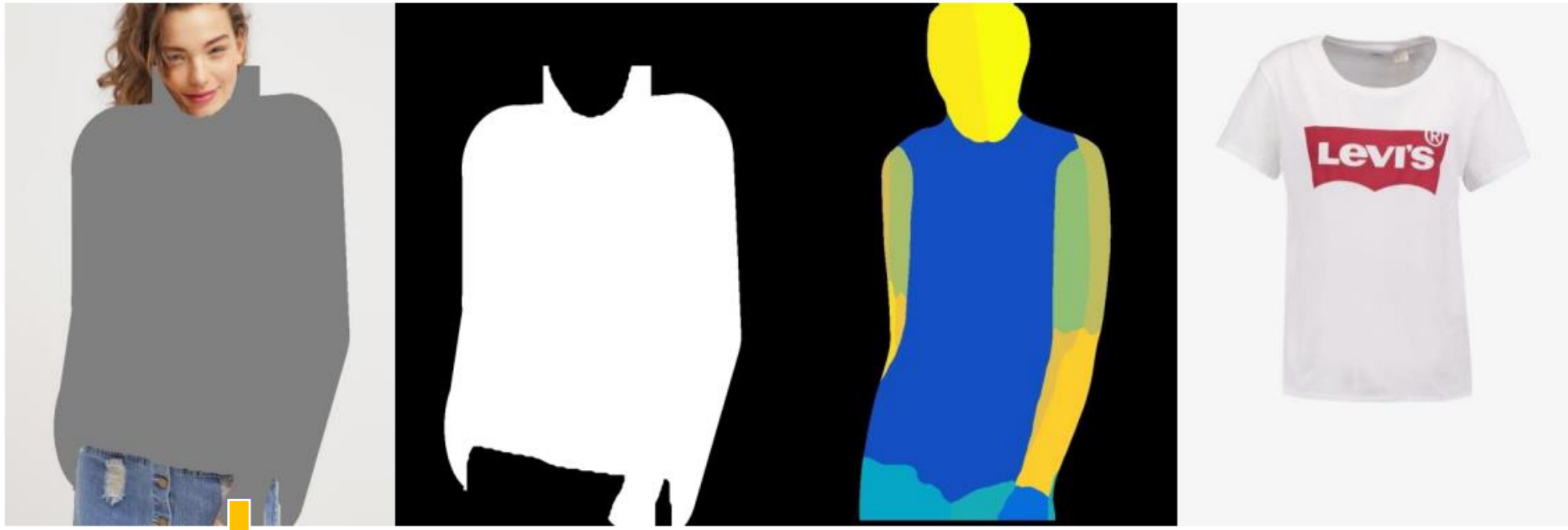
FLDM-VTON: Faithful Latent Diffusion Model for Virtual Try-On

IJCAI(2024), Fudan University, 0 citation

1 Introduction

- Task: Image-based VITON
  - Requirements: while maintaining the pose, body shape, and identity of the person, the clothing product must seamlessly deform to the desired clothing area
  - To meet the requirements, utilize various conditions: clothing, dense pose, etc.

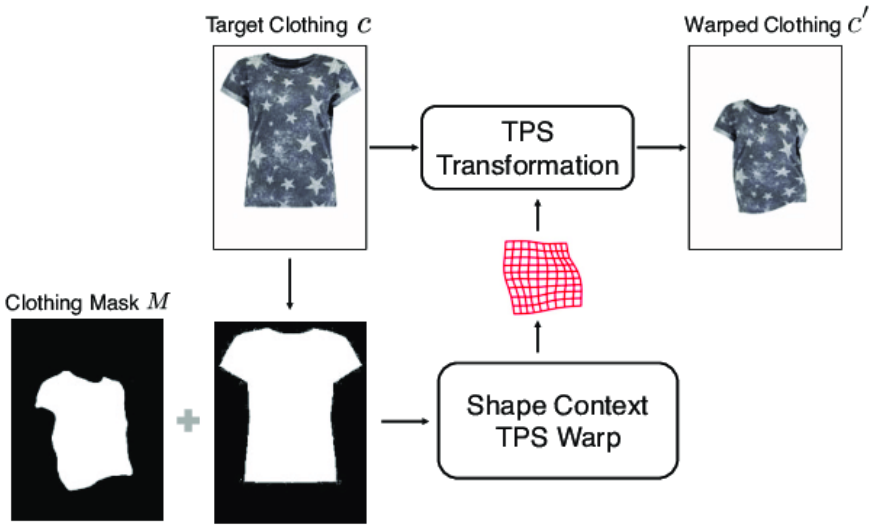


agnostic image: combine person and inpainting mask

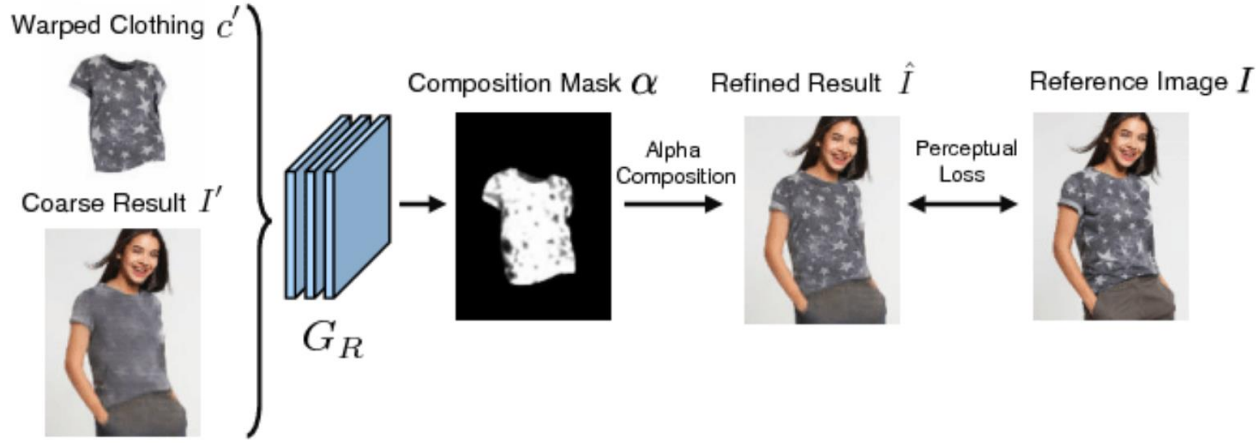
1 Introduction

- Previous works: GANs
  - It consists of two stages
  - Due to the model collapse issue, GAN failed to synthesis photo-realistic images and capture detail of clothing

Step 1. geometric matching



Step 2. refine the image



1 Introduction

- Current works: LDMs
  - Still lack faithfulness of clothing: style, pattern, and text
  - The cause of the problem:

    (1) stochastic property of diffusion model, (2) latent supervision

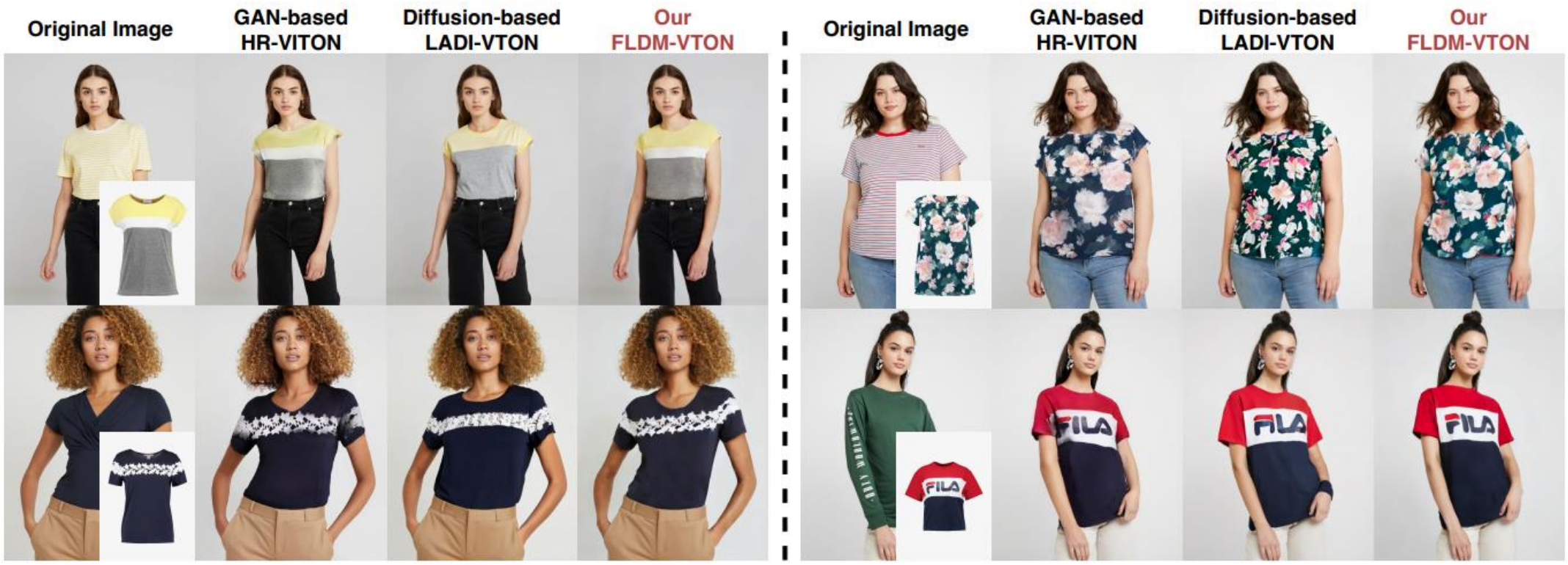
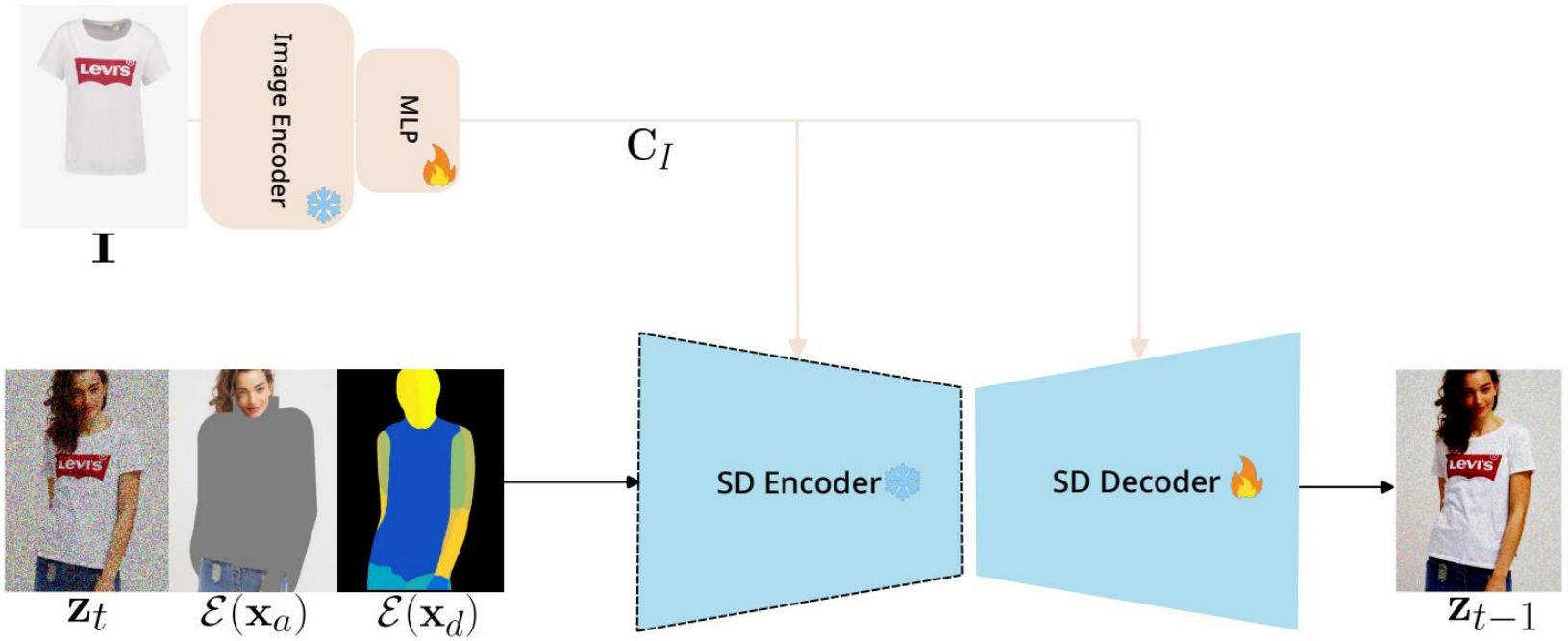


Figure 1: Comparison between state-of-the-art baselines and our FLDM-VTON on the VITON-HD dataset.

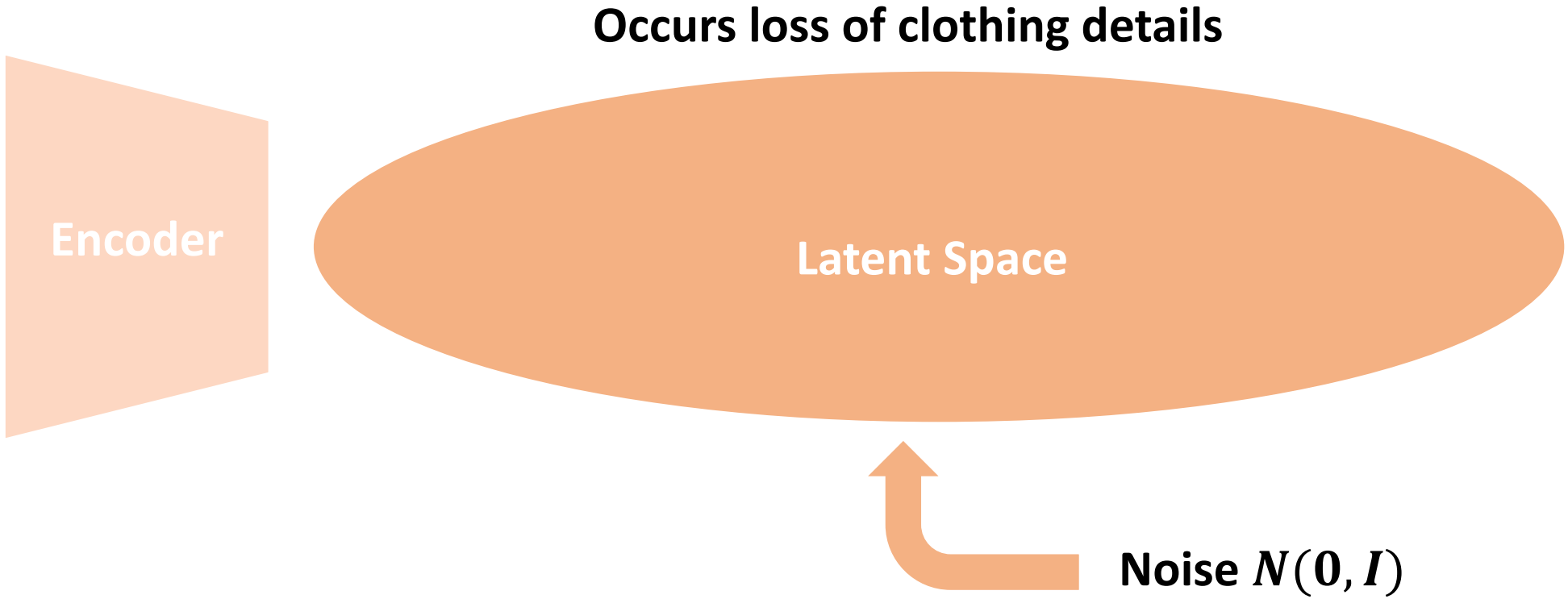
1 Introduction

- Basic generation process for LDMs performing VITON
  - Input clothing condition utilizing cross-attention
  - Concatenate various condition to noisy input
  - Model (U-Net) generates more clear image by predicting noise



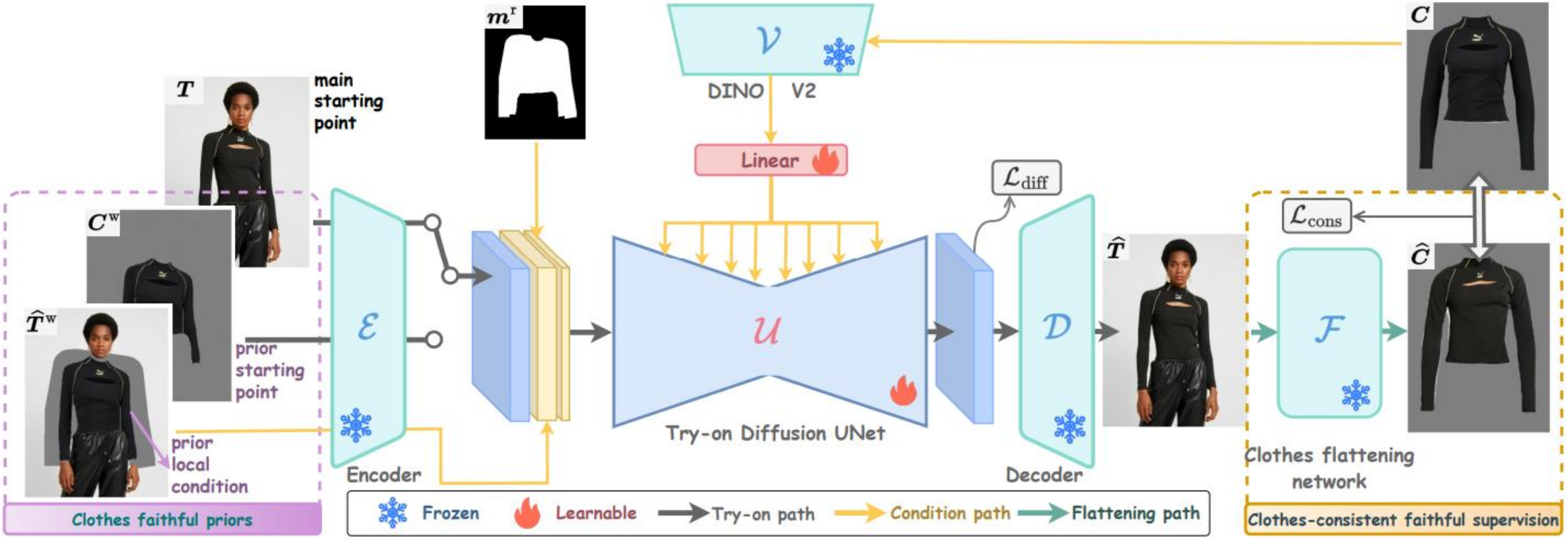
2 Definition of problem

- stochastic property of diffusion model
- latent supervision



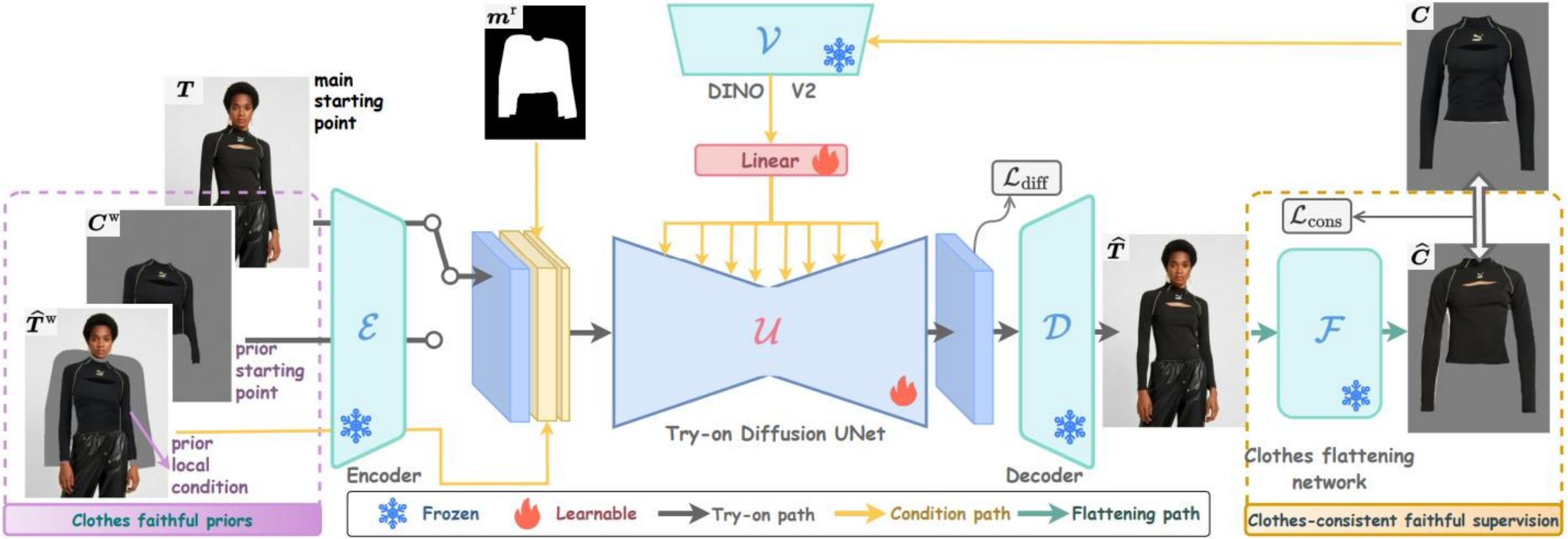
3 Methodology

- Similar components with other papers: image encoder, U-Net, resized mask
- Notable components: clothes faithful prior, clothes-consistent faithful supervision



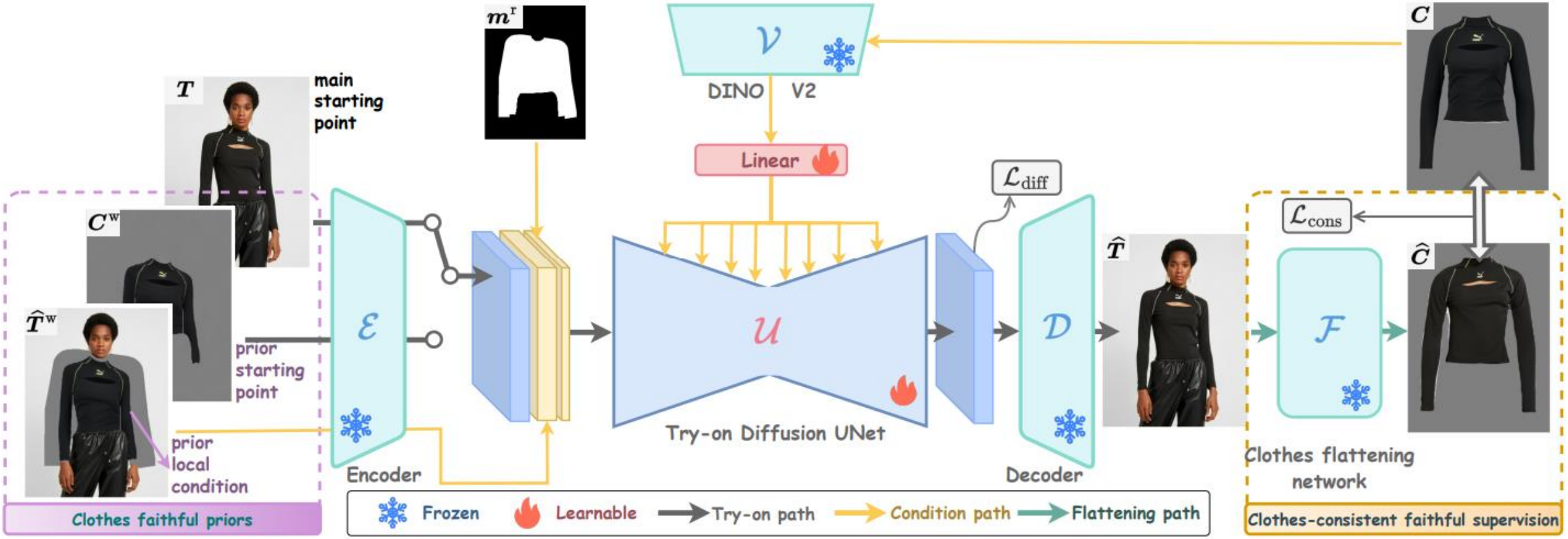
3 Methodology

- Clothes faithful priors: reduce stochastic property in diffusion process
  - Takes either *main starting point* or *prior starting point* as input
  - Encourage usage of *prior local condition*



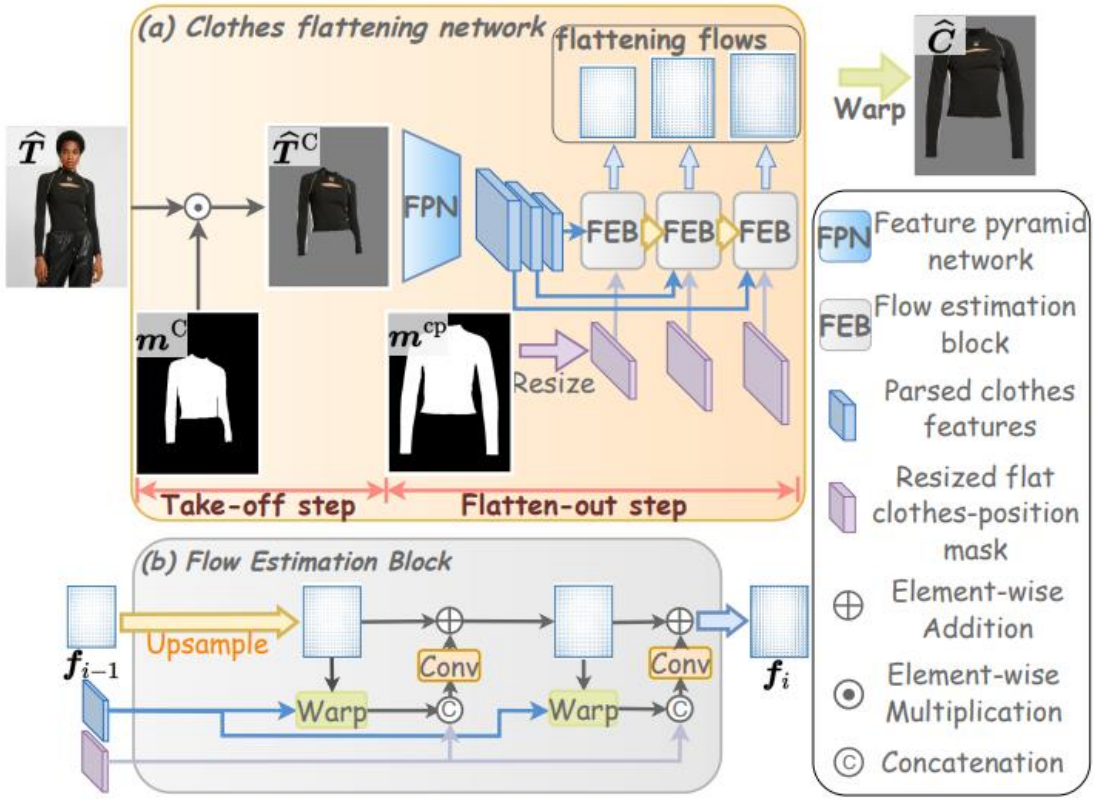
3 Methodology

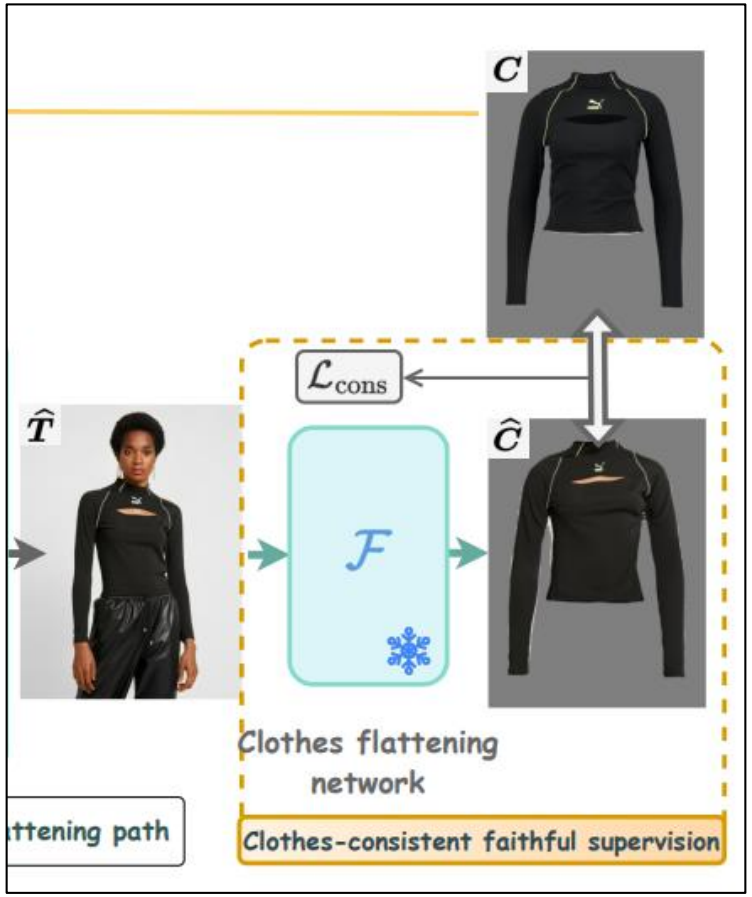
- Clothes-consistent faithful supervision
  - Preserve fine detail of clothing
  - Extract warped clothing in generated image, then generates original clothing



3 Methodology

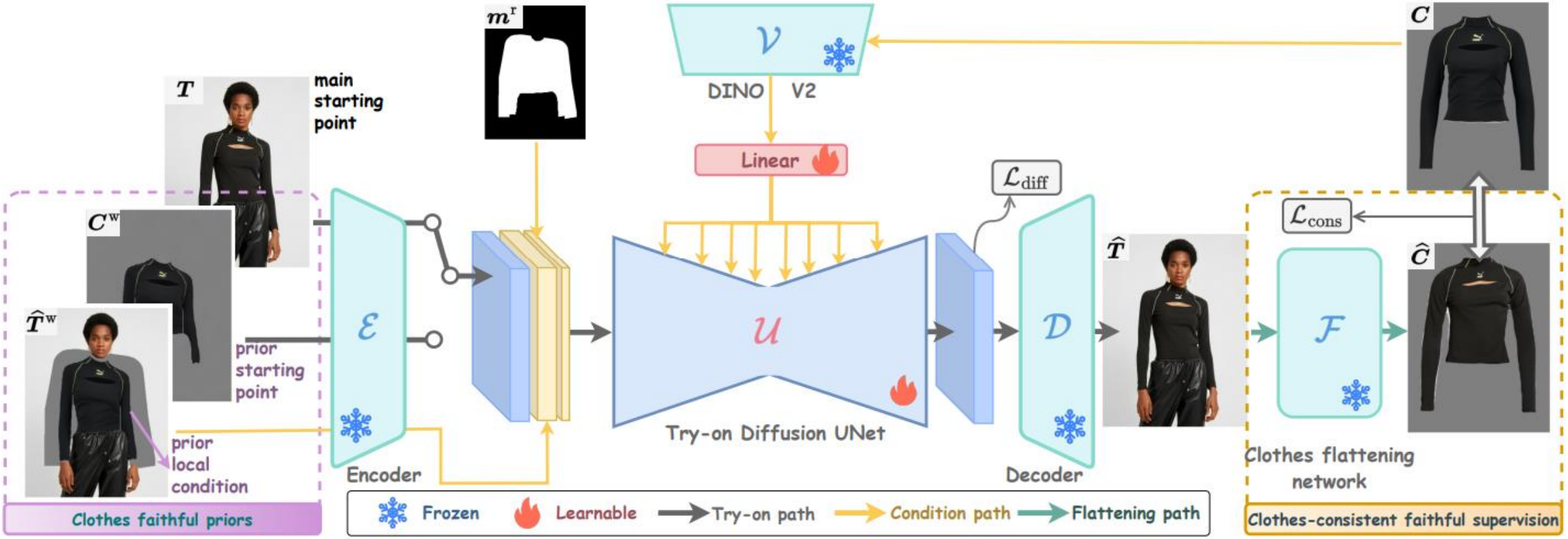
- Clothes flattening network
  - Take-off step: extract warped clothing
  - Flatten-out step: generate original clothing





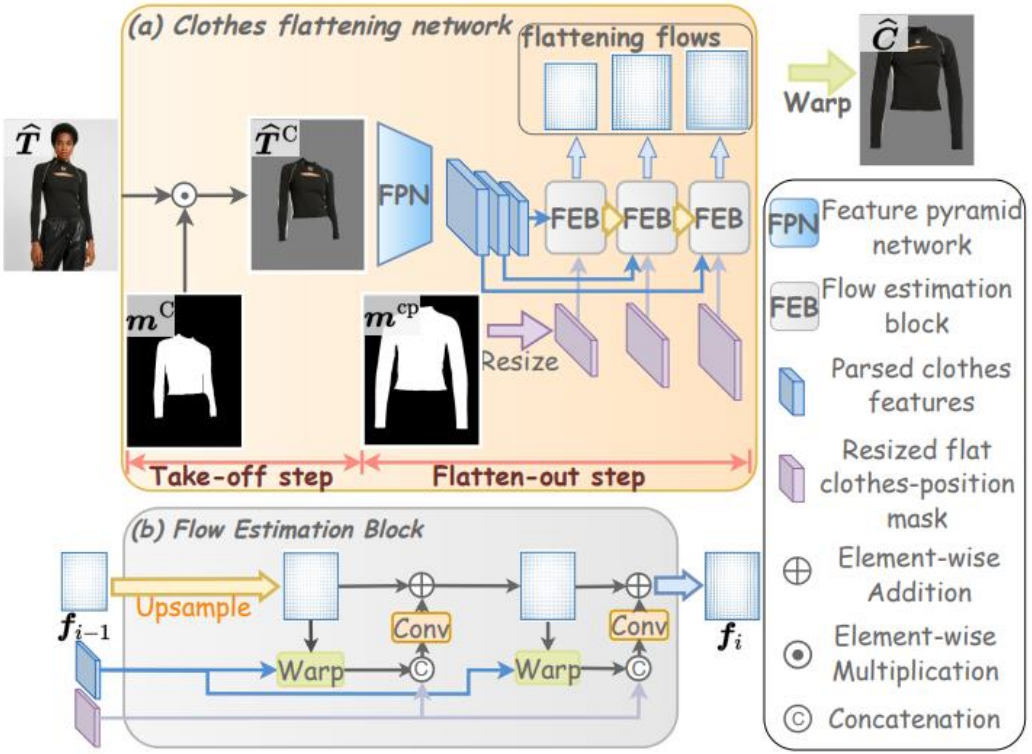
3 Methodology

- Loss
  - L_{diff}: existing loss for noise prediction
  - L_{cons}: L_1 loss between generated clothing and real clothing



3 Methodology

- Train flattening network
  - L_1: pixel-level
  - Perceptual loss: feature-level
  - Total variation loss: pixel-level, for smoothness
  - Second-order smooth loss: for smoothness



4 Experiments

- Setting
  - Dataset: VITON-HD (11,647/2,032)
    - ✓ frontal-view woman and top clothing image pairs
  - Paired setting: there’s an image of a person wearing clothes
    - ✓ SSIM ↑, LPIPS ↓: similarity between the two images
  - Unpaired setting
    - ✓ KID ↓, FID ↓: measure the statistical similarity real and generated images

4 Experiments

- More precise generation for text, style, etc.

4 Experiments

- I: CNN-based
- II: GAN-based
- III: Diffusion-based

Methods		Paired		Unpaired	
		LPIPS↓	SSIM↑	FID↓	KID↓
CP-VTON	I	0.160	0.831	31.34	2.37
CP-VTON+		0.131	0.847	22.79	1.55
VITON-HD	II	0.116	0.862	12.12	0.32
HR-VITON		0.104	0.878	11.27	0.27
GP-VTON		<u>0.081</u>	<u>0.884</u>	9.19	**0.09**
PAINT-BY-EXAMPLE	III	0.143	0.803	11.94	0.39
LADI-VTON		0.096	0.863	9.47	0.19
DCI-VTON		<u>0.081</u>	0.880	**8.76**	<u>0.11</u>
FLDM-VTON (**ours**)		**0.080**	**0.886**	<u>8.81</u>	0.13

4 Experiments

- Results on real-world data

4 Experiments

- Clothes-consistent results