Text-to-Hand-Image Generation Using Pose- and Mesh-Guided Diffusion

Supreeth Narasimhaswamy*, Uttaran Bhattacharya, Xiang Chen, Ishita Dasgupta, and Saayan Mitra
1Stony Brook University, USA, 2Adobe Research, USA

Abstract

Text-to-image generative models like Stable Diffusion can generate high-quality humans, but realism is lost when generating hands. The generated hands often have artifacts such as irregular hand poses, shapes, incorrect numbers of fingers, and physically implausible finger orientations. To overcome these issues, we propose to generate high-quality hands by injecting hand structure into a Stable Diffusion model using a ControlNet architecture. Specifically, our method uses rigid 2D hand poses and deformable 3D hand meshes to control the quality of generated hands. Our experiments show that using hand structure information during training enables the generation of realistic hand images. We also extend our method to generate plausible hands as part of full-body images.

1. Introduction

There has been a recent explosion of text-to-image generative methods [2, 6, 8, 11] for realistic and high-quality images. While there has been significant progress toward generating images with good composition, existing methods struggle to generate realistic human hands. For example, Fig. 1 shows some images obtained from Stable Diffusion [8], where the generated hands often have irregular hand shapes, incorrect numbers of fingers, physically implausible finger orientations, and poor hand-object interactions. And these issues are by no means unique to Stable Diffusion. Even current methods focused on generating human images [13] fail to reliably generate plausible hands. Controlled hand generation is more challenging than the rest of the human body because of some unique properties of hands that are difficult to learn from data alone:

• wide variety of flexibility, e.g., fingers can bend to different degrees in different hands;
• extremely articulate [1], e.g., combining finger transformations with palm deformations into myriad shapes;
• diverse configurations and interactions [3, 5] e.g., using different hand shapes even when interacting with the same object, such as holding a pen. Hand interactions become even more complicated due to different intended affordances, such as holding a pen at different angles (resulting in different mutual occlusions) and with different grasps for writing vs. passing.

To enable generative models to consistently synthesize realistic hands, we inject information on hand structure and hand configurations into the generative process. Specifically, we propose an end-to-end stable-diffusion-based architecture that leverages hand priors, as both rigid 2D hand poses and deformable 3D hand meshes, to guide hand image generation from text inputs. Our architecture follows the ControlNet paradigm [15] and treats hand priors as control signals for the generative process. Through quantitative evaluations and qualitative comparisons, we show that hand control signals help us generate plausible and diverse hand configurations. We also extend the hand image generation to a whole-body generation method that can generate high-quality hands together with other parts of the body by using 2D human poses, 2D hand poses, and 3D hand meshes as guidance.

2. Method

We base our method on the recent ControlNet paradigm [15]. ControlNet is a diffusion-based text-to-image generation method that uses spatial control signals to condition the generative process. We have developed two methods, one for generating hand-crop images and another for generating hands together with full-body images.

2.1. Hand Images with Pose & Mesh Guidance

We illustrate two variants to control hand generations in Fig. 2a and Fig. 2b. In the first method, given a hand crop image during training, we extract the 2D hand pose...
Hand Generation Pipelines. We use 2D hand poses (for rigid finger movements) and 3D hand meshes (for both rigid finger and deformable palm movements) as control signals.

Full-Body Generation Pipelines. We combine 2D hand poses and 3D hand meshes with full-body poses to get the control signals.

by estimating the 21 joint locations $h \in \mathbb{R}^{21 \times 2}$ using Mediapipe [12]. We use the hand pose image as a control signal to train our designed diffusion model. We provide a 2D hand pose image as an input control signal during the inference and sample the RGB hand image conditioned on the input pose image. While 2D hand poses provide structural information on the fingers and the overall shapes of the hands, they fail to capture the deformable palm configurations and how those deformations interplay with different finger transformations. To this end, we design a second method: given a hand crop image during training, we estimate the pose $\theta \in \mathbb{R}^{16 \times 3}$ and shape $\beta \in \mathbb{R}^{10}$ parameters of a 3D MANO hand mesh [9] and also the weak-perspective camera model $\Pi_h = (t_x, t_y, s_h)$ using FrankMocap [10]. We use the estimated MANO and camera parameters to render the hand as a control signal to train our designed diffusion model. We provide a rendered hand mesh as input during the inference and sample the RGB hand image conditioned on such input.

2.2. Full-Body Images with Pose & Mesh Guidance

We extend our hand-crop image generation methods to generate hands together with full-body images. We illustrate these methods in Fig. 3a and Fig. 3b. In the first method, given a ground-truth text-image pair during training, we first extract 2D human poses using YOLO-v8 [7] and 2D hand poses using MediaPipe [12]. We then use these body and hand poses to create a control image. We use the control image and text embeddings as conditions to train our designed diffusion model. Given a text, 2D body pose, and 2D hand poses, we sample a full-body image during the inference using the poses as a control image and text embeddings as conditions. Similar to our previous discussion, we also design a second method to enable learning from deformable hand shapes. Given a ground-truth text-image during training, we first extract 2D human poses using YOLO-v8 [7] and estimate 3D hand mesh using FrankMocap [10]. We then create a control image by rendering the 3D hand mesh and drawing in the 2D body pose. We finally use the control image and text embeddings as conditions to train our diffusion model. Given a text, a 2D body pose, and a rendered 3D hand image, we sample a full-body image during the inference using the pose and rendered mesh as the control image and text embeddings as conditions.

3. Experiments

We describe the data we use to train our method and the implementation details. We also highlight our quantitative and qualitative results.

3.1. Datasets and Implementation Details

We use an in-house stock dataset containing paired text instructions and images to train our methods. We curate our dataset to remove inappropriate and harmful content and validate the quality of the images through independent content creators. The dataset contains around 330K images
Using 2D Hand Pose as Guidance. Given an input 2D hand pose, we generate a hand corresponding to the hand pose.

Using 3D Hand Mesh as Guidance. Given an input 3D hand mesh, we generate a hand corresponding to the hand mesh.

Figure 4: Hand Generation Results. We show results using both 2D and pose and 3D hand mesh guidance. We generate hand images at size 256 × 256.

Figure 5: Hand Generation Comparisons. Using deformable 3D hand meshes, we can generate finer detailed hands with better finger shapes, articulations, and hand-object interactions compared to 2D hand poses. We generate hand images at size 256 × 256.

containing at least one person. We randomly split the dataset into 300K training and 30K test images.

We implement our method by incorporating our estimated 2D and 3D hand control signals into the publicly available ControlNet code [14]. We train our methods on eight A100 GPUs for 200 epochs. We resize the hand crops to 256 × 256 for training ControlNet for hand-crop images and a batch size of 128. For training ControlNet for full-body images, we resized the images to 512 × 512 and a batch size 16. We train all models with a learning rate of 1e − 5.

3.2. Quantitative Evaluations

We evaluate the quality of the generated hands using the standard evaluation metrics of Fréchet Inception Distance (FID) and Kernel Inception Distance (KID) [4]. Since FID and KID computed on the entire image do not explicitly focus on the hand generation quality, we follow [13] and compute FID and KID by comparing only the hand regions in the real and the generated images. We also evaluate the quality of the generated hands by computing the average hand detection confidence returned by a state-of-the-art hand detector [12]. The confidence score is higher if the detector is more confident about a hand. We report these results in Table 1. The results show that we can progressively improve the hand generation quality using progressively improving the sophistication of the hand control signals from rigid 2D hand poses to deformable 3D hand meshes.

Table 1: Hand Generation Evaluation. We show results using various types of hand control signals. Using the 3D hand mesh control produces more plausibility in the images compared to using 2D hand poses.

<table>
<thead>
<tr>
<th>Control Signal</th>
<th>FID↓</th>
<th>KID↓</th>
<th>Hand Det. Conf.↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Hands (Body Pose Only)</td>
<td>61.770</td>
<td>0.021</td>
<td>0.891</td>
</tr>
<tr>
<td>Body Pose + 2D Hand Pose</td>
<td>50.511</td>
<td>0.015</td>
<td>0.927</td>
</tr>
<tr>
<td>Body Pose + 3D Hand Mesh</td>
<td>49.324</td>
<td>0.014</td>
<td>0.934</td>
</tr>
</tbody>
</table>

3.3. Qualitative Comparisons

We show some qualitative results for generating hands using 2D hand pose and 3D hand mesh guidance in Fig. 4a and Fig. 4b, respectively. We also compare how 3D hand mesh guidance can generate more accurate hands than 2D hand pose guidance in Fig. 5. We generate hands at a resolution 256 × 256. Fig. 6 shows some qualitative results for text-to-full-body generation. We generate full-bodies at 512 × 512 resolution. We can see that using more accurate hand control signals such as 2D hand pose and 3D hand mesh as guidance improves the hand generation quality.

4. Conclusion

In this paper, we have developed a technique to generate plausible hands in text-to-image models by injecting hand priors into the generation process using a ControlNet architecture. We have developed the techniques to generate images of hands only as well as hands together with the rest of the body for various hand articulations and hand-object interactions. We have performed quantitative evaluations and qualitative comparisons and observed that hand priors can help generate plausible hand images. We also understand the potential risks of a generative method capable of creating realistic images. To this end, we have controlled the quality of the text descriptions and images in our training set through expert human annotators.
References


