Person in Place: Generating Associative Skeleton-Guidance Maps for Human-Object Interaction Image Editing

ChangHee Yang*[1,2] ChanHee Kang*[1] Kyeongbo Kong*[3] Hanni Oh[1] Suk-Ju Kang†[1]
Sogang University[1] AI Lab, CTO Division, LG Electronics[2] Pusan National University[3]
{yangchanghee2251,kkb4723,hannixxxoh}@gmail.com, jasper695@icloud.com, sjkang@sogang.ac.kr

(a) Initial image (b) Sequential editing process (c) Skeleton map + Result image

Figure 1. **Human-object interaction (HOI) image editing using generated skeleton:** We synthesize human interacting with objects for an initial image using the automated object-interactive diffuser. (a) an initial image to edit. (b) the sequential process of synthesizing human image with object-interactive skeletons using textual conditions. Given human bounding box and object bounding box our object-interactive diffuser generate a skeleton interacting with the object. Then a skeleton guided image editing model edit the image with the generated skeleton. (c) a final result image with the skeleton map. Our method generates the high quality object interactive skeleton map, and it can easily plug in to the skeleton guided generative model for HOI image editing.

Abstract

Recently, there were remarkable advances in image editing tasks in various ways. Nevertheless, existing image editing models are not designed for Human-Object Interaction (HOI) image editing. One of these approaches (e.g. ControlNet) employs the skeleton guidance to offer precise representations of human, showing better results in HOI image editing. However, using conventional methods, manually creating HOI skeleton guidance is necessary. This paper proposes the object interactive diffuser with associative attention that considers both the interaction with objects and the joint graph structure, automating the generation of HOI skeleton guidance. Additionally, we propose the HOI loss with novel scaling parameter, demonstrating its effectiveness in generating skeletons that interact better. To evaluate generated object-interactive skeletons, we propose two metrics, top-N accuracy and skeleton probabilistic distance. Our framework integrates object interactive diffuser that generates object-interactive skeletons with previous methods, demonstrating the outstanding results in HOI image editing. Finally, we present potentials of our framework beyond HOI image editing, as applications to human-to-human interaction, skeleton editing, and 3D mesh optimization. The code is available at https://github.com/YangChangHee/CVPR2024_Person-In-Place_RELEASE

1. Introduction

Look at a photograph of a room like Fig. 1 (a), without any human presence. While being inherently fancy, the scene gains vivacity when people are added. The addition of people should not be arbitrary, as only plausible person-in-place placements can enhance the scene's natural dynamics. For instance, in this case, a woman sitting on the bed, a man leaning against the bed, a boy sitting on the sofa and a child looking at the flowerpot would be candidates. Is it possible to synthesize these people altogether with initial image naturally? This is Human-Object Interaction (HOI) image editing that we want to solve.

Specifically, an image to edit, person bounding boxes

*These authors contributed equally to this work
†Corresponding author

and object bounding boxes are required to our method. With these inputs, it generates the skeletons which interact with the objects specified by bounding boxes. Finally, we use off-the-shelf skeleton guided image editing models to edit the given image. Our method can use multiple bounding boxes simultaneously but for ease of explanation, Fig. 1 shows the sequential editing process of our method.

There have been significant advancements in image editing models. Two notable developments are the text-free editing model [1–10] and the text-guided editing model [11–27]. Text-free approaches employ the overall context of an image to fill masked areas. The advancement of text-guided approaches have been accelerated with the rise of diffusion models. Prompt-to-Prompt [28] constructs image editing interfaces only using text prompts. InstructPix2Pix [18] edits images using text prompts in the form of instructions. The blended latent diffusion model [23] proposed a text-guided approach to edit the desired areas using user-provided masks. However, aforementioned image editing methods obtain HOI images of low quality. This is because there is no module for object interaction widely considered in the general HOI fields [29–38] (HOI detection, HOI 3D motion generation, etc.).

Recently, numerous studies of text-guided models such as ControlNet [39] employ controlling conditions, e.g., edge, segmentation map, skeleton, for precise image editing and generation. As the utilization of these conditions offer precise representation of the subjects within the image, a high-quality image can be generated using controlling conditions. Among several conditions, skeleton guidance is mainly utilized to generate human-centric images, as shown in HumanSD [40]. In order to edit HOI images, object-interactive skeleton is required.

This paper proposes a novel framework for HOI image editing. Our framework consists of two stages. The first stage generates object-interactive skeletons, while the second stage generates HOI images using existing image editing models with skeleton guidance. We employed ControlNet [39] as the second stage model. Our framework shows its flexibility, since the second stage model can be replaced with any skeleton-guided image editing models [39–42].

In the first stage, our framework employs object interactive diffuser with the novel associative attention (A.A.) module to generate a object-interactive skeleton automatically. Notably, the A.A. module is a key contributor to generate skeleton-involving object interaction. Through the denoise process, the module uses object conditioning as key and value, while adopting image conditioning combined with noise pose embedding as query. The A.A. network computes attention between object and joint, which enables the natural interaction with object by considering relationship between the joint and the object. Moreover, the network propagate joint-wise features, stabilizing the consistency between joints. We are the first to present an A.A. mechanism which enables propagation based on relations of skeleton joints.

Finally, we also discovered the potentials of our method as follows. First, it makes it possible to generate an user-desired output. The automatically generated skeleton can be adjusted by users. Second, the generated skeleton can be optimized by SMPLify [43] to generate more aligned pseudo SMPL [44] ground truths (GTs). Third, our method could be used to human-to-human interaction.

The overall contribution can be summarized as follows:

- To the best of our knowledge, we are the first to attempt HOI image editing.
- We propose an automatic skeleton generation module: object interactive diffuser. Moreover, we propose a novel A.A. mechanism which considers the graph structure of joints and the relationship between an object and joints. Additionally, we propose the HOI loss with novel scaling parameter $Joint_{param}$, demonstrating its effectiveness in generating skeletons that interact better.
- Our framework outperformed quantitatively and qualitatively in HOI image editing field. In addition, we qualitatively show the synthesized results of multiple persons interacting with objects. Moreover, we present two novel metrics to measure how naturally the generated skeleton interacts with objects.

2. Related Works

Image Editing: Techniques in image editing can be categorized into text-free and text-guided methods. Text-free methods [1–10] focus on filling masked areas of an image by utilizing the image's overall context, aiming for a natural synthesis. For example, CoModGAN [3] propose co-modulated generative adversarial networks (GANs), a new method to reduce the gap between image conditional and unconditional GANs [45]. Text-guided methods can be divided into two parts. The first approach involves editing images based on a single text porompts, simplifying the editing process by adjusting the images's overall context according to the provided instructions [11–20]. Thus it is useful to modify overall context of the image. For instance, Instruct-Pix2Pix [18] edits an input image using user-provided instructions what the model should do. The second approach combines a textual prompt with a local mask for precise editing of specific areas, leveraging both the prompt and the image's surrounding context for detailed modifications [21–27]. GLIDE [25] exemplifies this approach by using a two-stage diffusion model process, starting with a low-resolution generation followed by an upsampled refinement based on the text prompt. These works above do not target the HOI image editing task. Hence, these editing would not properly conducted.

Skeleton Guided Image Generation: The latest ad-

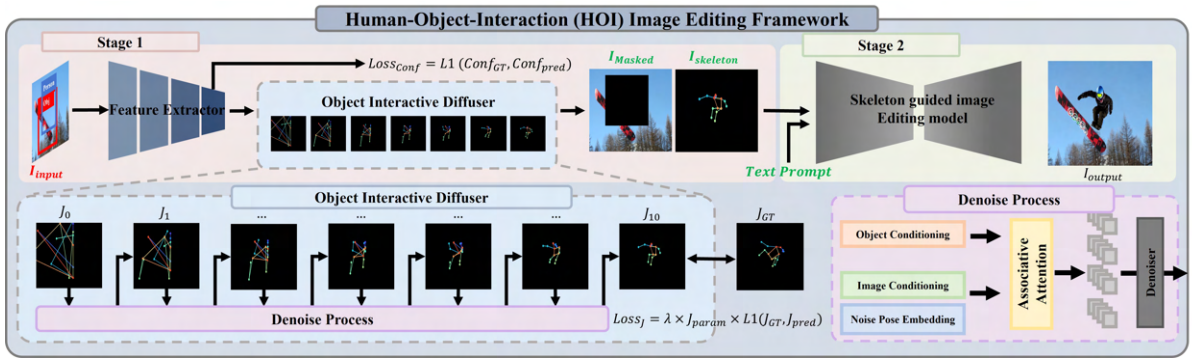

Figure 2. **Overview of proposed framework**: Our proposed framework uses a cropped image from a person bounding box as an input and the object bounding box. (**Left**) These are used to extract a image and an object features. (**Middle**) The extracted features are used as a image and object conditioning respectively in our object interactive diffuser. Using these conditionings, the object interactive diffuser comes to see the object-joint and joint-joint relationships then generate a denoised skeleton based on diffusion process. (**Right**) The synthesized skeleton together with a masked image using a person bounding box is used to edit image with off-the-shelf inpainting model.

vancements in image generation, especially diffusion models [6, 46, 47], have demonstrated remarkable capabilities in producing high-quality images across various domains such as [48–50]. Yet, accurately rendering human figures, which require detailed attention to both form and nuances, remains a challenge. To address this, recent works have been made to incorporate additional condition maps. Techniques like [39–41, 51–54] introduce additional condition maps, such as edges segmentation maps and skeletons to enhance the depiction of human figures. However, a critical limitation of these models lies in their inability to autonomously generate these additional information. This necessitates manual acquisition of such data so that these image generation models face constraints due to the need for manual input of supplementary details. In light of these challenges, we propose a method that amalgamates the realms of editing and human guidance image generation.

Human Object Interaction (HOI): The HOI domain emphasizes understanding and generating interactions between humans and objects. For instance, in the field of HOI detection [29, 55–60] which detect simultaneously human and object with their interaction, PaStaNet [29] is one example. The author of the PaStaNet [29] propose a method which considers a relationship between human body parts and objects. In HOI video detection which detect human and object relationship within videos, the author of [37] proposed graph parsing neural network (GPNN) which is a end-to-end framework that represents HOI graph structure explicitly with automatic optimal graph parsing. In addition, in HOI 3D motion generation [30, 61, 62] which reconstruct object interactive human motion using text, Humanise [30], uses self-attention to incorporate textual prompt with object point clouds. Finally, in HOI 3D reconstruction field [31, 63–67] which reconstruct a mesh or 3D primitives of object and human interaction are these examples, CHAIRS [31], estimate the root 6D pose of the object by considering the image feature and the SMPL-X [68] parameters. HOI is actively researched in various fields, and we are the first to attempt HOI image editing.

Attention Mechanism: The introduction of the transformer by [69] has revolutionized the use of transformers, significantly impacting various domain, including computer vision. Attention mechanisms enhance model performance by focusing on relevant features within an image, improving tasks like image classification, semantic segmentation and object localization. This approach has been integrated with convolutional neural networks (CNNs) to refine the extraction and processing of informaiton. Notable implementations include the squeeze and excitation [70], bottleneck attention module [71], convolutional block attention module [72], global context [73] and joint and triplet attention methods [74]. These mechanisms vary in their handling of feature inter-dependencies across channel and spatial dimensions, improving model accuracy and efficiency. Our work introduces an innovative application of attention mechanisms, specifically in the context of conditioning. This novel approach sets our method apart from existing techniques, demonstrating superior performance both quantitatively and qualitatively.

3. Proposed Method

In this section, we introduce our proposed image editing framework shown in Fig. 2. As shown at the top of the Fig. 2, our framework consists of three parts: feature extractor, object interactive diffuser, and skeleton guided image editing model. An image feature map, an object feature map and a joint confidence are extracted from feature extractor. Next, these feature maps are processed and fed into the object interative diffuser. Here, a noisy skeleton from the Gaussian distribution is used in denoising process as an initial skeleton. This process is repeated to generate an object-interactive denoised skeleton. Finally, through the skele-

ton guided image editing model, the generated skeleton and the masked image using a person bounding box are used to edit the image. Notably, this part can be combined with any image editing model that uses a skeleton as a condition. These part aforementioned parts are splited into two stage; the stage to generate objective-interactive skeleton and the stage to edit image with the generated skeleton guidance which is shown in Fig 2

3.1. Feature Extraction

In this section, we detail the feature extraction process which includes object and image features alongside joint confidences. Starting with an input image $I_{input} \in \mathbb{R}^{256\times256\times3}$ cropped to a person bounding box, we utilize a pretrained ResNet [75] for feature extraction. Specifically, object features $F_{Obj} \in \mathbb{R}^{8\times8\times1024}$ are derived from the third ResNet block's feature maps, chosen for their spatial information retention, via ROI pooling [76]. Image features F_{Img} are extracted from the fourth block's feature maps and processed through an MLP to estimate the confidence of each joint.

3.2. Object Interactive Diffuser

In this section, we introduce novel object interactive diffuser for object-interactive skeleton generation. This method is motivated by [77–79]. It iteratively creates a denoise embedding using our A.A. network. As shown in the right side of Fig. 3, the A.A. network uses the object and image conditioning to take into account the relationship between object and joints, as well as the relationship among joints. The A.A. first computes alignment using image and object conditionings. This alignment contains correlation of each joint. Unlike previous attention mechanisms that directly use this alignment to compute attention, our A.A. use GNN to propagate this per joint embedding. This propagated alignment is used to compute attention to estimate joint location.

3.2.1 Objective Interactive Conditioning

Before passing through A.A. network, we compute an image and object conditioning, using the image and object feature maps F_{Img} and F_{Obj} as followings:

$$I_{\text{condition}} = f(F_{Img}) \in \mathbb{R}^{N_J\times N_E}, \quad (1)$$

$$Obj_{\text{condition}} = \text{Conv}(F_{Obj}) \in \mathbb{R}^{N_p^2\times N_E}, \quad (2)$$

where $f(\cdot)$ is MLP network, N_p be the dimension of pooled feature and $\text{Conv}(\cdot)$ is a convolutional neural network. The image conditioning is summed with noise pose embedding (NPE) which are N_J noisy joints containing a embedding of a pose, where N_J is the number of joints and N_E be the embedding dimension as following formula. $I_{condition}$ and $Obj_{condition}$ stand for image conditioning and object conditioning respectively.

3.2.2 Associative Attention Network

Associative Joint Propagation

As mentioned at Eq. 2, the object conditioning $Obj_{condition}$ denotes the spatial information of the object contained in the image. This is important because, which part of the object to interact with is crucial when calculating meaningful attention to each joint. Consider a pixel on the snowboard and a pixel outside the snowboard in the Fig. 3. Our A.A. network computes alignment using both image conditioning and object conditioning and propagate this alignment with GNN. Therefore, the attention computed using a pixel inside the snowboard is stronger than using a pixel outside the snowboard. These process is depicted in the left side of Fig. 3 which is the process for computing alignment. In the context of the paper [69], we employ the image conditioning summed with noise pose embedding as a query Q and the object conditioning as a key K shown in Eq. 4 and Eq. 5, respectively. Then the alignment is computed as QK^T. The compute alignment A as followings:

$$A = Q \cdot K^T, \quad (3)$$

where NPE is the noise pose embeddings, Q and K are

$$Q = I_{\text{condition}} + \text{NPE}, \quad (4)$$

$$K = Obj_{\text{condition}}. \quad (5)$$

This section explains how the previously mentioned alignment can be used to discover the relationship between joints, using a GNN to compute attention unlike previous attention based mechanisms. We employ an adjacency matrix which stores the connectivity between joints. It is following the format of MSCOCO [80]. It preserves the relationship between connected joints and eliminates the unconnected ones. Let $\mathcal{W} \in \mathbb{R}^{N_p^2\times N_p^2}$ be a matrix of trainable parameters and $A_{adj} \in \mathbb{R}^{N_J\times N_J}$ be a adjacency matrix of a given skeleton. The GNN $\mathcal{G}$ as followings:

$$\mathcal{G}(A) = A_{adj} \cdot A \cdot \mathcal{W} \in \mathbb{R}^{N_J\times N_p^2}. \quad (6)$$

We named $\mathcal{G}(A)$ joint embeddings, since we compute attention to estimate joints of skeleton.

Consider a pixel on the snowboard in the Fig. 3 again. The attention between a pixel on the object and the joints should be higher than others. However, the joint that interacts with an object should also have a high degree of association. For this reason, our GNN is essential within the entire model.

Associative Attention Computation

The propagated joint embeddings are now used to compute attention score of each joint. We denote the output of the A.A. as $\mathcal{J}$ which is depicted as denoise process in Fig. 2 and Fig. 3. This process can be expressed with the following formula:

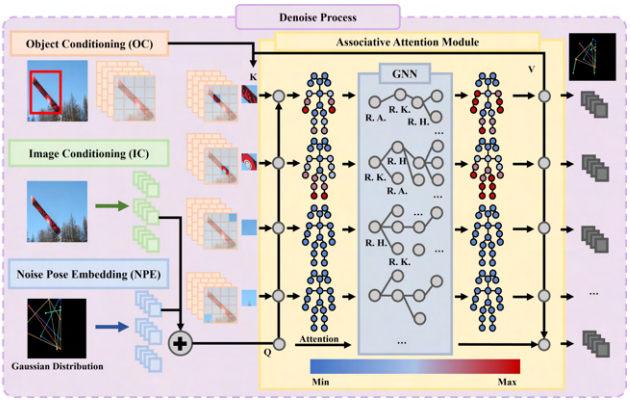

Figure 3. The denoise process first estimate the correlation between the object and the joints, and then it considers the relationship between the joint themselves using a GNN. After that, the object conditioning is used to predict which joints are most likely to interact with the object. The pixels located inside the snowboard have higher attention score on joints such as hands or foot.

$$\mathcal{J} = \text{softmax}(\mathcal{G}(A)) \cdot V \tag{7}$$

where $\mathcal{G}$ is a GNN. In Fig. 4, we illustrate the attention scores assigned to each joint. Notably, pixels associated with the tennis racket grip, marked in red, received higher attention for the hand joint compared to non-interacting parts. Additionally, neighboring joints that are not in direct contact with objects but interacting with the object showed higher attention scores relative to others far from interacting. For instance, the eyes have higher attention score due to their focus on the pixel highlighted in yellow. These observations suggest that employing a GNN could generate a more nuanced skeleton representation.

3.2.3 Iterative Denoising Process

The attention process described in Fig. 3 is repeated N_D times which is designed to gradually reduce noise from the skeleton. After the denoised skeleton is generated, it is used as an input to the next denoising block. This can be expressed with the following formula:

$$Q_i = I_{\text{condition}} + \mathcal{J}_{i-1}, \tag{8}$$

$$A_i = Q_i \cdot K^T, \tag{9}$$

$$\mathcal{J}_i = \text{softmax}(\mathcal{G}(A_i)) \cdot V. \tag{10}$$

We denote J_i be the i-th denoise embedding in Eq. 10 and A_i is the i-th alignment. Moreover, we formally write $J_0 = \text{NPE}$. In our experiments, we choose $N_D = 10$. In this way, our proposed method generates a skeleton interacting with an object than naïve attention methods. We will compare this in the experiment section.

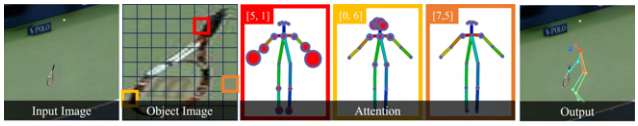

Figure 4. This figure visualizes which joint has the greatest association with features corresponding to selected pixels in the image colored red, yellow and orange. The size of the circle indicates the degree of association. The degree of association is computed as $\text{softmax}(\mathcal{G}(Q_{N_D} \cdot K^T))$ in Eq. 10.

3.3. Skeleton Guided Image Editing

In this section, we introduce how we edit the input image using the aforementioned generated skeleton. The image masked using a person bounding box and the skeleton generated from the previous module are used as an input for image editing. We use an off-the-shelf image inpainting model [81]. Existing inpainting models use an input image with a handmade skeleton image to fill predefined masked area by a person bounding box. However, our model directly uses the generated skeleton to inpaint the hole. In our experiment, we use ControlNet-Inpaint [81] as an inpainting model. The skeleton guided generation models could be altered into models such as HumanSD [40], T2I-Adpater [41] and Uni-ControlNet [42].

3.4. Network Training

From the feature extractor to object interactive diffuser with A.A. which creates an object-interactive skeleton. This is trained end-to-end. Our objective is defined as follows:

$$L_{HOI} = \lambda \times J_{param} \times L_{joint}^{init} + L_{conf}, \tag{11}$$

where λ is a hyper-parameter, L_{joint}^{init} is L1 distance between generated joints and GT joints, and L_{conf} is the L1 loss between predicted confidences and GT confidences. J_{param} is defined as

$$J_{param} = \text{softmax}\left(\frac{1}{\text{dist}(J_{GT}, \text{center}(\mathcal{B}_{object}))}\right). \tag{12}$$

where $\text{center}(\cdot)$ is a function which computes the center of a bounding box. This is designed to penalize loss as the distance from the center of the object bounding box to the GT joint increases, and conversely to reward loss as the distance decreases. We choose the Euclidean distance to measure a distance between the center of object bounding box and a joint location which is denoted as "$\text{dist}(\cdot,\cdot) \in \mathcal{M}$" in Eq. 12 where $\mathcal{M}$ is the metric space. We update the initial joint-wise loss using J_{param} as a scale factor. We use $\lambda = 10^4$ in our experiments.

4. Experiments

In this section, we quantitatively and quantitatively compare our method with existing methods, demonstrating the

effectiveness of our framework. We show that our A.A. module generates object-interactive skeletons well. Moreover, we conducted an ablation study comparing various methodologies with our method to demonstrate the necessity of our method. We use GT from V-COCO [82] training dataset and use LaMA [1] as an inpainting network. The details of user study, dataset and discussions on editing overlapped skeletons are on our supplementary materials.

4.1. Evaluation Metric

4.1.1 Metrics for assessing the quality of images

To quantitatively compare our framework with existing methods, we use Frèchet Inception distance (FID [83]), Kernel Inception distance (KID [84]) and CLIP score (CS [85]) as evaluation metrics. FID and KID measure how realistic generated images are in comparison to GT images. CS measures the extent to which the generated images are aligned with the text conditions. The detail explanations of these metrics are on our supplementary materials.

4.1.2 Metrics for assessing the quality of interaction

Object interaction top-n accuracy: This metric represents the extent to which the interacting joints in the generated image are similar to interacting joints in the real world. It is accuracy computed for each joint, such that it is one if n closest generated joints inside the object bounding box have the same index as the GT joint.

Skeleton Probability Distance (SPD): SPD measures the extent to which the joints interacting with an object are similar to the real world data per joints. The IoU of object bounding box and the bounding box covering joints is calculated. This IoU is computed for the bounding box covering GT joints and estimated joints, respectively. The size of bounding box is a manually defined. The joint-wise calculated IoUs are normalized by softmax. A distance between normalized joint-wise IoUs of GT and estimated joints is computed with Jensen-Shannon distance [86]. The SPD of bounding boxes of GT joints $\mathcal{B} = \{B_i\}$ and bounding boxes of predicted joints $\hat{\mathcal{B}} = \{\hat{B}_i\}$ is defined as:

$$\mathcal{B}_P = \text{softmax}(\text{IoU}(\mathcal{B}_{object}, \mathcal{B})), \tag{13}$$

$$\text{SPD}(\mathcal{B}, \hat{\mathcal{B}}; \mathcal{B}_{object}) = \text{dist}(\mathcal{B}_P, \hat{\mathcal{B}}_\mathcal{P}). \tag{14}$$

We discussed more of our SPD in the supplementary materials.

4.2. Quantitative Results

Table 1 shows quantitative results on various editing models. The average performance of text-guided editing model is better than that of text-free editing models. In addition, our method shows the best performance quantitatively. Our method uses the same diffusion backbone of SD-Inpainting and improved 4.12 in FID [83], 0.0033 in KID [84] and 1.24 in CS [85] than vanilla SD-Inpainting [26]. Moreover, SD-Inpainting [26] using our method outperforms SDXL-Inpainting [27] which is an enhanced model of SD-Inpainting [26]. This demonstrates the significance of our method in HOI image generation.

Table 1. **Quantitative results comparing our framework to previous image editing models**: Our framework outperforms others on the metrics indicating image quality FID [83], KID [84] and metric measuring prompt alignment to image CS [85].

Comparision Editing Model			
Evaluation Metric	FID [83] (↓)	KID [84] (↓)	CS [85] (↑)
Text-Free Editing Model			
LaMA [1]	59.30	0.0342	27.08
MAT [2]	77.55	0.0479	21.87
CoModGAN [3]	52.30	0.0282	26.18
Text-Guided Editing Model			
Instruct-Pix2Pix [18]	45.37	0.0200	28.44
MagicBrush [19]	60.01	0.0381	28.89
HIVE [20]	56.38	0.0346	27.70
Glide [25]	63.14	0.0344	25.70
BLDM [24]	25.52	0.0090	29.06
SDXL-Inpainting [27]	25.01	0.0082	29.63
SD-Inpainting [26]	28.16	0.0087	29.24
SD-Inpainting [81] + Ours	**24.04**	**0.0054**	**30.48**

4.3. Qualitative Results

Fig. 5 shows the qualitative results of image editing using ours and other models. We divided this figure into two parts: the upper and the lower parts. The upper part shows the results of generating a single person while the lower part shows the results of generating multiple people. The upper part shows other models generate incomplete or no humans using a textual prompt. Even when a human is generated, a human misaligned to a textual prompt or non-interactive human is synthesized.

Moreover, in case of editing multiple areas, SD-Inpainting [26] and SDXL-Inpainting [27] do not edit image properly. They tend to just fill in the inner areas using external information. This is demonstrated in the bottom of the figure. SD-Inpainting [26] does not generate any human, while SDXL-Inpainting [27] generates human in unsuitable size. However, our model edits the original image properly for object interaction. We experiment this task with three models on the same condition. Due to lack of space, additional qualitative results are in the supplementary materials.

4.4. Ablation study

In this section, we compare the results with and without our joint parameter on different methodologies. Moreover, we experiment with detailed designs for our object interactive diffuser. As a result of this experiment, we conclude our proposed method is best suited for the task.

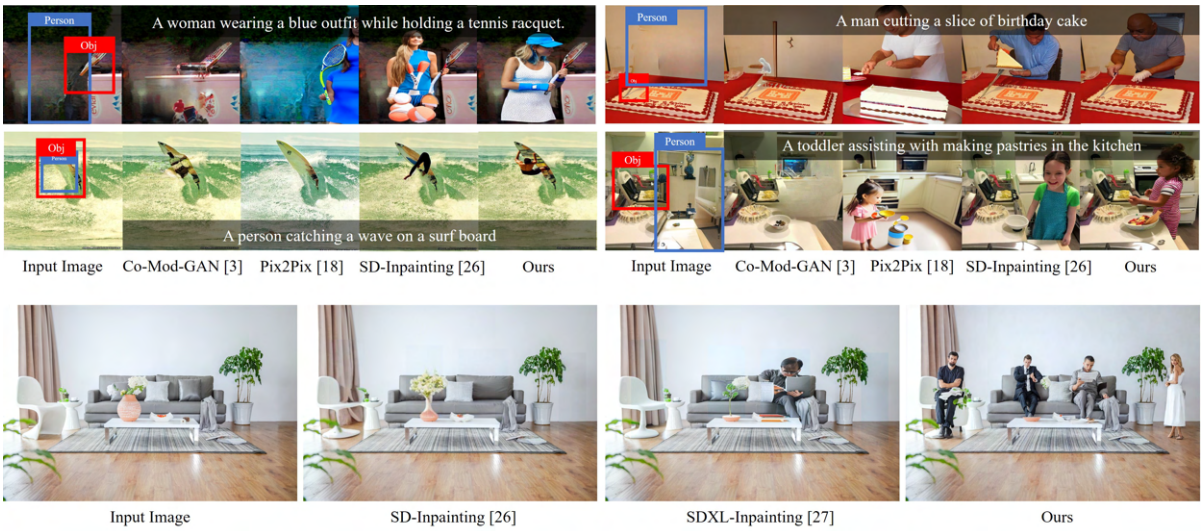

Figure 5. **Top**: Qualitative results when generating a single person using CoModGAN [3], Instruct-Pix2Pix [18], Stable-Diffusion Inpainting (SD-Inpainting) [26]. (Top left) Incomplete or no humans are generated using other models. (Top right) Even though humans are generated, the misaligned or non-interactive humans are synthesized. **Bottom** : Demonstration of image editing with SD-Inpainting [26], SDXL-Inpainting [27] and ours. Other models did not generate a human even using a guided skeleton.

Table 2. **Quantitative results with and without the proposed $Joint_{param}$ using different methodologies**: Numbers marked in red show the increments using our $Joint_{param}$.

ResNet [75] Backbone Comparision Using Our $Joint_{param}$				
Method	Object interaction			Skeleton
evaluation	Top 1 (↑)	Top 3 (↑)	Top 5 (↑)	distance (↓)
MLP (R 50)	58.9(+0.8) %	65.1(+0.5) %	68.2(+0.9) %	0.1336(-0.0032)
MLP (R 101)	60.8(+1.9) %	67.2(+1.9) %	68.6(+0.8) %	0.1309(-0.005)
MLP (R 152)	58.6(+1.8) %	64.8(+2.0) %	67.3(+1.9) %	0.1310(-0.003)
GNN (R 101)	58.6(+1.0) %	64.7(+0.8) %	67.2(+1.4) %	0.1312(-0.002)
Ours	**64.0(+1.5)** %	**69.3(+0.6)** %	**71.5(+0.3)** %	**0.1263(-0.003)**

Table 3. **Quantitative comparison based on structure of attention mechanisms**: Our A.A. module outperforms others using object conditioning. The method "Attention + GNN" stands for a method use attention to generate an initial skeleton and post-process using a GNN.

Comparison Specific Diffusion Module				
Method	Object Interaction			Skeleton
Evaluation	Top 1 (↑)	Top 3 (↑)	Top 5 (↑)	Distance (↓)
No Attention	61.4 %	67.4 %	69.6 %	0.128490
Attention	62.0 %	67.7 %	70.2 %	0.128255
Attention + GNN	60.6 %	67.1 %	69.3 %	0.129377
Associative Attention	**64.0** %	**69.3** %	**71.5** %	**0.126361**

4.4.1 Effect of scale factor

The experimental results demonstrating the effectiveness of our proposed joint parameter J_{param} are presented in Table. 2. To show that joint parameter is suitable for our task, even when applied to arbitrary models, we conducted experiments on three naïve models and our model. Even with the naïve models only consist of ResNet [75], the object interaction accuracy has been increased. Notably, the increment of top-1 accuracy with our method is remarkable. This is because the increase in this metric indicates that the proportion of joints most associated with the object has increased. In addition, the decrement of SPD demonstrate that our proposed J_{param} plays a role in bringing the interacting object and joints closer together.

4.4.2 Effect of attention mechanism

We conducted experiments from a methodological perspective shown in Table. 3. The model without using no attention, attention only, attention before GNN and ours. Our method outperforms others significantly. The highest top-1 accuracy and the SPD show that our method generates a skeleton which interacts better with objects. The detailed methodologies are provided in our supplementary materials.

Additionally, the comparison with and without our A.A. is shown in Fig. 6. Using our A.A., the joints that associate strong with the object get closer to the object than not using it. In addition, the generated skeletons are more natural because the relationship of each joint is considered using GNNs. For instance, Fig. 6 (a) which is the case that a skeleton is on the snowboard or skateboard, the holistic structure of a skeleton does not collapse using our proposed method. This is because, the A.A. contemplates the structural relationship of a skeleton using a GNN. Fig. 6 (b), the results from attention exhibit interaction with only one hand. However, our method generates skeletons of which both hands interact with object and generates more natural postures with object interaction. Fig. 6 (c) shows that joints required to interact with object get closer to it.

5. Applications

This section, we show that our framework could be extended or applied to various tasks. As shown in Fig. 7 (a),

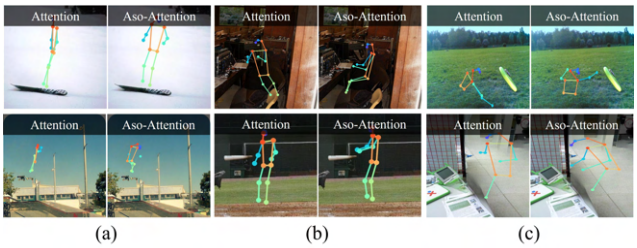

Figure 6. This figure shows the results with and without A.A. when generating a human skeleton. The "Attention" in figure means the attention module which is in Table. 3. Using A.A. the skeletons are generated towards the target object to interact with.

we show the potentials for expansion from object-to-human interaction to human-to-human interaction. We experiment with its applicability by simply changing an object bounding box to a person bounding box. We were able to get a skeleton who dances, step on people and surprise.

Additionally, we can manually rearrange the skeleton shown in Fig. 7 (b). Most editing models heavily rely on prompts so we have to modify prompts elaborately or might change random seeds until we get desirable results. However, using our framework we obtain an estimated skeleton from the network and users can manually rearrange the skeleton to obtain what they want. Therefore, more elaborate modification is possible. This alleviates heavy reliance on the prompt that existing editing models have. In addition, we experimented with automatic modification using PoseStylizer[48] as well as manually adjusting the skeleton. We discussed this in the supplementary materials.

Finally, obtaining 3D human mesh is possible shown in Fig. 7 (c). Owing to the recent development of image generative models, powerful data augmentation tools were used in face-related datasets [87]. This development has shown promise in a variety of task such as hand and human pose. However, most editing or generative models only rely on prompts to generate images. If the result is unsatisfying, users should accept or reject the output and there is no other option, since it is hard to change generated human in the image. They might compromise to use them even if there is a misalignment with the prompt. However, using our framework to optimize 3D human mesh with SMPLify [43], we would obtain a much elaborate and precise pseudo 3D human mesh dataset. This technique is a well-known method in the 3D human mesh estimation field. These extensive applications are our strength.

Our framework can be developed in various fields and is more practical than existing editing models. We believe that the development of this technology would be helpful on the field of computer vision in the future.

Figure 7. This figure shows three application cases using our framework. (a) demonstrates its extension to human-to-human interaction, (b) shows that manual editing of the predicted skeleton can enhance the image quality better. Moreover, (c) shows the results of 3D mesh optimization using SMPLify [43].

6. Conclusion

In this study, we proposed a method for generating associative skeleton guidance maps. It considered the relationship between an object and a skeleton to generate interactive skeleton guidance map. Through this method, it is possible to synthesize a human interacting with an object naturally. Moreover, it is easy to plug in skeleton guided generative models. We demonstrated that our method outperforms the others qualitatively and quantitatively. In addition, we showed the potentials of our framework in new applications.

7. 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), National R&D Program through the National Research Foundation of Korea(NRF) funded by Ministry of Science and ICT(2021M3H2A1038042) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1I1A1A01051225)

References

[1] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022. 2, 6

[2] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10758–10768, 2022. 6

[3] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428*, 2021. 2, 6, 7

[4] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[5] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022.

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[7] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European conference on computer vision (ECCV)*, pages 85–100, 2018.

[8] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4471–4480, 2019.

[9] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018.

[10] Raymond A Yeh, Chen Chen, Teck Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5485–5493, 2017. 2

[11] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European conference on computer vision*, pages 707–723. Springer, 2022. 2

[12] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021.

[13] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14297–14306, 2023.

[14] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022.

[15] Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22532–22541, 2023.

[16] Xingqian Xu, Zhangyang Wang, Eric Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. *arXiv preprint arXiv:2211.08332*, 2022.

[17] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *European Conference on Computer Vision*, pages 88–105. Springer, 2022.

[18] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 2, 6, 7

[19] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *arXiv preprint arXiv:2306.10012*, 2023. 6

[20] Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, et al. Hive: Harnessing human feedback for instructional visual editing. *arXiv preprint arXiv:2303.09618*, 2023. 2, 6

[21] David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. Paint by word. *arXiv preprint arXiv:2103.10951*, 2021. 2

[22] Katherine Crowson. Clip guided diffusion hq 256x256. *Colab Notebook. URL https://colab. research. google. com/drive/12a_Wrfi2_ gwwAuN3VvMTwVMz9TfqctNj*, 2021.

[23] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 2

[24] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transactions on Graphics (TOG)*, 42(4):1–11, 2023. 6

[25] 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. *arXiv preprint arXiv:2112.10741*, 2021. 2, 6

[26] Runway. Stable diffusion inpainting. In *https://huggingface.co/runwayml/stable-diffusion-inpainting*, 2022. 6, 7

[27] Suraj Patil. Sdxl inpainting. In *https://huggingface.co/spaces/diffusers/stable-diffusion-xl-inpainting/tree/main*, 2022. 2, 6, 7

[28] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2

[29] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Shiyi Wang, Hao-Shu Fang, Ze Ma, Mingyang Chen, and Cewu Lu. Pastanet: Toward human activity knowledge engine. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 382–391, 2020. 2, 3

[30] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. *Advances in Neural Information Processing Systems*, 35:14959–14971, 2022. 3

[31] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9365–9376, 2023. 3

[32] Jingwei Ji, Rishi Desai, and Juan Carlos Niebles. Detecting human-object relationships in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8106–8116, 2021.

[33] Hongsheng Li, Guangming Zhu, Wu Zhen, Lan Ni, Peiyi Shen, Liang Zhang, Ning Wang, and Cong Hua. Spatial parsing and dynamic temporal pooling networks for human-object interaction detection. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022.

[34] Ning Wang, Guangming Zhu, Liang Zhang, Peiyi Shen, Hongsheng Li, and Cong Hua. Spatio-temporal interaction graph parsing networks for human-object interaction recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4985–4993, 2021.

[35] Meng-Jiun Chiou, Chun-Yu Liao, Li-Wei Wang, Roger Zimmermann, and Jiashi Feng. St-hoi: A spatial-temporal baseline for human-object interaction detection in videos. In *Proceedings of the 2021 Workshop on Intelligent Cross-Data Analysis and Retrieval*, pages 9–17, 2021.

[36] Sai Praneeth Reddy Sunkesula, Rishabh Dabral, and Ganesh Ramakrishnan. Lighten: Learning interactions with graph and hierarchical temporal networks for hoi in videos. In *Proceedings of the 28th ACM international conference on multimedia*, pages 691–699, 2020.

[37] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 401–417, 2018. 3

[38] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8688–8697, 2019. 2

[39] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2, 3

[40] Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. Humansd: A native skeleton-guided diffusion model for human image generation. *arXiv preprint arXiv:2304.04269*, 2023. 2, 5

[41] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 3, 5

[42] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *arXiv preprint arXiv:2305.16322*, 2023. 2, 5

[43] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2252–2261, 2019. 2, 8

[44] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015. 2

[45] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018. 2

[46] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[47] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 3

[48] Siyu Huang, Haoyi Xiong, Zhi-Qi Cheng, Qingzhong Wang, Xingran Zhou, Bihan Wen, Jun Huan, and Dejing Dou. Generating person images with appearance-aware pose stylizer. In *IJCAI*, 2020. 3, 8

[49] Shuyuan Tu, Qi Dai, Zhi-Qi Cheng, Han Hu, Xintong Han, Zuxuan Wu, and Yu-Gang Jiang. Motioneditor: Editing video motion via content-aware diffusion. *arXiv preprint arXiv:2311.18830*, 2023.

[50] Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for referring human dance generation in real world. *arXiv preprint arXiv:2307.00040*, 2023. 3

[51] Sungnyun Kim, Junsoo Lee, Kibeom Hong, Daesik Kim, and Namhyuk Ahn. Diffblender: Scalable and composable multimodal text-to-image diffusion models. *arXiv preprint arXiv:2305.15194*, 2023. 3

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

[53] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*, 2023.

[54] Hongsuk Choi, Isaac Kasahara, Selim Engin, Moritz Graule, Nikhil Chavan-Dafle, and Volkan Isler. Finecontrolnet: Fine-level text control for image generation with spatially aligned text control injection. *arXiv preprint arXiv:2312.09252*, 2023. 3

[55] Liulei Li, Jianan Wei, Wenguan Wang, and Yi Yang. Neural-logic human-object interaction detection. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3

[56] Yichao Cao, Qingfei Tang, Xiu Su, Chen Song, Shan You, Xiaobo Lu, and Chang Xu. Detecting any human-object interaction relationship: Universal hoi detector with spatial prompt learning on foundation models. *arXiv preprint arXiv:2311.03799*, 2023.

[57] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. Hotr: End-to-end human-object interaction detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 74–83, 2021.

[58] Ting Lei, Fabian Caba, Qingchao Chen, Hailin Jin, Yuxin Peng, and Yang Liu. Efficient adaptive human-object interaction detection with concept-guided memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6480–6490, 2023.

[59] Yichao Cao, Qingfei Tang, Feng Yang, Xiu Su, Shan You, Xiaobo Lu, and Chang Xu. Re-mine, learn and reason: Exploring the cross-modal semantic correlations for language-guided hoi detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23492–23503, 2023.

[60] Frederic Z Zhang, Yuhui Yuan, Dylan Campbell, Zhuoyao Zhong, and Stephen Gould. Exploring predicate visual context in detecting of human-object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10411–10421, 2023. 3

[61] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14928–14940, 2023. 3

[62] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Imos: Intent-driven full-body motion synthesis for human-object interactions. In *Computer Graphics Forum*, volume 42, pages 1–12. Wiley Online Library, 2023. 3

[63] Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In *CVPR*, 2023. 3

[64] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *European Conference on Computer Vision*, pages 125–145. Springer, 2022.

[65] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15935–15946, 2022.

[66] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2282–2292, 2019.

[67] Xiang Xu, Hanbyul Joo, Greg Mori, and Manolis Savva. D3d-hoi: Dynamic 3d human-object interactions from videos. *arXiv preprint arXiv:2108.08420*, 2021. 3

[68] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 3

[69] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017. 3, 4

[70] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 3

[71] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514*, 2018. 3

[72] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 3

[73] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 3

[74] Diganta Misra, Trikay Nalamada, Ajay Uppili Arasanipalai, and Qibin Hou. Rotate to attend: Convolutional triplet attention module. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3139–3148, 2021. 3

[75] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 7

[76] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 4

[77] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao. Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. *arXiv preprint arXiv:2303.11579*, 2023. 4

[78] Hanbing Liu, Jun-Yan He, Zhi-Qi Cheng, Wangmeng Xiang, Qize Yang, Wenhao Chai, Gaoang Wang, Xu Bao, Bin Luo,

Yifeng Geng, et al. Posynda: Multi-hypothesis pose synthesis domain adaptation for robust 3d human pose estimation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5542–5551, 2023.

[79] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13147–13156, 2022. 4

[80] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014*, pages 740–755. Springer, 2014. 4

[81] Mikolaj Czerkawski. Stable diffusion controlnet inpainting. In *https://github.com/mikonvergence/ControlNetInpaint*, 2023. 5, 6

[82] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 6

[83] 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. *Advances in neural information processing systems*, 30, 2017. 6

[84] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 6

[85] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 6

[86] Dominik Maria Endres and Johannes E Schindelin. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860, 2003. 6

[87] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Matthew Johnson, Virginia Estellers, Thomas J. Cashman, and Jamie Shotton. Fake it till you make it: Face analysis in the wild using synthetic data alone. 2021. 8