EgoAnimate: Generating Human Animations from Egocentric top-down Views via Controllable Latent Diffusion Models

G. Kutay Türkoğlu
Sony Semiconductor Solutions Europe
Germany
Guerbuez.Tuerkoglu@sony.com

Iheb Belgacem
Sony Semiconductor Solutions Europe
Germany
Iheb.Belgacem@sony.com

Julian Tanke
Sony AI
United States
Julian.Tanke@sony.com

Lev Markhasin
Sony Semiconductor Solutions Europe
Germany
Lev.Markhasin@sony.com

Abstract

An ideal digital telepresence experience requires the accurate replication of a person's body, clothing, and movements. In order to capture and transfer these movements into virtual reality, the egocentric (first-person) perspective can be adopted, which makes it feasible to rely on a portable and cost-effective standalone device that requires no additional front-view cameras. However, this perspective also introduces considerable challenges, particularly in learning tasks, as egocentric data often contains severe occlusions and distorted body proportions. Human appearance and avatar reconstruction from egocentric views remains relatively underexplored, and approaches that leverage generative priors are rare. This gap contributes to limited out-of-distribution generalization and greater data and training requirements. We introduce a controllable latent-diffusion framework that maps egocentric inputs to a canonical exocentric (frontal T-pose) representation from which animatable avatars are reconstructed. To our knowledge, this is the first system to employ a generative diffusion backbone for egocentric avatar reconstruction. Building on a Stable Diffusion prior with explicit pose/shape conditioning, our method reduces training/data burden and improves generalization to in-the-wild inputs. The idea of synthesizing fully occluded parts of an object has been widely explored in various domains. In particular, models such as SiTH and MagicMan have demonstrated successful 360-degree reconstruction from a single frontal image. Inspired by these approaches, we propose a pipeline that reconstructs a frontal view from a highly occluded top-down image using ControlNet and a Stable Diffusion backbone enabling the synthesis of novel views. Our objective is to map a single egocentric top-down image to a canonical frontal (e.g., T-pose) representation that can be directly consumed by an image-to-motion model to produce an animatable avatar. This enables motion synthesis from minimal egocentric input and supports more accessible, data-efficient, and generalizable telepresence systems.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

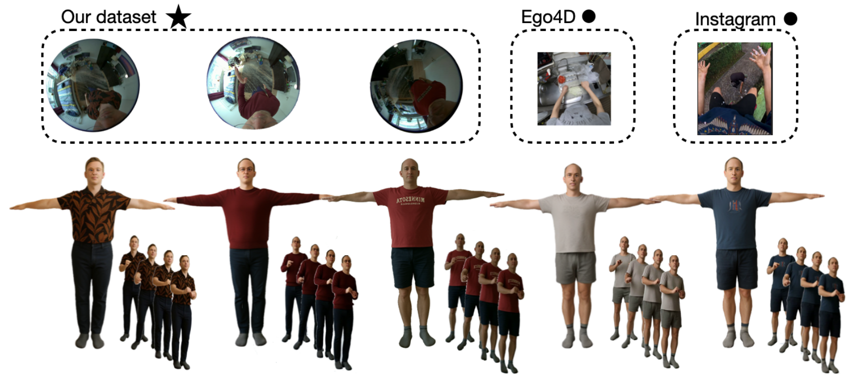

Figure 1: **EgoAnimate** synthesizes fully animatable avatars from a single egocentric top-down image, preserving clothing details while inferring occluded regions, such as pants and back views. This is accomplished by first converting the egocentric image into a frontal view utilizing our novel Egocentric-to-Frontal View Module and then applying off-the-shelf animation methods. The top row illustrates unseen egocentric inputs from our dataset ($\star$), one sample from Ego4D (Grauman et al., 2022), and one sample from Instagram. Notably, since the model was trained exclusively on our dataset ($\star$), all other inputs are considered out-of-distribution ($\bullet$). The bottom row demonstrates that, even with these egocentric inputs, our method produces consistent, animatable avatars. Unlike previous pipelines that require extensive capture setups and large datasets, EgoAnimate relies on only a single top-down image to enable downstream avatar reconstruction methods. Facial regions are not modeled, as egocentric head-mounted cameras do not capture faces (see Sec. A.3 for a discussion on dataset constraints and bias analysis).

1 Introduction

VR telepresence is an emerging technology that uses a head-mounted device (HMD) to enable people to communicate and interact with others while experiencing a sense of presence. Recent studies (Rzeszewski and Evans, 2020; Barreda-Ángeles and Hartmann, 2022) indicate the positive effects of this immersive experience during times of isolation, such as during the Covid-19 pandemic. To facilitate this feeling of presence, animatable full-body human avatars are an essential component of VR telepresence. To articulate the avatar, controllers or head-mounted cameras (HMCs) integrated into the HMD are used. Recently, EgoAvatar (Chen et al., 2024a) introduced HMC-driven full-body avatars with an impressive level of detail, where motion is driven only by the binocular HMC views. However, EgoAvatar requires a complex multi-camera studio setup to extract 3D avatars and cannot create avatars on the fly using only the HMCs.

Motivated by the recent successes in image- and video-to-avatar methods, we pose the following question: can a fully animatable avatar be created from just the HMD camera inputs alone? This would offer a simpler and more accessible solution to VR telepresence immersion that could be easily integrated while reducing the dependency on large-scale data capture.

While HMCs are cost-effective and readily available in existing HMDs, they pose a significant challenge for image-to-avatar pipelines due to the domain gap between egocentric HMC views and frontal views, which are typically used for image- or video-to-avatar systems such as ExAvatar (Moon et al., 2024), AnimateAnyone (Hu, 2024), or DreamPose (Karras et al., 2023). The domain gap stems from strong lens distortion, unusual camera perspectives, and significant body occlusion. To address this challenging task, we propose a simple yet effective two-step HMC-to-avatar baseline. As a first step and core contribution we introduce a diffusion-based canonization network which takes an egocentric image and generates a frontal T-pose view of the subject. This image is then used in the second step for avatar creation. Our pipeline is flexible and supports various methods for avatar generation.

Our egocentric-to-frontal view model draws inspiration from SiTH (Ho et al., 2024), but introduces key modifications. Specifically, we enhance the original loss function by combining the noise prediction loss with a perceptual loss in image space. This helps the model produce more visually plausible reconstructions. The model takes a 512×512 top-down frame as input, encodes it using a

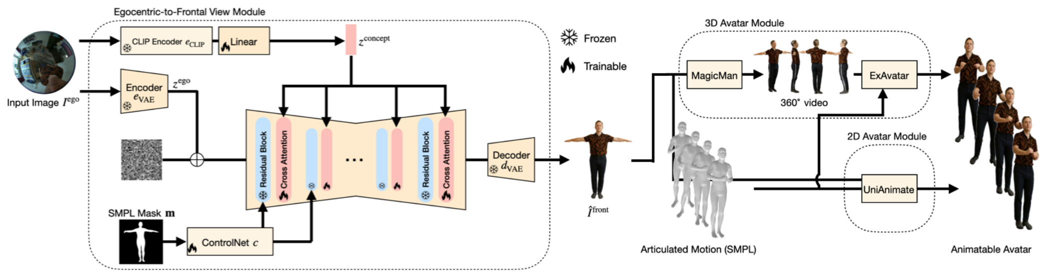

Figure 2: Method overview of **EgoAnimate**: Our approach is divided into two stages: 1. **Egocentric-to-Frontal View Module**: An egocentric top-down image, I^{ego}, is transformed into a canonical frontal view, $\hat{I}^{\text{front}}$, while preserving the subject's clothing. This transformation represents our primary contribution. 2. **Neural Avatar Integration**: In the second stage, we utilize pre-trained, off-the-shelf neural avatar methods, such as ExAvatar (Moon et al., 2024) or UniAnimate (Wang et al., 2025), which are driven by an articulated human pose model (Loper et al., 2015). This two-stage process demonstrates that our egocentric-to-frontal view module effectively bridges the domain gap between egocentric top-down views and conventional frontal-view images. Frontal-view images are more readily available and have been used to train the generic human avatar methods employed in the second stage. We conduct experiments with four neural avatar methods, balancing trade-offs between runtime performance and generation quality.

frozen Stable Diffusion VAE, and fuses it with noise-injected latent representations derived from ground truth samples. A U-Net with trainable transformer blocks performs denoising, guided by ControlNet (Zhang et al., 2023) pose input and additional cross-attention signals derived from CLIP embeddings (Radford et al., 2021).

In our experiments we explore two main directions: 3D-based avatars which allow for controllable animation and 2D-based avatars which offer higher visual quality. For the 3D avatar case we utilize ExAvatar (Moon et al., 2024), a state-of-the-art Gaussian Splatting (Kerbl et al., 2023) based method that can be animated using SMPL (Loper et al., 2015) sequences. As ExAvatar requires video input, we expand the generated frontal image into a 360° video using MagicMan (He et al., 2024), which generates multi-view human images from a single input. This preprocessing introduces minor artifacts and noise, reducing the final visual fidelity of the 3D avatar. For the 2D avatar case we experiment with UniAnimate (Wang et al., 2025), StableAnimator (Tu et al., 2024), and MimicMotion (Zhang et al., 2024), which directly animate from a single image. We find that UniAnimate performs best with our synthesized frontal view.

To summarize, our contributions are as follows:

- We present the first egocentric image-to-animatable avatar pipeline composed of simple building blocks that allow for easy extension and modification.
- We introduce an egocentric-to-frontal view synthesis module, which transforms heavily occluded egocentric images into plausible frontal views suitable for avatar creation.
- We evaluate several image-to-animation methods within this pipeline and identify which approaches best suit this novel task.
- We demonstrate that usable, animated avatars can be generated from minimal input and limited hardware using our dataset and simplified pipeline, providing a more generalizable and accessible solution to the egocentric-to-avatar problem.

2 Related Work

Diffusion Models: Diffusion models have achieved remarkable success in image generation. They have been applied to numerous tasks, including text-to-image generation (Nichol et al., 2021; Rombach et al., 2022), image editing (Brooks et al., 2023; Kawar et al., 2023; Couairon et al., 2023), novel view synthesis (Watson et al., 2022; Chan et al., 2023; Liu et al., 2023c; Kwak et al., 2024), and human body animation (Azadi et al., 2023; Xu et al., 2024). Trained on massive amounts of image-text data, they serve as image foundation models and have been used as powerful priors in many domains, particularly when data is scarce (Casas and Comino-Trinidad, 2023; Xiang et al., 2023; Song et al., 2021). Notably, Latent Diffusion (Rombach et al., 2022) enables efficient high-resolution image synthesis by operating in a compressed latent space. Another line of research has focused on improving the controllability of the produced images (Bar-Tal et al., 2023; Li et al., 2023; Mou et al., 2023). ControlNet (Zhang et al., 2023) and LoRA (Hu et al., 2021a) are very popular methods that have enabled conditioning the generation on a wide variety of modalities, such as depth maps or segmentation maps. In this paper, we base our work on a Stable Diffusion (Rombach et al., 2022) model and use ControlNet to control the pose of the generated image. Similarly to SiTH (Ho et al., 2024), we introduce an additional mechanism to condition the denoising process on the input image.

Generative Novel View Synthesis: The task of novel view synthesis (NVS) aims at generating images of a scene from unobserved viewpoints, given a set of input images. In this work, we address the task of egocentric-to-frontal view translation. This is a special case of NVS. Significant body occlusions and the use of a single input image make this a particularly challenging setting. In the era of deep learning, implicit representations such as Neural Radiance Fields (NeRF) (Mildenhall et al., 2021) and, more recently, 3D Gaussian Splatting (Kerbl et al., 2023) have achieved remarkable results. However, these methods typically operate in settings where a large number of input images are available and often perform poorly with sparse input.
In the sparse input setting, particularly in the challenging scenario of a single input image, NVS becomes a severely ill-posed problem. In this context, diffusion models have emerged as powerful generative priors, capable of producing plausible generations of unseen parts of the considered objects or scenes. Several works have leveraged diffusion models in various ways. One significant line of work, pioneered by DreamFusion (Poole et al., 2022), involves distilling the rich knowledge from pre-trained 2D diffusion models into implicit neural representations like NeRF, using a score distillation loss (Lin et al., 2022; Wang et al., 2023; Tang et al., 2023). Another line of research aims to incorporate 3D priors directly into the diffusion process (Wewer et al., 2024; Lan et al., 2024). More recently, enabled by the availability of large-scale synthetic 3D asset datasets such as Objaverse (Deitke et al., 2022), several works have adapted 2D diffusion models to directly predict novel views. This approach, pioneered by Zero-1-to-3 (Liu et al., 2023b), has shown an impressive level of generalization to unseen objects (Shi et al., 2023; Liu et al., 2023a,d).

Egocentric Avatars: While human avatars from third-person perspectives have received a lot of attention, the specific domain of egocentric avatars has been comparatively less explored. Much of the existing egocentric research has focused on aspects like 3D human pose estimation (Akada et al., 2022; Tomè et al., 2019, 2020; Akada et al., 2023; Millerdurai et al., 2024), or the creation of egocentric head avatars (Elgharib et al., 2019; Jourabloo et al., 2021; Chen et al., 2024b). To the best of our knowledge, very few works have attempted the challenging task of generating full-body egocentric avatars. Notably, EgoRenderer (Hu et al., 2021b) tackles this by decomposing the rendering into texture synthesis from fisheye inputs, egocentric pose construction, and a final neural image translation to a target view. Similarly, EgoAvatar (Chen et al., 2024a) learns a person-specific drivable avatar, represented by 3D Gaussians from multi-view videos. This model is then animated using a personalized egocentric pose detector trained on data from the same subject.
These methods have achieved remarkable results, however, they require extensive person-specific data for training their personalized models, making them difficult to scale. Our work aims to overcome this limitation by leveraging strong priors in the form of diffusion models to improve generalization to new subjects.

3 Method

In this work, we aim to address the challenging task of creating a fully animatable avatar from a single egocentric top-down image. To do so, we utilize a simple yet effective approach: we finetune a pre-trained latent diffusion model to convert an egocentric image I^{ego} into a frontal image $\hat{I}^{\mathrm{frontal}}$ which can then be used by off-the-shelf avatar animation methods. The flexibility of our pipeline allows us to experiment with two avatar paradigms, a 3D geometry approach based on Gaussian splats, and a 2D image-based approach.

3.1 Egocentric-to-Frontal Image Conversion

Figure 2 shows a more detailed overview of our Egocentric-to-Frontal image conversion pipeline. Taking the egocentric top-down image as input, it outputs a frontal image in T-pose preserving the clothing of the subject. Facial features are not part of this investigation, as most of the face lies outside the field of view of our camera. The reconstructed front image contains a synthesized face, unrelated to the input subject.

Latent backbone. We adopt the latent diffusion framework used in Stable Diffusion (Rombach et al., 2022). A frozen VAE encoder $\mathcal{E}_{\text{VAE}}$ compresses I^{ego} into a latent tensor z^{ego}. The forward diffusion noising process adds Gaussian noise following a linear variance schedule β_t:

$$z_t \;=\; \sqrt{\overline{\alpha}_t}\, z^{\text{ego}} + \sqrt{1-\overline{\alpha}_t}\,\epsilon, \quad t \in \{1,\ldots,T\},$$

with $\overline{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$ and $\epsilon\sim\mathcal{N}(0,\mathbf{I})$.

Pose-conditioned ControlNet. A ControlNet branch (Zhang et al., 2023) encodes the SMPL(Loper et al., 2015) pose map into spatial feature maps, which are integrated into the U-Net by adding them directly to the residual streams at both the downsampling and mid-block stages. For training phase, true SMPL masks of the subjects used. It is likewise for the testset we created. However for the O.O.D samples, we used a neutral coarse SMPL mask.

CLIP-guided egocentric top-down encoding. To extract semantically rich features from the egocentric image I^{ego}, we utilize a pretrained CLIP encoder (Radford et al., 2021). The resulting feature vector is projected and spatially expanded to match the resolution of the U-Net's latent space. These features are then fused with the VAE encoding and passed into the denoising U-Net via cross-attention. This design allows the model to leverage high-level visual semantics from the top-down input, improving its ability to synthesize plausible frontal appearances under severe occlusion.

Objective. Egocentric-to-Frontal is supervised by a compound loss:

$$\mathcal{L}_{\text{Ego2F}} = \lambda_{\text{diff}} \underbrace{\|\epsilon_\theta(\tilde{z}_t, \phi_P, t) - \epsilon\|_2^2}_{\mathcal{L}_{\text{diff}}} + \lambda_{\text{perc}} \underbrace{\text{LPIPS (Zhang et al., 2018)}(\hat{I}^{\text{front}}, I^{\text{front}})}_{\mathcal{L}_{\text{perc}}}$$

with $\lambda_{\text{diff}} = 1$ and $\lambda_{\text{perc}} = 0.2$.

3.2 Avatar Creation

We deliberately focus on a minimal solution under sparse egocentric input: translate to a canonical frontal representation and hand off to existing 2D/3D avatar modules. This isolates the contribution of the generative prior and quantifies how well the canonical view plugs into current state-of-the-art pipelines.

3.2.1 3D-Geometry-based Avatar

The 3D geometry-based avatar creation consists of 2 steps. In the first step we generate a full 360° view of the human subject from a given single ciew generated by our latent diffusion model(Rombach et al., 2022). To do this, we integrate the pre-trained multi-view synthesis module from MagicMan (He et al., 2024) directly into our pipeline. MagicMan is designed to produce consistent multi-view images from a single reference input and its corresponding 3D body mesh, in a single forward pass. We keep all weights frozen and do not fine-tune the model, allowing us to directly benefit from its state-of-the-art multi-view generation capabilities while focusing on improving the egocentric-to-frontal stage of our pipeline.

The second step of the avatar creation results in an animatable 3D avatar by transforming the synthesized multi-view image set. We use the off-the-shelf method ExAvatar (Moon et al., 2024). We do not modify, re-train, or re-implement any part of the ExAvatar pipeline. We directly apply its publicly released codebase on the 20 RGB images from the previous step and their corresponding normal maps.

Our goal in this step was to demonstrate that the frontal view synthesized from an egocentric top-down input enables downstream avatar reconstruction using existing, high-quality tools without any

Model Ablation		Egocentric Encoder				Full Body			Upper Body			Lower Body		
ControlNet	Perc.-Loss	SD VAE	Sapiens	CLIP [CLS]	CLIP grid	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
		✓			✓	12.11 ± 1.22	0.7294 ± 0.03	0.2694 ± 0.03	12.34 ± 1.35	0.7949 ± 0.025	0.1853 ± 0.027	11.90 ± 1.05	0.6589 ± 0.018	0.3188 ± 0.04
✓		✓			✓	17.05 ± 0.89	0.869 ± 0.0005	0.0987 ± 0.0001	17.73 ± 2.09	0.8143 ± 0.0511	0.1381 ± 0.0630	16.81 ± 0.56	0.8937 ± 0.012	0.0725 ± 0.017
✓			✓		✓	11.81 ± 0.60	0.7428 ± 0.006	0.200 ± 0.012	12.15 ± 0.85	0.8081 ± 0.015	0.1328 ± 0.020	11.50 ± 0.55	0.6772 ± 0.008	0.2613 ± 0.018
✓			✓	✓		15.66 ± 0.75	0.7730 ± 0.005	0.1858 ± 0.01	16.04 ± 0.95	0.8448 ± 0.012	0.1113 ± 0.011	15.30 ± 0.62	0.6998 ± 0.018	0.3188 ± 0.035
✓		✓	✓	✓		16.78 ± 0.67	0.8599 ± 0.003	0.1073 ± 0.05	17.16 ± 1.68	0.8147 ± 0.004	0.1456 ± 0.041	16.68 ± 0.62	0.8872 ± 0.011	0.0928 ± 0.015
✓	✓	✓		✓		17.73 ± 0.97	0.8743 ± 0.004	0.0835 ± 0.0006	18.32 ± 1.97	0.8260 ± 0.043	0.1174 ± 0.0429	17.53 ± 0.73	0.9026 ± 0.012	0.0725 ± 0.017

Table 1: Ablation results for different components of our pipeline. Our full model (bottom row) achieves the best performance across all metrics and body regions. We analyze the impact of ControlNet Zhang et al. (2023) conditioning, perceptual loss, Stable Diffusion VAE (SD VAE) Rombach et al. (2022), custom CLS-based CLIP encodingRadford et al. (2021), spatial grid-based CLIP encoding and Sapiens features. Results are reported for full body, upper body, and lower body regions using PSNR, SSIMWang et al. (2004), and LPIPSZhang et al. (2018) metrics.

Perc.-Loss	Encoder	Lower Body (%)
	SD VAE	62%
	Sapiens	21%
✓	SD VAE	**79%**

Table 2: **Clothing reconstruction accuracy on unseen samples for top-down to frontal synthesis.** Percentages reflect how often the generated outfit matches the ground truth in terms of coarse clothing type, such as correctly reproducing shorts versus pants, across 100 samples.

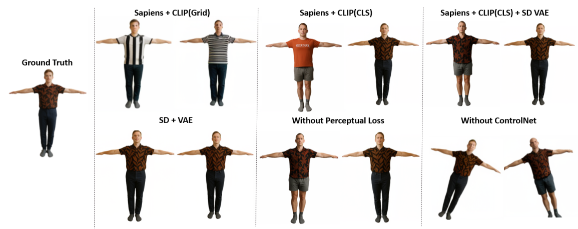

Figure 3: **Ablations on different Egocentric-to-frontal model configurations.** Our method consistently preserves both clothing type and pose across variations, demonstrating robustness compared to alternative configurations. Please zoom in for enhanced view.

additional training effort. By adopting ExAvatar as a drop-in module, we aim to isolate and evaluate the impact of our contribution: bridging the domain gap between sparse, occluded top-down images and the dense frontal inputs expected by state-of-the-art avatar generation methods. Our results show that once this gap is filled, avatar construction becomes straightforward using existing systems.

3.2.2 2D-Image-based Avatar

The 2D image-based avatar creation is a straight-forward application of UniAnimate (Wang et al., 2025) on our frontal T-pose input. UniAnimate synthesizes an animated sequence from the frontal image and some articulated motion.

During training, we apply the following augmentations:

- **Frontal perturbation (with probability** q**):** The frontal image and its corresponding body mask undergo light random transformations, including zoom-in cropping, additional center-shifts, and small rotations. This encourages robustness to variations in framing and pose alignment.
- **Independent top-down rotation (with probability** p**):** To prevent the model from overfitting to absolute positional cues or assuming fixed head/body orientation, we randomly apply a small global rotation to the top-down image independently of its ground-truth counterpart. This ensures that spatial layout is not trivially correlated with the camera-facing direction.

4 Experiments

4.1 Egocentric-to-Frontal Module Evaluation

4.1.1 Quantitative Results:

Baselines: We introduce the novel task of converting egocentric views to frontal T-posed images, establishing the first baseline for this challenging problem. To thoroughly investigate egocentric encoding strategies, we evaluate two alternative approaches: First, we replace the pre-trained stable diffusion VAE with Sapiens (Khirodkar et al., 2024) features—a model with strong human-specific priors—and apply convolutional downsampling to match the required latent resolution. This modification aims to leverage Sapiens' specialized human representation capabilities. Second, we enhance the CLIP embedding (Radford et al., 2021) pathway by incorporating a learnable transformer decoder that processes the CLIP ViT feature grid. This architecture theoretically enables more focused attention on human elements while suppressing background information.

Metrics: For our evaluation methodology, we adopt established metrics from the avatar generation literature (Ho et al., 2024; Moon et al., 2024; Wang et al., 2025) to assess image fidelity. Specifically, we quantify the correspondence between generated and ground-truth frontal images using PSNR (Peak Signal-to-Noise Ratio) for pixel-level accuracy, SSIM (Wang et al., 2004) (Structural Similarity Index) for perceptual structural alignment, and LPIPS (Zhang et al., 2018) (Learned Perceptual Image Patch Similarity) for feature-space similarity that better correlates with human visual perception.

We further introduce a novel evaluation metric specifically designed for egocentric-to-frontal view generation: coarse clothing type accuracy for both lower and upper body garments. This metric evaluates the model's ability to correctly infer fundamental clothing categories despite the challenging egocentric perspective. *Lower Body %* measures the accuracy in distinguishing between shorts and pants, while *Upper Body %* quantifies the accuracy in differentiating between t-shirts and sweaters. These metrics provide crucial insights into the model's capability to preserve semantic clothing attributes when reconstructing frontal views from partial egocentric observations.

Contrary to expectations, our experimental results indicate that neither modification yields performance improvements over our original architecture. We analyze the underlying factors for these outcomes in the following sections.

When comparing within-distribution reconstructions of printed tops (trained for the same number of epochs), 4, our method demonstrates that the integration of perceptual loss and SD VAE backbone yields significantly sharper and more faithful reproduction of logos, text, and other high-frequency details. This suggests that while OOD generalization remains an open challenge, our approach provides strong in-distribution fidelity and could be further improved by incorporating richer, more diverse datasets that include varied clothing styles and print patterns. We report our results on PSNR, SSIM (Wang et al., 2004) and LPIPS (Zhang et al., 2018) in Table 1 and our results on clothing reconstitution accuracy in Table 2. We find that both ControlNet and Perceptual loss improve overall generation performance, as can be furher seen in Figure 3.

Surprisingly, the strong human prior Sapiens (Khirodkar et al., 2024) is outperformed by the SD VAE. We hypothesis that this is due to the SD VAE being finetuned for the latent-diffusion architecture while Sapiens does not adhere to those expected distributions. Furthermore, Sapiens is not trained on egocentric views and thus tend to ignore lower-body parts, such as pants. This is further confirmed in its low lower-body accuracy in Table 2.

Similarly, learning features from the CLIP ViT grid (Radford et al., 2021) rather than directly utilizing the CLIP embedding (Radford et al., 2021) did not provide tangible improvements. We argue that this is due to the model only relying on high level contextual information in the CLIP (Radford et al., 2021) part of the model, for which the CLIP embedding is sufficient.

Hue & Saturation Consistency. We additionally evaluate color and texture preservation. On our held-out test split, the average ΔE_{00} color difference is 7.4 ± 1.9, and the 8-bin hue histogram intersection-over-union reaches 0.78. These metrics indicate that our model preserves approximately 80% hue consistency between predictions and ground truth, complementing the perceptual realism observed qualitatively.

4.1.2 Qualitative Results

Effect of the Encoder: When replacing the Stable Diffusion VAE with the Sapiens encoder, we observe that clothing fidelity and texture sharpness are significantly reduced. In particular, fine-grained features such as logos or garment folds are poorly reconstructed, highlighting the importance of a pretrained SD VAE backbone for robust appearance recovery.

Perceptual Loss: Removing the perceptual loss while keeping the noise prediction objective leads to random patterns on garments, and incorrect garment structure. This indicates that the perceptual loss

Figure 4: **Reconstructed Top-Wear:** Our method more accurately reconstructs printed designs and predicts clothing types compared to ablations. Please zoom in for details.

Option	BS (↑)	MR (↓)	Runtime (↓)
ExAvatar	33	3.20	**1.0**
MimicMotion	10	3.76	2.0
StableAnimator	86	1.90	3.75
UniAnimate	**117**	**1.15**	22.5

Table 3: User study results comparing ExAvatar Moon et al. (2024),MimicMotion Zhang et al. (2024),StableAnimator Tu et al. (2024),**UniAnimate** Wang et al. (2025). Runtime is relative to the fastest method. BS = Borda Score, MR = Mean Rank.

provides valuable image-space guidance, encouraging reconstructions that are visually closer to the ground truth, rather than only aligned in latent space.

ControlNet Conditioning: Without ControlNet, synthesized results fail to align with the target pose, often producing distorted limb proportions or incorrect body orientation. Conditioning on the SMPL-derived mask via ControlNet ensures both structural correctness and consistency across poses.

Full Model: Our full configuration (Sapiens + CLIP/SD features + perceptual loss + ControlNet + SD VAE) produces the most reliable outputs. Reconstructions preserve both clothing type and pose across variations, and generalize better to unseen samples. While subtle artifacts remain in challenging cases (e.g., loose clothing or extreme lighting conditions), this setup achieves the best trade-off between structural plausibility and visual realism.

4.2 Avatar Animation Module Evaluation

To animate our reconstructed frontal images, we evaluated four state-of-the-art methods: MagicMan He et al. (2024) + ExAvatar Moon et al. (2024) (a pipeline that first generates 360-degree views before creating a 3D avatar), MimicMotion Zhang et al. (2024), StableAnimator Tu et al. (2024), and UniAnimate-DiT Wang et al. (2025). This comparative analysis aimed to identify the most effective approach for generating high-quality animations from egocentric perspectives.

Our evaluation incorporated both computational performance metrics and perceptual quality assessments through a comprehensive user study with 41 participants. Given the absence of ground-truth animations with matching clothing, we employed a ranking-based evaluation protocol where participants ranked animations across five clothing variations according to three criteria: clothing consistency, motion realism, and animation smoothness. To maintain evaluation focus, participants were explicitly instructed to disregard heads and backgrounds while concentrating on motion integrity, limb movement coherence, and overall visual quality. Interestingly, the overall model performance negatively correlates with the method runtime.

As presented in Table 3, UniAnimate Wang et al. (2025) emerged as the preferred method, followed by StableAnimator Tu et al. (2024), ExAvatar Moon et al. (2024), and MimicMotion Zhang et al. (2024). Based on these findings, we implemented UniAnimate Wang et al. (2025) as our primary animation framework for the results presented throughout this paper.

5 Conclusion

In this paper, we present a novel modular pipeline that addresses the TopDown-to-Avatar problem using a combination of generative novel view synthesis and off-the-shelf avatar construction tools. By reframing the challenging task of egocentric avatar reconstruction into two simpler tasks- egocentric-to-frontal image translation and frontal image animation-we enable high-quality and efficient full-body avatar generation from a single egocentric image.

Our primary contribution is the development of a TopDown-to-Frontal (T2F) synthesis module built upon Stable Diffusion and ControlNet. This component produces realistic frontal views from top-down inputs, while being robust to both pose variation and occlusion. We also construct a small, easy-to-use, yet carefully curated dataset of top-down images paired with DALL·E-enhanced frontal views. We further show that by combining our frontal outputs with pre-trained multi-view generation

(MagicMan) and avatar modeling (ExAvatar), the full avatar pipeline generalizes to new identities without the need for additional fine-tuning or retraining.

We discuss limitations and future work in the A.3.

References

Hiroyasu Akada, Jian Wang, Soshi Shimada, Masaki Takahashi, Christian Theobalt, and Vladislav Golyanik. Unrealego: A new dataset for robust egocentric 3d human motion capture. In *European Conference on Computer Vision*, 2022.

Hiroyasu Akada, Jian Wang, Vladislav Golyanik, and Christian Theobalt. 3d human pose perception from egocentric stereo videos. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 767–776, 2023.

Samaneh Azadi, Akbar Shah, Thomas Hayes, Devi Parikh, and Sonal Gupta. Make-an-animation: Large-scale text-conditional 3d human motion generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15039–15048, 2023.

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. In *International Conference on Machine Learning*, 2023.

Miguel Barreda-Ángeles and Tilo Hartmann. Psychological benefits of using social virtual reality platforms during the covid-19 pandemic: The role of social and spatial presence. *Computers in Human Behavior*, 2022.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023.

Dan Casas and Marc Comino-Trinidad. SMPLitex: A Generative Model and Dataset for 3D Human Texture Estimation from Single Image. In *British Machine Vision Conference (BMVC)*, 2023.

Eric R. Chan, Koki Nagano, Matthew A. Chan, Alexander W. Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4217–4229, 2023.

Jianchun Chen, Jian Wang, Yinda Zhang, Rohit Pandey, Thabo Beeler, Marc Habermann, and Christian Theobalt. Egoavatar: Egocentric view-driven and photorealistic full-body avatars. In *SIGGRAPH Asia*, 2024a.

Zheng Chen, Zhiqi Zhang, Junsong Yuan, Yi Xu, and Lantao Liu. Show your face: Restoring complete facial images from partial observations for vr meeting. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8673–8682, 2024b.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13142–13153, 2022.

Mohamed A. Elgharib, R. MallikarjunB., Ayush Kumar Tewari, Hyeongwoo Kim, Wentao Liu, Hans-Peter Seidel, and Christian Theobalt. Egoface: Egocentric face performance capture and videorealistic reenactment. *ArXiv*, abs/1905.10822, 2019.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Conference on computer vision and pattern recognition*, 2022.

Xu He, Xiaoyu Li, Di Kang, Jiangnan Ye, Chaopeng Zhang, Liyang Chen, Xiangjun Gao, Han Zhang, Zhiyong Wu, and Haolin Zhuang. Magicman: Generative novel view synthesis of humans with 3d-aware diffusion and iterative refinement, 2024.

Hsuan-I Ho, Jie Song, and Otmar Hilliges. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021a.

Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Conference on Computer Vision and Pattern Recognition*, 2024.

T. Hu, Kripasindhu Sarkar, Lingjie Liu, Matthias Zwicker, and Christian Theobalt. Egorenderer: Rendering human avatars from egocentric camera images. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14508–14518, 2021b.

Amin Jourabloo, Fernando De la Torre, Jason M. Saragih, Shih-En Wei, Tenia Wang, Stephen Lombardi, Danielle Belko, Autumn Trimble, and Hernán Badino. Robust egocentric photo-realistic facial expression transfer for virtual reality. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20291–20300, 2021.

Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *International Conference on Computer Vision*, 2023.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Conference on Computer Vision and Pattern Recognition 2023*, 2023.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):75:1–75:14, 2023.

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.

Jeong-gi Kwak, Erqun Dong, Yuhe Jin, Hanseok Ko, Shweta Mahajan, and Kwang Moo Yi. Vivid-1-to-3: Novel view synthesis with video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6775–6785, 2024.

Yushi Lan, Fangzhou Hong, Shuai Yang, Shangchen Zhou, Xuyi Meng, Bo Dai, Xingang Pan, and Chen Change Loy. Ln3diff: Scalable latent neural fields diffusion for speedy 3d generation. In *European Conference on Computer Vision*, 2024.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22511–22521, 2023.

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 300–309, 2022.

Minghua Liu, Chao Xu, Haian Jin, Ling Chen, T. MukundVarma, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *ArXiv*, abs/2306.16928, 2023a.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9264–9275, 2023b.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9298–9309, 2023c.

Yuan Liu, Chu-Hsing Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *ArXiv*, abs/2309.03453, 2023d.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 65(1):99–106, 2021.

Christen Millerdurai, Hiroyasu Akada, Jian Wang, Diogo C. Luvizon, Christian Theobalt, and Vladislav Golyanik. Eventego3d: 3d human motion capture from egocentric event streams. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1186–1195, 2024.

Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3d gaussian avatar. In *European Conference on Computer Vision*, 2024.

Chong Mou, Xintao Wang, Liangbin Xie, Jing Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI Conference on Artificial Intelligence*, 2023.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. *CoRR*, abs/2112.10741, 2021.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *ArXiv*, abs/2209.14988, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

Michał Rzeszewski and Leighton Evans. Virtual place during quarantine–a curious case of vrchat. *Rozwój Regionalny i Polityka Regionalna*, 2020.

Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and X. Yang. Mvdream: Multi-view diffusion for 3d generation. *ArXiv*, abs/2308.16512, 2023.

Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. *ArXiv*, abs/2111.08005, 2021.

Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *ArXiv*, abs/2309.16653, 2023.

Denis Tomè, Patrick Peluse, Lourdes de Agapito, and Hernán Badino. xr-egopose: Egocentric 3d human pose from an hmd camera. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7727–7737, 2019.

Denis Tomè, Thiemo Alldieck, Patrick Peluse, Gerard Pons-Moll, Lourdes de Agapito, Hernán Badino, and Fernando de la Torre. Selfpose: 3d egocentric pose estimation from a headset mounted camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:6794–6806, 2020.

Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, and Zuxuan Wu. Stableanimator: High-quality identity-preserving human image animation. *arXiv preprint arXiv:2411.17697*, 2024.

Xiang Wang, Shiwei Zhang, Changxin Gao, Jiayu Wang, Xiaoqiang Zhou, Yingya Zhang, Luxin Yan, and Nong Sang. Unianimate: Taming unified video diffusion models for consistent human image animation. *Science China Information Sciences*, 2025.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *ArXiv*, abs/2305.16213, 2023.

Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. Available at `https://arxiv.org/abs/2210.04628`.

Christopher Wewer, Kevin Raj, Eddy Ilg, Bernt Schiele, and Jan Eric Lenssen. latentsplat: Autoencoding variational gaussians for fast generalizable 3d reconstruction. In *European Conference on Computer Vision*, 2024.

Tiange Xiang, Mahmut Yurt, Ali B. Syed, Kawin Setsompop, and Akshay S. Chaudhari. Ddm2: Self-supervised diffusion mri denoising with generative diffusion models. *ArXiv*, abs/2302.03018, 2023.

Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1481–1490, 2024.

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 (ICCV)*, pages 3813–3824, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024.

A Technical Appendices and Supplementary Material

A.1 Dataset

We curate a custom dataset tailored to the egocentric-to-frontal synthesis task. All data were collected by the first author using a single helmet-mounted RGB camera (Basler sensor) to capture the egocentric top-down view. Each motion sequence lasts approximately 7 seconds, recorded at 14 fps, yielding about 100 frames per sequence. From these, every 10th frame is sampled as a training input, resulting in $\sim$3,000 top-down frames overall. After each sequence, the helmet was removed and a static frontal T-pose was recorded to serve as ground truth for that clothing combination. In total, $\sim$300 frontal captures were obtained.

To ensure clothing diversity, we recorded 30 distinct upper-body garments (T-shirts, shirts, hoodies, jackets, sweaters) and 10 lower-body garments (pants and shorts). Lower-body garments are split roughly 40% shorts and 60% long pants. Upper-body garments include $\sim$40–60% short sleeves, with the remaining long sleeves consisting of $\sim$30% shirts, $\sim$10% hoodies/jackets, and the rest sweaters or long-sleeved T-shirts. No footwear was worn during data collection, resulting in all reconstructions showing socks.

The frontal T-pose images were post-processed to correct for inconsistent lighting and resolution. Specifically, we applied a diffusion-based enhancer (DALL-E family) with the following prompt: *"Enhance this image of the subject to look like it was taken in a professional studio. Keep the person, pose, angle, clothing, and background exactly the same. Do not make any artistic or stylistic changes — only improve the lighting, color balance, and clarity subtly."*

We use these enhanced frontals as appearance guidance rather than pristine ground truth. Each frontal capture is paired with $\sim$10 temporally aligned top-down frames based on timestamps and coarse pose cues. To further increase data variation, we generate photometric and mild geometric augmentations, and mirror all top-down samples to mitigate overfitting. This yields 600 frontal training instances (300 original + 300 augmented) and $\sim$3,000 top-down frames.

During training, we divided the dataset into 90% train and 10% test set, which leaves 540 images for the train set and 60 images for the test set. 60 images on the test set were not shown to the model during training phase.

Limitations of the dataset include demographic bias, as all sequences were recorded indoors by a single light-skinned male subject. Despite the limited diversity, we attempted to capture a wide variety of clothing styles and appearances.

A.2 Computation

Our egocentric-to-frontal model was trained for 200k steps on a single NVIDIA RTX A6000 GPU with 48GB memory. Training took approximately 48 hours using mixed precision (fp16) optimization. The network operates at a resolution of 512×512, and training batches consisted of 16 samples per step.

At inference time, generating a single frontal T-pose image from a top-down input requires $\sim$1.2 seconds on the same GPU. Memory usage is $\sim$11GB during inference. Runtime and memory requirements allow the model to be deployed on high-end consumer GPUs and optimized server-side setups.

For downstream animation, we benchmarked the integration of our synthesized frontal views with off-the-shelf animation modules. On the same hardware, UniAnimate achieves $\sim$0.03 fps, StableAnimator $\sim$0.23 fps, MimicMotion $\sim$0.44 fps, and ExAvatar $\sim$0.88 fps. These values highlight the trade-off between visual fidelity and computational cost across different modules.

Overall, our pipeline is computationally lightweight at the synthesis stage, and compatible with current image-to-animation systems, making it practical for research prototyping and future XR deployment.

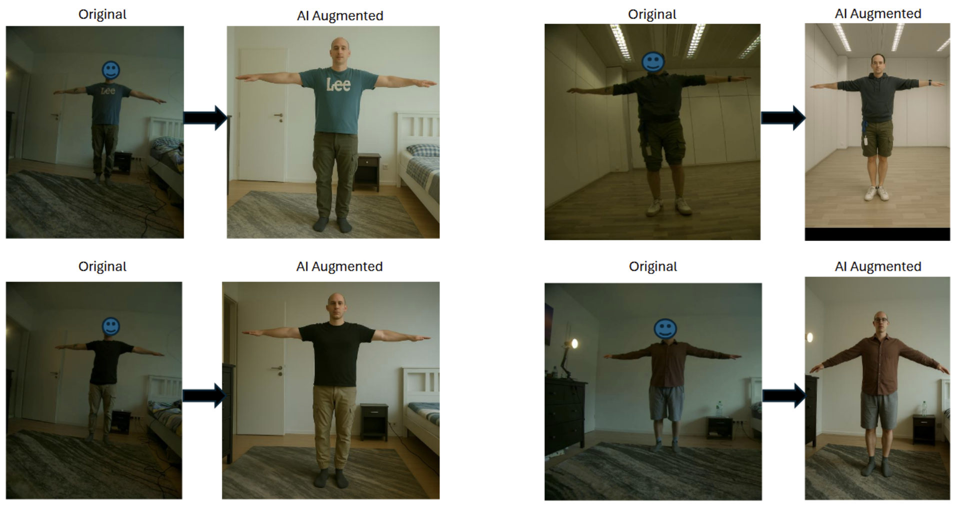

Figure 5: Examples of raw frontal captures and their enhanced counterparts. Since recordings were made in a home environment rather than a studio, the original images suffer from poor and inconsistent lighting, visible shadows, perspective distortions, and low photographic quality. To obtain more reliable supervision, we applied a diffusion-based enhancer (DALL·E family) with a prompt restricted to subtle photometric corrections (lighting, color balance, sharpness), without altering pose, angle, or clothing. The resulting outputs are substantially clearer and closer to studio quality, providing more stable appearance guidance for training.

A.3 Limitations and Future Work

- **Limited Demographic Diversity.** Our dataset primarily includes individuals with similar body types and light skin tones, recorded in home environments. As a result, reconstructed feet typically appear with socks, and the current model is biased toward light-skinned male appearances. Broader demographic representation—including different body shapes, skin tones, and clothing—would improve fairness and robustness.
- **OOD Generalization.** Although we demonstrate results on unseen datasets such as Ego4D and Instagram, the true robustness across diverse populations still remains limited, and improving OOD generalization is an important direction for future work.
- **No Face Modeling.** We intentionally exclude facial regions due to occlusion in top-down views. While this simplifies the task and avoids identity leakage, it limits realism in use cases like full telepresence.
- **Dependency on Clothing Visibility.** Although our pipeline infers occluded regions, highly ambiguous clothing geometries (e.g., long coats or flowing garments) can degrade reconstruction quality. Incorporating temporal cues from short top-down video clips may improve consistency.
- **Synthetic Appearance in Certain Cases.** Despite using perceptual and pose losses, the synthesized frontal images can exhibit subtle artifacts, especially in cases with extreme occlusion or unusual body posture. These artifacts are often smoothed out in downstream animation, but improving image realism remains an active research direction.
- **Garment Patterns:** While our model shows clear improvements in reconstructing clothing details within the distribution of our custom dataset (Figure 4), one of its current weaknesses

lies in handling out-of-distribution (OOD) scenarios. As highlighted earlier in Figure 1, when tested on Ego4D Grauman et al. (2022) or casual Internet images (e.g., Instagram samples), the model struggles to faithfully reconstruct fine-grained print patterns or logos on garments. This is likely due to the lack of diverse printed clothing in the training distribution, which limits the generalization ability of the latent representations.

Figure 6: Example image from Ego4D Grauman et al. (2022) of one of the failure cases that the model is not able to distinguish short sleeve and long sleeve shirt.

Besides print out and real life generalizability, as it can be seen from 6, model sometimes confuses with the short and long sleeve objects, particularly when they are not in the training dataset.

A.4 Avatar Reconstruction Module

To further evaluate the applicability of our canonical T-pose reconstructions, we tested whether a single synthesized T-pose combined with motion sequences is sufficient for downstream image-to-motion models. As illustrated in Figure 2, we extracted SMPL Loper et al. (2015) parameters (for ExAvatar Moon et al. (2024)) and keypoints (for StableAnimatorTu et al. (2024), UniAnimate Wang et al. (2025), and MimicMotionZhang et al. (2024)) from motion sequences recorded by the authors, and used these as inputs together with our reconstructed canonical T-poses.

This setup enables a direct comparison between 2D and 3D reconstruction pipelines. Our findings show that stacking modules, such as MagicManHe et al. (2024) followed by ExAvatar Moon et al. (2024), significantly decreases the visual quality of 3D motion reconstructions. In contrast, UniAnimate Wang et al. (2025) and StableAnimator Tu et al. (2024) better preserve image quality, with UniAnimate standing out as the only model capable of maintaining realistic clothing foldings. Interestingly, while stacked modules (e.g., MagicMan + ExAvatar) degrade performance, they sometimes benefit MimicMotion Zhang et al. (2024) and ExAvatar Moon et al. (2024), whereas UniAnimate Wang et al. (2025) and StableAnimator Tu et al. (2024) perform best when used directly without stacking. Overall, this analysis demonstrates that a single SMPL-based T-pose can suffice for driving both 2D and 3D avatar motion models, though reconstruction quality strongly depends on the chosen downstream method.

Figure 7: Motion reconstruction results of our synthesized frontal views evaluated with several state-of-the-art avatar animation models: StableAnimator Tu et al. (2024), UniAnimate Wang et al. (2025), MimicMotion Zhang et al. (2024), and MagicMan + ExAvatar Moon et al. (2024). The comparison highlights differences in clothing preservation, body consistency, and motion smoothness across methods when driven by our EgoAnimate frontal outputs.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction introducing a novel egocentric-to-frontal avatar generation pipeline, dataset, and evaluation match the contributions, experiments, and conclusions presented in the paper .

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A dedicated Limitations and Future Work section discusses dataset bias (single subject, light-skinned male), lack of face modeling, dependence on clothing visibility, out-of-distribution generalization issues, and synthetic artifacts .

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not present theoretical results or formal proofs; it focuses on empirical evaluation of generative and animation models.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes dataset composition, augmentations, training procedures, evaluation metrics, and baselines in detail, which are sufficient to replicate the main experimental results .

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: While dataset creation and model details are described, the paper does not mention releasing code or dataset publicly. Reproducibility relies on the methodological descriptions .

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification:The dataset splits, augmentations, training duration, optimizer setup, and evaluation metrics are described (e.g., 90/10 split, RTX A6000 GPU, PSNR/SSIM/LPIPS metrics) .

   Experiment statistical significance

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Results include quantitative metrics with comparisons across baselines and perceptual user studies (41 participants). While not all include error bars, statistical rigor is shown via multiple metrics and comparative ranking .

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: the paper specifies training hardware (RTX A6000, 200k steps, 48 hours), inference costs, and runtime comparisons across animation modules .

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work aligns with NeurIPS Ethics guidelines — sensitive areas like face modeling were intentionally excluded to reduce identity leakage, and limitations such as dataset bias are explicitly acknowledged .

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: While not a separate section, societal considerations are addressed: identity privacy is preserved by excluding faces, but limitations note potential demographic bias and generalization issues, which carry fairness implications .

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The models and dataset presented are relatively small-scale research contributions without direct high-risk misuse potential (e.g., not a large pretrained LLM or web-scraped dataset).

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Prior datasets (e.g., Ego4D, Instagram references) are cited appropriately; however, the main dataset was self-collected by the authors .

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: A new dataset (top-down captures with paired frontal images) is introduced. Its documentation includes collection setup, demographics, garment diversity, and augmentation methods

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: A user study with 41 participants was conducted for animation evaluation. Instructions and evaluation protocol (ranking animations by realism, smoothness, and clothing consistency) are described .

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not mention IRB approval. Since participants only evaluated rendered animations without personal risk, IRB approval may not have been required.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The paper explicitly states that DALL·E (a large diffusion model) was used to enhance lighting/clarity of frontal captures for training supervision .

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.