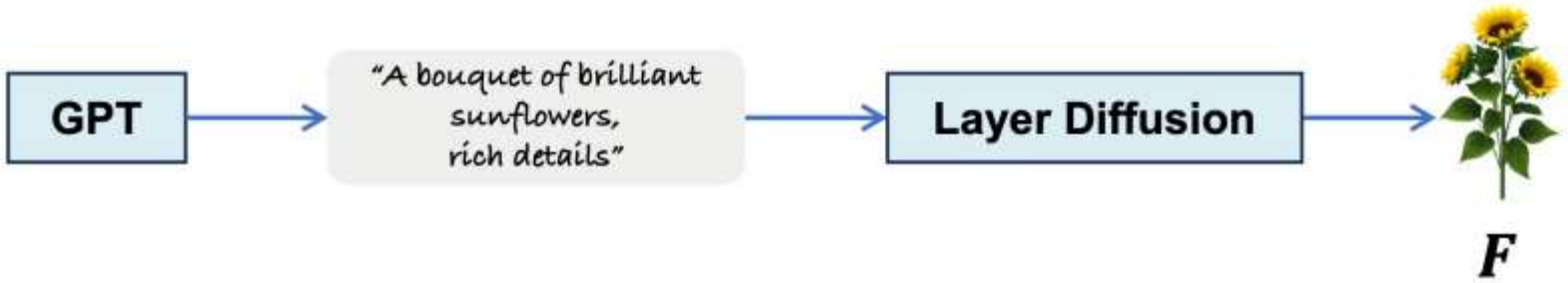

Table 1: Comparison of *SDXL* and older *Stable Diffusion* models.

| Model | *SDXL* | SD 1.4/1.5 | SD 2.0/2.1 |
|---|---|---|---|
| # of UNet params | 2.6B | 860M | 865M |
| Transformer blocks | [0, 2, 10] | [1, 1, 1, 1] | [1, 1, 1, 1] |
| Channel mult. | [1, 2, 4] | [1, 2, 4, 4] | [1, 2, 4, 4] |
| Text encoder | CLIP ViT-L & OpenCLIP ViT-bigG | CLIP ViT-L | OpenCLIP ViT-H |
| Context dim. | 2048 | 768 | 1024 |
| Pooled text emb. | OpenCLIP ViT-bigG | N/A | N/A |

# METHOD

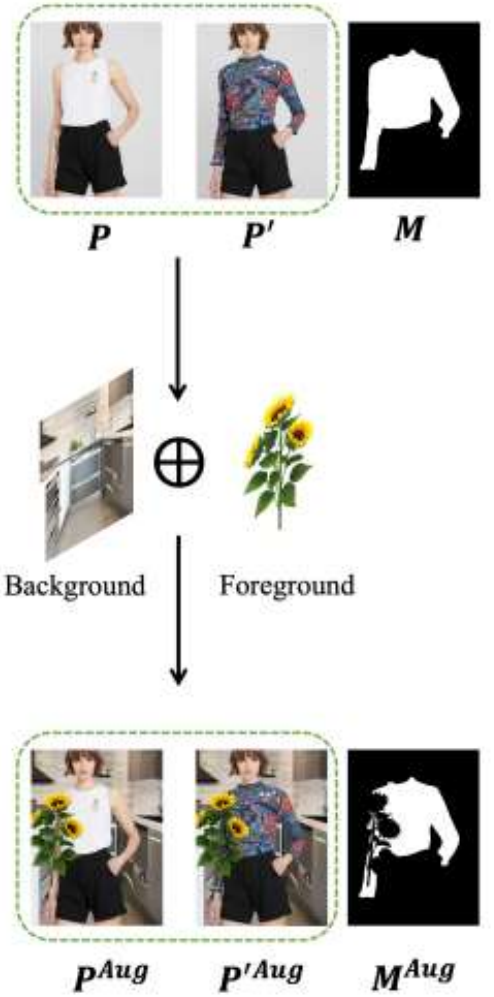
## (b-i) In-the-Wild Augmentation

- $\{P, P', M\}$: simple try-on samples
- Foreground: **LayerDiffusion**




(b) Generation of Foreground

Colorful balloons, helium-filled, shiny and reflective
Sunflower bouquet, vibrant yellow, lush green stems
Vintage bicycle, red paint, woven wicker basket



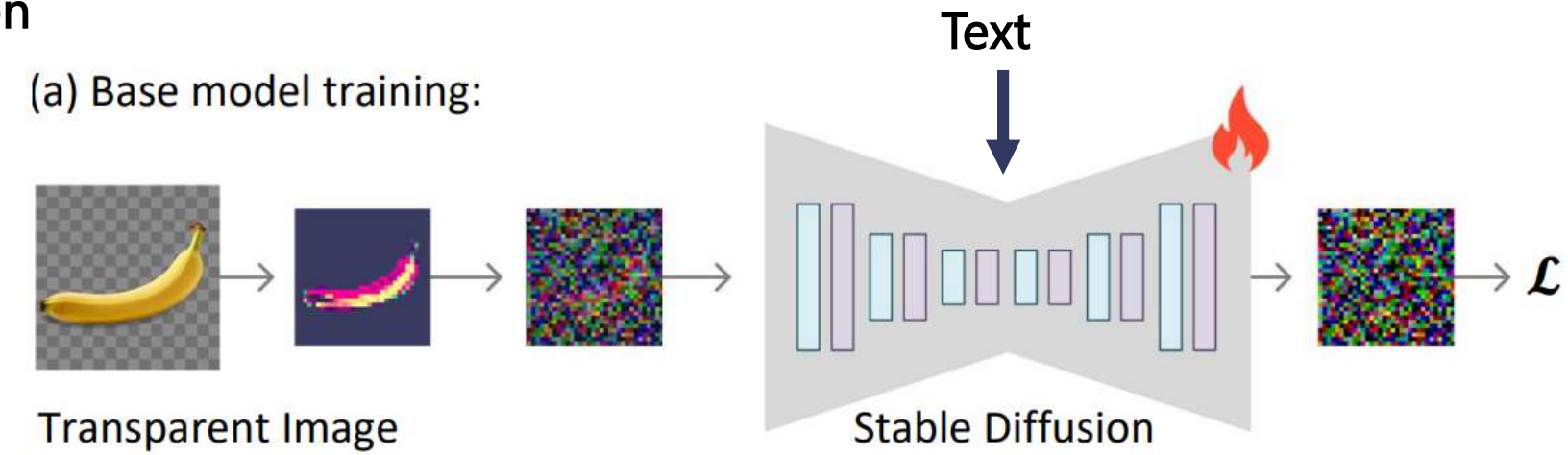

(b-i) In-the-Wild Augmentation

$P$: person image
$P'$: result of person $P$ wearing garment $G'$

# METHOD

## (b-i) In-the-Wild Augmentation

- LayerDiffusion




Inference:

# METHOD

## (a) & (b-ii)

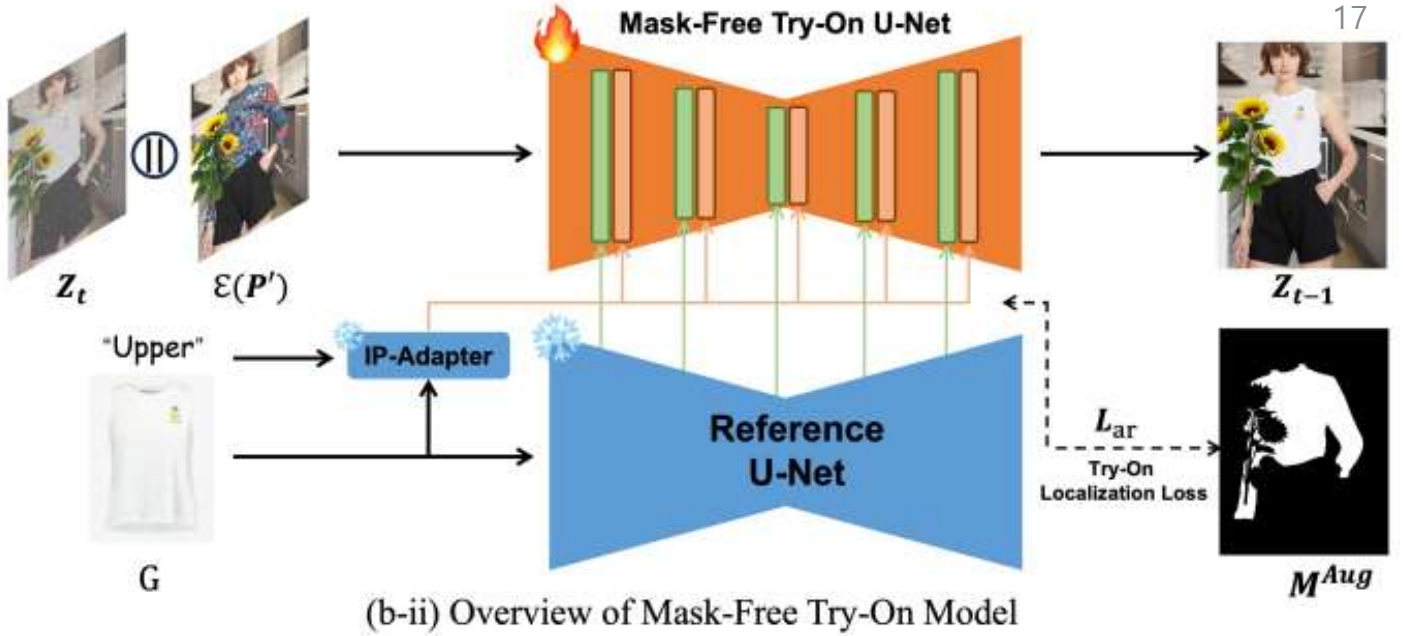
- Create in-the-wild unpaired dataset

$P^{Aug}$ $P'^{Aug}$ $M^{Aug}$

# METHOD

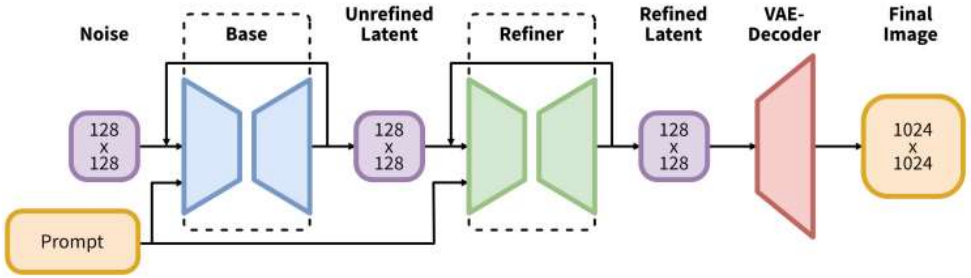
## (b-ii) Mask-Free Try-On Model

- Try-On U-Net: **SDXL**
- **IP-Adapter**: garment & text encoder
- Reference U-Net: garment encoder



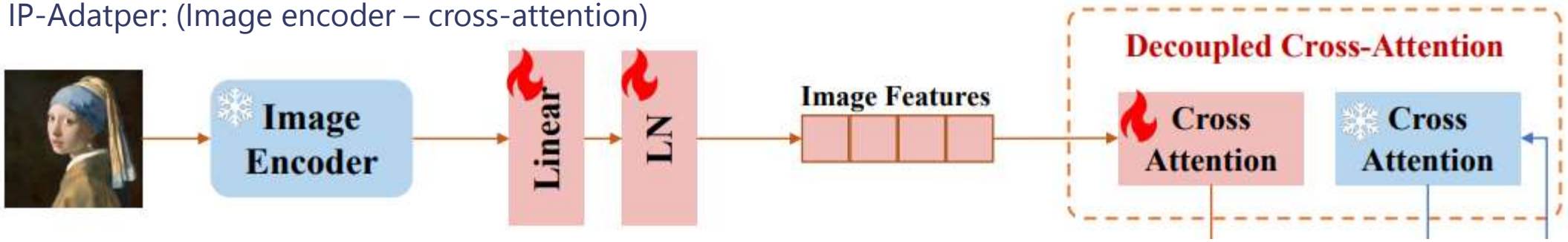

(b-ii) Overview of Mask-Free Try-On Model

SDXL




$P$: person image
$Z_t$: latent code at time-step $t$
$G$: garment worn by the $P$
$\boldsymbol{\varepsilon}$: encoder

IP-Adatper: (Image encoder – cross-attention)

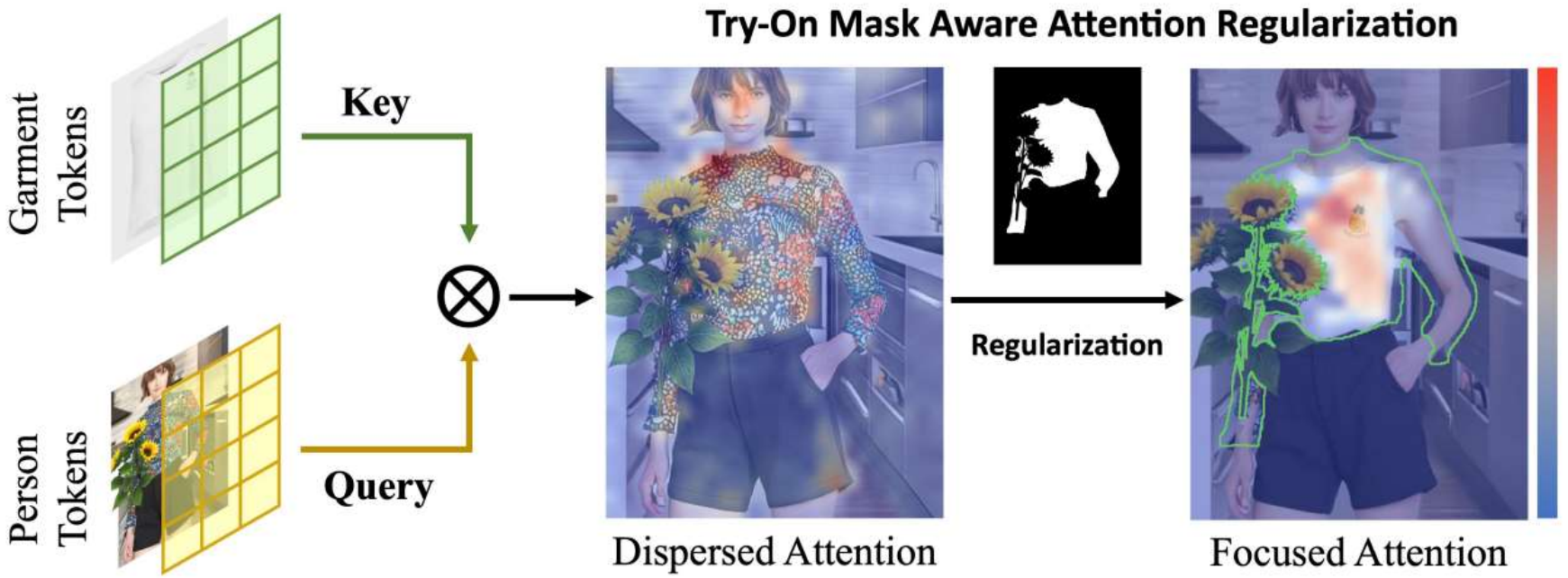


# METHOD

## Try-on localization loss

- Help the model correctly identify try-on areas

$$\mathcal{L}_{\text{ar}} = \frac{1}{n}\sum_{k=1}^{n} \text{mean}(A_k(1 - M^{Aug})).$$



# METHOD

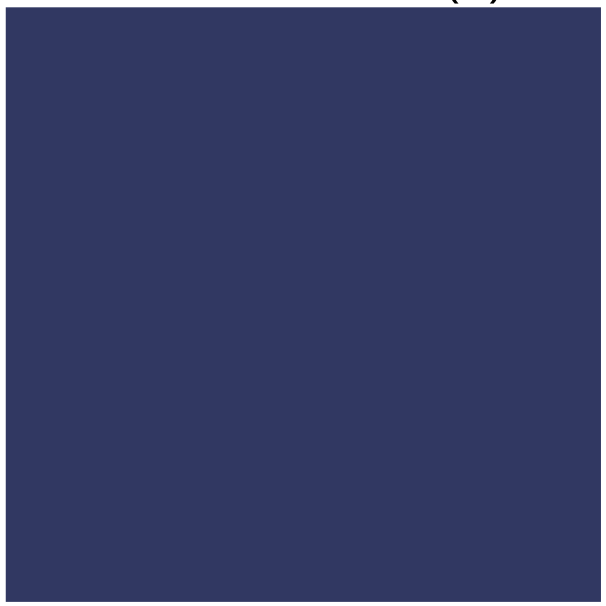
## Try-on localization loss

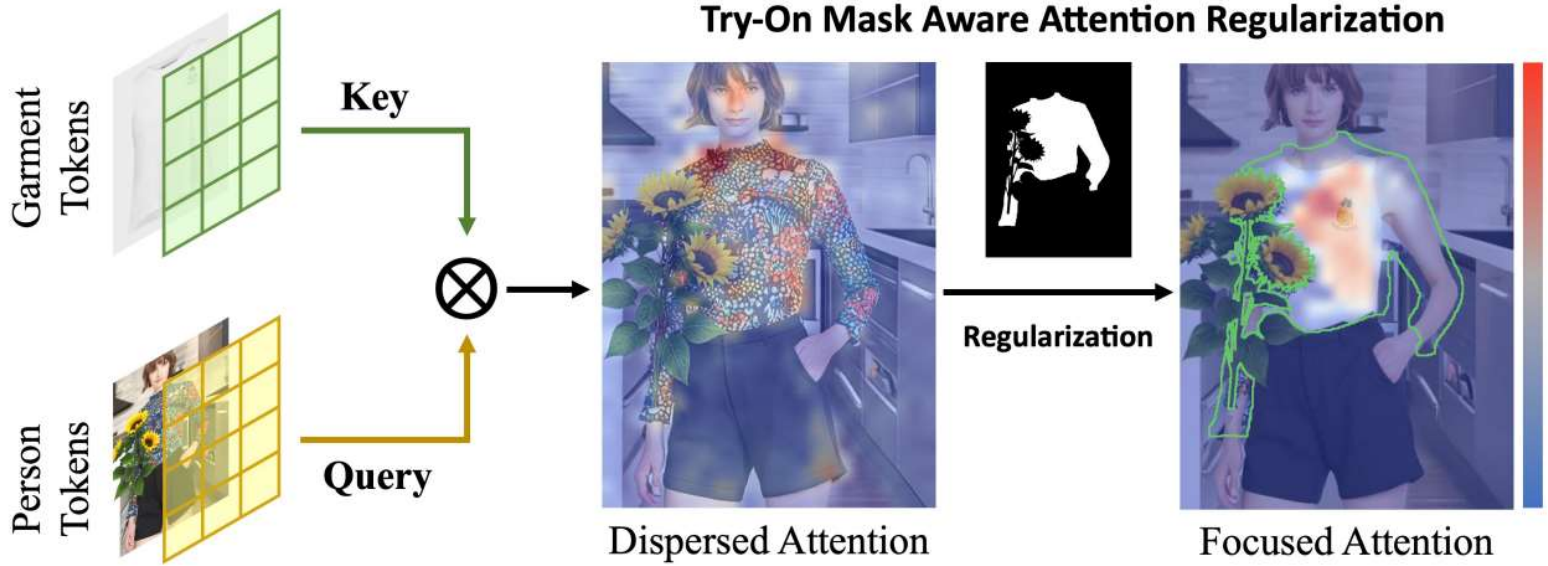
- Help the model correctly identify try-on areas

$$\mathcal{L}_{\text{ar}} = \frac{1}{n}\sum_{k=1}^{n} \text{mean}(A_k(1 - M^{Aug})).$$

Garment tokens (n)

Person tokens $(hw)$



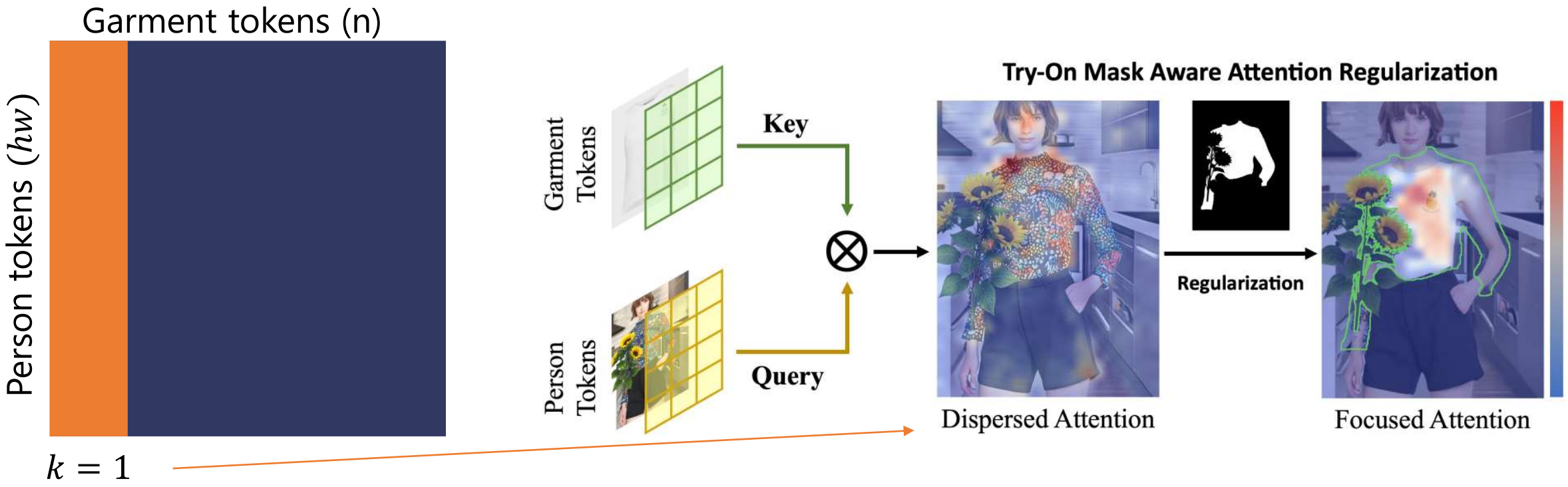


# METHOD

## Try-on localization loss

- Help the model correctly identify try-on areas

$$\mathcal{L}_{\text{ar}} = \frac{1}{n}\sum_{k=1}^{n} \text{mean}(A_k(1 - M^{Aug})).$$

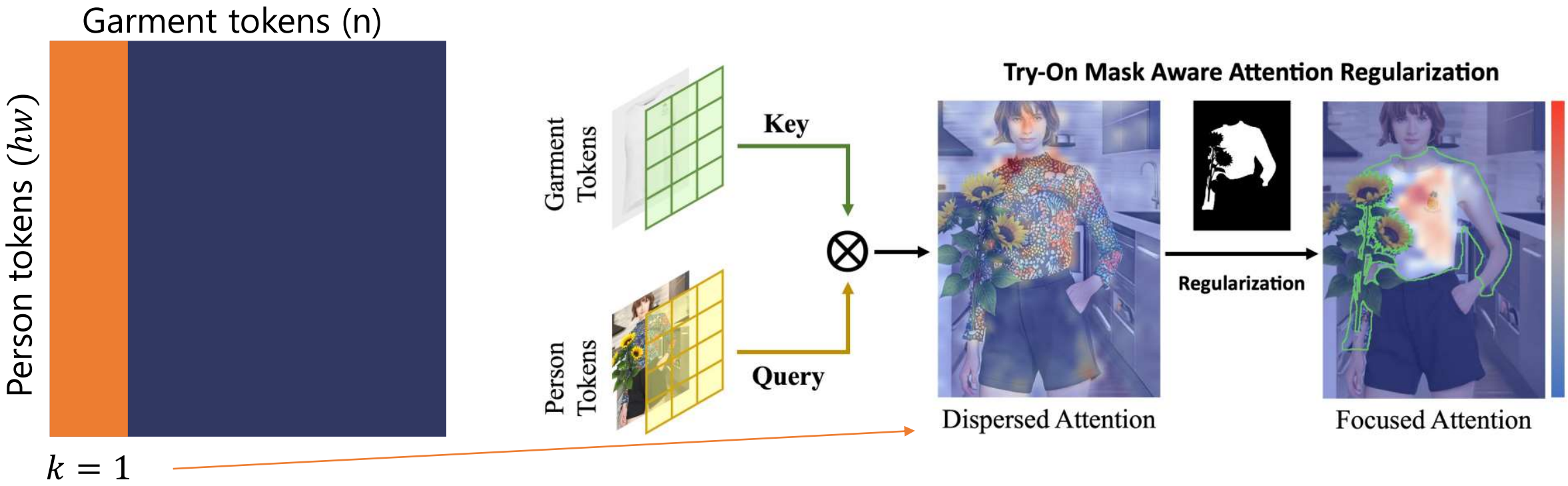


# METHOD

## Try-on localization loss

- Help the model correctly identify try-on areas

$$\mathcal{L}_{\text{ar}} = \frac{1}{n}\sum_{k=1}^{n} \text{mean}(A_k(1 - M^{Aug})).$$




The information of the garment should be within try-on areas

# EXPERIMENTS

## Baselines

- DCI-VTON, LaDI-VTON, CAT-DM, StableVITON, TPD, OOTD, IDM-VTON

## Datasets

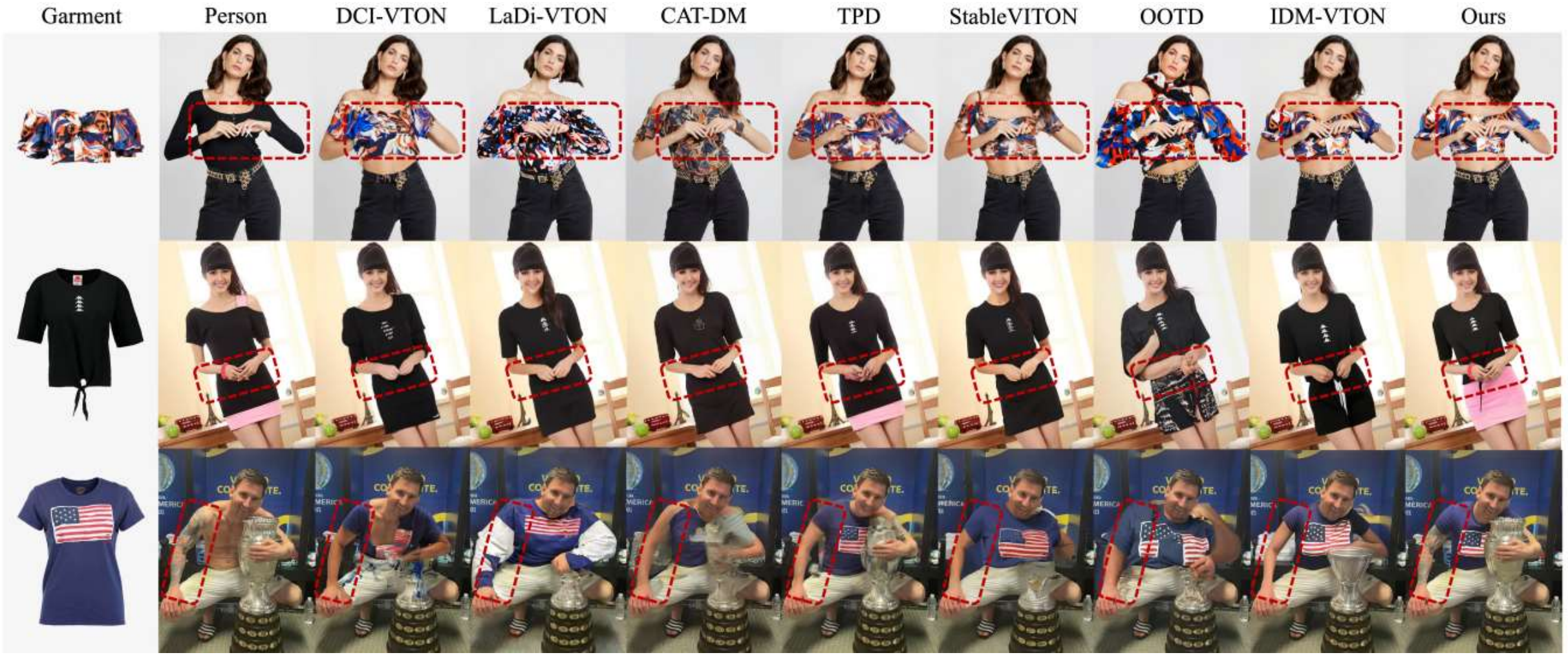
| Dataset | Images | Resolution | Pairs | Wild Back | Wild Fore |
|---|---|---|---|---|---|
| VITON-HD | 2032 | $1024 \times 768$ | ✓ | ✗ | ✗ |
| DressCode | $1800 \times 3$ | $1024 \times 768$ | ✓ | ✗ | ✗ |
| StreetVTON | 2089 | Various | ✗ | ✓ | ✗ |
| WildVTON | 1224 | Various | ✗ | ✓ | ✓ |

## Evaluation metric

- Paired setting: LPIPS, SSIM, PSNR
- Unpaired setting: FID, KID

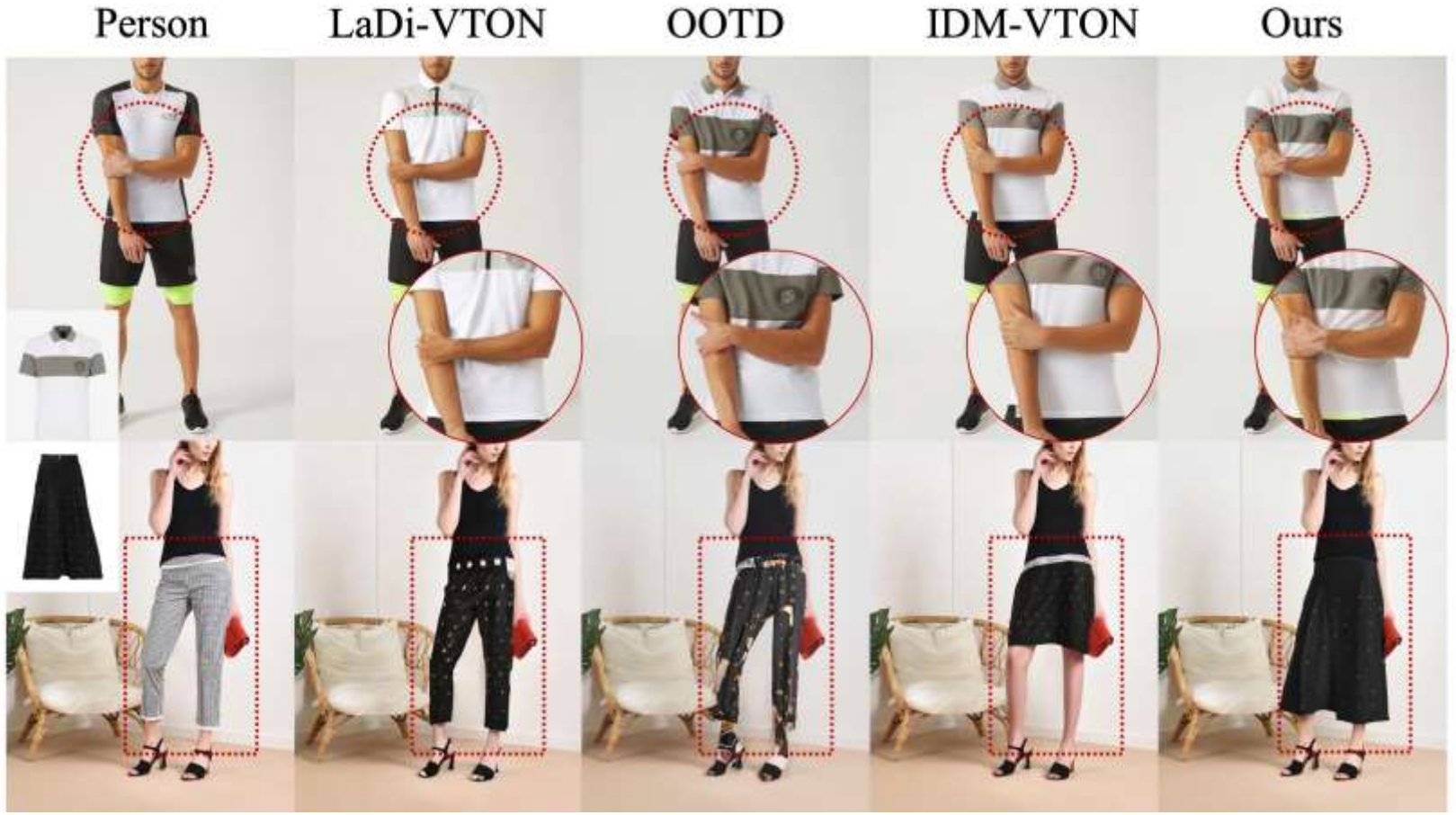
# EXPERIMENTS

- Occlusion (VITON-HD), accessories (StreetVITON), Tattoo (WildVTON)



# EXPERIMENTS

- DressCode
  - Muscle details
  - Characteristics of skirt



# EXPERIMENTS

- Base mask-free: distillation from **IDM-VTON** → in-the-wild

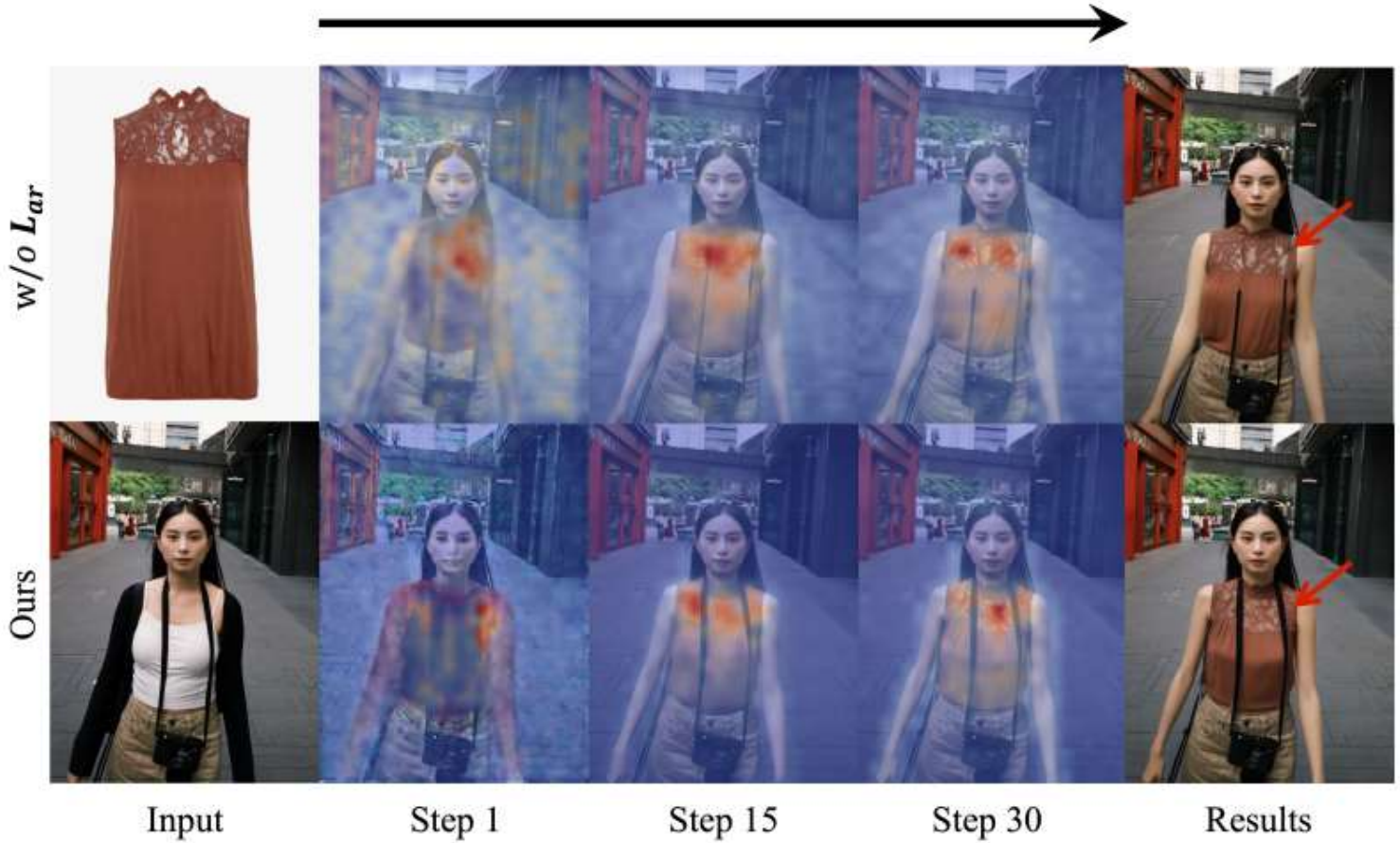
| Testsets | | VITON-HD | | | | | | | StreetVTON | | WildVTON | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | | **LPIPS ↓** | **SSIM ↑** | **PSNR ↑** | **$FID_p$ ↓** | **$KID_p$ ↓** | **$FID_u$ ↓** | **$KID_u$ ↓** | **$FID_u$ ↓** | **$KID_u$ ↓** | **$FID_u$ ↓** | **$KID_u$ ↓** |
| DCI-VTON | MM2023 | 0.1800 | 0.8545 | 19.27 | 24.05 | 16.44 | <u>8.998</u> | 1.187 | 20.95 | **3.470** | 35.66 | <u>4.649</u> |
| LaDI-VTON | MM2023 | 0.2014 | 0.8395 | 18.69 | 9.746 | 2.599 | 11.08 | 2.634 | 24.12 | 5.638 | 44.54 | 9.203 |
| CAT-DM | CVPR2024 | 0.1621 | 0.8391 | 20.45 | 9.336 | 2.294 | 10.28 | 1.980 | 25.84 | 6.879 | 42.16 | 8.374 |
| TPD | CVPR2024 | 0.1822 | 0.8516 | 20.75 | 13.07 | 7.880 | 13.82 | 6.641 | 23.02 | 4.671 | 45.37 | 14.67 |
| StableVITON | CVPR2024 | 0.1479 | 0.8519 | <u>21.72</u> | 8.926 | 2.538 | 9.851 | 1.727 | 23.15 | 4.628 | 42.32 | 8.194 |
| OOTD | 5.9k stars | 0.1420 | 0.8301 | 19.20 | 8.136 | 1.469 | 12.19 | 2.682 | 27.00 | 7.473 | 40.68 | 5.606 |
| IDM-VTON | ECCV2024 | 0.1223 | 0.8547 | 21.06 | 8.594 | 2.529 | 9.265 | 1.272 | 23.62 | 6.181 | 38.77 | 7.686 |
| Base mask-free | | 0.1206 | 0.8529 | 20.60 | 8.766 | 2.611 | 9.467 | 1.493 | 28.81 | 8.026 | 57.52 | 18.59 |
| +H.Q. pseudo data | | 0.1101 | <u>0.8597</u> | 21.37 | <u>6.896</u> | 1.580 | 9.191 | 1.120 | 27.26 | 7.616 | 56.14 | 18.31 |
| +Wild augmentation | | <u>0.1173</u> | 0.8589 | 21.23 | 7.405 | <u>1.405</u> | 9.204 | <u>1.089</u> | <u>21.70</u> | 5.170 | <u>35.62</u> | 6.180 |
| +$L_{ar}$ (Full Model) | | **0.1080** | **0.8618** | **21.80** | **6.885** | **1.366** | **8.809** | **0.8176** | **20.50** | <u>4.494</u> | **32.53** | **4.509** |

# EXPERIMENTS

- Ablation study
  - Wild Aug & Loss: preservation of non-try-on content

# EXPERIMENTS

- Ablation study
  - 1, 15, 30 denoising step
  - Self-attention layer in 35$^{th}$ attention block



# EXPERIMENTS

- Application


(a) Prompt Controlled Try-On (b) Full-Body Try-On

# EXPERIMENTS

- Limitation