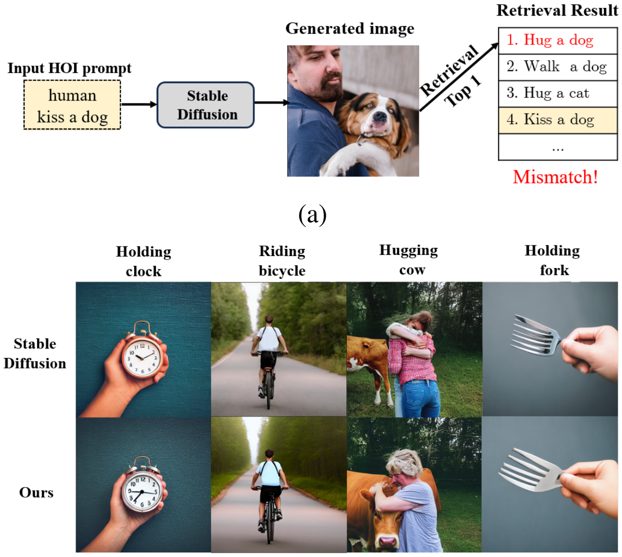
Semantic-Aware Human Object Interaction Image Generation

Zhu Xu [1] Qingchao Chen [2] Yuxin Peng [1] Yang Liu [1]

Abstract

Recent text-to-image generative models have demonstrated remarkable abilities in generating realistic images. Despite their great success, these models struggle to generate high-fidelity images with prompts oriented toward human-object interaction (HOI). The difficulty in HOI generation arises from two aspects. *Firstly*, the complexity and diversity of human poses challenge plausible human generation. *Furthermore*, untrustworthy generation of interaction boundary regions may lead to deficiency in HOI semantics. To tackle the problems, we propose a **S**emantic-**A**ware HOI generation framework SA-HOI . It utilizes human pose quality and interaction boundary region information as guidance for denoising process, thereby encouraging refinement in these regions to produce more reasonable HOI images. Based on it, we establish an iterative inversion and image refinement pipeline to continually enhance generation quality. Further, we introduce a comprehensive benchmark for HOI generation, which comprises a dataset involving diverse and fine-grained HOI categories, along with multiple custom-tailored evaluation metrics for HOI generation. Experiments demonstrate that our method significantly improves generation quality under both HOI-specific and conventional image evaluation metrics. The code is available at https://github.com/XZPKU/SA-HOI.git

1. Introduction



(a)

(b)

Figure 1: **Existing problems in HOI image generation.** (a) The generated image deviates from the prescribed HOI category as it retrieves to another one. (b) Qualitative comparisons between Stable Diffusion (Rombach et al., 2022) (top row) and our SA-HOI (bottom row). Images within each column are generated with the same text prompt describing the HOI category on top of them.

Recently, the diffusion-based models (HuggingFace, 2022; Song & Ermon, 2020) have shown great success in text-to-image generation tasks (Ramesh et al., 2022; Rombach et al., 2022; Saharia et al., 2022) through extensive training on large-scale datasets. Despite their impressive ability to produce realistic and diverse images closely aligned with provided object-oriented text descriptions, our empirical observations highlight challenges when they are exposed to human-object interaction (HOI) oriented descriptions. As shown in Figure 1(a), the generated image from the stable diffusion model (Rombach et al., 2022) deviates from the specified “kiss a dog” HOI category and aligns more with “hug a dog”. Through a human study, we empirically noted a 60.9% occurrence of semantic misalignment between the generated 1000 images and their specified texts.

The primary reasons for the difficulty in HOI generation can be attributed to two factors: **(1) Complex Human Structure and Pose Diversity:** Generating human images involves dealing with a complex structure comprising numerous interconnected body parts. As shown in Figure 1(b) (first row), the recent stable diffusion model sometimes results in deformations and unreliable poses, such as dis-

[1] Wangxuan Institute of Computer Technology, Peking University [2] National Institute of Health Data Science, Peking University. Correspondence to: Yang Liu <yangliu@pku.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

torted hands and missing right leg in the first two images. **(2) Semantic complexity and Variability in Interactions:** merely possessing a plausible human pose falls short of meeting the semantic demands inherent in HOI, the quality of interaction boundary region also plays a crucial role. As shown in Figure 1(b) first row, the third generated image from the Stable diffusion model lacks the portrayal of the human wrapping their arms around the cow at the interaction boundary, compromising the expression of "human *hugging* a cow". Similarly, the fourth generated image fails to express the "hold" semantic as hand and fork is not contacted.

To address the challenges above, we present a Semantic-Aware HOI (SA-HOI) generation framework as shown in Figure 2. Firstly, we use stable diffusion model to generate an initial image. We employ a human pose detector trained on a large dataset of real human poses, to assess the quality of each generated body part. A lower confidence score for a specific keypoint from the pose detector indicates a deviation from typical human poses, signaling a potential issue with the quality of the generated body part (*deformed body parts*). Additionally, we identify *interaction boundary regions* by assessing the proximity of human body parts to the contour of the object with the help of segmentation tool. Secondly, we utilize a blurring technique on the region containing deformed body parts and the interaction boundary, which intentionally removes fine-grained information from the corresponding area. Then the remaining information within blurred content is subsequently utilized to guide image refinement. Specifically, in the reverse process of diffusion models, we leverage pose and interaction boundary aware attention maps to enhance the overall quality and minimize artifacts around the highlighted body parts or interaction boundary area. Thirdly, we introduce an Iterative inversion and Image Refinement pipeline, allowing the model to continually improve itself based on the refined image from the last iteration. This process leads to a gradual enhancement of generation quality, all achieved without the need for additional training. Some generated samples from our approach is presented in Figure 1(b) second row.

To the best of our knowledge, we are the first work concentrating on HOI image generation from pure text descriptions. We introduce a comprehensive benchmark for HOI generation, consisting of 150 prompts covering human-object, human-animal, and human-human interactions. The benchmark includes diverse yet fine-grained scenarios, such as "holding" 49 different objects and 16 ways of interacting with a "horse". We further propose specific evaluation metrics tailored for HOI image generation. These metrics comprehensively assess the quality of generated HOI images in terms of *authenticity*, *plausibility*, and *fidelity*. They reflect the quality of the generated body pose, HOI spatial configuration, and the degree of semantic consistency with the provided text.

In summary, our contributions are threefold. (1) We introduce a semantic-aware method to enhance overall quality and reduce artifacts for HOI generation. Equipped with an iterative inversion and image refinement pipeline, our model can continually enhance itself in a step-by-step manner. (2) We propose the first HOI generation benchmark covering human-object, human-animal and human-human interaction, along with evaluation metrics that are specifically designed for HOI generation. (3) Extensive experiments demonstrate our method outperforms existing diffusion-based methods under both HOI-specific and conventional evaluation metrics for image generation.

2. Related Work

Sampling guidance for diffusion models. Multiple guidance schemes have been proposed for diffusion models recently. Classifier guidance (CG) (Dhariwal & Nichol, 2021) is proposed to use a classifier to guide the reverse process toward specific class distribution. (Ho & Salimans, 2022) proposes classifier-free guidance (CFG) as an alternate strategy for CG. DiffusionCLIP (Kim et al., 2022) expands text-to-image generation with CLIP guidance. SAG (Hong et al., 2023) further points out that self-attention maps within diffusion model can be adopted as guidance messages for reverse process, by utilizing which model can generate more high-quality images. However, self-attention maps mainly concentrate on high-frequency part within image, lacking targeted guidance toward human pose and interaction boundary region. Instead, our method explicitly concentrates on human pose and interaction boundary region, enhancing HOI image quality by refining these areas.

Posed-guided Human Image Generation Pose-guided human image generation (HIG) has been well explored (Men et al., 2020; Lv et al., 2021; Ma et al., 2018), which aims to generate images with source image's appearance and desired pose condition. The development of ControlNet (Zhang & Agrawala, 2023) further leads to more approaches (Ju et al., 2023) focusing on the accuracy and diversity of pose control. (Weng et al., 2023) proposes to utilize SMPL (Zhang et al., 2023) model to provide plausible human pose prior to refine images. However, pose-guided HIG only targets for the rationality of human pose, while HOI generation has further requirements on interaction expression and fidelity to HOI semantics.

Customized Image generation for diffusion models Customized image generation emerges as personalized applications of diffusion-based models. ControlNet (Zhang & Agrawala, 2023) introduces additional controls to generate images with customized signals like depth and skeleton information. Textual Inversion (Gal et al., 2022) generates images for specific unique concepts by optimizing corresponding concept embedding with a few exampler images.

Compared with them, HOI generation not only concerns instance-level generation for humans and objects, but also requires semantic-level interaction generation, which has been seldom explored before. ReVersion (Huang et al., 2023b) concentrates on relation modelling, targeting for generating relation-customized images by optimizing relation embedding in inversion manner. However, it can only apply to 10 specific relations and need extra training for each relation, which is not efficient. Instead, our method not only requires no additional training, but also can be applied to 150 HOI categories. T2I-CompBench (Huang et al., 2023a) proposes a comprehensive benchmark for the compositional image generation and evaluation, primarily focusing on attribute binding, spatial relationships, and complex object compositions, but the adopted evaluation metrics lack comprehensive justification for HOI image quality, thus not ideally suitable for HOI image generation task. Another attempt InteractGAN(Gao et al., 2020) targets HOI image generation with a different task formulation, which requires human images, object images, as well as action categories as inputs, and generates images with consistent identity for the input human and object by adjusting the human pose. Compared with it, our HOI image generation task relies solely on textual prompts describing HOI categories and can accommodate 150 HOI categories, embracing greater diversity and superior scalability.

3. Preliminaries

Denoising Diffusion Probabilistic Models Denoising Diffusion Probabilistic Model (DDPM) recovers an image from Gaussian distribution noise through an iterative denoising process. Formally, given an image x_0 along with variance schedule β_t at a timestep t ∈ {T, T-1, ..., 1}, we can obtain x_t through forward process, which is pre-defined as a Markovian process. For a trained diffusion model parameterized as $\epsilon_\theta(x_t, t)$ and $\Sigma_\theta(x_t, t)$, we define its reverse process as following: $\Sigma_\theta(x_t, t) = \beta_t = \alpha_t^2$. For a given $x_T \sim N(0, I)$, DDPM iteratively sample $x_{T-1}, x_{T-2}, ...x_0$ by computing:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)) + \sigma_t z \tag{1}$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=0}^{t} \alpha_i$, $z \sim N(0, I)$ and ϵ_θ is network parameterized by θ. By applying reparameterization trick, we can obtain $\hat{x}_0$, an intermediate reconstruction of x_0 at a timestep t, using the following equation:

$$\hat{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t)) \tag{2}$$

Generalized Guidance of Diffusion models Guidance schemes for diffusion models can be generalized as follows: the input for a diffusion model at timestep t is generalized condition h_t and perturbed sample $\overline{x}_t$ that lacks h_t, and guidance can be formulated through the utilization of an imaginary regressor, $p_{im}(h_t|\overline{x}_t)$, which aims to predict h_t from $\overline{x}_t$, whose formulation is:

$$\tilde{\epsilon}_{\overline{x}_t, h_t} = \epsilon_\theta(\overline{x}_t, h_t) - s\sigma_t \nabla_{\overline{x}_t} \mathrm{logp_{im}(h_t|\overline{x}_t)} \tag{3}$$

where s is the guidance scale. By the calculation of the gradient of p_{im}, samples generated under guidance are expected to be more suitable with that information stored in h_t. Further, with Bayes'rule, the gradient can be further formulated as:

$$\nabla_{\overline{x}_t} \mathrm{logp_{im}(c|\overline{X}_t)} = -\frac{1}{\mathrm{\sigma_t}}(\epsilon_\theta(\mathrm{\overline{x}_t, h_t}) - \epsilon_\theta(\mathrm{\overline{x}_t})) \tag{4}$$

then we acquire the final expression of $\tilde{\epsilon}_{\overline{x}_t, h_t}$ as follows:

$$\tilde{\epsilon}_{\overline{x}_t, h_t} = \epsilon_\theta(\overline{x}_t) + (1 + s)(\epsilon_\theta(\overline{x}_t, h_t) - \epsilon_\theta(\overline{x}_t)) \tag{5}$$

By storing different information inside h_t, we can provide corresponding guidance during the generation process. Specifically, applying Gaussian kernel G_σ convolution over x_t, i.e., $\tilde{x}_t = G_\sigma * x_t$, can be viewed as one simplified guidance scheme, as $h_t = x_t - \tilde{x}_t$ and $\overline{x}_t = x_t$ in Equation 5, which is proven (Hong et al., 2023) to effectively guide diffusion more appropriate to the salient information stored in h_t with a moderate σ in Gaussian kernel, thus harvesting more high-quality generation.

4. Method

4.1. Overview

The overall framework of SA-HOI is shown in Figure 2. It consists of two designs: Pose and Interaction Boundary Guidance (PIBG) and Iterative Inversion and Refinement (IIR). In PIBG, for given HOI prompt t_0 and noise n_0, we first utilize Stable Diffusion to generate I_0 as the initial image. To measure the pose quality of I_0, we adopt a pose detector to acquire human body joint positions $\{P_i^{pose}\}_{i=1}^{N^{pose}}$ and corresponding confidence score $\{S_i^{pose}\}_{i=1}^{N^{pose}}$, where N^{pose} is visible joint number for human in I_0. Considering that a low confidence score signals potentially deformed generation for specific body parts, we utilize S^{pose} and P^{pose} to construct pose mask M^{pose}, which highlights low-quality pose regions. Considering the spatial context of interactions, we identify interaction boundary regions by assessing the proximity of human body parts to the contour of the object with the help of segmentation tool. These identified regions are highlighted in the interaction boundary mask M^{inter} to enhance the semantic expression in the vicinity of interaction boundary areas. For each denoising step, to guide the model focusing on low pose quality or interaction boundary regions and minimizing artifacts in them, we adopt Gaussian Blurring on these regions by utilizing M^{pose} and M^{inter} as region constraints, and subsequently

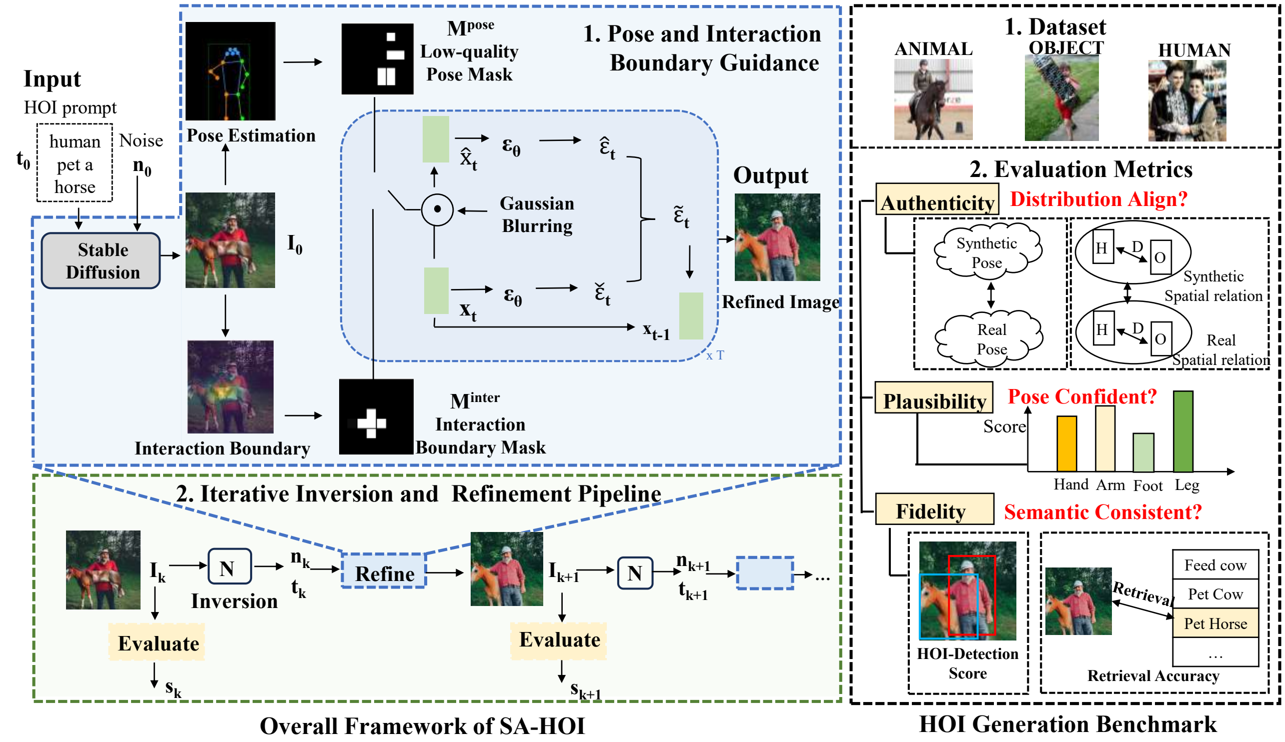


Figure 2: **The overall framework of our method SA-HOI (left) and HOI Generation benchmark (right).** Our method SA-HOI contains two designs: *Pose and Interaction Boundary Guidance (PIBG)*, *Iterative Inversion and Refinement (IIR)*. In PIBG, Low-quality Pose Mask M^{pose} and interaction boundary mask M^{inter} are adopted to refine low pose quality part and interaction boundary region of original image I_0 in each denoising step as detailed in Algorithm 1. IIR gradually enhances generation quality based on inversion model N and PIBG as shown in Algorithm 2. Our benchmark includes a *Dataset* of realistic images covering human-object, human-animal, and human-human interactions, as well as comprehensive *HOI Evaluation Metrics* reflecting authenticity, plausibility and fidelity of generated HOI images.

harvest more reasonable human pose and interaction boundary regions compared to I_0. More details about the pose and interaction boundary guidance process for HOI generation are provided in Section 4.2 and 4.3. Further, to continuously improve generation quality, we introduce IIR, which leverages inversion model N to extract noise n and text embedding t from the image to be further refined, and subsequently employs PIBG for the next refinement, then utilize the quality evaluator Q to assess the refined image quality. We iterate through <Inverse, Evaluate, Refine> operations to gradually enhance image quality. More details about IIR are provided in Section 4.4.

4.2. Pose Guidance for HOI Generation

To deal with the complex and diverse human structure across various scenarios, we propose to integrate pose guidance into the process of HOI generation. The pseudo code of our pose and interaction boundary guided sampling is shown in Algorithm 1. In each denoising step, we first acquire predicted noise ϵ_t and intermediate reconstruction $\hat{x}_0$ (line 5-6) following conventional design in Stable Diffusion. Then we apply Gaussian Blurring G on $\hat{x}_0$ to get degraded latent feature $\tilde{x}_0$ and $\tilde{x}_t$ (line 7-8). To apply the derivation presented in Equation 5, we incorporate pose quality information within h_t by utilizing the pose detection results P^{pose} and S^{pose}.

Pose Attention and Mask Generation P^{pose} and S^{pose} are utilized to generate A^{pose} and M^{pose}, which aims to highlight low pose quality regions and guide model to diminish deformed generation in these regions. For each joint $P_i^{pose} = (x_i, y_i)$, the score S_i^{pose} represents the reliability of corresponding joint generation, where a lower score denotes a higher possibility of low-quality generation. To guide the model to refine the low-quality areas, we highlight the regions with low pose scores by calculating

$$G_i(x,y) = \frac{1}{(S_i^{pose})^2} \exp\left(-\frac{(x-x_i)^2 + (y-y_i)^2}{2\sigma^2}\right) \tag{6}$$

where $G_i \in \mathbb{R}^{H\times W}$, x, y are pixel-wise coordinates of image, H, W are image size and σ is the deviation of Gaussian distribution. G_i represents the attention centered on the i^{th} joint, where a lower S_i^{pose} leads to increased emphasis on the regions centered around P_i^{pose}. By combining attention across all joints, we can form the ultimate attention map $A^{pose} \in \mathbb{R}^{H\times W}$, guiding the model to focus on all regions with potential generation problems aforementioned.

$$A^{pose}(x,y) = \sum_{i=1}^{N^{pose}} G_i(x,y) \tag{7}$$

Further, to preserve the majority of the generated content in I$_0$, we transform A^{pose} into a mask using a threshold. This allows us to selectively modify only the regions with high attention, indicating areas that require refinement. The mask M^{pose} is obtained by downsampling: $M^{pose} = \text{DownSample}(\mathbb{1}_{A_{pose}>\phi_t})$, where ϕ_t is the threshold to generate the mask at timestep t, and M^{pose} are downsampled to $\mathbb{R}^{H'\times W'}$ to fit the same spatial dimensions of x_t. To dynamically determine the ratio of refined areas as the image sample and timestep t vary, we formulate ϕ_t as

$$\phi_t = \phi_0 \cdot (1 - \alpha \sin\frac{\pi t}{2T}) \cdot \sum_{i=1}^{N_{pose}} \sum_{x,y} G_i(x,y) \tag{8}$$

ϕ_t concurrently considers two factors: (1) ϕ_t is scaled by multiplying with the sum of all joint pose attentions to accommodate varying sample content and modulate the scale of attention. (2) as t decreases, ϕ_t increases and leads to less activated areas in M^{pose}, consequently more subtle changes in the image content.

Pose Blurring Process Using M^{pose} to activate low pose quality regions , we sample h_t as follows:

$$h_t = M^{pose} \odot x_t - M^{pose} \odot \tilde{x}_t \tag{9}$$

By calculating the difference of x_t and $\tilde{x}_t$ with the constraint of M^{pose}, h_t only keeps fine-scale information in high attention areas, i.e., poor pose quality areas. Then degraded latent features $\check{x}_t^{pose}$ are constructed as follows:

$$\check{x}_t^{pose} = (1 - M^{pose}) \odot x_t + M^{pose} \odot \tilde{x}_t \tag{10}$$

$\check{x}_t^{pose}$ not only precisely applies blurring to the poor pose quality areas, but also mitigates unwarranted effects on unrelated regions across the entire image. Following Equation 5, we form the final prediction noise $\check{\epsilon}_t^{pose}$ (line 12 in Algorithm 1) under the guidance of pose quality.

4.3. Interaction Boundary Guidance for HOI generation

Given the significant impact of interactive boundary regions on HOI semantic expression, we detect these regions to form interaction boundary attention map A^{inter} and mask M^{inter} accordingly, then store interaction boundary information in h_t. We gauge the occurrence of interactive behavior by detecting the proximity between humans and objects. Formally, for a given human-object pair, we detect the joint coordinates of the person $\{C_i\}_{i=1}^{K}$ and obtain the outer contour points $\{O_t\}_{t=1}^{N_o}$ of the object through a segmentation tool, where N_o is a fixed number to describe object. Then we calculate $D \in \mathbb{R}^{K\times N_o}$ as distance matrix between C and O. Closer points between C and O proximately contain more interaction semantics, so we choose T pairs of (C_i, O_j) points with minimum distance within D as keypoints of interaction boundary, which are emsembled as $\{P_i^{inter}\}_{i=1}^{2T}$. The interactive score S_i^{inter} for keypoint P_i^{inter} is the minimum distance of this point within D. Then we form attention $A^{inter} \in \mathbb{R}^{H\times W}$ with the same formulation in Equation 6 and Equation 7, where S_i^{pose} changes to the S_i^{inter}, indicating points with smaller S^{inter}, i.e., closer distance, requires higher attention. Similarly following Equation 9 , we generate mask $M^{inter} \in \mathbb{R}^{H'\times W'}$ by A^{inter} and threshold ϕ_t, and form interaction boundary guided h_t with M^{inter}. Then we predict interaction boundary guided noise $\check{\epsilon}_t^{inter}$(line 12 of Algorithm 1). By combing $\check{\epsilon}_t^{inter}$ and $\check{\epsilon}_t^{pose}$, we generate final noise $\tilde{\epsilon}_t$ (line 13 in Algorithm 1) and finish one denoising step.

Algorithm 1 Pose and Interaction Boundary Guided Sampling

```
1:  Input: Initial image $I_0$, pose attention $A^{pose}$ and interaction boundary attention $A^{inter}$ on $I_0$. $\phi_t$ threshold for attention mask generation.
2:  Output: denoised image feature $x_0$
3:  Initialize $x_T \sim N(0, I)$.
4:  for $t = T, T-1, ..., 1$ do
5:      $\epsilon_t \leftarrow U(x_t)$ {U($x_t$) predicts noise $\epsilon_t$ for $x_t$. }
6:      $\hat{x}_0 \leftarrow \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t))$
7:      $\tilde{x}_0 \leftarrow G(\hat{x}_0)$ {G($\cdot$) is Gaussian Blurring Function.}
8:      $\tilde{x}_t \leftarrow \sqrt{\overline{\alpha}_t}\tilde{x}_0 + \sqrt{1-\overline{\alpha}_t}\epsilon_t$
9:      $M^{pose} \leftarrow \text{DownSample}(\mathbb{1}_{A^{pose}>\phi_t})$, $M^{inter} \leftarrow \text{DownSample}(\mathbb{1}_{A^{inter}>\phi_t})$
10:     $\check{x}_t^{pose} \leftarrow (1 - M^{pose}) \odot x_t + M^{pose} \odot \tilde{x}_t$
11:     $\check{x}_t^{inter} \leftarrow (1 - M^{inter}) \odot x_t + M^{inter} \odot \tilde{x}_t$
12:     $\check{\epsilon}_t^{pose} \leftarrow U(\check{x}_t^{pose})$, $\check{\epsilon}_t^{inter} \leftarrow U(\check{x}_t^{inter})$
13:     $\tilde{\epsilon}_t \leftarrow \frac{1}{2}(\check{\epsilon}_t^{pose}+\check{\epsilon}_t^{inter})+(1+s)[\epsilon_t-\frac{1}{2}(\check{\epsilon}_t^{pose}+\check{\epsilon}_t^{inter})]$
14:     $x_{t-1} \leftarrow N(\frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\tilde{x}_t), \Sigma_t)$
15: end for
```

4.4. Iterative Inversion and Image Refinement Pipeline

To obtain real-time quality assessment of generated images, we introduce quality evaluator Q, which serves as guideline for iterative <evaluate + refine> operations. The iterative refinement enables our model of continual enhancement of the generation quality in a step by step manner. The pipeline is shown in Algorithm 2. For image I_k at k^{th} round, we adopt evaluator Q to acquire its quality score S_k, which in

Algorithm 2 Iterative Image Refinement Pipeline

```
Input: Threshold $\theta$ to stop iterative refinement. Origin image $I_0$ and score $S_0$ evaluated by Q. Largest iterate rounds $K$. $\mathrm{I_{refine}} = \{\}$, $\mathrm{S_{refine}}=\{\}$ to store refined images and scores.
Output: refined HOI image $I$ in $\mathrm{I_{refine}}$ with highest score in $\mathrm{S_{refine}}$.
for $k = 0, 1, 2, ..., K-1$ do
    $n_k, t_k \leftarrow N(I_k)$ {Acquire noise and text embedding for $I_k$}
    $I_{k+1} \leftarrow$ SA-HOI $(n_k, t_k)$ {Refine image}
    $S_{k+1} \leftarrow Q(I_{k+1})$ {Assess Quality of $I_{k+1}$}
    $I_{refine} \leftarrow I_{refine} + I_{k+1}$ {Store $I_{k+1}$}
    $S_{refine} \leftarrow S_{refine} + S_{k+1}$
    if $S_{k+1} - S_k < \theta$ then {No significant change}
        break { Finish Iterative Refinement}
    end if
end for
```

this paper by default is the function to calculate PCS[1] , and then we generate I_{k+1} based on I_k. To preserve the main content of the I_k after refining, the corresponding noise is needed to serve as the initial value for denoising. However, such noise is not available, so we introduce null-text inversion (Mokady et al., 2022) N to acquire its noised latent feature n_k and text embedding t_k, which serves as input for SA-HOI to generate refined result I_{k+1}. By comparing quality scores at consecutive iterations, we judge whether to continue refining. When S_{k+1} and S_k show no significant difference, i.e., below threshold θ, we finish refinement and output the image with highest quality score.

5. HOI Generation Benchmark

5.1. Dataset

Our dataset consists of 150 HOI categoires, covering human-object, human-animal, and human-human interaction scenarios for comprehensive evaluation. The categories are all collected from public HOI detection data-set HICO-DET (Chao et al., 2015). *H-A*: <human, verb, *animal*>, where all 91 HOI categories concerning animal in HICO-DET are included. *H-O*: <human, *hold*, object>. We meticulously select 49 HOI categories with "hold" as verb, ensuring physical contact between human and object. *H-H*: <human, verb, *human*>, which includes all 10 human-human interaction categories in HICO-DET. We select real images from HICO-DET for these HOI categories as reference images, with each category randomly sampling up to most 200 images. To exclude the influence of multiple HOI co-existing in one image, we crop the corresponding HOI union part as real image samples. To sum up, we collect 5k images to serve as our HOI generation dataset, allowing us to estimate the distribution of real images. Exampler images of our dataset are shown in Figure. 5 in Appendix.

5.2. Evaluation Metrics

To better assess the generated HOI image quality, we tailored several metrics for HOI generation, enabling comprehensive evaluation of generated images from the perspectives of plausibility, authenticity, fidelity.

Authenticity

Pose Distribution Distance (PDD): Humans typically adopt similar poses for the same HOI category, thereby conforming to corresponding pose distribution. Consequently, we measure the distribution distance between real images and generated images to gauge the authenticity of the generated human poses. Formally, given N_{real} real images and N_{syn} generated images, we detect joints coordinates $\{J_i = (X_i, Y_i)\}_{i=1}^{K}$ for human in each image. We normalize J_i by the width W and height H of human bounding box to eliminate the influence of human size, and turn $\{J_i\}_{i=1}^{K}$ into $\{\hat{J}_i\}_{i=1}^{K}$ by transforming relative reference system with the human hip joint as the origin, with the aim to better capture the pose information. Then we utilize KL-Divergence to measure the pose distribution distance between real and generated pose utilizing $\hat{J}$. Given $P = \{\hat{J}_j^{syn}\}_{j=1}^{N_{syn}}$ and $Q = \{\hat{J}_j^{real}\}_{j=1}^{N_{real}}$, we calculate the distribution distance as $\mathrm{PDD}(P, Q) = \sum_i P(i) \log\left(\frac{P(i)}{Q(i)}\right)$, and average over all HOI categories to serve as PDD, with a lower PDD suggests a more precise matching to real poses.

Human-Object Distance Distribution (HODD): Similarly, human-object pairs involved in the same HOI category exhibit a consistent spatial configuration pattern, approximated by the distribution of distances between humans and objects. We utilize human joints and object outer contours to calculate their distances. For a given human object pair, we detect the joint coordinates of the person $\{C_i\}_{i=1}^{K}$ and obtain the outer contour points $\{O_t\}_{t=1}^{N_o}$ of the object through a segmentation model, where K is joint number for human and N_o is number for representing object outer contour. The exact distance between the person and the object is then represented by computing the closest distance between C_i and O_i, which is formulated as $\mathrm{DIS}(C, O) = \underset{i,j}{min} ||C_i - O_j||_2$. Then we measure the distribution distance between generated and real images with KL-Divergence following the same procedure as PDD.

Plausibility

Pose Confidence Score (PCS): The confidence score of pose

[1] More details and ablations about PCS are provided in 5.2 and 6.3. In principle, we can adjust the function of Q based on user requirements.

showcases the plausibility and rationality of the detected human joints. Formally, for a detected person with K body joints along with confidence scores $\{S_i\}_{i=1}^{K}$ within image I, the pose confidence for I is $\frac{1}{K}\sum_{i=1}^{K} S_i$, and we average pose confidence across all generated images to serve as PCS. Given that the human pose detector is trained on an extensive dataset of real human poses, a higher Pose Confidence Score (PCS) indicates increased plausibility in the human generation. Conversely, a low score suggests a departure from typical real human poses used in detector training, signaling a potential issue with the quality of the generated body part, termed as deformed body parts.

Fidelity

Human Object Interaction Fidelity (HOIF): To measure the semantic fidelity between images and HOI categories, we introduce HOI detector as oracle model to examine the alignment of images. An off-the-shelf HOI detector $g(\cdot)$ (Lei et al., 2023) trained on HICO-DET (Chao et al., 2015) is adopted to detect possible HOI triplets in images. Since no available ground truth bounding boxes can be treated as annotation, we simplify the metric as the confidence score on the target HOI category detection. Given the image I, we detect all possible HOI triplets for given target HOI category c and a threshold θ for activate HOI examination, $S_c = g(I, c, \theta)$, where S_c is the confidence score for the detected HOI. We average over all generated images as HOIF, with higher HOIF signifying better semantic consistency between images and HOI text prompts.

R-accuracy: We measure the retrieval accuracy of images to their HOI categories to measure fidelity. Formally, we utilize CLIP (Radford et al., 2021) to encode image embedding E_i and HOI categories text embedding E_t, and calculate the accuracy of images retrieved to their own HOI categories within E_t. Higher accuracy is positively correlated with higher semantic consistency between text prompts and images.

6. Experiment

6.1. Implementation Details

We employ the Stable Diffusion v1.5 (Rombach et al., 2022) as the base text-to-image model and apply our SA-HOI on it. Hyperparameters θ, δ, ϕ_0, α, T are set as 0.01, 1, 1, 0.6 and 4. For evaluation, we adopt general image evaluation metrics, including FID, KID, and our tailored metrics for HOI generation as introduced in Sec. 5.2. We also include a subjective evaluation for more comprehensive evaluation of our method. We utilize our iterative image refinement pipeline to generate HOI images for evaluation. More details can be found in Appendix.

Table 1: Comparison on general text-to-image metrics with other methods.

MODEL	SCENARIO	FID↓	KID(10^{-2})↓
STABLE DIFFUSION	H-A	58.84	2.022
SAG	H-A	56.91	2.019
DIFFUSION_HPC	H-A	59.42	2.043
OURS	H-A	**54.55**	**1.812**
STABLE DIFFUSION	H-O	78.28	2.896
SAG	H-O	75.99	2.845
DIFFUSION_HPC	H-O	77.92	3.012
OURS	H-O	**74.70**	**2.784**
STABLE DIFFUSION	H-H	137.66	3.5792
SAG	H-H	138.06	3.6213
DIFFUSION_HPC	H-H	139.43	3.9578
OURS	H-H	**134.72**	**3.3218**

6.2. Comparison with other methods

General Metrics: We compare our methods with Stable Diffusion (Rombach et al., 2022), SAG (Hong et al., 2023) and Diffusion_HPC (Weng et al., 2023) in Table 1. We observe: (1) Our method has lower FID and KID than Stable Diffusion under each scenario (H-A, H-O, H-H), which signifies by rectifying pose and interaction boundary regions, our method stably harvests quality improvement by statistical significance. (2) We still harvest lower FID and KID compared with SAG under each scenario , indicating our tailored guidance provides more accurate and suitable information for the refinement of HOI images than self-attention maps, thus gaining further improvement. (3) We outperform Diffusion_HPC by 4.87%/ 3.22%/ 4.71% on FID (line 3/ 7/ 11), as well as KID. This showcases rectifying interaction boundary regions further improves overall image quality.

HOI metrics: The results for HOI-specific metrics are shown in Table 2 , from which we can observe (1) *Plausibility:* Credit to our pose guidance refining on low-quality poses, both PCS for body and hand joints are improved under all scenarios compared to SD, indicating SA-HOI enhances the credibility for joint poses. Also, our PCS is comparable with Diffusion_HPC (line11-12), who utilize extra SMPL (Pavlakos et al., 2019) model to provide human pose prior, indicating our plausibility in human pose generation. (2) *Authenticity:* Our method surpasses other comparison approaches on all splits in terms of HODD and PDD metrics. This suggests that both the distribution of body/hand/animal pose and the distribution of inter-distance (spatial configuration pattern) in our generated images closely resemble realistic distributions. (3) *Fidelity:* Thanks to the interaction boundary guidance to improve generation quality on interaction boundary regions, we achieve the highest HOIF,

Table 2: **Comparison on tailored metrics of HOI Generation with other methods**. Each split under the metrics denotes the evaluation for corresponding region pose, "JOINT" denotes combined pose of human and animal under H-A scenario or combined pose of two humans under H-H scenario.

Model	Scenario	PCS(%)↑		HODD(10^{-2})↓		PPD(10^{-2})↓				HOIF(%) ↑	R-Acc(%) ↑	
		BODY	HAND	BODY	HAND	BODY	HAND	ANIMAL	JOINT		A@1	A@5
SD	H-A	69.54	27.83	15.71	18.82	8.14	12.22	9.24	7.63	30.06	38.06	83.38
SAG	H-A	71.35	27.93	14.33	17.90	7.92	11.34	7.99	7.06	32.14	38.92	87.50
D_HPC	H-A	**74.58**	28.17	15.82	19.91	8.05	11.13	8.67	7.22	31.55	36.18	84.52
Ours	H-A	73.88	**29.01**	**12.70**	**15.77**	**7.63**	**11.01**	**7.19**	**6.94**	**34.38**	**42.48**	**96.57**
SD	H-O	67.48	28.86	12.10	14.48	12.87	10.86	-	-	25.21	93.08	99.04
SAG	H-O	67.05	28.37	12.23	14.37	12.31	10.99	-	-	26.93	94.57	99.91
D_HPC	H-O	67.92	27.53	12.87	14.21	13.55	10.94	-	-	27.50	92.16	99.41
Ours	H-O	**68.69**	**29.65**	**11.50**	**13.97**	**11.19**	**10.81**	-	-	**32.77**	**94.69**	**99.91**
SD	H-H	70.67	23.57	11.49	12.89	8.77	13.54	10.18	8.62	27.08	47.83	85.66
SAG	H-H	73.40	25.50	11.43	12.77	8.65	13.07	10.18	8.55	28.33	49.10	88.77
D_HPC	H-H	**76.99**	**27.01**	11.03	12.41	8.34	13.76	10.54	8.47	26.01	43.80	86.43
Ours	H-H	75.31	26.77	**10.82**	**11.56**	**7.91**	**12.00**	**9.25**	**7.92**	**29.32**	**52.10**	**93.12**

indicating our approach preserves rich HOI semantic information and more accurate expression for it. R-Acc are all improved compared to SD under all scenarios and we reach the best accuracy among all methods. Both in detection and retrieval manner, SA-HOI harvests better fidelity towards HOI semantics, indicating that our model attains better semantic consistency between text prompts and images.

Subjective Evaluation: We carry out a user study for a subjective evaluation. We invited 26 participants to rate the quality of the provided images through a questionnaire. Ten textual prompts describing different HOI categories were randomly selected, and images were generated accordingly using different models. Additionally, realistic images from our benchmark were included for each prompt to enhance credibility. Participants were instructed to rate images on a scale from 1 to 5, representing bad quality to perfect quality. Evaluation criteria included (1) Human Pose Realism, which assesses the credibility of the length, number, and angles of human limbs. (2) Object Appearance: Evaluating the appearance of objects in the generated image. (3) Interaction Semantics: Judging the semantic relevance between the generated images and the given textual prompts. (4) Overall Quality: Rating the overall quality of the generated image. With a total of 5.2k responses, insights into the performance comparison among the different models were obtained. Table. 3 shows the performance, from which we can conclude: (1) our approach outperforms other methods in subjective evaluation across all metrics; (2) there exists a positive correlation between our proposed evaluation metric HOIF and human preference, suggesting its reliability. Qualitative comparisons are shown in Figure. 4 in Appendix.

Table 3: **Subjective evaluation of generation results.**

Model	Human↑	Object↑	Inter↑	Overall↑
SD	2.48	3.24	3.67	2.98
D_HPC	3.01	2.27	2.97	2.15
SAG	3.08	3.37	3.35	3.04
Ours	3.31	3.66	4.03	3.54
GT	4.35	4.34	4.09	4.25

6.3. Ablations and Analysis

In this section, we investigate how the performance of the proposed method is affected by different model settings. We study mainly in three aspects: contribution of network components, the formation of pose guidance and masking parameter sensitivity analysis.

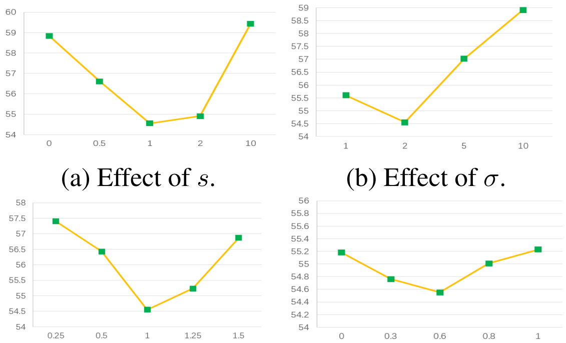
Contribution of Network Components: We conduct a detailed ablation study by examining the effectiveness of each proposed component in our network structure, and the result are presented in Table 4, from which we can observe: (1) separately introducing Pose and Interaction boundary guidance (line 3-4) along with iterative refinement outperform Gaussian Blurring (line 2) and SD (line 1) under all HOI metrics (such as 3.61% body PCS for pose guidance) and FID (separately minimizes 3.87% and 3.14%), demonstrating that our guidance schemes are well-suited for HOI targeted refinement. (2) without iterative refinement based on quality assessment criteria (line 5), we outperform SD by only adopting pose and interaction boundary guidance, indicating their effectiveness. (3) incorporating iterative refinement (line 6) harvests continual enhancement over only one-step refinement (line 5) under all HOI metrics (such

Table 4: **Experiments on contribution of network components**. "P", "IB" and "R" denotes Pose guidance, interaction boundary guidance and iterative refinement."SD" represents Stable Diffusion, "GB" denotes Gaussian Blurring. Experiments are conducted under H-A scenario.

MODEL	FID↓	PCS(%)↑		HODD(10^{-2})↓		PPD(10^{-2})↓				HOIF(%) ↑	R-ACC(%) ↑	
		BODY	HAND	BODY	HAND	BODY	HAND	ANIMAL	JOINT		A@1	A@5
SD	58.84	69.54	27.83	15.71	18.82	8.14	12.22	9.24	7.63	30.06	38.06	83.38
GB	57.40	70.08	26.54	15.11	18.83	8.02	12.15	9.11	7.45	30.15	37.34	82.88
P+R	54.97	73.15	**29.04**	13.70	16.92	7.88	11.84	8.60	7.33	31.26	39.57	86.10
IB+R	55.70	72.43	28.55	13.15	15.99	7.85	11.61	7.89	7.12	33.00	39.11	88.93
P+IB	54.56	73.40	28.77	13.08	16.31	8.11	11.07	8.05	7.09	32.31	41.91	92.73
P+IB+R	**54.55**	**73.88**	29.01	**12.70**	**15.77**	**7.63**	**11.01**	**7.19**	**6.94**	**34.38**	**42.48**	**96.57**

as 0.44% Body PCS and 2.07% HOIF), indicating it can continually improve refined image quality. Notably, all metrics are boosted when we adopt PCS as quality assessment criteria, indicating the strong correlations between PCS and overall HOI generation quality.

Formation of pose guidance: As shown in Table 5, our body pose guidance (line 2) surpasses SD (line 1) with higher PCS and lower PPD. To test the generalization ability of our pose guidance, we implement more pose guidance beyond human body. Considering hands hold crucial part for various HOI categories, we adopt hand pose guidance, which follows the same process in Algorithm 1, except the pose attention is generated from hand pose detector rather than body pose detector. We also adopt animal pose guidance to improve animal generation. Experiments are conducted on H-A scenario. From results in Table 5, we observe: (1) separately applying pose guidance on hand part (line 3) and animal part (line 4) contributes to not only higher PCS (1.88% on hand), but also minimizes PDD for corresponding regions (1.50% on hand and 2.35% on animal). (2) By incorporating different guidance schemes, we harvest better performance than single pose guidance (line 5-7). This underscores the robust extensibility of our approach, allowing flexible adjustment of applied guidance types and combinations according to the specific HOI categories.



(a) Effect of s.

(b) Effect of σ.

(c) Effect of ϕ_0.

(d) Effect of α.

Figure 3: **Masking Parameter Sensitivity Analysis.**

Masking Parameter Sensitivity Analysis: We analyze the sensitivity of parameters concerning attention masking: guidance scale s, kernel deviation σ, masking threshold ϕ_0 and α (noted in Equation 8). As shown in Figure 3, $s = 1$, σ = 2, ϕ_0=1 and α=0.6 separately reaches best performance (54.44 in FID) in guidance scheme. We notice that extreme value choice for parameters will affect the performance. For instance, with regards to σ, exceedingly small value will lead to over-uniform attention distribution, thereby compelling the model to rectify inconsequential regions. Conversely, excessively large value can cause model to focus solely on areas with the poorest quality, neglecting other regions that also require refinement, thus resulting in a degradation of performance.

Table 5: **Experiments on formation of pose guidance**. "SD" represents Stable Diffusion, "B", "H" and "A" separately present pose guidance of body, hand and animal. Experiments are conducted under H-A scenario.

MODEL	PCS(%)↑	PPD(10^{-2})↓		
	HAND	HAND	ANIMAL	JOINT
SD	27.83	12.22	9.24	7.63
B	29.01	11.01	7.19	6.94
H	29.71	10.72	8.35	7.44
A	28.05	13.11	6.89	6.83
B+H	**31.50**	**10.56**	7.67	6.97
B+A	29.67	11.14	6.73	**6.40**
B+A+H	31.37	10.71	**6.67**	6.42

7. Conclusion

We present a method SA-HOI that utilizes pose quality and interaction boundary region information as guidance to generate high-quality human object interaction (HOI) images. We further introduce a HOI generation benchmark including a dataset and multiple tailored metrics for comprehensive quality evaluation. Experiments show the effectiveness of our method. With the method and benchmark, our work contributes novel insights to the HOI image generation field.

Impact Statement

Human Object Interaction Image Generation carries significantly broader impact potential from both positive and negative sides. By generating realistic images depicting human-object interactions, our method contributes to diverse applications like immersive virtual environment creation. However, it also raises ethical concerns related to privacy and misinformation, which must be carefully considered and addressed to ensure responsible deployment and minimize potential negative consequences.

Acknowledgements

This work was supported by the grants from National Natural Science Foundation of China (62372014,61925201, 62132001).

References

Abdulla, W. Mask r-cnn for object detection and instance segmentation on keras and tensorflow. https://github.com/matterport/Mask_RCNN, 2017.

Chao, Y.-W., Wang, Z., He, Y., Wang, J., and Deng, J. HICO: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C. C., and Lin, D. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Diller, C. and Dai, A. Cg-hoi: Contact-guided 3d human-object interaction generation, 2023.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022.

Gao, C., Liu, S., Zhu, D., LIU, Q., Cao, J., He, H., He, R., and Yan, S. Interactgan: Learning to generate human-object interaction. *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. URL https://api.semanticscholar.org/CorpusID:222278643.

Ho, J. and Salimans, T. Classifier-free diffusion guidance, 2022.

Hong, S., Lee, G., Jang, W., and Kim, S. Improving sample quality of diffusion models using self-attention guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7462–7471, 2023.

Huang, K., Sun, K., Xie, E., Li, Z., and Liu, X. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation, 2023a.

Huang, Z., Wu, T., Jiang, Y., Chan, K. C., and Liu, Z. ReVersion: Diffusion-based relation inversion from images. *arXiv preprint arXiv:2303.13495*, 2023b.

HuggingFace, 2022. URL https://huggingface.co/CompVis/stable-diffusion-v1-4.

Ju, X., Zeng, A., Zhao, C., Wang, J., Zhang, L., and Xu, Q. Humansd: A native skeleton-guided diffusion model for human image generation, 2023.

Kim, G., Kwon, T., and Ye, J. C. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2426–2435, June 2022.

Lei, T., Caba, F., Chen, Q., Ji, H., Peng, Y., and Liu, Y. Efficient adaptive human-object interaction detection with concept-guided memory. 2023.

Lv, Z., Li, X., Li, X., Li, F., Lin, T., He, D., and Zuo, W. Learning semantic person image generation by region-adaptive normalization, 2021.

Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., and Gool, L. V. Pose guided person image generation, 2018.

Men, Y., Mao, Y., Jiang, Y., Ma, W.-Y., and Lian, Z. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5084–5093, 2020.

Mokady, R., Hertz, A., Aberman, K., Pritch, Y., and Cohen-Or, D. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022.

Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A. A., Tzionas, D., and Black, M. J. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Peng, X., Xie, Y., Wu, Z., Jampani, V., Sun, D., and Jiang, H. Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv e-prints*, art. arXiv:2204.06125, April 2022. doi: 10.48550/arXiv.2204.06125.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Song, Y. and Ermon, S. Improved techniques for training score-based generative models. *arXiv preprint arXiv:2006.09011*, 2020.

von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., and Wolf, T. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

Weng, Z., Bravo-Sánchez, L., and Yeung-Levy, S. Diffusion-hpc: Synthetic data generation for human mesh recovery in challenging domains, 2023.

Xu, S., Li, Z., Wang, Y.-X., and Gui, L.-Y. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *ICCV*, 2023.

Zhang, L. and Agrawala, M. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

Zhang, M., Wang, X., Decardi-Nelson, B., Bo, S., Zhang, A., Liu, J., Tao, S., Cheng, J., Liu, X., Yu, D., Poon, M., and Garg, A. Smpl: Simulated industrial manufacturing and process control learning environments, 2023.

A. Additional Implementational Details

Parameter Setting For the experiments, we use two A100 80G GPUs to sample images from the pre-trained models Stable Diffusion v1.5 (Rombach et al., 2022). Our SA-HOI is built upon it, with all the available pre-trained weights from its publicly available repository. For CFG, we adopt guidance scale of 7.5, the text prompt is "A photo of a person *verbing* a/an *object*." for HOI category <*verb*, *object*>, and the negative prompt is set as "". We adopt DDIMScheduler(von Platen et al., 2022) with 50 steps for the denoising process, and all the generated images are with size 512×512. For pose detection, we apply pose detectors from RTMPose toolbox(Chen et al., 2019). For the object segmentation mask, we adopt Mask-RCNN(Abdulla, 2017) as the segmentation model. For each HOI category, we utilize templated description "A photo of a human *verbing* a/an *object*" as a text prompt for category <human,*verb*,*object*>, and sample 50 images with different seeds.

Combination of pose and interaction boundary guidance with CFG To incorporate our tailored guidance scheme for HOI into text-2-image models like Stable Diffusion(Rombach et al., 2022), we need to combine our guidance with CFG(Ho & Salimans, 2022). In practice, we formulate the final guided noise as

$$\begin{aligned}&\tilde{\epsilon}(x_t) = \epsilon_\theta(x_t, c) + (1 + s_c)(\epsilon_\theta(x_t, c) - \epsilon_\theta(x_t, c)) +\\&(1 + s_p)(\epsilon_\theta(x_t, M_t^{pose}) - \epsilon_\theta(\overline{x_t})) +\\&(1 + s_i)(\epsilon_\theta(x_t, M_t^{inter}) - \epsilon_\theta(\overline{x_t}))\end{aligned} \tag{11}$$

where s_c, s_p, s_i are guidance scale for CFG, pose guidance and interaction boundary guidance. c denotes text prompts describing HOI category, and M_t^{pose} and M_t^{inter} are masks for corresponding guidance. We simply adopt identical value for s_p and s_i in practice.

B. Additional Related Work

3D Human Object Interaction Generation Several attempts(Xu et al., 2023; Diller & Dai, 2023; Peng et al., 2023) exist for 3D Human Object Interaction Generation, which aims to generate motion sequence for HOI instance. 3D HOI generation differs from 2D HOI generation from following perspectives: (1) their generation target is sparse keypoint information including coordinates and velocity, while image generation requires dense pixel-wise generation. (2) their main focus lies on the temporal modeling over motion sequence, while we focus on the realness and fidelity of images. (3) applicable HOI categories for existing 3D HOI generation approaches are quite limited, while 2D HOI generation could expand to most common HOI categories. In summary, there are significant differences between 2D and 3D HOI generation in various aspects. Therefore, we did not compare them in subsequent experiments.

C. Qualitative results

Visualization comparison of different methods We provide several qualitative comparisons for our methods and other model, which is shown in Figure. 4. Comparing with other three recent methods, our approach showcases enhancements in human pose, object appearance, and HOI semantic expression simultaneously, resulting in improved overall image quality.

D. Human Object Interaction Image Generation Benchmark

D.1. Exampler dataset images

Some exampler images of our dataset are shown in Figure 5. These images not only exhibit a diverse range of poses and realistic spatial distances between human object pairs, but also effectively convey the semantic information associated with their HOI categories, which makes them well-suited for the evaluation of image generation quality.

D.2. HOI category List

We list detailed HOI categories for our H-A, H-O, H-H scenarios as follow,

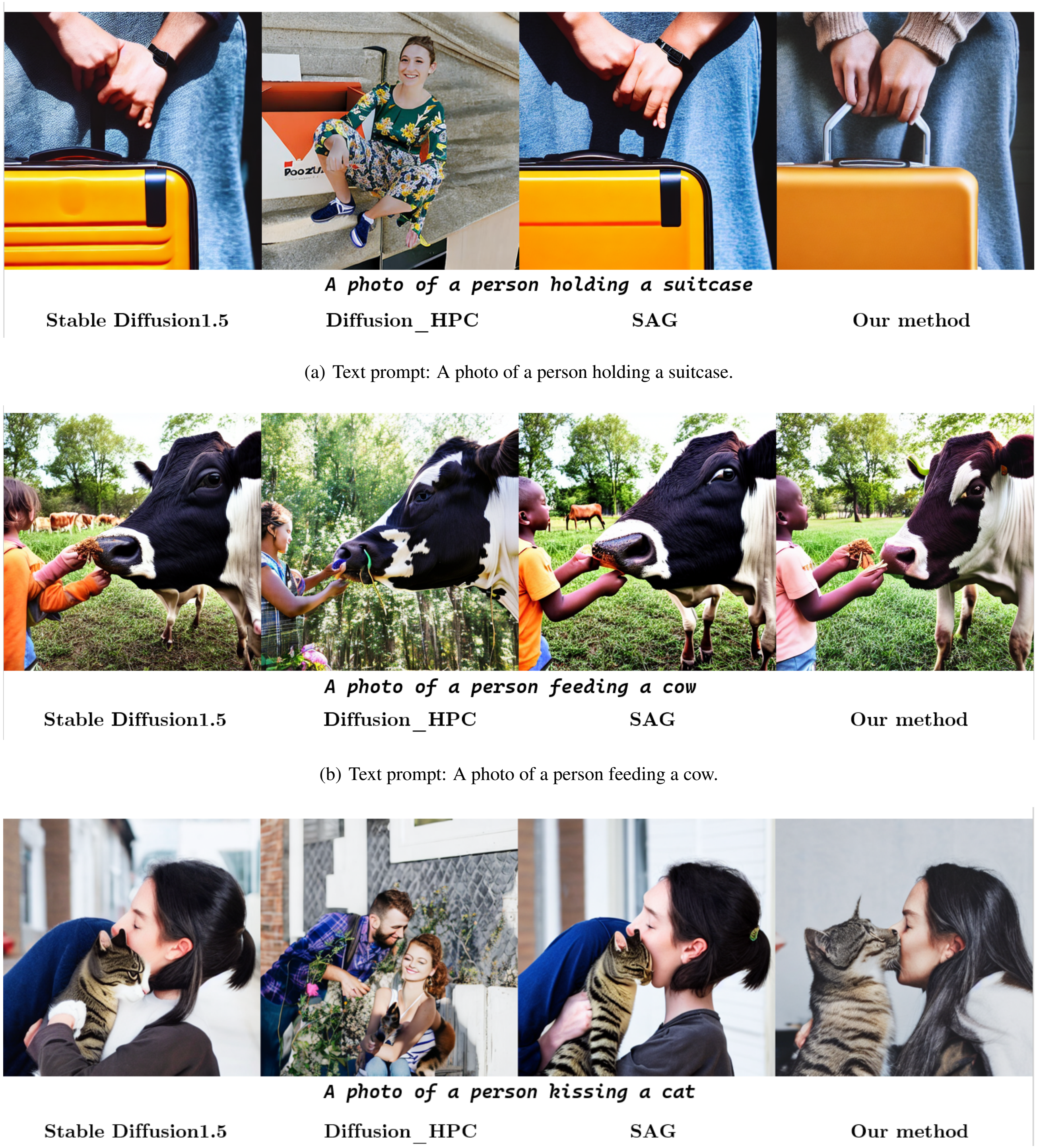
H-A: <chase, bird>, <feed, bird>, <pet, bird>, <release, bird>, <watch, bird>, <feed, cow>, <herd, cow>, <hold, cow>, <hug, cow>, <kiss, cow>, <lasso, cow>, <milk, cow>, <pet, cow>, <ride, cow>, <walk, cow>, <carry, dog>, <dry, dog>, <feed, dog>, <groom, dog>, <hold, dog>, <hose, dog>, <hug, dog>, <inspect, dog>, <kiss, dog>, <pet, dog>, <run, dog>, <scratch, dog>, <straddle, dog>, <train, dog>, <walk, dog>, <wash, dog>, <chase, dog>, <feed,

horse>, <groom, horse>, <jump, horse>, <kiss, horse>, <load, horse>, <hop, horse>, <pet, horse>, <race, horse>, <ride, horse>, <run, horse>, <straddle, horse>, <train, horse>, <walk, horse>, <wash, horse>, <, horse>, <carry, sheep>, <feed, sheep>, <herd, sheep>, <hold, sheep>, <hug, sheep>, <kiss, sheep>, <pet, sheep>, <ride, sheep>, <shear, sheep>, <walk, sheep>, <wash, sheep>, <feed, bear>, <hunt, sheep>, <watch, sheep>, <feed, elephant>, <hold, elephant>, <hose, elephant>, <hug, elephant>, <kiss, elephant>, <hop, elephant>, <pet, elephant>, <ride, elephant>, <walk, elephant>, <wash, elephant>, <watch, elephant>, <feed, giraffe>, <kiss, giraffe>, <pet, giraffe>, <ride, giraffe>, <watch, giraffe>, <feed, zebra>, <hold, zebra>, <pet, zebra>, <watch, zebra>.

H-O: <hold, bicycle>, <hold, bird>, <hold, bottle>, <hold, cat>, <hold, chair>, <hold, cow>, <hold, dog>, <hold, horse>, <hold, motorcycle>, <hold, potted plant>, <hold, sheep>, <hold, apple>, <hold, backpack>, <hold, banana>, <hold, baseball bat>, <hold, baseball glove>, <hold, cake>, <hold, carrot>, <hold, cellphone>, <hold, clock>, <hold, cup>, <hold, frisbee>, <hold, handbag>, <hold, keyboard>, <hold, kite>, <hold, knife>, <hold, laptop>, <hold, mouse>, <hold, orange>, <hold, oven>, <hold, pizza>, <hold, refrigerator>, <hold, remote>, <hold, sandwich>, <hold, scissors>, <hold, skateboard>, <hold, skis>, <hold, spoon>, <hold, suitcase>, <hold, surfboard>, <hold, teddy bear>, <hold, tie>, <hold, toothbrush>, <hold, umbrella>, <hold, vase>.

H-H: <greet, person>, <hold, person>, <hug, person>, <kiss, person>, <stab, person>, <tag, person>, <teach, person>, <lick, person>, <carry, person>, <no interaction, person>

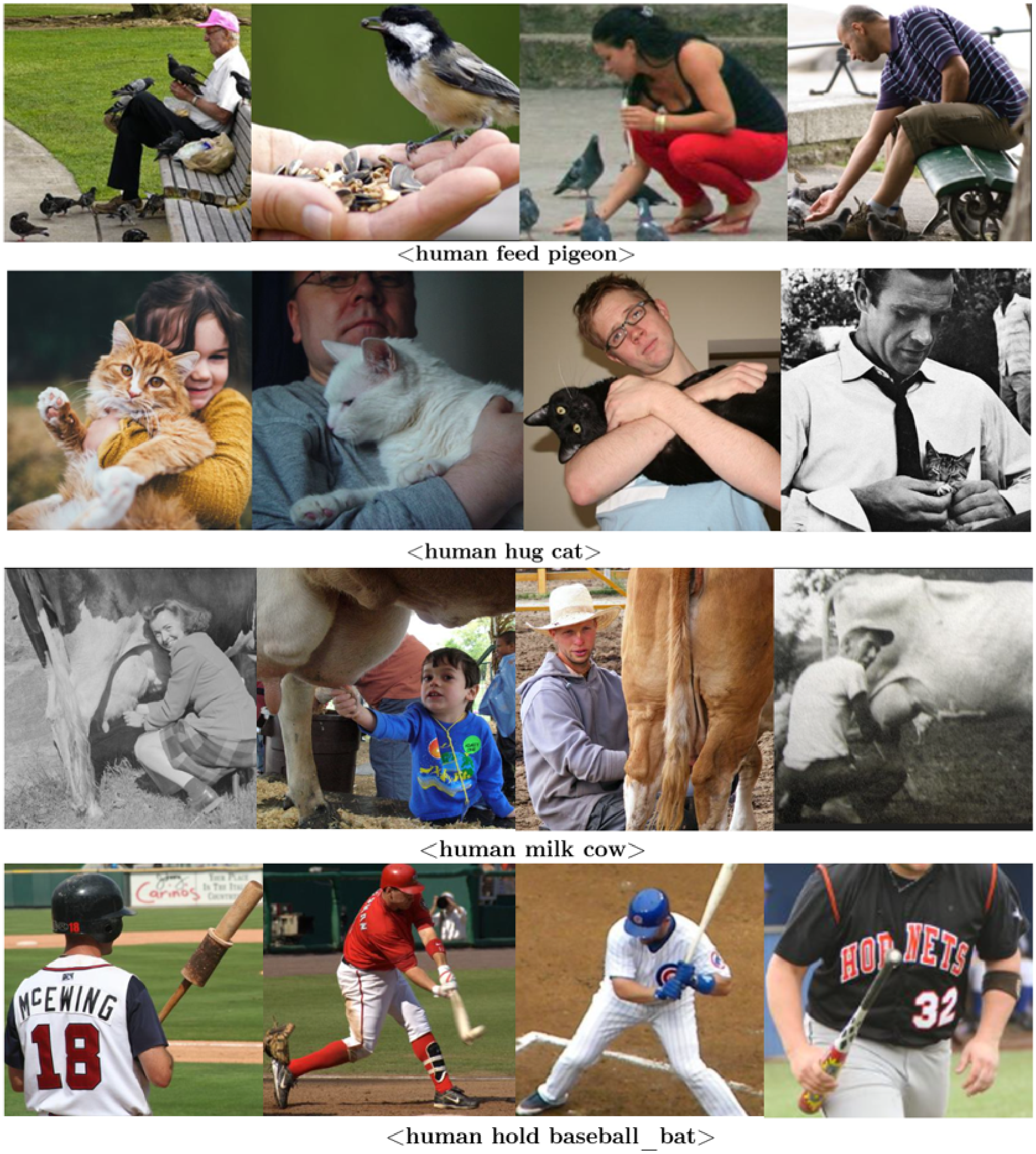
Figure 4: **Qualitative comparisons among Stable Diffusion v1.5(Rombach et al., 2022), Diffusion_HPC(Weng et al., 2023) and SAG(Hong et al., 2023) and our method.**



(a) Text prompt: A photo of a person holding a suitcase.

(b) Text prompt: A photo of a person feeding a cow.

(c) Text prompt: A photo of a person kissing a cat.



<human feed pigeon>

<human hug cat>

<human milk cow>

<human hold baseball_bat>

Figure 5: **Exampler images of our dataset**. Images within each HOI category show diverse poses and coherent spatial correlations between human and object, and they all adhere to their HOI categories semantics tightly.