Pro-Pose: Unpaired Full-Body Portrait Synthesis via Canonical UV Maps

Sandeep Mishra[1*†] Yasamin Jafarian[2] Andreas Lugmayr[2]
Yingwei Li[2‡] Varsha Ramakrishnan[2] Srivatsan Varadharajan[2]
Alan C. Bovik[1] Ira Kemelmacher-Shlizerman[2]

[1]The University of Texas at Austin, USA [2]Google, USA
https://pro-pose-portrait.github.io
sandy.mishra@utexas.edu, bovik@ece.utexas.edu
{jafarian, alugmayr, yingweili, vio, srivatsanv, kemelmi}@google.com

Fig. 1: Pro-Pose Canonicalization and Downstream Virtual Try-On. Given a single in-the-wild photo (left), Pro-Pose synthesizes high-fidelity, pose-controllable avatars (middle) driven by arbitrary SMPL-X [38] poses. By canonicalizing the subject into a minimal black outfit, our method preserves the user's visible identity – facial features, skin textures, and body shape, while stripping away the occlusions of the original clothing. This standardized representation provides an optimal, clean geometric canvas for downstream tasks like Virtual Try-On (right), enabling off-the-shelf VTO [9] models to apply novel garments without interference from the source image.

Abstract. Photographs of people taken by professional photographers typically present the person in beautiful lighting, with an interesting pose, and flattering quality. This is unlike common photos people take of themselves in uncontrolled conditions. In this paper, we explore how to canonicalize a person's "in-the-wild" photograph into a controllable, high-fidelity avatar—reposed in a simple environment with standardized minimal clothing. A key challenge is preserving the person's unique whole-body identity, facial features, and body shape while stripping away the complex occlusions of their original garments. While a large paired dataset of the same person in varied clothing and poses would simplify this, such data does not exist. To that end, we propose two key insights:

*Work done during an internship at Google. †Corresponding author. ‡Yingwei Li was affiliated with Google at the time of this research.

1) Our method transforms the input photo into a canonical full-body UV space, which we couple with a novel reposing methodology to model occlusions and synthesize novel views. Operating in UV space allows us to decouple pose from appearance and leverage massive unpaired datasets. 2) We personalize the output photo via multi-image finetuning to ensure robust identity preservation under extreme pose changes. Our approach yields high-quality, reposed portraits that achieve strong quantitative performance on real-world imagery, providing an ideal, clean biometric canvas that significantly improves the fidelity of downstream applications like Virtual Try-On (VTO).

Keywords: Pose-guided Person Image Synthesis, Virtual Try-On, Avatar Generation

1 Introduction

Images have become the foundation of digital identity, yet everyday photos are often captured under uncontrolled conditions with cluttered backgrounds, imperfect lighting, and arbitrary poses and clothing. Consequently, obtaining a clean, controllable "photoshoot-style" representation is difficult, primarily due to scarce high-quality training data. Existing reposing methods such as MCLD [26] and LEFFA [57] rely on the small-scale paired DeepFashion dataset [27] (≈100 identities), leading to overfitting and limited generalization to unseen individuals, and they cannot modify garments. In contrast, virtual try-on (VTO) models [11] typically rely on paired data of a source person and a target garment. Crucially, during training, they derive the person input from the target image and thus are trained to preserve the original pose rather than control or modify it. Consequently, neither method can jointly repose a person and generalize to real-world imagery - motivating our goal to generate photorealistic, reposed portraits that preserve full-body identity (face, body shape, and skin features) in standardized conditions. Beyond generating standalone digital avatars, this controlled canonical representation acts as a powerful intermediate foundation for downstream tasks like VTO. Current VTO models often suffer from unwanted structural leakage or geometric guidance from the subject's original clothing, which frequently leads to failures on unconstrained in-the-wild photos. By deliberately canonicalizing the subject into a minimal black tank top and shorts, Pro-Pose provides VTO pipelines with an effective, form-fitting 'blank canvas' that robustly preserves body shape and pose while eliminating the geometric interference of source garments. As shown in Figure 1, our approach produces a versatile portfolio of reposed, identity-faithful portraits from a single input.

In this paper, we address the limited identity diversity by leveraging both abundant single images [18,59,60] and scarce paired data [27]. We propose a self-supervised framework utilizing a canonical UV-space [38], which theoretically decouples pose from texture. However, practical extraction remains ill-posed, since the boundaries of the visible texture map inevitably leak pose information. To eliminate these clues, we introduce *Donor-based UV Reposing*, which uses occlusion masks from unrelated "donor" images to disguise the original visibility

patterns, forcing the model to learn robust geometric warping. This formulation enables seamless joint training on 30K paired samples and ≈470K unpaired single images. Finally, we employ Gemini 2.5 Flash Image [7] to standardize the garments in the target images, creating a **B**ase **C**lothing (**BC**) dataset featuring subjects in black tops and shorts without altering their identity or pose.

Our method employs a dual-branch training strategy to leverage both paired and unpaired data. The paired branch conditions the model on a SMPL-X [38] target, the partial UV texture map, and a source face crop to synthesize the subject in the standardized base garment. Conversely, for unpaired data, we utilize a self-supervised approach where the input also serves as the target. Here, we extract a donor-based UV map and drop out the face condition. This forces the network to reconstruct identity solely from the warped texture, preventing it from trivially copying input pixels. Finally, to address the overfitting risks inherent in single-shot personalization methods trained on limited data, we introduce a lightweight finetuning strategy. This adapts the pre-trained priors to new subjects using minimal reference images, ensuring high identity preservation without requiring large-scale multi-image datasets.

We summarize our contributions as follows: (1) a scalable joint training framework that integrates scarce paired data [27] with abundant unpaired single images [18, 59, 60] to overcome the limited identity diversity of existing benchmarks; (2) a novel self-supervised Donor-based UV Reposing mechanism that reduces coupling between pose and full-body texture to prevent boundary-based pose leakage, enabling effective learning from unpaired data; (3) a strategy to create a large-scale **B**ase **C**lothing (**BC**) dataset [18, 27], synthesized via Gemini 2.5 Flash Image [7] to ensure consistent visual conditions; (4) a subject-specific adaptation mechanism that utilizes our unified training objective to fine-tune LoRA layers at test-time, significantly reducing identity drift in challenging scenarios; and (5) extensive evaluations - including comparisons against Gemini 2.5 Flash Image [7], Gemini 3 Pro Image [10] and demonstrations of downstream VTO - showing that our approach achieves state-of-the-art, identity-faithful synthesis on challenging real-world imagery.

2 Related Work

2D Human Image Synthesis. Research in 2D human image synthesis has historically split into related tracks: pose-guided generation and disentangled attribute editing. Pose-guided methods aim to generate a novel pose of a person from a single image [23, 39, 44, 61]. To this aim, PG^2 [31] adopted a coarse-to-fine pipeline, while subsequent single-stage architectures improved end-to-end training by incorporating spatial alignment via deformable skip connections [44], progressive attention [61], or dense appearance flow [23]. Parallel work focused on disentangling attributes, such as appearance and shape [5], or attribute-specific editing [33, 54]. This line of work was further generalized by motion models for arbitrary object animation [43]. To reduce paired data needs, [42] introduced cyclic self-supervision. However, these 2D models struggle with spatial consistency and entangle pose with texture, leading to artifacts and identity loss.

3D-Aware Human Generation. To overcome spatial inconsistency, later methods incorporated 3D reasoning, implicit surface modeling, or neural rendering [14, 40, 41, 49, 56]. 3DHumanGAN [53] introduced a 3D-pose mapping module to enforce geometric plausibility, and EG3D [2] established efficient hybrid explicit/implicit 3D GANs for view-consistent generation. Other work such as PIFu [40] and PIFuHD [41] recovered pixel-aligned implicit fields, while ConTex-Human [8] focused on achieving texture consistency. More recent approaches have explored 3D Gaussian Splatting for real-time avatars [19, 34, 35] and self-supervised learning from social media videos [15, 16]. However, these approaches rely on strict multi-view or 3D supervision. While such datasets may be voluminous in terms of total frames, they remain severely limited in identity diversity compared to 2D images, preventing generalization to the distribution of real-world human appearances.

Diffusion-Based Animation. Recent diffusion-based techniques have significantly advanced fidelity and controllability [1,13,26,28,30,47,57]. Coarse-to-Fine Latent Diffusion (CFLD) [28] separates semantic and texture modeling; Animate Anyone [13] and DisCo [47] extend this paradigm to temporal consistency and disentangled control; and follow-up works such as MCLD [26] and LEFFA [57] enhance pose controllability. While high-quality, these models remain highly data-hungry, relying on large paired datasets like DeepFashion. Consequently, they are prone to overfitting and "identity drift" on unseen individuals.

Virtual Try-On and Garment Manipulation. Virtual try-on (VTO) methods transfer garments onto target bodies while preserving identity [11, 21, 36, 46, 50, 51, 59, 60]. VITON [11] and CP-VTON [46] pioneered geometric matching for in-shop clothing. M&M VTO [59] extends this with multi-garment diffusion control. However, these methods assume garment–person pairs rather than person-pose pairs, limiting reposing ability.

Summary. Existing 2D models lack consistency, 3D methods require expensive supervision, and diffusion animators rely on scarce paired data. To address these limitations, we propose a self-supervised framework formulated in UV space. By decoupling pose from texture and introducing donor-based reposing, we leverage abundant unpaired data alongside scarce paired data to generate photorealistic, reposed, and garment-neutral avatars from single images.

3 Method

We address the challenge of synthesizing highly realistic and pose-controllable human avatars by proposing a unified framework capable of learning from both abundant single-view images and scarce paired data. This integration is achieved through a novel supervision strategy formulated directly in canonical UV texture space, which naturally disentangles pose from appearance.

3.1 Overview and Problem Formulation

The fundamental goal is to generate an image, $\mathbf{A_p}$, of a specific person in a novel target pose $\mathbf{p}$ (represented as a SMPL-X [38] rendering), *standardized into*

minimal Base Clothing (BC) – a black sleeveless tank top and shorts. This canonicalization strips away source-garment occlusions while preserving the subject's identity, body shape, and skin appearance. Given a single reference image $\mathbf{I}_{\mathbf{p}'}$ of that person in a source pose $\mathbf{p}'$, we formulate this as a conditional generation problem:

$$\mathbf{A}_{\mathbf{p}} = f_\theta(\mathbf{I}_{\mathbf{p}'}, \mathbf{p}) \tag{1}$$

where f_θ is the parameterized model we aim to learn.

A standard approach to train such a model involves minimizing a reconstruction loss, such as $\|\mathbf{A}_{\mathbf{p}} - \mathbf{I}_{\mathbf{p}}\|$, where $\mathbf{I}_{\mathbf{p}}$ is a ground-truth image of the same person in the target pose $\mathbf{p}$. This requires a large dataset of pose-paired images $(\mathbf{I}_{\mathbf{p}'}, \mathbf{I}_{\mathbf{p}})$. However, collecting such data at scale presents significant challenges due to scarcity, limited identity diversity, and inconsistencies (e.g. clothing changes) between pairs.

These limitations motivate leveraging abundant unpaired single images. However, naively setting $\mathbf{p} = \mathbf{p}'$ leads to a trivial identity function. The central challenge is therefore to formulate a self-supervised manner that forces the model to learn meaningful representations from a single image.

3.2 Self-Supervision in Canonical UV Space

To enable effective learning from abundant single-view images, we shift our representation from image space to a canonical UV texture space. Ideally, we would extract a complete, identity-specific texture map $\mathbf{T}$ from any input image, perfectly decoupled from the subject's pose. This would allow us to formulate a unified generator g_θ for any target pose $\mathbf{p}$:

$$\mathbf{A}_{\mathbf{p}} = g_\theta(\mathbf{T}, \mathbf{p}) \tag{2}$$

In practice, single-view unwrapping is ill-posed due to self-occlusions. We can only obtain a *partial* texture map, $\mathbf{T}_{\mathbf{p}} = \mathbf{T} \odot \mathbf{M}_{\mathbf{p}}$, where $\mathbf{M}_{\mathbf{p}} \in \{0,1\}^{H \times W}$ is the binary visibility mask uniquely defined by pose $\mathbf{p}$ (see Figure 2 a). We must therefore approximate the generator using partial information:

$$\mathbf{A}_{\mathbf{p}} \approx g_\theta(\mathbf{T}_{\mathbf{p}}, \mathbf{p}) \tag{3}$$

Pose Leakage via Occlusion Boundaries. While Eq. 3 enables self-supervision, it introduces a critical challenge: the partial texture $\mathbf{T}_{\mathbf{p}}$ is highly correlated with the pose $\mathbf{p}$ via the mask $\mathbf{M}_{\mathbf{p}}$. As visualized in Figure 2 a, occlusion boundaries in $\mathbf{T}_{\mathbf{p}}$ perfectly align with the target pose. This allows the network to minimize reconstruction loss trivially by "copy-pasting" visible pixels based on these boundaries, rather than learning the desired 3D geometric warping.

Donor-Based UV Reposing To mitigate pose leakage, we must break the correlation between the input texture boundaries and the target pose. While

rendering the textured SMPL-X mesh into novel poses generates pseudo-pairs in image space (Figure 2 b), this process is computationally prohibitive for online training. We instead bypass explicit rendering by formulating the problem directly in UV space via *Donor-based UV Reposing* (Figure 2 c), an efficient 2D strategy that synthetically simulates novel pose occlusions.

We mask the input texture using visibility mask $\mathbf{M}_{\tilde{\mathbf{p}}}$ from a random 'donor' image with pose $\tilde{\mathbf{p}}$. Because the UV parameterization is canonical across all subjects, this donor mask can be sourced from *any* other person in the training set, rather than being synthetically generated; in practice we sample donors whose visibility masks share a 40–80% overlap (IoU) with the source (Sec. 3.6). This yields a hybrid texture $\mathbf{T}_{\mathbf{p}\to\tilde{\mathbf{p}}} = \mathbf{T}_{\mathbf{p}} \odot \mathbf{M}_{\tilde{\mathbf{p}}}$. Mathematically, this represents the source–donor visibility intersection: $\mathbf{T} \odot (\mathbf{M}_{\mathbf{p}} \odot \mathbf{M}_{\tilde{\mathbf{p}}})$. Because this intersection is commutative, the resulting texture boundaries no longer uniquely characterize the source pose $\mathbf{p}$. This ambiguity prevents trivial boundary-based leakage, forcing the generator to recover the original view via geometric inpainting:

$$\mathbf{A}_{\mathbf{p}} \approx g_\theta(\mathbf{T}_{\mathbf{p}\to\tilde{\mathbf{p}}}, \mathbf{p}) \tag{4}$$

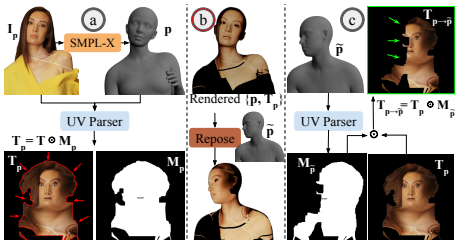

Fig. 2: Pose Leakage Mitigation. (a) Standard partial textures $\mathbf{T}_{\mathbf{p}}$ leak source pose information via occlusion boundaries, allowing trivial reconstruction shortcuts. (b) Generating pseudo-pairs via image-space rendering is computationally prohibitive for online training. Moreover, SMPL-X lacks fine details—such as hair—leading to unrealistic renderings. (c) Our **Donor-based Reposing** efficiently bypasses rendering by applying random donor masks $\mathbf{M}_{\tilde{\mathbf{p}}}$ directly in UV space. This simulates novel occlusions, preventing leakage and forcing the network to learn robust geometric warping. Note: To clearly visualize the subtle boundary differences (indicated by arrows), here we display a crop of just the face region from the full texture map; in practice, donor-reposing applies to all body parts.

3.3 Generative Framework

We parameterize our generator g_θ as a Latent Rectified Flow model [25], instantiated using the state-of-the-art Flux.1 [dev] transformer backbone [20]. Consistent with this architecture, we operate in the compressed 16-channel latent space defined by the pre-trained Flux Variational Autoencoder (VAE), comprising an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$.

Target images $\mathbf{I}_{\mathbf{p}}$ are compressed into latent representations $\mathbf{x}_0 = \mathcal{E}(\mathbf{I}_{\mathbf{p}})$. Following the Flow Matching formulation [25], we model the generative process as an Ordinary Differential Equation (ODE) [3] that transports samples from a standard Gaussian noise distribution $\mathbf{x}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to the data distribution $\mathbf{x}_0$.

We train a velocity prediction network v_θ to estimate the *vector field* driving this transport using the standard Flow Matching objective:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_1,\mathbf{c}} \left[\| v_\theta(\mathbf{x}_t, t, \mathbf{c}) - (\mathbf{x}_1 - \mathbf{x}_0) \|^2 \right] \tag{5}$$

where $t \sim \mathcal{U}[0,1]$ is the timestep, $\mathbf{c}$ is the condition set, and $\mathbf{x}_t = t \cdot \mathbf{x}_1 + (1-t) \cdot \mathbf{x}_0$ represents the linear interpolation between noise ($\mathbf{x}_1$) and data ($\mathbf{x}_0$).

Once trained, our high-level generator $g_\theta(\mathbf{c})$ (Eq. 4) is defined by numerically integrating this field backwards from noise ($t = 1$) to data ($t = 0$), followed by decoding:

$$g_\theta(\mathbf{c}) = \mathcal{D}\left(\mathbf{x}_1 + \int_1^0 v_\theta(\mathbf{x}_t, t, \mathbf{c})dt\right) \tag{6}$$

Conditioning Architecture. Following OminiControl [45], we inject the condition set $\mathbf{c}$ by concatenating the tokenized conditions (UV texture, SMPL-X pose render, and face crop) with the noisy image and text tokens along the sequence dimension, allowing them to interact through the DiT's multi-modal attention. We keep the pre-trained image- and text-stream parameters frozen and apply a LoRA mask so that updates only affect the condition tokens. We additionally inject rank-128 LoRA adapters into the attention projections ($\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{O}$) and the MLP layers of all DiT blocks. This adapts the frozen Flux.1 [dev] backbone to our task while training only a small fraction of the parameters.

3.4 Training Data and Conditioning Strategy

We enable joint training via a unified condition vector $\mathbf{c} = \{\mathbf{T}_{\text{in}}, \mathbf{p}_{\text{target}}, \mathbf{I}_{\text{in}}^{FC}\}$, comprising the input texture, target pose, and an optional identity-anchoring face crop extracted via MediaPipe [29]. An overview of our unified training framework is presented in Figure 3. We instantiate $\mathbf{c}$ based on the data source:

Paired Supervision For datasets with ground-truth pairs (reference $\mathbf{I}_{\mathbf{p}'}$, target $\mathbf{I}_{\mathbf{p}}$), we condition on the reference partial texture and face crop to reconstruct the target view ($\mathbf{x}_0 = \mathcal{E}(\mathbf{I}_{\mathbf{p}})$):

$$\mathbf{c}_{\text{paired}} = \{\mathbf{T}_{\mathbf{p}'}, \mathbf{p}, \mathbf{I}_{\mathbf{p}'}^{FC}\} \tag{7}$$

The face crop $\mathbf{I}_{\mathbf{p}'}^{FC}$ is included to boost fidelity in facial regions where the partial texture $\mathbf{T}_{\mathbf{p}'}$ may lack resolution.

Single-View Self-Supervision For single-view images ($\mathbf{I}_{\mathbf{p}}$), we employ Donor-based Reposing (Eq. 4). We condition on the donor-masked texture $\mathbf{T}_{\mathbf{p}\to\tilde{\mathbf{p}}}$ and the target pose. Crucially, we explicitly drop the face crop condition ($\mathbf{I}_{\text{in}}^{FC} = \emptyset$) to prevent pixel-perfect information leakage:

$$\mathbf{c}_{\text{single}} = \{\mathbf{T}_{\mathbf{p}\to\tilde{\mathbf{p}}}, \mathbf{p}, \emptyset\} \tag{8}$$

This constraint forces the network to recover identity from the warped texture, ensuring robust geometric learning.

In both cases the supervision target (and its rendered target pose) is the BC-standardized image, while the input texture comes from the original image. For paired data these are two different images of the same person, whereas for single images the original and its own BC version form the pair, with the face crop dropped to prevent copy-paste shortcutting.

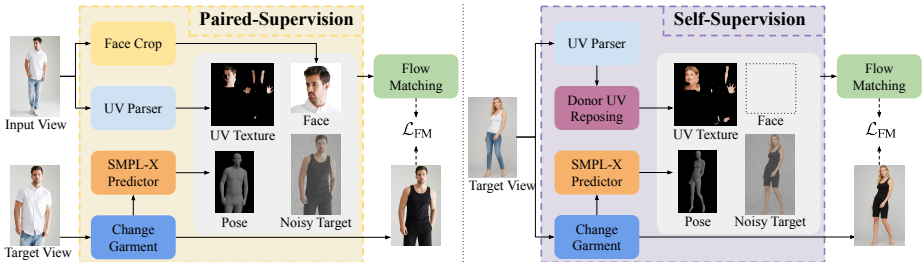

Fig. 3: Overview of our Avatar Generation Framework. Our approach leverages single-view datasets by operating in a canonical UV space, extracting UV texture and pose [38]. **Left (Paired Supervision):** When ground-truth pose pairs are available, we condition the Flow Matching model on the partial UV texture, target pose, and face crop. **Right (Single-View Self-Supervision):** To prevent "pose leakage" from occlusion boundaries when training on single images, we introduce a **Donor-based UV Reposing** module (Sec. 3.2). This synthetically re-poses the input texture using a random donor visibility mask, forcing the model to learn robust identity representations. Furthermore, we drop-out the face crop condition in this branch to prevent trivial reconstruction via pixel-perfect information leakage.

3.5 Test-Time Personalization via Few-Shot Adaptation

While our method already achieves high identity preservation in a feed-forward manner, we can further enhance fidelity through test-time personalization via few-shot adaptation. Given a small set $\mathcal{S}$ comprising N images of a person in various poses, we create pairs by using one view as the reference and another as the target. We then fine-tune our LoRA adapters using the paired objective, as illustrated in Figure 4. This personalizes the model to the specific identity, improving consistency, particularly for extreme pose generation.

Fig. 4: Finetuning pipeline. We build input–target pairs from a few-shot subject set and apply a facially masked paired Flow Matching loss to personalize the model at test time.

We apply the personalization loss to the full visible foreground using a binary mask $\mathbf{M}_j$, which represents the target's latent-space skin segmentation mask:

$$\mathcal{L}_{\text{FT}} = \mathbb{E}_{t,(i,j)} \left[\| \mathbf{M}_j \odot \left(v_\theta(\mathbf{x}_{j,t}, t, \mathbf{c}_{\text{paired}}) - v_{\text{target}} \right) \|^2 \right] \tag{9}$$

where $\mathbf{c}_{\text{paired}} = \{\mathbf{T}_i, \mathbf{p}_j, \mathbf{I}_i^{FC}\}$ is the conditions extracted from a sample pair (i, j) drawn from $\mathcal{S}$ (such that $i \neq j$), and $v_{\text{target}} = \mathbf{x}_{j,1} - \mathbf{x}_{j,0}$ is the target transport guidance.

3.6 Implementation Details

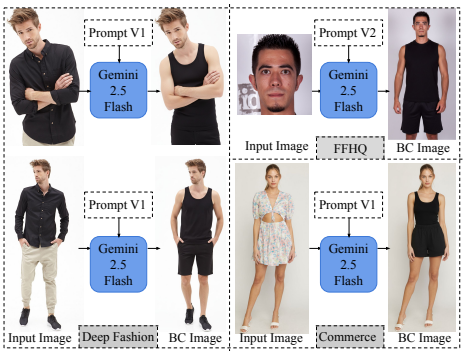

Fig. 5: Base Clothing (BC) Standardization. We apply different strategies based on the data. For DeepFashion [27] and Commerce images [59, 60], we generate the base garment while preserving pose and identity via pixel-aligned editing (Prompt V1). For FFHQ [18], we use generative outpainting to expand limited face crops into full-body samples (Prompt V2).

Base Clothing (BC) Dataset. We construct a standardized dataset by processing paired DeepFashion [27] data, unpaired FFHQ [18] faces, and commercial images [59,60] using Gemini 2.5 Flash Image [7]. As shown in Figure 5, we employ distinct strategies: *pixel-aligned editing* for full-body inputs to standardize garments to a black tank top and shorts, and *generative outpainting* to expand face crops into full-body portraits. The prompts used for this standardization are provided in the Appendix. While we do not explicitly handle target-pose lighting, BC standardization either re-lights subjects into even studio shading or matches the source illumination; by distilling this data, the model learns to attenuate baked-in highlights from the input texture and render lighting consistent with the input, yielding realistic illumination under novel poses.

Training Setup. Our dataset totals approximately 500k samples: 470K single unpaired images [18, 59, 60], and 30K paired images from the DeepFashion dataset [27] (We restrict training to a subset of DeepFashion data to prevent overfitting to its limited number of unique identities (≈ 100), while ensuring all identities are still represented.)

Optimization. We use Flux.1 [dev] [20] as our backbone, training rank-128 LoRA adapters with AdamW with a learning rate of 10^{-4}. Training runs for 100K iterations on 128 TPUv5 chips with a batch size of 128. At inference, our model synthesizes a single reposed avatar in approximately 51 seconds.

Donor Pool & Dropout. For every single-view sample, we pre-compute a pool of 10 valid donor masks $\mathbf{M}_{\tilde{\mathbf{p}}}$ that satisfy an IoU constraint of $[0.4, 0.8]$ with the source mask. To prevent modality over-reliance, we apply mutually exclusive conditioning dropout: we either drop all conditions ($p = 0.05$), or individual components: $\mathbf{T}_{\text{in}}$ ($p = 0.3$), $\mathbf{I}_{\text{in}}^{FC}$ ($p = 0.3$), or $\mathbf{p}_{\text{target}}$ ($p = 0.1$).

4 Experiments

In this section, we lay out the details of our evaluation strategy, including datasets, baselines, and metrics used to compare our method against state-of-the-art approaches. We also conduct extensive ablations to highlight the effectiveness of each of our contributions.

4.1 Datasets and Metrics for Evaluation

Following prior work [26, 28, 57], we use the DeepFashion In-Shop Clothes Retrieval test split [27] (8570 image pairs). To assess generalization to in-the-wild scenarios, we additionally evaluate on WPose dataset [22] (2305 image pairs).

Since our method canonicalizes clothing, direct comparison against ground-truth images (which include diverse garments and backgrounds) is not meaningful. For image similarity (PSNR, SSIM [48]) and perceptual metrics (FID [12], LPIPS [55]), we therefore compare our outputs with the BC version. We further report these metrics on the original image only face regions in the appendix.

Pose, Identity, Semantic and Perceptual Metrics. To measure how well the generated avatar matches the target **pose**, we compute Object Keypoint Similarity (OKS) [24] between the predicted and ground-truth keypoints. **Identity** fidelity is evaluated within the facial region. We compute Face Similarity (FaceSim) using cosine similarity of ArcFace [4] identity embeddings. We report DINOv2 similarity [37] as a measure of **semantic** and structural alignment. For reference-free **perceptual** quality, we report HPSv3 [32].

4.2 Quantitative and Qualitative Comparison

We compare our method with recent diffusion-based pose-conditioned avatar generation models, including MCLD [26], CFLD [28], LEFFA [57], OnePose-Trans [6], UniHuman [22], Gemini 2.5 Flash Image (Nano Banana) [7], and Gemini 3 Pro Image (Nano Banana Pro) [10].

As detailed in Table 1, our method achieves SOTA performance on both benchmarks. On the in-domain DeepFashion set [27], our model sets a new SOTA, significantly outperforming all baselines across image fidelity (PSNR, SSIM, LPIPS), identity (FaceSim), and perceptual quality (HPSv3). For the in-the-wild WPose dataset [22], where we use foreground-masked metrics (M-PSNR, M-SSIM, M-LPIPS), our full model (*Unpaired + Paired*) conclusively dominates all SOTA baselines. This highlights the critical role of our unpaired data strategy in achieving robustness for challenging, in-the-wild data.

Figure 12 provides a qualitative comparison with SOTA methods. As demonstrated, our method generates high-fidelity avatars that accurately preserve the input person's identity, facial features, and body characteristics. For instance, the eighth row shows robust preservation of face and beard shape. Furthermore, our approach more robustly follows the target pose than any other method (e.g., sixth row) and maintains body shape fidelity relative to the original image (e.g., ninth row).

Stratified Identity Analysis. To quantify identity robustness under varying difficulty, we stratify FaceSim scores by the magnitude of pose change between source and target. As shown in Figure 6, Pro-Pose closely tracks the inherent identity variance of Ground Truth images across all pose-change bins, whereas all baselines exhibit rapid degradation as pose difficulty increases. Defining identity breakage as a FaceSim score below 0.4, Pro-Pose exhibits a failure rate of only 13.5%, while all competing methods (LEFFA, MCLD, UniHuman, OnePose-Trans) fail to preserve identity in over 97% of generated samples.

DeepFashion (In-Domain)								
Method	PSNR↑	FID↓	SSIM↑	LPIPS↓	OKS↑	FaceSim↑	DINO↑	HPSv3↑
CFLD [28]	17.65	7.15	0.748	0.182	**0.48**	0.3180	0.9731	4.15
MCLD [26]	18.21	7.08	0.756	0.176	**0.49**	0.3440	0.9654	4.29
LEFFA [57]	14.02	**4.23**	0.755	0.119	0.44	0.5794	0.9409	4.41
OnePoseTrans [6]	13.57	8.74	0.605	0.307	0.46	0.5750	0.9476	4.32
UniHuman [22]	14.05	6.25	0.796	0.156	0.46	0.5810	0.9434	4.03
Gemini 2.5 Flash Image [7]	16.98	4.59	0.738	0.179	0.43	0.5815	0.9691	7.19
Gemini 3 Pro Image [10]	17.51	4.30	0.775	0.109	0.45	0.5856	0.9705	7.22
Unpaired Only	15.66	6.54	0.715	0.201	0.47	0.3585	0.9259	4.25
Paired Only	**19.38**	**4.19**	**0.815**	**0.071**	**0.48**	**0.6255**	**0.9761**	**7.25**
Ours (Unpaired + paired)	**19.36**	4.24	**0.818**	**0.075**	**0.48**	**0.6047**	**0.9759**	**7.24**
WPose (Out-of-Domain)								
Method	M-PSNR↑	FID↓	M-SSIM↑	M-LPIPS↓	OKS↑	FaceSim↑	DINO↑	HPSv3↑
CFLD [28]	15.43	96.07	0.744	0.208	0.31	0.0885	0.6412	1.94
MCLD [26]	15.64	94.23	0.759	0.201	0.35	0.0995	0.6478	1.96
LEFFA [57]	16.71	67.85	0.776	0.193	0.32	0.0914	0.5725	2.01
OnePoseTrans [6]	17.23	27.43	0.818	0.151	0.33	0.1735	0.7205	4.44
UniHuman [22]	17.64	27.75	0.807	0.161	0.34	0.1121	**0.7207**	2.89
Gemini 2.5 Flash Image [7]	16.67	9.55	0.779	0.149	0.32	0.4713	0.7005	7.35
Gemini 3 Pro Image [10]	17.19	7.15	0.795	**0.145**	0.33	**0.5241**	0.7119	**7.4**
Unpaired Only	16.13	6.95	0.761	0.215	**0.37**	0.3805	0.6779	4.37
Paired Only	**18.30**	**6.65**	**0.820**	0.155	0.34	0.4959	0.6972	7.30
Ours (Unpaired + paired)	**19.95**	**5.99**	**0.860**	**0.121**	**0.38**	**0.5571**	**0.7394**	**7.55**

Table 1: Quantitative evaluation and ablation study. We report results on DeepFashion (top) and WPose (bottom). The upper section of each block compares our method against state-of-the-art baselines, while the lower section details our ablation study (Unpaired Only, Paired Only, and our full hybrid model). All Pro-Pose results are zero-shot (without test-time fine-tuning). **Red** indicates the best performance and **Blue** indicates the second best across all methods.

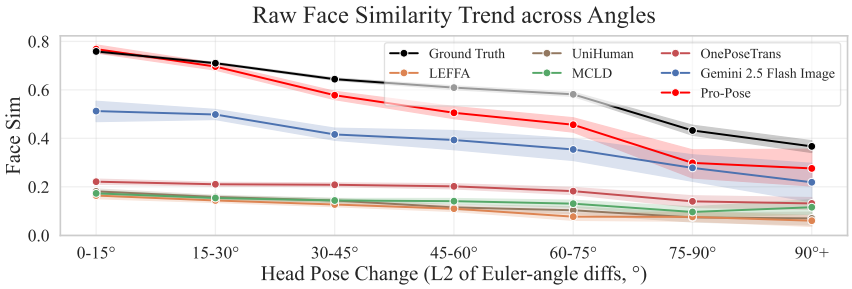

Fig. 6: Identity Preservation vs. Pose Difficulty. FaceSim scores stratified by pose-change magnitude. Pro-Pose closely tracks Ground Truth identity variance across all difficulty bins, while baselines degrade rapidly. With identity breakage defined as FaceSim < 0.4, Pro-Pose fails in only 13.5% of cases vs. $> 97\%$ for all baselines.

4.3 Ablation Study

Hybrid Data Strategy. We validate our hybrid data strategy by training three variants of our model: (i) *Unpaired Only*, trained exclusively on unpaired single images; (ii) *Paired Only*, trained exclusively on paired images; and (iii) *Unpaired + Paired*, our full model.

As shown in Table 1, while the *Paired Only* model performs well on in-domain data (DeepFashion [27]), it suffers on the out-of-domain WPose set [22],

Fig. 7: Ablation Study of Training Data Sources. We evaluate the impact of different training dataset combinations on generated avatar quality. The result columns display models trained on Unpaired data only, Paired data only, and the full Unpaired + Paired combination, respectively. One can observe the improvement in fidelity with the full combined dataset.

confirming its poor generalization. Figure 7 demonstrates that training solely on the limited paired dataset leads to overfitting and identity leaking from the training data. Alternatively, the model trained exclusively on unpaired data fails when subjected to significant pose changes. Our full method (*Unpaired + Paired*) synergizes these sources, effectively utilizing the larger dataset to learn robust identity priors while maintaining stable pose guidance.

Test-Time Personalization.

Furthermore, we validate our test-time personalization strategy, detailed in Section 3.5. As shown in Table 2 and qualitatively in Figure 8, applying this few-shot adaptation further improves performance over our base model across all metrics. The most significant gain is in identity preservation, with FaceSim improving by 18.3%. This confirms that our finetuning approach is highly effective at enhancing model personalization and identity fidelity for specific subjects. We share more finetuning results in the Appendix.

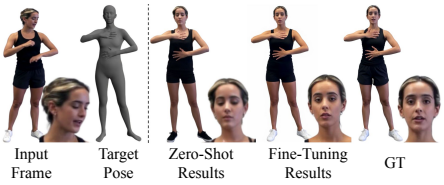

Fig. 8: Finetuning Qualitative results. Although the input image suffered from motion blur, leading to blurry base model results, finetuning successfully restored crisp face features, demonstrating the model's ability to learn robustly from multiple images rather than relying solely on one input.

Method	PSNR	FID	SSIM	FaceSim	DINO	HPSv3
Ours (Zero-Shot)	18.59	6.88	0.823	0.4837	0.689	7.15
Ours (Fine-Tuned)	**19.34**	**6.76**	**0.835**	**0.5722**	**0.705**	**7.17**

Table 2: Ablation study (Fine-Tuning). Test-time fine-tuning with a few reference images improves identity preservation over our zero-shot base model across all metrics.

4.4 Downstream Application: Virtual Try-On

A key advantage of our canonical representation is its utility as a preprocessing stage for downstream tasks. We demonstrate this by combining Pro-Pose with the state-of-the-art Google Virtual Try-On tool [9]. As shown in Figure 9, applying VTO directly to in-the-wild images can lead to erroneous results due to challenging poses, cluttered backgrounds, and complex clothing. In contrast, first generating a clean, reposed canonical avatar with Pro-Pose significantly improves VTO fidelity.

Fig. 9: Downstream Application: Virtual Try-On. Applying VTO [9] directly to an unconstrained in-the-wild image (VTO on RI) yields limited fidelity. Using Pro-Pose to first generate a clean, reposed canonical avatar (Pro-Posed RI) significantly improves VTO quality (VTO on PRI).

To fairly quantify this benefit, we conducted a user study on paired WPose samples in which both branches target the same pose: Branch-A applies VTO directly to the original image, while Branch-B applies VTO to the Pro-Posed avatar. Across 30K votes (10 raters over 3K pairs; 500 images × 6 garments), raters preferred VTO on Pro-Posed avatars in 71.98% of comparisons, with substantially fewer garment errors (Table 3).

Error / Method	Orig.	Pro-Pose
VTO pref. ↑	28.02%	**71.98%**
Garment err. ↓	31.04%	**17.34%**
Person err. ↓	**2.51%**	3.02%

Table 3: VTO User Study. VTO on Pro-Posed avatars is strongly preferred over VTO on originals under a matched pose. The slight person-error uptick is attributed to SMPL-X limitations (Sec. 5).

4.5 Emergent Model Properties: Direct Texture Editing.

Our framework unlocks other powerful features without explicit training such as direct texture editing. Because Pro-Pose extracts a canonical UV map, any manual edits to this texture (e.g., adding a tattoo) propagate with strict geometric consistency to the final reposed avatar, preserving identity without distortion (Figure 10 adding a heart tattoo on texture map can transfer to the reposed person).

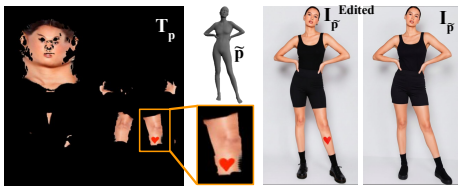

Fig. 10: Emergent Capabilities. T_p: Edited texture map. By introducing a localized modification — in this case, adding a heart tattoo to the base texture — the edit geometrically propagates with high fidelity to novel target poses $\tilde{p}$ as shown in $I_{\tilde{p}}^{Edited}$. $I_{\tilde{p}}$ shows the output using the unedited texture map. This demonstrate our model's capability to apply texture changed such as make up or tattoo in synthesizing the output portraits.

5 Limitations and Future Work

While our framework significantly advances single-image avatar reposing,

some limitations remain. First, the base model struggles with extreme pose changes, such as synthesizing a frontal pose from a single rear-view image. Consequently, the model inevitably hallucinates facial features (Figure 11, left). However, this limitation can be mitigated by providing multiple reference images of the subject and applying our fine-tuning approach. Second, our method relies heavily on the accuracy of the underlying SMPL-X body estimation. During zero-shot in-the-wild evaluation, we extract body shape parameters exclusively from the single reference image and combine them with the pose parameters of the target. However, recovering 3D body shape from a single image can be erroneous. Figure 11 (right) showcases an example from the WPose dataset where the reference body shape is inaccurately estimated. When this flawed shape geometry is reposed into a novel target view, the errors propagates directly to the generated canonical avatar and resulting in incorrect body shape. This highlights that our synthesis fidelity is ultimately bounded by the robustness of single-view 3D body priors. To mitigate such failures in practice, we leverage the pose-conditioning dropout used during training to drop unreliable pose estimates, or replace them with the $\hat{\beta}$ of a similar-bodied reference subject; this reduces gross geometric artifacts at the cost of exact body-shape fidelity. Third, while our full-body UV conditioning preserves visible skin details, self-occluded body regions cannot be faithfully reconstructed from a single view. Future work will explore more robust body estimation techniques and extend the framework to zero-shot multi-reference consistency.

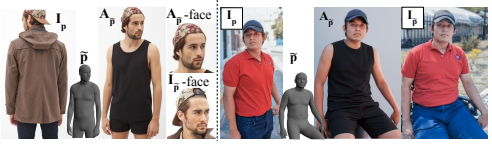

Fig. 11: Limitations of the base model. Left: Extreme pose changes (e.g., back-to-front) result in minimal UV texture overlap, forcing the model to hallucinate unseen regions. While general attributes like beard shape or hair color may be inferred from partial cues, precise facial structure can lose fidelity. **Right:** Zero-shot synthesis is constrained by the accuracy of single-view 3D body estimation. Errors in SMPL-X fitting can propagate during the reposing stage, leading to inaccurate body proportions (e.g., a noticeable reduction in thigh volume in the ProPose results compared to the ground truth).

6 Conclusion

In this paper, we presented a novel framework for generating photo-realistic, reposed human avatars from unconstrained single images. By introducing a self-supervised *Donor-based UV Reposing* strategy operating on full-body UV textures, we effectively decoupled pose from appearance, enabling our Flow Matching generator to learn robust full-body identity preservation from massive unpaired image collections. We further showed that this strong base model can be rapidly specialized via multi-image fine-tuning to handle more extreme reposing scenarios. By standardizing data into a base clothing and formulating generation in canonical UV space, our approach provides a scalable foundation for creating high-fidelity, controllable digital avatars from everyday photography, and serves as a powerful canonical intermediate for downstream applications such as virtual try-on.

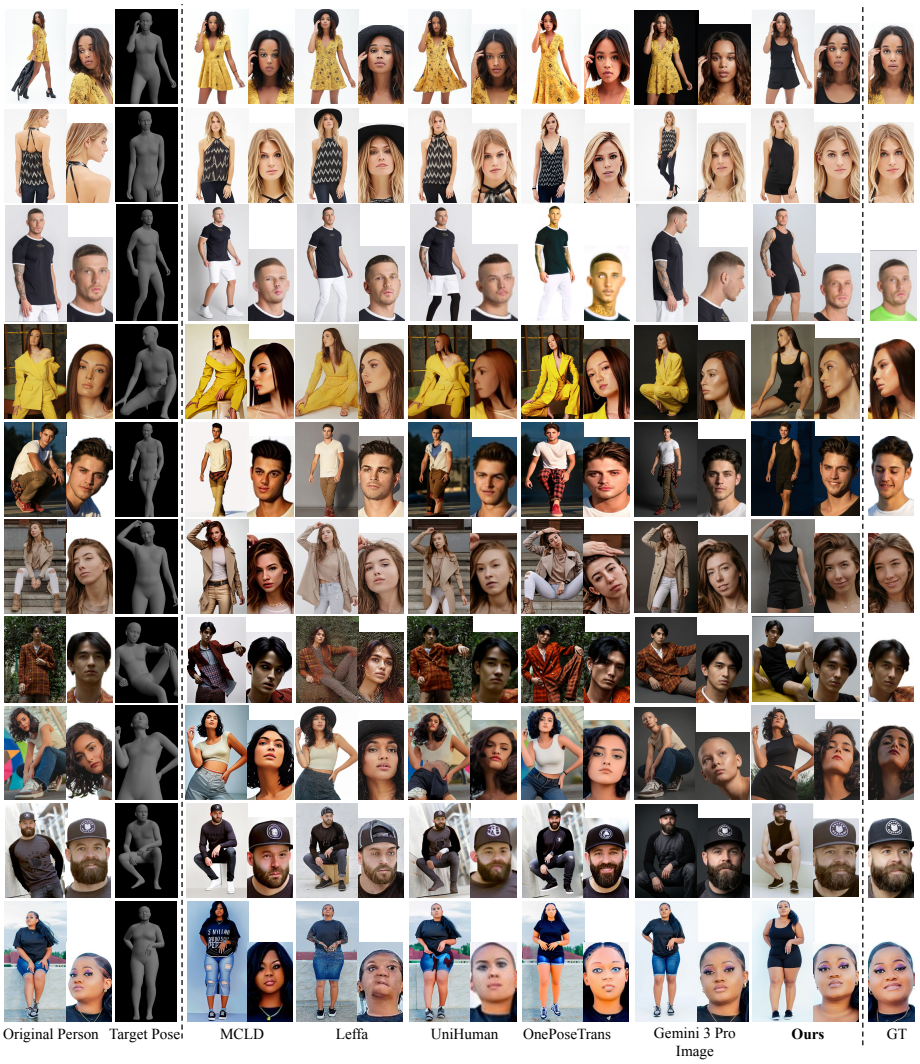

Fig. 12: Qualitative Comparison. Results are shown for the DeepFashion [27] (rows 1-3) and WPose [22] (bottom rows) datasets. The columns show the Original Person, Target Pose, and results from state-of-the-art methods (MCLD [26], Leffa [57], UniHuman [22], OnePoseTrans [6], Gemini 3 Pro Image [10]) compared to **Ours (Zero-Shot)** and the Ground Truth (GT). Our method demonstrates superior performance in preserving the person's identity, particularly the facial characteristics, across varying poses. For instance, in the third and fifth rows, previous methods introduce significant identity distortion and loss of facial likeness (e.g., changes in jawline or features). In contrast, **Ours** robustly preserves the unique face structure and identity of the original person, closely matching the GT result. (prompt used for Gemini 3 Pro Image: `A professional studio portrait of the person from IMAGE_1, maintaining their exact facial features, hairstyle, build, and identity. They are captured in the pose and camera angle shown in IMAGE_2`).

References

1. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: CVPR (2023)
2. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., Mello, S.D., Gallo, O., Guibas, L., Tremblay, J., Khamis, S., Karras, T., Wetzstein, G.: Efficient geometry-aware 3D generative adversarial networks. In: CVPR (2022)
3. Chen, R.T.Q., Rubanova, Y., Bettencourt, J., Duvenaud, D.K.: Neural ordinary differential equations. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018), https://proceedings.neurips.cc/paper_files/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf
4. Deng, J., Guo, J., Yang, J., Xue, N., Kotsia, I., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 5962–5979 (Oct 2022). https://doi.org/10.1109/tpami.2021.3087709, http://dx.doi.org/10.1109/TPAMI.2021.3087709
5. Esser, P., Sutter, E., Ommer, B.: A variational u-net for conditional appearance and shape generation. In: CVPR (2018)
6. Fan, D., Chen, T., Wang, M., Ma, R., Tang, Q., Yi, Z., Wang, Q., Chang, L.: One-shot learning for pose-guided person image synthesis in the wild. In: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2025). https://doi.org/10.1109/ICASSP49660.2025.10890784
7. Fortin, A., Vernade, G., Kampf, K., Reshi, A.: Introducing gemini 2.5 flash image, our state-of-the-art image model. Google for Developers Blog (Aug 2025), https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/, accessed: 2025-11-04
8. Gao, X., Li, X., Zhang, C., Zhang, Q., Cao, Y., Shan, Y., Quan, L.: Contex-human: Free-view rendering of human from a single image with texture-consistent synthesis. In: CVPR (2024)
9. Google: How the Google Try-On tool works. https://support.google.com/googleshopping/answer/16253678 (2023), accessed: 2026-03-02
10. Google: Gemini 3 pro image. Gemini API Documentation (Nov 2025), https://ai.google.dev/gemini-api/docs/models/gemini-3-pro-image, accessed: 2026-06-27
11. Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: Viton: An image-based virtual try-on network. In: CVPR (2018)
12. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
13. Hu, L., Gao, X., Zhang, P., Sun, K., Zhang, B., Bo, L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In: CVPR (2024)
14. Huang, Z., Xu, Y., Lassner, C., Li, H., Tung, T.: ARCH: Animatable Reconstruction of Clothed Humans . In: CVPR (2020)
15. Jafarian, Y., Park, H.: Self-supervised 3d representation learning of dressed humans from social media videos. IEEE Transactions on Pattern Analysis & Machine Intelligence (2022). https://doi.org/10.1109/TPAMI.2022.3231558

16. Jafarian, Y., Park, H.S.: Learning high fidelity depths of dressed humans by watching social media dance videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12753–12762 (June 2021)
17. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
18. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019)
19. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics (2023)
20. Labs, B.F.: Flux. https://github.com/black-forest-labs/flux (2024), accessed: 2026-06-30
21. Lee, S., Gu, G., Park, S., Choi, S., Choo, J.: High-resolution virtual try-on with misalignment and occlusion-handled conditions. arXiv preprint arXiv:2206.14180 (2022)
22. Li, N., Liu, Q., Singh, K.K., Wang, Y., Zhang, J., Plummer, B.A., Lin, Z.: Unihuman: A unified model for editing human images in the wild. In: CVPR (2024)
23. Li, Y., Huang, C., Loy, C.C.: Dense intrinsic appearance flow for human pose transfer. In: CVPR (2019)
24. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
25. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling (2023), https://arxiv.org/abs/2210.02747
26. Liu, J., Zhang, J., Rota, P., Sebe, N.: Multi-focal conditioned latent diffusion for person image synthesis. In: CVPR (2025)
27. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: CVPR (June 2016)
28. Lu, Y., Zhang, M., Ma, A.J., Xie, X., Lai, J.H.: Coarse-to-fine latent diffusion for pose-guided person image synthesis. In: CVPR (2024)
29. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M., Lee, J., Chang, W.T., Hua, W., Georg, M., Grundmann, M.: Mediapipe: A framework for perceiving and processing reality. In: Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019 (2019), https://mixedreality.cs.cornell.edu/s/NewTitle_May1_MediaPipe_CVPR_CV4ARVR_Workshop_2019.pdf
30. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Patterm Recognition (CVPR) (2022)
31. Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., Van Gool, L.: Pose guided person image generation. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)
32. Ma, Y., Wu, X., Sun, K., Li, H.: Hpsv3: Towards wide-spectrum human preference score. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15086–15095 (October 2025)
33. Men, Y., Mao, Y., Jiang, Y., Ma, W.Y., Lian, Z.: Controllable person image synthesis with attribute-decomposed gan. In: CVPR (2020)
34. Moreau, A., Song, J., Dhamo, H., Shaw, R., Zhou, Y., Pérez-Pellitero, E.: Human gaussian splatting: Real-time rendering of animatable avatars. arXiv:2311.17113 [cs.CV] (2023)

35. Moreau, A., Song, J., Dhamo, H., Shaw, R., Zhou, Y., Pérez-Pellitero, E.: Human gaussian splatting: Real-time rendering of animatable avatars. In: CVPR (2024)
36. Morelli, D., Fincato, M., Cornia, M., Landi, F., Cesari, F., Cucchiara, R.: Dress Code: High-Resolution Multi-Category Virtual Try-On. In: Proceedings of the European Conference on Computer Vision (2022)
37. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
38. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: CVPR (2019)
39. Ren, Y., Yu, X., Chen, J., Li, T.H., Li, G.: Deep image spatial transformation for person image generation. In: CVPR (2020)
40. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: ICCV (2019)
41. Saito, S., Simon, T., Saragih, J., Joo, H.: Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In: CVPR (2020)
42. Sanyal, S., Vorobiov, A.E., Bolkart, T., Loper, M., Mohler, B.J., Davis, L.S., Romero, J., Black, M.J.: Learning realistic human reposing using cyclic self-supervision with 3d shape, pose, and appearance consistency. In: ICCV (2021)
43. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2019)
44. Siarohin, A., Sangineto, E., Lathuilière, S., Sebe, N.: Deformable gans for pose-based human image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
45. Tan, Z., Liu, S., Yang, X., Xue, Q., Wang, X.: Ominicontrol: Minimal and universal control for diffusion transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2025)
46. Wang, B., Zheng, H., Liang, X., Chen, Y., Lin, L., Yang, M.: Toward characteristic-preserving image-based virtual try-on network. In: ECCV (2018)
47. Wang, T., Li, L., Lin, K., Lin, C.C., Yang, Z., Zhang, H., Liu, Z., Wang, L.: Disco: Disentangled control for referring human dance generation in real world. In: CVPR (2024)
48. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
49. Xiu, Y., Yang, J., Tzionas, D., Black, M.J.: ICON: Implicit Clothed humans Obtained from Normals. In: CVPR (2022)
50. Xu, Y., Gu, T., Chen, W., Chen, C.: Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. AAAI (2025)
51. Yang, H., Zhang, R., Guo, X., Liu, W., Zuo, W., Luo, P.: Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In: CVPR (2020)
52. Yang, X., Taketomi, T., Endo, Y., Kanamori, Y.: Freeuv: Ground-truth-free realistic facial uv texture recovery via cross-assembly inference strategy. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 326–337 (2025)
53. Yang, Z., Li, S., Wu, W., Dai, B.: 3dhumangan: 3d-aware human image generation with 3d pose mapping. In: ICCV (2023)
54. Zhang, J., Li, K., Lai, Y.K., Yang, J.: PISE: Person image synthesis and editing with decoupled gan. In: CVPR (2021)

55. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
56. Zhao, F., Yang, W., Zhang, J., Lin, P.Y., Zhang, Y., Yu, J., Xu, L.: Humannerf: Efficiently generated human radiance field from sparse inputs. In: CVPR (2022)
57. Zhou, Z., Liu, S., Han, X., Liu, H., Ng, K.W., Xie, T., Cong, Y., Li, H., Xu, M., Pérez-Rúa, J.M., Patel, A., Xiang, T., Shi, M., He, S.: Learning flow fields in attention for controllable person image generation. In: CVPR (2025)
58. Zhu, H., Wu, W., Zhu, W., Jiang, L., Tang, S., Zhang, L., Liu, Z., Loy, C.C.: CelebV-HQ: A large-scale video facial attributes dataset. In: ECCV (2022)
59. Zhu, L., Li, Y., Liu, N., Peng, H., Yang, D., Kemelmacher-Shlizerman, I.: M&m vto: Multi-garment virtual try-on and editing. In: CVPR (2024)
60. Zhu, L., Yang, D., Zhu, T., Reda, F., Chan, W., Saharia, C., Norouzi, M., Kemelmacher-Shlizerman, I.: Tryondiffusion: A tale of two unets. In: CVPR (2023)
61. Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., Bai, X.: Progressive pose attention transfer for person image generation. In: CVPR (2019)

Supplementary Material

A Base Clothing (BC) Dataset.

To train our model, we utilize paired data from DeepFashion [27], unpaired single images from FFHQ [18], and commercial data derived from e-commerce websites [59,60]. To ensure visual uniformity across these diverse sources, we generate the **Base Clothing (BC)** dataset using an identity-preserving I2I model (Gemini 2.5 Flash Image [7]), as shown in Figure 5 (main paper). Depending on the nature of the input data, we employ either Prompt V1 or Prompt V2:

Prompt V1:

```
Precisely edit IMAGE_0 to change the garment worn by the person in
the photo to a black sleeveless tank top and 3-inch black shorts.
Do not change the person in the image or the framing. Remove or
replace any accessories such as bags, hats, jewelry and sunglasses
in IMAGE_0. Do not alter IMAGE_0 in any other way. Keep it
pixel-aligned.
```

Prompt V2:

```
Precisely edit IMAGE_0 to generate a full body image of this
person (head to toes) while keeping the face and head pose same in
a black sleeveless tank top and 3-inch black shorts. Generate the
body pose suiting the existing face pose. The person is posing
naturally against a clean and plain background. Even, soft studio
lighting illuminates the subject from the front, highlighting the
form and texture of the garments. No accessories (no bags, hats,
jewelry, sunglasses). Professional product photography style, no
harsh shadows or distracting elements. High resolution, sharp
focus.
```

B Canonical UV-Space Formulation

To formulate our generative objective in the canonical UV space, we first extract the parametric body model and the corresponding partial texture map $\mathbf{T_p}$ from the input image $\mathbf{I}$.

Pose and Shape Estimation. We utilize the SMPL-X body model [38] to represent the human subject. Given an input image $\mathbf{I}$, we estimate the shape $\hat{\beta}$, pose $\hat{\theta}$, and expression $\hat{\psi}$ parameters via SMPLify-X optimization [38]. The pose parameters $\hat{\theta}$ are the control variable for the target pose. The SMPL-X function $M(\hat{\beta}, \hat{\theta}, \hat{\psi})$ maps these parameters to a triangular mesh $\mathcal{M} = (\mathbf{V}, \mathcal{F})$, where $\mathbf{V} \in \mathbb{R}^{N\times 3}$ represents the vertices and $\mathcal{F}$ the fixed topology faces. We obtain the target pose condition $\mathbf{p}$ by rendering the estimated mesh $\mathcal{M}$ into the image space.

Partial Texture Extraction. To obtain the partial texture $\mathbf{T}_p$, we perform an inverse rendering operation via barycentric sampling in UV space. Let $\mathbf{V}^{uv} \in$

$\mathbb{R}^{N\times 2}$ denote the fixed UV coordinates of the SMPL-X topology, and $\mathbf{V}^{img} \in \mathbb{R}^{N\times 2}$ denote the projection of the 3D vertices $\mathbf{V}$ onto the coordinate space of input image $\mathbf{I}$.

Computing $\mathbf{T_p}$ directly via ray-casting is computationally expensive. Instead, we adopt an efficient rasterization-based approach. We rasterize the mesh in UV space, utilizing $\mathbf{V}^{uv}$ as the position attributes. During rasterization, we interpolate the image-space coordinates $\mathbf{V}^{img}$ across the triangle faces. For a given texel $\mathbf{u}$ in the canonical UV space Ω_{uv}, located within a triangle $f \in \mathcal{F}$ with barycentric weights $\mathbf{b} = (b_1, b_2, b_3)$, the corresponding source image coordinate $\mathbf{x}_{src}$ is computed as:

$$\mathbf{x}_{src}(\mathbf{u}) = \sum_{k=1}^{3} b_k \cdot \mathbf{V}^{img}_{f,k} \tag{10}$$

where $\mathbf{V}^{img}_{f,k}$ corresponds to the image-space coordinate of the k-th vertex of face f. The pixel value for the partial texture at location $\mathbf{u}$ is then obtained by bilinear sampling of the input image:

$$\mathbf{T}_p(\mathbf{u}) = \begin{cases} \mathbf{I}(\mathbf{x}_{src}(\mathbf{u})) & \text{if } \mathbf{M_p}(\mathbf{u}) = 1 \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

Here, $\mathbf{M_p} \in \{0,1\}^{H\times W}$ is the binary visibility mask determined by the z-buffer during the initial estimation. To ensure high-fidelity unwrapping and avoid artifacts at UV seams, we filter triangles where the edge lengths in UV space exceed a threshold τ_{dist}.

Design Choice: Partial UVs vs. Generative UV Recovery. An alternative to our barycentric partial-texture extraction is generative UV recovery, such as FreeUV [52]. FreeUV, however, operates only on the facial (FLAME) region and recovers texture through generative completion. Adopting such a model upstream risks altering the subject's facial identity, and is in any case restricted to the face, whereas our pipeline requires full-body textures. We therefore retain the original partial UV maps, despite their incompleteness, to prevent generative identity loss and preserve whole-body skin appearance, and leave the joint generation of a complete UV texture and the final avatar to future work.

C Ablation Study: Donor-based UV Reposing vs Random Patch Masking

In Section 3.2 (main paper), we introduced *Donor-based UV Reposing* to prevent the model from learning shortcuts. We argued that using a real mask from a "donor" creates realistic occlusions that force the network to learn geometry. In this section, we test if this is actually necessary. We compare our method against a simpler baseline where we just block out random parts of the texture to create a synthetic pair.

Implementation Details. For the "Random masking" baseline, we take the input texture ($\mathbf{T_p}$) and cover it with random black squares. We use up to 6 non-overlapping patches, each 64×64 pixels. All other training settings remain identical to our main method.

Qualitative Results.

Visual comparisons in Figure 13 show that random masking is insufficient. The problem is that random squares do not change the overall "shape" of the visible texture. The boundaries of the valid texture still look like the source pose. Because the model can still recognize the source pose from these boundaries, it tries to find a shortcut. However, the random black squares create confusing holes that don't match natural occlusions. To explain these unnatural missing regions, the model, more frequently, hallucinates accessories.

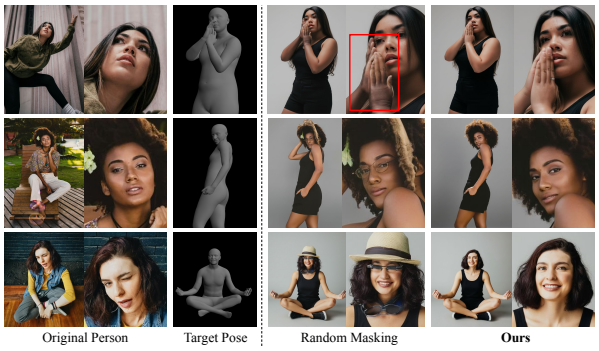

Fig. 13: Random Patch Masking vs. Ours (Donor-Based UV Reposing). Random masking preserves UV boundary shape, leaking pose information and causing artifacts (e.g., hallucinated accessories). Our donor-based reposing changes texture boundaries completely, forcing the model to learn geometry rather than shortcuts.

Our Donor-based method avoids this because the donor mask looks like a real human pose. It changes the texture boundaries completely, so the model cannot rely on the visible pose outline and must learn beyond simple "copy-pasting".

Quantitative Results. Table 4 confirms these visual findings. The Random Masking baseline performs worse on all metrics when evaluated on the WPose dataset. Most notably, identity preservation drops (FaceSim decreases by 6.75%) and overall image quality degrades. This proves that simply hiding pixels is not enough; the occlusion must mimic a realistic pose to train the model effectively.

Masking	PSNR↑	FID↓	SSIM↑	FaceSim↑	DINO↑	HPSv3↑
Random	18.56	6.45	0.821	0.5187	0.6965	7.35
Ours	**19.95**	**5.99**	**0.860**	**0.5571**	**0.7394**	**7.55**

Table 4: Donor Masking Ablation. Random masking (64×64 patches) vs. our donor-based reposing. Our method outperforms on all metrics, confirming that structurally semantic masks are essential.

D Additional Results

D.1 Extended Evaluation In-The-Wild

To demonstrate the robustness of our approach, we evaluate Pro-Pose on "in-the-wild" images sourced from CelebA-HQ [17]. These samples feature diverse

DeepFashion (In-Domain) – BC Inputs								
Method	**PSNR↑**	**FID↓**	**SSIM↑**	**LPIPS↓**	**OKS↑**	**FaceSim↑**	**DINO↑**	**HPSv3↑**
CFLD [28]	18.51	7.09	0.757	0.180	0.48	0.3182	0.9733	4.21
MCLD [26]	19.18	7.01	0.769	0.173	**0.49**	0.3437	0.9657	4.35
LEFFA [57]	14.95	**4.20**	0.767	0.117	0.45	0.5781	0.9415	4.40
OnePoseTrans [6]	14.35	8.59	0.618	0.287	0.46	0.5761	0.9481	4.37
UniHuman [22]	15.95	6.16	0.803	0.149	0.46	0.5805	0.9436	4.24
Gemini 2.5 Flash Image [7]	17.59	4.44	0.757	0.167	0.43	0.5817	0.9697	7.22
Ours	**19.36**	4.24	**0.818**	**0.075**	0.48	**0.6047**	**0.9759**	**7.24**
WPose (Out-of-Domain) – BC Inputs								
Method	**M-PSNR↑**	**FID↓**	**M-SSIM↑**	**M-LPIPS↓**	**OKS↑**	**FaceSim↑**	**DINO↑**	**HPSv3↑**
CFLD [28]	15.61	91.67	0.753	0.201	0.31	0.0888	0.6418	1.92
MCLD [26]	15.80	88.76	0.764	0.196	0.35	0.0997	0.6479	1.93
LEFFA [57]	16.82	64.44	0.786	0.186	0.33	0.0911	0.5734	2.04
OnePoseTrans [6]	17.35	25.87	0.825	0.145	0.33	0.1734	0.7214	4.39
UniHuman [22]	17.75	24.13	0.815	0.153	0.34	0.1121	0.7211	2.95
Gemini 2.5 Flash Image [7]	16.74	8.18	0.788	0.144	0.32	0.4727	0.7021	7.38
Ours	**19.95**	**5.99**	**0.860**	**0.121**	**0.38**	**0.5571**	**0.7394**	**7.55**

Table 5: Quantitative Comparison with BC Inputs. All baselines receive BC-preprocessed input images at test time, removing garment variation as a confounding factor. Pro-Pose maintains its advantage, confirming that our gains are methodological.

lighting conditions, backgrounds, and subjects. The qualitative results, presented in Figure 14, confirm that our method generalizes effectively to unseen real-world data, maintaining high identity fidelity and pose accuracy.

D.2 Extended Qualitative Comparisons

We provide more extensive qualitative comparisons against state-of-the-art methods (MCLD [26], Leffa [57], UniHuman [22], OnePoseTrans [6], and Gemini 2.5 Flash [7]) in Figure 15. Our method consistently produces higher fidelity identity preservation and fewer geometric artifacts than competing approaches, particularly for challenging poses and diverse identities. Our method performs equally well on both in-domain (DeepFashion [27]) and in-the-wild (WPose [22]) scenarios, maintaining consistent quality across datasets.

D.3 Comparison with Gemini 3 Pro Image and Lighting Analysis

We specifically highlight a few examples in Figure 16 which shows representative qualitative results of our method alongside Gemini 3 Pro Image [10] and Gemini 2.5 Flash Image [7]. While the Gemini family produces visually compelling images, both models struggle with faithful reposing to the target SMPL-X pose, often failing to match the requested head and body pose. In contrast, Pro-Pose follows the target pose accurately while preserving identity. The same figure also illustrates our lighting behavior: because our Base Clothing standardization either re-lights subjects into even studio shading or matches the source illumination, Pro-Pose renders realistic, consistent lighting under novel poses without explicitly modeling target-pose illumination.

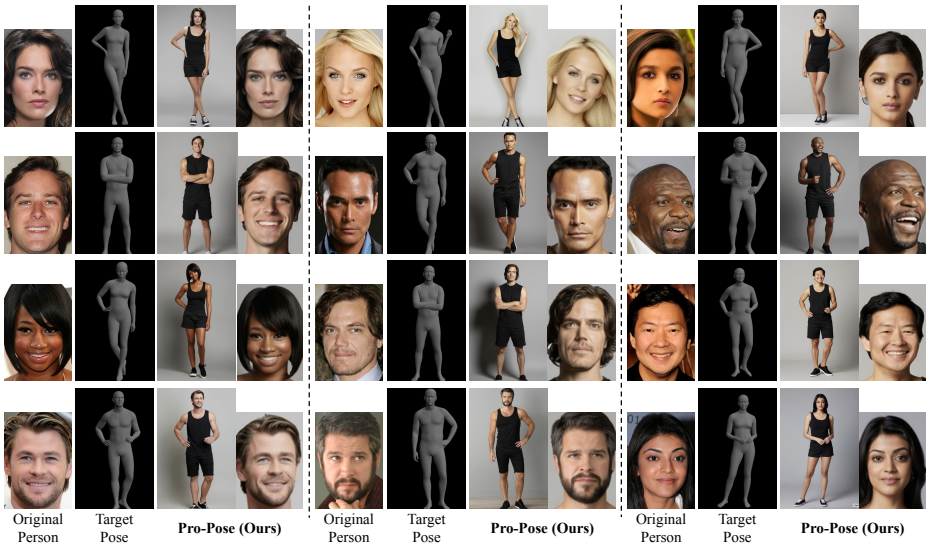

Fig. 14: Pro-Pose In-the-wild. We apply our method to real-world images characterized by varying poses and environments. The columns display (from left to right): (1) the original input image; (2) the target pose represented by a SMPL-X mesh (reposed using ground truth parameters); (3) the output generated by Pro-Pose with a corresponding face crop for detailed visual comparison. Our method consistently preserves identity, including facial details, body shape, and visible skin features, compared to the ground truth.

D.4 Evaluation with BC Inputs for Baselines

A key difference between Pro-Pose and prior methods is that our model is trained on Base Clothing (BC) standardized inputs, while all baselines are trained on original images with diverse garments. To ensure a fair comparison, we additionally evaluate all baselines when provided with BC-preprocessed input images at test time. This removes garment variation as a confounding factor and isolates each method's ability to preserve identity and follow the target pose.

Quantitative Results. Table 5 reports the results. Even when baselines receive the same BC inputs that align with our training distribution, Pro-Pose maintains a significant advantage across identity (FaceSim) and perceptual quality (HPSv3) metrics, confirming that our gains stem from the method itself rather than the data preprocessing.

Qualitative Results. Figure 17 shows representative examples. Despite receiving the same standardized inputs, baselines still exhibit characteristic failures: identity drift (CFLD, MCLD), geometric artifacts in extreme poses (OnePose-Trans), and over-smoothed facial details (UniHuman). Pro-Pose consistently produces sharper, more identity-faithful results.

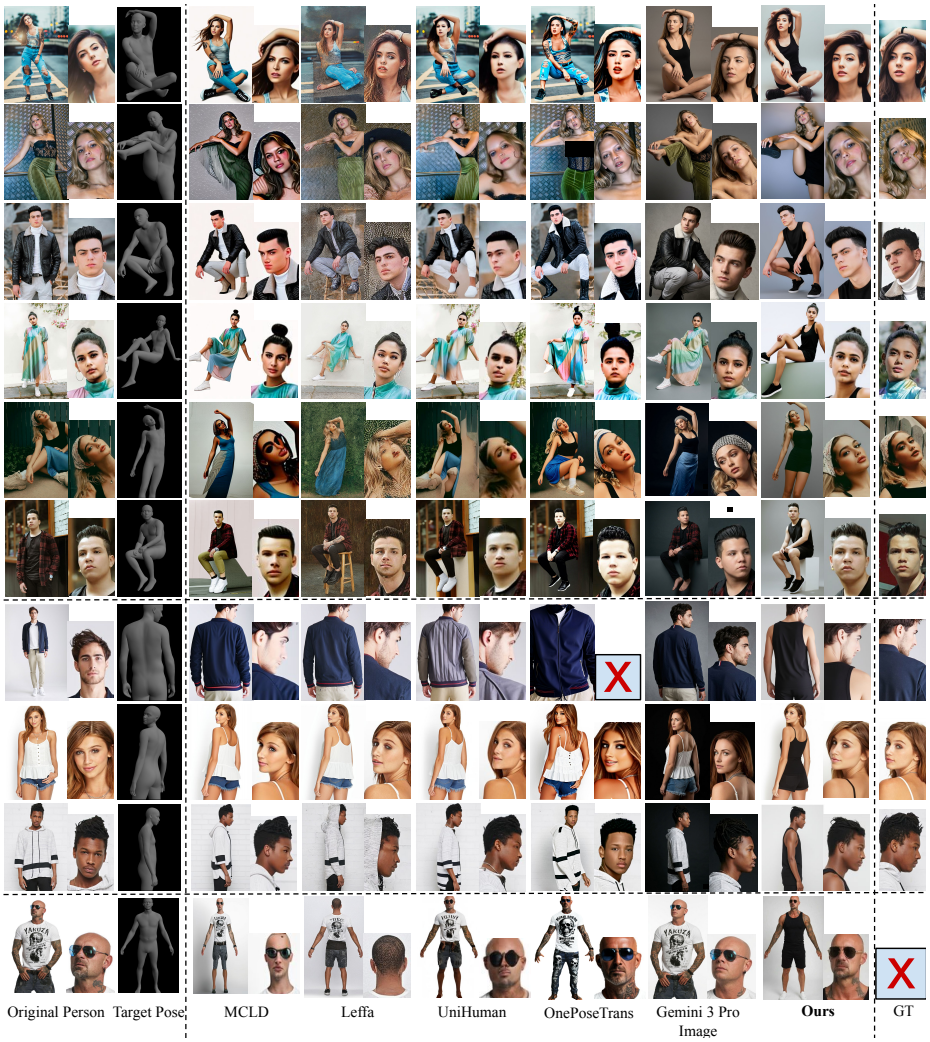

Fig. 15: Qualitative Comparison. The last three rows display results from the DeepFashion dataset [27], and all previous rows utilize real images from the WPose dataset [22]. The columns show the Original Person, Target Pose, and results from state-of-the-art methods (MCLD [26], Leffa [57], UniHuman [22], OnePoseTrans [6], Gemini 3 Pro [10]) compared to **Ours** and the Ground Truth (GT). Our method demonstrates superior performance in preserving the person's identity, including facial features, body shape, and visible skin characteristics, across varying poses. For instance, in the third and fifth rows, previous methods introduce significant identity distortion and loss of facial likeness (e.g., changes in jawline or features). In contrast, **Ours** robustly preserves the unique face structure and identity of the original person, closely matching the GT result. For the final row, no ground truth (GT) exists because the input is a single image reposed into a target A-pose. Although a GT face crop cannot be provided, comparing the result to the face crop of the original image demonstrates our strong identity preservation.

Fig. 16: Comparison with Gemini 3 Pro Image [10] and lighting analysis. Columns show the original image, the target pose, Pro-Pose (ours), Gemini 2.5 Flash Image [7], and Gemini 3 Pro Image. Pro-Pose follows the target pose more faithfully while preserving identity and rendering realistic, consistent illumination.

D.5 Extended Quantitative Evaluation

To further validate the fidelity of our generated avatars, we report metrics computed exclusively on the facial regions in Table 6. This complements the full-body metrics in the main paper by providing a stricter assessment of identity and facial detail preservation, where the large, uniform areas of background and body cannot inflate pixel-level scores.

Metric Selection.

We focus on PSNR, SSIM, LPIPS, and DINO to assess reconstruction quality and semantic alignment. We exclude OKS from this specific analysis, as it relies on full-body keypoints and becomes unstable when restricted to the sparse keypoints available in tight facial crops. Similarly, we omit HPSv3, as it is designed to assess global image aesthetics; when applied to small facial crops, it lacks the necessary global context to yield meaningful discriminative trends. Note that FaceSim is excluded from this table as it was already reported in the main paper (where it is inherently computed on facial embeddings).

DeepFashion (In-Domain)				
Method	**PSNR↑**	**SSIM↑**	**LPIPS↓**	**DINO↑**
CFLD [28]	14.34	0.7012	0.1823	0.9543
MCLD [26]	15.76	0.7151	0.1798	0.9429
LEFFA [57]	13.95	0.6956	0.12	0.9316
OnePoseTrans [6]	13.12	0.5738	0.322	0.9319
UniHuman [22]	13.87	0.7557	0.161	0.9534
Gemini 2.5 Flash [7]	15.59	0.7107	0.1018	0.9531
Unpaired Only	12.87	0.6766	0.2489	0.9001
Paired Only	**16.45**	**0.7765**	**0.087**	**0.9555**
Ours (Unpaired + Paired)	**16.38**	**0.7785**	**0.089**	**0.9554**
WPose (Out-of-Domain)				
Method	**PSNR↑**	**SSIM↑**	**LPIPS↓**	**DINO↑**
CFLD [28]	15.12	0.741	0.279	0.6109
MCLD [26]	15.29	0.757	0.24	0.6108
LEFFA [57]	16.47	0.77	0.236	0.5541
OnePoseTrans [6]	17.1	**0.807**	0.174	0.6719
UniHuman [22]	17.49	0.802	0.179	**0.6814**
Gemini 2.5 Flash [7]	16.96	0.781	0.198	0.6776
Unpaired Only	15.81	0.748	0.255	0.6445
Paired Only	**17.98**	0.7866	**0.165**	0.6443
Ours (Unpaired + Paired)	**19.57**	**0.832**	**0.138**	**0.7043**

Table 6: Extended Quantitative Evaluation. Complementing the full-body metrics in the main paper, we compute PSNR, SSIM, LPIPS, and DINO for the face region.

Analysis. The results reinforce the generalization gap observed in our main experiments. On the in-domain DeepFashion dataset, the "Paired Only" model performs marginally better, consistent with its tendency to overfit to the training distribution. However, on the challenging, out-of-distribution WPose dataset, Pro-Pose significantly outperforms all baselines and the "Paired Only" variant across all metrics. This confirms that our hybrid training strategy yields superior

generalization on fine-grained identity features, avoiding the identity degradation observed in baselines.

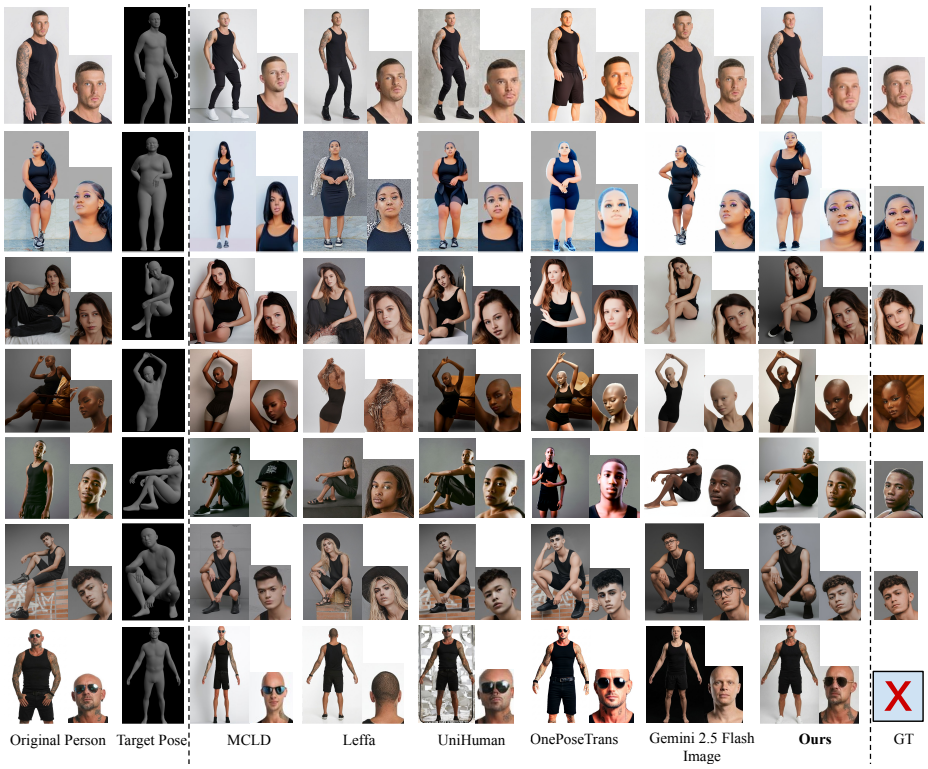

Fig. 17: Qualitative Comparison with BC Inputs. All methods receive BC-preprocessed images as input, isolating the effect of the generation method from the data preprocessing. Pro-Pose preserves identity and pose fidelity more consistently than all baselines. For the final row, no ground truth (GT) exists because the input is a single image reposed into a target A-pose. Although a GT face crop cannot be provided, comparing the result to the face crop of the original image demonstrates our strong identity preservation.

D.6 Full-Body Identity: Skin and Body-Shape Preservation

While the main paper evaluates identity primarily through facial metrics, we additionally quantify how well Pro-Pose preserves *body-skin appearance* and *body shape*, directly assessing the full-body identity claim. Table 7A reports image and perceptual metrics computed exclusively on segmented body-skin pixels (excluding face, hair, and clothing). Table 7B measures body-shape consistency via the SMPL-X shape parameter $\hat{\beta}$, comparing the target and generated bodies

using L2 distance and cosine similarity. As shown, Pro-Pose outperforms both OnePoseTrans [6] and Gemini 3 Pro Image across all metrics, confirming that our gains extend beyond the face to whole-body skin and shape fidelity.

WPose data	**A. Body-skin preservation**				**B. Body-shape preservation**	
Method	PSNR↑	SSIM↑	LPIPS↓	DINO↑	$\hat{\beta}$ L2 ↓	$\hat{\beta}$ Cosine-Similarity ↑
OnePoseTrans [6]	18.01	0.821	0.105	0.7292	2.156	0.6584
Gemini 3 Pro Image	19.13	0.841	0.091	0.7119	1.981	0.6921
Ours	**22.34**	**0.892**	**0.079**	**0.7456**	**0.892**	**0.9487**

Table 7: Full-body identity preservation on WPose. A. Metrics on body-skin pixels only (excluding face, hair, and clothing). **B.** SMPL-X $\hat{\beta}$ L2 distance and cosine similarity between target and generated bodies.

D.7 Impact of Training Data Source

Complementing the quantitative ablation in the main paper, Figure 18 visually demonstrates the impact of our hybrid data strategy. Models trained on "Unpaired Only" data struggle with extreme poses, while "Paired Only" models suffer from identity drift. Our combined approach ("Unpaired + Paired") leverages the strengths of both, achieving robust pose control and identity preservation.

D.8 Ablation: UV as the Control Signal

A natural question is whether Pro-Pose's improvements stem specifically from using the canonical UV texture as the control signal. In the unpaired regime this cannot be ablated, as the UV map is the *sole* source of texture information; omitting it leaves the self-supervised objective undefined. We therefore ablate the UV condition in the *paired-only* setting, where a face crop still provides identity information. As shown in Table 8 (on WPose, directly comparable to Table 1 of the main paper), removing the UV map from paired-only training yields only a modest degradation and closely matches the paired-only baseline with UV. This confirms that our primary gains stem not from the UV control signal alone, but from our novel use of large-scale unpaired data via donor-based reposing.

Pro-Pose (paired only)	**M-PSNR↑**	**FID↓**	**M-SSIM↑**	**M-LPIPS↓**	**OKS↑**	**FaceSim↑**	**DINO↑**	**HPSv3↑**
with UV	18.30	6.65	0.820	0.155	0.34	0.4959	0.6972	7.30
without UV	18.05	7.11	0.801	0.192	0.34	0.4721	0.6888	6.98

Table 8: Ablation of UV as the control signal. Removing the UV map from paired-only training (comparable to Table 1, main paper, WPose) produces only a modest drop, confirming that Pro-Pose's gains stem primarily from our use of unpaired data rather than the UV control signal alone.

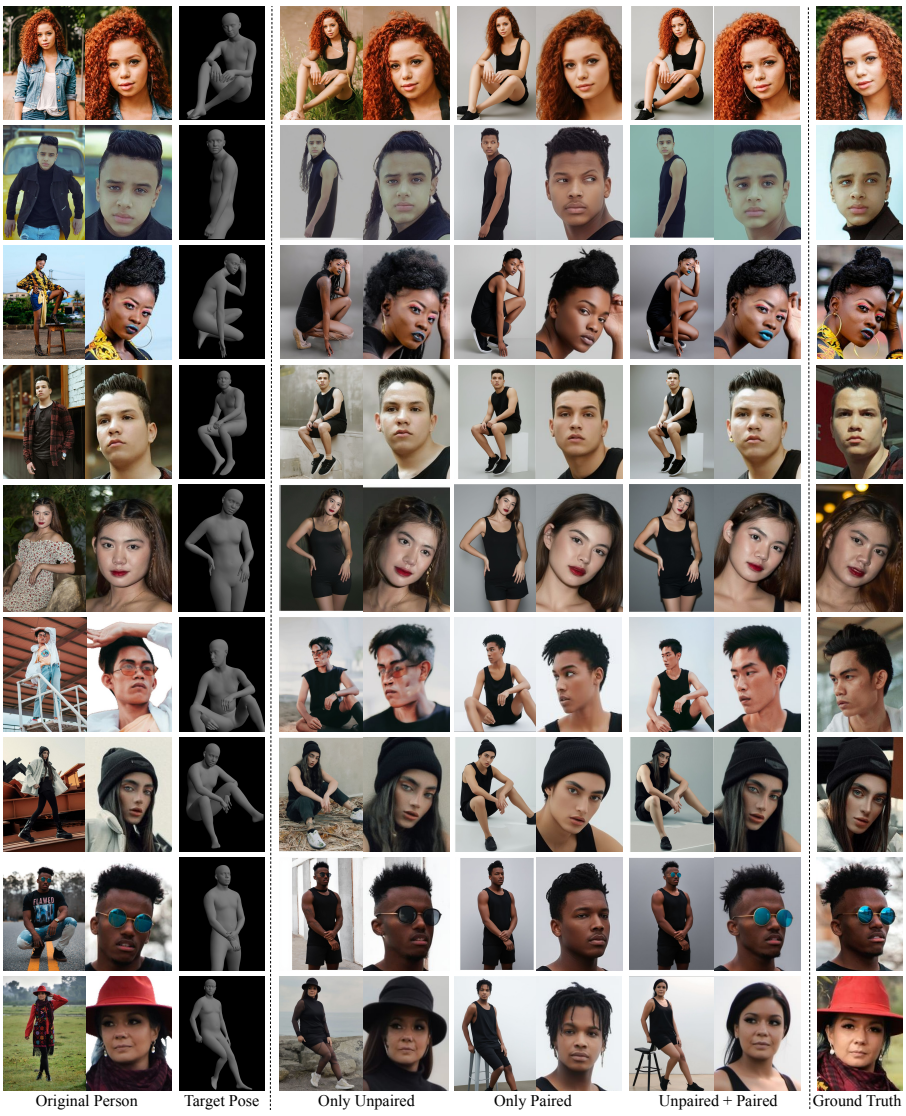

Fig. 18: Ablation Study of Training Data Sources. We evaluate the impact of different training dataset combinations on generated avatar quality. The result columns display models trained on Unpaired data only, Paired data only, and the full Unpaired + Paired combination, respectively. One can observe the improvement in fidelity with the full combined dataset.

D.9 Downstream Virtual Try-On: Same-Pose Comparison

To evaluate Pro-Pose's benefit as a VTO pre-processing stage under a fair, pose-matched setting, Figure 19 compares applying an off-the-shelf VTO model [9] directly to the original in-the-wild image versus applying it to the Pro-Posed avatar, with both branches targeting the same pose. VTO on the original image frequently fails due to source-garment interference, whereas VTO on the Pro-Posed avatar succeeds. This qualitatively complements the user study reported in the main paper (Section 4.4).

Fig. 19: Virtual Try-On under matched target pose. Despite the same target pose, VTO fails on original images due to source-garment interference, but succeeds on Pro-Posed avatars.

D.10 Personalization with Few-Shot Learning

As outlined in Section 3.5 of the main paper, we perform few-shot adaptation to specialize the model for a specific target identity. Here, we provide the specific implementation details for this process.

Data Source and Preprocessing. While the visualization (Figure 4) in the main paper utilizes frames from a video sequence, our fine-tuning protocol is data-agnostic. It requires only a set of paired images of the subject in different poses; these can be sourced from a video or a collection of independent multi-view photographs. Crucially, before fine-tuning, we process all reference images to match our **Base Clothing (BC)** standard (black tank top and shorts). This ensures that the input data aligns with the canonical distribution the model was trained on, allowing the optimization to focus on identity features rather than clothing discrepancies.

Implementation Details. We fine-tune the pre-trained LoRA adapters for 5,000 iterations with a batch size of 4. To prevent the model from overfitting to the limited few-shot examples or forgetting its general geometric priors, we include a small set of regularization images in the training batches. For consistency and reproducibility, we maintain all other hyperparameters (including learning rate and optimizer settings) identical to the main training phase. While subject-specific hyperparameter tuning could theoretically yield higher fidelity results, we report all metrics using this fixed configuration.

More Qualitative Evaluation. We demonstrate how our few-shot learning strategy generalizes to in-the-wild scenarios [58] in Figure 20. Notably, the identi-

Fig. 20: Personalized Pro-Pose. We fine-tune our method using the images shown in the leftmost column (rows 1-4). Column (5) displays the target pose represented by a SMPL-X mesh (reposed using ground truth parameters). Columns (6) and (7) present the generated output without and with fine-tuning, respectively. The subsequent columns repeat this arrangement for a second target pose.

ties of the generated new poses become significantly closer to the original subject after fine-tuning, validating the effectiveness of this adaptation step.

E Analysis of Identity Overfitting

A primary motivation for our work is the limited identity diversity in existing paired datasets like DeepFashion. As visualized in Figure 21, DeepFashion contains limited (≈ 100) unique identities. Training generative models exclusively on such limited data leads to severe overfitting, where the model biases generated faces toward the training identities rather than the input subject.

E.1 DeepFashion Identity Analysis

To empirically validate the limited identity diversity of the DeepFashion In-Shop Clothes Retrieval benchmark [27], we performed a graph-based clustering analysis on the training set. Our pipeline proceeds as follows:

1. **Embedding Extraction:** We utilized ArcFace [4] to detect faces and extract identity embeddings. To ensure embedding reliability, we explicitly filtered out faces with extreme head poses (yaw or pitch $> 45°$).
2. **Graph Construction:** We constructed an undirected graph where nodes represent images. Edges were established via a dual-criteria strategy:
   - *Visual Similarity:* An edge is created if the cosine similarity between two embeddings exceeds a strict threshold of 0.6.
   - *Structural Verification:* We leveraged the dataset's directory structure (where images in the same garment folder typically depict the same subject) to add ground-truth edges. However, to filter labeling errors

Fig. 21: DeepFashion. We highlight the top 100 identity clusters found in the DeepFashion dataset. We also exclusively highlight the first person on the top-left, who can be repeatedly observed in the generations of the overfitted models.

or occlusions, these structural edges were only added if the similarity exceeded a verification threshold of 0.4.

3. **Clustering:** We computed the connected components of this graph. This process yielded approximately 100 distinct clusters (identities), confirming that the dataset is heavily dominated by a small number of professional models appearing across thousands of garment items.

E.2 Identity Analysis of Generated Images

We diagnose this phenomenon in Figure 22. We reposed subjects from the out-of-distribution - WPose dataset using two models: one trained only on paired DeepFashion data, and our full model trained on both paired and unpaired

data. We then measured the face similarity between the *generated* outputs and the *DeepFashion training set*.

- The **Paired-Only** model (red histograms in Figure 22 (a) and (b)) produces outputs with high similarity to the training data, indicating it is "pulling" novel identities toward the few identities it memorized during training.
- **Our Full Model** (blue histograms in Figure 22 (a) and (b)) maintains lower similarity to the training set, proving that the inclusion of abundant unpaired data enables better generalization to unseen identities.

The box-plot in Figure 22 (c) further highlights the extensive number of samples generated by the paired-only model that have very high similarity scores to DeepFashion training samples.

The cumulative distribution (Figure 22 (d)) highlights that our model trained on both paired and unpaired data generated samples with a maximum Face Similarity score of 0.4128, while almost 15% of the total samples generated by the paired only model scored an even higher face similarity score.

E.3 DeepFashion Identity Limitation Leading to Overfitting

We highlight a dominant identity from the DeepFashion training set in the red box of Figure 21. Comparing this to the "Paired Only" results in Figure 18 (and Figure 7 of main paper) reveals severe overfitting: As we can see in some cases the model tends to collapse to this specific training identity (for example second row in Figure 18). This bias is so strong that it overrides semantic gender cues, frequently substituting female subjects with this specific male face, confirming that the paired-only model relies on memorization rather than generalization.

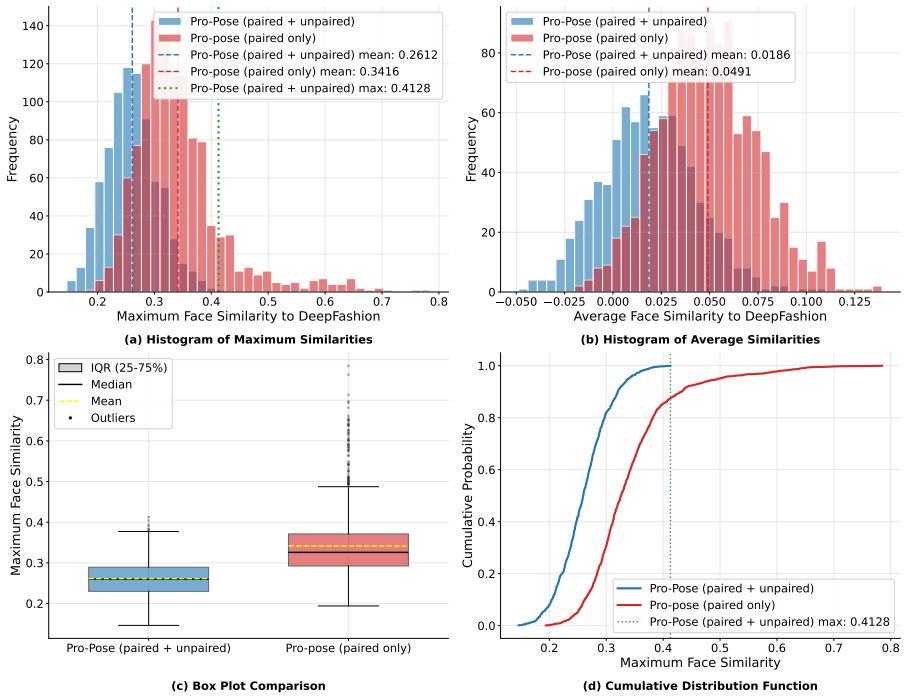

Fig. 22: Diagnosing Identity Overfitting via Cross-Dataset Generation. We evaluate the extent of identity overfitting by comparing the face similarity between images generated on the out-of-distribution WPose dataset and the identities present in the DeepFashion training set. Two Pro-Pose models are tested: Pro-Pose (paired only), trained on limited DeepFashion paired data, and Pro-Pose (unpaired + paired), trained on abundant unpaired data plus the limited paired data. The Pro-Pose (paired only) model exhibits a significantly higher face similarity to the DeepFashion training identities when reposing WPose identities. This result acts as a diagnostic, demonstrating that the paired-only model has overfit to the limited DeepFashion identities, causing it to generate new faces that are biased towards its training set, whereas the combined training approach ensures better generalization.