This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

InteractDiffusion: Interaction Control in Text-to-Image Diffusion Models

Jiun Tian Hoe[1] Xudong Jiang[1] Chee Seng Chan[2] ✉ Yap-Peng Tan[1] Weipeng Hu[1] ✉

[1]School of EEE, Nanyang Technological University, Singapore [2]CISiP, Universiti Malaya, Malaysia

jiuntian001@e.ntu.edu.sg {exdjiang,eyptan,weipeng.hu}@ntu.edu.sg cs.chan@um.edu.my

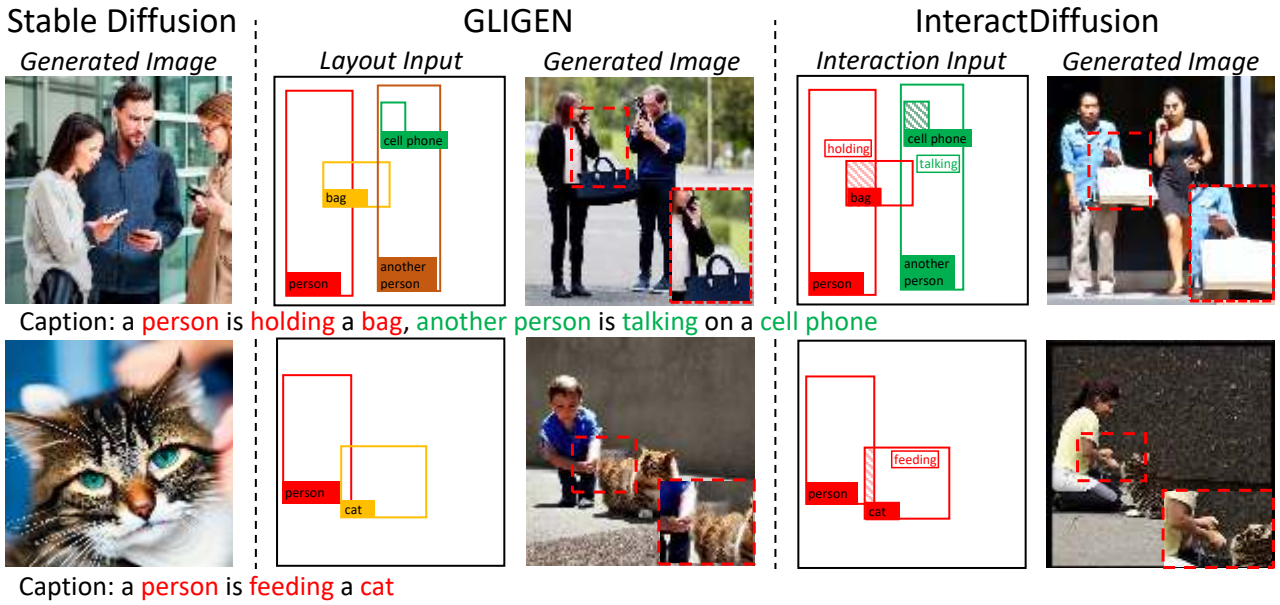

Figure 1. Generated samples of size 512x512. Stable Diffusion conditions on text caption only, while GLIGEN conditions on extra layout input. Our proposed InteractDiffusion conditions on extra interaction label and its location shown by the shaded area.

Abstract

Large-scale text-to-image (T2I) diffusion models have showcased incredible capabilities in generating coherent images based on textual descriptions, enabling vast applications in content generation. While recent advancements have introduced control over factors such as object localization, posture, and image contours, a crucial gap remains in our ability to control the interactions between objects in the generated content. Well-controlling interactions in generated images could yield meaningful applications, such as creating realistic scenes with interacting characters. In this work, we study the problems of conditioning T2I diffusion models with Human-Object Interaction (HOI) information, consisting of a triplet label (person, action, object) and corresponding bounding boxes. We propose a pluggable interaction control model, called InteractDiffusion that extends existing pre-trained T2I diffusion models to enable them being better conditioned on interactions. Specifically, we tokenize the HOI information and learn their relationships via interaction embeddings. A conditioning self-attention layer is trained to map HOI tokens to visual tokens, thereby conditioning the visual tokens better in existing T2I diffusion models. Our model attains the ability to control the interaction and location on existing T2I diffusion models, which outperforms existing baselines by a large margin in HOI detection score, as well as fidelity in FID and KID. Project page: https://jiuntian.github.io/interactdiffusion.

1. Introduction

The advent of diffusion generative models recently opens up new creative task opportunities. While diffusion models could generate diverse high quality images that reconstruct the original data distributions, it is important to control the content generated. Numerous literatures have

✉ corresponding authors

since extensively studied how to control the image generation of the diffusion models via *e.g*. class [10, 37], text [24, 26, 27, 29], image (including edge, line, scribble and skeleton) [2, 16, 36] and layout [2, 7, 19, 34, 38]. However, these are insufficient to effectively express the nuanced intentions and desired outcomes, especially the interactions between objects. Our work introduces another important control in image generation: *interaction*.

Interaction refers to a reciprocal action between two entities or individuals. Without a doubt, interaction is an integral part of describing our daily activities. However, we find that existing diffusion models work well on static images such as paintings or scenic photos but face great challenges in generating images involving interactions. For instance, GLIGEN [19] adds layout as a condition to help specify the location of objects, but controlling the relationship or interaction between the objects remains an open difficult problem, as shown in Fig. 1. Control at the interaction level in text-to-image (T2I) diffusion models has countless applications, *e.g*. e-commerce, gaming, interactive storytelling etc.

This paper studies the interaction-conditioned image generation, *i.e*. how to specify the interaction in the image generation process. It faces three main challenges:

a) **Interaction representation**: How to represent interaction information in a meaningful token representation.
b) **Intricate interaction relationship**: The relationship among objects with interaction is complex, and generating coherent images remains a great challenge.
c) **Integrating conditions into existing models**: Current T2I diffusion models excel in image generation quality but lack interaction control. A pluggable module that can be seamlessly integrated into them is imperative.

To address the aforementioned issues, we propose an interaction control model called **InteractDiffusion** as a pluggable module to existing T2I diffusion model as illustrated in Fig. 2, aiming to impose interaction control. First, to provide conditioning information to the diffusion model, we treat each interacting pair as a HOI triplet and transform its information into a meaningful token representation that contains information about position, size, and category label. Particularly, we generate three different tokens for each HOI triplet, *i.e*. *subject*, *action*, and *object* tokens. While both *subject* and *object* tokens contain information about location, size, and object category, the *action* token includes the location of the interaction and its category label.

Secondly, the challenge of representing intricate interaction lies in encoding the relationship between the tokens of multiple interactions where tokens are from different interaction instances and have different role within an interaction instance. To address this challenge, we propose instance embedding and role embedding to group the tokens of the same interaction and embed their role semantically. Thirdly, as the existing transformer block consists of a self-attention and a cross-attention layer [27], we add a new Interaction Self-Attention layer in between to incorporate interaction tokens into the existing T2I model. This helps to preserve the original model during training, while simultaneously incorporating additional interaction conditioning information.

Our main contributions are summarized as follows:

(i) We address the interaction-mismatch problem in existing T2I models and raise a new challenge: controlling interaction in T2I diffusion models. Our proposed framework, **InteractDiffusion**, is pluggable to existing T2I model. It incorporates interaction information as additional conditions for training an interaction-controllable T2I diffusion model, enhancing the precision of interactions in generated images.
(ii) To effectively capture intricate interaction relationships, we introduce a novel method where we tokenize the localization and category information of ⟨subject, action, object⟩ into three distinct tokens. These tokens are then grouped together and specified in their roles of interaction through an embedding framework. This innovative approach enhances the representation of complex interactions.
(iii) InteractDiffusion significantly outperforms the baseline methods in HOI Detection Scores and maintains generation quality with slight improvements in both FID and KID metrics. To the best of our knowledge, this work is the first attempt to introduce *interaction control* to diffusion models.

2. Related Work

Human-Object Interactions Recent advancements in Human-Object Interactions (HOI) have focused on detecting HOIs in images. It aims to locate interacting human and object pairs via bounding boxes and categorize these objects and their interactions in a triplet form (*e.g*., person, feeding, cat). While recent HOI detection works [6, 17, 21, 32, 35] show promising results, data scarcity hampers detection performance for rare interactions. Conversely, HOI image synthesis, an inverse task of HOI detection, is relatively underexplored. InteractGAN [12] proposed HOI image generation via human pose and reference images of humans and objects. However, this approach is complicated as it requires a pose-template pool and reference images of humans and objects. A more closely related work is the layout-proposal-based method [15], which focuses on scene layout proposals according to HOI triplets to synthesize images but is limited to generating "object placement" proposals. Our work focuses on a new problem, namely, controlling the interaction in existing T2I diffusion models using simple bounding box and interaction relations in an end-to-end manner, without human pose and reference images. This approach efficiently addresses the data scarcity in HOI detection tasks and opens a wide range of applications.

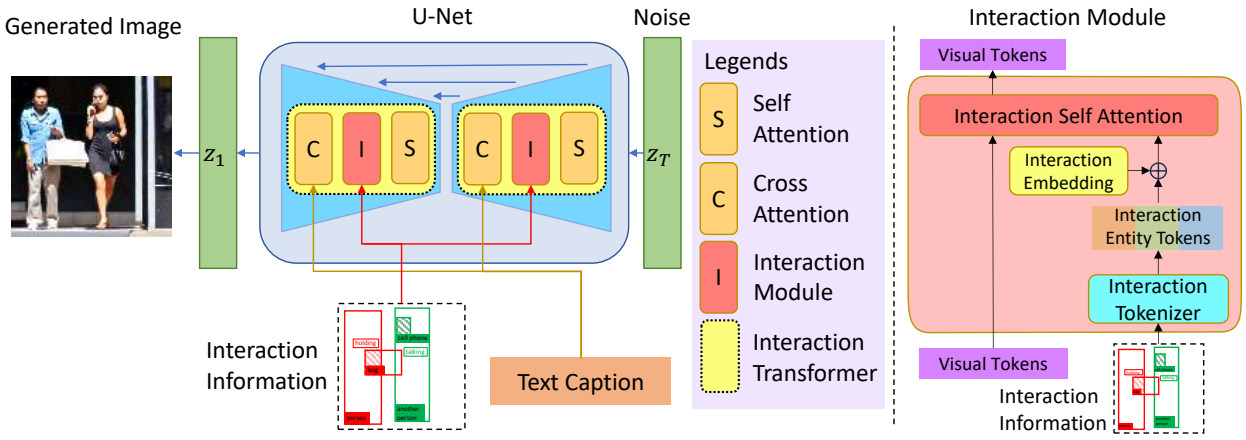

Figure 2. The overall framework of InteractDiffusion. Our proposed pluggable Interaction Module I seamlessly incorporates interaction information into an existing T2I diffusion model (left). The proposed module I (right) consists of Interaction Tokenizer (Sec. 3.2) that transforms interaction information into meaningful tokens, Interaction Embedding (Sec. 3.3) that incorporates intricate interaction relationship, and Interaction Self-Attention (Sec. 3.4) that integrates interaction control information into Visual Tokens of the existing T2I model.

Diffusion Models The diffusion probabilistic model was first proposed in [30], and further improved in training and sampling methods by [14, 31]. Training and evaluating diffusion models in pixel space could be costly and slow, and training on high-resolution images always requires calculating expensive gradients. Latent Diffusion Model (LDM) [27] compresses the image into a latent representation of lower dimensionality [11] and carries out the diffusion process in latent space to reduce the computation which was further extended to Stable Diffusion. Our work adds interaction control to the Stable Diffusion Model.

Controlling Image Generation T2I diffusion models [24, 26, 27, 29] often utilize a pretrained language model like CLIP [25] to guide the image diffusion process. This allows the generated image's content to be controlled by a provided text caption. However, a text caption alone often provides insufficient control over the generated content, particularly when aiming to create specific content such as object location and layout, scene depth maps, human poses, boundary lines, and interactions. To address this issue, several models have proposed different methods for controlling the generated content, including object layout [19, 33, 38], segmentation maps [1, 3, 8, 18] and images [23, 36]. Although controlling image generation via object layout and images can generally yield better results, one essential aspect of image has been largely ignored, namely, the interaction between objects. Our work extends the capabilities of the current T2I model by strengthening the control of interactions in the generated content.

3. Method

We first formulate the problem and then detail our InteractDiffusion model, as illustrated in Fig. 2. It comprises four parts: (a) *interaction tokenizer* that transforms interaction conditions into tokens, (b) *interaction embedding* that links the relationship between tokens of interacting triplets, (c) *interaction transformer* that constructs attention between image patches and interaction information, and (d) *interaction-conditional diffusion model* that generates images with interaction conditions.

3.1. Preliminary

We study the problem of incorporating interaction conditions $\mathbf{d}$ into existing T2I diffusion model alongside with text caption condition $\mathbf{c}$. Our aim is to train a diffusion model $f_\theta(\mathbf{z}, \mathbf{c}, \mathbf{d})$ to generate images conditioned on interaction $\mathbf{d}$ and text caption $\mathbf{c}$, where $\mathbf{z}$ is the initial noise.

Stable Diffusion, one of the best models, is a scale-up of the LDM [27] with a larger model and data size. Unlike other diffusion models, LDM splits into two stages to reduce computational complexity. It first learns a bi-directional projection to project image $\mathbf{x}$ from pixel space to a latent space as latent representation $\mathbf{z}$ and then trains a diffusion model $f_\theta(\mathbf{z}, \mathbf{c})$ in the latent space with latent $\mathbf{z}$. Our work focuses on the second stage as we are only interested in conditioning the diffusion model with interaction.

LDM learns a reverse process of a fixed Markov Chain of length T. It can be interpreted as an equally weighted sequence of denoising autoencoders $\epsilon_\theta(\boldsymbol{z}_t, t); t = 1, \cdots, T$, which are trained to predict a denoised version of their input $\boldsymbol{z}_t$, where $\boldsymbol{z}_t$ is a noisy version of the input $\boldsymbol{z}$.

The unconditional objective can be viewed as

$$\min_\theta \mathcal{L}_{\text{LDM}} = \mathbb{E}_{\boldsymbol{z}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t} \left[\|\epsilon - \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t)\|_2^2 \right], \quad (1)$$

with t uniformly sampled from $\{1, \cdots, T\}$. The model iteratively produces less noisy samples from noise $\boldsymbol{z}_T$ to

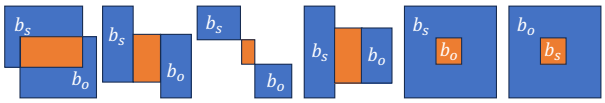

Figure 3. "Between" operation obtains the action focus area (highlighted in orange) between subject and object bounding boxes.

$\boldsymbol{z}_{T-1}, \boldsymbol{z}_{T-2}, \cdots, \boldsymbol{z}_0$, where the model $\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t)$ is realized by a UNet [28]. The final image is obtained by projecting $\boldsymbol{z}_0$ in latent space back into image space in a single pass through the decoder trained in the first stage.

Conditioning In LDM, to condition the diffusion model with various modalities like text captions, a cross-attention mechanism was added on top of the UNet backbone. The conditional input of various modalities is denoted as y and a domain specific encoder $\tau_\theta(\cdot)$ is used to project y to an intermediate token representation $\tau_\theta(y)$.

In StableDiffusion, text captions represented by y are used to condition the model. It uses a CLIP encoder denoted as $\tau_\theta(\cdot)$ to project the text caption y into 77 text embeddings, *i.e.* $\tau_\theta(y)$. In particular, the conditioned objective for StableDiffusion can be viewed as

$$\min_\theta \mathcal{L}_{\text{LDM}} = \mathbb{E}_{\boldsymbol{z}, \epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}), t}\left[\|\epsilon - \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \tau_\theta(y))\|_2^2\right], \quad (2)$$

where $\tau_\theta(\cdot)$ represents the CLIP text encoder and y represents the text caption.

3.2. Interaction Tokenizer (InToken)

We define interaction $\boldsymbol{d}$ as a triplet label consisting of $\langle$subject s, action a, and object $o\,\rangle$, as well as their corresponding bounding boxes denoted as $\langle \boldsymbol{b}_s, \boldsymbol{b}_a, \text{ and } \boldsymbol{b}_o \rangle$, respectively. We use the subject and object bounding boxes to describe their location and sizes, and introduce an action bounding box to specify the spatial location of the action. For example, a subject (*e.g.* women, boy) performing a specific action (*e.g.* carrying, kicking) toward a particular object (*e.g.* handbag, ball).

To obtain the action bounding box, we define a "between" operation, applied to the subject and object bounding boxes. Suppose $\boldsymbol{b}_s$ and $\boldsymbol{b}_o$ be specified by their corner coordinates $[\alpha_i, \beta_i], i = 1, 2, 3, 4$, the "between" operation on $\boldsymbol{b}_s$ and $\boldsymbol{b}_o$ to obtain $\boldsymbol{b}_a$ is:

$$\begin{aligned}\boldsymbol{b}_a &= \boldsymbol{b}_s \text{ between } \boldsymbol{b}_o \\ &= [R_2(\alpha_i), R_2(\beta_i)], [R_3(\alpha_i), R_3(\beta_i)],\end{aligned} \quad (3)$$

where $R_k(\cdot)$ is the k^{th} ascending rank of its arguments. Some examples of the "between" operation results are shown in Fig. 3. With this, our interaction condition inputs of an image is:

$$\begin{aligned}\mathcal{D} = [\boldsymbol{d}_1, \ldots, \boldsymbol{d}_N] = [(s_1, a_1, o_1, \boldsymbol{b}_{s_1}, \boldsymbol{b}_{a_1}, \boldsymbol{b}_{o_1}), \ldots, \\ (s_N, a_N, o_N, \boldsymbol{b}_{s_N}, \boldsymbol{b}_{a_N}, \boldsymbol{b}_{o_N})],\end{aligned} \quad (4)$$

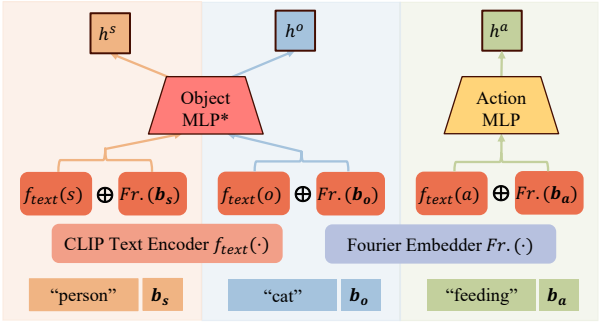

Figure 4. Interaction Tokenizer. View bottom-up.

where N is the number of interaction instances.

Subject and Object tokens We first pre-process the text label and the bounding box into an intermediate representation. In particular, we use the pre-trained CLIP text encoder to encode the text of subject, action and object as a representative text embedding and use Fourier embedding [22] to encode their respective bounding boxes following GLIGEN [19]. To generate the subject and object tokens, h^s, h^o, we use a multi-layer perceptron ObjectMLP$(\cdot)$ to fuse them as:

$$h^s = \text{ObjectMLP}([f_{\text{text}}(s), \text{Fourier}(\boldsymbol{b}_s)]) \quad (5)$$

$$h^o = \text{ObjectMLP}([f_{\text{text}}(o), \text{Fourier}(\boldsymbol{b}_o)]). \quad (6)$$

Action token For action token, we train a separate multi-layer perceptron ActionMLP$(\cdot)$ since action is semantically apart from the subject and object,

$$h^a = \text{ActionMLP}([f_{\text{text}}(a), \text{Fourier}(\boldsymbol{b}_a)]). \quad (7)$$

For each interaction, we transform the interaction condition input $\boldsymbol{d}$ into a triplet of tokens $\boldsymbol{h}$:

$$\boldsymbol{h} = (h^s, h^a, h^o) = \text{InToken}(s, a, o, \boldsymbol{b}_s, \boldsymbol{b}_a, \boldsymbol{b}_o), \quad (8)$$

where InToken$(\cdot)$ is a combination of Eqs. (5) to (7) as shown in Fig. 4.

3.3. Interaction Embedding (InBedding)

Interaction is an intricate relationship between subject, object and their action. From Eq. (8), tokens h^s, h^a, h^o are individually embedded (as shown in Fig. 2). For multiple interaction instances, all tokens $h_i^s, h_i^a, h_i^o; i = 1, \cdots, N$, are individually embedded. Therefore, it is necessary to group these tokens by interaction instance and specify different role of tokens within the interaction instance. Segment Embedding [9], has demonstrated its effectiveness in capturing relationships between segments in a text sequence by adding a learnable embedding to tokens to group a sequence of words into segments. In our work, we extend this concept to group the tokens into triplets. Specifically, we add

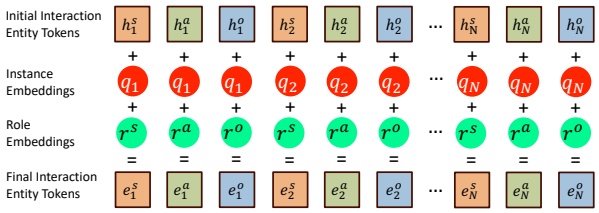

Figure 5. Interaction Embeddings. Learnable instance embedding q and role embedding r are added to tokens to represent intricate interaction relationships between subject s, action a and object o.

a new *instance embedding* denoted as $q \in \{q_1, \ldots, q_N\}$ to interaction instances $\boldsymbol{h} \in \{\boldsymbol{h}_1, \cdots, \boldsymbol{h}_N\}$ as:

$$\boldsymbol{e}_i = \boldsymbol{h}_i + q_i, \tag{9}$$

where all tokens in the same instance share the same instance embedding. This groups all tokens into interaction instances or triplets.

Besides, each token in the triplet has different role. So, we embed their roles with three *role embeddings* $r \in \{r^s, r^a, r^o\}$ to form final entity token $\boldsymbol{e}_i$:

$$\begin{aligned}\boldsymbol{e}_i &= \boldsymbol{h}_i + q_i + r \\ &= (h_i^s + q_i + r^s,\ h_i^a + q_i + r^a,\ h_i^o + q_i + r^o),\end{aligned} \tag{10}$$

where r^s, r^a and r^o represent the role embeddings for subject, action and object respectively. From Eq. (10) we see that tokens of the same role in all instances share the same role embedding. Adding instance and role embedding to the interaction entity token $\boldsymbol{h}_i$ (as in Fig. 5) encodes the intricate interaction relationship, *i.e.* specifies a token's role and interaction instance, which results in significantly improved image generation, especially in scenarios with multiple interaction instances.

3.4. Interaction Transformer (InFormer)

Large-scale T2I models such as Stable Diffusion have been trained on massive-scale image-text pairs and demonstrated remarkable capabilities in generating highly realistic images, owing to the knowledge acquired during large-scale pre-training. In this paper, we aim to incorporate the interaction control into these T2I models with minimal cost. Therefore, it is crucial to preserve the valuable knowledge embedded in them.

Lets denote $\boldsymbol{v} = [v_1, \cdots, v_M]$ as the visual feature tokens of an image, and $\boldsymbol{c}$ as the caption tokens where $\boldsymbol{c} = \tau_\theta(y)$. In LDM models, a Transformer block consists of two attention layers, *i.e.* (i) self-attention layer for the visual tokens and (ii) cross-attention layers that model the attention between visual tokens and caption tokens:

$$\boldsymbol{v} = \boldsymbol{v} + \text{SelfAttn}(\boldsymbol{v}); \quad \boldsymbol{v} = \boldsymbol{v} + \text{CrossAttn}(\boldsymbol{v}, \boldsymbol{c}) \tag{11}$$

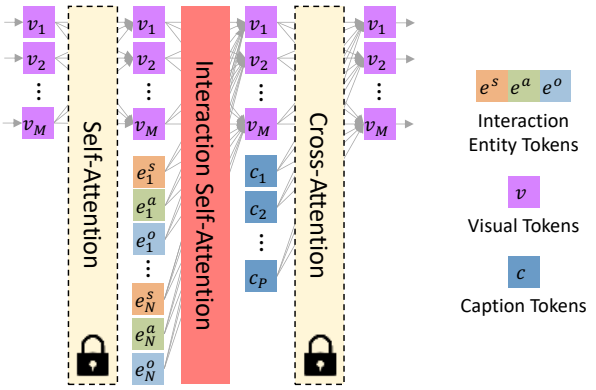

Figure 6. Interaction Transformer. An Interaction Self-Attention is added between the visual token self-attention and the visual-caption cross-attention to incorporate the interaction conditions.

Interaction Self-Attention Following GLIGEN [19], we freeze the two original attention layers and introduce a new gated self-attention layer namely *Interaction Self-Attention* (see Fig. 6) between them. This is to add the interaction condition onto the existing Transformer block. Different from [19], we perform self-attention over the concatenation of visual and interaction tokens $[\boldsymbol{v}, \boldsymbol{e}^s, \boldsymbol{e}^a, \boldsymbol{e}^o]$, which focuses on the relationship of interactions as:

$$\boldsymbol{v} = \boldsymbol{v} + \eta \cdot \tanh\gamma \cdot \text{TS}(\text{SelfAttn}([\boldsymbol{v}, \boldsymbol{e}^s, \boldsymbol{e}^a, \boldsymbol{e}^o])), \tag{12}$$

where $\text{TS}(\cdot)$ is a Token Slicing operation to keep only the output of visual tokens and slice off the others as shown in Fig. 6, η is a hyper-parameter for scheduled sampling that controls the activation of Interaction Self-Attention and γ is a zero-initialized learnable scale that gradually controls the flow of the gate. Note that Eq. (12) performs in between the two parts of Eq. (11). As a summary, our Interaction Self-Attention layer transforms the interaction information, including the interaction, subject and object bounding boxes, into visual tokens.

Scheduled Sampling We set $\eta = 1$ in Eq. (12) during training and standard inference scheme as to [19]. However, in some occasional situations, the newly added Interaction Self-Attention layer could cause sub-optimal effects (poor rendering of non-natural concepts) on existing T2I models. Thus we include a control on sampling interval on the Interaction Self-Attention layer, which can balance out the level of text caption and interaction control.

Technically, our scheduled sampling scheme is controlled during the inference time by a hyper-parameter $\omega \in [0, 1]$. It defines the proportion of diffusion steps influenced by the interaction control as follow:

$$\eta = \begin{cases} 1, & t \leq \omega * T \quad \text{\# Text + Interaction} \\ 0, & t > \omega * T \quad \text{\# Text only} \end{cases} \tag{13}$$

where T is total number of diffusion steps.

3.5. Interaction-conditional Diffusion Model

We combine InToken, InBedding and InFormer to form the pluggable Interaction Module, enabling interaction control in existing T2I diffusion models. The LDM training objective (Eq. (2)) is adopted. Denoting the newly added parameters as θ', the diffusion model is now defined as $\epsilon_{\theta,\theta'}(\cdot)$ where the extra interaction information is processed by the interaction tokenizer $\tau_{\theta'}(\cdot)$. As such, the overall training objective of our model is:

$$\min_{\theta'} \mathcal{L}_{\text{InteractDiffusion}} = \mathbb{E}_{\boldsymbol{z},\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t}\left[\|\epsilon - \epsilon_{\theta,\theta'}(\boldsymbol{z}_t, t, \tau_\theta(y), \tau_{\theta'}(\mathcal{D}))\|_2^2\right]. \quad (14)$$

4. Experiments

We train and evaluate models at 512x512 resolution. We initialize our model with the pre-trained GLIGEN model based on StableDiffusion v1.4. Training uses a constant learning rate of 5e-5 with Adam optimization and a linear warm-up for the initial 10k iterations. It ran for 500k iterations with a batch size of 8 ($\approx$ 106 epochs), taking around 160 hours on 2 NVIDIA GeForce RTX 4090 GPUs. We use a gradient accumulate step of 2, resulting in an effective batch size of 16. For inference, we employ diffusion sampling steps of 50 with the PLMS [20] sampler. More details are given in Sec. 6 of supplementary.

4.1. Datasets

Our experiments were conducted on the widely-used HICO-DET dataset [5], which comprises 47,776 images: 38,118 for training and 9,658 for testing. The dataset includes 151,276 HOI annotations: 117,871 in training and 33,405 in testing. HICO-DET includes 600 types of HOI triplets constructed from 80 object categories and 117 verb classes. We extracted the annotations in the testing set as input to generate interaction images and subsequently performed HOI detection on the generated images using FGAHOI [21].

Following the evaluation methodology outlined in HICO-DET [5], we evaluated the generation results in both Default and Known Object settings. In the Default setting, the average precision (AP) is computed across all testing images for each HOI class. The Known Object setting, on the other hand, calculates the AP of an HOI class solely over the images containing the object in the corresponding HOI class (e.g., the AP of the HOI class 'riding bicycle' is calculated exclusively on the images containing the 'bicycle' object). We reported the HOI detection results in the Full and Rare subsets. The Full and Rare subsets consist of 600 and 138 HOI classes, respectively, with a rare class defined as one represented by less than 10 training samples.

4.2. Evaluation Metrics

We evaluate the quality and controllability of interaction in generation with three metrics.

Fréchet Inception Distance [13] measures the Fréchet distance in distribution of Inception feature between the real-images and the generated images (FID).

Kernel Inception Distance [4] measures the squared Maximum Mean Discrepancy (MMD) between the Inception features of the real and generated images using a polynomial kernel. It relaxes the Gaussian assumption in FID and requires fewer samples.

HOI Detection Score is proposed as a measure of the controllability of interaction in generation models. We utilize the pretrained state-of-the-art HOI detector, FGAHOI [21], to detect the HOI instances in generated images and compare them against the ground truth from the original annotations in HICO-DET. This process quantifies the models' controllability in interaction generation. We report the HOI Detection Score based on the FGAHOI protocol in two categories, namely *Default* and *Known* Object. Default setting is more challenging as it requires distinguishing the non-related images. FGAHOI is implemented with Swin-Tiny and Swin-Large backbones, and we evaluate with the both.

In summary, FID and KID assess generation quality, while HOI Det. Score evaluates interaction controllability.

4.3. Qualitative results

Fig. 7 presents a qualitative comparison with existing methods. The results demonstrate that our model renders the interaction relationship between objects better than others, aligning better with the provided interaction instructions. Other models often exhibit either mismatched actions or inaccurate interactions. For instance, while GLIGEN incorporates layout control to precisely position objects within an image, it fails to capture their intricate interactions. Especially, when multiple interaction instances occur within an image, GLIGEN's rendering of interaction relationships is often mismatched. This challenge persists even in the case of GLIGEN* which is fine-tuned on HICO-DET. While the individual placement (location) of objects is accurate, the interactions between objects appear perplexing.

Our proposed method facilitates improved control over object interaction in image generation. For instance, in Fig. 7(a)-(c), although the interaction appears to be correct in existing works, the interaction details are inaccurate. Our method better renders these details. Moreover, when multiple interacting pairs are involved, as shown in Fig. 7(d), only our method is capable of correctly rendering all pairs of interactions. In Fig. 7(e)-(i), while the interactions (*e.g.* directing airplane, sitting at the dining table, blowing cake, eating pizza, flushing the toilet) in images were inaccurately generated in existing works, our InteractDiffusion well renders these interactions. Our model's capability stems from

Model	Quality ↓		FGAHOI Swin-Tiny (mAP) ↑				FGAHOI Swin-Large (mAP) ↑			
			Default		Known Object		Default		Known Object	
	FID	KID	Full	Rare	Full	Rare	Full	Rare	Full	Rare
StableDiffusion	35.85	0.01297	0.63	0.68	0.66	0.70	0.64	0.83	0.65	0.84
GLIGEN	29.35	0.01275	21.73	15.35	23.31	17.24	23.99	19.56	24.99	20.37
GLIGEN*	18.82	0.00694	25.23	17.45	26.66	18.78	26.45	18.93	27.32	19.90
InteractDiffusion	**18.69**	**0.00676**	**29.53**	**23.02**	**30.99**	**24.93**	**31.56**	**26.09**	**32.52**	**27.04**
HICO-DET	-	-	29.94	22.24	32.48	24.16	37.18	30.71	38.93	31.93

Table 1. Comparison between InteractDiffusion and existing baselines in terms of generated image quality scores in FID and KID and HOI detection score in mAP. GLIGEN* is HICO-DET fine-tuned GLIGEN model. The last row shows the Detection Score from real images.

Model	Tr.	To.	Em.	Quality		Default ↑		Kn. Obj. ↑	
				FID ↓	KID ↓	Full	Rare	Full	Rare
StableDiffusion				35.85	0.01297	0.63	0.68	0.66	0.70
GLIGEN	✓*			29.35	0.01275	21.73	15.35	23.31	17.24
GLIGEN*	✓*			18.82	0.00694	25.23	17.45	26.66	18.78
InteractDiffusion	✓	✓		18.88	0.00686	28.73	21.93	30.15	23.38
	✓	✓	✓	**18.69**	**0.00676**	**29.53**	**23.02**	**30.99**	**24.93**
HICO-DET				-	-	29.94	22.24	32.48	24.16

Table 2. Ablation study of InteractDiffusion. Tr., To., and Em. represent Interaction Transformer, Interaction Tokenizer, and Interaction Embedding respectively. ✓* indicate Gated Self-Attention in GLIGEN.

Figure 7. Visual comparison with existing baselines. In all methods, we use the text caption format of "a person {*action*} a {*object*}". Input and Caption rows represent the interaction conditions, each interaction pair shown by a line link them and is colored differently. GT represents the ground truth images. Ours gains better control to interaction, and renders images matching the text instructions better.

two key components: the InToken for translating interaction conditions into meaningful tokens, and the InBedding for modeling complex interaction relationships.

Fig. 8 shows how InteractDiffusion renders different actions with the same object, in comparison to StableDiffusion and GLIGEN*. This shows that our model can generate various combinations of interactions that maintain the coherence and naturalness of interactions between people and objects. More qualitative results are shown in Sec. 8 and Figs. 13 and 14 of the supplementary, while user preference study is in Sec. 9.

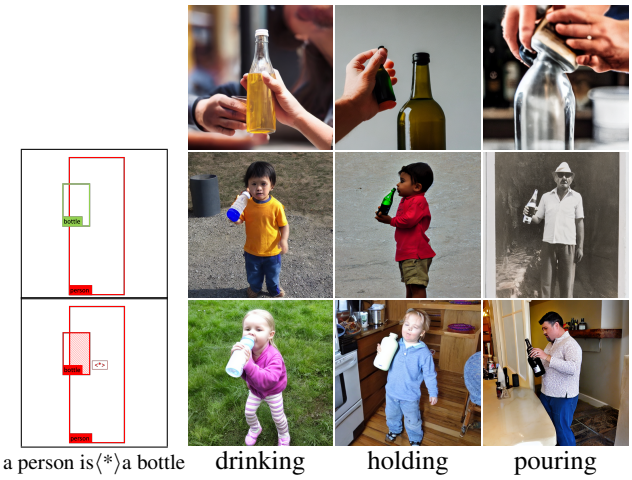

Figure 8. Visualization comparison between StableDiffusion (top), GLIGEN* (middle), and InteractDiffusion (bottom) demonstrates the generation of *different actions* for the same object.

4.4. Quantitative results

Tab. 1 compares our proposed with existing baselines in terms of the quality and interaction controllability, specifically FID, KID, and HOI Detection Score. Compared to the existing baselines, our proposed achieves the best result.

For image generation quality, our proposed produces slightly higher quality than the baselines. It shows that despite additional parameters incorporated into the original model to control interactions, the image generation quality remains unaffected. It is even improved marginally. GLIGEN* exhibits higher image generation quality than StableDiffusion and GLIGEN because we fine-tuned it on the HICO-DET dataset in the same way as InteractDiffusion.

In the HOI Detection Score benchmark, StableDiffusion performs poorly because it does not consider the object's location and size. Comparing GLIGEN and GLIGEN* that only consider the object's location and size, our method encodes interaction control information along with localization information, leading to a significant performance gain.

Using the Tiny backbone for detection, the slight disparity in mAP between the generated images by our method and the real image dataset demonstrates that our approach can generate realistic interactions nearly indistinguishable from real-world interactions by a detection algorithm, such as FGAHOI with a Swin-Tiny backbone. Yet, we have observed that the gap between the real dataset and the generated samples widens when a detector of a large model is used. This indicates that although our generation process outperforms existing baselines, it still has room for further improvement in rendering finer details.

Empirically, the results demonstrate that our proposed enhances interaction controllability while maintaining high-quality image generation capability, thereby significantly outperforming the existing methods in all metrics. This superior performance can be attributed to the proposed components within InteractDiffusion, which include the *InToken* that incorporates new interaction conditions, *InBedding* that encode intricate interaction relationships, and the *InFormer* that injects interaction control into the existing transformer blocks. Collectively, these components constitute a pluggable Interaction Module seamlessly integrated into existing T2I diffusion models.

4.5. Ablation studies

There are three key components that constitute the proposed InteractDiffusion, namely, InToken, InBedding, and InFormer. We conducted an ablation study on these components and tabulated the results in Tab. 2. GLIGEN introduced a gated self-attention layer into the transformer block of the Stable Diffusion model to incorporate additional layout conditions, resulting in a significant performance improvement from 0.63 to 21.73 in mAP. Upon further finetuning on HICO-DET, it achieved an mAP of 25.23.

In InteractDiffusion, we include interaction conditions, alongside layout conditions, to enable the interaction control. With InToken, we convert the interaction conditions (consisting of bounding boxes, object labels, action labels, and relationships) into meaningful interaction entity tokens. Compared to GLIGEN, the incorporation of additional action tokens introduces new information that enhances interaction generation and provides greater interaction control. The inclusion of InToken as a key component further improved the detection score from 25.23 to 28.73, thereby demonstrating its effectiveness. Lastly, we include InBedding to encode the complex interactions relationship, which further improved detection score from 28.73 to 29.53. More ablation studies are shown in Sec. 7 of the supplementary.

5. Conclusion

This paper proposes an interaction-conditioned T2I diffusion model, namely InteractDiffusion, which addresses problems of conditioning generated images beyond the text caption. In existing T2I diffusion models, although several controls (*e.g.* text, images, layout, etc) have been imposed, controlling the interaction in the generated image remains a formidable challenge. Our contributions can be unified as a pluggable interaction module being seamlessly integrated into existing T2I models. The quantitative and qualitative evaluations demonstrate the effectiveness of our method in controlling the interaction of generated content, which significantly outperforms the state-of-the-art approaches.

Acknowledgement This research is supported in part by the National Research Foundation, Singapore, under the NRF Medium Sized Centre Scheme (CARTIN). Any opinions, findings and conclusions expressed in this material are those of the authors and do not reflect the views of National Research Foundation, Singapore.

References

[1] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18370–18380, 2023. 3

[2] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 843–852, 2023. 2

[3] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113*, 2023. 3

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

[5] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE, 2018. 6

[6] Junwen Chen and Keiji Yanai. Qahoi: Query-based anchors for human-object interaction detection. *arXiv preprint arXiv:2112.08647*, 2021. 2

[7] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. *arXiv preprint arXiv:2304.03373*, 2023. 2

[8] Guillaume Couairon, Marlène Careil, Matthieu Cord, Stéphane Lathuilière, and Jakob Verbeek. Zero-shot spatial layout conditioning for text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2174–2183, 2023. 3

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 4

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[11] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3

[12] Chen Gao, Si Liu, Defa Zhu, Quan Liu, Jie Cao, Haoqian He, Ran He, and Shuicheng Yan. Interactgan: Learning to generate human-object interaction. In *ACM MM*, page 165–173, New York, NY, USA, 2020. Association for Computing Machinery. 2

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

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

[15] Tianyu Hua, Hongdong Zheng, Yalong Bai, Wei Zhang, Xiao-Ping Zhang, and Tao Mei. Exploiting relationship for complex-scene image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1584–1592, 2021. 2

[16] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*, 2023. 2

[17] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J. Kim. Hotr: End-to-end human-object interaction detection with transformers. In *CVPR*, 2021. 2

[18] Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7701–7711, 2023. 3

[19] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *CVPR*, pages 22511–22521, 2023. 2, 3, 4, 5

[20] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In *International Conference on Learning Representations*, 2022. 6

[21] Shuailei Ma, Yuefeng Wang, Shanze Wang, and Ying Wei. Fgahoi: Fine-grained anchors for human-object interaction detection. *arXiv preprint arXiv:2301.04019*, 2023. 2, 6

[22] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65:99–106, 2022. 4

[23] Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. *arXiv preprint arXiv:2312.07536*, 2023. 3

[24] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2, 3

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 4

[26] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2, 3

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2, 3, 4

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 4

[29] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, pages 36479–36494, 2022. 2, 3

[30] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 3

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

[32] Suchen Wang, Yueqi Duan, Henghui Ding, Yap-Peng Tan, Kim-Hui Yap, and Junsong Yuan. Learning transferable human-object interaction detector with natural language supervision. In *CVPR*, pages 939–948, 2022. 2, 4

[33] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7452–7461, 2023. 3

[34] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, et al. Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14246–14255, 2023. 2

[35] Hangjie Yuan, Jianwen Jiang, Samuel Albanie, Tao Feng, Ziyuan Huang, Dong Ni, and Mingqian Tang. Rlip: Relational language-image pre-training for human-object interaction detection. *Advances in Neural Information Processing Systems*, 35:37416–37431, 2022. 2

[36] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 2, 3

[37] Guangcong Zheng, Shengming Li, Hui Wang, Taiping Yao, Yang Chen, Shouhong Ding, and Xi Li. Entropy-driven sampling and training scheme for conditional diffusion generation. In *European Conference on Computer Vision*, pages 754–769. Springer, 2022. 2

[38] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *CVPR*, 2023. 2, 3