Visual Persona: Foundation Model for Full-Body Human Customization

Supplementary Material

In the following, we additionally explain the details of implementation in Sec. A, the detailed curation process of Visual Persona-500K in Sec. B, and the evaluation details in Sec. C, covering comparison studies, existing metric analysis, GPT-based evaluation, human evaluation on facial expressions, and human evaluation details. Further analysis of Visual Persona, including a detailed comparison with StoryMaker [44], is presented in Sec. D. We further provide more application results and analysis in Sec. E. Additional qualitative results are included in Sec. F, and limitations are discussed in Sec. G.

A. Implementation Details

We used a pre-trained SDXL model [35] for text-to-image generation at a resolution of 1024×1024. We first trained our model in a reconstruction manner on an unpaired human dataset, using the same image for X and Y in Equation 6, for 35,000 steps with a batch size of 32 and a learning rate of $1e-4$. This was followed by fine-tuning on the 580K paired human dataset, Visual Persona-500K, using paired images for X and Y for 35,000 additional steps with a batch size of 8 and a learning rate of $5e-6$. For GPT-based evaluation [34], we used GPT-4o-mini [32] for all evaluations. We set $\lambda = 1$ for training and $\lambda = 0.7$ for all evaluations. All experiments were conducted on 8 NVIDIA A100 GPUs using the Adam optimizer [22]. All input images for character customization are AI-generated images [7, 35].

B. Visual Persona-500K Data Curation Details

The pipeline for curating consistent full-body identities is illustrated in Figure A.1. From the collected unpaired human pool, comprising multiple images per individual that only guarantee facial identity consistency, we aim to further evaluate body consistency using the VLM [29]. For efficiency, we begin with two randomly selected images of the same individual and prompt LLAVA [29] to assess whether the individual in both images is wearing the same outfit (Figure A.1(a)). A simple prompt—*"Are they wearing exactly the same clothes?"*—enables the model to provide a binary decision with high precision. If the model returns a positive response, the individual is retained for further processing; otherwise, the individual is excluded from the dataset.

To ensure full-body consistency across all images for each retained individual, we further refine the dataset using a sliding window approach (Figure A.1(b)). Given our observation that LLaVA [29] can compare up to three images, we concatenate consecutive sets of three images with a window size of 3 and a stride of 2, evaluating all sets for the same individual. If consistency is maintained across all sets, the individual is retained; otherwise, they are excluded. Ultimately, we curated a dataset of 580k paired human images across 100k unique individuals.

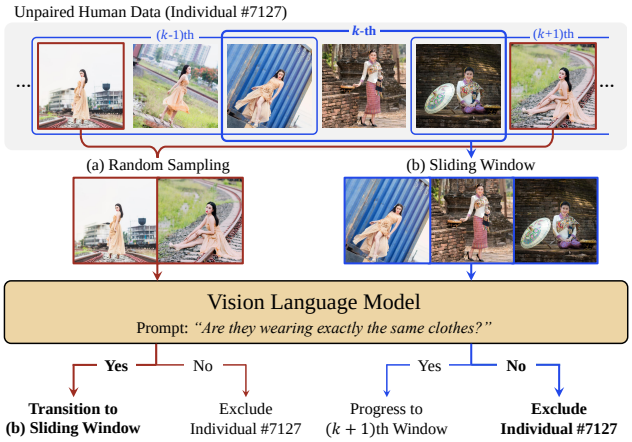


Figure A.1. **Curating Consistent Full-Body Identities.**

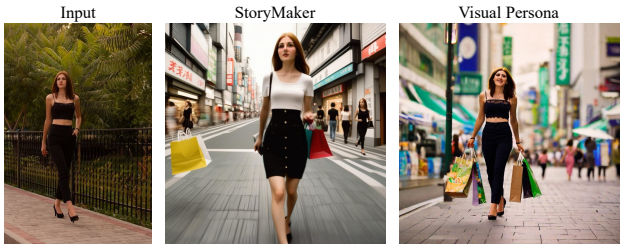
C. Evaluation

C.1. Comparison

We benchmark our method against recent encoder-based zero-shot human customization models [26, 41, 42, 44]. These methods often focus on human face generation (IP-Adapter-FaceID [42], InstantID [41], PhotoMaker [26]) or attempt to generate full-body images but are limited to reconstructing a single image per individual (IP-Adapter [42], StoryMaker [44]).

Specifically, we compare the following open-source models built on SDXL [35]:

- **IP-Adapter-FaceID-SDXL (IP-Adapter-FaceID)** [42]: Embeds facial features extracted from a face recognition model [9, 12] into small identity token embeddings, conditioning the pre-trained T2I diffusion model through a decoupled attention mechanism.
- **InstantID** [41]: Extends IP-Adapter-FaceID by incorporating ControlNet [43] to add spatial control using facial keypoints. InstantID is trained on a dataset of 50M LAION-Face [38] images and 10M face-annotated images collected internally from the web.
- **PhotoMaker** [26]: Stacks CLIP [36] features from multiple face images and combines them with text embeddings to condition the T2I diffusion model. PhotoMaker is trained on a curated dataset of 112K images featuring 13K celebrities collected from the web.



Prompt: An excited person walks ahead, carrying shopping bags on a busy Japanese street

Metric	StoryMaker	Visual Persona
I_{DINO}	**0.531**	0.512
I_{CLIP}	**0.638**	0.628
D-I	2	**8**
Human-I	0.40	**0.97**
T_{CLIP}	**0.326**	0.319
D-T	**8**	**8**
Human-T	0.67	**0.97**

Figure A.2. **GPT-based Metrics Align Better with Human Preferences:** The upper part of the table presents evaluations for identity preservation (I_{DINO} [33], I_{CLIP} [18], D-I, Human-I), while the lower part presents evaluations for text alignment (T_{CLIP}, D-T, Human-T). Prior metrics (I_{DINO}, I_{CLIP}, T_{CLIP}) fail to align with human preferences (Human-I, Human-T) because they calculate cosine distances only between global feature vectors from generated images and given conditions. In contrast, GPT-based evaluations (D-I, D-T) better align with human preferences (Human-I, Human-T).

- **IP-Adapter-Plus-SDXL (IP-Adapter)** [42]: Extends the original IP-Adapter [42] by using patch image embeddings from OpenCLIP-ViT-H-14 [16].
- **StoryMaker** [44]: Combines facial features from ArcFace [8] and portrait features from CLIP [36], mapping them into small identity embeddings while fine-tuning a subset of parameters in the diffusion U-Net. StoryMaker is trained on an internally collected unpaired dataset of 500K human images, including 300K single-character and 200K two-character images. StoryMaker is concurrent work with ours.

C.2. Dataset

To evaluate our method on SSHQ [10], following its instructions, we completed the data release agreement and obtained permission for the non-commercial use of the dataset. To minimize the influence of off-the-shelf foreground mask generators [15, 21, 25], we used the foreground masks provided by SSHQ [10] and PPR10K [27] for evaluating all methods in this paper.

To assess text alignment, we augmented 17 prompts for live objects in Dreambooth [37] using ChatGPT [2], specifically tailored for full-body human customization to follow the template *"A photo of a {facial expression} person, {pose}, {action}, and {surrounding}"*. The generated prompt list is provided in Figure A.11. This prompt list was used for all evaluations.

C.3. Metric

GPT-based Evaluation. As discussed in [6, 19, 24, 28, 39], existing metrics, including identity preservation metrics, DINO image similarity (I_{DINO}) [33], CLIP image similarity (I_{CLIP}) [18], and the text alignment metric, CLIP image-text similarity (T_{CLIP}) [18], often fail to align with human preferences, struggling to accurately evaluate local appearance transfer (I_{DINO}, I_{CLIP}) and the alignment of complex human body structures with the given prompts (T_{CLIP}). This limitation is demonstrated in Figure A.2, where Visual Persona achieves higher human preference scores in identity preservation (Human-I) and text alignment (Human-T), yet existing metrics (I_{DINO}, I_{CLIP}, T_{CLIP}) assign higher scores to StoryMaker [44] across all three metrics. This discrepancy arises because these metrics extract global vectors from the generated images and the given conditions (input image or text prompt) and calculate the distances between them, thereby ignoring local appearance details, intricate human poses, actions, and surrounding elements in the images. For human evaluation (Human-I, Human-T) in this comparison, 30 human raters were recruited to assess identity preservation and text alignment using a scale of $\{0, 0.5, 1\}$ for not aligned, partially aligned, and fully aligned, respectively.

To address this issue, we adopt Dreambench++ [34], a human-aligned, automated, GPT [2]-based evaluation benchmark designed for customized image generative models. Figure A.2 shows that GPT-based evaluation scores for identity preservation and text alignment, denoted as D-I and D-T respectively, align more closely with human preferences compared to previous metrics. Specifically, Dreambench++ [34] provides evaluation instructions as user prompts to GPT, which include the task description, scoring criteria, scoring range, and format specifications. We tailored the task description and scoring criteria for full-body human customization and adjusted the scoring range from [0, 4] to [0, 9] to enable a more comprehensive evaluation. The complete evaluation instructions for identity preservation and text alignment are provided in Figure A.12 and Figure A.14, respectively.

To align the user's instructions with GPT's pre-trained knowledge, Dreambench++ [34] asks GPT to confirm its understanding of the task and to summarize the task itself. This process facilitates GPT's internal reasoning, enhancing task understanding and alignment with user instructions. Dreambench++ achieves this by incorporating GPT's summary and planning responses as assistant prompts, which summarize the user instructions and outline the evaluation protocol based on the given instructions. The complete assistant prompts for identity preservation and text alignment are presented in Figure A.13 and Figure A.15. Note that we can further prompt GPT to output the analysis process for the scores. In Figure A.16, A.17, A.18 and A.19, we also

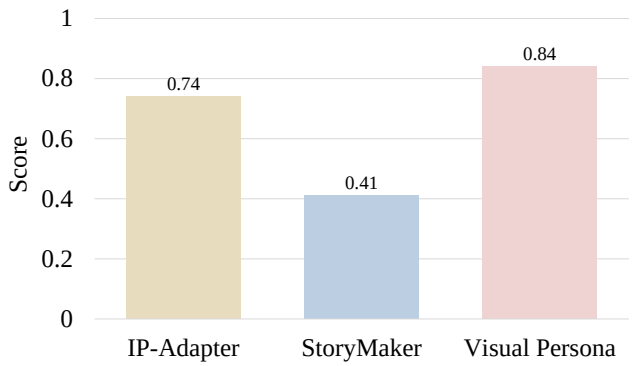


Figure A.3. **Human Evaluation on Facial Expression:** Visual Persona outperforms prior works [42, 44] in text alignment related to facial expression.

provide GPT's analysis procedure for evaluating the samples generated from StoryMaker [44] and Visual Persona shown in Figure A.2.

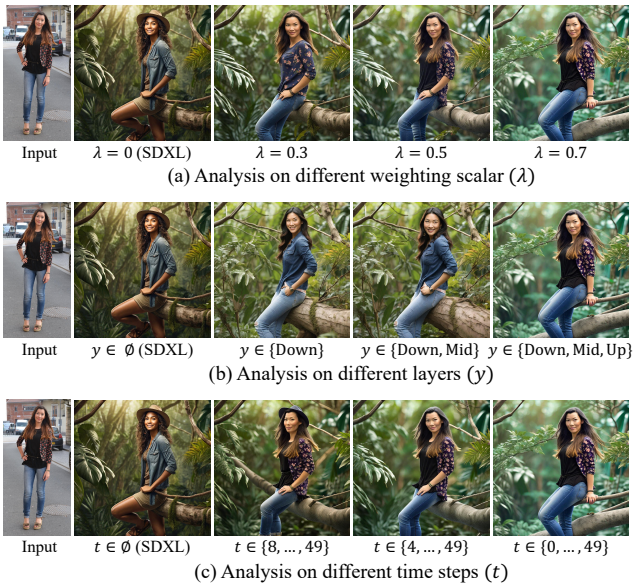
Human Evaluation on Facial Expression. We observed that GPT [32] often fail to detect facial expressions when the subject is positioned far from the foreground center. To evaluate text alignment for facial expression-related prompts, we conducted a human evaluation, with the results presented in Figure A.3. Eight human raters assessed whether the subject's facial expression in the generated image aligned with the given prompts. Scores were assigned as follows: 0 for no alignment, 0.5 for partial alignment, and 1 for perfect alignment.

The raters were divided into two groups, each assessing 150 images generated by three different methods: IP-Adapter [42], StoryMaker [44], and Visual Persona. The same input images and prompts were used across all methods to ensure intra-rater reliability.

Compared to StoryMaker [44], IP-Adapter [42] and Visual Persona demonstrate superior alignment with facial expression prompts. This difference arises because StoryMaker [44] employs ArcFace loss [8], which often leads to overfitting to the pose and expression of the input image, while IP-Adapter [42] does not account for facial expression in the text prompt during training. In contrast, Visual Persona captures facial expressions through detailed text descriptions generated by Phi-3 [1] (Sec. 4.1), without relying on facial loss, enabling it to generate diverse facial expressions while maintaining facial identity consistency.

C.4. Human Evaluation

Human Evaluation Metrics. For rigorous human evaluation, we followed the ImagenHub [23] evaluation protocol, which standardizes the assessment of conditional image generative models. ImagenHub [23] defines two human evaluation scores: Semantic Consistency (SC) and Perceptual Quality (PQ).



(a) Analysis on different weighting scalar (λ)

(b) Analysis on different layers (y)

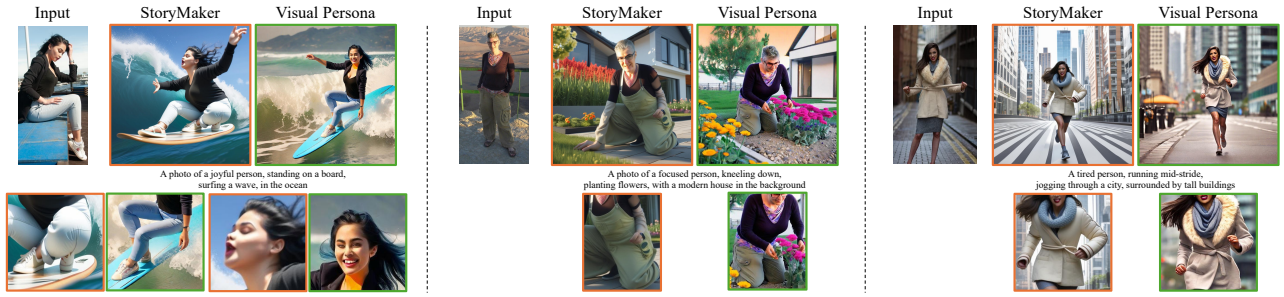
(c) Analysis on different time steps (t)

Figure A.4. **Analysis: Identity Cross-Attention Module.** Users can balance identity preservation and text alignment by adjusting the weighting scalar λ, layers y, and time steps t. Increasing the weighting scalar λ and using later layers y and time steps t better preserve the image structure and layout from the pre-trained SDXL [35], while slightly compromising identity preservation from the input.

Semantic Consistency (SC) evaluates how well a generated image aligns with the provided conditions. Since our method uses an input image and a text prompt as conditions, human raters assess SC based on identity preservation relative to the input image and text alignment with the prompt. SC scores each condition independently as "inconsistent" (0 points), "partially consistent" (0.5 points), or "mostly consistent" (1 point). The final SC score for an image is the lowest score across the two conditions. We highlight that this metric avoids bias toward either the input image or the text prompt, as it prioritizes the lowest score, aligning with our goal of achieving both identity preservation and text alignment. Note that we applied foreground masks [10, 27] to the input images to assist human raters in focusing on the human parts.

Perceptual Quality (PQ) measures how visually convincing and natural the generated image appears, considering artifacts, distortions, and overall realism. Human raters assign 0 points if the image contains obvious artifacts or distortions, 0.5 points if the image appears unnatural with minor artifacts, and 1 point if the image looks genuine and realistic.

Number of Human Raters. Based on ImagenHub's analysis [23] showing that involving more than four human raters increases standard deviation and decreases score reliability, as measured by Krippendorff's Alpha [11], we recruited eight raters, split into two groups, each including four raters.



(a) Overfitting to Identity-Unrelated Attributes (Pose, Facial Expression) (b) Cloth Details Preservation (c) Unnatural Synthetic Cloth Texture

Figure A.5. **Comparison between StoryMaker [44] (orange) and Visual Persona (green), including full and zoomed-in images:** Compared to StoryMaker, Visual Persona enables large deformations, including pose and facial expression variations, preserves clothing details, and generates realistic clothing textures.

Method	(a) Training		(b) Model Architecture		
	Dataset	Strategy	Body Part Decomposition	Encoder	Decoder
StoryMaker [44]	Unpaired	Reconstruction	2 (Face, Body)	CLIP [36], buffalo_l [8]	Resampler, Linear
Visual Persona	Paired	Cross-Image	5 (Full-Body, Face, Torso, Legs, Shoes)	DINO [33]	Transformer Decoder

Table A.1. **Comparison between StoryMaker [44] and Visual Persona.**

Each group was assigned the same evaluation sheet to ensure inter-rater consistency.

Human Evaluation Setting. Each evaluation sheet includes 150 images generated from 50 individuals sampled from the SSHQ [10] and PPR10K [27], with 25 individuals from each dataset. We used two distinct evaluation sheets, covering a total of 300 unique images with no overlap between sheets. For each individual, one prompt was randomly sampled from the 17 pre-defined prompts, and images were generated using three different methods: IP-Adapter [42], StoryMaker [44], and Visual Persona. This setup ensures intra-rater reliability, as the same rater evaluates all three methods on the same input images and text prompts. We provided detailed evaluation guidelines to the human raters. The guidelines and an example of an evaluation question are shown in Figures A.20 and A.21.

D. Analysis

D.1. Identity Cross-Attention Module

Weighting Scalar. Figure A.4(a) presents an ablation study on the weighting scalar λ in Equation 5. $\lambda = 0$ indicates that the identity cross-attention is disabled, which is identical to the original SDXL [35]. For a fair comparison, we fix the layers and time steps for identity cross-attention to include all cross-attention layers in SDXL and all 50 sampling time steps.

The results show that increasing λ enhances identity preservation from the input but slightly degrades the original image structure in pre-trained SDXL, including background details and human pose. This suggests that users can control λ to balance the degree of identity preservation from the input and the text alignment derived from SDXL.

Layers. Figure A.4(b) presents an ablation study on different layers in SDXL [35], denoted as y, where the identity cross-attention module is applied. Here, Down, Mid, and Up refer to the down blocks, mid block, and up blocks in the diffusion U-Net, respectively. For a fair comparison, we set $\lambda = 0.7$ across all 50 sampling time steps.

As discussed in [4, 31], the down blocks primarily capture image structure and layout, including background details and human pose, while the up blocks focus on image appearance. In line with this, applying identity cross-attention in the down and mid blocks significantly limits identity injection, while preserving the original image structure generated by the pre-trained SDXL. On the other hand, adding identity cross-attention to the up blocks effectively injects identity while maintaining the pre-trained SDXL image structure. Based on this observation, we utilize all cross-attention layers in the identity cross-attention module for all evaluations presented in this paper.

Time Steps. Figure A.4(c) shows the results of using the identity cross-attention module at different sampling time steps t. Since earlier time steps focus on producing image structure, while later time steps refine image appearance [4, 31], applying identity cross-attention at later time steps better preserves the original SDXL [35] image structure but compromises identity preservation. In our experiment, $t \in \{4, ..., 49\}$ achieves strong identity preservation while maintaining the original image structure of the pre-trained SDXL. This demonstrates that users can adjust the sampling time steps to balance identity injection and generative fidelity. In our main paper, for a fair comparison with other works, we used all sampling time steps $t \in \{0, ..., 49\}$

Figure A.6. **Comparison for multi-person customization between StoryMaker [44] (orange) and Visual Persona (green):** Compared to StoryMaker, Visual Persona generates more realistic interactions between multiple individuals while preserving the full-body identity of each person. Notably, Visual Persona is not trained with a multi-person dataset, as used in StoryMaker, yet our method enables multi-person customization through a simple inference modification.

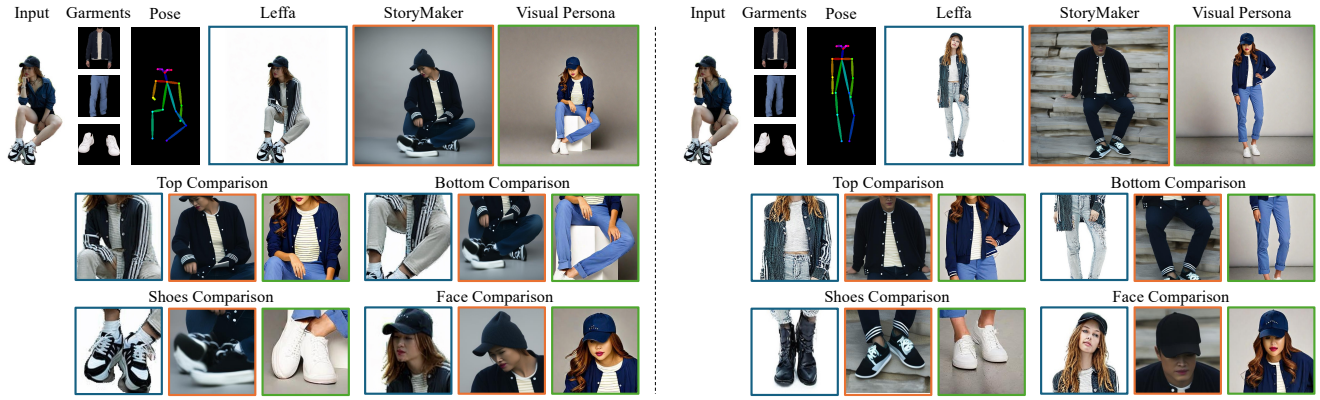


Figure A.7. **Comparison for VTON between Leffa [45] (blue), StoryMaker [44] (orange), and Visual Persona (green), including full and zoomed-in images:** Compared to Leffa and StoryMaker, Visual Persona enables more flexible VTON, including top, bottom, and shoes, preserves the details of each garment, and allows accurate pose control.

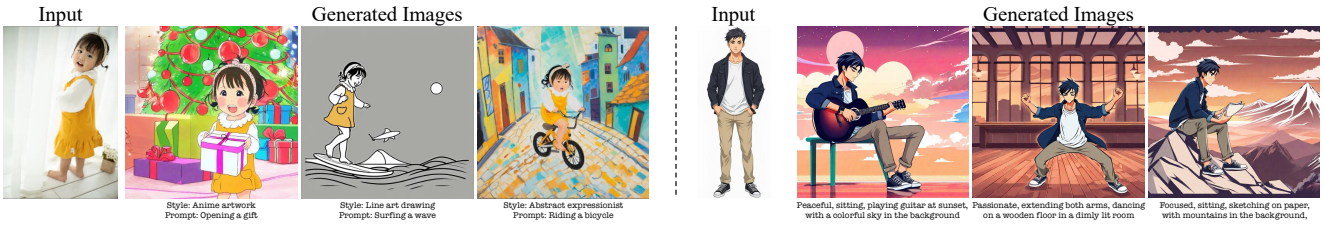
for all evaluations.

D.2. Detailed Comparison with StoryMaker

In Figure A.5 and Table A.1, we provide detailed comparisons of Visual Persona with StoryMaker, which is a concurrent work to ours. As presented in Figure A.5(a) and Table A.1(a), StoryMaker relies on reconstruction training with an unpaired dataset, which often leads to overfitting to human location, pose, and facial expressions. In contrast, our method uses cross-image training on a curated paired dataset, enabling large deformations, including pose and facial expressions, aligned with the given text. As presented in Figure A.5(b) and Table A.1(b), StoryMaker encodes two-part inputs with semantic encoders and then compresses them using a resampler and a linear layer, which often lose local details in clothing and fail to disentangle different body parts. In contrast, our fine-grained decomposition and transformer encoder-decoder better preserve each part of the full-body identity. This also limits StoryMaker to top-garment Virtual Try-On (VTON), while ours supports more flexible VTON, which is further discussed in Section E. Additionally, as displayed in Figure A.5(c), StoryMaker often produces synthetic-looking outputs, possibly due to the dataset quality, while our method can generate realistic cloth textures, benefiting from our curated dataset quality.

E. Application

Multi-Person Customization. Figure A.6 shows that Visual Persona supports multi-person customization without requiring the additional multi-person training used by StoryMaker [44]. This is achieved through a simple inference modification, which involves concatenating identity embeddings from multiple inputs, extracting foreground masks for each individual using text cross-attention, and augmenting identity cross-attention with these masks. StoryMaker struggles to generate interactions between multiple individuals (e.g., eye contact between two people). This is because StoryMaker is trained in a reconstruction manner, which often leads to overfitting identity-unrelated attributes from the input images (e.g., face pose, body pose, facial expression) and results in foreground-biased outputs. In contrast, Visual Persona employs cross-image training to mitigate



(a) Human Stylization (b) Character Customization

Figure A.8. **Human Stylization and Character Customization.**

Figure A.9. **Part-Guided Full-Body Generation:** Users can select a single body part from the human image as input, allowing the pre-trained T2I diffusion model to synthesize the remaining body parts, without requiring additional training.

overfitting, producing natural interactions between individuals, seamlessly integrated into the generated scenes. Additionally, StoryMaker often fails to accurately preserve the full-body appearance of each individual, while Visual Persona better retains them, benefiting from the proposed transformer architecture.

Virtual Try-On (VTON). In Figure A.7, we also compare our method with Leffa [45], the state-of-the-art VTON approach, and StoryMaker [44]. Leffa supports only top and bottom garments and requires sequential processing, which often blends garment identities. StoryMaker supports only top garments, as it decomposes the input into only two parts, the face and the whole body. Additionally, StoryMaker often struggles with pose changes and face identity preservation. In contrast, Visual Persona enables fine-grained VTON with parallel body part decomposition and better preserves full-body identity under large pose variations, benefiting from cross-image training and a transformer architecture.

Human Stylization. Figure A.8(a) display human stylization results based on text prompts by our Visual Persona, effectively altering the image style while maintaining the full-body appearance.

Character Customization. Figure A.8(b) showcase the robustness of Visual Persona with out-of-domain inputs (e.g., animation domain) not included in the training set, successfully producing visually consistent outputs for anime-style input.

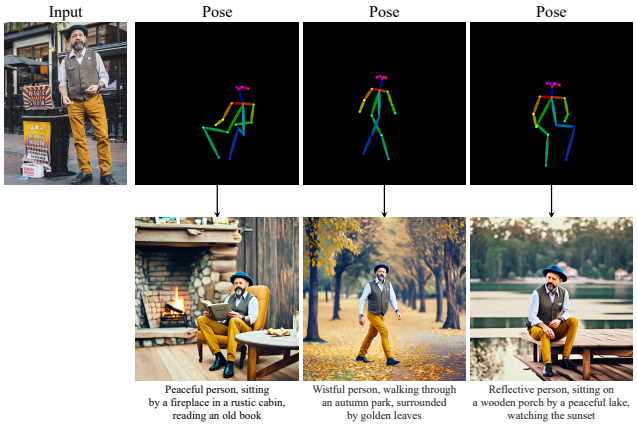


Figure A.10. **Pose-Guided Consistent Story Generation:** Users can generate a consistent story for a given human, guided by an external human pose using ControlNet [43].

Part-Guided Full-Body Generation. Figure A.9 illustrates qualitative results for part-guided full-body generation. In this experiment, we use only one body part image from the given human image as input and allow the pre-trained T2I diffusion model to synthesize the remaining body parts. The results demonstrate that Visual Persona effectively generates diverse human images while preserving the given body part, without requiring additional training, suggesting future applications such as fashion advertisements.

Pose-Guided Consistent Story Generation. Figure A.10 illustrates the consistent story generation of a given human, following the narrative text and guided by the human pose using ControlNet [43]. This further highlights the practicality of our method in film production [20, 30] or book illustration [3, 40].

F. More Results

More qualitative results of Visual Persona on SSHQ [10] and PPR10K [27] are provided in Figure A.22 and Figure A.23. Additional qualitative results on applications, including text-guided virtual try-on, human stylization, and character customization, are provided in Figure A.24 and Figure A.25. Additional qualitative comparison results with [26, 41, 42, 44] are presented in Figure A.26.

A photo of a happy person, standing on a branch, climbing a tree, surrounded by a dense jungle
A photo of an excited person, leaning into a turn, skiing downhill, surrounded by snowy mountains
A photo of a joyful person, standing on a board, surfing a wave, in the ocean
A photo of an angry person, leaning forward, riding a bicycle on a cobblestone street, surrounded by old buildings
A photo of a calm person, clapping hands while doing yoga, surrounded by a lush garden
A photo of a passionate person, extending both arms, dancing on a wooden floor in a dimly lit room
A photo of a tired person, running mid-stride, jogging through a city, surrounded by tall buildings
A photo of a focused person, sitting on a rock, sketching on paper, with mountains in the background
A photo of a focused person, kneeling down, planting flowers, with a modern house in the background
A photo of a satisfied person, sitting on a park bench, eating a sandwich, surrounded by trees
A photo of an excited person, walking ahead, carrying shopping bags on a busy Japanese street
A photo of an amazed person, jumping, placing a flag, surrounded by a barren lunar landscape
A photo of a happy person, standing, taking selfies in New York, surrounded by tall skyscrapers
A photo of a peaceful person, sitting, playing guitar at sunset, with a colorful sky in the background
A photo of a sad person, reclining in a theater, eating popcorn, surrounded by comfortable seats
A photo of a sad person, lounging, reading on a soft couch in a luxurious room
A photo of a calm person, stretching out both arms, floating in space, in a dream of a distant galaxy

Figure A.11. **Evaluation Prompts for Full-Body Human Customization:** To evaluate full-body human customization, we generated 17 text prompts by augmenting the original DreamBooth prompts [37] using ChatGPT [2]. These prompts were utilized for all evaluations in this paper.

Task Definition
You will be provided with an image generated based on a reference image.
As an experienced evaluator, your task is to assess how well the appearance of the human subject is preserved in the generated image compared to the reference image, based on the scoring criteria.
Focus solely on the human subject. Regardless of whether the subject in the generated image differs in size, pose, action, or surroundings compared to the one in the reference image, your evaluation should prioritize the subject's visual appearance.

Scoring Criteria
Assess whether the human subject in the generated image remains consistent with the one in the reference image, focusing on the preservation of fine details across the following five visual features:
1. Clothing Types: Check whether the clothing types in the generated image match those in the reference image. This includes distinctions like short vs. long sleeves, short vs. long pants, and the presence of accessories.
2. Design: Evaluate whether the design of the subject's clothing in the generated image matches that in the reference image. This includes the pattern (e.g., floral, striped, or solid) and decorative elements (e.g., logos, zippers, or pockets). Focus on fine-grained details in the design.
3. Texture: Assess whether the texture of the fabrics worn by the subject in the generated image matches that in the reference image. This includes the material's appearance and quality. Focus on fine details that contribute to realism.
4. Color: Compare the primary colors of the subject's clothing and body in both images, considering hue, saturation, brightness, and overall color distribution.
5. Face Identity: Evaluate whether the subject's face in the generated image resembles the face in the reference image. It is acceptable for the subject in the generated image to have a different expression or pose than in the reference image. The focus should be on whether the facial identity aligns, without expecting an exact replica.

Scoring Range
You need to give a specific integer score based on the comprehensive performance of the visual features above, ranging from 0 to 9:
- Very Poor (0): No resemblance. The generated image's subject has no relation to the reference.
- Poor (1-2): Minimal resemblance. The subject falls within the same broad category but differs significantly in appearance.
- Fair (3-4): Moderate resemblance. The subject shows some likeness to the reference but has notable variances.
- Good (5-6): Strong resemblance. The subject closely matches the reference with only minor discrepancies.
- Very Good (7-8): Very close resemblance. The subject of the generated image is similar to the reference, with few differences in details.
- Excellent (9): Near-identical resemblance. The subject of the generated image is virtually indistinguishable from the reference.

Input format
Every time you will receive two images, the first image is a reference image, and the second image is the generated image.

Please carefully review each image of the subject.

Output Format
Score: [Your Score]

You must adhere to the specified output format, which means that only the scores need to be output, excluding your analysis process.

Figure A.12. **User Prompts for Evaluating Identity Preservation:** To assess identity preservation, we provide evaluation instructions to GPT [32] as user prompts.

Yes, I understand the task. It involves evaluating the visual consistency between a reference image and a generated image based on specific criteria. The evaluation will focus on five key aspects: clothing types, design, texture, color, and face identity. The goal is to assess how closely the subject in the generated image matches the one in the reference image across these aspects. Regardless of the subject's size, pose, action, or surroundings, the evaluation will prioritize the visual appearance of the human subject. The result will be a specific score ranging from 0 (no resemblance) to 9 (near-identical resemblance).

To evaluate the images, I plan to follow these steps:

1. **Clothing Types**: Check whether the clothing types in the generated image match those in the reference image. This includes distinctions like short vs. long sleeves, short vs. long pants, and the presence of accessories.
2. **Design**: Evaluate whether the design of the subject's clothing in the generated image matches that in the reference image. This includes the pattern (e.g., floral, striped, or solid) and decorative elements (e.g., logos, zippers, or pockets). Focus on fine-grained details in the design.
3. **Texture**: Assess the texture of the fabrics worn by the subject in the generated image matches that in the reference image. This includes the material's appearance and quality. Focus on fine details that contribute to realism.
4. **Color**: Compare the primary colors of the subject's clothing and body in both images, considering hue, saturation, brightness, and overall color distribution.
5. **Facial Features**: Evaluate whether the subject's face in the generated image resembles the face in the reference image. It is acceptable for the subject in the generated image to have a different expression or pose than in the reference image. The focus should be on whether the facial identity aligns, without expecting an exact replica.

After analyzing these aspects, I will assign a score based on the overall performance of the generated image in relation to the reference image. The score will reflect how similar the generated image is to the reference, strictly adhering to the evaluation criteria provided.

My output format should be Score: [0-9], and I don't need to write out the specific analysis process.

Please provide me with the samples I need to evaluate.

Figure A.13. **Assistant Prompts for Evaluating Identity Preservation:** To assess identity preservation, we provide summary and planning responses for GPT [32] as assistant prompts.

Task Definition
You will be provided with an image and a text prompt.
As an experienced evaluator, your task is to evaluate the semantic consistency between the image and the text prompt, focusing on human pose, human action, surroundings, composition and image quality, according to the criteria below.

Scoring Criteria
Assess how well the visual content of the image aligns with the text prompt based on the following five key aspects:
1. Human Pose: Assess whether the body pose of the human subject aligns with the pose described in the text (e.g., "stand" or "stretch out arms"). Focus on the subject's pose regardless of their size and position.
2. Human Action: Examine the action or movement of the human subject as described in the text prompt (e.g., "jogging," "climbing," or "walking"). Focus on the subject's action regardless of their size and position.
3. Surroundings: Evaluate whether the environment and background elements in the image are consistent with the text prompt. The surroundings should match the described context, including location, props, and overall atmosphere.
4. Composition: Assess how naturally the arrangement of the human subject in the generated image aligns with the description, considering variations in the subject's placement, position, and size.
5. Image Quality: Evaluate whether the overall image exhibits realistic fidelity, clarity, and visual appeal, avoiding an overly synthetic or artificial look.

Scoring Range
Based on these criteria, a specific integer score from 0 to 9 can be assigned to determine the level of semantic consistency:
- Very Poor (0): No correlation. The image does not reflect any of the key points or details of the text.
- Poor (1-2): Weak correlation. The image addresses the text in a very general sense but misses most details and nuances.
- Fair (3-4): Moderate correlation. The image represents the text to an extent but lacks several important details or contains some inaccuracies.
- Good (5-6): Strong correlation. The image accurately depicts most of the information from the text with only minor omissions or inaccuracies.
- Very Good (7-8): Very strong correlation. The image captures nearly all relevant details from the text, with very few omissions or inaccuracies.
- Excellent (9): Near-perfect correlation. The image captures the text's content with high precision and detail, leaving out no significant information.

Input format
Every time you will receive a text prompt and an image.

Output Format
Score: [Your Score]

You must adhere to the specified output format, which means that only the scores need to be output, excluding your analysis process.

Figure A.14. **User Prompts for Evaluating Text Alignment:** To assess text alignment, we provide evaluation instructions to GPT [32] as user prompts.

Yes, I understand the task. It involves evaluating the semantic consistency between an image and its accompanying text prompt based on five key criteria: human pose, human action, surroundings, composition and image quality. The goal is to assess how well the visual content of the image aligns with the textual description, including both direct and subtle connections. The evaluation will result in a score ranging from 0 to 9, where 0 indicates no correlation and 9 indicates near-perfect correlation.

To evaluate the semantic consistency, I plan to follow these steps:

1. **Human Pose**: Assess whether the body pose of the human subject aligns with the pose described in the text (e.g., "stand" or "stretch out arms"). Focus on the subject's pose regardless of their size and position.
2. **Human Action**: Examine the action or movement of the human subject as described in the text prompt (e.g., "jogging," "climbing," or "walking"). Focus on the subject's action regardless of their size and position.
3. **Surroundings**: Evaluate whether the environment and background elements in the image are consistent with the text prompt. The surroundings should match the described context, including location, props, and overall atmosphere.
4. **Composition**: Assess how naturally the arrangement of the human subject in the generated image aligns with the description, considering variations in the subject's placement, position, and size.
5. **Image Quality**: Evaluate whether the overall image exhibits realistic fidelity, clarity, and visual appeal, avoiding an overly synthetic or artificial look.

After analyzing these aspects, I will assign a score that reflects the overall semantic consistency between the image and text. The score will reflect how well the image matches the text prompt, strictly adhering to the evaluation criteria provided.

My output format should be Score: [0-9], and I don't need to write out the specific analysis process.

Please provide me with the samples I need to evaluate.

Figure A.15. **Assistant Prompts for Evaluating Text Alignment:** To assess text alignment, we provide summary and planning responses to GPT [32] as assistant prompts.

Clothing Types:
- **Reference Image**: The subject is wearing a black crop top and black fitted pants.
- **Generated Image**: The subject is wearing a white short-sleeved top and a black skirt.
- **Assessment**: The clothing types differ significantly. The reference features a crop top and pants, while the generated image has a different top style and a skirt instead of pants.

Design:
- **Reference Image**: The design of the crop top is simple, and the pants are plain.
- **Generated Image**: The top in the generated image has a different design with sleeves, and the skirt has buttons and a different cut.
- **Assessment**: The designs do not match at all, indicating a significant disparity in design elements.

Texture:
- **Reference Image**: The fabric appears smooth and form-fitting.
- **Generated Image**: The texture of the clothing looks different, particularly the skirt's fabric, which appears less form-fitting than the pants in the reference
- **Assessment**: The textures are not consistent between the two images.

Color:
- **Reference Image**: Predominantly black.
- **Generated Image**: The top is white, and the skirt is black.
- **Assessment**: Although the skirt is black, the addition of a white top introduces a significant color difference when compared to the all-black outfit in the reference image.

Face Identity:
- **Reference Image**: The face has a specific structure and features.
- **Generated Image**: The face has similarities but is not identical due to differences in expression and possibly features.
- **Assessment**: While there might be some resemblance, the identity is not strongly preserved.

Overall Score

Given the evaluation, there are notable discrepancies across all five visual features. The clothing types, design, and textures differ significantly, while the color scheme is partially aligned. The facial identity shows some resemblance but is not sufficient for a strong correlation.

Taking all of this into account, I would assign the following score:

Score: 2

Figure A.16. **Analysis Process of GPT for Evaluating Identity Preservation in StoryMaker [44] Sample (Figure A.2):** GPT [32] provides a detailed analysis procedure to evaluate identity preservation based on the given scoring criteria.

The image captures a person walking ahead, which aligns well with the action described in the prompt. The subject is shown carrying shopping bags, consistent with the description of a busy shopping scenario. The surroundings depict a busy Japanese street, enhancing the contextual backdrop. The composition effectively frames the individual in a natural walking pose, contributing to a strong visual alignment with the prompt. Overall, the image quality is good, exhibiting clarity and detail, though there's a slight motion blur that suggests a busy atmosphere.

Score: 8

Figure A.17. **Analysis Process of GPT for Evaluating Text Alignment in StoryMaker [44] Sample (Figure A.2):** GPT [32] provides a detailed analysis procedure to evaluate text alignment based on the given scoring criteria.

Clothing Types:The clothing types in both images are consistent. The subject in both images is wearing a crop top and fitted pants, which are the same style.

Design: The design of the clothing matches as both images feature a similar style of crop top and pants, although the top in the second image has lace details that are not explicitly visible in the first image due to lighting and angles.

Texture: The texture of the fabrics appears to be similar, with both outfits looking sleek and form-fitting. The material seems to be comparable in both images.

Color: The color of the clothing is consistent as both subjects are dressed in black. The saturation and brightness are also similar, maintaining a cohesive color scheme.

Face Identity: Although expressions and poses vary slightly, the facial characteristics are recognizable, maintaining a strong resemblance to the reference image.

Overall, the generated image exhibits a very close resemblance to the reference image, with only minor differences in detail and clarity.

Score: 8

Figure A.18. **Analysis Process of GPT for Evaluating Identity Preservation in Visual Persona Sample (Figure A.2):** GPT [32] provides a detailed analysis procedure to evaluate identity preservation based on the given scoring criteria.

The image shows a person walking on a busy street, consistent with the prompt. The individual is carrying multiple shopping bags, which aligns well with the described action. The setting appears to be in Japan, indicated by signage and overall urban feel. The composition effectively places the subject in a manner that draws attention without obstructions. The image quality is high, with good clarity and visual appeal. Overall, the image captures the essence of the text prompt with very few omissions or inaccuracies.

Score: 8

Figure A.19. **Analysis Process of GPT for Evaluating Text Alignment in Visual Persona Sample (Figure A.2):** GPT [32] provides a detailed analysis procedure to evaluate text alignment based on the given scoring criteria.

You will be provided with an input image, a text prompt, and a generated image based on the input image and text prompt.
Your task is to evaluate the generated image based on three metrics:

1. Text Alignment between the text prompt and the generated image.
2. Identity Preservation between the input image and the generated image.
3. Perceptual Quality of the generated image.

Please follow the detailed guidelines below for each metric.

1. Text Alignment

Task Definition:
Your task is to evaluate how well the generated image corresponds to the details specified in the text prompt.

Scoring Criteria:
Your score should reflect how well the generated image aligns with all the elements of the text prompt, including facial expression, pose, action, and surrounding environment.

Example:
Consider the following text prompt:
"A photo of an angry person, leaning forward, riding a bicycle on a cobblestone street, surrounded by old buildings."

• **"angry"** indicates the **facial expression.**
• **"leaning forward"** specifies the person's **pose.**
• **"riding a bicycle"** denotes the person's action.
• **"on a cobblestone street, surrounded by old buildings"** describes the **surrounding environment.**

Scoring Range:
• **0 points:** The generated image does not follow the text prompt at all.
• **0.5 points:** The generated image partially follows the text prompt.
• **1 point:** The generated image fully follows the text prompt.

2. Identity Preservation

Task Definition:
Your task is to evaluate how well the visual appearance of the person in the generated image retains the visual appearance of the person in the input image.

Scoring Criteria:
Your score should reflect how well the following aspects of the input image are preserved in the generated image:

• **Clothing Types:** Refers to the type of clothing, such as short sleeves, long pants, or rounded necklaces.
• **Clothing Design:** Includes the pattern of the clothing (e.g., floral, striped, or solid) and decorative elements (e.g., logos, zippers, or pockets).
• **Clothing Texture:** Indicates the texture of the fabrics worn by the person.
• **Clothing Color:** Represents the primary colors of the person's clothing and body, including hue, saturation, brightness, and overall color distribution.
• **Face Identity:** Refers to the unique characteristics of the person, such as hairstyle, race, age, and other distinctive features.

Scoring Range:
• **0 points:** The person's appearance is not preserved at all.
• **0.5 points:** The person's appearance is partially preserved.
• **1 point:** The person's appearance is mostly preserved.

3. Perceptual Quality

Task Definition:
Your task is to evaluate how natural and realistic the generated image appears.

Scoring Criteria and Range:
• **0 points:** The generated image appears highly unnatural due to obvious artifacts or distortions.
• **0.5 points:** The generated image is slightly unnatural with minor artifacts.
• **1 point:** The generated image looks entirely genuine and realistic.

Figure A.20. **Evaluation Guidelines for Human Evaluation.** We provide detailed evaluation guidelines to each human rater.

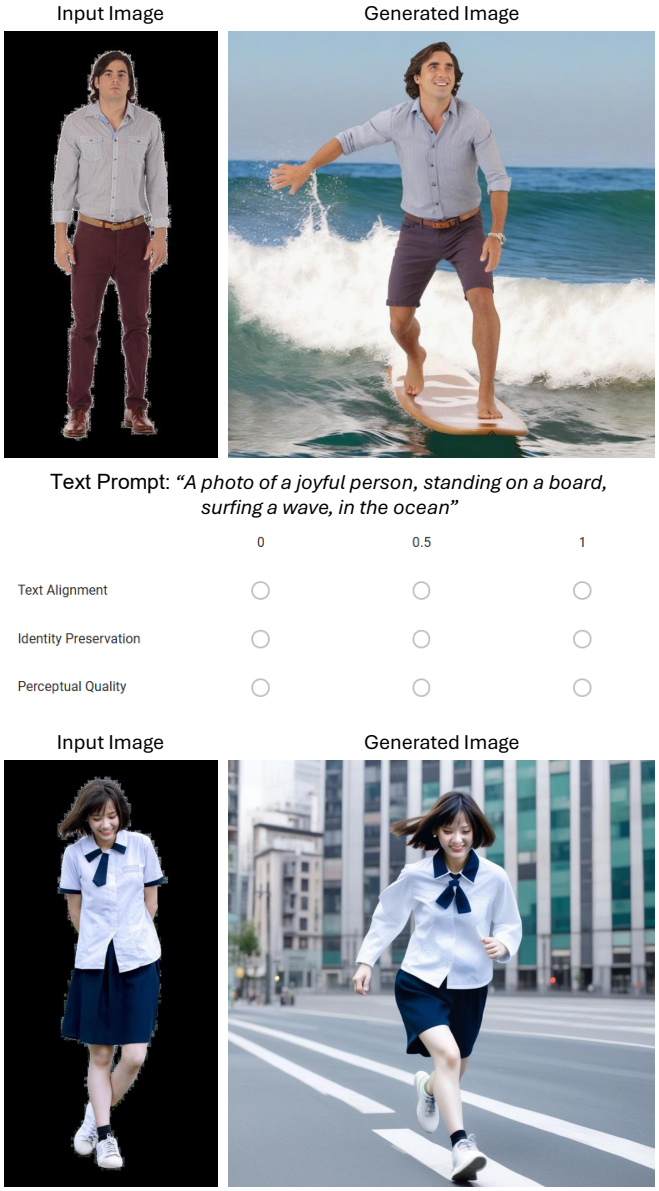


Figure A.21. **Examples of Human Evaluation Questions.**

Figure A.22. **Qualitative Results of Visual Persona on SSHQ [10] and PPR10K [27]:** The first row includes input human images from SSHQ and PPR10K. The second to last rows include the generated images by Visual Persona based on the input images and the given prompts. Visual Persona generates full-body consistent, customized images of the input human, while closely aligning with the diverse text prompts.

Figure A.23. **Qualitative Results of Visual Persona on SSHQ [10] and PPR10K [27]:** The first row includes input human images from SSHQ and PPR10K. The second to last rows include the generated images by Visual Persona based on the input images and the given prompts. Visual Persona generates full-body consistent, customized images of the input human, while closely aligning with the diverse text prompts.

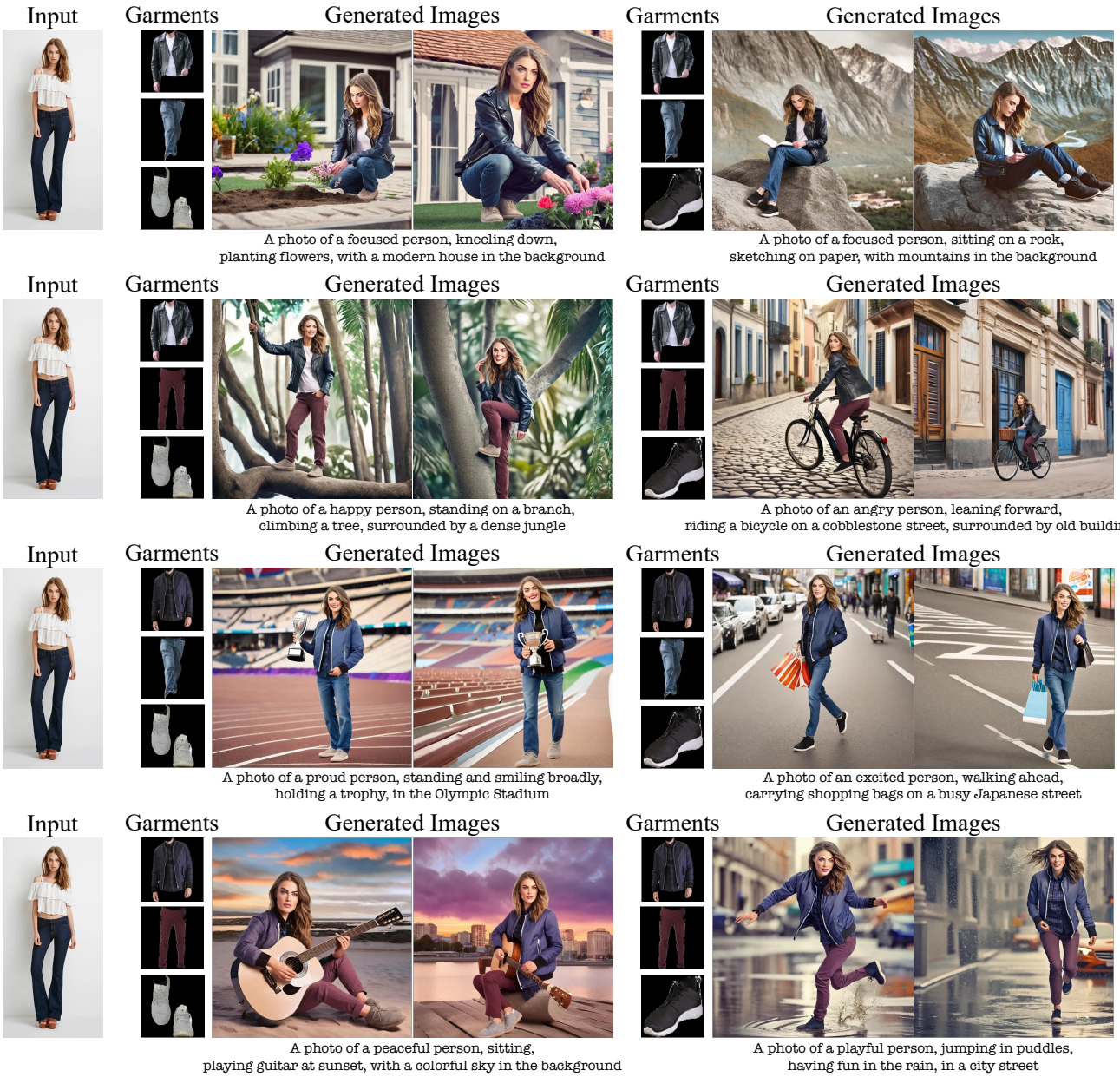
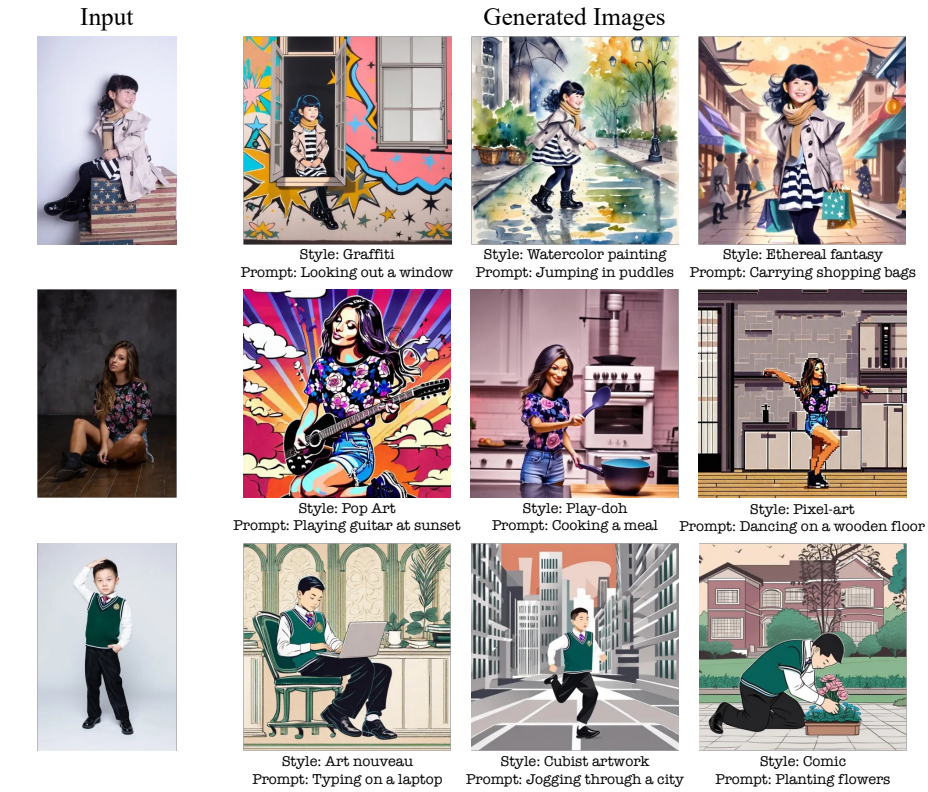
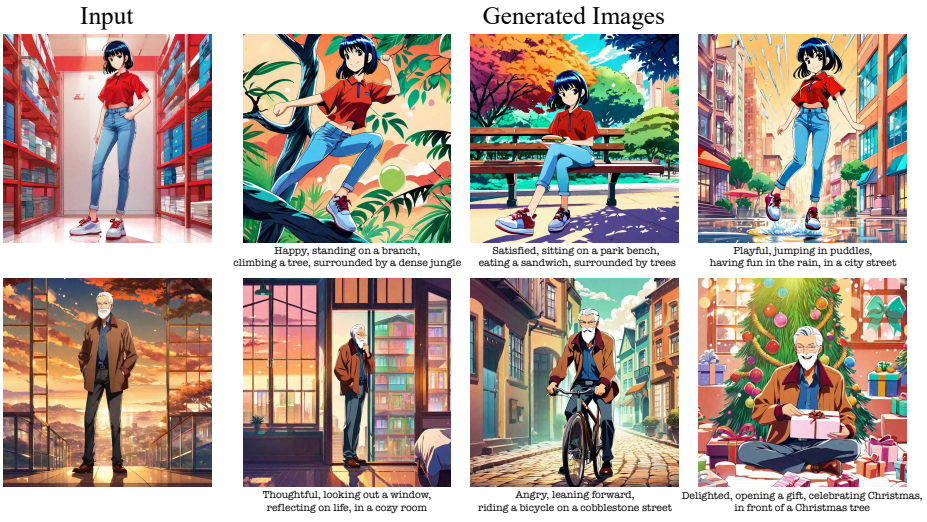


Figure A.24. **Qualitative Results of Visual Persona for Text-Guided Virtual Try-On (VTON):** Although Visual Persona is not specifically designed for VTON, our method naturally supports text-guided VTON, whereas existing VTON models [5, 13, 17, 45] are limited to minor scene and pose changes due to the absence of text control.



(a) Human Stylization



(b) Character Customization

Figure A.25. **Qualitative Results of Visual Persona for Human Stylization and Character Customization:** (a) Visual Persona can adapt to various stylization prompts while preserving the input's full-body identity. (b) Although Visual Persona is not trained for the character domain, our method can generalize to the character domain.

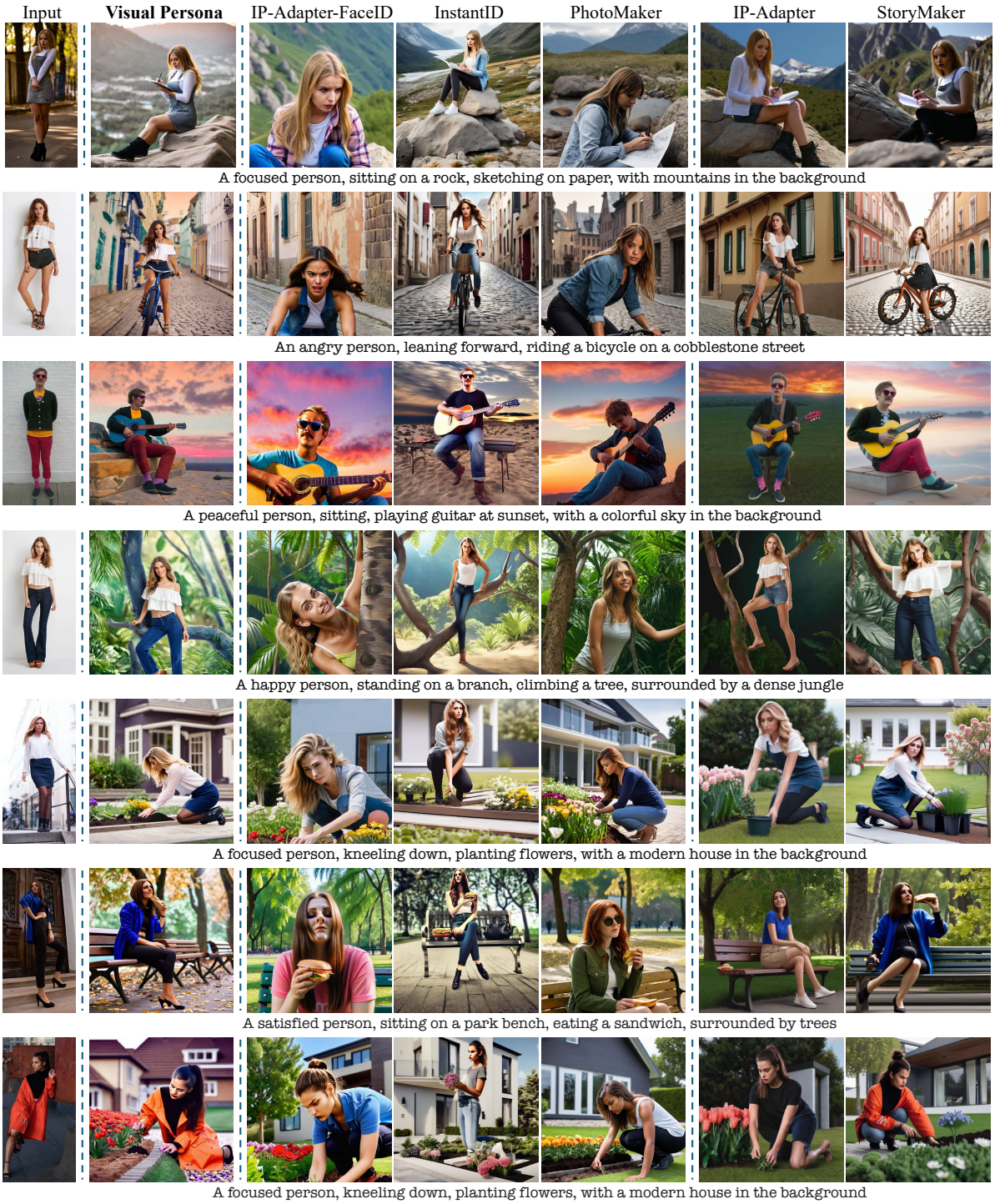


Figure A.26. **Qualitative Comparison on SSHQ [10] and PPR10K [27]:** We compare Visual Persona with IP-Adapter-FaceID [42], InstantID [41], PhotoMaker [26], IP-Adapter [42], and StoryMaker [44].

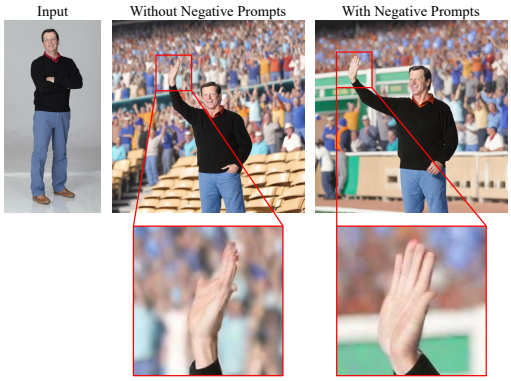


Figure A.27. **Limitation: Inaccurate Body Proportions.** The inherent challenge of generating bad body proportions in pre-trained T2I diffusion models can be alleviated through negative prompting.

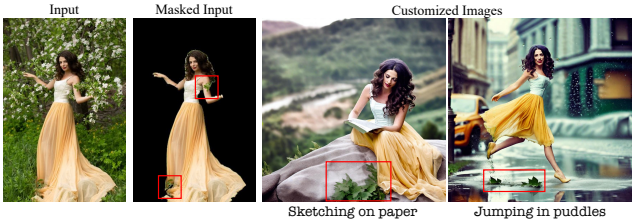


Figure A.28. **Limitation: Identity-Unrelated Attribute Leakage from Input.** When the input human is occluded by identity-unrelated elements, these elements are often included in the customized images.

G. Limitation

Inaccurate Body Proportions. SDXL [35] inherently struggles to generate human images with accurate body proportions, often resulting in artifacts such as fused fingers or extra arms and legs. Figure A.27 provides an example of fused fingers. Since we leverage the pre-trained SDXL to maximize its generative capabilities, our model also inherits these issues. To alleviate this, we incorporate a negative prompt, including terms such as *"disfigured, deformed, three arms, three legs, fused fingers, cloned face, bad proportions, bad anatomy."* This negative prompt guides the pre-trained T2I diffusion model away from generating such undesired features through classifier-free guidance [14]. Figure A.27 shows that this negative prompt effectively enables the model to generate more natural and anatomically accurate human body proportions, without the need for additional training.

Identity-Unrelated Attribute Leakage from Input. Figure A.28 shows that when the input human is occluded by identity-unrelated elements (e.g., background leaves or grass), the customized image includes these elements instead of filtering them out.In future work, we plan to address this by refining the foreground mask using body parsing models [21, 25], which separate each part of the human body individually and more accurately isolate only the foreground, in contrast to the human matting method [15], which directly detects the whole human body.

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. A.3

[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. A.2, A.7

[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. A.6

[4] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22560–22570, 2023. A.4

[5] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for virtual try-on. *arXiv preprint arXiv:2403.05139*, 2024. A.16

[6] 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. A.2

[7] DeepAI. Text-to-image api. `https://deepai.org/machine-learning-model/text2img`, 2025. Accessed: 2025-03-13. A.1

[8] 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. A.2, A.3, A.4

[9] 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. A.1

[10] 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. A.2, A.3, A.4, A.6, A.14, A.15, A.18

[11] Andrew F Hayes and Klaus Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89, 2007. A.3

[12] 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. A.1

[13] 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. A.16

[14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. A.19

[15] 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. A.2, A.19

[16] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. A.2

[17] 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. A.16

[18] 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. A.2

[19] 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. A.2

[20] 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. A.6

[21] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. *arXiv preprint arXiv:2408.12569*, 2024. A.2, A.19

[22] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. A.1

[23] 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. A.3

[24] 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. A.2

[25] 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. A.2, A.19

[26] 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. A.1, A.6, A.18

[27] 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. A.2, A.3, A.4, A.6, A.14, A.15, A.18

[28] 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. A.2

[29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. A.1

[30] Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. Videodrafter: Content-consistent multi-scene video generation with llm. *arXiv preprint arXiv:2401.01256*, 2024. A.6

[31] 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. A.4

[32] OpenAI. GPT-4o: Advancing Cost-Efficient Intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. Accessed: 2024-11-05. A.1, A.3, A.8, A.9, A.10, A.11

[33] 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. A.2, A.4

[34] 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. A.1, A.2

[35] 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. A.1, A.3, A.4, A.19

[36] 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. A.1, A.2, A.4

[37] 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. A.2, A.7

[38] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo

Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. A.1

[39] 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. A.2

[40] 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. A.6

[41] 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. A.1, A.6, A.18

[42] 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. A.1, A.2, A.3, A.4, A.6, A.18

[43] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. A.1, A.6

[44] 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. A.1, A.2, A.3, A.4, A.5, A.6, A.10, A.18

[45] 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. A.5, A.6, A.16