AttentionHand: Text-driven Controllable Hand Image Generation for 3D Hand Reconstruction in the Wild

Junho Park[1,2*], Kyeongbo Kong[3*], and Suk-Ju Kang[1✉]

[1] Department of Electronic Engineering, Sogang University, South Korea
[2] AI Lab, CTO Division, LG Electronics, South Korea
[3] Department of Electrical & Electronics Engineering, Pusan National University, South Korea

junho18.park@gmail.com kbkong@pusan.ac.kr sjkang@sogang.ac.kr

https://github.com/redorangeyellowy/AttentionHand

Abstract. Recently, there has been a significant amount of research conducted on 3D hand reconstruction to use various forms of human-computer interaction. However, 3D hand reconstruction in the wild is challenging due to extreme lack of in-the-wild 3D hand datasets. Especially, when hands are in complex pose such as interacting hands, the problems like appearance similarity, self-handed occclusion and depth ambiguity make it more difficult. To overcome these issues, we propose AttentionHand, a novel method for text-driven controllable hand image generation. Since AttentionHand can generate various and numerous in-the-wild hand images well-aligned with 3D hand label, we can acquire a new 3D hand dataset, and can relieve the domain gap between indoor and outdoor scenes. Our method needs easy-to-use four modalities (i.e, an RGB image, a hand mesh image from 3D label, a bounding box, and a text prompt). These modalities are embedded into the latent space by the encoding phase. Then, through the text attention stage, hand-related tokens from the given text prompt are attended to highlight hand-related regions of the latent embedding. After the highlighted embedding is fed to the visual attention stage, hand-related regions in the embedding are attended by conditioning global and local hand mesh images with the diffusion-based pipeline. In the decoding phase, the final feature is decoded to new hand images, which are well-aligned with the given hand mesh image and text prompt. As a result, AttentionHand achieved state-of-the-art among text-to-hand image generation models, and the performance of 3D hand mesh reconstruction was improved by additionally training with hand images generated by AttentionHand.

Keywords: 3D Hand Mesh Reconstruction · Text-to-Image Generation

[*] Equal contribution.
[✉] Corresponding author.

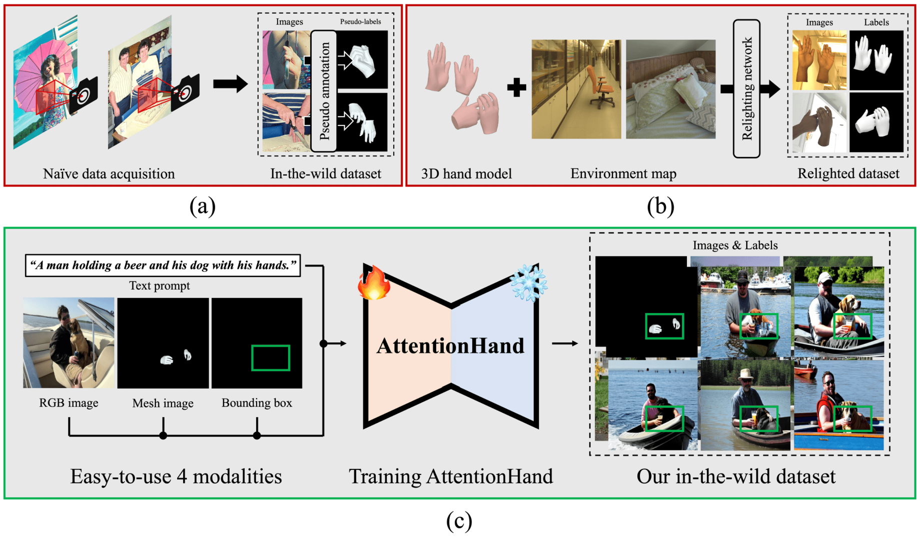

Fig. 1: Various acquisition types of 3D hand datasets. (a) In-the-wild dataset (i.e., MSCOCO [6]) is naively acquired with inaccurate pseudo annotation, (b) relighted dataset (i.e., Re:InterHand [5]) consists of unnatural hands with inharmonious background, and (c) our in-the-wild dataset from AttentionHand, which is annotated with accurate 3D labels, contains natural hands with harmonious background, easy to generate, and can be made infinitely.

1 Introduction

The goal of 3D hand mesh reconstruction is to recover the 3D hand mesh from a single RGB image. It becomes difficult when hands are in the wild, due to insufficiency of in-the-wild 3D hand datasets. Compared to in-the-lab datasets [1–3], acquisition in-the-wild datasets is challenging due to unpredictable conditions such as weather, lighting, cost of sensors, and safety issues on crowded roads and public places. Even if an in-the-wild dataset is collected, data diversity would be poor due to the aforementioned severe constraints. Although arbitrary labels can be obtained through pseudo annotation, the precision and accuracy is still poor compared to in-the-lab datasets as shown in Fig. 1(a). To tackle this problem, several synthetic datasets [4, 5] have introduced. However, since the hand and background images are synthesized out of harmony, they consist of unnatural and unrealistic hand images as shown in Fig. 1(b). Hence, it is difficult to overcome the domain gap between indoor and outdoor scenes with synthetic datasets.

Moreover, when hands are in a complex pose like interacting hands, it becomes even more challenging to reconstruct 3D hand meshes due to the appearance similarity, self-handed occclusion and depth ambiguity. Starting with InterHand2.6M [7], several works [8–15] have emerged to solve the complex hand

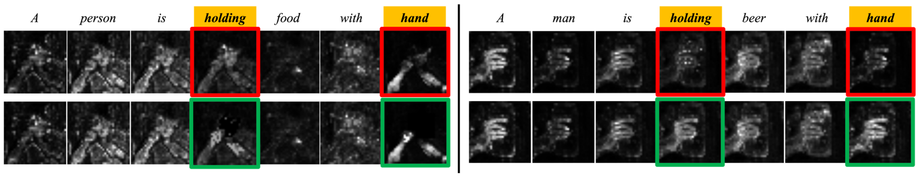

Fig. 2: Visualization of attention maps with corresponding tokens from given text prompts. Red and green boxes represent attention maps without and with AttentionHand, respectively.

pose. However, they have been employed and evaluated primarily on in-the-lab scenes except for InterWild [16]; it tried to relieve the domain gap by leveraging the geometric features of the hand, which is not affected by the domain. Nevertheless, since InterWild was trained with MSCOCO [6], which is extremely lack of in-the-wild hand images with inaccurate 3D labels, there is earnestly need of more diverse, numerous, and accurately annotated in-the-wild datasets.

To address aforementioned issues, we propose AttentionHand, a new method for the text-driven controllable hand image generation. AttentionHand is designed based on Stable Diffusion (SD) [17] to create accurate, natural, realistic and harmonious in-the-wild hand images easily and infinitely as shown in Fig. 1(c). AttentionHand has a huge advantage: we can simply generate images with only four modalities – an RGB image, the corresponding hand mesh image, bounding box, and text prompt. Therefore, we can generate (1) various in-the-wild hand images with flexible text prompts, and (2) well-aligned hand images with 3D hand label. By generating new samples with AttentionHand, we can alleviate aforementioned issues of the 3D hand mesh reconstruction in the wild.

To train AttentionHand, we need to additionally prepare a local RGB image and local mesh image for attention of hand-focused region of the image. The preparation of local information is essential because hands commonly occupy relatively small region in the image. Hence, we obtain local RGB and mesh images by cropping and resizing original RGB and mesh images (i.e., we define them as global information) with the bounding box of hand region. After encoding prepared information in the encoding phase, encoded latent embeddings are fed to the conditioning phase, which is composed of the text attention stage (TAS) and visual attention stage (VAS).

TAS attends on hand-related tokens from the given text prompt by leveraging attention maps as shown in Fig. 2. Specifically, TAS extracts hand-related attention maps (i.e., *holding* and *hand*), and these attention maps are updated to highlight hand-related regions by the refinement based on the softmax operation and Gaussian filter. With TAS, we can obtain more hand-focused images than before. On the other hand, VAS attends on hand-related regions by conditioning global and local hand mesh images with the SD-based pipeline. With global and local information, AttentionHand can be jointly optimized to reflect the global context (i.e., in-the-wild background) and local context (i.e., hand-focused foreground.) In the end of the conditioning phase, we finally get the diffusion feature, which is decoded to new hand images in the decoding phase. Hence, AttentionHand can generate well-aligned hand images with the given mesh image and text prompt for the 3D hand mesh reconstruction in the wild.

To prove the excellence of AttentionHand, we conducted extensive experiments for the text-to-hand image generation and 3D hand mesh reconstruction. As a result, AttentionHand achieved state-of-the-art in the text-to-hand image generation, and the performance of 3D hand mesh reconstruction was considerably improved by additionally training with new hand samples generated by AttentionHand. Especially, the performance was enhanced significantly on in-the-wild datasets, which implies AttentionHand can generate various and well-annotated in-the-wild hand images. The summary of our contributions is as follows:

- We propose a novel method, AttentionHand, which generates well-aligned in-the-wild hand images in a simple manner without laborious data acquisition.
- AttentionHand is designed based on a generative model that attends on hand-related tokens from the text prompt and hand-related regions from the hand mesh image, for generating hand-focused images.
- AttentionHand achieved state-of-the-art in the text-to-hand image generation, and we verified that utilizing the dataset generated from AttentionHand improves the performance on 3D hand mesh reconstruction in the wild.

2 Related Work

2.1 Text-to-Image Generation

Text-to-image generation aims to synthesize high-resolution image from natural language descriptions. With the advent of diffusion models, various studies on text-to-image generation have been conducted in recent years [17–21]. Specifically, ControlNet [18] and T2I-Adapter [19] proposed novel approaches to incorporate arbitrary condition into the generation process. Recently, Uni-ControlNet [20] presented a novel approach that allows for the simultaneous utilization of various conditions in a flexible and composable manner. Nevertheless, aforementioned models exhibited common limitations in generating hand images, due to the relatively small size of hands within the overall image resolution.

2.2 Generative Models for Hand

GANs for Hand. There are several works [22–25] to tackle the hand image generation problem with the generative adversarial network (GAN) [26]. Specifically, a novel network for image-to-image translation [22] was proposed to make generated images follow the same statistical distribution as real-world hand images. GestureGAN [23] was designed to translate hand gesture-to-gesture with the explicit hand skeleton information through the color loss and the cycle-consistency loss. Moreover, the first model-aware gesture-to-gesture translation framework [24] was introduced with hand prior as the intermediate representation. Recently, a new method [25], which employs the expressive model-aware hand-object representation and leverages its inherent topology to build the unified surface space, was proposed. However, these works have a common limitation; they are confined to target gestures in generating new hand images. In other words, they are inappropriate to generate in-the-wild images focused on various hands.

Diffusion Models for Hand. Recently, some works [27–29] have been addressed hand-related problems with diffusion models. DiffHand [27] introduced the first diffusion-based framework that approaches hand mesh reconstruction as a denoising diffusion process. HandDiffuse [28] proposed a strong baseline for the controllable motion generation of interacting hands using various controllers by designing a diffusion-based model. HandRefiner [29] presented an inpainting pipeline to rectify malformed human hands in generated images with diffusion-based models. However, since these models are not text-driven methods, they cannot generate various in-the-wild hand images conditioned on language instructions.

3 Method

We introduce AttentionHand, a novel framework for creating various and plausible hand images. AttentionHand is a SD-based framework that can generate new RGB images infinitely conditioned on hand mesh images and text prompts. The overall pipeline is shown in Fig. 3.

3.1 Data Preparation Phase

As shown in the first box of Fig. 3, it just requires four inputs to train AttentionHand: (1) a global RGB hand image $I^G_{RGB} \in \mathbb{R}^{3\times512\times512}$, (2) a global hand mesh image $I^G_{mesh} \in \mathbb{R}^{3\times512\times512}$, (3) a bounding box of the hand region $B \in \mathbb{R}^{1\times4}$, and (4) a hand-related text prompt U. However, since hands typically occupy small areas on in-the-wild scenes, we also obtain a local RGB hand image $I^L_{RGB} \in \mathbb{R}^{3\times512\times512}$ and a local hand mesh image $I^L_{mesh} \in \mathbb{R}^{3\times512\times512}$ by cropping and resizing I^G_{RGB} and I^G_{mesh} with B. This combination of local and global information enhances hand image conditioning. Details will be explained in the supplementary materials.

3.2 Encoding Phase

For the diffusion process in latent space, encoding phase for I^G_{RGB}, I^L_{RGB} and U is implemented by the encoder $\mathcal{E}$. It makes global and local latent image embeddings $X^G_0, X^L_0 \in \mathbb{R}^{4\times64\times64}$ for I^G_{RGB} and I^L_{RGB}, and a latent text embedding

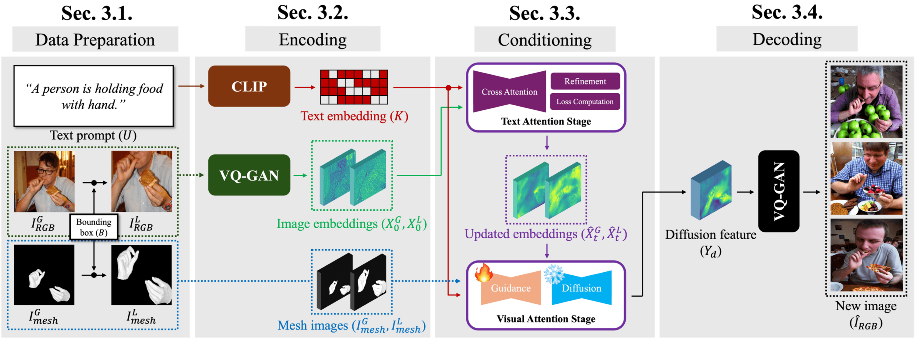

Fig. 3: Overall pipeline of AttentionHand. In the data preparation phase, we prepare global and local RGB images, global and local hand mesh images, bounding box, and text prompt. In the encoding phase, we get global and local latent image embeddings through VQ-GAN [30], and text embedding through CLIP [31]. In the conditioning phase, we refine image embeddings through the text attention stage, and obtain the diffusion feature through the visual attention stage. In the decoding phase, we generate a new hand image $\hat{I}_{RGB}$ from Y_d through VQ-GAN.

$K \in \mathbb{R}^{77\times768}$ for U. Specifically, X^G_0 and X^L_0 are obtained by VQ-GAN [30], and K is obtained by CLIP [31] as shown in the second box of Fig. 3. These latent embeddings are fed as inputs to the conditioning phase, which will be introduced by the next subsection. The encoding phase is expressed as follows:

$$X^G_0, X^L_0, K = \mathcal{E}(I^G_{RGB}, I^L_{RGB}, U). \tag{1}$$

3.3 Conditioning Phase

For generating new hand images conditioned by given text prompt and mesh images, we design the text attention stage (TAS) and the visual attention stage (VAS) in the conditioning phase, as shown in the third box of Fig. 3. TAS is a stage of paying attention to tokens for the hand and its corresponding gesture in a given text. VAS is a stage of training the SD-based model specialized for hand image generation by conditioning global and local mesh images.

Text Attention Stage (TAS). TAS is a stage of attending tokens which represent hand or gestures in a given text prompt as shown in Fig. 4(a). First, by adding Gaussian noise to X^G_0 and X^L_0 with t diffusion steps, the global noisy embedding X^G_t and local noisy embedding X^L_t are obtained. For simplicity, we define as $X_0 = (X^G_0, X^L_0)$ and $X_t = (X^G_t, X^L_t)$. Then, X_t and K are fed to TAS as inputs. For the text attention of TAS, we utilize the cross attention [32]. Specifically, an attention map $A \in \mathbb{R}^{H\times W\times N}$ is obtained by calculating the key (i.e., K) and query (i.e., Q, which is the linear projection of intermediate feature

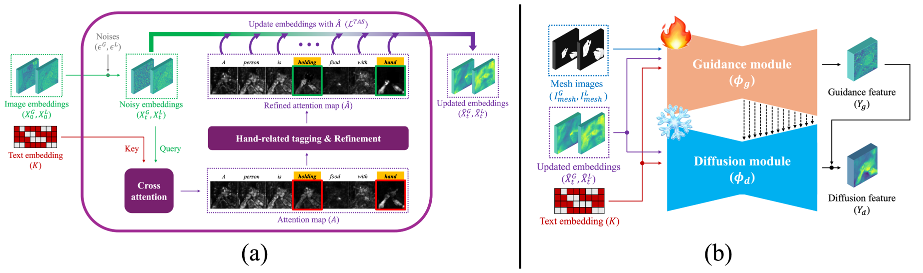
Update embeddings with $\hat{A}$ ($\mathcal{L}^{TAS}$)
Noises (ϵ^G, ϵ^L)
Image embeddings (X_0^G, X_0^L)
Noisy embeddings (X_t^G, X_t^L)
A
person
is
holding
food
with
hand
Refined attention map ($\hat{A}$)
Updated embeddings ($\hat{X}_t^G, \hat{X}_t^L$)
Text embedding (K)
Key
Query
Hand-related tagging & Refinement
Cross attention
Attention map (A)
(a)
Mesh images (I_{mesh}^G, I_{mesh}^L)
Updated embeddings ($\hat{X}_t^G, \hat{X}_t^L$)
Text embedding (K)
Guidance module (ϕ_g)
Diffusion module (ϕ_d)
Guidance feature (Y_g)
Diffusion feature (Y_d)
(b)

Fig. 4: (a) Overall process of the text attention stage (TAS). By leveraging the hand-related tagging and refinement, we can highlight hand-related attention maps, which leads to update noisy embeddings with $\mathcal{L}^{TAS}$. (b) Overall process of the visual attention stage (VAS). By utilizing the global and local information, we can obtain the harmonious diffusion feature, which leads to generate high-fidelity hand images.

map from X_t of U-Net [33] in SD). H and W denote the height and width of A, and N denotes the number of all tokens of K.

Next, to extract hand-related attention maps $A_{k \in K} \in \mathbb{R}^{H \times W \times N_k}$ in A, we design the hand-related tagging $\mathcal{H}_{tag}$, which is based on part-of-speech tagging [34], where N_k denotes the number of hand-related tokens in K. Specifically, $\mathcal{H}_{tag}$ determines if the input token indicates the hand-related word (i.e., *holding*, *taking*, or *hand*). With $\mathcal{H}_{tag}$, we can attend hand-related tokens k to generate more hand-focused images. More details are in the supplementary materials.

Then, we employ the softmax operation and Gaussian smoothing to maximize the effect of A_k. Since the Gaussian filter effectively removes noise from images and preserves detailed information by using the average value of surrounding pixels, we fully exploit these advantages. Hence, A_k is updated to $\hat{A}_k \in \mathbb{R}^{H \times W \times N_k}$ by refining hand-related attention maps as follows:

$$\hat{A}_k = Gaussian(Softmax(A_k)). \tag{2}$$

For simplicity, we define $\hat{A} \in \mathbb{R}^{H \times W \times N}$ as the concatenation of $\hat{A}_{k \in K}$ and $A_{l \notin K} \in \mathbb{R}^{H \times W \times N_l}$, where N_l denotes the number of not hand-related tokens in K, and $N = N_k + N_l$.

Moreover, optimization for evenly reflecting the image features of all attention maps is necessary. In other words, it is required to design an objective to prevent poor generation of the image feature for a specific token. Specifically, for arbitrary token $n \in K$, the highest value s_n among all patches in the n-th refined attention map $\hat{A}_n$ is extracted, and it is subtracted from 1. This operation is implemented for all tokens in K, and a novel loss, which named $\mathcal{L}^{TAS}$, is computed the largest value among them as follows:

$$\mathcal{L}^{TAS} = max_{n \in K}(1 - s_n). \tag{3}$$

Based on $\mathcal{L}^{TAS}$, X_t is updated to $\hat{X}_t = (\hat{X}_t^G, \hat{X}_t^L)$ as follows:

$$\hat{X}_t = X_t - \alpha_t \nabla_{X_t} \mathcal{L}^{TAS}, \tag{4}$$

where α_t indicates the learning rate, and $\hat{X}_t^G$ and $\hat{X}_t^L$ indicate updated global and local noisy embedding, respectively.

Visual Attention Stage (VAS). VAS is a stage of training SD-based model by conditioning the aforementioned global and local mesh image. VAS is composed of two modules: one is the guidance module ϕ_g, and the other is the diffusion module ϕ_d as shown in Fig. 4(b). First, the diffusion module ϕ_d is designed based on an U-Net network, consisting of 25 blocks: 8 blocks are downsampling and upsampling convolution layers, and the remaining 17 blocks consist of four ResNet [35] layers and two Vision Transformers [36]. We define the parameter set of ϕ_d as θ_d, which is fixed frozen to maintain the image generation performance of SD.

On the other hand, the guidance module ϕ_g is also based on an U-Net network with 25 blocks of ϕ_d. We define the parameter set of ϕ_g as θ_g, which is a copied version of θ_d. Different from θ_d, θ_g is set to be learnable for generating images conditioned to I_{mesh}^G and I_{mesh}^L. Specifically, ϕ_g has zero convolution $\mathcal{Z}$ [18] at the front of the network, and last 12 blocks of the network consist of $\mathcal{Z}$. Since $\mathcal{Z}$ is defined as a 1×1 convolution layer whose weights and bias are initialized to zero, the gradients of the weight and bias progressively grow from zeros to optimized parameters in a learnable manner. Hence, $\mathcal{Z}$ helps generated images to be conditioned on I_{mesh}^G and I_{mesh}^L, while maintaining the quality of image generation. More specifically, ϕ_d and ϕ_g share weights at the beginning of training, because parameter sets of both modules, i.e, θ_d and θ_g, are initialized with the pre-trained SD. However, while continuing with training process, θ_g is updated to learn I_{mesh}^G and I_{mesh}^L, whereas θ_d is fixed frozen to preserve the performance of image generation. At the end of training, θ_d and θ_g are completely different from the beginning. Hence, ϕ_g is formulated as follows:

$$Y_g = \phi_g(\hat{X}_t, I_{mesh}, K, t; \theta_g), \tag{5}$$

where $I_{mesh} = (I_{mesh}^G, I_{mesh}^L)$ denotes the concatenation of the global and local mesh image, t denotes the diffusion step obtained by positional encoding, $Y_g = (Y_g^G, Y_g^L) \in \mathbb{R}^{2\times4\times64\times64}$ denotes the concatenation of the global guidance feature Y_g^G and local guidance feature Y_g^L. Next, ϕ_d is formulated as follows:

$$Y_d = \phi_d(\hat{X}_t, K, t; \theta_d) + Y_g, \tag{6}$$

where $Y_d = (Y_d^G, Y_d^L) \in \mathbb{R}^{2\times4\times64\times64}$ indicates the concatenation of the global diffusion feature Y_d^G and local diffusion feature Y_d^L.

Optimization. Since the diffusion model typically involves both forward and reverse processes, our AttentionHand also employs two processes. For the forward process, the noisy embedding $X_t = (X_t^G, X_t^L)$ is obtained by progressively perturbing Gaussian noise $\epsilon = (\epsilon^G, \epsilon^L)$ to the initial embedding $X_0 = (X_0^G, X_0^L)$ by t diffusion steps. ϵ^G and ϵ^L denote the global and local noise added to X_0^G and X_0^L, respectively. Then, since X_t is updated to $\hat{X}_t$ by TAS, ϵ is also updated to $\hat{\epsilon} = (\hat{\epsilon^G}, \hat{\epsilon^L})$. In other words, $\hat{\epsilon}$ is considered the residual noise between X_0 and $\hat{X}_t$. Thus, $\hat{\epsilon^G}$ and $\hat{\epsilon^L}$ denote the global and local residual noise. For the reverse process, AttentionHand learns to gradually remove residual noises with global and local denoising processes. Therefore, given text embedding K, diffusion steps t, and mesh images I^G_{mesh} and I^L_{mesh}, the diffusion training network ϵ_θ is optimized to predict $\hat{\epsilon^G}$ and $\hat{\epsilon^L}$ jointly through the following objectives:

$$\mathcal{L}^G = \mathbb{E}_{X_0^G, I^G_{mesh}, K, t, \hat{\epsilon^G} \sim \mathcal{N}(0,1)}[\|\hat{\epsilon^G} - \epsilon_\theta(\hat{X}_t^G, I^G_{mesh}, K, t)\|_2^2], \tag{7}$$

$$\mathcal{L}^L = \mathbb{E}_{X_0^L, I^L_{mesh}, K, t, \hat{\epsilon^L} \sim \mathcal{N}(0,1)}[\|\hat{\epsilon^L} - \epsilon_\theta(\hat{X}_t^L, I^L_{mesh}, K, t)\|_2^2], \tag{8}$$

where $\mathcal{L}^G$ and $\mathcal{L}^L$ indicate the cost function of global and local features, respectively. Thus, the final objective is defined as follows:

$$\mathcal{L} = \lambda^G \mathcal{L}^G + \lambda^L \mathcal{L}^L, \tag{9}$$

where λ^G and λ^L are weighted coefficients of $\mathcal{L}^G$ and $\mathcal{L}^L$.

3.4 Decoding Phase

In the decoding phase, we can generate a new RGB hand image $\hat{I}_{RGB} \in \mathbb{R}^{3\times512\times512}$ by passing Y_d through the decoder $\mathcal{D}$, as shown in the fourth box of Fig. 3. While $\mathcal{E}$ encodes X_0 by downsampling I_{RGB} in the latent space, $\mathcal{D}$ decodes $\hat{I}_{RGB}$ by upsampling Y_d in the pixel space, conditioned to given text prompt and mesh images. The structure of $\mathcal{D}$ is similar to the decoder of VQ-GAN. The decoding phase is expressed as follows:

$$\hat{I}_{RGB} = \mathcal{D}(Y_d). \tag{10}$$

4 Experiments

4.1 Datasets

For the text-to-image generation, we adopted MSCOCO [6]. For the 3D hand mesh reconstruction, we adopted Hands-In-Action (HIC) [37], Re:InterHand (ReIH) [5], InterHand2.6M (IH2.6M) [7], and MSCOCO. Due to the page limit, details will be explained in the supplementary materials.

Table 1: Quantitative comparisons with state-of-the-art text-to-image generation models.

Methods	FID↓	KID↓	FID-H↓	KID-H↓	Hand Conf.↑	MSE-2D↓	MSE-3D↓	User Pref.(%)↑
Stable Diffusion [17]	40.52	0.00684	50.78	0.02554	0.651	2.932	4.591	5.864
Uni-ControlNet [20]	30.34	0.00744	37.77	0.02004	0.855	2.105	3.039	8.796
T2I-Adapter [19]	22.00	0.00761	32.08	0.01568	0.914	1.546	2.451	19.676
ControlNet [18]	21.67	0.00658	40.32	0.02098	0.810	1.252	2.182	7.948
AttentionHand (w/o TAS)	21.27	0.00331	28.56	0.01390	0.955	1.211	2.042	20.734
AttentionHand (w/ TAS)	**20.71**	**0.00301**	**27.09**	**0.01287**	**0.965**	**1.026**	**1.986**	**36.905**

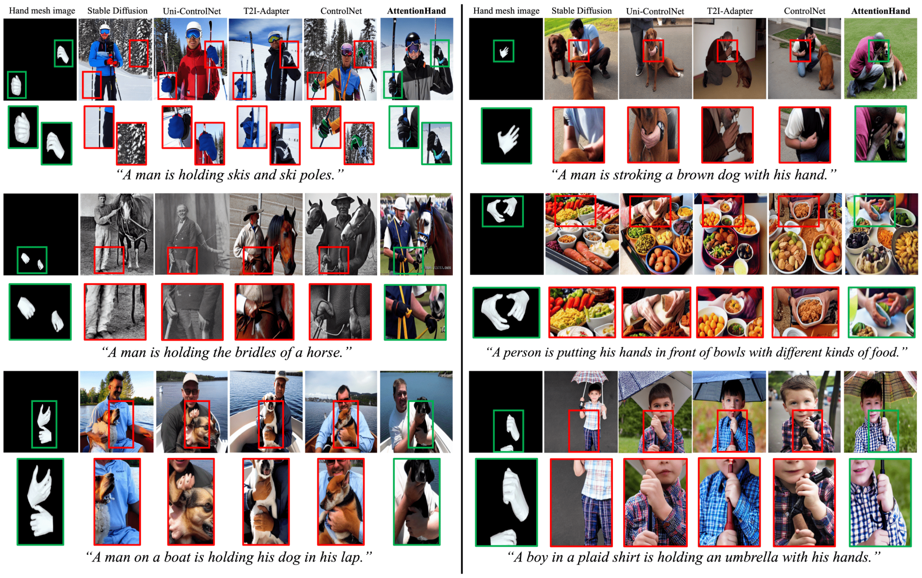
Hand mesh image
Stable Diffusion
Uni-ControlNet
T2I-Adapter
ControlNet
AttentionHand
"A man is holding skis and ski poles."
"A man is stroking a brown dog with his hand."
"A man is holding the bridles of a horse."
"A person is putting his hands in front of bowls with different kinds of food."
"A man on a boat is holding his dog in his lap."
"A boy in a plaid shirt is holding an umbrella with his hands."

Fig. 5: Qualitative comparisons with state-of-the-art text-to-image generation models. Red and green boxes in each sample indicate the wrong and corrent hand bounding box, respectively.

4.2 Evaluation Protocol

For the text-to-image generation, we adopted FID [38], FID-Hand (FID-H), KID [39], KID-Hand (KID-H), the hand confidence score (Hand Conf.) [40], the mean square error of 2D and 3D keypoints (MSE-2D, 3D), and the user preference (User Pref.). For the 3D hand mesh reconstruction, we adopted the mean per-vertex position error (MPVPE), the right hand-relative vertex error (RRVE), and the mean relative-root position error (MRRPE). Due to the page limit, details will be explained in the supplementary materials.

4.3 Comparisons with State-of-the-arts

Text-to-Image Generation. As shown in Table 1, our AttentionHand exhibited the highest performance in all metrics among state-of-the-arts [17–20]. This

Table 2: Quantitative comparisons with state-of-the-art 3D hand mesh reconstruction methods with and without AttentionHand. The red subscripts indicate the difference in performance with and without AttentionHand.

Datasets	In-the-wild						In-the-lab		
	HIC [37]			ReIH [5]			IH2.6M [7]		
Methods	MPVPE↓	RRVE↓	MRRPE↓	MPVPE↓	RRVE↓	MRRPE↓	MPVPE↓	RRVE↓	MRRPE↓
IHMR [8]	38.57	45.51	119.64	30.90	45.55	98.45	16.94	21.98	33.39
IHMR+AttentionHand	**36.73**$_{-1.84}$	**44.10**$_{-1.41}$	**94.63**$_{-25.01}$	**29.11**$_{-1.79}$	**43.12**$_{-2.43}$	**87.07**$_{-11.38}$	**15.09**$_{-1.85}$	**20.55**$_{-1.43}$	**32.21**$_{-1.18}$
InterShape [9]	27.66	34.69	110.25	27.87	38.56	80.04	12.97	17.35	31.56
InterShape+AttentionHand	**25.04**$_{-2.62}$	**33.33**$_{-1.36}$	**80.17**$_{-30.08}$	**26.44**$_{-1.43}$	**36.54**$_{-2.02}$	**61.41**$_{-18.63}$	**11.90**$_{-1.07}$	**16.22**$_{-1.13}$	**30.04**$_{-1.52}$
IntagHand [10]	23.07	28.74	52.46	25.90	30.05	42.22	12.34	17.32	29.31
IntagHand+AttentionHand	**21.87**$_{-1.20}$	**27.09**$_{-1.65}$	**47.11**$_{-5.35}$	**23.39**$_{-2.51}$	**28.77**$_{-1.28}$	**33.98**$_{-8.24}$	**11.42**$_{-0.92}$	**15.81**$_{-1.51}$	**29.18**$_{-0.13}$
DIR [13]	21.89	26.11	43.11	21.82	29.66	37.01	10.26	17.11	28.98
DIR+AttentionHand	**20.66**$_{-1.23}$	**25.87**$_{-0.24}$	**40.54**$_{-2.57}$	**19.91**$_{-1.91}$	**26.67**$_{-2.99}$	**35.05**$_{-1.96}$	**10.09**$_{-0.17}$	**16.99**$_{-0.12}$	**28.02**$_{-0.96}$
InterWild [16]	15.30	21.35	31.26	13.99	20.07	22.38	11.52	19.77	26.87
InterWild+AttentionHand	**14.74**$_{-0.56}$	**21.10**$_{-0.25}$	**29.26**$_{-2.00}$	**13.95**$_{-0.04}$	**19.94**$_{-0.13}$	**22.05**$_{-0.33}$	**10.62**$_{-0.90}$	**19.09**$_{-0.68}$	**25.74**$_{-1.13}$

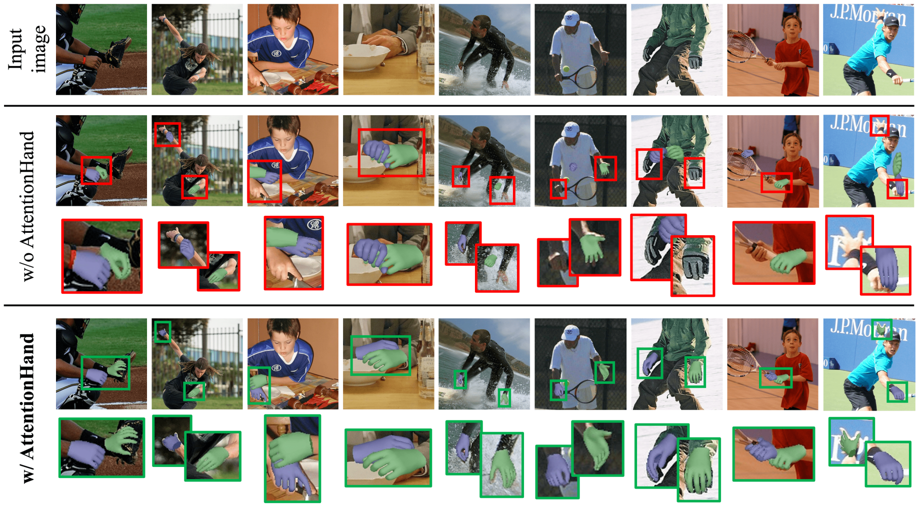

Fig. 6: Qualitative comparisons on MSCOCO [6]. Red and green boxes indicate wrong and correct region of the reconstructed hand, respectively.

is particularly evident in the comparison of FID(-H) and KID(-H), which signify the quality of the generated images being on par with real RGB images. Furthermore, the lowest MSE-2D and MSE-3D indicates the remarkable alignment between the generated images and the corresponding hand mesh images. With respect to the user preference, AttentionHand scored the highest compared to other methods. It implies that most users acknowledged the outstanding quality of hand images generated by AttentionHand. In addition, as shown in Fig. 5, our AttentionHand generated the high-quality hand image which is well-corresponding with the given mesh image and fully reflected the given text prompt. Specifically, even when two-hands mesh image is given, which is more challenging than in the case of single-hand mesh image, AttentionHand generated the hand image robustly. It implies our AttentionHand is proper to generate well-aligned hand images with given mesh images and text prompt. Additional qualitative results are in the supplementary materials.

3D Hand Mesh Reconstruction. To verify our AttentionHand extensively, we trained state-of-the-art hand pose networks [8–10, 13, 16] by additionally adding new data generated by AttentionHand. As shown in Table 2, the performance of all methods increased for all metrics. Specifically, with respect to the MPVPE, AttentionHand showed the dramatic performance improvement with InterWild [16] about 3.66% and 7.81% on HIC and ReIH, respectively. With respect to the RRVE, it increased by about 1.17% and 0.65% on HIC and ReIH, respectively. With respect to the MRRPE, it increased by about 6.40% and 1.47% on HIC and ReIH, respectively. These imply generated hand images help increasing the accuracy of the 3D hand mesh reconstruction. In addition, the qualitative performance for in-the-wild scenes is also verified as shown in Fig. 6. Although MSCOCO mainly contains in-the-wild situations, 3D hand mesh is reconstructed robustly. It implies that even for difficult situations, the performance of reconstruction can be improved by utilizing AttentionHand. Additional qualitative results are in the supplementary materials.

4.4 Ablation Studies

Text Attention Stage (TAS). We deeply dived into TAS to verify its superiority. Firstly, as in the last two rows in Table 1, TAS showed its effectiveness in all metrics. In addition, as shown in Fig. 7, attention maps are well described their corresponding tokens in the case of with TAS. It implies that with TAS, AttentionHand can reflect hand-related tokens enough. Additional qualitative results are in the supplementary materials.

Secondly, we conducted more experiments about Gaussian filters as follows: (1) no Gaussian filter, (2) random Gaussian filter, and (3) fixed Gaussian filter. As shown in the second, third, and fourth columns of Fig. 8, we found interesting results: in the case of (1), the hand was disappeared or its shape became strange. In the case of (2), generated images are not well-aligned with given hand mesh images. However, in the case of (3), generated images are well-aligned with given hand mesh images and look natural. Hence, we determined fixed Gaussian filter makes the generated image plausibly regardless of diffusion timestep t.

Thirdly, we compared our loss, $\mathcal{L}^{TAS}$, with the load balancing loss ($\mathcal{L}^{LB}$) [41, 42]. Since $\mathcal{L}^{LB}$ is an auxiliary loss for balancing loads among experts, it plays a similar role with $\mathcal{L}^{TAS}$, which evenly reflects the image features of all the attention maps. Therefore, we replaced $\mathcal{L}^{TAS}$ to $\mathcal{L}^{LB}$ and considered its feasibility as shown in the fifth column of Fig. 8. Unfortunately, in the case of $\mathcal{L}^{LB}$, generated images are not fit at all with given hand mesh images. We guess while $\mathcal{L}^{TAS}$ updates the image embedding based on the spatial information of the attention map, $\mathcal{L}^{LB}$ flattens the 2D attention map as 1D representation, leading to distort spatial knowledge.

Last but not least, we explored the range of updated noise ($\hat{\epsilon}$). According to [43], we set α_t of Eq. 4 as gradually decreasing according to timestep t (i.e.,

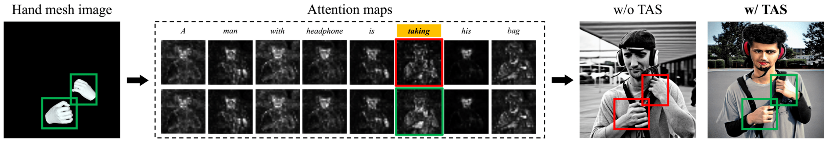

Fig. 7: Ablation studies on the text attention stage (TAS). Attention maps with red and green box are results without and with TAS, respectively. Red and green bounding boxes indicate wrong and correct hand poses, respectively.

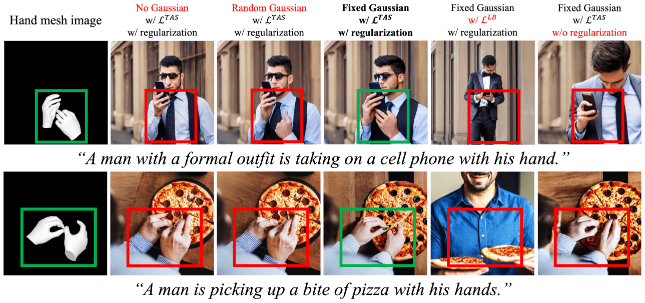
Hand mesh image
No Gaussian w/ $\mathcal{L}^{TAS}$ w/ regularization
Random Gaussian w/ $\mathcal{L}^{TAS}$ w/ regularization
Fixed Gaussian w/ $\mathcal{L}^{TAS}$ w/ regularization
Fixed Gaussian w/ $\mathcal{L}^{LB}$ w/ regularization
Fixed Gaussian w/ $\mathcal{L}^{TAS}$ w/o regularization
"A man with a formal outfit is taking on a cell phone with his hand."
"A man is picking up a bite of pizza with his hands."

Fig. 8: Ablation studies on Gaussian filter, losses (i.e. $\mathcal{L}^{TAS}$ and $\mathcal{L}^{LB}$), and the regularization of $\hat{\epsilon}$ for the text attention stage (TAS). Red and green bounding boxes indicate wrong and correct hand poses, respectively.

from 20 to 10) for regularization of $\hat{\epsilon}$. However, if α_t is randomly set, $\hat{\epsilon}$ tends to be out of distribution (i.e., Gaussian distribution) as shown in the sixth column of Fig. 8: in the case of w/o regularization, generated images are not aligned with given mesh images, or missed some hands. Therefore, it is necessary to regularize $\hat{\epsilon}$ for faithful hand image generation.

Model Design Justification. To justify our model's superiority, we compared the characteristics of prior works including our model. As shown in Table 3, our model's distinctive and potential features compared to prior works are (1) harmonious preservation of locality (i.e., hand) with globality (i.e., in-the-wild scene), and (2) selective attention on hand-related tokens by cross attention. Specifically, to harmonize globality and locality, we developed global and local designs for the visual attention stage (VAS). Moreover, since the global and local branches are designed structurally same, we set them to share their weights for reducing the number of training parameters (about 20.2% ↓) and improving the generalizability (see two shaded rows in Table 4). We experimentally verified the effectiveness of our design as shown in Table 4.

Table 3: Network comparisons with prior works.

Methods	Text Prompt	Visual Prompt	Locality	Hand-related Token Attention
Stable Diffusion	✓			
Uni-ControlNet	✓	✓		
T2I-Adapter	✓	✓		
ControlNet	✓	✓		
AttentionHand (Ours)	✓	✓	✓	✓

Table 4: Ablation studies on the visual attention stage (VAS).

Globality	Locality	Weights	FID↓	KID↓	FID-H↓	KID-H↓	Hand Conf.↑	MSE-2D↓	MSE-3D↓
		Shared	40.52	0.00684	50.78	0.02554	0.651	2.932	4.591
✓		Shared	21.67	0.00658	40.32	0.02098	0.810	1.252	2.182
	✓	Shared	52.98	0.00713	32.11	0.01604	0.911	1.539	2.397
✓	✓	**Shared**	**20.71**	0.00301	27.09	**0.01287**	**0.965**	**1.026**	**1.986**
✓	✓	Separated	21.90	**0.00293**	**26.89**	0.01340	0.960	1.108	2.017

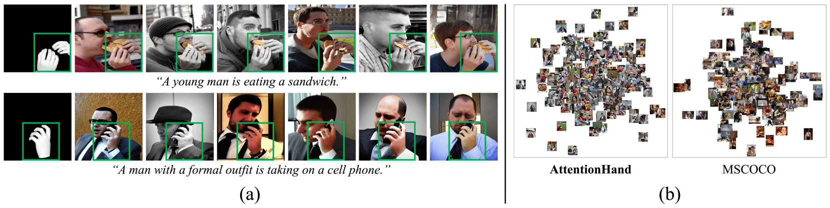
"A young man is eating a sandwich."
"A man with a formal outfit is taking on a cell phone."
(a)
AttentionHand
MSCOCO
(b)

Fig. 9: (a) Multiple generated hand images from same modalities. Green boxes indicate correct hand poses. (b) t-SNE distribution of AttentionHand and MSCOCO [6].

Robustness of Generated Dataset. To verify robustness of our generated dataset, we generated multiple hand images from same modalities as shown in Fig. 9(a). As a result, all generated images are perfectly well-aligned with given hand mesh images. Moreover, we found the t-SNE distribution [44] of AttentionHand is broader than MSCOCO as shown in Fig. 9(b). As a result, we believe that AttentionHand can contribute to the downstream task with our extensive in-the-wild hand images, leading to alleviate the domain gap between indoor and outdoor scenes.

5 Conclusion

In this paper, we introduced a novel text-to-hand image generation model, AttentionHand, which pays attention to the hand-related tokens from the text prompt and global and local mesh images. AttentionHand achieved state-of-the-art performance in text-to-hand image generation, and we demonstrated that training with the dataset generated by our AttentionHand improved the performance of 3D hand mesh reconstruction. However, the diversity may decrease as the generative model is trained to optimize hand mesh images. We expect for the emergence of outstanding diffusion model to improve the diversity and quality of the hand image.

Acknowledgements. This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2024-RS-2023-00260091) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation) and Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (MOTIE) (P0020535, The Competency Development Program for Industry Specialist) and National Supercomputing Center with supercomputing resources including technical support (KSC-2023-CRE-0444).

References

1. Hampali, Shreyas and Rad, Mahdi and Oberweger, Markus and Lepetit, Vincent. Honnotate: A method for 3D annotation of hand and object poses. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3196–3206, 2020.
2. Chao, Yu-Wei and Yang, Wei and Xiang, Yu and Molchanov, Pavlo and Handa, Ankur and Tremblay, Jonathan and Narang, Yashraj S and Van Wyk, Karl and Iqbal, Umar and Birchfield, Stan and Kautz, Jan and Fox, Dieter. DexYCB: A benchmark for capturing hand grasping of objects. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9044–9053, 2021.
3. Ohkawa, Takehiko and He, Kun and Sener, Fadime and Hodan, Tomas and Tran, Luan and Keskin, Cem. Assemblyhands: Towards egocentric activity understanding via 3D hand pose estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12999–13008, 2023.
4. Lin, Fanqing and Wilhelm, Connor and Martinez, Tony. Two-hand global 3d pose estimation using monocular rgb. pages 2373–2381, 2021.
5. Moon, Gyeongsik and Saito, Shunsuke and Xu, Weipeng and Joshi, Rohan and Buffalini, Julia and Bellan, Harley and Rosen, Nicholas and Richardson, Jesse and Mize, Mallorie and De Bree, Philippe and Simon, Tomas and Peng, Bo and Garg, Shubham and McPhail, Kevyn and Shiratori, Takaaki. A dataset of relighted 3d interacting hands. *Adv. Neural Inform. Process. Syst.*, 36, 2023.
6. Lin, Tsung-Yi and Maire, Michael and Belongie, Serge and Hays, James and Perona, Pietro and Ramanan, Deva and Dollár, Piotr and Zitnick, C Lawrence. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis.*, pages 740–755, 2014.
7. Moon, Gyeongsik and Yu, Shoou-I and Wen, He and Shiratori, Takaaki and Lee, Kyoung Mu. Interhand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image. In *Eur. Conf. Comput. Vis.*, pages 548–564, 2020.
8. Rong, Yu and Wang, Jingbo and Liu, Ziwei and Loy, Chen Change. Monocular 3D reconstruction of interacting hands via collision-aware factorized refinements. In *3DV*, pages 432–441, 2021.
9. Zhang, Baowen and Wang, Yangang and Deng, Xiaoming and Zhang, Yinda and Tan, Ping and Ma, Cuixia and Wang, Hongan. Interacting two-hand 3D pose and shape reconstruction from single color image. In *Int. Conf. Comput. Vis.*, pages 11354–11363, 2021.
10. Li, Mengcheng and An, Liang and Zhang, Hongwen and Wu, Lianpeng and Chen, Feng and Yu, Tao and Liu, Yebin. Interacting attention graph for single image two-hand reconstruction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2761–2770, 2022.

11. Hampali, Shreyas and Sarkar, Sayan Deb and Rad, Mahdi and Lepetit, Vincent. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3D pose estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11090–11100, 2022.
12. Meng, Hao and Jin, Sheng and Liu, Wentao and Qian, Chen and Lin, Mengxiang and Ouyang, Wanli and Luo, Ping. 3D interacting hand pose estimation by hand de-occlusion and removal. In *Eur. Conf. Comput. Vis.*, pages 380–397, 2022.
13. Ren, Pengfei and Wen, Chao and Zheng, Xiaozheng and Xue, Zhou and Sun, Haifeng and Qi, Qi and Wang, Jingyu and Liao, Jianxin. Decoupled Iterative Refinement Framework for Interacting Hands Reconstruction from a Single RGB Image. In *Int. Conf. Comput. Vis.*, pages 8014–8025, 2023.
14. Zuo, Binghui and Zhao, Zimeng and Sun, Wenqian and Xie, Wei and Xue, Zhou and Wang, Yangang. Reconstructing interacting hands with interaction prior from monocular images. In *Int. Conf. Comput. Vis.*, pages 9054–9064, 2023.
15. Li, Lijun and Tian, Linrui and Zhang, Xindi and Wang, Qi and Zhang, Bang and Bo, Liefeng and Liu, Mengyuan and Chen, Chen. Renderih: A large-scale synthetic dataset for 3d interacting hand pose estimation. In *Int. Conf. Comput. Vis.*, pages 20395–20405, 2023.
16. Moon, Gyeongsik. Bringing inputs to shared domains for 3D interacting hands recovery in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 17028–17037, 2023.
17. Rombach, Robin and Blattmann, Andreas and Lorenz, Dominik and Esser, Patrick and Ommer, Björn. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10684–10695, 2022.
18. Zhang, Lvmin and Rao, Anyi and Agrawala, Maneesh. Adding conditional control to text-to-image diffusion models. In *Int. Conf. Comput. Vis.*, pages 3836–3847, 2023.
19. Mou, Chong and Wang, Xintao and Xie, Liangbin and Zhang, Jian and Qi, Zhongang and Shan, Ying and Qie, Xiaohu. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.
20. Zhao, Shihao and Chen, Dongdong and Chen, Yen-Chun and Bao, Jianmin and Hao, Shaozhe and Yuan, Lu and Wong, Kwan-Yee K. Uni-ControlNet: All-in-one control to text-to-image diffusion models. *Adv. Neural Inform. Process. Syst.*, 2023.
21. Podell, Dustin and English, Zion and Lacey, Kyle and Blattmann, Andreas and Dockhorn, Tim and Müller, Jonas and Penna, Joe and Rombach, Robin. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.
22. Mueller, Franziska and Bernard, Florian and Sotnychenko, Oleksandr and Mehta, Dushyant and Sridhar, Srinath and Casas, Dan and Theobalt, Christian. GANerated hands for real-time 3D hand tracking from monocular RGB. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 49–59, 2018.
23. Tang, Hao and Wang, Wei and Xu, Dan and Yan, Yan and Sebe, Nicu. GestureGAN for hand gesture-to-gesture translation in the wild. In *ACM Int. Conf. Multimedia*, pages 774–782, 2018.
24. Hu, Hezhen and Wang, Weilun and Zhou, Wengang and Zhao, Weichao and Li, Houqiang. Model-aware gesture-to-gesture translation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16428–16437, 2021.
25. Hu, Hezhen and Wang, Weilun and Zhou, Wengang and Li, Houqiang. Hand-object interaction image generation. *Adv. Neural Inform. Process. Syst.*, 35:23805–23817, 2022.

26. Goodfellow, Ian and Pouget-Abadie, Jean and Mirza, Mehdi and Xu, Bing and Warde-Farley, David and Ozair, Sherjil and Courville, Aaron and Bengio, Yoshua. Generative adversarial nets. *Adv. Neural Inform. Process. Syst.*, 27:2672–2680, 2014.
27. Li, Lijun and Zhuo, Li'an and Zhang, Bang and Bo, Liefeng and Chen, Chen. DiffHand: End-to-end hand mesh reconstruction via diffusion models. *arXiv preprint arXiv:2305.13705*, 2023.
28. Lin, Pei and Xu, Sihang and Yang, Hongdi and Liu, Yiran and Chen, Xin and Wang, Jingya and Yu, Jingyi and Xu, Lan. HandDiffuse: Generative controllers for two-hand interactions via diffusion models. *arXiv preprint arXiv:2312.04867*, 2023.
29. Lu, Wenquan and Xu, Yufei and Zhang, Jing and Wang, Chaoyue and Tao, Dacheng. HandRefiner: Refining malformed hands in generated images by diffusion-based conditional inpainting. *arXiv preprint arXiv:2311.17957*, 2023.
30. Esser, Patrick and Rombach, Robin and Ommer, Bjorn. Taming transformers for high-resolution image synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12873–12883, 2021.
31. Radford, Alec and Kim, Jong Wook and Hallacy, Chris and Ramesh, Aditya and Goh, Gabriel and Agarwal, Sandhini and Sastry, Girish and Askell, Amanda and Mishkin, Pamela and Clark, Jack and Krueger, Gretchen and Sutskever, Ilya. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn.*, pages 8748–8763, 2021.
32. Chen, Chun-Fu Richard and Fan, Quanfu and Panda, Rameswar. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Int. Conf. Comput. Vis.*, pages 357–366, 2021.
33. Ronneberger, Olaf and Fischer, Philipp and Brox, Thomas. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.
34. Chiche, Alebachew and Yitagesu, Betselot. Part of speech tagging: a systematic review of deep learning and machine learning approaches. *Journal of Big Data*, 9(1):1–25, 2022.
35. He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016.
36. Dosovitskiy, Alexey and Beyer, Lucas and Kolesnikov, Alexander and Weissenborn, Dirk and Zhai, Xiaohua and Unterthiner, Thomas and Dehghani, Mostafa and Minderer, Matthias and Heigold, Georg and Gelly, Sylvain and Uszkoreit, Jakob and Houlsby, Neil. An image is worth 16x16 words: Transformers for image recognition at scale. In *Int. Conf. Learn. Represent.*, 2020.
37. Tzionas, Dimitrios and Ballan, Luca and Srikantha, Abhilash and Aponte, Pablo and Pollefeys, Marc and Gall, Juergen. Capturing hands in action using discriminative salient points and physics simulation. *Int. J. Comput. Vis.*, 118:172–193, 2016.
38. Heusel, Martin and Ramsauer, Hubert and Unterthiner, Thomas and Nessler, Bernhard and Hochreiter, Sepp. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Adv. Neural Inform. Process. Syst.*, 30:6626–6637, 2017.
39. Bińkowski, Mikołaj and Sutherland, Danica J and Arbel, Michael and Gretton, Arthur. Demystifying MMD GANs. In *Int. Conf. Learn. Represent.*, 2018.
40. Narasimhaswamy, Supreeth and Bhattacharya, Uttaran and Chen, Xiang and Dasgupta, Ishita and Mitra, Saayan and Hoai, Minh. Handiffuser: Text-to-image generation with realistic hand appearances. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2468–2479, 2024.

41. Zhou, Yanqi and Lei, Tao and Liu, Hanxiao and Du, Nan and Huang, Yanping and Zhao, Vincent and Dai, Andrew M and Le, Quoc V and Laudon, James. Mixture-of-experts with expert choice routing. *Adv. Neural Inform. Process. Syst.*, 35:7103–7114, 2022.
42. Fedus, William and Zoph, Barret and Shazeer, Noam. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.*, 23(120):1–39, 2022.
43. Chefer, Hila and Alaluf, Yuval and Vinker, Yael and Wolf, Lior and Cohen-Or, Daniel. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Trans. Graph.*, 42(4):1–10, 2023.
44. Van der Maaten, Laurens and Hinton, Geoffrey. Visualizing data using t-SNE. *J. Mach. Learn. Res.*, 9(86):2579–2605, 2008.