Visual Persona: Foundation Model for Full-Body Human Customization

Jisu Nam[1*], Soowon Son[1], Zhan Xu[2], Jing Shi[2], Difan Liu[2], Feng Liu[2],
Seungryong Kim[1†], Yang Zhou[2†]

[1]KAIST AI [2]Adobe Research

https://cvlab-kaist.github.io/Visual-Persona

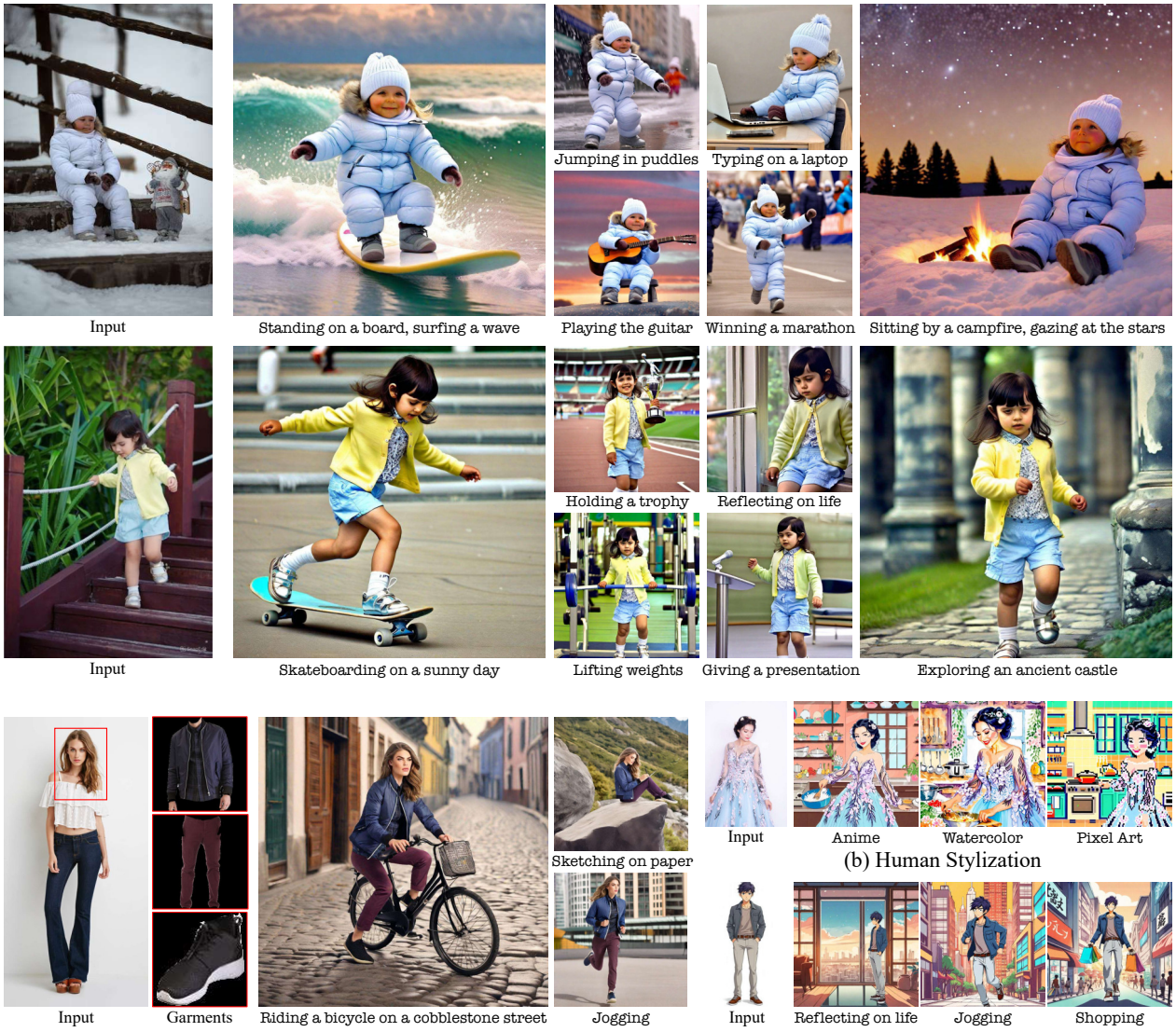


Figure 1. **Visual Persona enables Text-to-Image Full-Body Human Customization.** Given a single human image, our method generates diverse customized images of the individual that are closely aligned with text descriptions while accurately preserving full-body appearance, which unlocks a wide range of applications, including (a) text-guided virtual try-on, (b) human stylization, and (c) character customization.

Abstract

We introduce Visual Persona*, a foundation model for text-to-image full-body human customization that, given a single in-the-wild human image, generates diverse images of the individual guided by text descriptions. Unlike prior methods that focus solely on preserving facial identity, our approach captures detailed full-body appearance, aligning with text descriptions for body structure and scene variations. Training this model requires large-scale paired human data, consisting of multiple images per individual with consistent full-body identities, which is notoriously difficult to obtain. To address this, we propose a data curation pipeline leveraging vision-language models to evaluate full-body appearance consistency, resulting in* Visual Persona-500K*—a dataset of 580k paired human images across 100k unique identities. For precise appearance transfer, we introduce a transformer encoder-decoder architecture adapted to a pre-trained text-to-image diffusion model, which augments the input image into distinct body regions, encodes these regions as local appearance features, and projects them into dense identity embeddings independently to condition the diffusion model for synthesizing customized images.* Visual Persona *consistently surpasses existing approaches, generating high-quality, customized images from in-the-wild inputs. Extensive ablation studies validate design choices, and we demonstrate the versatility of* Visual Persona *across various downstream tasks.*

1. Introduction

Recent advances in large-scale generative models [23, 55, 57, 60, 66] have made remarkable strides in creating photo-realistic images. The success of these models enables users to customize pre-trained models with their own personal data [15, 49, 58, 81]. More recently, such methods have focused on human data [16–18, 41, 52, 53, 65, 72, 75, 79, 81, 84], driven by various practical applications, including film production [30, 46], book illustration [3, 71], and virtual/augmented reality [8, 35, 40, 48]. These methods allow users to synthesize novel renditions of specific individuals.

However, most human-customized models primarily focus on generating human face images [16–18, 41, 52, 75, 79, 81], restricting their applicability to the face domain. This limitation motivates us to broaden the scope of previous studies to the full-body human domain, which we refer to as *Full-Body Human Customization*, a largely underexplored area. In this paper, we aim to develop a foundational model for full-body human customization, unlocking a wide range of in-the-wild applications.

We argue that such a foundational model should satisfy two key criteria: text alignment and identity preservation. Text alignment refers to the ability to generate diverse images aligned with given text descriptions, adjusting to variations in facial expressions, poses, actions, and surroundings. Identity preservation ensures that the model produces a consistent full-body appearance that accurately matches the input human image, including facial identity, clothing, and accessories. We present the development of this model, which embodies the above criteria, dubbed **Visual Persona**.

Large-scale paired human data, comprising multiple images of the same individuals with consistent full-body identities, is crucial for preserving full-body appearance consistency and achieving text alignment. However, obtaining such data is notoriously difficult. As a result, recent works use single human datasets that contain only one image per individual [65, 84]. To overcome this, we design a data curation pipeline that constructs paired human data from a large pool of unpaired human collections. In light of recent progress in Vision Language Models (VLMs) [1, 2, 5, 44, 47, 70, 76], we propose that VLMs can serve as simple yet effective tools for evaluating full-body visual consistency, even when considerable variations exist across images. Furthermore, we generate detailed text captions for each image to disentangle the individual's identity from intra-individual variations depicted in the text prompts. As a result, we collected a total of 580k images with text captions across 100k unique full-body identities, which we denote as **Visual Persona-500K**. In Table 1, we highlight that Visual Persona-500K surpasses the scope of earlier datasets.

For identity preservation, it is important to accurately transfer the input full-body human's appearance to the customized image while retaining the large geometric deformations introduced by pre-trained Text-to-Image (T2I) models. Recent works [16–18, 41, 52, 53, 65, 72, 75, 79, 81, 84] solved human customization using pre-trained im-

Methods	Type	Database	Size	Res.	Domain
IP-Adapter [81]	Single	LAION-2B, COYO-700M	10M	-	General
E4T [16]	Single	FFHQ, CelebA-HQ	100K	1024	Face
PortraitBooth [53]	Single	CelebV-T	70K	512+	Face
InstantID [75]	Single	LAION-Face, Internal	60M	-	Face
PuLID [18]	Single	Internal	1.5M	-	Face
InstantBooth [65]	Single	Internal	1.4M	1024+	Human
StoryMaker [84]	Single	Internal	500K	-	Human
LCM-Look. [17]	Paired	Synthetic	500K(100K)	1024	Face
PhotoMaker [41]	Paired	VoxCeleb1, VGGFace2	112K(13K)	256+	Face
Arc2Face [52]	Paired	WebFace42M	21M(1M)	448	Face
Visual Persona (Ours)	Paired	Visual Persona-500K	580K(100K)	1024+	Human

Table 1. **Comparison of datasets in state-of-the-art customized models.** This table outlines the data type, database, data size, image resolution, and data domain used for each method. Unlike existing works that focus on the face domain or use a single image per individual, our approach aims to explore large-scale, real paired human data with full-body appearance consistency.

†Co-corresponding author.

*Work done during an internship at Adobe Research.

age encoders to accelerate the tuning process. These approaches commonly leverage semantic representations from CLIP [56] or facial recognition models [11, 12, 62, 63] to encode the input image, then globally map the encoded features to compact identity embeddings through simple linear projection layers [16, 18, 41, 53, 79, 81]. While effective in the face domain, they often fail to preserve the detailed appearance of in-the-wild human inputs.

In this paper, we propose a transformer encoder-decoder architecture, adapted to a pre-trained T2I diffusion model, specifically designed for full-body human customization. To effectively transfer each body part of the input human into the complex body structure of the customized images, we augment the input image into distinct body images using an off-the-shelf parsing method [39]. A pre-trained image transformer encoder [51] encodes the set of body images for the individual and extracts local appearance features from each. The body-partitioned transformer decoder then projects each body feature into its corresponding identity embedding through a cross-attention layer, further refining intra-relationships within the embeddings using a self-attention layer. The output embeddings are concatenated along the token length to condition the pre-trained T2I diffusion model through a learnable identity cross-attention module, while text embeddings condition the model via a text cross-attention module. Finally, the diffusion model synthesizes diverse images with a consistent identity matching the input. Notably, we freeze the pre-trained T2I model to maximize its generative capability, preserving both image quality and text alignment.

We evaluate the effectiveness of our method on a GPT-based evaluation [54] and a human evaluation, with comparison to state-of-the-art works [41, 75, 81, 84]. Comprehensive ablation studies validate our design choices and the contribution of each component. Furthermore, we highlight the versatility of our method across various applications.

2. Related Work

Learning Domain Priors for T2I Customization. The objective of T2I customization [15, 37, 58] is to adapt pre-trained T2I generative models to user-provided subject images. While earlier tuning-based approaches [4, 13, 15, 20, 37, 58, 59, 74] involve a laborious tuning process for each subject, recent methods have shifted toward learning domain-specific priors from large-scale data within a single domain [7, 16–18, 34, 41, 53, 65, 72, 75, 77, 79, 81, 84]. In particular, human face customization has gained significant attention in this area [16–18, 41, 52, 53, 65, 72, 75, 79, 81], driven by extensive face datasets, such as FFHQ [32], CelebA-HQ [31], CelebV-T [82], and LAION-Face [61]. However, these datasets are confined to the face domain, leaving the full-body human domain mostly unexplored. Although some works [65, 84] seek to use in-the-wild human images, it is challenging to obtain paired human data consisting of multiple images per individual with consistent full-body identity. As a result, these methods compromise by training on a single image per individual, which struggles with the trade-off between identity preservation and text alignment [17, 18, 52].

Encoder-based T2I Customization. To initialize the tuning process, domain-tuning approaches for face customization [16–18, 41, 52, 53, 65, 72, 75, 79, 81] often extract semantic representations from pre-trained image encoders such as CLIP [56] or face recognition models [11, 25, 67, 68], which are trained with global supervision on text or identity classes. These methods commonly embed features from the image encoder into a small number of token embeddings [16–18, 41, 52, 53, 65, 72, 75, 79, 81], often using linear projection layers [16, 18, 41, 53, 79, 81]. Although effective in the face domain, these approaches often struggle with full-body human inputs, as globally mapping into a small number of tokens fails to condense key visual information and tends to blend local details from each body part, resulting in poor full-body appearance preservation.

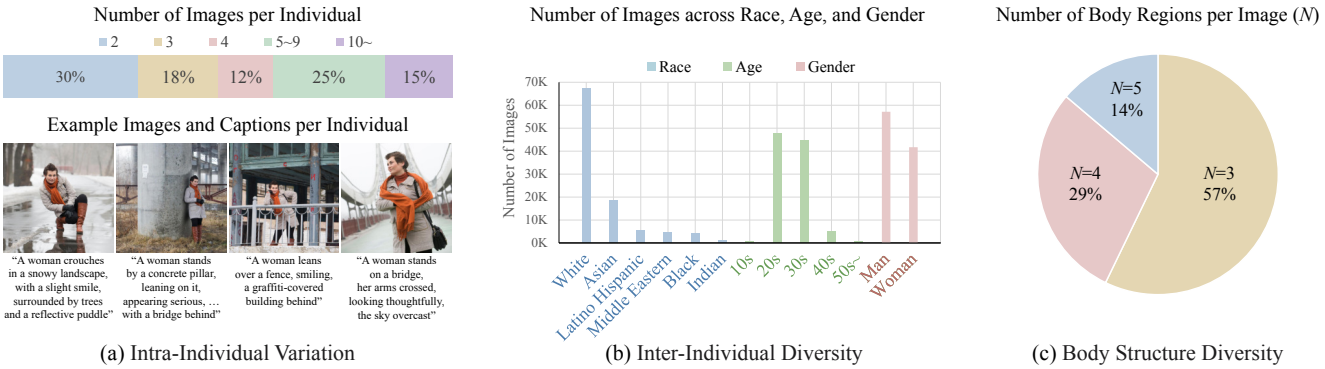
3. Method

3.1. Visual Persona-500K Dataset

A paired human dataset, comprising multiple images per individual with consistent identities, is essential for concurrently achieving identity preservation and text alignment. However, due to the challenges of assembling consistent full-body human pairs, previous works have compromised by training on a single image per individual [65, 84]. In this work, we curated our paired human dataset, Visual Persona-500K, from a large pool of unpaired human collections that include visually inconsistent images of the same individuals. Furthermore, we generate detailed captions for each image to disentangle an individual's identity from input and intra-individual variations driven by text prompts. The overall statistics and scope of our dataset are presented in Figure 2 and Table 1.

Curating Consistent Facial Identities. We first collect unpaired human data, consisting of multiple images per each individual. To further ensure facial identity consistency, we calculate the cosine similarity between facial embeddings for each image pair using a face recognition model [11]. We then select an anchor image with the highest average similarity score and discard any images whose similarity scores with the anchor image fall below a predefined threshold. We further filter out images with unidentifiable faces, watermarks, or duplicates. For high image quality, we discard images where the shortest side is less than 1024 pixels.

Curating Consistent Full-Body Identities. From the unpaired human data, which only guarantees facial identity consistency, we further evaluate body identity consistency by assessing whether each individual is wearing identical



(a) Intra-Individual Variation (b) Inter-Individual Diversity (c) Body Structure Diversity

Figure 2. **Data Statistics:** Our curated training dataset, **Visual Persona-500K**, consists of 580k images representing 100k individuals. (a) illustrates the distribution of the number of images per individual, with over 50% of individuals having more than four images, and shows example image-caption pairs from the same individual. (b) highlights the diversity of individuals based on facial attributes, including race, age, and gender, which are estimated by DeepFace [64]. (c) showcases body structure diversity, segmented into five clusters—full-body, face, torso, legs, and shoes—categorized using a body-parsing method [39].

clothing. Specifically, we suggest that VLMs [1, 2, 5, 44, 47, 70, 76] are simple yet powerful tools for evaluating the visual consistency of the human body. For each image subset of the same individual, we prompt LLAVA [44] to assess whether the individuals in the subset are wearing the same clothing. A simple prompt—*"Are they wearing exactly the same clothes?"*—triggers the model to provide a binary decision with high precision. If the model returns a positive response for all subsets of the individual, the individual is retained; otherwise, the individual is excluded from the dataset. In total, we curated a dataset of 580k full-body paired human images across 100k unique individuals, referred to as Visual Persona-500K. A detailed description of the curation method is provided in Appendix B.

Captioning. We aim to disentangle the individual's identity in the input from intra-individual variations depicted in the text prompts. To achieve this, we use Phi-3 [1] to generate detailed text descriptions for each image that contain identity-irrelevant information. The prompt used is: *"Describe the image in detail in one sentence, focusing on facial expression, pose, actions, and surroundings."* Notably, our data curation and captioning pipeline is compatible with any off-the-shelf VLM [1, 2, 5, 44, 47, 70, 76].

3.2. Model Architecture

Given a single human image and a text prompt, our goal is to transfer the full-body appearance from the input to customized images generated by the pre-trained T2I model, while preserving its generative capabilities for image quality and text alignment. Prior studies [16–18, 41, 52, 53, 65, 72, 75, 79, 81, 84] commonly leverage semantic representations from pre-trained image encoders, such as CLIP [56], by mapping them into compact identity embeddings, often through linear projection layers [16, 18, 41, 53, 79, 81]. However, these methods frequently struggle to preserve the detailed appearance of in-the-wild human inputs.

In this paper, we introduce a novel transformer encoder-decoder architecture [73], adapted to a pre-trained T2I diffusion model, specifically designed for full-body human customization. We decompose the input human body into distinct body regions as separate images. The image transformer encoder then encodes each body image into localized appearance features, while the body-partitioned transformer decoder projects each feature into the corresponding dense identity embeddings. These embeddings then guide the pre-trained T2I model to synthesize diverse customized images that accurately preserve the input's full-body appearance. The overall architecture is illustrated in Figure 3.

Body Part Decomposition. Figure 4 provides an overview of body part decomposition. To effectively transfer the input human appearance to the complicated body structure in the customized image, it is important for the diffusion model to attend independently to each distinct body part from the input and map it to the corresponding part in the synthesized image. To achieve this, we augment the input image $I \in \mathbb{R}^{H\times W\times 3}$ (Figure 4(a)), where H, W, and 3 represent the height, width, and RGB channels, respectively, into N distinct body images. Here, we apply a foreground mask [26] to the input image to focus solely on the human part. Specifically, we leverage an off-the-shelf body parsing method [39] to parse body regions (Figure 4(b)) and group them into $N-1$ categories (Figure 4(c)). We then extract bounding box coordinates for each category (Figure 4(d)), crop the corresponding regions from I, zero-pad along the largest side to preserve image scale, and resize them to the original input size (Figure 4(e)). This results in an input set of the individual, $X \in \mathbb{R}^{N\times H\times W\times 3}$, comprising one full-body image and $N-1$ body part images. Note that N can represent any number of body regions. In this paper, we empirically set N to 5, corresponding to the full body, face, torso, legs, and shoes. Table 3 demonstrates the effectiveness of body part decomposition.

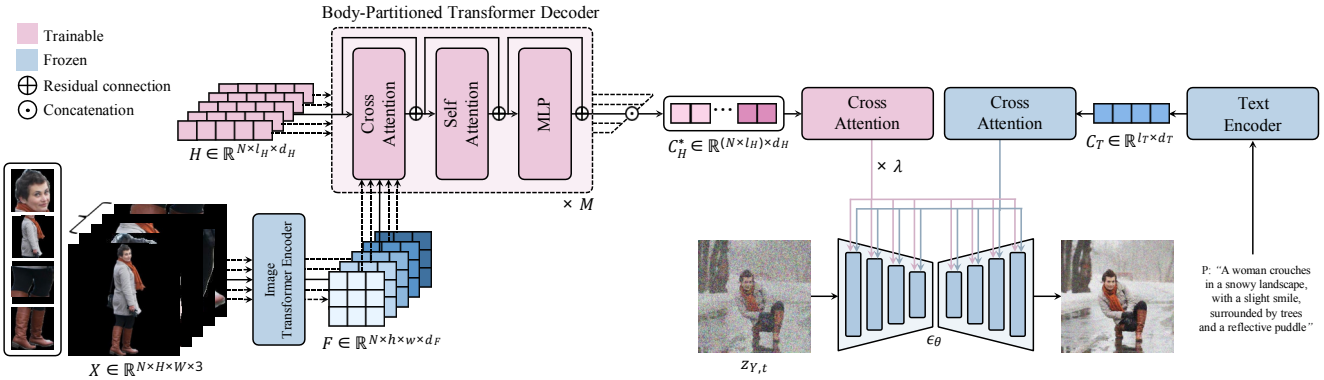


Figure 3. **Overall Architecture:** Our network augments the input human image into body regions, which are encoded into local features by an image transformer encoder. A body-partitioned transformer decoder projects these features into learnable identity embeddings via cross-attention, followed by self-attention and MLP. After M iterations, the embeddings are concatenated to form a stacked identity embedding. The identity embedding and text embedding from detailed captions condition a pre-trained T2I diffusion model to synthesize a new image with the input identity. Only the body-partitioned transformer decoder and identity cross-attention module are trained.

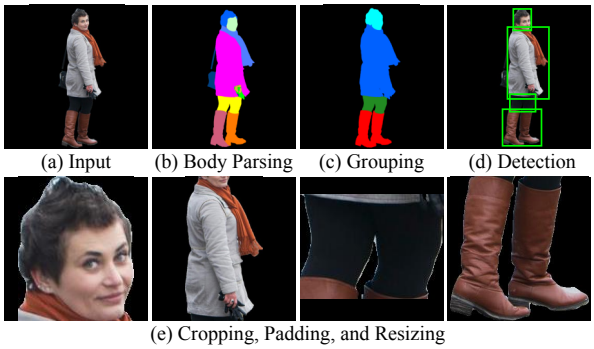


Figure 4. **Body Part Decomposition.**

Image Transformer Encoder. For detailed appearance preservation, we extract local image features from the input set of the individual. Specifically, we use the pre-trained vision transformer DINOv2 [51] as our image encoder. In contrast to other image encoders [21, 56] trained with weak supervision on image-text or class alignment, which focus on the semantics of images, the self-supervised vision transformer DINOv2 [51] can capture fine-grained, localized features, including structural and texture information across images [28, 83]. This makes DINOv2 [51] particularly well-suited for our task of preserving visual details, such as clothing patterns and textures. Specifically, we encode the input set X into a set of image features $F \in \mathbb{R}^{N\times h\times w\times d_F}$, where h, w, and d_F represent the height, width, and channel dimensions of each feature. Notably, instead of using the pre-defined class token `CLS`, we leverage the full set of local tokens to focus on spatial information from the input.

Body-Partitioned Transformer Decoder. A straightforward approach to mapping the input features onto the pre-trained T2I model involves using Multi-Layer Perceptron (MLP) layers, as in prior works [16, 18, 41, 53, 79, 81]. However, as discussed in Table 3, simply using an MLP is insufficient to capture key visual attributes from complex input features, particularly from in-the-wild human images, and tends to blend local details across the body into a global appearance. To address this, we propose a body-partitioned transformer decoder that projects local features from each body region into corresponding dense identity embeddings, preserving detailed visual appearance. Each transformer layer comprises a cross-attention layer, a self-attention layer, and an MLP layer.

Let $H \in \mathbb{R}^{N\times l_H\times d_H}$ represent the learnable hidden identity embeddings for each body image, whose dimensions match those of the output identity embeddings. Here, l_H and d_H denote the token length and the channel dimension, respectively. We highlight that, unlike prior methods [16–18, 41, 52, 53, 65, 72, 75, 79, 81, 84] that map image features into a small number of tokens (typically $l_H = 16$), we set $l_H = h \times w$, where h and w represent the height and width of the DINOv2 feature. Table 4 shows that dense embeddings preserve finer full-body details that a small number of tokens may overlook. For the i-th input body image and the j-th transformer layer, the cross-attention layer links each hidden embedding $H^{i,j}$ to its corresponding feature F^i from the i-th body image X^i. The updated embedding, $H^{i,j}_{ca}$, is then passed through the self-attention layer, which learns internal relationships within the embedding. The resulting embedding, $H^{i,j}_{sa}$, is further processed by the MLP layer, mapping it to the output embedding $H^{i,j+1}$, which is recursively fed into the next transformer layer. This process is formulated as:

$$H^{i,j}_{ca} = \texttt{C-Att}(\texttt{LN}(H^{i,j}), F^i, F^i) + H^{i,j}, \quad (1)$$

$$H^{i,j}_{sa} = \texttt{S-Att}(\texttt{LN}(H^{i,j}_{ca}), \texttt{LN}(H^{i,j}_{ca}), \texttt{LN}(H^{i,j}_{ca})) + H^{i,j}_{ca}, \quad (2)$$

$$H^{i,j+1} = \texttt{MLP}(\texttt{LN}(H^{i,j}_{sa})) + H^{i,j}_{sa}, \quad (3)$$

where LN stands for layer normalization. $\texttt{C-Att}(\cdot)$ represents the cross-attention layer, where $\texttt{LN}(H^{i,j})$ is projected as the queries and F^i is projected as the keys and values. $\texttt{S-Att}(\cdot)$ denotes the self-attention layer, where $\texttt{LN}(H^{i,j}_{ca})$ is projected as the queries, keys, and values. MLP refers to the multi-layer perceptron layer. After M iterations, we obtain the refined identity embeddings, $C_H \in \mathbb{R}^{N \times l_H \times d_H}$. Finally, these embeddings are concatenated along the token length to form a stacked identity embedding, C^*_H:

$$C^*_H = \texttt{Concat}([C^1_H, \ldots, C^N_H]) \in \mathbb{R}^{(N \times l_H) \times d_H}, \quad (4)$$

where $\texttt{Concat}(\cdot)$ represents the concatenation operation.

Decoupled Cross-Attention. Given C^*_H, the pre-trained T2I model $\epsilon_\theta(\cdot)$ portrays the target image Y, aligned with the text embedding C_T derived from text prompt T. Specifically, we adopt a decoupled cross-attention mechanism [81]. C^*_H is processed through an additional learnable identity cross-attention module, where target queries Q_Y for Y attend to keys K^*_H and values V^*_H projected from C^*_H. The text embedding $C_T \in \mathbb{R}^{l_T \times d_T}$ is incorporated through the text cross-attention module, projected as keys K_T and values V_T, where l_T and d_T denote the token length and the channel dimension, respectively. The output Z' is formulated as:

$$Z' = \texttt{C-Att}(Q_Y, K_T, V_T) + \lambda \cdot \texttt{C-Att}(Q_Y, K^*_H, V^*_H), \quad (5)$$

where $\texttt{C-Att}(\cdot)$ is the text or identity cross-attention module, and λ is the weighting scalar.

3.3. Training

We solely train the body-partitioned transformer decoder and the identity cross-attention module, while freezing all other parameters, including the image transformer encoder, body parsing model, and T2I diffusion model. For training, we leverage our paired human dataset, Visual Persona-500K, where the input set of the individual, X, guides the frozen diffusion model $\epsilon_\theta(\cdot)$ to render another image of the same individual, Y, aligned with the target prompt T. The training objective is formulated by minimizing the prediction error L [23, 66]:

$$L := \mathbb{E}_{z_{Y,t}, \epsilon, t, C_T, C^*_H} \left[\|\epsilon - \epsilon_\theta(z_{Y,t}, t, C_T, C^*_H)\|^2_2 \right], \quad (6)$$

where ϵ is the added noise in the forward pass and $z_{Y,t}$ represents the noisy latent variable of Y at time step t. L is backpropagated through the diffusion model, updating the parameters in the body-partitioned transformer decoder and the identity cross-attention module. Note that prior works [16, 18, 53, 65, 72, 75, 79, 81, 84] were trained on identical images for X and Y, leading to overfitting on identity-unrelated elements like background, lighting, and composition [17, 18, 52].

4. Experiments

4.1. Experimental Settings

Dataset. We evaluated our method on two human datasets: SSHQ [14] and PPR10K [42]. SSHQ [14] includes high-quality full-body images featuring various identities, poses, and clothing, sourced from DeepFashion [45] and African images from InFashAI [19]. PPR10K [42] consists of a wide range of in-the-wild human images with significant geometric variations and diverse viewpoints. We randomly selected 50 individuals each from the SSHQ and PPR10K test splits [65] for testing. To assess text alignment, we augmented 17 prompts for live objects in Dreambooth [58] using ChatGPT [2] to include facial expressions, poses, actions, and surroundings. For evaluation, all methods generated 4 samples for each input image and prompt pair, resulting in 3,400 samples per evaluation dataset. Additional details on the evaluation data are provided in Appendix C.2.

Evaluation Metric. As discussed in [10, 29, 38, 43, 69] and Appendix C.3, obtaining reliable quantitative evaluations for customized image generation is challenging, as they heavily rely on human preferences. In this paper, we evaluated our method using Dreambench++ [54], an automated, human-aligned, GPT [2]-based benchmark.

To evaluate identity preservation, we adapted Dreambench++ [54], prompting GPT [2] to assess face identity, clothing type, design, texture, and color in generated images compared to input images, denoted as D-I. To evaluate text alignment, we instructed GPT to assess pose, actions, surroundings, composition, and overall quality to ensure prompt alignment, denoted as D-T. Each metric assigns an integer score from 0 (no resemblance or correlation) to 9 (near-perfect resemblance or correlation). As we aim for high scores on both D-I and D-T, following [6, 78], we also calculate their harmonic mean per sample, denoted as D-H. As GPT has difficulty detecting facial expressions when the subject is distant from the foreground, we also conducted a human evaluation for text alignment on facial expressions, detailed in Appendix C.3.

Human Evaluation Setting. We conducted a rigorous human study, strictly following the ImagenHub evaluation protocol [36]. Each generated sample, along with its input image and text prompt, was assessed by two metrics: semantic consistency (SC) and perceptual quality (PQ), both scored as $\{0, 0.5, 1\}$. For SC, 0 indicates that either the text or the identity does not align with the generated sample, 0.5 indicates partial alignment of both, and 1 indicates full alignment. This ensures a balanced evaluation, penalizing bias toward either text alignment or identity preservation. PQ measures the visual realism of the generated image, where higher is better. The final score is calculated as $O = \sqrt{SC \times PQ}$. Following [24], eight human raters were recruited and trained according to ImagenHub [36] guidelines. Eight raters were divided into two groups, with each

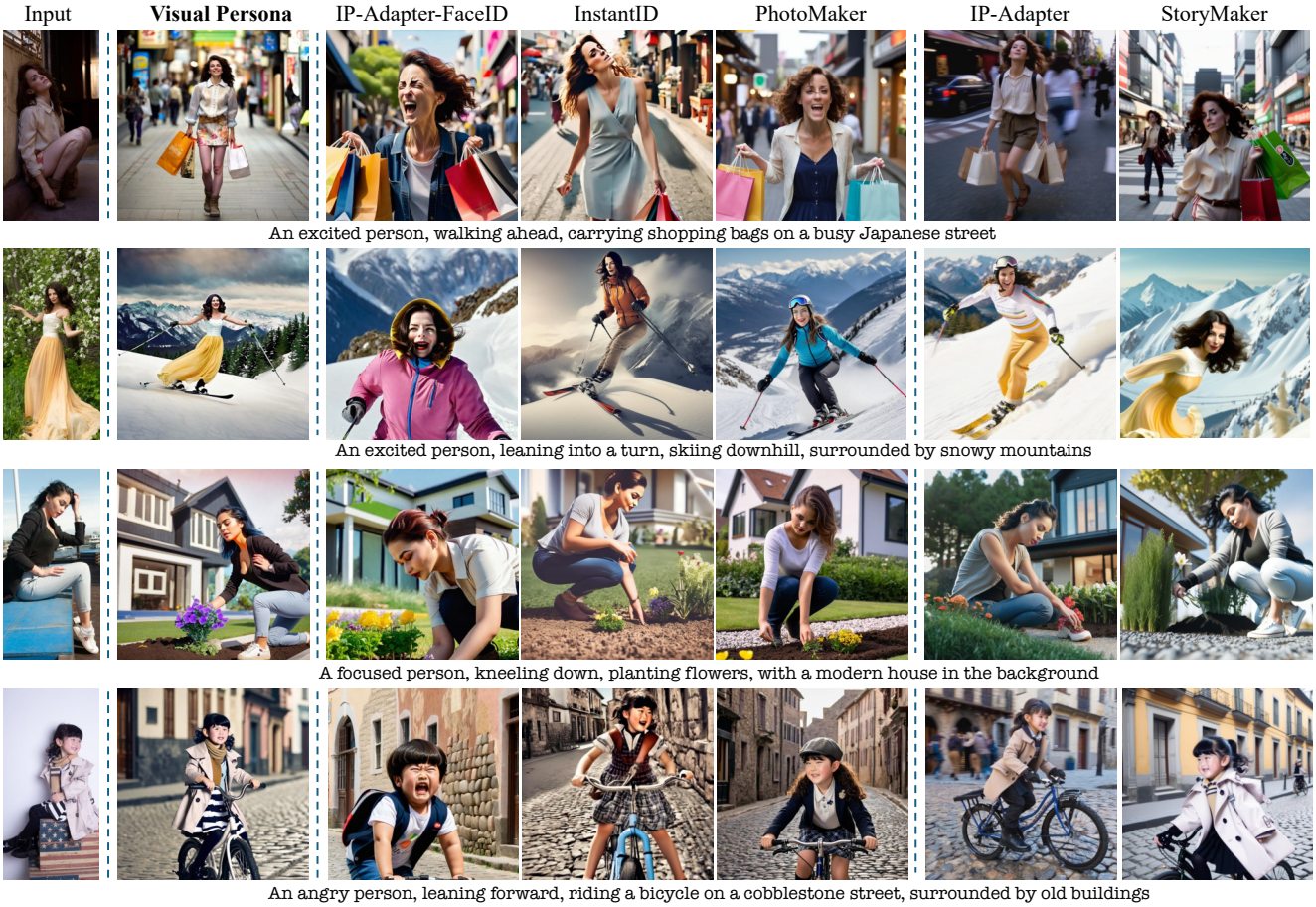


Figure 5. **Qualitative Comparison on PPR10K [42]:** Compared to prior works that focus on face identity preservation [41, 75, 81] or fail to capture the input's detailed appearance [81, 84], Visual Persona accurately preserves the full-body appearance while generating diverse images based on text prompts.

Method	SSHQ			PPR10K		
	D-I ↑	D-T ↑	D-H ↑	D-I ↑	D-T ↑	D-H ↑
IP-Adapter-FaceID [81]	1.78	7.50	2.76	1.86	7.49	2.81
InstantID [75]	1.52	6.94	2.37	1.70	7.12	2.63
PhotoMaker [41]	1.70	**7.72**	2.64	2.03	**7.64**	3.03
IP-Adapter [81]	5.44	7.26	5.96	5.04	7.38	5.75
StoryMaker [84]	6.74	7.08	6.71	6.80	6.77	6.63
Visual Persona (Ours)	**7.10**	7.15	**6.99**	**7.30**	6.67	**6.85**

Table 2. **Quantitative Comparison.**

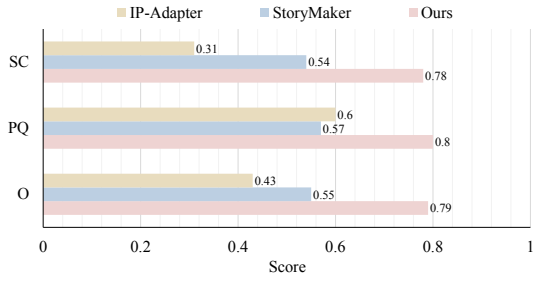


Figure 6. **Human Evaluation.**

group evaluating 150 samples generated by three methods using the same input images and prompts to ensure rating consistency. Further details are in Appendix C.4.

4.2. Results

Comparison. Figure 5 and Table 2 summarize the qualitative and quantitative comparisons with state-of-the-art customized models [41, 75, 81, 84]. As shown in Figure 5, face customization models [41, 75, 81] exhibit strong text alignment but are limited to the face domain, while full-body customization models [81, 84] often fail to preserve the detailed appearance of inputs and produce foreground-biased outputs. In contrast, Visual Persona accurately retains the input's full-body appearance while adapting to complex body deformations and scene changes driven by the text prompts. Table 2 further shows that Visual Persona significantly outperforms previous methods in identity preservation (D-I), while maintaining comparable text alignment (D-T), ultimately achieving the best harmonic mean (D-H). More results are provided in Appendix F.

Human Evaluation. We present the human evaluation re-

Figure 7. **Component Analysis.**

	Component	PPR10K D-I ↑	PPR10K D-T ↑	PPR10K D-H ↑
(I)	MLP	6.66	**7.11**	6.74
(II)	(I) + Self-Attention	6.54	7.01	6.63
(III)	(II) + Cross-Attention	**7.47**	6.13	6.40
(IV)	(III) + Body Part Decomposition	7.30	6.67	**6.85**

Table 3. **Component Analysis.**

sults in Figure 6. Unlike previous methods that are biased toward text alignment [81] or produce artificial outputs [84], Visual Persona surpasses these approaches in SC, which concurrently measures identity preservation and text alignment, as well as in PQ, which evaluates image quality, achieving the highest overall score O.

4.3. Ablation Study

Component Analysis. Figure 7 and Table 3 summarize the effectiveness of different configurations for the body-partitioned transformer decoder. (I) presents the results solely using an MLP layer on DINOv2 [51] features, while (II) shows the results with the addition of a self-attention layer. (III) indicates the results of adding a cross-attention layer, which maps the features onto identity embeddings. Compared to (I) and (II), which fail to condense detailed appearance from redundant input features, (III) shows that the transformer enhances identity preservation (D-I 6.66 vs. 7.47) but sacrifices text alignment (D-T 7.11 vs. 6.13). Compared to (III), (IV) demonstrates that body part decomposition retains the transformer's identity preservation (D-I 7.47 vs. 7.30) while significantly improving text alignment (D-T 6.13 vs. 6.67), achieving the highest D-H (6.85). This indicates that breaking down body regions is essential for the diffusion model to attend to each part separately, enabling diverse body structures in customized images.

Identity Embedding Token Length Analysis. Table 4 presents an ablation study on the identity embedding token length (l_H), showing that increasing token length proportionally enhances identity preservation (D-I) while keeping text alignment (D-T) consistent, ultimately improving the harmonic mean (D-H). This underscores the importance of dense identity embeddings for full-body identity preservation, setting our approach apart from previous works [16, 18, 41, 53, 79, 81, 84] that compress input features into small token embeddings, typically with $l_H = 16$.

Robustness to Model Components. In Table 5, we further demonstrate Visual Persona's robustness to body parsing by comparing results obtained using body parsing masks from the evaluation dataset [14, 42] with those generated by the off-the-shelf method SCHP [39], showing that our method generalizes across different body parsing masks.

Token Length	PPR10K D-I ↑	PPR10K D-T ↑	PPR10K D-H ↑
$l_H = 4 \times 4$	5.51	**6.90**	5.81
$l_H = 8 \times 8$	6.56	6.50	6.52
$l_H = 16 \times 16$ (Ours)	**7.30**	6.67	**6.85**

Table 4. **Identity Embedding Token Length (l_H) Analysis.**

Method	Body Parsing Method	SSHQ D-I ↑	SSHQ D-T ↑	SSHQ D-H ↑	PPR10K D-I ↑	PPR10K D-T ↑	PPR10K D-H ↑
StoryMaker	SSHQ [14] / PPR10K [42]	6.74	7.08	6.71	6.80	6.77	6.63
Visual Persona	SSHQ [14] / PPR10K [42]	7.10	**7.15**	6.99	**7.30**	6.67	6.85
Visual Persona	SCHP [39]	**7.20**	**7.15**	**7.10**	7.10	**7.12**	**7.00**

Table 5. **Analysis of Robustness to Different Body Parsing.**

4.4. Applications

Figure 1 demonstrates the versatility of Visual Persona across various downstream tasks. Although not designed for VTON, Figure 1(a) demonstrates that Visual Persona naturally supports text-guided VTON, unlike existing VTON models [9, 22, 27, 33, 50, 80, 85, 86], which are limited to minor scene and pose changes due to the absence of text-based control. Specifically, given an input image for facial identity and additional images for garments, we apply a body parsing model [39] to segment the face and garments and use them as inputs to our model. In this experiment, full-body images are not used. We further provide a comparison of our method with state-of-the-art VTON model, Leffa [85] in Appendix E. (b) displays human stylization results based on text prompts, effectively altering the image style while retaining full-body appearance. (c) showcases the robustness of Visual Persona with out-of-domain inputs not included in the training set, successfully producing visually consistent outputs for the anime-style input. In Appendix E, we present additional application results, including multi-person customization, part-guided generation, and consistent story generation.

5. Conclusion

In this paper, we introduce **Visual Persona**, a foundation model for full-body human customization. To address the difficulty in obtaining paired human datasets, we propose a data curation pipeline to collect paired data with full-body consistency from a large pool of unpaired images. We further design a transformer encoder-decoder adapted to a pre-trained T2I diffusion model, enabling precise visual transfer by projecting full-body appearance into dense identity representations. Combined, Visual Persona surpasses state-of-the-art customization methods in both GPT-based and human evaluations, while demonstrating versatility across various downstream tasks.

6. Acknowledgement

We would like to express our gratitude to Aashish Misraa at Adobe for collecting the unpaired human dataset and for his valuable insights and discussions on using DINO features. This research was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190075, RS-2024-00509279, RS-2025-II212068, RS-2023-00227592) and the Culture, Sports, and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism (RS-2024-00345025, RS-2024-00333068), and National Research Foundation of Korea (RS-2024-00346597).

References

[1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 2, 4

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 4, 6

[3] Kiymet Akdemir and Pinar Yanardag. Oracle: Leveraging mutual information for consistent character generation with loras in diffusion models. *arXiv preprint arXiv:2406.02820*, 2024. 2

[4] Yuval Alaluf, Elad Richardson, Gal Metzer, and Daniel Cohen-Or. A neural space-time representation for text-to-image personalization. *ACM Transactions on Graphics (TOG)*, 42(6):1–10, 2023. 3

[5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 2, 4

[6] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019. 6

[7] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6593–6602, 2024. 3

[8] Xiyi Chen, Marko Mihajlovic, Shaofei Wang, Sergey Prokudin, and Siyu Tang. Morphable diffusion: 3d-consistent diffusion for single-image avatar creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10359–10370, 2024. 2

[9] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for virtual try-on. *arXiv preprint arXiv:2403.05139*, 2024. 8

[10] Dimitrios Christodoulou and Mads Kuhlmann-Jørgensen. Finding the subjective truth: Collecting 2 million votes for comprehensive gen-ai model evaluation. *arXiv preprint arXiv:2409.11904*, 2024. 6

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 3

[12] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020. 3

[13] Ziyi Dong, Pengxu Wei, and Liang Lin. Dreamartist: Towards controllable one-shot text-to-image generation via positive-negative prompt-tuning. *arXiv preprint arXiv:2211.11337*, 2022. 3

[14] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. In *European Conference on Computer Vision*, pages 1–19. Springer, 2022. 6, 8

[15] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2, 3

[16] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)*, 42(4):1–13, 2023. 2, 3, 4, 5, 6, 8

[17] Rinon Gal, Or Lichter, Elad Richardson, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Lcm-lookahead for encoder-based text-to-image personalization. *arXiv preprint arXiv:2404.03620*, 2024. 2, 3, 6

[18] Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, and Qian He. Pulid: Pure and lightning id customization via contrastive alignment. *arXiv preprint arXiv:2404.16022*, 2024. 2, 3, 4, 5, 6, 8

[19] Gilles Hacheme and Noureini Sayouti. Neural fashion image captioning: Accounting for data diversity. *arXiv preprint arXiv:2106.12154*, 2021. 6

[20] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7323–7334, 2023. 3

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[22] Sen He, Yi-Zhe Song, and Tao Xiang. Style-based global appearance flow for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3470–3479, 2022. 8

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 6

[24] Hexiang Hu, Kelvin CK Chan, Yu-Chuan Su, Wenhu Chen, Yandong Li, Kihyuk Sohn, Yang Zhao, Xue Ben, Boqing Gong, William Cohen, et al. Instruct-imagen: Image generation with multi-modal instruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4754–4763, 2024. 6

[25] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2020. 3

[26] Chuong Huynh, Seoung Wug Oh, Abhinav Shrivastava, and Joon-Young Lee. Maggie: Masked guided gradual human instance matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3870–3879, 2024. 4

[27] Boyuan Jiang, Xiaobin Hu, Donghao Luo, Qingdong He, Chengming Xu, Jinlong Peng, Jiangning Zhang, Chengjie Wang, Yunsheng Wu, and Yanwei Fu. Fitdit: Advancing the authentic garment details for high-fidelity virtual try-on. *arXiv preprint arXiv:2411.10499*, 2024. 8

[28] Dongsheng Jiang, Yuchen Liu, Songlin Liu, Jin'e Zhao, Hao Zhang, Zhen Gao, Xiaopeng Zhang, Jin Li, and Hongkai Xiong. From clip to dino: Visual encoders shout in multi-modal large language models. *arXiv preprint arXiv:2310.08825*, 2023. 5

[29] Dongfu Jiang, Max Ku, Tianle Li, Yuansheng Ni, Shizhuo Sun, Rongqi Fan, and Wenhu Chen. Genai arena: An open evaluation platform for generative models. *arXiv preprint arXiv:2406.04485*, 2024. 6

[30] Siyoon Jin, Jisu Nam, Jiyoung Kim, Dahyun Chung, Yeong-Seok Kim, Joonhyung Park, Heonjeong Chu, and Seungryong Kim. Appearance matching adapter for exemplar-based semantic image synthesis. *arXiv preprint arXiv:2412.03150*, 2024. 2

[31] Tero Karras. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 3

[32] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3

[33] Jeongho Kim, Guojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8176–8185, 2024. 8

[34] Kihong Kim, Yunho Kim, Seokju Cho, Junyoung Seo, Jisu Nam, Kychul Lee, Seungryong Kim, and KwangHee Lee. Diffface: Diffusion-based face swapping with facial guidance. *Pattern Recognition*, page 111451, 2025. 3

[35] Seyeon Kim, Siyoon Jin, Jihye Park, Kihong Kim, Jiyoung Kim, Jisu Nam, and Seungryong Kim. Moditalker: Motion-disentangled diffusion model for high-fidelity talking head generation. *arXiv preprint arXiv:2403.19144*, 2024. 2

[36] Max Ku, Tianle Li, Kai Zhang, Yujie Lu, Xingyu Fu, Wenwen Zhuang, and Wenhu Chen. Imagenhub: Standardizing the evaluation of conditional image generation models. *arXiv preprint arXiv:2310.01596*, 2023. 6

[37] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 3

[38] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, et al. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024. 6

[39] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3260–3271, 2020. 3, 4, 8

[40] Shaoxu Li. Aniartavatar: Animatable 3d art avatar from a single image. *arXiv preprint arXiv:2403.17631*, 2024. 2

[41] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8640–8650, 2024. 2, 3, 4, 5, 7, 8

[42] Jie Liang, Hui Zeng, Miaomiao Cui, Xuansong Xie, and Lei Zhang. Ppr10k: A large-scale portrait photo retouching dataset with human-region mask and group-level consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 653–661, 2021. 6, 7, 8

[43] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer, 2025. 6

[44] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2, 4

[45] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 6

[46] Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. Videodrafter: Content-consistent multi-scene video generation with llm. *arXiv preprint arXiv:2401.01256*, 2024. 2

[47] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 2, 4

[48] Jisu Nam, Gyuseong Lee, Sunwoo Kim, Hyeonsu Kim, Hyoungwon Cho, Seyeon Kim, and Seungryong Kim. Diffusion model for dense matching. *arXiv preprint arXiv:2305.19094*, 2023. 2

[49] Jisu Nam, Heesu Kim, DongJae Lee, Siyoon Jin, Seungryong Kim, and Seunggyu Chang. Dreammatcher: Appearance matching self-attention for semantically-consistent text-to-image personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8100–8110, 2024. 2

[50] Shuliang Ning, Duomin Wang, Yipeng Qin, Zirong Jin, Baoyuan Wang, and Xiaoguang Han. Picture: Photorealistic virtual try-on from unconstrained designs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6976–6985, 2024. 8

[51] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3, 5, 8

[52] Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Jiankang Deng, Bernhard Kainz, and Stefanos Zafeiriou. Arc2face: A foundation model of human faces. *arXiv preprint arXiv:2403.11641*, 2024. 2, 3, 4, 5, 6

[53] Xu Peng, Junwei Zhu, Boyuan Jiang, Ying Tai, Donghao Luo, Jiangning Zhang, Wei Lin, Taisong Jin, Chengjie Wang, and Rongrong Ji. Portraitbooth: A versatile portrait model for fast identity-preserved personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27080–27090, 2024. 2, 3, 4, 5, 6, 8

[54] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. *arXiv preprint arXiv:2406.16855*, 2024. 3, 6

[55] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 4, 5

[57] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[58] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 2, 3, 6

[59] Simo Ryu. Low-rank adaptation for fast text-to-image diffusion finetuning. https://github.com/cloneofsimo/lora, 2023. 3

[60] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2

[61] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 3

[62] Sefik Serengil and Alper Ozpinar. A benchmark of facial recognition pipelines and co-usability performances of modules. *Journal of Information Technologies*, 17(2):95–107, 2024. 3

[63] Sefik Ilkin Serengil and Alper Ozpinar. Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 23–27. IEEE, 2020. 3

[64] Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4. IEEE, 2021. 4

[65] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8543–8552, 2024. 2, 3, 4, 5, 6

[66] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 6

[67] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 3

[68] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, 2017. 3

[69] Zhiyu Tan, Xiaomeng Yang, Luozheng Qin, Mengping Yang, Cheng Zhang, and Hao Li. Evalalign: Evaluating text-to-image models through precision alignment of multimodal large models with supervised fine-tuning to human annotations. *arXiv preprint arXiv:2406.16562*, 2024. 6

[70] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2, 4

[71] Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *ACM Transactions on Graphics (TOG)*, 43(4):1–18, 2024. 2

[72] Dani Valevski, Danny Lumen, Yossi Matias, and Yaniv Leviathan. Face0: Instantaneously conditioning a text-to-image model on a face. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–10, 2023. 2, 3, 4, 5, 6

[73] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 4

[74] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. p+: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023. 3

[75] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 2, 3, 4, 5, 6, 7

[76] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 2, 4

[77] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15943–15953, 2023. 3

[78] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018. 6

[79] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023. 2, 3, 4, 5, 6, 8

[80] Xu Yang, Changxing Ding, Zhibin Hong, Junhao Huang, Jin Tao, and Xiangmin Xu. Texture-preserving diffusion models for high-fidelity virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7017–7026, 2024. 8

[81] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2, 3, 4, 5, 6, 7, 8

[82] Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. Celebv-text: A large-scale facial text-video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14805–14814, 2023. 3

[83] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36, 2024. 5

[84] Zhengguang Zhou, Jing Li, Huaxia Li, Nemo Chen, and Xu Tang. Storymaker: Towards holistic consistent characters in text-to-image generation. *arXiv preprint arXiv:2409.12576*, 2024. 2, 3, 4, 5, 6, 7, 8

[85] Zijian Zhou, Shikun Liu, Xiao Han, Haozhe Liu, Kam Woh Ng, Tian Xie, Yuren Cong, Hang Li, Mengmeng Xu, Juan-Manuel Pérez-Rúa, et al. Learning flow fields in attention for controllable person image generation. *arXiv preprint arXiv:2412.08486*, 2024. 8

[86] Luyang Zhu, Yingwei Li, Nan Liu, Hao Peng, Dawei Yang, and Ira Kemelmacher-Shlizerman. M&m vto: Multi-garment virtual try-on and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1346–1356, 2024. 8