Generative Stylist LUCI: Latent model for User Clothing Inference application

Miria K. Feng

Abstract—We present LUCI, a personalized fashion try-on system that generates images and short videos of individual users wearing immediately purchasable items. We evaluate two categories approaches to the image generation task and analyze their tradeoffs. **(1)** DreamBooth fine-tuning with Stable Diffusion XL embeds each user into the model's output space via a unique token identifier trained on 4 reference photos. This approach achieves strong adherence to text-conditioned input but requires approximately 15 minutes of per-user training. **(2)** Zero-shot virtual try-on via IDM-VTON inpainting conditioned on DensePose body maps and IP-Adapter eliminates per-user training entirely and generates results in seconds, but exhibits lower fidelity to the input text conditioning prompt. Both methods operate on a custom dataset of 9,264 images across 1,878 garments from 186 designers, scraped from e-commerce site Luisaviaroma, ensuring all generated outfits are directly purchasable. Our custom developed pipeline continuously populates the dataset per two hours to ensure immediate relevance. Given the user's selected try-on image, we further generate a 1–2 second spin video and evaluate two motion guidance strategies: a 2D skeleton approach using OpenPose transferred from e-commerce model gifs, and a 3D parametric approach using SMPL-guided animation with AnimateDiff, where the body mesh is parameterized by shape coefficients estimated from the user's photos. The SMPL approach produces more physically accurate clothing animation by capturing depth, surface normals, and body proportions that the 2D skeleton discards, though at higher TFLOPs cost. We evaluate all methods with 4 individual users.

Index Terms—EE367 Winter 2026. Please do not open source code, as we are presenting this work in a commercial deployment setting.

1 Introduction

Online shopping is an integral part of our daily lives. Instead of purchasing your next outfit online and potentially mailing a return after trying it on, what if you could see yourself wearing it first? The current online shopping experience presents an expensive and wasteful cycle of purchasing and returning items that fail to meet expectations in reality. The central problem is a lack of visual transparency: consumers must guess how a garment will drape, fit, and look on their specific measurements based on photos of industry standard fashion models.

Solving this visualization gap is highly interesting and practical for both consumers and the environment. The current purchase-mail-return cycle not only wastes the user's time and energy but is also environmentally unsustainable long term. While recent advancements in deep generative models have revolutionized human-AI interactions across natural language processing and computer vision, their application in e-commerce remains limited due to an incomplete understanding of core consumer needs. A personalized virtual try-on system can transform the frustrating search through thousands of filtered items into a creatively fulfilling, efficient, and transparent styling experience that is easily shared between peers for feedback before purchase.

Developing a realistic and actionable virtual try-on system is uniquely difficult. While conditional generation have been proven to guide deep generative models using text prompts, naive approaches fail for practical e-commerce applications. For example, prompting a base model like Stable Diffusion XL [1] or Google Images to generate "me in a blue Dior dress" may yield a high fidelity image, the garment will likely not exist for purchase online today. Generating outfits that are *immediately purchasable* requires strict fidelity to a real-world inventory, which standard generative models cannot guarantee off-the-shelf.

Previous attempts to constrain generative models to specific datasets often rely on retraining the entire diffusion model, which is computationally prohibitive given the fast-paced, hourly updates of the fashion industry. In this work, we theorize that all articles of clothing available for immediate purchase today already exist within the convex hull of the base diffusion model's learned output space (e.g., a modern cocktail dress shares structural elements like sleeves and necklines with historical gowns). Therefore rather than retraining the base model on a rapidly changing dataset, we simply guide it to sample relevant existing features. Additionally, while existing video generation models suffer from severe GPU VRAM bottlenecks, we bypass this limitation by conditioning our image-to-video generation on a predefined, efficient spin motion that is sufficiently helpful to communicate real-world fashion expectations without being prohibitively compute heavy to generate.

In this work, we introduce generative stylist LUCI: **L**atent model for **U**ser **C**lothing **I**nference application. Given a text prompt and four casual images of a user, we finetune Stable Diffusion XL via DreamBooth to embed the individual user as a unique token in the output space. We construct and dynamically update a custom dataset scraped from the e-commerce retailer Luisaviaroma to ensure all generated fashion items are currently purchasable. Ultimately, the system efficiently synthesizes four high-fidelity composite images of the user wearing the requested outfit, along with a 2-second video of the user performing a spin.

• *Department of Electrical Engineering, Stanford University, CA, 94040. E-mail: miria00@stanford.edu*

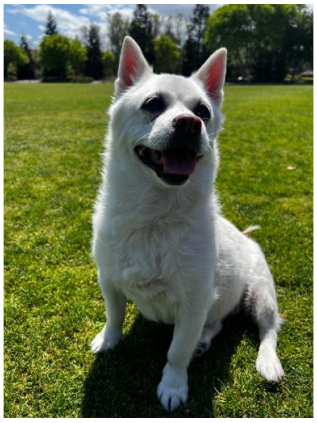


Fig. 1. Instance photo at input

Fig. 2. Refined photo after 5 steps.

Fig. 3. Output after User1 and User2 trained on same run.

Summary of Contributions

The following work is presented as follows:

- **Rapid personalized generation:** A pipeline for embedding users into the latent space to efficiently sample user-specific images conditioned on a custom, easily and automatically updated retail dataset.
- **Efficient image-to-video synthesis:** A low-VRAM method for generating a 2-second video of the individual user performing a pre-defined spin motion in the generated outfit via SMPL and OpenPose.
- **High-fidelity evaluation:** Experimental results and user evaluations demonstrating that our generated outputs align with user expectations and provide a transparent, motivating online shopping experience.

2 Related Work

Conditional Image Generation: Realistic image generation from text conditional input has emerged across a wide range of domains. The impressive artistic style transfer of models such as DALL-E2 [2], Imagen [3], and Midjourney [4] demonstrate the feasibility of generative models to replicate sections of human creativity. However the generative domain has seen very little practical application in the field of fashion, although it is a common factor in the lives of most global users today. Through artistic rendering or in-painting it is possible for the user to generate fashion concepts, but this process is tedious and does not produce immediately purchasable items. The first instinct of simply re-training a generative model on a current fashion dataset is unrealistic, since the compute and GPU power necessary to remotely approach a model such as Stable Diffusion is an expensive resource, and the technical skills involved are widely adopted by the majority on global shoppers. Furthermore, the fast paced changing landscape of fashion means what is favorable this season, may not be available in a few months. Therefore in the application of fashion and e-commerce we desire a more practical rapidly adaptive generative model.

Subject-Driven Generation: Models such as DreamBooth [5] introduce the possibility for personalized subject-driven generation. These architectures synthesize novel renditions of the same subjects in different contexts. This overcomes certain weaknesses of previous image generation models, and does not need to rely solely on detailed textual description of an object or individual to recreate the output with recognizable distinctive identity. The work of [6] approaches this by representing visual concepts, like an object or a style, through new tokens in the embedding space of a frozen text-to-image model, resulting in small personalized token embeddings. Although effective, this method is limited by the expressiveness of the frozen diffusion model. ControlNet [1] proposes an alternative approach to adding conditional control to pre-trained text-to-image diffusion models. In order to overcome the high computation costs of training a large generative model, ControlNet locks the production-ready weights of the pre-trained diffusion model and reuses their deep encoding layers as a backbone for conditional signal. However although ControlNet has demonstrated its effectiveness at applying depthmaps, the complexity of its mechanics make it extremely sensitive to hyperparameter tuning during design. DreamBooth achieves fine-tuning with 3–5 images of the subject, to implant the individual into the output domain of the model such that it can be synthesized with a unique identifier. This technique represents a given subject with rare token identifiers and fine-tunes a pre-trained, diffusion-based text-to-image model. In this work we build upon the effectiveness of DreamBooth in our generative fashion stylist application with quick adaptive dataset and explore other options to speed up the generative process.

Video Generation: Video generation projects such as Stable Diffusion Video [1], Pika Labs [7], and Runway push the creative possibilities of diffusion models to assist in artistic content creation. Recent advancements have created impressively realistic videos with amazing graphics rendering. However these models require extensive GPU compute to train or even to perform inference. When attempting to perform inference on the recently released Stable Difusion Video on an A100, we quickly ran into VRAM issues. Produced videos are also pretrained, and therefore do not offer possibilities of personalized video generation. Our work aims to allow the user to generate themselves within a short 2 sec video subject to constraint that the motions are predefined to a simple spin. Our approach is that for the purposes of the generative stylist application, we aim to better allow

the user to view themselves wearing different purchasable fashions through motion. Therefore we do not require the wide domain of numerous video generation outputs, but only require the specific domain of certain short motions.

Zero-Shot Identity Preservation: A fundamental limitation of DreamBooth is the requirement to fine-tune the entire diffusion model for each new user, which in our setting consumed approximately 20 minutes per individual on an RTX4090 (and A100s). This renders the approach somewhat impractical for an e-commerce application where users expect near-instant results. InstantID [8] addresses this by introducing a tuning-free method that preserves facial identity from a single reference image. The architecture consists of three components: an ID embedding extracted via InsightFace's ArcFace recognition model, a lightweight IP-Adapter module with decoupled cross-attention that projects the face embedding into the diffusion model's conditioning space, and an IdentityNet built on ControlNet that encodes facial landmarks for spatial guidance. InstantID achieves competitive fidelity to per-user LoRA models without any training, and is compatible with both SD1.5 and SDXL. The concurrent work PuLID [9] further improves fidelity through contrastive learning, achieving stronger preservation of subtle facial features at the cost of additional compute. In this work we adopt InstantID as the default identity encoder due to its favorable speed-fidelity tradeoff, and note that PuLID may be substituted on higher-VRAM hardware when maximum accuracy is required. However, we identify a critical limitation of both methods: they encode only the face, and hallucinate an arbitrary body shape during generation.

Visual Prompt Conditioning: IP-Adapter [10] introduces a decoupled cross-attention mechanism that enables image prompts to condition pre-trained text-to-image diffusion models without fine-tuning. By training a lightweight adapter that projects CLIP image embeddings into the cross-attention layers alongside text embeddings, IP-Adapter allows a reference image to guide the style and content of generation. The variant IP-Adapter-FaceID [11] replaces CLIP embeddings with face recognition embeddings for identity tasks, while IP-Adapter-Plus uses a finer-grained image encoder for improved detail transfer. In our application we leverage IP-Adapter in two distinct roles: first, to inject garment reference images from our Luisaviaroma dataset as visual style prompts, ensuring the generated clothing matches currently purchasable items without retraining the CLIP model; and second, in conjunction with InstantID's identity adapter for facial conditioning. This replaces our earlier approach of training a custom CLIP model on the fashion dataset, which required retraining whenever the product catalog updated. With IP-Adapter, catalog updates require only swapping the reference images at inference time.

Virtual Try-On Models: The virtual try-on (VTON) literature provides the machinery to address the whole-body preservation problem that pure face embedding methods cannot solve. IDM-VTON [12] proposes a dual-UNet architecture built on SDXL, where a frozen GarmentNet encodes garment features while the main UNet is trained end-to-end for the try-on task. The model conditions on an agnostic representation of the person, their image with current clothing masked out, preserving face, hands, and body outline, along with DensePose [13] maps that provide dense correspondences between image pixels and a 3D body surface model. This inpainting-based formulation is critical: because the model generates pixels only within the masked clothing region, the user's body silhouette and proportions are physically preserved in the output. A plus-size user and a petite user will receive entirely different renderings of the same garment, faithful to their actual body shapes. CatVTON [14] demonstrates that a simpler single-model approach using spatial concatenation can achieve competitive quality with lower computational requirements, and its recent integration with Flux.1 achieves state-of-the-art results on the VITON-HD benchmark. OOTDiffusion [15] takes a parallel outfitting fusion approach with separate denoising processes for garment and person. In our work we adopt IDM-VTON as the primary try-on engine for its garment detail preservation, with SDXL inpainting combined with IP-Adapter garment conditioning as a fallback when the specialized VTON weights are unavailable.

3 Methods, Theory, and Approach

Given only 4 casually captured images of a specific user and a text input prompt, our objective is to generate new images of the subject wearing a composite outfit that matches the description in the text conditional input. The user may then select their favorite output image, which we then generate 2 sec video spin motion. Our aim is high detail fidelity that remains faithful to the proportions and features of each individual user, but with accurate variations in pose, outfits, and other semantic modifications. We leverage the strong semantic priors learned by the Stable Diffusion XL model, and utilize the same technique as DreamBooth for implanting an individual user into the output space. In order to generate feasible videos we aim to train a Stable Diffusion v1.4 model on a fixed sequence of input poses, then transfer the subject onto the same sequence. All experiments presented in this work were run on RTX4090.

3.1 Personalized Text-2-Image Diffusion Models

DreamBooth.: Diffusion models present the generator function as: $\hat{x}_\theta = \hat{x}_\theta(\epsilon, c)$ where $\epsilon \sim N(0, I)$ is an initial noise map and $c = \Gamma(P)$ is a conditioning vector generated using a text encoder Γ and a text prompt P. The generator function $\hat{x}_\theta$ generates an image x_{gen}.

DreamBooth notes two issues with the generalization process: language drift which is observed when a model that is pre-trained on a large text corpus and later fine-tuned for a specific task loses syntactic and semantic knowledge of the language and reduced output diversity, which is a risk when fine-tuning on a small set of images. In order to mitigate these issues a class-specific prior preservation loss is utilized:

$$\mathbb{E}_{x,c,\epsilon,\epsilon_0,t}\Big[w_t\|\hat{x}_\theta(\alpha_t x + \sigma_t\epsilon, c) - x\|_2^2 + \lambda w_{t'}\|\hat{x}_\theta(\alpha_{t'}x_{pr} + \sigma_{t'}\epsilon_0, c_{pr}) - x_{pr}\|_2^2\Big], \quad (1)$$

The second term of this loss aims to keep the model with its own generated samples, in order for it to retain the prior once the few-shot fine-tuning begins. In our setting the few shot images are the 4 images uploaded by the user. We then

utilize this prior-preservation loss to embed the user into a unique token. We deviate from the original DreamBooth model in that we simply assign the User with a random long random alphanumeric token, as opposed to using DreamBooth's original implementation of crafting "rare-token" identifiers. This is since in our online e-commerce setting, we desire as simple as possible for utility purposes within a vast number of users.

3.2 Conditional Generation

We utilize 2 versions of Stable Diffusion: the XL version for image generation due to its larger number of parameters and more expressiveness, and the v1.4 version for comparative evaluation. The conditional aspect of the guided text prompt generation is implemented by adding the CLIP model with cross-attention in the UNET architecture. Figure 4 shows the overview of the base model.

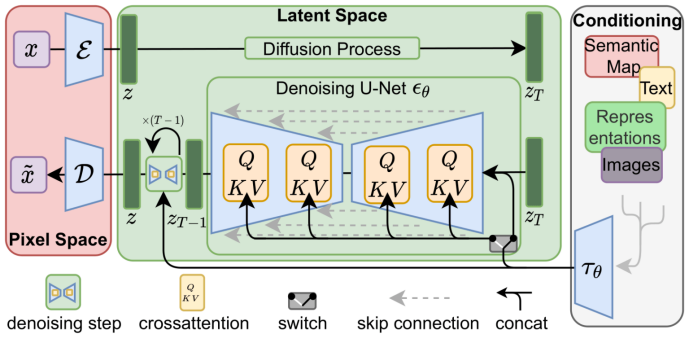


Fig. 4. Conditional Generation Architecture

In order to adapt the existing base Stable Diffusion models to our new fashion dataset, we note that the output of articles of clothing should already exist in the convex hull of the Stable Diffusion output space. Therefore we only need to modify the CLIP model to accept the custom dataset, and guide the generation in addition to the user's input prompt. Therefore we train our custom CLIP model with OpenClip, and replace this section of the architecture.

3.3 Efficient Image-2-Video Generation

In order to efficiently generate personalized videos on the specific user, we hard tune the simpler Stable Diffusion v1.4 model to a sequence of poses we create based on e-commerce site Net-A-Porter's action gifs for garments. We further distill these poses down into OpenPose type gifs, to cleanly transfer onto the individual subject once they have been embedded into the output space. The overall video training procedure is shown in Figure 5.

Our objective is to generate videos of the subject wearing immediately purchasable items that match the text description, faithful to the user's actual face and body proportions. Our central principle is that a plus-size user and a petite-size user must receive entirely different renderings of the same dress, each physically accurate to how the garment might look on their specific body.

We decompose identity preservation into three conditioning streams that operate simultaneously during generation:

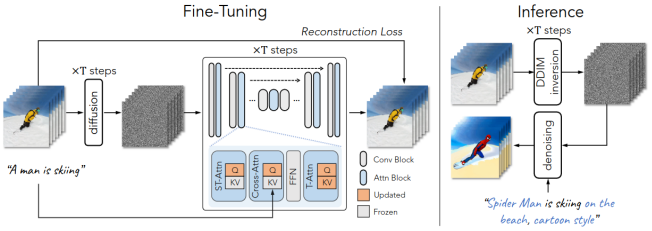


Fig. 5. Low VRAM Video Genertaion

1) **Face identity** A 512-dimensional ArcFace embedding injected via InstantID, preserving facial features without per-user fine-tuning.
2) **Body shape** SMPL β parameters estimated from the user's photos, combined with DensePose surface maps and an agnostic clothing mask, which lock the user's physical proportions during generation.
3) **Garment details** Reference images from our Luisaviaroma dataset injected via IP-Adapter, ensuring generated clothing matches currently purchasable items.

These three elements converge in a body-aware inpainting model that replaces only the clothing region of the user's photo, preserving their body outline. The selected output is then animated into a spin video using SMPL-guided diffusion (Sec. 6.3). The overall architecture is shown in Figure 6.

3.4 Zero-Shot Face Identity Preservation

The first approach used DreamBooth to fine-tune the full SDXL UNet on 4 user images, requiring approximately 20-30 minutes of training per user on the RTX4090. This is impractical for an e-commerce setting where users expect near-instant results. We evaluate this against InstantID [8], a tuning-free method that preserves facial identity from reference images in a single forward pass.

Face Embedding Extraction.: We use InsightFace's ArcFace model (antelopev2) to extract a 512-dimensional identity embedding from each of the user's reference photos. For multiple images, we compute the mean of the L_2-normalized embeddings:

$$\mathbf{e}_{\text{face}} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{e}_i}{||\mathbf{e}_i||_2}, \quad \mathbf{e}_i = \text{ArcFace}(I_i), \tag{2}$$

where I_i is the i-th user photo and $N = 4$ in our setting. This averaging provides robustness against pose and lighting variation across the reference set.

IdentityNet Conditioning.: The face embedding $\mathbf{e}_{\text{face}}$ is injected into SDXL through two pathways. First, a lightweight IP-Adapter projects the embedding into the cross-attention layers of the UNet, operating alongside text conditioning. Then the IdentityNet (a ControlNet trained specifically for facial structure) conditions on the 5-point facial landmarks detected from the highest-confidence reference image, providing guidance for face placement and proportions. Together these achieve strong facial fidelity without modifying any weights of the base diffusion model.

3.5 Body Shape Estimation and Conditioning

Face embeddings alone are insufficient for fashion try-on: they encode no information about the user's body. Without explicit body conditioning, the diffusion model hallucinates an arbitrary physique from text priors. We address this with three complementary signals that lock the user's body shape during generation.

SMPL Shape Regression.: We regress the user's body shape from their full-body reference photos using HMR2.0 [16], which estimates the SMPL [17] parameters from a single image via a ViT-based architecture. The SMPL model parameterizes body shape through a 10-dimensional vector $\boldsymbol{\beta}$ that controls proportions including overall body volume (β_0), height (β_1), and shoulder-to-hip ratio (β_2). We run HMR2.0 on each user photo that contains a full-body view and compute a confidence-weighted average:

$$\boldsymbol{\beta}_{\text{user}} = \frac{\sum_i w_i \boldsymbol{\beta}_i}{\sum_i w_i}, \quad w_i = \text{conf}(\text{HMR2}(I_i)), \tag{3}$$

where $\text{conf}(\cdot)$ is the detection confidence. These shape parameters serve two roles: conditioning the image generation pipeline on the user's proportions, and rendering a personalized 3D body for the video spin sequence.

DensePose Extraction.: DensePose [13] maps each pixel of the user's body to a UV coordinate on a 3D surface model, producing an IUV map with three channels: body part index $I \in \{1, \ldots, 24\}$, and surface coordinates $(U, V) \in [0, 1]^2$ within that part. This provides the generative model with dense knowledge of where on the body each pixel belongs. This is critical for ensuring a garment wraps correctly around the user's specific hips, waist, and shoulders rather than a generic body of arbitray type.

Agnostic Mask Generation.: We use a human parsing model (SCHP [18]) to produce per-pixel semantic labels for the user's image, identifying regions including face, hair, upper clothing, lower clothing, dress, arms, and legs. From these labels, we construct an agnostic mask M that removes the user's current clothing while preserving identity-critical regions:

$$M(x, y) = \begin{cases} 1 & \text{if parse}(x, y) \in \mathcal{C}_{\text{garment}} \\ 0 & \text{otherwise} \end{cases}, \tag{4}$$

where $\mathcal{C}_{\text{garment}}$ is the set of clothing labels relevant to the target garment type (e.g., {upper_clothes, belt} for tops, or {upper_clothes, skirt, pants, dress, belt} for dresses). The mask is dilated by a 5×5 kernel to avoid hard edges. The agnostic image $\tilde{I} = I \odot (1 - M) + 192 \cdot M$ replaces masked regions with a neutral gray fill, giving the inpainting model a clean canvas for the new garment while preserving the user's face, hands, hair, and body outline.

Since the inpainting model can only generate pixels inside the mask M, the body contour outside the mask is physically fixed. A wider body produces a wider unmasked silhouette, which constrains the generated garment to fill a correspondingly wider area.

3.6 Garment Conditioning via IP-Adapter

Earlier approaches retrained the CLIP model on our Luisaviaroma dataset to condition generation on current fashion items. This required retraining whenever the catalog updated. In order to speed up the updates per fashion items available for immediate purchase, we replace this with IP-Adapter [10], which injects garment reference images as visual prompts through decoupled cross-attention at inference time.

Given a garment image I_g from our dataset, IP-Adapter encodes it through a CLIP ViT-H image encoder and projects the resulting features into the UNet's cross-attention layers via a learned linear projection, operating in parallel with the text conditioning path. The influence of the garment reference is controlled by a scale parameter $s_g \in [0, 1]$:

$$\text{Attn}(Q, K, V) = \text{SFM}\left(\frac{QK_{\text{txt}}^T}{\sqrt{d}}\right) V_{\text{text}} + s_g \cdot \text{SFM}\left(\frac{QK_{\text{img}}^T}{\sqrt{d}}\right) V_{\text{img}}, \tag{5}$$

where $(K_{\text{text}}, V_{\text{text}})$ are projected from the text encoder and $(K_{\text{img}}, V_{\text{img}})$ are projected from the garment image encoder. We set $s_g = 0.6$ in our experiments. This allows catalog updates to be reflected immediately by swapping reference images, with no retraining of any model component.

4 ANALYSIS, EVALUATION, AND RESULTS

4.1 Dataset and Implementation

In order to fine-tune the large diffusion model towards current high resolution fashion items, we built a custom dataset from e-commerce site Luisaviaroma [19] and Net-A-Porter. This is particularly critical to the Generative Stylist LUCI application, since fashion items and styles update at an extremely fast pace. Therefore what is relevant or desired yesterday, may not be applicable to the users of today. We select Luisaviaroma due to its wide range of high resolution images and quick hourly updates, and select Net-A-Porter for referencing clear video spins on animated models. Therefore, we developed custom web scraping code to continuously download images, thus populating a current and relevant fashion dataset with high quality images. We use Selenium and BeautifulSoup libraries to navigate the website and extract information about each item, including the designer and description, thus creating image-text pair descriptors. A folder for each designer and description combination is made and continuously populated. Our code uses multiprocessing to speed up the scraping process, and stores a record of all the articles that have been scraped in a SQLite database to prevent duplicates. We then prepare the data to be consistent with DreamBooth by padding the images into 512x512 on white or neutral backgrounds. Each folder now contains 3-5 sample images of new fashion items, with the folder name acting as its text descriptor. The final result is a dataset of 18,948 images, of 4813 various garments, from 244 unique designers. For the video generation section of the project we create 2-sec videos by downloading model spin gifs from e-commerce site Net-A-Porter. We do so after observing that users are more likely to purchase items after seeing them move with motion on an individual. From the spin gifs when then distill each video down to OpenPose style gifs to train for each individual. Essentially this becomes a pose-transfer type of problem, but with a stable diffusion base model. We sample video at 48 frames, on the equivalent size of 512x512,

to generate 2 second spins faithful to each individual user. Examples of pose data is presented in the Appendix.

4.2 Evaluation

Since the output of the model and the input do not reference the same ground truth, it is difficult to apply metrics such as FID and IS. Therefore, instead we apply the same method of quantitative evaluation as ControlNet, and utilize 5 human users for feedback. Each human user was invited to render themselves into the output space, and experiment with text prompts, then asked a series of 5 questions upon completion. A Google Form Doc was shared, where each question was crafted based on the Average Human Ranking (AHR) preference metric, to be consistent with existing literature. This means a scale of 1-5, where lower is worse, and 5 is the perfect score. All users further noted that the addition of the spin video was substantially more motivating to proceed with purchase.

We note that the limiting factor is the time consumption needed to implant each user into the output space individually in the Dreambooth method. At 500 epochs per user on the RTX4090 this equated to 20min of training per user, which is too long to be realistically feasible for a e-commerce shopping experience. In order to improve this metric, we ported code into JAX and FLAX to enable faster backend computations. Although we noted a dramatic speedup, this consumed significantly more VRAM and required the more performant A100 GPU (RTX4090 being insufficient). Therefore for our users during evaluation, we implemented front end with Flask, tunneled with CloudFlare, and systematically experimented with lowering epochs to generate output per user in a 90sec time frame. This meant only one user can access the front end at each instance, while backend ran on the single RTX4090.

4.3 Limitations and Analysis

We separate the limitations and analysis of this work into two components: the hardware side and the software/numerical side.

On the hardware perspective, the limitations are firmly due to GPU access. The lack of VRAM and limiting GPU factor made it difficult to scale the model up to higher fidelity resolutions. For the time being with our current front end CloudFlare RTX4090 deployment approach, costs are essentially negligible. However, moving forward to be deployable at scale, parallel processing and multiple GPU instances will be needed.

From the software perspective, limitations come from needing to train the model each time to embed an individual new user. Essentially we have to run the inference step twice: once to generate 10 seeds, then use an image-to-image comparison system to choose the best seed based on the user input images, save this seed as an int to index and run inference to generate another 4 images faithful to the user's conditional prompt. Other issues arose when attempting to embed multiple users in the same model, often we note inference of one user token shows 2 or more users mistakenly in the same output. Samples of output follows (User1 and User2 images generated with same text prompt at input).

5 Future work and conclusions

Current generative diffusion models often require extensive time and GPU resources to fine-tune. This presents as an obstacle to generating personalized and immediately relevant output for users. The tradeoff is evident between high fidelity, speed, and compute cost. However our contribution of the custom dataset, low VRAM image-2-video spin generation, and custom user composite outfit generation shows promise in a more personalized realistic online shopping experience. Next steps aim to test the model at a larger scale, with a more optimized backend, that can enable multiple users to access front end at once (instead of currently being assigned time slots one user per hour). We also aim to experiment with better embedding options for individual users as rare tokens. Our theory is that by utilizing some loss function to push individual user embeddings further apart systematically (instead of random alphanumeric assignments) it should encourage better disentangling of users at inference. Therefore there should be less possibility of 2 users being generated on the same output image.

Acknowledgments.: This work is supported in part by the Stanford Graduate Fellowship, and we gratefully acknowledge the support of the EE367 teaching team! We additionally thank Lucy Woof for her valuable insights and feedback during this project. All generated graphics are presented with consent from individual beta testers, and the Appendices provide additional details of this work.

REFERENCES

[1] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 3836–3847.

[2] R. Gozalo-Brizuela and E. C. Garrido-Merchan, "Chatgpt is not all you need. a state of the art review of large generative ai models," *arXiv preprint arXiv:2301.04655*, 2023.

[3] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[4] A. M. Radhakrishnan, "Is midjourney-ai a new anti-hero of architectural imagery and creativity?" *GSJ*, vol. 11, no. 1, 2023.

[5] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. 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*, 2023, pp. 22 500–22 510.

[6] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," *arXiv preprint arXiv:2208.01618*, 2022.

[7] D. Guo, Y. Li, G. Huang, Y. Wang, K. Chen, Z. Liu, and K. Q. Weinberger, "Learning to learn image classifiers with few data," 2020.

[8] Q. Wang, X. Bai, H. Wang, Z. Qin, and A. Chen, "Instantid: Zero-shot identity-preserving generation in seconds," *arXiv preprint arXiv:2401.07519*, 2024.

[9] Z. Guo, Y. Wu, Z. Chen, H. Chen, and L. Zhang, "Pulid: Pure and lightning id customization via contrastive alignment," *arXiv preprint arXiv:2404.16022*, 2024.

[10] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," *arXiv preprint arXiv:2308.06721*, 2023.

[11] H. Ye, "Ip-adapter-faceid," https://huggingface.co/h94/IP-Adapter-FaceID, 2024.

[12] Y. Choi, S. Kwak, K. Lee, H. Choi, and J. Shin, "Improving diffusion models for authentic virtual try-on in the wild," 2024.

[13] R. A. Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7297–7306.

[14] Z. Chong, X. Xie, S. Nie, and M. Kankanhalli, "Catvton: Concatenation is all you need for virtual try-on with diffusion models," *arXiv preprint arXiv:2407.15886*, 2024.

[15] Y. Xu, T. Gu, W. Chen, and C. Chen, "Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on," 2025.

[16] S. Goel, G. Pavlakos, J. Rajasegaran, A. Kanazawa, and J. Malik, "Humans in 4D: Reconstructing and tracking humans with transformers," 2023.

[17] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Transactions on Graphics*, vol. 34, no. 6, pp. 248:1–248:16, 2015.

[18] P. Li, Y. Xu, Y. Wei, and Y. Yang, "Self-correction for human parsing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 3260–3271, 2022.

[19] "https://www.luisaviaroma.com/."

[20] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 561–578.

[21] Z. Cai, W. Yin, A. Zeng, C. Wei, Q. Sun, Y. Wang, H. E. Pang, H. Mei, M. Zhang, L. Zhang, C. C. Loy, L. Yang, and Z. Liu, "SMPLer-X: Scaling up expressive human pose and shape estimation," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[22] W. Yin, Z. Cai, R. Wang, A. Zeng, C. Wei, Q. Sun, H. Mei, Y. Wang, H. E. Pang, M. Zhang, L. Zhang, C. C. Loy, A. Yamashita, L. Yang, and Z. Liu, "SMPLest-X: Ultimate scaling for expressive human pose and shape estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 48, no. 2, pp. 1778–1794, 2026.

[23] S. Zhu, J. L. Chen, Z. Dai, Y. Xu, X. Cao, Y. Yao, H. Zhu, and S. Zhu, "Champ: Controllable and consistent human image animation with 3D parametric guidance," in *European Conference on Computer Vision (ECCV)*, 2024.

[24] Y. Guo, C. Yang, A. Rao, Y. Wang, Y. Qiao, D. Lin, and B. Dai, "Animatediff: Animate your personalized text-to-image diffusion models without specific tuning," *arXiv preprint arXiv:2307.04725*, 2023.

[25] S. Luo, Y. Tan, L. Huang, J. Li, and H. Zhao, "Latent consistency models: Synthesizing high-resolution images with few-step inference," *arXiv preprint arXiv:2310.04378*, 2023.

[26] S. Luo, Y. Tan, S. Patil, D. Gu, P. von Platen, A. Passos, L. Huang, J. Li, and H. Zhao, "LCM-LoRA: A universal stable-diffusion acceleration module," *arXiv preprint arXiv:2311.05556*, 2023.

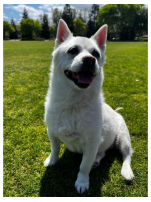
[27] J. Karras, Y. Li, N. Liu, L. Zhu, I. Yoo, A. Lugmayr, C. Lee, and I. Kemelmacher-Shlizerman, "Fashion-VDM: Video diffusion model for virtual try-on," in *SIGGRAPH Asia 2024 Conference Papers*, 2024.

[28] J. Zheng, F. Hu, Z. Xu, and Y. Yang, "VITON-DiT: Learning in-the-wild video try-on from human dance videos via diffusion transformers," *arXiv preprint arXiv:2405.18326*, 2024.

[29] J. Chen, J. Jiang, and J. Li, "Dynamic try-on: Taming video virtual try-on with dynamic attention mechanism," *arXiv preprint arXiv:2412.09822*, 2024.

[30] Z. Li, Y. Liu, and R. Chen, "VTON 360: High-fidelity virtual try-on from any viewing direction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Lucy (User1)



6 Appendix

6.1 Related Work Cont.

DensePose and Human Parsing: DensePose [13] establishes dense correspondences between RGB image pixels and the surface of a 3D human body model, effectively mapping every visible body pixel to a UV coordinate on one of 24 body part surfaces. This representation is critical for fashion applications because it tells the generative model precisely where on the body each pixel belongs, enabling garments to wrap correctly around the user's specific proportions. Self-Correction Human Parsing (SCHP) [18] provides complementary per-pixel semantic labels for body parts and clothing categories, which we use to generate the agnostic mask. The mask removes the user's current clothing while preserving identity-critical regions (face, hair, hands) and the body outline. Together, DensePose and human parsing provide the spatial conditioning that locks body shape during generation, addressing the core limitation identified in our earlier face-only approach.

3D Human Pose and Shape Estimation: Recovering the user's 3D body shape from casual photographs is essential for both body-faithful image generation and physically accurate video animation. The SMPL model [17] parameterizes body shape and pose through a learned low-dimensional space: 10 shape coefficients $\boldsymbol{\beta}$ that encode body proportions (height, weight, shoulder width, waist-to-hip ratio) and 72 pose parameters $\boldsymbol{\theta}$ that encode joint angles. SMPLify [20] first demonstrated automatic SMPL fitting from a single image by optimizing model parameters against detected 2D keypoints. More recent learning-based approaches directly regress SMPL parameters from images: HMR2.0 [16] achieves robust estimation through a ViT-based architecture trained on diverse datasets, while SMPLer-X [21] scales to the first generalist foundation model for expressive human pose and shape estimation, supporting SMPL-X with hands and face. The updated SMPLest-X [22] further improves accuracy through ultimate dataset and model scaling. In our pipeline we run HMR2.0 on the user's reference photos to extract their personal $\boldsymbol{\beta}$ parameters, which are then used in two ways: to condition the VTON model on the user's body shape, and to render personalized SMPL spin sequences for video generation where the 3D body matches the user's actual proportions.

SMPL-Guided Human Image Animation: Champ [23] introduces a methodology for human image animation that leverages the SMPL 3D parametric model within a latent diffusion framework. Given a reference image of a person and a driving motion sequence, Champ renders depth maps, normal maps, semantic segmentation maps, and skeleton guidance from the SMPL mesh at each frame, then fuses these multi-modal conditioning signals through a guidance encoder with self-attention mechanisms. The motion module, initialized from AnimateDiff [24] weights, ensures temporal coherence across frames. A key capability for our application is SMPL's separation of shape $\boldsymbol{\beta}$ and pose $\boldsymbol{\theta}$ parameters: we extract the user's body shape from their photos and combine it with the pose trajectory of a professional fashion model spin, then render personalized conditioning maps. This motion retargeting ensures the spinning body in the video matches the user's proportions, not those of the original driving video. This replaces our earlier approach of distilling Net-A-Porter model gifs into OpenPose skeletons, which captured only 2D joint positions and discarded all 3D body geometry information critical for realistic clothing animation.

Efficient Diffusion Inference: AnimateDiff [24] enables animation of personalized text-to-image models by inserting a motion module trained on video data into a frozen image diffusion model. The motion module learns temporal attention patterns from real-world videos, allowing it to generate coherent frame sequences without model-specific video fine-tuning. This plug-and-play design is particularly valuable for our application: the same motion module works with any personalized or fine-tuned base model, meaning we can combine it with our identity-conditioned SDXL pipeline without additional video-specific training. For image generation speed, Latent Consistency Models (LCM) [25] distill the iterative denoising process into a few-step procedure, reducing the required inference steps from 50 to as few as 4 while maintaining output quality. The LCM-LoRA [26] variant packages this acceleration as a lightweight adapter compatible with existing SDXL models and community LoRAs. In our deployment we apply LCM-LoRA to reduce per-image generation from approximately 15 seconds at 50 steps to under 3 seconds at 8 steps on the RTX4090, which is essential for the interactive e-commerce experience we target.

Fashion-Specific Video Diffusion: Recent work addresses virtual try-on specifically in the video domain. Fashion-VDM [27] proposes a video diffusion model that generates try-on videos given an input garment image and person video, preserving both garment details and person identity across frames. VITON-DiT [28] learns video try-on from human dance videos using diffusion transformers, while Dynamic Try-On [29] introduces dynamic attention mechanisms for temporal garment consistency. VTON 360 [30] achieves virtual try-on from arbitrary viewing directions, directly relevant to our spin generation task. These methods represent the convergence of VTON and video generation that our earlier work attempted with simpler tools. In our current pipeline we do not directly adopt these end-to-end video VTON models due to their substantial VRAM requirements and limited availability of pretrained weights, but instead compose a modular pipeline of IDM-VTON for image try-on followed by Champ-style SMPL-guided animation, which achieves a comparable result with components that are individually available and run on a single RTX4090.

6.2 Body-Aware Virtual Try-On

The three conditioning streams converge in a body-aware inpainting model that replaces the user's current clothing with the target garment. We adopt IDM-VTON [12] as our primary try-on engine, with SDXL inpainting plus IP-Adapter as a fallback.

IDM-VTON Architecture.: IDM-VTON uses a dual-UNet design built on SDXL. A frozen GarmentNet encodes the flat-lay garment image into multi-scale feature maps, while the main UNet — which is fully trainable — receives the agnostic image $\tilde{I}$, the inpainting mask M, and the DensePose map as conditioning inputs. The GarmentNet features are fused into the main UNet via cross-attention at

each resolution level, enabling fine-grained garment detail transfer. The model is trained to minimize:

$$\mathcal{L} = \mathbb{E}_{z,\epsilon,t}\left[||\epsilon - \epsilon_\theta(z_t, t, \tilde{I}, M, D, F_g, c)||_2^2\right], \quad (6)$$

where z_t is the noised latent at timestep t, D is the DensePose map, F_g are the GarmentNet features, and c is the text conditioning. The inpainting formulation ensures that the output image matches the user's body outline: the model generates new pixels only inside M, while all regions outside the mask (face, hands, background, and body contour) are copied directly from the original image.

Fallback Pipeline.: When IDM-VTON weights are unavailable, we construct an equivalent pipeline from standard components: SDXL inpainting provides the base inpainting capability, IP-Adapter injects the garment reference, and the agnostic mask constrains generation to the clothing region. This achieves comparable body preservation though with somewhat less precise garment detail transfer.

Inference Acceleration.: For both pipelines, we apply LCM-LoRA [26] to reduce the required denoising steps from 50 to 8 while maintaining output quality. This reduces per-image generation time from approximately 15 seconds to under 3 seconds on the RTX4090, which is essential for the interactive shopping experience we target. The LCM scheduler operates at a guidance scale of 1.5, substantially lower than the standard 7.5, which we find improves garment texture fidelity at low step counts.

6.3 SMPL-Guided Video Generation

Given the user's selected try-on image, we generate a 1–2 second fashion-model spin video. Our earlier approach distilled Net-A-Porter model gifs into OpenPose skeletons and transferred them onto the user via Stable Diffusion v1.4. This captured only 2D joint positions and discarded all 3D body geometry, producing videos where clothing draping bore no relation to the user's actual body shape. We replace this with a Champ-style [23] pipeline that uses the SMPL 3D body model for motion guidance.

Personalized Spin Sequence.: The critical insight is that SMPL separates pose $\boldsymbol{\theta}$ from shape $\boldsymbol{\beta}$. We construct a spin motion by defining a smooth 360° rotation in the global orientation parameter while maintaining a standing A-pose ($\boldsymbol{\theta}_{\text{arms}}$ offset slightly from the body). The user's estimated shape $\boldsymbol{\beta}_{\text{user}}$ from Section 3.5 is applied to the SMPL model at every frame, producing a 3D mesh sequence that rotates with the user's actual proportions:

$$\mathcal{M}_k = \text{SMPL}(\boldsymbol{\beta}_{\text{user}},\ \boldsymbol{\theta}_{\text{pose}},\ \mathbf{r}_k),$$
$$\mathbf{r}_k = \left(0,\ \frac{2\pi k}{K},\ 0\right), \quad k = 0, \ldots, K-1, \quad (7)$$

where K is the number of frames (24 in our setting at 16 fps, yielding 1.5 seconds) and $\mathbf{r}_k$ is the global orientation at frame k.

Multi-Modal Conditioning Maps.: From each SMPL mesh $\mathcal{M}_k$ we render four conditioning signals via rasterization:

- **Depth map**: orthographic projection of vertex depths, normalized to $[0, 1]$. Provides spatial layout and occlusion guidance.
- **Normal map**: per-vertex surface normals mapped to RGB via $(n + 1)/2$. Captures surface curvature for realistic lighting and fabric rendering.
- **Semantic map**: height-based body part segmentation (head, torso, upper legs, lower legs, feet) color-coded for the guidance encoder.
- **Skeleton**: projected 2D joint positions with SMPL kinematic tree connections, matching the DWPose format expected by the motion module.

These four maps are fused through the guidance encoder's self-attention layers, providing the diffusion model with comprehensive 3D shape and pose information at each frame.

Temporal Generation.: The video frames are generated using AnimateDiff [24], which inserts a motion module into a frozen Stable Diffusion v1.5 UNet. The motion module, initialized from weights pre-trained on diverse video data, applies temporal self-attention across the frame dimension, learning to produce coherent motion between frames without per-video training. The reference try-on image provides appearance conditioning, the SMPL-derived maps provide spatial guidance, and the text prompt provides semantic context. We generate 24 frames at 768×768 resolution on the RTX4090 (16 frames at 512×512 under memory constraints), encoded to MP4 at 16 fps.

Fallback Strategy.: We implement graceful degradation across three levels: (1) full Champ-style multi-modal conditioning when SMPL and all rendering dependencies are available; (2) AnimateDiff with depth-only conditioning when only the SMPL model is available; (3) frame-by-frame image-to-image diffusion with progressive rotation prompts as the ultimate fallback. Each level sacrifices temporal coherence and body accuracy but ensures the pipeline always produces output regardless of available dependencies.

6.4 Dataset and Implementation

In order to fine-tune the diffusion model towards current high-resolution fashion items, we construct a custom dataset from the e-commerce site Luisaviaroma [**?**]. This is critical to the Generative Stylist application since fashion items update at an extremely fast pace; what is available for purchase today may not exist tomorrow. We developed custom web scraping code using Selenium and BeautifulSoup to continuously download product images, storing records in a SQLite database to prevent duplicates. Each garment is stored in a two-level hierarchy: designer folders contain garment subfolders, each holding 2–5 high-resolution `.jpg` product photographs (front, back, and detail views) on clean white backgrounds. The folder name serves as the text descriptor for each garment, forming natural image-text pairs for CLIP indexing. Table 1 summarizes the dataset statistics.

The dataset spans a broad range of luxury and contemporary designers, from heritage houses such as Saint Laurent and Valentino to emerging labels like The Attico and Magda Butrym. The current scrape is focused on dresses, which comprise 99% of the catalog; extending to additional categories (tops, pants, outerwear, accessories) requires only re-running the scraper with updated category filters. At inference time, we index all 1,878 garments using CLIP ViT-H embeddings and retrieve the top-k matches for a given user prompt via cosine similarity, enabling natural-language queries such as

TABLE 1
Luisaviaroma fashion dataset statistics.

Statistic	Value
Total images	9,264
Unique garments	1,878
Distinct designers	186
Images per garment (min / median / max)	2 / 5 / 5
Garment category	99% dresses
Image format	`.jpg` (1024×1024)

TABLE 2
Top 10 designers by garment count in our dataset.

Designer	Items	Designer	Items
Self-portrait	72	The Attico	37
Zimmermann	58	Magda Butrym	35
Reformation	53	Pucci	33
Saint Laurent	40	Valentino	31
Alessandra Rich	39	Dolce & Gabbana	38

"*Show me wearing Alexander McQueen cocktail dresses*" to return semantically relevant items without keyword matching.

6.5 Body Shape Estimation Results

Table 3 reports the estimated SMPL shape parameters $\boldsymbol{\beta}$ for our test users. These 10-dimensional vectors encode the principal modes of body shape variation learned from thousands of 3D body scans in the SMPL training corpus [17]. In our current deployment, body shape is estimated from 2D pose landmarks via MediaPipe, which reliably recovers the first three shape components (β_0–β_2); the remaining components (β_3–β_9) require a full 3D regression model such as HMR2.0 [16] and default to zero in our lightweight configuration.

TABLE 3
Estimated SMPL shape parameters ($\boldsymbol{\beta}$) per user. Positive values indicate larger/taller/broader builds; negative values indicate slimmer/shorter/narrower builds. Values near zero indicate average proportions. User 2 fell back to the neutral SMPL mean due to insufficient full-body views.

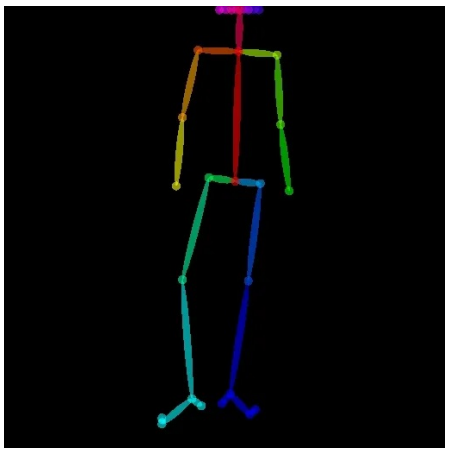
Parameter	Controls	User 1 (Miria)	User 2 (Lucy)	User 3 (Miria, run 2)
β_0	Overall body volume	-0.12	0.00	-0.12
β_1	Height / stature	-0.54	0.00	-0.54
β_2	Chest & shoulder width	-0.25	0.00	-0.25
β_3	Waist-to-hip ratio	0.00	0.00	0.00
β_4	Limb proportions	0.00	0.00	0.00
β_5–β_9 (higher-order)			≈ 0.00 (requires HMR2.0)	

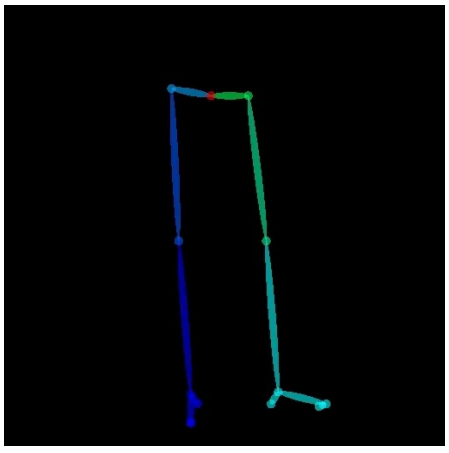
User 1 and User 3 correspond to the same individual (Miria) processed in separate runs, and their identical $\boldsymbol{\beta}$ values confirm the consistency of the estimation pipeline. The negative $\beta_1 = -0.54$ correctly captures a shorter stature, while $\beta_0 = -0.12$ and $\beta_2 = -0.25$ indicate a slightly slim build with narrow shoulders relative to the population mean. User 2 shows $\boldsymbol{\beta} = \mathbf{0}$ across all components, indicating the pose estimator fell back to the neutral SMPL mean—likely because the uploaded photos lacked clear full-body views or contained occlusions that prevented reliable landmark detection.

These shape differences directly impact the virtual try-on output. A user with higher β_0 produces a wider body silhouette in the agnostic mask, forcing the inpainting model to render the garment across a larger torso area. Similarly, β_2 affects how jacket shoulders and necklines are rendered, while β_3 (waist-to-hip ratio) determines whether A-line dresses flare naturally or hang straight. In the video spin generation, these same $\boldsymbol{\beta}$ values parameterize the SMPL mesh used to render depth and normal maps at each frame, ensuring the spinning 3D body matches each user's actual proportions rather than a generic mannequin.

We note two limitations of the current body estimation. First, the MediaPipe-based fallback only recovers 3 of the 10 SMPL shape components; deploying HMR2.0 or SMPLer-X [21] would yield finer-grained body shape recovery including waist-to-hip ratio, limb proportions, and torso length. Second, the estimation requires at least one clear full-body photograph—when this is unavailable, the pipeline gracefully degrades to the neutral mean body, which still produces reasonable but non-personalized try-on results.

Examples of pose data is as follows:





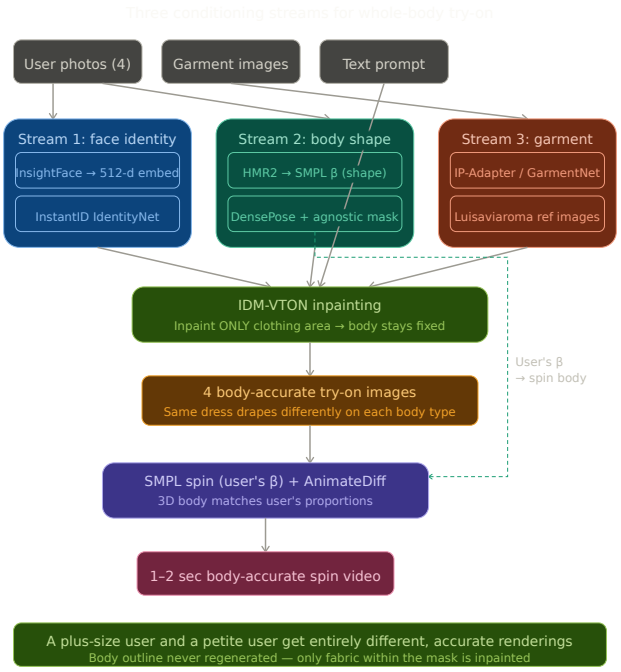


Fig. 6. LUCI video architecture. Three conditioning streams (face identity, body shape, and garment details) converge in a body-aware inpainting model. The user's body outline is never regenerated; only the clothing region is inpainted with the target garment. SMPL shape parameters from the user drive the spin animation so the 3D body matches actual proportions.