EDITHOI: A FRAMEWORK FOR HOI IMAGE EDITING WITH SELF-GENERATED SKELETON GUIDANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, there were remarkable advances in image editing tasks. This could be categorized into text-guided global editing, local editing, text-guided local editing. To resolve the flawed human generation problem in prior image editing models, we propose a novel skeleton and text-guided local editing framework, EditHOI. Our goal is to edit an image by synthesizing an object-interactive human in the image. To do this, our framework consists of two stages: the first stage generates object-interactive skeleton using diffusion-based module, while the second stage outputs a Human and Object Interaction (HOI) image based on skeleton and text guidance. For effective evaluation on a object-interactive skeleton, we designed joint parameter and two evaluation metrics; object interaction top-n accuracy and skeleton probability distance. The excellent performance of our framework is demonstrated through experiments qualitatively and quantitatively. Lastly, we show its applicability such as user controllable editing, generating pseudo SMPL ground truth and scalability to human-to-human interaction. The corresponding code is available at `https://anonymous.4open.science/r/HOI_editing_image-43F1/`

1 INTRODUCTION

Imagine kicking a ball on a playground. You could kick the ball with your left or right foot. You might kick it gently like a pass, or you could kick it hard as if you aim to score a goal. If you edit an image of a soccer ball to become the image you imagine, can your imagination be the same as everyone else's? It is definitely not, since it is a highly ill-posed problem. There are plenty of possibilities how human could interact with objects. To realize your imagination among various scenarios, we define Human and Object Interaction(HOI) image editing task and proposed a novel framework.

Text-guided global editing [1, 2, 3] basically edit images using an input prompt and an image. Models designed for this task alter the style of an image, apply colorization or generate objects based on textual prompts. However, as illustrated in the first row of Figure 1, we observed absence of human and incomplete multi-person generation, which can be critical in HOI image editing. Next, local editing models [4, 5, 6] are designed to fill in masked areas in consideration of the contexts of their surroundings. We attempted HOI editing using local editing models with a bounding box which represents the expected location of a person. However, as depicted in the second row of Figure 1, absence of human is observed, which can lead to a critical issue in HOI editing. Text-guided local editing models [7, 8, 9, 10] fill in a mask of an image, using both surrounding contexts and a text prompt. We tried HOI editing with text-guided local editing models using a text prompt and a person's bounding box. Unlike former methods, its performance looks relatively fine. However, four problems are still observed in third row at Figure 1. First, absence of human that a human is not generated. Second, incomplete human generation that improper human is generated. Third, absence of interaction that a human not interacting with object is generated. Last, incomplete multi-human generation that more than two people are generated improperly. As stated above, we could find the flawed human generation problem in prior works, which our framework overcomes with additional skeleton guidance.

We propose a novel skeleton and text-guided local editing framework which generates a skeleton interacting with an object and then uses this skeleton to inpaint local area in the image. Our framework

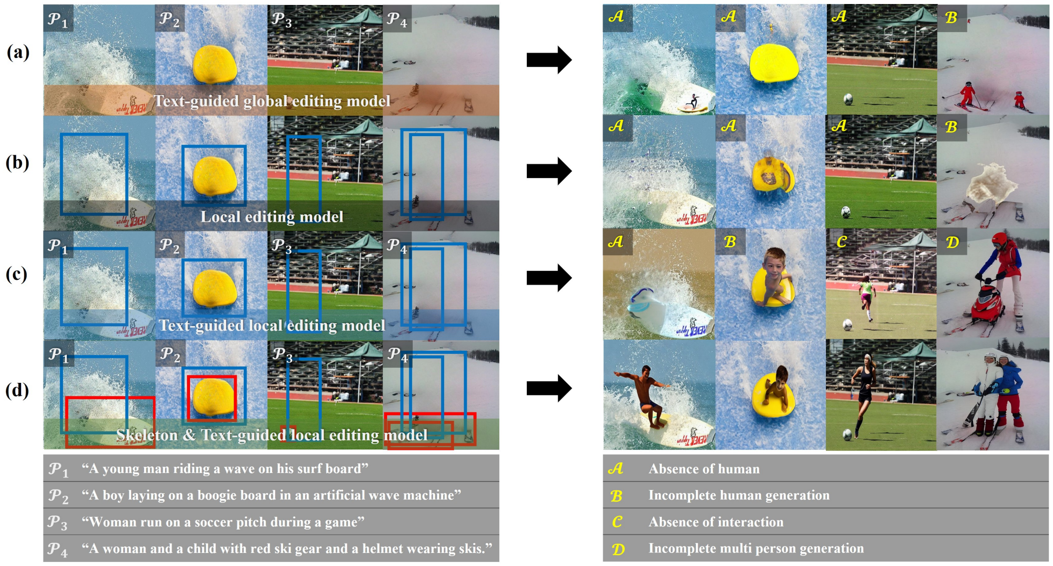

(a)
(b)
(c)
(d)
$\mathcal{P}_1$
$\mathcal{P}_2$
$\mathcal{P}_3$
$\mathcal{P}_4$
Text-guided global editing model
Local editing model
Text-guided local editing model
Skeleton & Text-guided local editing model
$\mathcal{A}$
$\mathcal{B}$
$\mathcal{C}$
$\mathcal{D}$
$\mathcal{P}_1$ "A young man riding a wave on his surf board"
$\mathcal{P}_2$ "A boy laying on a boogie board in an artificial wave machine"
$\mathcal{P}_3$ "Woman run on a soccer pitch during a game"
$\mathcal{P}_4$ "A woman and a child with red ski gear and a helmet wearing skis."
$\mathcal{A}$ Absence of human
$\mathcal{B}$ Incomplete human generation
$\mathcal{C}$ Absence of interaction
$\mathcal{D}$ Incomplete multi person generation

Figure 1: **Comparison of previous models with ours** : Images edited by (a) text-guided global editing, (b) local editing, (c) text-guided local editing, and (d) ours(skeleton and text-guided local editing). In the left Figure, the input condition and prompts(P_1, P_2, P_3, P_4) are visualized. As shown in the right Figure, four problems, such as (A) Absence of human, (B) Incomplete human generation, (C) Absence of Interaction, (D) Incomplete multi-person generation, are observed in existing models. Our results exhibit more object-interactive human images.

consists of two stages. At first stage, a skeleton is generated with a diffusion-based object interaction skeleton generation module. Unlike fields of HOI classification [11, 12] or scene-interactive human motion generation [13, 14], in which only restricted indoor scenes are applicable, our approach can utilize various in-the-wild images to generate a skeleton interacting with object. At the second stage, the skeleton-guided editing model synthesizes object-interactive human on the masked image, using skeleton of first stage and text prompt.

Our approach solves these four existing problems in Figure 1. In the left side of Figure 1, the last row exhibits edited images using our framework. The first image shows that our framework have solved the absence of human problem that a human is not generated. This is because our framework directly generate a skeleton guidance. The second image shows that the incomplete human generation problem does not occur using our framework. The third image demonstrates the problem of absence and incomplete interaction is solved using our framework. The fourth image shows that our framework solved incomplete multi-person generation problem.

We also discover the potentials of our skeleton and text-guided local editing framework: EditHOI. In existing methods, there are two options for users when the results of image editing are unsatisfying. First, adjusting random seed iteratively until they obtain a satisfying output. Second, trying various prompts until the desired image is generated. Compared to previous works, our framework is more controllable, allowing users to edit the self-generated skeleton as the way they want. This makes it possible to generate a user-desired output using EditHOI. Additionally, more aligned pseudo SMPL [15] ground truth can be generated, since it can be optimized by SMPLify [16] using our self-generated skeleton. Moreover, a skeleton generated by our framework can be applied to human-to-human interaction.

We summarize our contributions below.

- We are the first to address the task of HOI image editing, synthesizing object-interacting realistic human on an image containing objects.
- We propose a novel skeleton and text-guided local editing framework, EditHOI. Our framework solved four problems in HOI image editing, demonstrated to outperform existing image editing models through experiments quantitatively and qualitatively.
- We suggest a diffusion-based object interaction module which generates object-interactive skeletons by itself. Additionally, we introduce new metrics and joint parameters for effec-

tive evaluation on a object-interactive skeleton. The effectiveness of the module and joint parameters is shown in ablation study.

- Our self-generated skeleton could be applied in various ways. Since our framework consists of two stages, users can choose to use the self-generated skeleton without modifications or manually edit it in order to get the desired output. More aligned pseudo SMPL [15] ground truth optimized by SMPLify [16] can be constructed with our self-generated skeleton. In addition, there appears to be the potential of scalability to human-to-human interaction

2 RELATED WORKS

Text-guided global editing : Text-guided global editing is a task which modifies the whole input image using text prompts. Specially, text prompts are provided in the form of a single noun or a combination of multiple words [17, 18, 19, 20], a sentence [21, 22, 23], and an instruction[1, 2, 3]. StyleCLIP [18] applied Contrastive Language-Image Pre-training(CLIP) models to StyleGAN, which enables intuitive text-guided image editing without additional manual controls. On the other hand, VQGAN-CLIP [23] is the first work to introduce a unified framework for both semantic image generation and image editing based on text prompts. Furthermore, Text2LIVE [17] extended its work of image editing to video. Rather than directly generating the output image, its key idea is to generate an edit layer which can be synthesized over the original image. However, as Text2LIVE [17] is designed to edit existing objects, it shows certain limitations in generating new objects. Instruct-Pix2Pix [1] edits an input image using user-provided instructions which inform the model of what to do. In the process of preparing a large-sized dataset of image editing examples, it integrated a language model (GPT-3) and a text-to-image model (Stable Diffusion) [24]. In addition, MagicBrush [2] introduces the first large-scale and manually annotated dataset for instruction-guided real image editing. It fine-tuned aforementioned Instruct-Pix2Pix [1] on MagicBrush [2] and demonstrated their new model achieves better results through human evaluation. A framework to utilize human feedback in instruction-guided image editing was introduced by Hive [3]. It obtained human feedbacks from annotators and fine-tuned diffusion models based on collected human preferences. Despite the significant advances in text-guided global editing, previous works are limited to replacing objects, changing the color of an image or modifying the background. Creating new objects remains either impossible or poorly processed in text-guided global editing.

Local editing : Local editing is a task which aims to edit the input image locally, filling in the masked or removed space in the image. Numerous works have achieved high-quality image synthesis quality with no guidance except for the surrounding contexts in the image, also known as inpainting [4, 5, 6, 25, 26, 24, 27, 28, 29, 30]. In the early stages of applying deep learning to inpainting task, [29] proposed a generative model that utilizes surrounding context around masks. With free-form mask and guidance, [28] presents a generative image inpainting network based on gated convolutions. For handling large-scale masks, CoModGAN [6] proposes co-modulated generative adversarial networks, a new method to reduce the gap between image conditional and unconditional GANs [31]. While existing inpainting models lack a large effective receptive field, LAMA [4] suggests an architecture called large mask inpainting (LaMa). MAT [5] integrates the merits of transformers and convolutions for large hole inpainting and high-resolution image generation. Similar to text-guided global editing, it is nearly impossible to create new object by local editing models, as they rely only on the surrounding contexts to fill in the missing region.

Text-guided local editing : Text-guided local editing is a relatively recently introduced task in computer graphics, filling in missing regions of an input image in consideration of both the surrounding context and additional textual descriptions [32, 33, 34, 8, 7, 10, 9]. Paint by Word [32] is the first method to investigate the problem of local zero-shot semantic image editing by pairing CLIP [35] with StyleGAN2 [36] and BigGAN [37]. GLIDE [7] utilizes diffusion models in two-stage approach for text-guided local editing : the first text-conditional diffusion model generates a row-resolution version of image, while the second stage processes upsampling using both the low-resolution version and the text prompt. SD-Inpainting [10] is a inpainting version of Stable Diffusion, while advanced Stable Diffusion XL [38] performs inpainting in SDXL-Inpainting [9] as well. BDM [34] was suggested by Avrahami et. al., performing local editing based on a textual description and an ROI [39] mask. It combined a Contrastive Language-Image Pre-training (CLIP) [35] model and a Denoising Diffusion Probablistic Model (DDPM) [40] to control the edit using a user-provided text prompt and generate generic natural images. Developing the prior work, Avrahami et. al proposed BLDM [8] ,

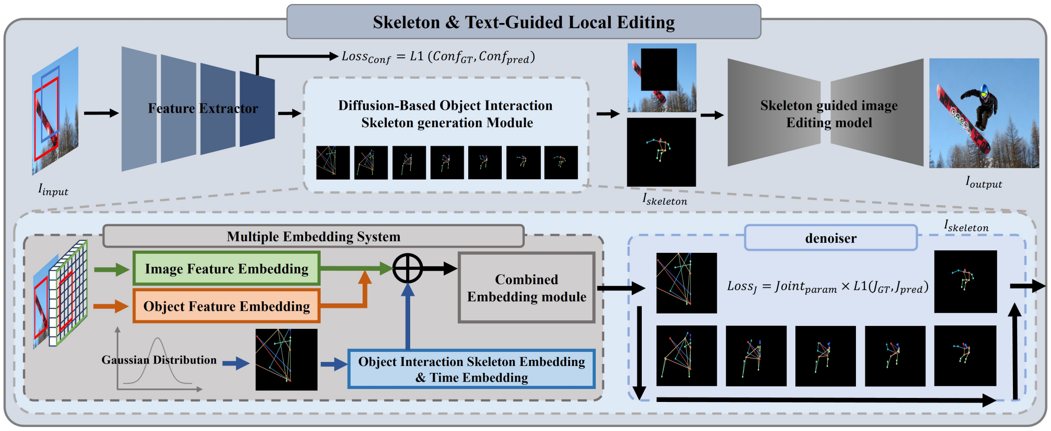

Skeleton & Text-Guided Local Editing
$Loss_{Conf} = L1\ (Conf_{GT}, Conf_{pred})$
Feature Extractor
Diffusion-Based Object Interaction Skeleton generation Module
Skeleton guided image Editing model
I_{input}
$I_{skeleton}$
I_{output}
Multiple Embedding System
Image Feature Embedding
Object Feature Embedding
Combined Embedding module
Gaussian Distribution
Object Interaction Skeleton Embedding & Time Embedding
denoiser
$I_{skeleton}$
$Loss_J = Joint_{param} \times L1(J_{GT}, J_{pred})$

Figure 2: **Overview of proposed framework** : Our framework uses ResNet [41] backbone to extract features of an input image. In multiple embedding system, feature embeddings of an image and an object are embedded through an embedding network, using bounding boxes of a person and an object. Merging these embeddings with a noisy skeleton sampled from Gaussian distributions, combined embedding module feed the merged output to the denoiser network. Finally, a denoiser network reconstructs a skeleton interacting with an object.

an architecture with a text-to-image Latent Diffusion Model (LDM) [24] that works much faster than BDM [34]. Even though text-guided local editing is advancing rapidly, existing works still take a lot of time, making them hard to apply to real-time applications. More seriously, they fail to interpret text prompts describing not-familiar objects or scenarios, which causes failure in human and object interaction addressed in our paper.

Unlike previous works mentioned above, we suggest a framework employing additional skeleton guidance interacting with objects in the input image. With our method, higher-quality images can be generated than with prior works.

3 PROPOSED METHOD

At this section, we explain the overall process of our skeleton and text-guided local editing framework. Given an image, bounding boxes of a person and an object, a skeleton guidance is generated after passing them through a feature extractor and diffusion-based object interaction skeleton estimation module. Using the self-generated skeleton with the aforementioned inputs, our framework outputs the edited image which contain a human interacting with objects

3.1 SKELETON & TEXT-GUIDED LOCAL EDITING ARCHITECTURE

At this section, we propose a two-stage architecture for HOI image editing. The overall framework is visualized in Figure 2. and it consists of feature extractor, diffusion-based interaction skeleton estimation module and skeleton guided image editing model.

Feature extractor : We use a bounding box to define boundaries of an object interacting with a person. Without specifying locations of a person and an object, the network would struggle with where to locate the person and the object. Therefore, shown at the left side of Figure 2., we locate a person's bounding box $\mathcal{B}_{person} \in \mathbb{R}^{1\times4}$ as a hard decision. After that, the input image is fed to the backbone network.

Extracting a feature map from a backbone network, we obtain confidence score $Conf_{pred}$ which indicates whether each skeleton is visible or not. The ground truth confidence score of joints is 0 if the joint is invisible and 1 if the joint is visible. The predicted confidence is supervised by the ground truth confidence as followings:

$$Loss_{conf} = |p - \hat{p}| \tag{1}$$

where, p and $\hat{p}$ indicate the ground truth confidence and estimated confidence respectively.

Diffusion-based object interaction skeleton estimation module : The feature map $F_{backbone}$ is fed through the multiple embedding system (MES). Shown in the left bottom of Figure 2., an image embedding $E_{image} \in \mathbb{R}^{17\times32}$ and an object embedding $E_{object} \in \mathbb{R}^{17\times32}$ are obtained from the feature map. We apply a region of interest (ROI [39]) pooling to the feature map with the object bounding box $\mathcal{B}_{object} \in \mathbb{R}^{1\times4}$ to obtain $F_{object} \in \mathbb{R}^{2048\times N\times N}$ by the object feature embedding network. Next, using Gaussian distribution, we generate noisy joint $\mathcal{J}_{noised} \in \mathbb{R}^{17\times2}$. Adding those embeddings and $\mathcal{J}_{noised}$ together, we obtain $E_{skel} \in \mathbb{R}^{17\times32}$. And we add time embedding for timestep T to get $E_{time} \in \mathbb{R}^{17\times32}$. All these embeddings together, using combined embedding module $\mathcal{C}$, we obtain object interaction noised skeleton $\mathcal{J}_{embedding} \in \mathbb{R}^{17\times2}$.

$$\mathcal{J}_{embdding} = \mathcal{C}(E_{image} \oplus E_{object} \oplus E_{skel} \oplus E_{time}) \tag{2}$$

A denoiser network gradually denoise $\mathcal{J}_{embedding}$ with timestep T to obtain a object interactive skeleton $J_{pred} \in \mathbb{R}^{17\times2}$. An initial joint-wise L1 loss would guide predicted joints close to the ground truth joints. Moreover, we do not use this initial joint-wise L1 loss in naive manner. We consider how close the joint in ground truth is to the object bounding box, using our joint distance parameter $Joint_{param}$ as below:

$$Loss^{\text{init}}_{joint} = |J - \hat{J}| \tag{3}$$

$$Joint_{param} = \text{softmax}\left(\frac{1}{\text{dist}(J, \text{center}(\mathcal{B}_{object}))}\right) \tag{4}$$

$$Loss_{joint} = \lambda \times Joint_{param} \times Loss^{\text{init}}_{joint} \tag{5}$$

where J and $\hat{J}$ indicates the ground truth joint and the predicted joints respectively. And center$(\cdot)$ is a function which computes the center location of a bounding box. We use Euclidean distance to measure distance between the center of the object bounding box and a joint location. We update initial joint-wise loss using $Joint_{param}$ as a weight. λ is a scale factor we set $\lambda = 10^{-4}$ in experiments.

After that, we choose D3DP [42] as a diffusion structure which reconstruct noisy joint at timestep T.

Skeleton-guided Image Editing Model : In this stage, we edit an input image $\mathcal{I} \in \mathbb{R}^{H\times W\times3}$ using $\mathcal{I}_{skel} \in \mathbb{R}^{H\times W\times3}$ and $\mathcal{B}_{person}$ as additional conditions. We modify the predicted skeleton from the previous network to MSCOCO [43] format. We note that skeleton guidance to $\mathcal{I}_{skel}$.

Unlike previous image editing model, we use generated $\mathcal{I}_{skel}$ as a condition. Using our $\mathcal{I}_{skel}$, we could solve the aforementioned four problems of previous editing models. In addition, users could modify our predicted joints manually which is impossible in previous models. Moreover, changing a object bounding box to a person bounding box generates skeletons interacting each other. Using the predicted skeleton to generate pseudo SMPL [15] ground truth demonstrates its applicability. We employ ControlNet-Inpainting [44] as the skeleton guided imaged editing model.

4 EXPERIMENTS

At this section, we compare our method with existing methods in qualitative and quantitative ways, demonstrating the effectiveness of our framework. We conduct ablation study to show the effectiveness of our newly developed $Joint_{param}$ for object interaction. Moreover, we experiment various methods to obtain a human skeleton: (1) using feature embeddings of an image and an object and MLP, (2) using these embeddings and GNN, (3) using these embeddings, Gaussian noise and diffusion algorithm. We demonstrate that (3) works most effectively compared to the others. Implementation details are on our supplementary materials.

4.1 DATASETS

V-COCO [45] is well-known dataset in HOI field. Different from datasets such as HICO [46] and Bongard-HOI [47], it contains ground truths of segmentation, skeletons and a person's bounding

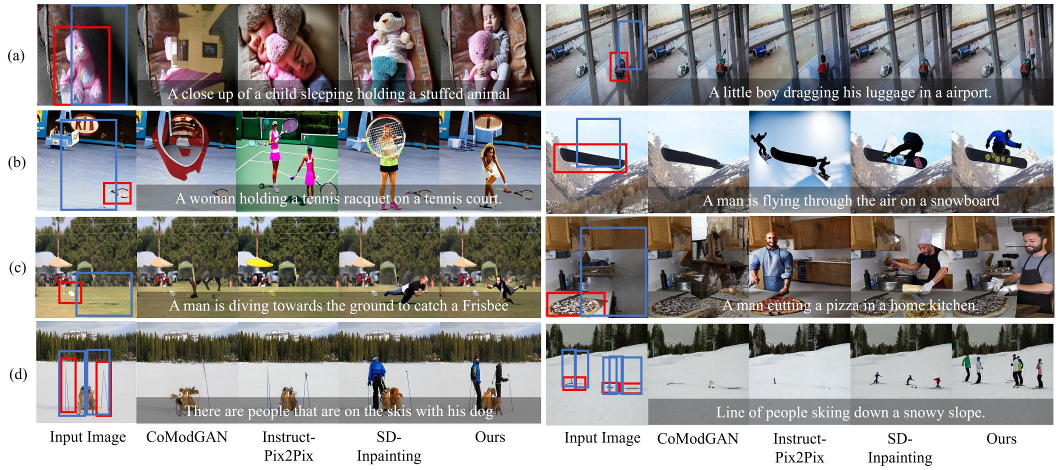

(a)
A close up of a child sleeping holding a stuffed animal
A little boy dragging his luggage in a airport.
(b)
A woman holding a tennis racquet on a tennis court.
A man is flying through the air on a snowboard
(c)
A man is diving towards the ground to catch a Frisbee
A man cutting a pizza in a home kitchen.
(d)
There are people that are on the skis with his dog
Line of people skiing down a snowy slope.
Input Image
CoModGAN
Instruct-Pix2Pix
SD-Inpainting
Ours
Input Image
CoModGAN
Instruct-Pix2Pix
SD-Inpainting
Ours

Figure 3: This figure shows comparison of three editing frameworks to ours. CoModGAN [6], Instruct-Pix2Pix [1], Stable-Diffusion Inpainitng (SD-Inpainting) [10] were used for comparison. The results of CoModGAN[6] and Instruct-Pix2Pix [1] failed to generate a human in most cases. In edited images using SD-Inpainting [10], aforementioned four problems occurred in order: (a) absence of human, (b) incomplete human generation, (c) absence of interaction, (d) incomplete multi-person generation

box. We made a masked area based on the segmentation ground truth of a person and filled the mask using LAMA [4]. To select images containing HOIs, we collected images of which Intersection over Union (IoU) are greater than zero. We used V-COCO [45] protocol for training and testing.

4.2 Evaluation Metric

To quantitatively compare our framework with existing methods, we use Frèchet Inception distance (FID [48]), Kernel Inception distance (KID [49]) and CLIP score (CS [50]) as evaluation metrics.

Frèchet Inception Distance (FID [48]) : FID [48] aims to compare the distributions of generated images to images from a real dataset. Assuming two datasets follow Gaussian distribution $\mathcal{N}(\mu, \Sigma), \mathcal{N}\hat{\mu}, \hat{\Sigma})$, FID [48] is defined as:

$$\text{FID}(\mathcal{N}(\mu, \Sigma), N(\hat{\mu}, \hat{\Sigma})) = ||\mu - \hat{\mu}||_2^2 + \text{Tr}(\Sigma + \hat{\Sigma} - 2(\Sigma\hat{\Sigma})^{1/2}) \tag{6}$$

Kernel Inception Distance (KID [49]) : KID [49] measures the squared maximum mean discrepancy (MMD) between the feature of inception network of the real and generated images using a polynomial kernel. Since it is a non-parametric test, it does not need the strict Gaussian assumption.

CLIP score (CS [50]) : CS [50] measures the extent to which the generated images are aligned with the text conditions. In precise manner, it is a metric that represents the extent to which a text condition matches an images without relying on human annotations. Let I be an input image, C be a corresponding text condition, and E_I, E_C be embeddings within the image and text condition, respectively. Then, the CLIP score [50] is defined as follows :

$$\text{CLIPScore}(C, I) = \max(100 \times \cos(E_C, E_I), 0) \tag{7}$$

where the CLIP score [50] is between [0, 100].

FID [48] and KID [49] are indicators of how generated image is realistic. And CS [50] measures how well synthesized image is well aligned with a prompt describing interactions. Moreover, to evaluate the generated skeleton interacting with object, we define two evaluation metrics.

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. Specifically, it is 1 when the predicted joint is inside the object bounding box where its index is same as the n^{th} closest joint in the ground truth skeleton to the object bounding box and 0 otherwise.

Skeleton Probability Distance (SPD) : SPD measures the extent to which the joints interacting with an object are similar to the real world data. The IoU of object bounding box and the bounding box covering joints is calculated. This IoU is computed for the bounding box covering ground truth joints and estimated joints. The size of bounding box is a manually defined. The joint-wise calculated IoUs are normalized by softmax. And a distance between normalized joint-wise IoUs of ground truth and estimated joints is computed with Jensen-Shannon distance [51]. The SPD of bounding boxes of ground truth joints $\mathcal{B} = \{B_i\}$ and bounding boxes of predicted joints $\hat{\mathcal{B}} = \{\hat{B}_i\}$ is defined as:

$$\text{SPD}(\mathcal{B}, \hat{\mathcal{B}}; \mathcal{B}_{object}) = \text{dist}(\text{softmax}(\text{IoU}(\mathcal{B}_{object}, \mathcal{B}), \text{softmax}(\text{IoU}(\mathcal{B}_{object}, \hat{\mathcal{B}}))) \tag{8}$$

4.3 Quantitative Results

Table. 1 shows quantitative results on various editing models. Text-guided local editing and skeleton and text-guided local editing models show the best performance on average among text-guided global editing, local editing, text-guided local editing, skeleton and text-guided local editing models. Our framework skeleton and text-guided local editing model outperform others. Our framework uses the same diffusion backbone of SD-inpainting and improved 4.14 in FID [48], 0.0035 in KID [49] and 1.22 in CS [50] than vanilla SD-inpainting [10]. Moreover, SD-inpainting [10] using our framework outperforms SDXL-inpainting [9] which is an enhanced model of SD-inpainting [10]. This demonstrates the significance of our framework in HOI image generation.

Table 1: **Quantitative results comparing our framework to the previous image editing model** : our framework outperform others on the metrics indicating image quality FID [48], KID [49] and metric measuring prompt alignment to image CS [50].

Comparision Editing Model			
Evaluation Metric	FID [48] (↓)	KID [49] (↓)	CS [50] (↑)
Text-Guided Global Editing Model			
Instruct-Pix2Pix [1]	45.37	0.0200	28.44
MagicBrush [2]	60.01	0.0381	28.89
HIVE [3]	56.38	0.0346	27.70
Local Editing Model			
LAMA [4]	59.30	0.0342	27.08
MAT [5]	77.55	0.0479	21.87
CoModGAN [6]	52.30	0.0282	26.18
Text-Guided Local editing model			
Glide [7]	63.14	0.0344	25.70
BLDM [8]	25.52	0.0090	29.06
SDXL-Inpainting [9]	25.01	0.0082	29.63
SD-Inpainting [10]	28.16	0.0087	29.24
Skeleton & Text-guided Local editing model			
SD-Inpainting [44] + Ours	**24.02**	**0.0052**	**30.46**

4.4 QUALITATIVE RESULT

Figure 3. is the qualitative comparison between our models to Instruct-Pix2Pix [1] which is a text-guided global editing model, CoModGAN [6] which is a local editing model, stable diffusion inpainting (SD-Inpainting) [10] which is text-guided local editing model. In most cases, CoModGAN [6] and instruct pix2pix [1] do not generate human properly. So we concentrate on comparing with SD-inpainting [10]. (a) shows the absence of humans in generated images. In the case of SD-inpainting [10], a child and a little boy are contained in the prompts but no human is generated. (b) shows the generation of incomplete or awkward human and object interaction. In the case of SD-inpainting [10], the generated tennis racket and a person overlapped which could not happen in the real world. On the right side of (b) only black objects are generated except for ours. (c) shows the results of a human and an object not well interacting. In the case of SD-inpainting [10], there is no interaction between a person and a frisbee but ours interact well. In addition on the right side of (c), despite the prompt including the phrase ”cutting a pizza” our image is the only one that generates proper interaction. Last, (d) shows the examples of other models that do not generate multi-person properly. However, our model uses additional explicit skeleton guidance so that generates natural images with multi-person. Additional comparisons between SDXL and ours are on supplementary materials.

4.5 ABLATION STUDY

Table. 2 is the results from our model in the presence and absence of our proposed $Joint_{parm}$. We experiment with ResNet [41] 50, 101, 152 as the backbone network with using a object bounding

Table 2: **Quantitative results on the absence and presence of proposed** $Joint_{param}$: Left side of the arrow shows the result without using $Joint_{param}$ while right side shows the results using it. Overall results enhanced using our proposed parameter.

ResNet [41] Backbone Comparision Using Our Hyper parameter				
Backbone	Object interaction			Skeleton
evaluation	Top 1 (↑)	Top 3 (↑)	Top 5 (↑)	distance (↓)
ResNet [41] 50	58.1 → 58.9 %	64.6 → 65.1 %	67.3 → 68.2 %	0.135838 → 0.132561
ResNet [41] 101	58.9 → 60.8 %	65.3 → 67.2 %	67.8 → 68.6 %	0.131394 → 0.130857
ResNet [41] 152	56.8 → 58.6 %	62.8 → 64.8 %	65.4 → 67.3 %	0.133987 → 0.131039

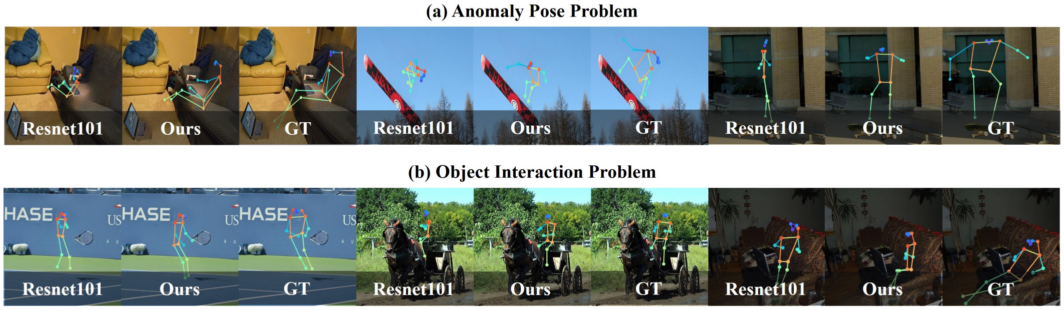


Figure 4: This figure shows the results of absence and the presence of our diffusion-based object interaction skeleton estimation module generating a human skeleton. (a) show the problem that generating a skeleton where joints are squeezed which is an anomaly pose. (b) shows the problem of generating invalid interaction with an object.

box. Every results in object interaction top 1, 3, 5 have increased for all backbone. Moreover, skeleton probability has been increased through all backbone networks.Especially ResNet [41] 101 scores highest in perspective of object interaction. After these, we experiment predicting joints with MLP, GNN and diffusion with image and object embeddings using the object bounding box. Shown in Table. 3, the results using MLP and GNN are worse than the naive models in Table. 2 As a result, the model using diffusion achieve the best results among them. We not only quantitatively compare them but also visualize the results at Figure 4. We visually compare our model with the model using ResNet [41] 101, since the model using ResNet [41] 101 is the best among others except for our model.

Two major problems are shown in Figure 4. on the models without using our framework using the ResNet [41] 101 as a backbone network. First is the anomaly pose problem. Second is the object interaction problem. Our framework using diffusion module, gradually de-noise on the noisy skeleton so that obtain plausible skeleton. However, using only the ResNet [41] 101 to predict joint shows non-interactive skeletons generation shown at the Figure 4. bottom, since its architecture is not complex enough to consider object embeddings well. Theses shows that our diffusion-based object to object interaction human pose estimation module is effective.

Table 3: **Quantitative results applying** E_{object} **on various methods** : using diffusion based method shows the best quantitative results than others.

Comparison Object Embedding Module				
Method	Object interaction			Skeleton
evaluation	Top 1 (↑)	Top 3 (↑)	Top 5 (↑)	distance (↓)
MLP	60.8%	67.2%	68.6%	0.130857
GNN	58.6%	64.7%	67.2%	0.131156
Diffusion	**62.6%**	**68.2%**	**70.6%**	**0.126317**

4.6 User Study

We survey using the image generated by the text-guided global editing model, local editing model, text-guided local editing model and our proposed framework on image generation quality, prompt relevance, image with the best editing and image well interacting with the object. 83.2% of people agree that the image quality generated with our framework is better than others. And 85.7% of people think that our framework has the highest prompt relevance. 86.7% of people think our framework shows the best image editing quality. 83.6% of people agree that our framework is the best model showing plausible interaction with an object. Considering that these four criteria have significant meaning in image editing, we conclude that 84.8%

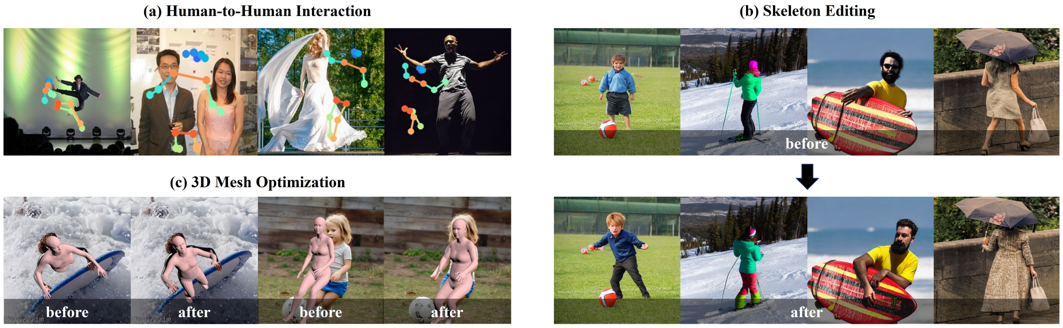


Figure 5: This figure shows three application cases using our framework. (a) demonstrates its extension to human-to-human interaction, (b) shows that manual editing of the predicted skeleton could enhance the image quality better. And (c) shows the results of 3D mesh optimization using SMPLify [16].

5 APPLICATIONS

In this section, we show that our framework could be extended or applied to various tasks. Shown in Figure 5. (a), we confirmed the potential for expansion from object-to-human interaction to human-to-human interaction. We experiment with its possibility by simply changing the object bounding box to the person bounding box. We were able to get a skeleton who dances, step on people and surprise. Using this skeleton guidance, we generated images with our framework but failed to synthesize plausible images. This is because using a person bounding box for masking eliminates most part of the given images. So, there would be not enough information to infer. Additionally, under trained diffusion model might be the reason. We left these problems to our future work to solve.

Next, we could manually edit the skeleton shown in Figure 5. (b). Most editing models heavily rely on prompts so we have to modify prompts elaborately or might change random seeds until get what we want. However, using our framework we obtain an estimated skeleton from the network and users could manually modify these skeletons to what they want. So, a more elaborate modification is possible. This solves the heavy reliance on the prompt that existing editing models have.

Finally, obtaining 3D human mesh is possible shown in Figure 5. (c). Owing to the recent development of image generative models, powerful data augmentation tools were used in face-related datasets [52]. These 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. This is the critical problem of an existing model. Because if the result is unsatisfying then users should accept or reject the output and there is no other option. Or 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 [16], 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 could be developed in various fields and is more practical than existing editing models. We believe that the development of this technology will have a huge impact on the field of computer vision in the future.

6 CONCLUSIONS

In this study, we define a HOI image editing task and propose a novel framework for HOI image editing, EditHOI. Our framework solves the four critical problems in existing editing models by generating skeleton guidance to edit an image by itself. We demonstrated that our framework outperforms than the others in quantitative and qualitative ways. In addition, we show the potentials of our framework in new applications. Although our framework is still limited to object interaction, it can be applied to the fields of human-to-human and human pose estimation in future developments.

REFERENCES

[1] 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] 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.

[3] 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.

[4] 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.

[5] 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] 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.

[7] 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.

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

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

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

[11] Sanghyun Kim, Deunsol Jung, and Minsu Cho. Relational context learning for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2925–2934, 2023.

[12] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. Exploring structure-aware transformer over interaction proposals for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19548–19557, 2022.

[13] 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.

[14] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Chairs: Towards full-body articulated human-object interaction. *arXiv preprint arXiv:2212.10621*, 2022.

[15] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

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

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

[34] 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.

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[36] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[37] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[38] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

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

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

[41] 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.

[42] 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.

[43] 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.

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

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

[46] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE international conference on computer vision*, pages 1017–1025, 2015.

[47] Huaizu Jiang, Xiaojian Ma, Weili Nie, Zhiding Yu, Yuke Zhu, and Anima Anandkumar. Bongard-hoi: Benchmarking few-shot visual reasoning for human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19056–19065, 2022.

[48] 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.

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

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

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

[52] 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.

[53] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[54] Zijun Zhang. Improved adam optimizer for deep neural networks. In *2018 IEEE/ACM 26th international symposium on quality of service (IWQoS)*, pages 1–2. Ieee, 2018.

A Supplementary materials

A.1 detailed of user study

Figure 6 shows a comparison of text-guided local editing, text-guided global editing, local editing, and ours. It visualizes the results for image quaility prompt relevance, image with the best editing, and image well interacting with object. We could see that our framework overwhelmingly outperformed the other framewors. Moreover, Fig 7 shows a questionnaire via goole-form. (a), (b), (c) and (d) show a rando mix of the four frameworks, and a total of 15 visualizations are shown.

A.2 Additional object interaction skeleton qualitative results

8 shows additional visualization results which are not on the main paper. In case of the results of anomaly pose problem for example, images lying on a bed or jumping on a bench or using a laptop, unknown skeleton is generated but not in the results using our framework. We could obtain better object interacting skeleton with object feature embedding than without using it so that more natural skeletons are generated which better describe the situation.

A.3 Additional object interaction image comparision SDXL qualitative results

We only demonstrate qualitative results by SD-inpainting [10] as a comparison in our main paper. We more qualitative results with SDXL [9]. Similarly, using SDXL [9] there were four problems, (a) absence of human, (b) incomplete human generation, (c) absence of interaction, (d) incomplete multi person generation. In the first row of figure 9, we have an absence of human problem where the multi person disappears or the chef preparing the food is not properly visible. In the second row, we have an incomplete human generation where the woman in bed is not properly created or the person skiing is not properly created. In the third row, we have an absence of interaction where the wooden sppon is not properly interacted with or the tennis racket is not properly interacting with. In the last row, we have an incomplete multi person generation problem where the multi person is not properly created. However, you could see taht our method sovles all of these problems.

A.4 Implementation Detail

We use Pytorch [53] in training and evaluating our framework. Moreover, we use ImageNet pre-trained ResNet backbone from torchvision [53]. And we use Adam optimizer [54] in training with batch size 64. Initially, we set learning rate to 10^{-4} and reduce by $\frac{1}{10}$ at epoch 70 and 120. A single Nvidia RTX-3090 is used to train and inference our framework. We use D3DP [42] for diffusion-based object interaction skeleton estimation module. A simple MLP network is used in combined

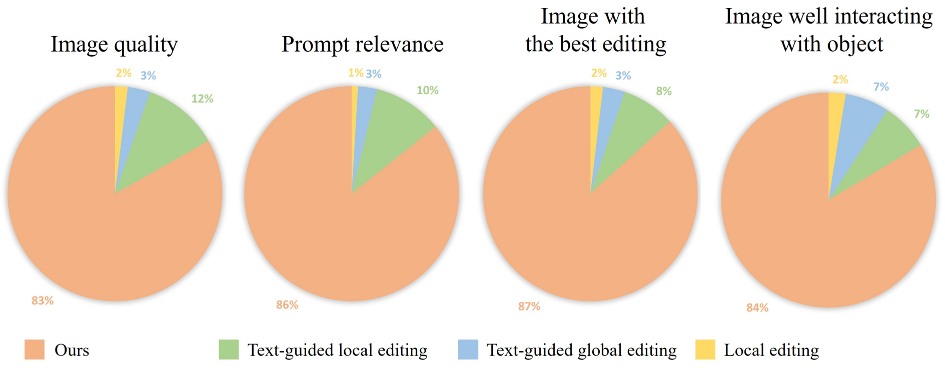
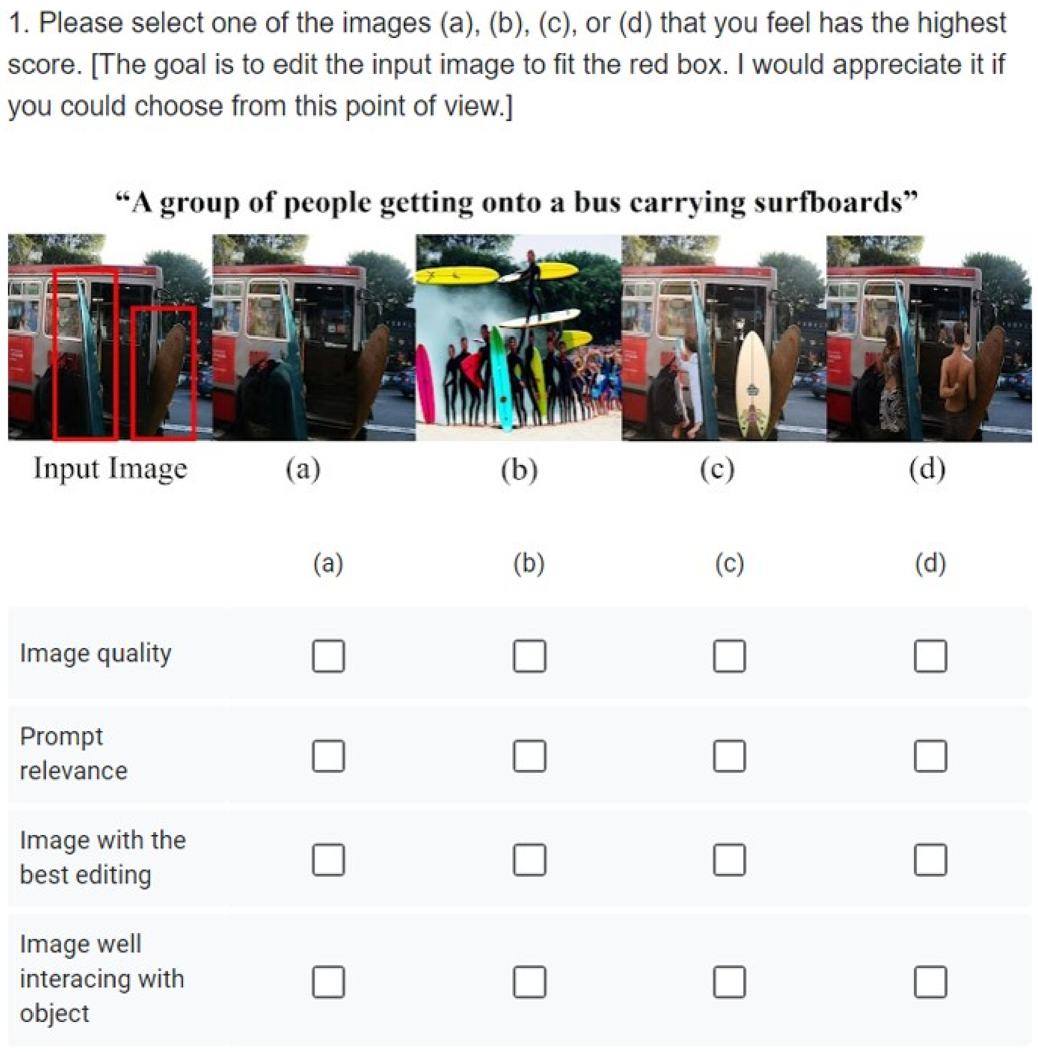


Figure 6: A user study result showing that our framework outperform others in all four categories: image quality, prompt relevance, image with the best editing and image well interacting with object.



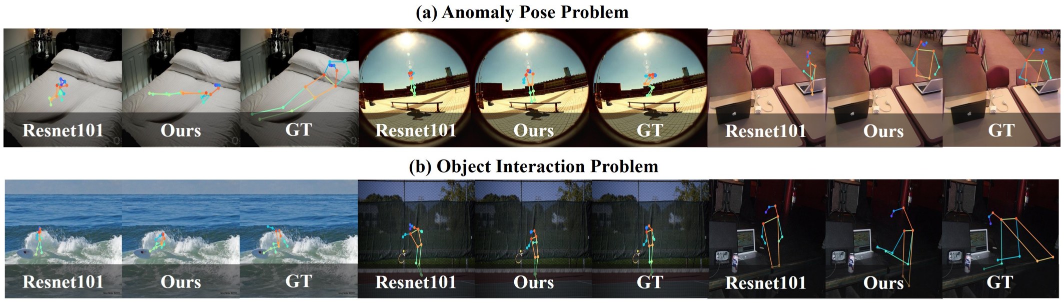
1. Please select one of the images (a), (b), (c), or (d) that you feel has the highest score. [The goal is to edit the input image to fit the red box. I would appreciate it if you could choose from this point of view.]

"A group of people getting onto a bus carrying surfboards"

Input Image (a) (b) (c) (d)

	(a)	(b)	(c)	(d)
Image quality	□	□	□	□
Prompt relevance	□	□	□	□
Image with the best editing	□	□	□	□
Image well interacing with object	□	□	□	□

Figure 7: Examples of our survey format. Surveyee can choose which is better for each of the four categories.



(a) Anomaly Pose Problem

(b) Object Interaction Problem

Figure 8: This figure shows the results of absence and the presence of our diffusion-based object interaction skeleton estimation module generating a human skeleton. (a) show the problem that generating a skeleton where joints are squeezed which is an anomaly pose. (b) shows the problem of generating invalid interaction with an object.

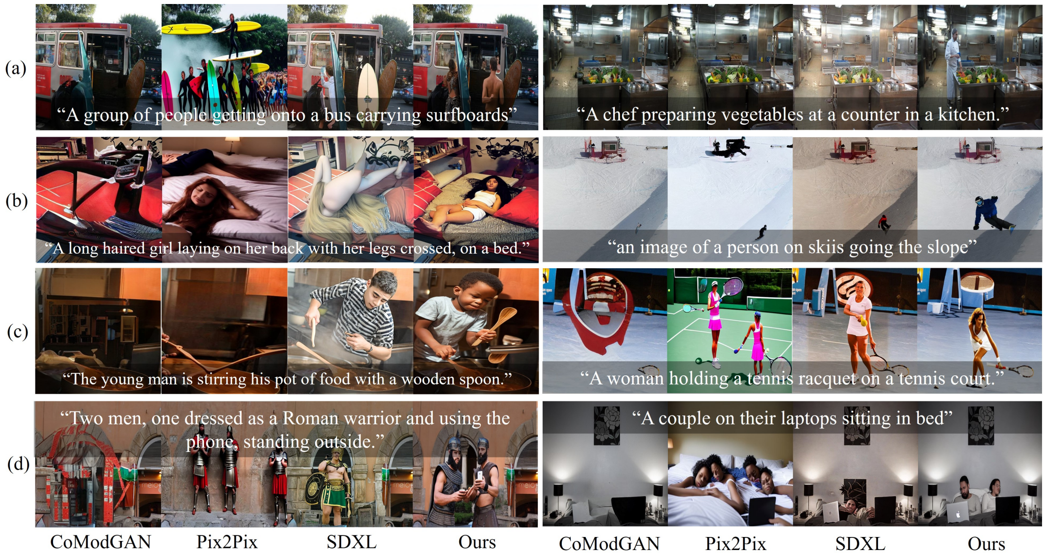

(a)
“A group of people getting onto a bus carrying surfboards”
“A chef preparing vegetables at a counter in a kitchen.”
(b)
“A long haired girl laying on her back with her legs crossed, on a bed.”
“an image of a person on skiis going the slope”
(c)
“The young man is stirring his pot of food with a wooden spoon.”
“A woman holding a tennis racquet on a tennis court.”
(d)
“Two men, one dressed as a Roman warrior and using the phone, standing outside.”
“A couple on their laptops sitting in bed”
CoModGAN
Pix2Pix
SDXL
Ours
CoModGAN
Pix2Pix
SDXL
Ours

Figure 9: This figure demonstrate the comparison between representative editing framework to ours. CoModGAN [6], Instruct-Pix2Pix [1], Stable-Diffusion XL inpainitng (SDXL) [9] were used for comparison. Compare with SDXL [9], (a) shows the absence of human that needs to be generated. (b) shows the incompleteness of human generation. (c) shows the absence of object interaction. (d) shows the results of incomplete generation of multi-person.

embedding module. In addition, various time embeddings were used. In inference stage, we set timestep to 2000 and employ denoiser. Stable-diffusion controlnet inpainitng [44] is used for skeleton guided image editing model.