HanDiffuser: Text-to-Image Generation With Realistic Hand Appearances

Supreeth Narasimhaswamy*[1], Uttaran Bhattacharya[2], Xiang Chen[2],
Ishita Dasgupta[2], Saayan Mitra[2], and Minh Hoai[1]

[1]Stony Brook University, USA, [2]Adobe Research, USA

Abstract

Text-to-image generative models can generate high-quality humans, but realism is lost when generating hands. Common artifacts include irregular hand poses, shapes, incorrect numbers of fingers, and physically implausible finger orientations. To generate images with realistic hands, we propose a novel diffusion-based architecture called HanDiffuser that achieves realism by injecting hand embeddings in the generative process. HanDiffuser consists of two components: a Text-to-Hand-Params diffusion model to generate SMPL-Body and MANO-Hand parameters from input text prompts, and a Text-Guided Hand-Params-to-Image diffusion model to synthesize images by conditioning on the prompts and hand parameters generated by the previous component. We incorporate multiple aspects of hand representation, including 3D shapes and joint-level finger positions, orientations and articulations, for robust learning and reliable performance during inference. We conduct extensive quantitative and qualitative experiments and perform user studies to demonstrate the efficacy of our method in generating images with high-quality hands. Project page: https://supreethn.github.io/research/handiffuser/index.html

1. Introduction

Text-to-Image (T2I) generative models have shown impressive advancement in recent years. Generative models such as Stable Diffusion [56], Imagen [58], and GLIDE [45] can generate high quality, photorealistic images. However, these methods often struggle to synthesize high-quality and realistic hands. The generated hands often have improbable hand poses, irregular hand shapes, incorrect number of fingers, and poor hand-object interactions (Fig. 1).

Generating images with high-quality hands is a challenging problem since hands often take up a small part of the image, but are highly articulate. They have high degrees of freedom, with a wide variety of flexibility where fingers can bend to various degrees relatively independently. Hands can also occur in various shapes, sizes, and orientations and can be occluded with other human body parts. Further, hands often interact with objects and can have a wide range of grasps depending on the object's size, shape, and affordance. Therefore, capturing such a vast range of articulations and interactions directly from text inputs remains challenging. Despite having billions of parameters and several millions of trainable images, T2I models struggle to generate realistic hands.

A central challenge in hand image generation is learning diverse hand poses and configurations at scale. Existing hand representations based on keypoint skeletons and shape formats [34, 57] are useful for generative tasks in pose animation [66] and hand-object interactions [12]. These representations provide a grounded understanding of plausible hand shapes and postures, especially in relation to the rest of the body and different interacting objects. However, the necessary steps to incorporate these hand representations into T2I pipelines, in terms of both learning these representations from text prompts and mapping these representations into the pixel space of images, remain open problems. These problems are exacerbated when we consider naturally constructed prompts, which often imply rather than specify hand postures and articulations (*e.g.*, all prompts in Fig. 1). Prompt engineering [16, 33], focusing on hand descriptions, can potentially improve the generation quality. But it comes with the cost of distilling and learning appropriate prompts from large-scale data, and with the caveats of learning spurious inter-relationships between the prompt and the hands or between the hands and the rest of the images.

In this paper, we propose a learning-based model to generate images containing realistic hands in an end-to-end fashion from text prompts. Our model, called HanDiffuser, consists of two key trainable components. The first component, Text-to-Hand-Params (T2H), generates parameters of a hand model [34, 57] conditioned on the input text prompts. The second component, Text-Guided Hand-Params-to-Image (T-H2I), uses the hand parameters and the input text prompts as conditions to generate images. By

*Work started when Supreeth was an intern at Adobe Research

Figure 1. **Generating realistic hands.** Text-to-Image generative models, *e.g.*, [56], often produce various hand artifacts *(top row)*. We inject hand embeddings, capturing hand shapes, poses, and articulations, in the generation process to generate realistic hands *(bottom row)*.

conditioning the image generation on accurate hand models, HanDiffuser can generate high-quality hands with plausible hand poses, shapes, and finger articulations. Specifically, we consider three aspects of hand representation, each serving a unique purpose. These include the spatial locations of hand joints to capture the hand pose, the joint rotations to capture the finger orientations and articulations, and the hand vertices to capture the overall hand shape. We design a novel Text+Hand Encoder by extending the CLIP encoder [54] to obtain joint embeddings for these three representations together with the text. We use the proposed joint embeddings to condition the image generation, allowing us to generate images by conditioning on both the hand parameters and the text.

We train the two components of HanDiffuser independently. We train T2H using around 450K text and 3D human pairs and fine-tune T-H2I using around 900K text and image pairs. Once trained, we use the two components end-to-end in a single inference pipeline to generate images from text prompts. We conduct extensive experiments and user studies to show the effectiveness of the HanDiffuser in generating images with high-quality hands.

In short, the contributions of our paper are:

- **HanDiffuser**, a generative model to synthesize images with high-quality hands by conditioning on text and hand embeddings. It has two novel components: Text-to-Hand-Params and Text-Guided Hand-Params-to-Image.
- **Text-to-Hand-Params**, a diffusion model to generate SMPL-Body and MANO-Hand parameters from text inputs. The generated MANO-Hands are used to further condition the image generation.
- **Text-Guided Hand-Params-to-Image**, a diffusion model to generate images with high-quality hands by conditioning on hand and text embeddings. We design hand embeddings to capture hand shape, pose, and finger orientations and articulations.

2. Related Work

We briefly summarize related work on text-to-image generation, concurrent work on text-to-human generation, and commonly used hand representations.

Text-to-Image Generation. Text-guided image generation is a well-studied problem, with modern approaches ranging from GANs [65, 68, 70, 78], autoregressive generation approaches [37, 55], and VQ-VAE transformers [13] to state-of-the-art diffusion models [11, 21, 22, 61]. Text-to-image generation using diffusion models often bootstrap the generative pipeline with pre-trained language models, such as BERT [10] or CLIP [54], to efficiently learn from the text information [4, 17, 38, 45, 73]. More recently, Stable Diffusion [56] performs diffusion in the latent image space to generate high-quality images at low computational costs. Imagen [58], by contrast, diffuses the pixels directly in a hierarchical fashion. ControlNet [74] provides additional controllability in the image generation process in the form of conditioning signals ranging from sketches to pose priors. Latest software products for text-to-image generation include Midjourney [39], DALL-E 3 [46], and Firefly [1]. While the advancements in this area have been rapid and significant, generating highly articulate hands remains prone to unrealistic artifacts.

Text-to-Human Generation. Alongside image generation, there has also been considerable progress in human pose and motion generation from text prompts. Recent generative methods typically follow various skeletal joint formats, such as OpenPose [5], or combined joint and mesh formats, such as SMPL [34], to represent the human body. They train on various large-scale pose and motion datasets, including KIT [51], AMASS [36], BA-

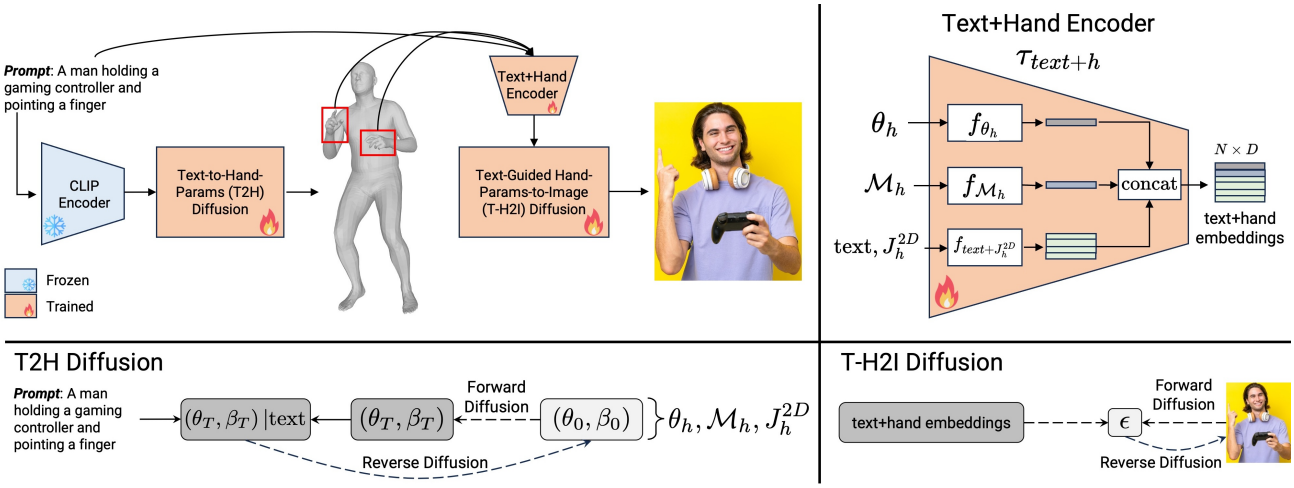


Figure 2. **HanDiffuser architecture.** Our architecture consists of two components. The first component, Text-to-Hand-Params (T2H), takes the text as input and generates body and hand parameters. The second component, Text-Guided Hand-Params-to-Image (T-H2I), uses the hand parameters from the first component and the text to generate images with high-quality hands. The Text+Hand encoder jointly encodes hand parameters and text, and captures hand pose, articulation, and shape.

BEL [53], and HumanML3D [18]. To efficiently map text prompts to motion sequences, text-to-motion generation methods learn combined representations of language and pose using techniques ranging from recurrent neural networks [2, 30], hierarchical pose embeddings [3, 14], and VQ-VAE transformers [49, 50] to motion diffusion models [7, 28, 66, 75]. Some approaches even generate 3D meshes on top of pose motion synthesis to synthesize fully rendered humans [23, 71]. Separate from motion synthesis, there are works on generating parametric pose models from text [8, 9, 47]. However, these methods focus on the human body and ignore the hand regions. As a result, they cannot generate articulate hands. There is a method [43] to generate plausible hands using ControlNet but it requires a hand skeleton or mesh as an additional input. A concurrent work [35] proposes an inpainting approach to refine hands. Given a generated image, the method reconstructs 3D meshes for hands and further refines hand regions. As a result, the quality of the reconstructed hand mesh, and consequently the final refined hand, depends on the quality of the initially generated hand. On the contrary, our method first generates hand mesh parameters from the prompt and further conditions the image generation on such intermediate hand parameters. Moreover, [35] ignores the hand-object interactions in the initial image and might not preserve hand-object occlusions and interactions when refining hands.

Hand Representations. Available datasets on hand configurations, gestures, and hand-object interactions offer hand representations in a variety of formats, including bounding boxes, silhouettes, depth maps [25, 29, 31, 40–42, 44, 60], keypoints and parametric models [12, 26, 27, 57]. These representations are useful for multiple hand-centric tasks, including detection [67], gesture and pose recognition [76], motion generation [64, 77], and hand-object interactions [15, 24, 63]. Our work combines representations based on keypoint and parametric models to efficiently encode diverse hand shapes and highly articulated finger movements.

3. HanDiffuser

Fig. 2 illustrates the proposed HanDiffuser architecture. Given a text input, HanDiffuser first uses a novel Text-to-Hand-Params diffusion model to generate the parameters of the human body and hand models. The second component is the Text-Guided Hand-Params-to-Image diffusion model that generates the output image by conditioning on the hand model and the text. This section provides more detailed insights into the Text-to-Hand-Params and Text-Guided Hand-Params-to-Image models, following a brief introduction to the fundamentals of human models and stable diffusion.

3.1. Preliminaries

SMPL-H. Our Text-to-Hand-Params model generates parameters of human body and hand models from text inputs. We use SMPL [34] and MANO [57] as our body and hand model, respectively. The SMPL is a differentiable function $\mathcal{M}_b(\theta_b, \beta_b)$ that takes a pose parameter $\theta_b \in \mathbb{R}^{69}$ and shape parameter $\beta_b \in \mathbb{R}^{10}$, and returns the body mesh $\mathcal{M}_b \in \mathbb{R}^{6890\times3}$ with 6890 vertices. Similarly, MANO is a differentiable function $\mathcal{M}_h(\theta_h, \beta_h, s)$ that takes the hand pose parameter $\theta_h \in \mathbb{R}^{48}$, hand shape parameter $\beta_h \in \mathbb{R}^{10}$, and the hand side $s \in \{\text{left}, \text{right}\}$, and returns hand mesh $\mathcal{M}_h \in \mathbb{R}^{778\times3}$ with 778 vertices. The 3D hand joint locations $J_h \in \mathbb{R}^{k\times3} = \mathcal{W}_h\mathcal{M}_h$ can be regressed from ver-

tices using a pre-trained linear regressor $\mathcal{W}_h$. The SMPL-H model combines the body, left hand, and right hand model into a single differntiable function $\mathcal{M}(\theta, \beta)$ with pose parameters θ=$(\theta_b, \theta_{lh}, \theta_{rh})$ and shape parameters β. The pose parameters θ_b, θ_{lh}, and θ_{rh} captures the root-relative joint rotations for body, left hand, and right hand, respectively. The shape parameter β captures the scale of the person.

Stable Diffusion. Our Text-Guided Hand-Params-to-Image model is built upon Stable Diffusion [56]. Stable Diffusion is a latent diffusion model consisting of an auto-encoder, a U-Net for noise estimation, and a CLIP text encoder. The encoder $\mathcal{E}$ encodes an image x into a latent representation $z = \mathcal{E}(x)$ that the diffusion process operates on. The decoder $\mathcal{D}$ reconstructs the image from $\hat{x} = \mathcal{D}(z)$ from the latent z. The U-Net is conditioned on the denoising step t and the text $\tau_{text}(text)$, where $\tau_{text}(text)$ is a CLIP [54] text encoder that projects a sequence of tokenized texts into an embedding space. To jointly condition the image generation on hand parameters and the text, we replace the text encoder $\tau_{text}(text)$ with a novel Text+Hand encoder $\tau_{text+h}(text, hand)$ that jointly embeds the text and hand parameters into a common embedding space.

3.2. Text-to-Hand-Params Diffusion

The Text-to-Hand-Params diffusion model takes a text as input and generates the pose parameters $\theta = (\theta_b, \theta_{lh}, \theta_{rh})$ and shape parameters β for the SMPL-H model by conditioning on the text.

We define $x := (\theta, \beta)$ and model the forward diffusion process by iteratively adding Gaussian noise to x for T time steps:

$$q\left(x_t|x_{t-1}\right) = \mathcal{N}\left(\sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I\right), \qquad (1)$$

where $\alpha_t \in (0, 1)$ are constant hyper-parameters.

We model the text-conditioned SMPL-H generation distribution $p(x_0|c)$ as the reverse diffusion process of gradually denoising x_T. Following [66], we learn the denoising by directly predicting $\hat{x}_0 = G(x_t, t, c)$ using a model G. We train the reverse diffusion using the training objective:

$$\mathcal{L}_1 = \mathbb{E}_{x_0 \sim q(x_0|c), t\sim[1,T]}||x_0 - G(x_t, t, c)||_2^2. \qquad (2)$$

We get the conditional text embeddings c by encoding the text using CLIP [54]. We implement G using a transformer encoder-only architecture similar to MDM [66].

Given a text during inference, we conditionally sample $x = (\theta, \beta)$. We use the shape and pose parameters to obtain the joints J_{lh}, J_{rh} and vertices $\mathcal{M}_{lh}, \mathcal{M}_{rh}$ for left and right hands using MANO-Hand model. We also choose camera parameters and project J_{lh}, J_{rh} into an image space and obtain the corresponding image-space joint locations J_{lh}^{2D}, J_{rh}^{2D}. We use the joint rotations θ_{lh}, θ_{rh}, hand vertices $\mathcal{M}_{lh}, \mathcal{M}_{rh}$, and spatial joint locations J_{lh}^{2D}, J_{rh}^{2D} to condition the image generation in the next stage.

3.3. Text-Guided Hand-Params-to-Image Diffusion

The Text-Guided Hand-Params-to-Image diffusion model is built upon Stable Diffusion [56] and conditions the image generation on hand parameters generated from the Text-to-Hand-Params model and the text. Specifically, Text-Guided Hand-Params-to-Image uses a novel Text+Hand Encoder τ_{text+h} to first obtain joint embeddings for text and hand parameters. It then uses the joint hand and text embeddings to condition the image generation. We provide more details on this below.

Text+Hand Encoder. Given the provided text, along with the spatial joint locations J_h^{2D}, vertices $\mathcal{M}_h$, and joint rotations θ_h of the hand, our goal is to generate D-dimensional embeddings to encode both the text and hand parameters. Here D denotes the CLIP [54] token embedding dimension. To encode hand joint locations in the image space, we follow [6, 69] and introduce additional positional tokens. We quantize the image height and width uniformly into N_{bins} bins. This allows us to approximate and tokenize any normalized spatial coordinate into one of N_{bins} tokens. We then encode the text tokens and the hand joint spatial tokens into a D-dimensions using $f_{text+J_h^{2D}}$. Specifically, we construct $f_{text+J_h^{2D}}$ by introducing an additional $N_{\text{bins}} \times D$ embedding layer into an existing CLIP token embedder and fine-tune it during training. To encode hand vertices, we transform them to basis point set (BPS) [52] representations and pass through $f_{\mathcal{M}_h}$, a Multi-Layer Perceptron (MLP) consisting of fully-connected linear and ReLU layers. Similarly, we encode 6D hand joint rotations θ_h using an MLP f_{θ_h} consisting of fully-connected linear and ReLU layers. Finally, we concatenate embeddings from text, spatial hand joints, hand vertices, and hand joint rotations to produce joint text and hand embeddings.

Diffusion. We instantiate the Text-Guided Hand-Params-to-Image using Stable Diffusion [56] and train using the following objective

$$\mathcal{L}_2 = \mathbb{E}_{\mathcal{E}(x), \epsilon\sim\mathcal{N}(0,1), t, y}||\epsilon - F\left(z_t, t, \tau_{text+h}(y)\right)||_2^2. \quad (3)$$

In the above equation, the condition $y = (\text{text}, J_h^{2D}, \mathcal{M}_h, \theta_h)$ denotes the combination of the text and the hand parameters, which include spatial joint locations, vertices, and joint rotations. The function F is a denoising U-Net to predict the noise, τ_{text+h} is the trainable Text+Hand encoder. We refer the readers to [56] for more details regarding Eq. (3).

Generating SMPL-H vs Skeletons in Text-to-Hand-Params Diffusion. We design the first component of HanDiffuser to generate pose and shape parameters of SMPL-H instead of keypoints or skeletons since SMPL-H encodes topological and geometric priors about humans and encodes richer information than skeletons. Also, SMPL-H parameters tend to be more robust to noise than skeletons; we

Method	FID ↓	KID ↓	FID-H ↓	KID-H ↓	Hand Conf. ↑
Stable Diffusion	29.005	9.63×10^{-3}	34.372	4.63×10^{-2}	0.887
Stable Diffusion Fine-tuned	20.056	7.91×10^{-3}	31.219	3.09×10^{-2}	0.913
ControlNet	18.694	5.93×10^{-3}	28.091	2.19×10^{-2}	0.969
HanDiffuser w/o 2D hand joints	16.839	5.21×10^{-3}	29.902	2.46×10^{-2}	0.953
HanDiffuser w/o 3D joint rotation and vertices	14.586	4.14×10^{-3}	28.186	2.21×10^{-2}	0.961
HanDiffuser (proposed)	**13.918**	$\mathbf{4.07\times10^{-3}}$	**27.550**	$\mathbf{2.11\times10^{-2}}$	**0.978**

Table 1. **Quantitative results.** We report the scores of current baselines and ablated versions of our method on multiple evaluation metrics. ↑ indicates higher values are better, ↓ indicates lower values are better.

Figure 3. **Qualitative results.** We compare the quality of hands in images generated by different methods from the same text prompts. (Images are generated at 512x512 resolution)

can still get plausible poses even with noisy SMPL-H parameters, whereas noisy joint locations lead to implausible poses. Since we are generating *parameters* (51 joint rotations and 10 shape parameters) of SMPL-H mesh, the Text-to-Hand-Params component is computationally lighter compared to the second component, Text-Guided Hand-Params-to-Image .

4. Experiments

This section describes the datasets used to train HanDiffuser , implementation details, and evaluation metrics used. We also present qualitative and quantitative results, and user studies to show the efficacy of HanDiffuser in generating images with high-quality hands.

4.1. Datasets

We train the two components of HanDiffuser using our own curated datasets. We start with 930K paired text and images, then curate it to remove inappropriate and harmful content and validate the quality of the images through independent content creators. We randomly split the dataset to obtain 900K train and 30K test text-image pairs. We further preprocess the dataset to obtain SMPL-H parameters. Specifically, we use [62] to obtain SMPL parameters for the body and [57] to obtain MANO parameters for hands. We reject estimated SMPL body and MANO hands that have low confidence scores. Finally, we curate two datasets using the estimated hand and body parameters. The first dataset consists of tuples of the form (text, SMPL-H). We keep such tuples only for images where we can reliably estimate SMPL-H parameters. This dataset has 450K tuples (text, SMPL-H) and is used to train the first component of the HanDiffuser, Text-to-Hand-Params. The second dataset consists of triplets of the form (text, image, SMPL-H), and we keep all 930K triplets. We use this dataset to train the second component of HanDiffuser, Text-Guided Hand-Params-to-Image. During training, we only conditioned the image generation on hand parameters when the SMPL-H parameters were reliably estimated.

4.2. Implementation Details

We train the Text-to-Hand-Params diffusion to generate the SMPL-H pose θ and shape β by conditioning on the text. We encode the text using a frozen CLIP-ViT-B/32 model [54]. We train the Text-to-Hand-Params model using a classifier-free guidance [20] by randomly setting 10% of the text conditions to be empty. We train this model for 100 epochs on a single A100 GPU using a batch size of 64. We use 1000 steps and a guidance scale of $s = 2.5$ during the inference. We fine-tune Text-Guided Hand-Params-to-Image starting from the Stable Diffusion v1.4 checkpoint. To implement the Text+Hand encoder τ_{text+h}, we start with the CLIP ViT-L/14 model and introduce additional $N_{bins} = 1000$ positional tokens for spatial hand joints. We choose simple three-layer MLPs $f_{\mathcal{M}_h}$ and f_{θ_h}

to encode hand vertices and joint rotations, respectively. We fine-tune Text-Guided Hand-Params-to-Image, including the Text+Hand encoder τ_{text+h}, for 20 epochs on eight A100 GPUs using a batch size of 8 and AdamW optimizer with a constant learning rate of 10^{-4}. We perform inference with 50 PLMS [32] steps using a classifier-free guidance [20] of 4.0.

HanDiffuser Inference. Given a text input, we first sample SMPL-H parameters using our trained Text-to-Hand-Params model. We then extract the MANO hand parameters from SMPL-H and choose camera parameters randomly with some constraints to make hands somewhat visible in the image and obtain spatial hand joint joints. Finally, we use these spatial hand joints, MANO parameters, and the text to conditionally sample an image from our trained Text-Guided Hand-Params-to-Image model.

4.3. Evaluation Metrics

We access the quality of generated images from HanDiffuser using the Frechet Inception Distance (FID) and Kernel Inception Distance (KID) [19, 48]. Since FID and KID measures the overall quality of the image, we also compute FID-H and KID-H to measure the quality of images only in the hand regions. We perform this by first extracting crops using hand bounding boxes and then computing FID and KID using such hand crops. We also measure the quality of hands using average hand detection confidence scores. Specifically, we run an off-the-shelf hand detector [72] on generated images and compute the confidence scores for detection. Higher confidence scores mean that the hand detector is more confident of a region being a hand, indicating higher-quality hand generations.

4.4. Quantitative Results and Ablation Studies

We compare the proposed HanDiffuser with three different methods and report these results in Table 1. First, we use an off-the-shelf Stable Diffusion [56] model pre-trained on the LAION-5B [59] dataset. The LAION-5B dataset is a general-purpose text and image pairs dataset and does not necessarily focus on humans. Therefore, a Stable Diffusion model trained on the LAION-5B dataset does not perform well on images solely focused on humans. Second, we fine-tuned Stable Diffusion on our dataset and observed that it generated better images than the pre-trained model. While this generates better performance than the pre-trained model, the performance is still low compared to the proposed HanDiffuser. Third, we experiment with ControlNet [74], a popular latent diffusion model that uses spatial control images to condition the image generation process. We train a ControlNet architecture on our dataset using hand-pose skeleton images as controls. However, unlike HanDiffuser, which generates images directly from text input, ControlNet requires an additional hand pose skeleton image as input during inference. To address this, we directly use the *ground-truth* skeleton images from the test data as control images. Despite employing these *ground-truth* control images, ControlNet does not perform as well as HanDiffuser.

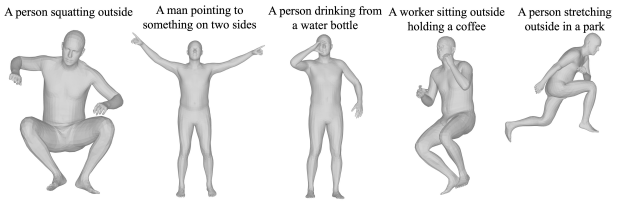


Figure 4. **Illustrative SMPL-H results**, generated from our Text-to-Hand-Params model.

It is important to note that the reported FID-H, KID-H, and hand confidence scores for Stable Diffusion and Stable Diffusion Fine-tuned in Table 1 are optimistic performance measures. To evaluate the performance of these two methods, we first ran a hand detector [72] to obtain hand crops. However, this approach is biased towards rejecting bad hand generations since the hand detector cannot localize low-quality hand generations, leaving out unrealistic-looking generated hands from evaluation. On the contrary, the corresponding metrics for ControlNet and HanDiffuser in Table 1 are the true performance measures since both the methods generate images conditioned on hands, allowing us to crop every generated hand precisely.

We also study the benefits of different hand representations that are used to condition the hand generation in HanDiffuser. First, we evaluate HanDiffuser by omitting the spatial hand joint locations J_h^{2D} in hand embeddings. Second, we evaluate HanDiffuser by omitting the hand joint rotations θ_h and hand vertices $\mathcal{M}_h$ in hand embeddings. We report these results in the fourth and fifth row of Table 1. These results show that all three hand representations help in generating quality hands.

4.5. Qualitative Results and Failure Cases

We report some good qualitative results in Fig. 3. We compare results from Stable Diffusion, ControlNet [74], the proposed HanDiffuser without hand joint rotations and vertices embeddings, and HanDiffuser. Stable Diffusion does not generate realistic hands even after fine-tuning on human-centric datasets. It generates an incorrect number of hand fingers, poor hand-object interactions, implausible finger orientations, and hand shapes. ControlNet generates better-looking results but requires hand skeleton control images as additional input. We can see that HanDiffuser generates hands with plausible hand poses by conditioning the image generation on spatial hand joint locations. Further conditioning on hand joint rotations and vertices enables HanDiffuser to generate high-quality, detailed hands with plausible

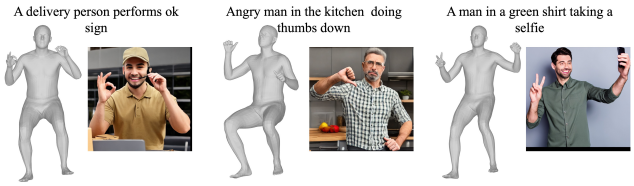


Figure 5. **Generating images from text via SMPL-H.** The intermediate SMPL-H representations are essential in generating realistic hand appearances.

Figure 6. **HanDiffuser failure cases.** We note some failure cases when the given action may be unclear (*e.g.*, "adjusts" in the first image from left), the model does not exactly follow the hand pose description (second image from left), the model does not fully realize the finger dexterity when handling small and thin objects (third image from left), and model does not strictly obey the intended affordance of the object (fourth image from left).

orientations and shapes.

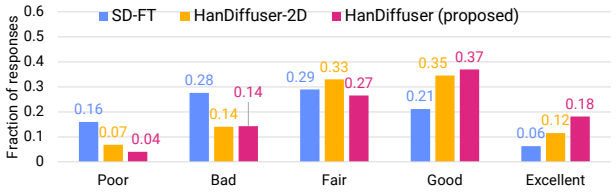
Fig. 4 shows a few SMPL-H results generated from text inputs using our Text-to-Hand-Params. While we only use hand parameters from these SMPL-H outputs, our Text-to-Hand-Params can be directly used in other applications that require generating SMPL-H models from text inputs. We also show how Text-Guided Hand-Params-to-Image maps these SMPL-H results to generated images in Fig. 5. Fig. 6 shows some failure cases of HanDiffuser.

4.6. User Studies

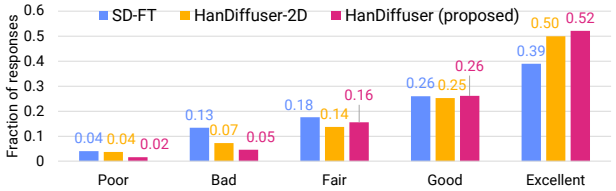
We evaluate the quality of both our generated images and the intermediate outputs of our approach through two different user studies. We evaluate the generated images in two aspects. The first is *(A) plausibility*, which considers how natural the hands look, for example, in terms of hand shapes, finger orientations, number of fingers, hands, and how clearly the hands are in focus in the image. The second is *(B) relevance*, which considers how natural the hand poses or gestures appear given the prompt, for example, holding objects or gesticulating conventionally (unless otherwise specified in the prompt).

We evaluate the intermediate SMPL-H outputs for generating the images in three aspects: *(A) plausibility* of the pose, *(B) relevance* to prompt, and *(C) consistency* with the generated image.

Setup. We compare three methods in the user study to evaluate the generated images: fine-tuned Stable Diffusion (SD-FT), HanDiffuser trained with only 2D hand



(a) Results on the plausibility of generated images



(b) Results on the relevance of generated images to given prompts

Figure 7. **User study results for generated images.** We report the mean fraction of responses for each point on the Likert scale.

Image Aspect	SD-FT	HanDiffuser-2D	HanDiffuser (proposed)
Plausibility ↑	2.74 ± 0.08	3.30 ± 0.11	$\mathbf{3.51 \pm 0.11}$
Relevance ↑	3.83 ± 0.12	4.11 ± 0.17	$\mathbf{4.23 \pm 0.18}$

Table 2. **User study for generated images score summary.** We compute the mean and standard deviation of the Likert-scale scores of the three evaluated methods across all the 700 responses of 35 participants. Higher scores are better.

joints (HanDiffuser-2D), and HanDiffuser trained with all its model components. We show participants 20 sets of images. Each set consists of a unique prompt randomly selected from the test partition of LAION-5B [59] and the images generated by the three methods given that prompt. We arrange the three images within each set in a random order not known to the participants. For each image in each set, we ask them to respond to two questions: *"How is the visual quality of the hands?"* (image plausibility) and *"How well do the hands follow the prompt?"* (image relevance). We collect responses to the two questions on a 5-point Likert scale, consisting of the following choices: *"Poor (e.g., too many or severe mistakes)"*, *"Bad (e.g., some aspects reasonable but still many or severe mistakes)"*, *"Fair (e.g., some aspects are plausible but some mistakes visible)"*, *"Good (e.g., most aspects are plausible but a few mistakes visible)"*, and *"Excellent (e.g., everything looks good, no visible mistakes)"*. Note that we perform the user study on methods that only require text prompts to generate image outputs at test time. Therefore, we exclude methods such as ControlNet [74], which additionally requires pose information to generate similar images. Moreover, the overall performance of ControlNet is also at the same level as HanDiffuser-2D (Table 1 rows 3 and 4) even if we manually provide the ground-truth poses, leading to no meaningful differences between their responses in a pilot study.

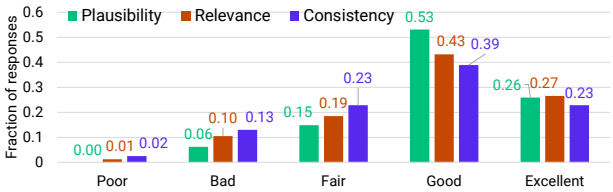


Figure 8. **User study results for SMPL-H poses.** We report the mean fraction of responses for each point on the Likert scale.

For the user study to evaluate the intermediate SMPL-H outputs, we show participants 9 random triplets of (prompt, SMPL-H poses, generated image), where the prompts are randomly selected from the test partition of LAION-5B [59] and the SMPL-H poses and generated images come from our approach. For each triplet, we ask the participants to respond to three questions: *"How plausible is the pose?"* (SMPL-H plausibility), *"How relevant are the hands in the pose given the prompt?"* (SMPL-H relevance), and *"How consistent are the hands in the pose with the hands in the image?"* (SMPL-H consistency). We collect responses to the three questions on the same 5-point Likert scale as above. ask them to evaluate the poses on 5-point Likert scales for each of the three aspects of plausibility, relevance, and consistency. To evaluate consistency, we additionally ask participants to focus primarily on the hand configurations and gestures described or implied in the text. We ask them to ignore distractors, such as the quality of any facial expressions (or lack thereof), any component of the 3D pose that is not visible in the image, and the differences in orientations of the body between the pose and the image.

Results. Our user study to evaluate the generated images was completed by 35 participants, resulting in a total of 700 responses over the 20 image sets. We did not observe any notable response differences across genders and age groups. We report the distribution of scores for the two aspects across all the responses in Fig. 7. We also summarize the mean and the standard deviation of the scores of the two aspects for each of the three methods in Table 2. To compute these values, we assign numbers 1 through 5 to the response choices *Poor* through *Excellent*. Consequently, higher scores indicate better performance. HanDiffuser outperforms the other methods in both aspects for generated images. Looking at the distribution of image plausibility scores (Fig. 7a), we observe the mode of SD-FT on *"Fair"*, while the modes of both HanDiffuser versions are a point higher, on *"Good"*. Overall, 55% of HanDiffuser scores are *"Good"* or better, compared to 47% of HanDiffuser-2D scores and 27% of SD-FT scores. Looking at the distribution of image relevance scores (Fig. 7b), we observe the modes of all the three methods on *"Excellent"*, indicating their efficacy in generating hand appearances aligned with text prompts. Among the three methods, we note a relatively higher distribution of good responses for HandDiffuser variants. Specifically, 78% of HanDiffuser scores are *"Good"* or better, compared to 75% of HanDiffuser-2D scores and 65% of SD-FT scores. Looking at the mean scores across all the 700 responses (Table 2), we note marked improvements for HanDiffuser. Its image plausibility scores are 0.77 points (or 15% on the 5-point scale) higher than SD-FT and 0.21 points (or 4%) higher than HanDiffuser-2D. Correspondingly, its image relevance scores are 0.40 points (or 8%) higher than SD-FT and 0.12 points (or 2%) higher than HanDiffuser-2D.

Plausibility ↑	Relevance ↑	Consistency ↑
4.00 ± 0.81	3.83 ± 0.98	3.67 ± 1.04

Table 3. **User study for SMPL-H score summary of HanDiffuser.** We compute the mean and standard deviation of the Likert-scale scores for the three evaluation aspects across all 162 responses of 18 participants. Higher scores are better.

Our user study to evaluate the intermediate SMPL-H poses was completed by 18 participants, resulting in a total of 171 responses over the 9 triplets. We did not observe any notable response differences across genders and age groups. We report the distribution of scores for the two aspects across all the responses in Fig. 8. We also summarize the mean and the standard deviation of the scores of the two aspects for each of the three methods in Table 3. To compute these values, we assign numbers 1 through 5 to the response choices *Poor* through *Excellent*. Consequently, higher scores indicate better performance.

5. Conclusions and Limitations

We have presented HanDiffuser, an end-to-end model to generate images with realistic hand appearances from text prompts. Our model explicitly learns hand embeddings based on hand shapes, poses and finger-level articulations, and combines them with text embeddings to generate images with high-quality hands. We demonstrate the state-of-the-art performance of our method on the benchmark T2I dataset both quantitatively, through multiple evaluation metrics, and qualitatively, through a user study.

In the future, we plan to extend our model to more unexplored territories of hand generation. These include images consisting of multiple people, complex hand-object interactions, prompts describing highly specialized hand activities (*e.g.*, origami), the same person handling multiple objects simultaneously, hand-hand interactions of two or more people, and non-anthropomorphic hands (*e.g.*, a dog a using a computer). A concurrent future direction is to make the hand generation pipeline style- and shape-aware, such that it can consistently generate the same hands when asked to generate the same person in different images.

Acknowledgements. This project was partially supported by US National Science Foundation Award NSDF DUE-2055406.

References

[1] Adobe. *Firefly, https://www.adobe.com/sensei/generative-ai/firefly.html*, 2023. 2

[2] Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. Text2action: Generative adversarial synthesis from language to action. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5915–5920, 2018. 3

[3] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*, pages 719–728, 2019. 3

[4] Alembics. *Disco-Diffusion, https://github.com/alembics/disco-diffusion*, 2023. 2

[5] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2

[6] Ting Chen, Saurabh Saxena, Lala Li, David J. Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. In *International Conference on Learning Representations (ICLR)*, 2022. 4

[7] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9760–9770, 2023. 3

[8] Delmas, Ginger and Weinzaepfel, Philippe and Lucas, Thomas and Moreno-Noguer, Francesc and Rogez, Grégory. PoseScript: 3D Human Poses from Natural Language. In *ECCV*, 2022. 3

[9] Delmas, Ginger and Weinzaepfel, Philippe and Moreno-Noguer, Francesc and Rogez, Grégory. PoseFix: Correcting 3D Human Poses with Natural Language. In *ICCV*, 2023. 3

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. 2

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, pages 8780–8794. Curran Associates, Inc., 2021. 2

[12] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12943–12954, 2023. 1, 3

[13] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *Computer Vision – ECCV 2022*, pages 89–106, Cham, 2022. Springer Nature Switzerland. 2

[14] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1396–1406, 2021. 3

[15] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Imos: Intent-driven full-body motion synthesis for human-object interactions. *Computer Graphics Forum*, 42(2):1–12, 2023. 3

[16] Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*, 2023. 1

[17] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10696–10706, 2022. 2

[18] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161, 2022. 3

[19] 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 6

[20] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 5, 6

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851. Curran Associates, Inc., 2020. 2

[22] Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23(1), 2022. 2

[23] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *ACM Transactions on Graphics (TOG)*, 41(4):1–19, 2022. 3

[24] Hezhen Hu, Weilun Wang, Wengang Zhou, and Houqiang Li. Hand-object interaction image generation. In *Advances in Neural Information Processing Systems*, pages 23805–23817. Curran Associates, Inc., 2022. 3

[25] Mingzhen Huang, Supreeth Narasimhaswamy, Saif Vazir, Haibin Ling, and Minh Hoai. Forward propagation, backward regression and pose association for hand tracking in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[26] Juntao Jian, Xiuping Liu, Manyi Li, Ruizhen Hu, and Jian Liu. Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose. In *Proceedings of*

the IEEE/CVF International Conference on Computer Vision (ICCV), pages 14713–14724, 2023. 3

[27] Alexander Kapitanov, Andrey Makhlyarchuk, and Karina Kvanchiani. Hagrid - hand gesture recognition image dataset. *arXiv preprint arXiv:2206.08219*, 2022. 3

[28] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis and editing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7): 8255–8263, 2023. 3

[29] Oscar Koller, Jens Forster, and Hermann Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125, 2015. 3

[30] Angela S Lin, Lemeng Wu, Rodolfo Corona, Kevin Tai, Qixing Huang, and Raymond J Mooney. Generating animated videos of human activities from natural language descriptions. *Proceedings of the Visually Grounded Interaction and Language Workshop at NeurIPS*, 2018(1), 2018. 3

[31] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2684–2701, 2020. 3

[32] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In *International Conference on Learning Representations (ICLR)*, 2022. 6

[33] Vivian Liu and Lydia B Chilton. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2022. Association for Computing Machinery. 1

[34] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. In *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 2015. 1, 2, 3

[35] Wenquan Lu, Yufei Xu, Jing Zhang, Chaoyue Wang, and Dacheng Tao. Handrefiner: Refining malformed hands in generated images by diffusion-based conditional inpainting, 2023. 3

[36] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[37] Elman Mansimov, Emilio Parisotto, Jimmy Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. In *International Conference on Learning Representations (ICLR)*, 2016. 2

[38] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 2

[39] Midjourney. *https://www.midjourney.com*, 2023. 2

[40] Supreeth Narasimhaswamy, Zhengwei Wei, Yang Wang, Justin Zhang, and Minh Hoai. Contextual attention for hand detection in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 3

[41] Supreeth Narasimhaswamy, Trung Nguyen, and Minh Hoai. Detecting hands and recognizing physical contact in the wild. In *Advances in Neural Information Processing Systems*, 2020.

[42] Supreeth Narasimhaswamy, Thanh Nguyen, Mingzhen Huang, and Minh Hoai. Whose hands are these? hand detection and hand-body association in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[43] Supreeth Narasimhaswamy, Uttaran Bhattacharya, Xiang Chen, Ishita Dasgupta, and Saayan Mitra. Text-to-hand-image generation using pose- and mesh-guided diffusion. In *IEEE/CVF International Conference on Computer Vision (ICCV), International Workshop on Observing and Understanding Hands in Action*, 2023. 3

[44] Supreeth Narasimhaswamy, Huy Nguyen, Lihan Huang, and Minh Hoai. Hoist-former: Hand-held objects identification, segmentation, and tracking in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[45] 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. In *International Conference on Machine Learning(ICML)*, 2022. 1, 2

[46] OpenAI. *Dall-E 3, https://openai.com/dall-e-3*, 2023. 2

[47] Boris N Oreshkin, Florent Bocquelet, Felix G Harvey, Bay Raitt, and Dominic Laflamme. Protores: Proto-residual network for pose authoring via learned inverse kinematics. In *The Tenth International Conference on Learning Representations (ICLR)*, 2022. 3

[48] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, 2022. 6

[49] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10985–10995, 2021. 3

[50] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[51] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. 2

[52] Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 4

[53] Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. Babel: Bodies, action and behavior with english la-

bels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 722–731, 2021. 3

[54] 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. In *International Conference on Machine Learning (ICML)*, 2021. 2, 4, 5

[55] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2

[56] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 1, 2, 4, 6

[57] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 2017. 1, 3, 5

[58] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems(NeurIPS)*, 2022. 1, 2

[59] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 6, 7, 8

[60] Roy Shilkrot, Supreeth Narasimhaswamy, Saif Vazir, and Minh Hoai. WorkingHands: A hand-tool assembly dataset for image segmentation and activity mining. In *Proceedings of British Machine Vision Conference*, 2019. 3

[61] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2256–2265, Lille, France, 2015. PMLR. 2

[62] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. Monocular, One-stage, Regression of Multiple 3D People. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 5

[63] Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. Goal: Generating 4d whole-body motion for hand-object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13263–13273, 2022. 3

[64] Omid Taheri, Yi Zhou, Dimitrios Tzionas, Yang Zhou, Duygu Ceylan, Soren Pirk, and Michael J Black. Grip: Generating interaction poses using latent consistency and spatial cues. *arXiv preprint arXiv:2308.11617*, 2023. 3

[65] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16515–16525, 2022. 2

[66] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 1, 3, 4

[67] Ultralytics. *YOLOv8, https://github.com/ultralytics/ultralytics*, 2023. 3

[68] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[69] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14246–14255, 2023. 4

[70] Hui Ye, Xiulong Yang, Martin Takac, Rajshekhar Sunderraman, and Shihao Ji. Improving text-to-image synthesis using contrastive learning. *The 32nd British Machine Vision Conference (BMVC)*, 2021. 2

[71] Kim Youwang, Kim Ji-Yeon, and Tae-Hyun Oh. Clip-actor: Text-driven recommendation and stylization for animating human meshes. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[72] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*, 2020. 6

[73] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 833–842, 2021. 2

[74] 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 3836–3847, 2023. 2, 6, 7

[75] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 3

[76] Jiangbin Zheng, Yile Wang, Cheng Tan, Siyuan Li, Ge Wang, Jun Xia, Yidong Chen, and Stan Z. Li. Cvt-slr: Contrastive visual-textual transformation for sign language recognition with variational alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23141–23150, 2023. 3

[77] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Toch: Spatio-temporal object-to-hand correspondence for motion refinement. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[78] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2