Pippo: High-Resolution Multi-View Humans from a Single Image

Yash Kant[1,2,3], Ethan Weber[1,4], Jin Kyu Kim[1], Rawal Khirodkar[1], Su Zhaoen[1], Julieta Martinez[1], Igor Gilitschenski[*,2,3], Shunsuke Saito[*,1], Timur Bagautdinov[*,1]
[1]Meta Reality Labs, [2]University of Toronto, [3]Vector Institute, [4]UC Berkeley
https://yashkant.github.io/pippo/
[*]Joint Advising

Figure 1. **Pippo produces high-resolution, multi-view, studio-quality images from a single image.** First row shows full-body generation of an unseen subject captured from an iPhone picture. Second and third rows show face-only and partial full-body input images from unseen studio subject(s) to novel views respectively. All visuals are at 1K resolution.

Abstract

We present Pippo, a generative model capable of producing 1K resolution dense turnaround videos of a person from a single casually clicked photo. Pippo is a multi-view diffusion transformer and does not require any additional inputs — e.g., a fitted parametric model or camera parameters of the input image. We pre-train Pippo on 3B human images without captions, and conduct multi-view mid-training and post-training on studio captured humans. During mid-training, to quickly absorb the studio dataset, we denoise several (upto 48) views at low-resolution, and encode tar-

get cameras coarsely using a shallow MLP. During post-training, we denoise fewer views at high-resolution and use pixel-aligned controls (e.g., Spatial anchor and Plucker rays) to enable 3D consistent generations. At inference, we propose an attention biasing technique that allows Pippo to simultaneously generate greater than 5× as many views as seen during training. Finally, we also introduce an improved metric to evaluate 3D consistency of multi-view generations, and show that Pippo outperforms existing works on multi-view human generation from a single image.

1. Introduction

Creating photorealistic human representations with the ability to control viewpoints has numerous applications in entertainment, healthcare, fashion and social media. Building such representations, first and foremost, requires high-quality multi-view studio data [2, 29], which is costly to acquire. This significantly limits the scalability in terms of the number identities for high-quality studio data [29, 43, 84]. In contrast, large-scale, unstructured, and diverse human images and videos are available online. However, the raw data of such in-the-wild images does not offer ground-truth 3D or multi-view representations of humans.

In this work, we present a novel approach leveraging the best of two worlds: generalizability from in-the-wild unstructured images, and fidelity and view-controllability from studio capture data. Specifically, our model Pippo is a diffusion transformer, which can generate several 1K resolution multi-view consistent images jointly during inference. Pippo takes as input a single image of an individual, camera poses of the target viewpoints to be generated. Since the scale and placement of the subject is ambiguous from a single image, Pippo uses a Spatial Anchor which roughly specifies the location and orientation of the subject in 3D space. Our model does not rely on additional conditioning such as body priors or camera parameters of the input images, to scale our training pipeline to in-the-wild data and support unconstrained inputs at test time.

We employ a multi-stage training recipe to train Pippo. First, we pre-train the model for latent-to-image generation task, similar to [54], on a large human-centric dataset of in-the-wild images. Next, in Mid-training stage, we jointly generate multiple consistent images of the subject, conditioned on target viewpoints and a single input image using high-quality studio dataset. Finally, in Post-training stage, the model is provided with a minimal placement signal - *spatial anchor* - encoding rough head orientation - which further improves 3D consistency. Our architecture is also carefully tailored to the conditional multi-view generation - we propose several simple but effective modifications to the basic DiT architecture, including self-attention-based conditioning, lightweight spatial controls and camera conditioning with Plücker coordinates.

For evaluating such a multi-view diffusion model, 3D consistency is a critical metric for understanding the geometric correctness of the generation. For generation task in an ill-posed setup (e.g., a single image input), there may be multiple possibilities that are equally plausible. Existing multi-view generation methods typically report reconstruction metrics (i.e., PSNR, SSIM and LPIPS) [19] or FID [41], which either (1) penalize any new content that is actually 3D consistent or (2) are unable to measure the 3D consistency of the generation. To address this problem, we design a metric measuring 3D consistency: we compute 2D keypoint matches with an off-the-shelf method [59], triangulate them, and then reproject back into the other views to measure the reprojection error in pixels. Some generation methods measure reprojection errors [17] from SfM [60] or epipolar distance [45, 89], but our metric uses camera pose as input and is more precise than measuring distance to a epipolar line. We show our metric helps to quantify our results, finding that our method is favorable compared with existing approaches and other baseline methods.

To summarize, we make the following contributions:

- a generative model capable of generating high-resolution and multi-view consistent humans from a single image and its effective training strategy.
- a diffusion transformer architecture designed to enhance multi-view generation and viewpoint control.
- a novel 3D consistency metric tailored to accurately measure the level of 3D consistency in generative tasks.

2. Related Work

We review multi-view human datasets and generative models for human synthesis, categorizing methods by their data requirements and prior constraints—such as parametric human models or explicit 3D structures. In this work, we minimize reliance on complex priors by training the model on large amounts of human-centric data and impose minimal camera and spatial controls.

Multi-view Human Datasets. While many large 2D human datasets exist [12, 39, 62], 3d captures of people (i.e., captures using dozens of views at a time) are still relatively rare, since they are expensive to collect and are thus mostly obtained in specialized research labs. Nevertheless, this kind of data has recently become more common, going from hundreds to thousands of publicly-available captures. For face-only captures, a few datasets [43, 47, 87, 88, 90] provide up to 600 subjects captured from over 60 views each, while for full-body captures, 4D-Dress [77] provides 32 subjects with 2 outfits each, and detailed annotations for their garments, while DNA-rendering [8] provides 500 subjects from 60 views. MVHumanNet [84] stands apart by

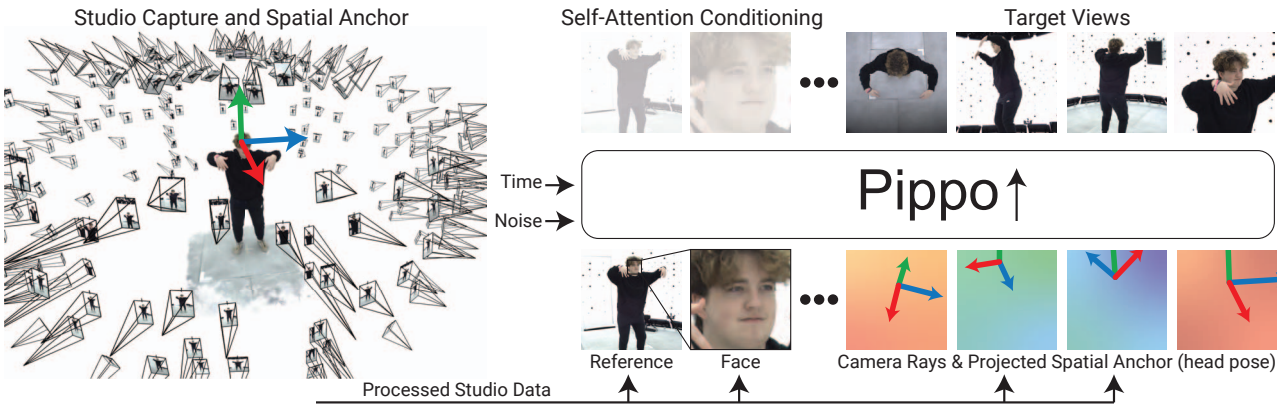

Figure 2. **Pipeline overview**. This is an illustration of how we train our model. (Left) we use data from a studio capture and train our multi-view diffusion model (right). We condition on a full reference photo and a cropped face, as well as the target view cameras and 2D projected spatial anchor indicating head position and orientation. Our diffusion model also takes in noisy target views and a timestep in order to predict the denoised views (top). In practice, we apply a segmentation mask around the person.

providing 4,500 people captured from 48 views in and over 9,000 sets of clothing, outpacing other datasets by an order of magnitude in terms of people diversity. We use internal multi-view data of $\sim$1000 identities with dense ($\sim$160) camera setup, but we expect our model to produce reasonable results with publicly-available captures.

Generating Novel Poses and Expressions. Although different from our task, this related work builds 3D animatable models of people and faces. These methods use neural radiance fields [70, 71, 81], or 3d Gaussians [34, 51, 80, 94], using canonical "T"-poses and a corresponding forward and backward mapping to model local appearance changes. Many of these methods, however, rely on accurate human shape and pose estimation, and only achieve high photorealism for personalized models from studio captures. Relatedly, a family of methods focus on generating animatable faces from a single image by disentangling viewpoints and expressions from large collections of 2d portraits [14, 15, 32, 48, 86, 92], allowing photorealistic realtime reenactment, while being limited to faces and small viewpoint variations. Unlike these methods, we complete missing parts of a person given either one or a few partial views. We do not repose the person or animate the faces but can instead recover entirely missing views, which is orthogonal to these methods. Moreover, our approach addresses both full-body and facial reconstruction, rather than specializing in either domain.

Generations with Explicit 3D Structures. Similar to early methods that extracted 3d representations from 2d models via score distillation [7, 11, 40, 50, 53, 74, 78], 3d representations of human appearance have been extracted from GANs by co-training a neural renderer from learned TriPlane [5] or HexPlane [1, 37] latent spaces. Others train generative models that operate directly in 3D space with diffusion or GAN objectives [18, 44, 72, 76, 82, 83] or regression models to directly predict 3D representations [56, 57, 63]. While these methods produce 3d consistent faces by design, their quality is limited by the relatively small amounts of 3d training data and/or the limitations of their respective 3d representations. In contrast, we focus on generating 3d-consistent 2d images and thus avoid the downsides of explicit 3d modelling.

Generations with 2D Models. Another option is to model viewpoint changes in 2d, without an underlying 3d representation [41], sometimes modelling either photometric [30] or epipolar [28] constraints explicitly. DiffPortrait3D [20] is a portrait-generation method that falls roughly within this category, by fine-tuning a 2d diffusion model for 3d-aware face generation using ControlNet-style [93] conditioning for viewpoint control, as well as cross-view attention [21] and initialization from a 3d-consistent GAN [5]. Other methods are more general and not specific to humans, focusing on single-view to 3D generation [19, 41, 58, 64, 65, 73, 85]. In addition to these single-view methods, video models have been fine-tuned for camera control [23, 79].

3. Method

We train our models following a three-stage strategy:

- **Image-only Pre-training (P1).** We pretrain on a large-scale human-centric dataset with image conditioning.
- **Multiview Mid-training (M2).** We train models at a low-resolution of 128×128 to denoise 48 target views with coarse camera control (no pixel-aligned spatial control).
- **Multiview Post-training (P3).** We train at a high-resolution of 1024×1024 to denoise $1 - 3$ target views with spatial control injected via ControlMLP layers.

We denote the training stage and resolution of any given model as {stage}@{resolution}. For instance, M2@128 represents a mid-trained model at 128 resolution.

3.1. Base Model

Architecture. We adopt a DiT-like [49] architecture with scale, shift, and gate modulation for timestep conditioning inspired by Stable Diffusion 3 [16] and Flux [35]. We simplify the architecture by employing MLP and attention in parallel [91], and removing second LayerNorm after attention layers. We use VAE for training in latent space with 8x spatial compression, and patchify latent images via a linear layer and patch size of 2. We use fixed sinusoidal positional encoding during training. We provide more details in **??**, and the exact design of our DiT block in Fig. 3.

Image-only Pre-training. During pre-training, the model learns to denoise an image conditioned on its corresponding image embedding from DINOv2 [9, 46], this is similar in principle to the image decoder of DALL-E 2 [54]. We project both embeddings to the model dimension using with a linear layer to create a joint conditioning. Importantly, our pretraining setup does not require any annotations or captions for the images, and is well-aligned with our downstream objective of generating consistent multiview images given a single reference image as input.

Formally, given an image $\mathbf{y} \in \mathbb{R}^{H\times W\times C}$, and the joint conditioning as $\mathbf{e}^{\text{img}} \in \mathbb{R}^{N\times D}$. We pre-train our diffusion model ϵ_θ with the following objective:

$$\mathcal{L}_{\text{DM}} = ||\boldsymbol{\epsilon}^t - \epsilon_\theta(\mathbf{y}^t, \mathbf{e}^{\text{img}}, t)||^2 \quad (1)$$

Where $t \in [0, T]$ is diffusion timestep, $\boldsymbol{\epsilon}^t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ is the noise added at the given timestep. We use the DDPM [26, 69] formulation to define discrete timesteps and set $T = 1000$. We first pretrain our model at 256×256 and then at 512×512 on a large corpus of human-centric images. Exact training details can be found in **??**.

3.2. Multiview Model

Our goal is to generate many high resolution and unseen novel viewpoints of a human subject (akin to a studio capture) given a single input image.

Input Reference. We denote the input image as $\mathbf{x}^{\text{ref}}$ and corresponding face crop as $\mathbf{x}^{\text{face}}$. The face crop is obtained using FaceNet [61] and resized to the same size as input image, such that: $\mathbf{x}^{\text{face}}, \mathbf{x}^{\text{ref}} \in \mathbb{R}^{H\times W\times C}$.

Target Cameras. We denote the target viewpoints to be synthesized using distinct cameras (intrinsics and extrinsics), represented as $\mathbf{c}_{1:N}$. Each camera is used to generate its Plücker coordinates $\mathbf{P}_i \in \mathbb{R}^{H\times W\times 6}$.

Target Spatial Anchor. In addition to target cameras, we provide an oriented 3D point denoted as $\mathbf{a}_i = [\mathbf{R}_i|\mathbf{t}_i]$, which roughly defines the center of the subject's head, as well as their gaze direction. We show an example of our Spatial Anchor in Fig. 2 in the studio on the left. This anchor is color-coded and projected into a 2D image, which is used as conditioning for our target view generations. During inference, the Spatial Anchor can be placed at any given point that lies within the field-of-view of the target cameras.

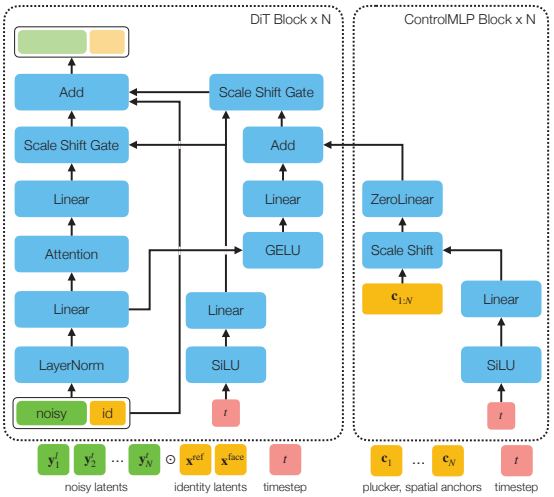

Figure 3. **DiT and ControlMLP Block**. Our DiT block (left) loosely follows [55], with a AdaIn-based timestep modulation. We apply attention and MLP blocks in parallel [10], and jointly apply self-attention to the noisy generated and identity conditioning tokens. ControlMLP block (right) is used to provide lightweight spatially-aligned conditioning - Plücker and Spatial Anchor.

Multiview Diffusion Model. Given the above inputs, we train a multiview diffusion model ϵ_θ to *jointly* denoise all the target views $\mathbf{y}_{1:N} \in \mathbb{R}^{N\times H\times W\times C}$ with the objective:

$$\mathcal{L}_{\text{DM}} = ||\boldsymbol{\epsilon}^t - \epsilon_\theta(\mathbf{y}^t_{1:N}, \mathbf{c}_{1:N}, \mathbf{x}^{\text{ref}}, \mathbf{x}^{\text{face}}, t)||^2 \quad (2)$$

where $\mathbf{y}^t_{1:N}$ are noisy target images. We condition the base model on the provided reference image and its face crop by concatenating their patchified latent tokens with the noisy input latent tokens to the model. This is shown in Fig. 2

Mid-training. In mid-training stage, we want to train a strong multiview model that can denoise several images together, and absorb the dataset quickly at a lower resolution. During this phase, we do not use any pixel-aligned spatial control such as Plücker or Spatial Anchor. We use an MLP to encode the flattened 16-dimensional target camera intrinsics and extrinsics into a single token . We fuse this camera token into each noisy latent token (for the corresponding view) as positional encoding, which makes our multiview model 3D-aware of the target viewpoints. We mid-train our model at 128×128 resolution to jointly denoise 24 views.

Post-training. In the post-training stage, our objective is to create a high-resolution model which is 3D-consistent starting with a low-resolution and a 3D-aware (but not consistent) model. For this, we design a lightweight Control-

Net [93]-inspired module, which takes as input the pixel-aligned Plücker and Spatial Anchor controls, and the denoising timestep to create a separate modulation signal for the multiview model. We name this module ControlMLP, as it uses a single MLP to generate scale-and-shift modulated control for each multiview-DiT block as shown in Fig. 3. Each layer of ControlMLP is zero-initialized at the start. We find that the Post-training phase is crucial in reducing flicker and 3D inconsistencies in generations. We post-train models at 512×512 and 1024×1024 resolutions to jointly denoise 10 and 2 views respectively. Increasing the number of views further lead to GPU out-of-memory issues.

Encoding Plücker and Spatial Anchor. We notice that the relative differences between neighboring pixels in Plücker coordinates are tiny. To amplify these differences better, we use a SIREN [67] layer to first process the 6D grid into a 32D feature grid. Then, we downsample it by 8x to match the size of latent tokens and feed it as input to the ControlMLP. In addition, use the Spatial Anchor to fix the position and orientation of the subject's head in 3D. We only use the Spatial Anchor for generations, and not for the input reference view. We encode the Spatial Anchor image into the latent space of our model via VAE, and concatenate it with Plücker input and pass it through an MLP to create the modulation signal at each layer.

3.3. Understanding and Improving Spatial Control

This section presents our design choices and examines alternative approaches for injecting pixel-aligned spatial controls during Post-training stage. We demonstrate the effectiveness of spatial control through a focused overfitting experiment with quantitative evaluations in Tab. 1.

Scene Overfitting Task. We use 160 frames from a fixed 3D scene of a given subject and timestamp, split into 100 training and 60 validation views. We overfit our mid-trained model to the training views while testing various spatial control methods, training only the control modules while keeping other weights frozen. After overfitting for 10K iterations, we evaluate the model on validation views for novel view synthesis. Strong generalization to validation viewpoints indicates effective spatial control and appropriate camera viewpoint sensitivity. Through this task, we evaluate different spatial control injection methods in Tab. 1, starting with simple to advanced modulation designs.

- **No overfitting (Row 1).** The Mid-trained model without scene-specific overfitting achieves comparable PSNR of 19.2 and 19.7 on train and validation views respectively. We treat this setup as baseline to improve over.
- **Encoding Camera with MLP (Row 2).** We encode camera using an MLP similar to prior works [41, 66] and our Mid-training stage (Sec. 3.2). After overfitting, the model achieves slightly better PSNR on training views as expected, however the validation PSNR drops by 1.28 points to 17.95. This suggests that an MLP does not provide enough modulation for camera control.
- **Plücker as Positional Encoding (Row 3).** In this setup, we use downsampled and patchified Plücker coordinates processed through MLP to create positional encoding, which is added to the noisy latent tokens. This setup is inspired from prior works [3, 6, 23, 31, 68], and it further improves the validation PSNR compared to MLP at 18.89, but lags behind the non-overfitted baseline.
- **Plücker with ControlMLP and SIREN (Row 4, 5).** Here, we use our ControlMLP module to inject spatial control at each multiview-DiT block output. Moreover, encoding Plücker coordinate with a SIREN [67] amplifies the relative differences between neighboring pixels (Sec. 3.2). This setup achieves PSNR of 20.13 with an improvement of 0.9 over baseline.
- **Adding Spatial Anchor (Row 6).** Finally using the Spatial Anchor gives validation PSNR of 22.6 (gain of 3.3 points over baseline) and enables strong spatial control. Thus, we adopt this configuration for Post-training stage.

#	**Method (M2@ 128)**	$\text{PSNR}_{\text{val}} \uparrow$	$\text{PSNR}_{\text{train}} \uparrow$
1.	Mid-trained (No Overfitting)	19.23	19.70
2.	+ Camera (w/ MLP) [41, 66]	$17.95_{-1.28}$	$19.92_{+0.22}$
3.	+ Plücker (w/ MLP) [6, 23, 31, 68]	$18.89_{-0.34}$	$20.74_{+1.04}$
4.	+ ControlMLP	$19.45_{+0.22}$	$29.36_{+9.66}$
5.	+ SIREN	$20.13_{+0.90}$	$30.19_{+10.49}$
6.	+ Spatial Anchor (Ours)	$22.60_{+3.37}$	$30.49_{+10.79}$

Table 1. **Evaluating and Designing Spatial Controls (Sec. 3.3).** We overfit our multi-view model for 10K iterations on 100 views of a single scene from our Head-only dataset, and evaluate on 60 novel views of this scene by only varying the spatial controls (camera and Spatial Anchor). We use this setup to test the modulation strength of different spatial controls. Subscript values green/red show deviations from Row 1.

3.4. Enhanced 3D Consistency Metric

Traditionally, the 3D consistency of multiview generation models is evaluated using 2D image metrics such as PSNR, LPIPS, and SSIM against a fixed set of ground truth images. However, this approach unfairly penalizes models that generate plausible and 3D-consistent novel content deviating from the fixed ground truth images. Some works try to address this by measuring SfM [17, 60] or epipolar error [45, 89], but these methods solve for pose or are not robust since they measure against the entire epipolar line. To address these limiations, we use our GT camera poses as input and compute Reprojection Error (RE) given our known camera poses and predicted correspondences. Our computation of RE involves the following steps:

1. **Landmarks and Correspondence estimation.** We use SuperPoint [13] to detect landmarks in the generated images and employ SuperGlue [59] to establish pairwise correspondences between landmarks across images.

2. **Triangulation.** Given the correspondences and camera parameters, we apply Triangulation based on Direct Linear Transformation (DLT) [22] to obtain the corresponding 3D points for each landmark.
3. **Reprojection and Error calculation.** We reproject these 3D points onto each image and compute the RE as the L2 distance between the original landmark and the reprojected 3D point normalized by image resolution, and average error across all images.

This approach evaluates multiview generation models by focusing on their ability to produce 3D-consistent results rather than adhering to a fixed ground truth. The Reprojection Error provides a valuable basis for comparison across different methods. Furthermore, by computing RE on a set of real-world images that are independent of our generated images, we can establish a baseline that quantifies the error due to the noisy predictions from SuperGlue and SuperPoint rather than the quality of our generated images.

Naming Convention (RE@SG). Please note that the use of SuperPoint [13] and SuperGlue [59] estimation modules is a particular instantiation of our metric, and these could be replaced in the future with stronger counterparts such as MAST3R [36, 75] or domain-specific keypoint detectors such as Sapiens [33]. Thus we use the naming convention of RE@SG to denote the Reprojection Error (RE) under SuperGlue (SG) estimation, which could be modified accordingly in the future for different estimators.

4. Experiments

We present details of the datasets used in all training and validation stages, followed by discussion of evaluation metrics – with particular emphasis on the 3D consistency metric and conclude with core experimental results and ablations.

4.1. Data

Humans-3B Dataset. We utilize a large proprietary dataset of approximately 3 billion human-centric in-the-wild images for pretraining. We provide more details on data filtering and curation in **??**.

Head and Full-body Studio Datasets. We rely on high-quality proprietary studio captures as our primary source of data for learning 3D consistency. Our model comes in two variants: head-only and full-body, each trained (for Mid-training and Post-training stages) on corresponding datasets. For our full-body model, we utilize a dataset of 861 subjects (811 train, 50 test), nearly 1000 frames per subject. For our head-only model, we use a dataset of 1450 subjects (1400 train, 50 test), nearly 40000 frames per subject. The studio setup is similar to [43], with two capture domes for capturing full-body and head-only high-resolution 4K images with 230 and 160 cameras.

iPhone Dataset. To evaluate real-world performance, we collect casual images of 50 test subjects in an indoor office environment using an iPhone 13 Pro. We preprocess these images with Sapiens-2B [33] for background segmentation before model inference. This dataset serves exclusively for evaluating our model's performance on in-the-wild inputs.

4.2. Evaluation Setup and Metrics

3D Consistency. Following prior works, we report standard metrics like PSNR, SSIM ($\times 100$), and LPIPS ($\times 100$) metrics. However, these metrics unfairly penalize plausible novel views generated under incomplete inputs. We therefore introduce the Reprojection Error metric, described in Sec. 3.4, which validates 3D consistency without directly relying on ground truth. Our evaluation generates 4 randomly selected views from the test split.

Identity Preservation. We use two metrics that measure identity preservation across generated views by computing cosine distance between features extracted via FaceNet [4, 61] for face similarity; and CLIP [52] vision encoder for full-body similarity.

Pretrained Model Evaluation. We measure the effectiveness of our pre-training strategy by reporting the FID [24]. We use a smaller annotated 30M subset of Humans-3B for training image and text-conditioned P1@128 models, a 30M subset of unfiltered dataset for training No Filtering P1@128, and a 1000 sample test set from iPhone dataset.

4.3. Results

Pretraining and Data Filtering. Tab. 2 presents our pretrained model results (Row 1) and ablations on Human-centric data filtering and Image-conditioned pretraining (Rows 2-5). Human-centric filtering and image-based conditioning are both critical to achieve high-quality generation. The qualitative results are shown in **??**.

#	Method (Stage @ Resolution)	FID ↓
1.	Pippo (P1@512)	**51.164**
2.	Human-centric Filtering (P1@128)	**75.639**
3.	No Filtering (P1@128)	86.838
4.	Image-conditioned (P1@128)	**75.639**
5.	Text-conditioned (P1@128)	109.720

Table 2. **Pretraining and Data Filtering.** We report results of the full pretrained model P1@512, and compare several variants of P1@128 models. We report FID on iPhone dataset (1k samples).

High-resolution Multi-view Generation. In Tab. 3, we evaluate 3D reconstruction and identity preservation for unseen subjects from the studio datasets. We show that increasing the output resolution of generation in our approach does not hurt 3D consistency or similarity. We put corresponding visuals in **??**, Rows 2 and 3.

	#	Split & Resolution	3D Consistency RE@SG ↓	PSNR ↑	SSIM ↑	LPIPS ↓	Similarity Face ↑	Body ↑
Head-only	1.	Studio (128)	$3.3_{+0.8}$	21.0	67.6	14.8	$62.0_{-13.3}$	$61.5_{-16.2}$
	2.	Studio (1K)	$3.4_{+0.5}$	20.3	72.0	26.2	$73.5_{-2.5}$	$79.4_{-0.1}$
	3.	iPhone Face (1K)	3.0	-	-	-	67.6	-
Full-body	4.	Studio (128)	$3.6_{+0.1}$	22.8	84.0	10.0	$41.7_{-8.5}$	$67.1_{-9.4}$
	5.	Studio (1K)	$1.5_{+0.1}$	22.4	91.7	11.1	$64.7_{-6.0}$	$74.1_{-2.2}$
	6.	iPhone (1K)	1.7	-	-	-	58.0	68.1

Table 3. **Results on Unseen Studio and iPhone data (Sec. 4.3.)** We report results on Post-trained Pippo models at three different resolutions. We report 3D metrics PSNR, SSIM, and LPIPS; as well as our proposed Reprojection Error (RE@SG) under SuperGlue estimation which does not require ground truth (Sec. 3.4). We report 2D Face and Body similarities. Red subscript show deviation against ground truth value.

Generations from Casual iPhone Photos. We present in Tab. 3 (Rows 3,6) Reprojection Error and similarity scores for casually taken images from the iPhone dataset with 1K resolution model. In this scenario, the standard reconstruction error metrics cannot be assessed due to missing ground truth. We find that the reprojection error on iPhone captures remains comparable to the studio dataset – demonstrating 3D consistency. This illustrates the generalizability of Pippo beyond the multi-view training data domain, where our pretraining with large-scale in-the-wild human data is critical. We put corresponding visuals in **??**, Row 1 (Appendix).

Comparisons with External Benchmarks. In Fig. 4, we compare Pippo with individual state-of-the-art baselines in Full-body and Head-only generation. SiTH [25] reconstructs textured human-mesh using ControlNet paired with a SDF representation. Compared to SiTH, Pippo facilitates high-resolution and accurate multiview synthesis. DiffPortrait3D [20] inverts a 3D-GAN based on a given input image. Compared to it, our model supports greater viewpoint variability and ensures closer adherence to the input image.

Quantitative comparisons and baselines. Existing SoTA Human methods [20, 25] use explicit SMPL priors, and thus are difficult to compare with directly. Qualitatively, we found that they cannot handle novel views or preserve details (Fig. 4), and hence do not quantitatively compare against them. In Pippo, we focus on creating a strong multi-view human generator, and we benchmark four state-of-the-art multi-view diffusion models on the iPhone full-body dataset in Tab. 4. We find that Pippo preserves identity (i.e., face and body similarity) and 3D consistency (RE) better while also operating at a higher resolution compared to baselines.

Benchmarking Pippo on public datasets. In Tab. 5, we benchmark Pippo on the public Codec Avatar datasets [43] to aid future comparisons. Specifically, we use Ava-256 containing 256 head-only captures and Goliath containing 4 full-body. During evaluation, we only use subsets of these datasets that were not used for training. We find that Pippo's performance on these datasets is inline with its performance on internal studio datasets.

#	Method & Resolution	RE@SG ↓	Face ↑	Body ↑
1.	MV-Adapter [27] (768)	4.7	43.0	64.1
2.	Era3D [38] (512)	4.1	38.1	64.4
3.	Wonder3D [42] (256)	5.3	34.7	58.8
4.	Pippo (P3@1K)	**3.0**	**58.0**	**68.1**

Table 4. **Quantitative comparison with SoTA multi-view models.** We quantitatively compare Pippo against state-of-the-art multi-view diffusion models. We find that Pippo preserves identity (i.e., face and body similarity) and 3D consistency (RE) better while operating at a higher resolution compared to baselines.

#	Dataset (P3@1K)	RE@SG ↓	PSNR ↑	SSIM ↑	LPIPS ↓	Face ↑	Body ↑
1.	Ava-256 (Head-only)	3.8	20.1	68.7	26.6	72.9	76.8
2.	Goliath (Full-body)	1.2	20.4	89.7	15.5	87.5	77.7

Table 5. **Benchmarking Pippo on public Ava-256 and Goliath datasets.** We find that Pippo's performance on these datasets is inline with its performance on internal studio datasets. This will aid future comparisons against Pippo.

4.4. Ablations

We examine design choices at each training stage (Sec. 3) and present our ablation results in Tab. 6. Experiments are conducted at 128×128 resolution on the Head-only dataset.

#	Method (Stage @ Res.)	3D Consistency RE@SG ↓	PSNR ↑	SSIM ↑	LPIPS ↓	Similarity Face ↑	Body ↑
1.	Pippo (P3@512)	$2.7_{+0.7}$	21.7	71.7	21.5	$74.1_{-2.0}$	$77.4_{-2.4}$
2.	Pippo Face-only (P3@512)	$2.4_{+0.4}$	20.5	70.3	25.3	$74.3_{-1.7}$	$75.8_{-4.0}$
3.	No Mid-train (P3@512)	$5.6_{+3.6}$	14.5	59.5	44.2	$20.4_{-55.6}$	$63.4_{-16.4}$
4.	Pippo (P3@128)	$3.3_{+0.8}$	21.0	67.6	14.8	$62.0_{-13.3}$	$61.5_{-16.2}$
5.	No Anchor (P3@128)	$11.5_{+9.0}$	17.1	54.0	21.1	$63.1_{-12.2}$	$68.3_{-9.3}$
6.	No Plücker (P3@128)	$4.2_{+1.7}$	20.2	64.5	16.1	$63.5_{-11.8}$	$66.1_{-11.5}$
7.	Pippo (M2@128)	$3.4_{+0.9}$	19.1	61.4	17.2	$60.0_{-15.3}$	$62.6_{-15.0}$
8.	No Pre-train (M2@128)	$5.9_{+3.4}$	13.1	39.3	44.9	$19.5_{-55.8}$	$49.0_{-28.6}$
9.	Cross-Attn (M2@128)	$3.5_{+1.0}$	18.0	59.2	22.1	$66.2_{-9.1}$	$70.7_{-7.0}$
10.	Non-frontal (M2@128)	$5.8_{+3.3}$	15.2	52.4	30.1	$49.2_{-26.1}$	$60.7_{-16.9}$

Table 6. **Ablation on design choices and training stages.** We evaluate several multi-view models at 128×128 and 512×512 resolution at different training stages on Head-only dataset. We ablate the choice of doing Mid-training and Pre-training, along with different spatial controls and reference conditioning methods..The training stages are labeled as M2 (Mid-training) and P3 (Post-training), followed by @ specified resolution.

Significance of Pre-training and Mid-training. Pre-training our model on the Humans-3B dataset enables robust generalization to novel identities, as demonstrated in

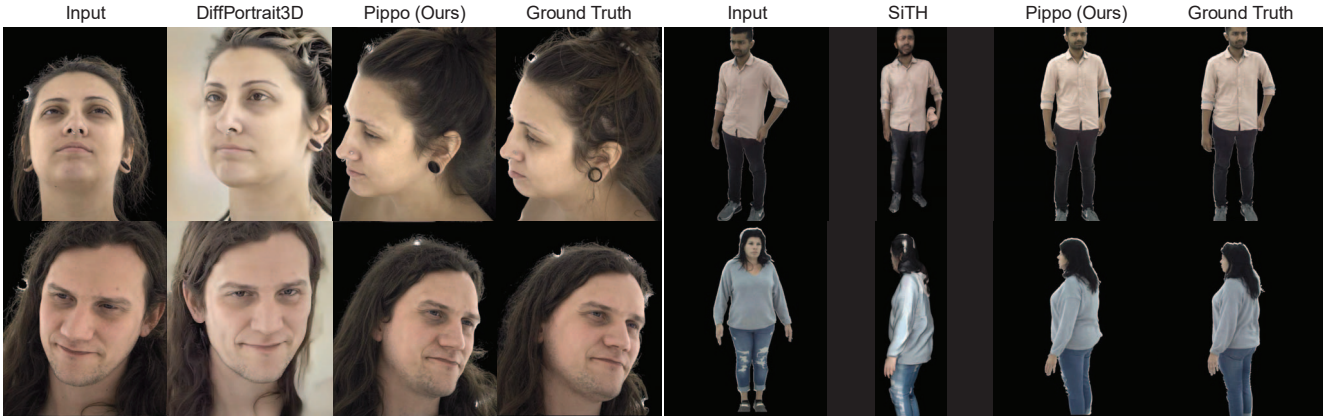

Figure 4. **Visual comparison with state-of-the-art methods**. We visually compare Pippo with state-of-the-art baselines DiffPortrait3D [20] (head-only) and SiTH [25] (full-body) generation.

Tab. 6, Row 8. Without pre-training, model generalization deteriorates, resulting in unclear facial features. Skipping mid-training at lower resolution impairs consistent multi-view generation, as shown in Row 2.

Importance of Frontal Input Reference. Our ablation in Tab. 6, Row 10 demonstrates that completely randomizing the view point of an input reference image leads to overfitting to training identities. Non-frontal views, especially rear views, have very limited information about the identity which forces the network to pick up spurious correlations.

#	Pretrain Data (M2@128)	RE@SG ↓	PSNR ↑	SSIM ↑	LPIPS ↓	Face ↑	Body ↑
1.	30M (1% HQ)	3.8	21.3	81.8	13.3	30.6	66.2
2.	900M (30% Random)	4.0	21.6	82.0	**12.1**	30.6	67.2
3.	2.1B (70% Random)	**3.7**	21.6	82.1	**12.1**	37.6	**67.5**
4.	Humans-3B (100%)	**3.7**	**21.9**	**82.3**	12.7	**41.7**	67.2

Table 7. **Ablating the size of pre-training dataset on Humans-3B.** We conduct Full-body mid-training with pretrained checkpoints trained on 1%, 30%, 70%, and 100% subsets of Humans-3B dataset. We found large-scale data is crucial for generalization to novel identities (i.e., face similarity).

Importance of Self-attention. Replacing self-attention with cross-attention for reference image encoding, using the same routing as image-conditioned pretraining from Sec. 3.1, leads to degraded performance as shown in Tab. 6, Row 9. We observe that this setup causes the model to ignore input conditioning, generating images that only vaguely resemble training subjects.

Role of large scale Humans-3B pre-training dataset. We utilized intermediate checkpoints from pre-training stage (P1) trained on 30% and 70% of the data, and a separate checkpoint trained on 1% high-quality subset of Humans-3B. Starting from these checkpoints, we mid-trained Pippo to denoise 4 views at 128×128 resolution on Full-body dataset for two days. We report their respective results in Tab. 7. We found large-scale data is crucial for generalization to novel identities – indicated by high gains in face similarity metric.

5. Conclusion

We present Pippo, a diffusion transformer model that generates a dense set of high-resolution multi-view consistent images of a person from a single image. Our experiments show that our multi-stage training strategy, which combines large-scale in-the-wild data with high-quality multi-view studio captures, enables generalizable high-resolution multi-view synthesis. Analysis of the diffusion transformer architecture reveals that self-attention with the reference image, Plücker coordinates with SIREN, and Spatial Anchor are all essential for high-fidelity multi-view human generation. Pippo achieves, for the first time, consistent multi-view human image generation at 1K resolution. We also show that our proposed 3D consistency metric enables evaluation of 3D consistency without paired ground-truth data. One of the limitations of our approach is the limited number of simultaneously generated views, originating from large context length and memory constraints, which can be tackled with parallelization techniques and autoregressive generation. Extending our approach to multi-view consistent video generation will be addressed in future work.

References

[1] Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y. Ogras, and Linjie Luo. Panohead: Geometry-aware 3d full-head synthesis in 360deg. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20950–20959, 2023. 3

[2] Timur Bagautdinov, Chenglei Wu, Tomas Simon, Fabián Prada, Takaaki Shiratori, Shih-En Wei, Weipeng Xu, Yaser Sheikh, and Jason Saragih. Driving-signal aware full-body avatars. *ACM Transactions on Graphics (TOG)*, 40(4):1–17, 2021. 2

[3] Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. Vd3d: Taming large video diffusion transformers for 3d camera control. *arXiv preprint arXiv:2407.12781*, 2024. 5

[4] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018)*, pages 67–74, 2018. 6

[5] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, 2022. 3

[6] Eric Ming Chen, Sidhanth Holalkere, Ruyu Yan, Kai Zhang, and Abe Davis. Ray conditioning: Trading photo-consistency for photo-realism in multi-view image generation. *arXiv preprint arXiv:2304.13681*, 2023. 5

[7] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *ICCV*, 2023. 3

[8] Wei Cheng, Ruixiang Chen, Wanqi Yin, Siming Fan, Keyu Chen, Honglin He, Huiwen Luo, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, Daxuan Ren, Lei Yang, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, Bo Dai, and Kwan-Yee Lin. Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. *arXiv preprint*, arXiv:2307.10173, 2023. 2

[9] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2023. 4

[10] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme Ruiz, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd Van Steenkiste, Gamaleldin Fathy Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Collier, Alexey A. Gritsenko, Vighnesh Birodkar, Cristina Nader Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetic, Dustin Tran, Thomas Kipf, Mario Lucic, Xiaohua Zhai, Daniel Keysers, Jeremiah J. Harmsen, and Neil Houlsby. Scaling vision transformers to 22 billion parameters. In *Proceedings of the 40th International Conference on Machine Learning*, pages 7480–7512. PMLR, 2023. 4

[11] Congyue Deng, Chiyu Jiang, Charles R Qi, Xinchen Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. In *CVPR*, 2023. 3

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2

[13] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 5, 6

[14] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. In *ACM MM*, 2022. 3

[15] Nikita Drobyshev, Antoni Bigata Casademunt, Konstantinos Vougioukas, Zoe Landgraf, Stavros Petridis, and Maja Pantic. Emoportraits: Emotion-enhanced multimodal one-shot head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8498–8507, 2024. 3

[16] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 4

[17] Rafail Fridman, Amit Abecasis, Yoni Kasten, and Tali Dekel. Scenescape: Text-driven consistent scene generation. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 5

[18] Matheus Gadelha, Rui Wang, and Subhransu Maji. Multiresolution tree networks for 3d point cloud processing. In *ECCV*, 2018. 3

[19] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024. 2, 3

[20] Yuming Gu, Hongyi Xu, You Xie, Guoxian Song, Yichun Shi, Di Chang, Jing Yang, and Linjie Luo. Diffportrait3d: Controllable diffusion for zero-shot portrait view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10456–10465, 2024. 3, 7, 8

[21] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *International Conference on Learning Representations*, 2024. 3

[22] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004. 6

[23] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 3, 5

[24] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. 2017. 6

[25] I Ho, Jie Song, Otmar Hilliges, et al. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 538–549, 2024. 7, 8

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. 2020. 4

[27] Zehuan Huang, Yuanchen Guo, Haoran Wang, Ran Yi, Lizhuang Ma, Yan-Pei Cao, and Lu Sheng. Mv-adapter: Multi-view consistent image generation made easy. *arXiv preprint arXiv:2412.03632*, 2024. 7

[28] Zehuan Huang, Hao Wen, Junting Dong, Yaohui Wang, Yangguang Li, Xinyuan Chen, Yan-Pei Cao, Ding Liang, Yu Qiao, Bo Dai, et al. Epidiff: Enhancing multi-view synthesis via localized epipolar-constrained diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9784–9794, 2024. 3

[29] Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. Humanrf: High-fidelity neural radiance fields for humans in motion. *ACM Transactions on Graphics (TOG)*, 42(4):1–12, 2023. 2

[30] Yash Kant, Aliaksandr Siarohin, Michael Vasilkovsky, Riza Alp Guler, Jian Ren, Sergey Tulyakov, and Igor Gilitschenski. invs: Repurposing diffusion inpainters for novel view synthesis. In *SIGGRAPH Asia 2023 Conference Papers*, 2023. 3

[31] Yash Kant, Aliaksandr Siarohin, Ziyi Wu, Michael Vasilkovsky, Guocheng Qian, Jian Ren, Riza Alp Guler, Bernard Ghanem, Sergey Tulyakov, and Igor Gilitschenski. Spad: Spatially aware multi-view diffusers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10026–10038, 2024. 5

[32] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 3

[33] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, pages 206–228. Springer, 2025. 6

[34] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023. 3

[35] Black Forest Labs. Flux model, 2024. Accessed: 2024-11-10. 4

[36] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r, 2024. 6

[37] Heyuan Li, Ce Chen, Tianhao Shi, Yuda Qiu, Sizhe An, Guanying Chen, and Xiaoguang Han. Spherehead: Stable 3d full-head synthesis with spherical tri-plane representation. In *European Conference on Computer Vision*, pages 324–341. Springer, 2025. 3

[38] Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wenhan Luo, Ping Tan, Wenping Wang, Qifeng Liu, and Yike Guo. Era3d: High-resolution multiview diffusion using efficient row-wise attention, 2024. 7

[39] Chieh Hubert Lin, Chia-Che Chang, Yu-Sheng Chen, Da-Cheng Juan, Wei Wei, and Hwann-Tzong Chen. Coco-gan: Generation by parts via conditional coordinating. In *ICCV*, 2019. 2

[40] 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. In *CVPR*, 2023. 3

[41] 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 *ICCV*, 2023. 2, 3, 5

[42] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, and Wenping Wang. Wonder3d: Single image to 3d using cross-domain diffusion, 2023. 7

[43] Julieta Martinez, Emily Kim, Javier Romero, Timur Bagautdinov, Shunsuke Saito, Shoou-I Yu, Stuart Anderson, Michael Zollhöfer, Te-Li Wang, Shaojie Bai, Chenghui Li, Shih-En Wei, Rohan Joshi, Wyatt Borsos, Tomas Simon, Jason Saragih, Paul Theodosis, Alexander Greene, Anjani Josyula, Silvio Mano Maeta, Andrew I. Jewett, Simon Venshtain, Christopher Heilman, Yueh-Tung Chen, Sidi Fu, Mohamed Ezzeldin A. Elshaer, Tingfang Du, Longhua Wu, Shen-Chi Chen, Kai Kang, Michael Wu, Youssef Emad, Steven Longay, Ashley Brewer, Hitesh Shah, James Booth, Taylor Koska, Kayla Haidle, Matt Andromalos, Joanna Hsu, Thomas Dauer, Peter Selednik, Tim Godisart, Scott Ardisson, Matthew Cipperly, Ben Humberston, Lon Farr, Bob Hansen, Peihong Guo, Dave Braun, Steven Krenn, He Wen, Lucas Evans, Natalia Fadeeva, Matthew Stewart, Gabriel Schwartz, Divam Gupta, Gyeongsik Moon, Kaiwen Guo, Yuan Dong, Yichen Xu, Takaaki Shiratori, Fabian Prada, Bernardo R. Pires, Bo Peng, Julia Buffalini, Autumn Trimble, Kevyn McPhail, Melissa Schoeller, and Yaser Sheikh. Codec Avatar Studio: Paired Human Captures for Complete, Driveable, and Generalizable Avatars. *NeurIPS Track on Datasets and Benchmarks*, 2024. 2, 6, 7

[44] Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulo, Peter Kontschieder, and Matthias Nießner. Diffrf: Rendering-guided 3d radiance field diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4328–4338, 2023. 3

[45] Norman Müller, Katja Schwarz, Barbara Rössle, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. Multidiff: Consistent novel view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10258–10268, 2024. 2, 5

[46] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4

[47] Dongwei Pan, Long Zhuo, Jingtan Piao, Huiwen Luo, Wei Cheng, Yuxin Wang, Siming Fan, Shengqi Liu, Lei Yang, Bo Dai, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, and Kwan-Yee Lin. RenderMe-360: Large digital asset library and benchmark towards high-fidelity head avatars. In *Thirty-seventh Conference on Neural Information*

Processing Systems Datasets and Benchmarks Track, 2023. 2

[48] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *ICCV*, 2021. 3

[49] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 4

[50] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 3

[51] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5020–5030, 2024. 3

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. 2021. 6

[53] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, et al. Dreambooth3d: Subject-driven text-to-3d generation. *arXiv preprint arXiv:2303.13508*, 2023. 3

[54] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2, 4

[55] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 4

[56] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *CVPR*, 2019. 3

[57] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 84–93, 2020. 3

[58] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single real image. *arXiv preprint arXiv:2310.17994*, 2023. 3

[59] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 2, 5, 6

[60] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 2, 5

[61] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 4, 6

[62] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. 2022. 2

[63] Akash Sengupta, Thiemo Alldieck, Nikos Kolotouros, Enric Corona, Andrei Zanfir, and Cristian Sminchisescu. Diffhuman: Probabilistic photorealistic 3d reconstruction of humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1439–1449, 2024. 3

[64] Junyoung Seo, Kazumi Fukuda, Takashi Shibuya, Takuya Narihira, Naoki Murata, Shoukang Hu, Chieh-Hsin Lai, Seungryong Kim, and Yuki Mitsufuji. Genwarp: Single image to novel views with semantic-preserving generative warping. *arXiv preprint arXiv:2405.17251*, 2024. 3

[65] Qiuhong Shen, Xingyi Yang, and Xinchao Wang. Anything-3d: Towards single-view anything reconstruction in the wild. *arXiv preprint arXiv:2304.10261*, 2023. 3

[66] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 5

[67] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. 2020. 5

[68] Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. 2021. 5

[69] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. 2015. 4

[70] Shih-Yang Su, Timur Bagautdinov, and Helge Rhodin. Danbo: Disentangled articulated neural body representations via graph neural networks. In *European Conference on Computer Vision*, pages 107–124. Springer, 2022. 3

[71] Shih-Yang Su, Timur Bagautdinov, and Helge Rhodin. Npc: Neural point characters from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14795–14805, 2023. 3

[72] Qingyang Tan, Lin Gao, Yu-Kun Lai, and Shihong Xia. Variational autoencoders for deforming 3d mesh models. In *CVPR*, 2018. 3

[73] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. 2023. 3

[74] Christina Tsalicoglou, Fabian Manhardt, Alessio Tonioni, Michael Niemeyer, and Federico Tombari. Textmesh: Generation of realistic 3d meshes from text prompts. 2023. 3

[75] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 6

[76] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *CVPR*, 2023. 3

[77] Wenbo Wang, Hsuan-I Ho, Chen Guo, Boxiang Rong, Artur Grigorev, Jie Song, Juan Jose Zarate, and Otmar Hilliges. 4d-dress: A 4d dataset of real-world human clothing with semantic annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[78] 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. 2023. 3

[79] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3

[80] Jing Wen, Xiaoming Zhao, Zhongzheng Ren, Alexander G Schwing, and Shenlong Wang. Gomavatar: Efficient animatable human modeling from monocular video using gaussians-on-mesh. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2059–2069, 2024. 3

[81] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *CVPR*, 2022. 3

[82] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. 2016. 3

[83] Jianwen Xie, Zilong Zheng, Ruiqi Gao, Wenguan Wang, Song-Chun Zhu, and Ying Nian Wu. Learning descriptor networks for 3d shape synthesis and analysis. In *CVPR*, 2018. 3

[84] Zhangyang Xiong, Chenghong Li, Kenkun Liu, Hongjie Liao, Jianqiao Hu, Junyi Zhu, Shuliang Ning, Lingteng Qiu, Chongjie Wang, Shijie Wang, et al. Mvhumannet: A large-scale dataset of multi-view daily dressing human captures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19801–19811, 2024. 2

[85] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360deg views. In *CVPR*, 2023. 3

[86] Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. Vasa-1: Lifelike audio-driven talking faces generated in real time. *arXiv preprint arXiv:2404.10667*, 2024. 3

[87] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. FaceScape: A large-scale high quality 3D face dataset and detailed riggable 3D face prediction. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[88] T Yenamandra, A Tewari, F Bernard, HP Seidel, M Elgharib, D Cremers, and C Theobalt. i3DMM: Deep implicit 3D morphable model of human heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2

[89] Jason J Yu, Fereshteh Forghani, Konstantinos G Derpanis, and Marcus A Brubaker. Long-term photometric consistent novel view synthesis with diffusion models. In *ICCV*, 2023. 2, 5

[90] Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. HUMBI: A large multiview dataset of human body expressions. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[91] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113, 2022. 4

[92] Bowen Zhang, Chenyang Qi, Pan Zhang, Bo Zhang, Hsiang-Tao Wu, Dong Chen, Qifeng Chen, Yong Wang, and Fang Wen. Metaportrait: Identity-preserving talking head generation with fast personalized adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22096–22105, 2023. 3

[93] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 3, 5

[94] Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. Drivable 3d gaussian avatars. *arXiv preprint arXiv:2311.08581*, 2023. 3