DreamRelation: Bridging Customization and Relation Generation

Qingyu Shi[1] Lu Qi[3]* Jianzong Wu[1] Jinbin Bai[5] Jingbo Wang[4] Yunhai Tong[1] Xiangtai Li[2,4]*
[1] PKU [2] NTU [3] Insta360 Research [4] Shanghai AI Laboratory [5] NUS

Project page: https://shi-qingyu.github.io/DreamRelation.github.io

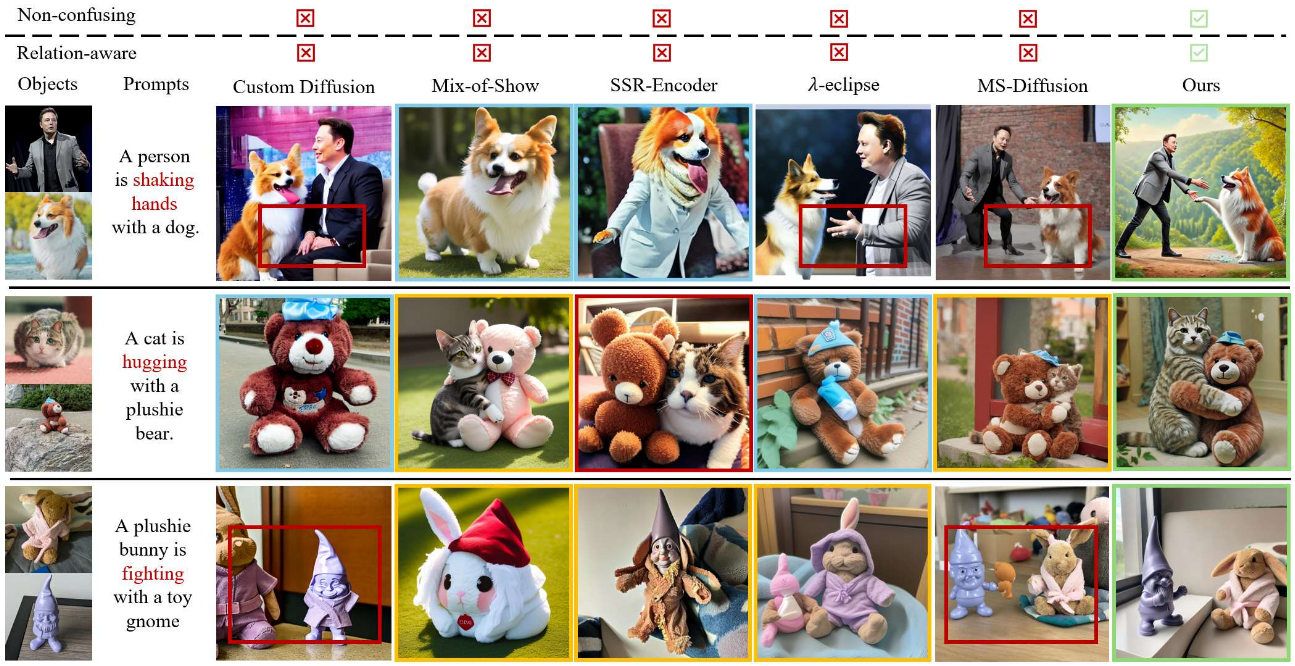

Non-confusing ☒ ☒ ☒ ☒ ☒ ☑
Relation-aware ☒ ☒ ☒ ☒ ☒ ☑
Objects
Prompts
Custom Diffusion
Mix-of-Show
SSR-Encoder
λ-eclipse
MS-Diffusion
Ours
A person is shaking hands with a dog.
A cat is hugging with a plushie bear.
A plushie bunny is fighting with a toy gnome

Figure 1. In our Relation-Aware Image Customization task, the generated images must accurately preserve the relationships between objects and maintain their identity. We highlight the limitations of previous approaches using three color codes: red indicates **failure to capture relationships**, blue marks **missing objects**, and orange represents **object confusion**. Each image is annotated to reflect its specific issue. Our results, highlighted by green boxes, demonstrate the advantages of our proposed method.

Abstract

Customized image generation is essential for creating personalized content based on user prompts, allowing large-scale text-to-image diffusion models to more effectively meet individual needs. However, existing models often neglect the relationships between customized objects in generated images. In contrast, this work addresses this gap by focusing on relation-aware customized image generation, which seeks to preserve the identities from image prompts while maintaining the relationship specified in text prompts. Specifically, we introduce DreamRelation, a framework that disentangles identity and relation learning using a carefully curated dataset. Our training data consists of relation-specific images, independent object images containing identity information, and text prompts to guide relation generation. Then, we propose two key modules to tackle the two main challenges—generating accurate and natural relationships, especially when significant pose adjustments are required, and avoiding object confusion in cases of overlap. First, we introduce a keypoint matching loss that effectively guides the model in adjusting object poses closely tied to their relationships. Second, we incorporate local features of the image prompts to better distinguish between objects, preventing confusion in overlapping cases. Extensive results on our proposed benchmarks demonstrate the superiority of DreamRelation in generating precise relations while preserving object identities across a diverse set of objects and relationships.

[1]* Corresponding authors: Lu Qi, Xiangtai Li.

1. Introduction

Driven by large-scale text-to-image diffusion models [26, 27, 30, 32], customized image generation has recently made significant strides [1, 14, 15, 31, 35, 36, 43]. This task focuses on generating images that preserve the identity of objects from user-provided inputs, enabling the creation of personalized and meaningful content. It has shown value in numerous applications, including personalized artwork, branding, virtual fashion try-ons, social media content creation, augmented reality experiences, and more.

Despite the success of many methods for customizing single or multiple objects [8, 16, 18, 20, 23, 24, 35, 42], they often overlook the relationships between objects and the corresponding text prompts. For instance, when two user-provided objects are paired with a text prompt specifying a particular relationship, the generated output should not only preserve their identities but also accurately reflect the intended relationship, such as a 'hug'. This introduces new challenges in what we refer to as relation-aware customized image generation, which focuses on preserving multiple identities while adhering to relationship prompts.

An intuitive solution to this issue is to adapt existing relation-aware generation methods [11, 21, 22] to tuning-based or training-based customized methods for customizing objects while maintaining the relation. However, both approaches face challenges. Tuning-based methods struggle to preserve multiple identities, as they invert objects into specific tokens. While training-based schemes often fail to balance image and text prompts, frequently overlooking key textual elements and hindering the generation of relationships between objects. As illustrated in Fig. 1, previous methods fail to capture the relationships described in the text prompt and lose identity preservation.

We attribute this failure to two key factors: a lack of relevant data and an ineffective model design. Unlike data augmentation techniques such as flipping or rotation, commonly used to create paired training data in object customization methods [5, 40], our approach requires a triplet of images: two image prompts and one target image. The image prompts should contain similar objects but exhibit distinct poses compared to the target image. To collect these triplets, we propose a data engine to curate our fine-tuning set. We leverage an advanced text-to-image generation model [2] to generate triplets where the same object pair is shared across the images. Through text prompt guidance, the image prompt provides strong identity information. This enables the decoupling of the relationship in the target image, which enhances the relation learning.

For the model design, we propose DreamRelation, which applies the Low-Rank Adaptation (LoRA) [9] strategy to the text cross-attention layers of existing diffusion models to process user-provided text prompts. In DreamRelation, two key modules are introduced during training to enhance the relation generation in customized generation. First, we introduce a keypoint matching loss (KML) as additional supervision to explicitly encourage the model to manipulate object poses, since relationships between objects are closely tied to their poses. Importantly, the KML operates on the latent representation rather than the original image space, aligning with the default diffusion loss. Second, since relation-aware customization requires local features from image prompts, such as the "hands" features for generating "shaking hands"—which are not captured by CLIP's coarse image-level features, we introduce dense features from CLIP. Through partitioning and pooling, we obtain local tokens that contain detail and local information. To further enhance the compatibility between dense features and image-level features, we employ a self-distillation method [39] to improve their alignment.

To more comprehensively evaluate relation-aware customized image generation, we constructed our Relation-Bench based on three established benchmarks [15, 31]. Our methods demonstrate strong performance compared to existing approaches, achieving significant improvements in visual quality and quantitative results.

The contributions of this work are:

- We explore a novel task called relation-aware customized generation, which aims to preserve multiple identities from image prompts while adhering to the relationships specified in text prompts. This task can enhance various user-driven applications by enabling more control.
- We introduce a data engine that uses the recent text-to-image generation model to generate triplet images where the same object is present with distinct poses. This well-curated dataset enables the model to focus on relation learning during fine-tuning, leveraging the identity information provided by the image prompts.
- Our proposed DreamRelation has two key modules, including keypoint matching loss and local token injection, to enhance the relation generation in customized generation. The extensive experiments on three benchmarks demonstrate the effectiveness of our method.

2. Related Work

Diffusion-Based Text-to-Image Generation. Diffusion-based text-to-image models [30, 32] generate high-quality images based on user-provided text prompts. These works encode the text prompt through the text encoder [6, 28] and inject the text embedding into the U-Net's cross-attention layers. Some methods [26, 27, 29, 38] upscale diffusion models and incorporate additional conditions as priors to generate high-resolution images. Meanwhile, several methods [2, 17] introduce stronger text encoders or large language models (LLMs) to enhance the text comprehension capabilities of diffusion models. Recent works [4, 25] have replaced the U-Net denoiser with the Transformers [33].

However, these models primarily focus on text conditions and struggle to handle other forms of guidance, such as user-provided images. In contrast, our work facilitates customized generation while following relation conditions given by text inputs.

Diffusion-Based Customized Generation. Customized image generation aims to produce new images based on user-provided objects. Tuning-based methods accomplish this by fine-tuning specific parameters of diffusion models to learn new objects. Some of these methods [7, 20, 34] employ text embeddings to represent the object identity. On the other hand, DreamBooth [31] fine-tunes the entire U-Net and introduces a prior preservation loss to mitigate language drift. Additionally, several approaches [8, 15, 18, 23] integrate text embeddings and U-Net parameters for image customization. These methods fine-tune both components during object learning, achieving impressive results. Considering the substantial cost of fine-tuning for commercial uses, training-based methods [5, 16, 24, 36, 43] have been proposed. This approach primarily utilizes an encoder to extract the object's identity and inject it into the U-Net. For instance, MS-Diffusion [35] integrates grounding tokens with a Resampler to maintain detailed identity and employs layout guidance to explicitly locate the objects. Despite their efficiency during inference, training-based methods often face challenges in balancing identity preservation and text control. A common issue is that the relationships described in text prompts are frequently overlooked. To address this limitation, we introduce DreamRelation, a framework designed to bridge the customization and relation generation.

Relation-Aware Text-to-Image Generation. Inspired by tuning-based customization methods [7], some recent works [11, 22, 37] represent a "neglected word" by learnable parameters. These methods fine-tune part of the parameters on the content co-existing images. For instance, Reversion [11] fine-tunes the text embedding on a set of relation co-existing images and introduces a relation-steering contrastive loss. Reversion gains a better alignment between relational words and generated images by injecting new text embedding into prompts. While effective with prepositions and adjectives, Reversion struggles with relationships that involve significant overlaps, and cannot customize user-provided objects. Moreover, this method still struggles to generate vivid relationships while maintaining the fidelity of multiple customized objects, likely due to the inherent disregard for text embeddings in customization methods. Our work bridges the gap in the predicate relation-following capability of training-based customization methods. This enables efficient inference that naturally and accurately generates relationships in the text prompts.

3. Bridging Customization and Relation Generation

3.1. Problem definition

Different from conventional customization tasks, we explore a new setting that focuses on image generation by both image prompts $c_i \in \mathcal{R}^{N\times 3\times H\times W}$ and text prompts c_t. We call this setting relation-aware customized image generation due to the requirements that the generated image $\hat{x}$ should strictly preserve each identity and keep the relationship among those identities provided by c_i and c_t.

We define this task as:

$$\hat{x} = \Phi_\theta(c_i, c_t) \tag{1}$$

where Φ_θ is the network parameterized by θ. For brevity clarification, we set $N = 2$ due to the basic triplet element defined in the DreamRelation.

3.2. Discussion and Motivation

While previous works [8, 15, 43] have explored the customization of multiple objects under text control, they lack the ability to effectively manage the relationships between these objects. Specifically, the state-of-the-art multi-object customization model [35] utilizes bounding boxes to enhance multi-object generation. Given the image prompt c_i and the text prompt c_t, it leverages CLIP [28] and the Resampler [35] to extract text tokens tok_{text} and image tokens tok_{image}, respectively. These tokens are then injected into the U-Net $\boldsymbol{\epsilon_\theta}$ through parallel cross-attention layers:

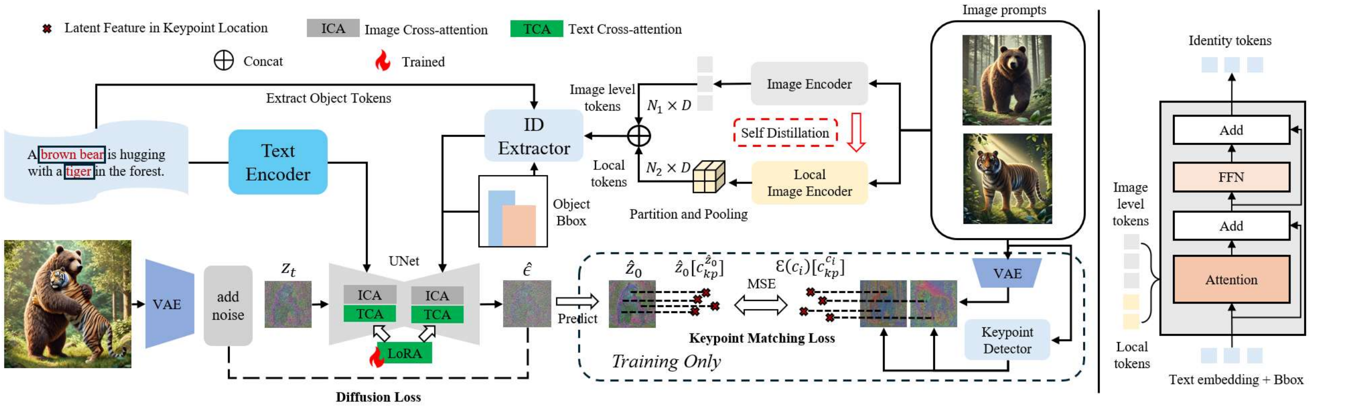
$$h = \text{Softmax}(\frac{QK_i}{\sqrt{d}} + \mathcal{M})V_i + \text{Softmax}(\frac{QK_t}{\sqrt{d}})V_t \tag{2}$$

where Q represents the query, while K_i, V_i and K_t, V_t denote the keys and values obtained from tok_{image} and tok_{text}, respectively. The mask $\mathcal{M}$, generated from the bounding boxes, assists the model in localizing objects during multi-object customization. However, it encounters difficulties when bounding boxes overlap, frequently resulting in object confusion, as shown in Fig. 1. Consequently, relation-aware generation in customized models remains a largely unexplored area of research.

On the other hand, we revisit the relation-aware generation task. Existing relation inversion methods [11, 22] focus on inverting a relationship into a text embedding R^* within pre-trained text-to-image diffusion models. Given a set of images $\{x_k\}_{k=1}^n$ that share the same relationship, these methods use a denoising loss $\mathcal{L}_{RI}$ to fine-tune R^*, ensuring alignment with the specific relationship:

$$\mathcal{L}_{RI} = \mathbb{E}_{z_t,t,\epsilon,c}[\|\epsilon_t - \boldsymbol{\epsilon_\theta}(z_t, t, c_t)\|_2^2] \tag{3}$$

Directly applying relation-aware generation methods to customized models leads to suboptimal performance, as illustrated in Fig. 4. We attribute this to two main reasons.



(a) DreamRelation Pipeline (b) Details of ID Extractor

Figure 2. The overview of DreamRelation. DreamRelation utilizes the off-the-shelf identity extractor to decouple the relation and identity information in relation-specific images. After getting the U-Net output $\hat{\epsilon}$, we predict $\hat{z}_0$ and calculate the keypoint matching loss. The part in the dotted box is only for training.

First, the balance between the control of c_i and c_t is not addressed in relation-aware generation tasks. During inference, the model tends to overlook the target relationship R^* due to the influence of c_i, resulting in incorrect relationship generation. Second, even if the shared relationship exists across $\{x_k\}_{k=1}^n$, the disentangled object identity information complicates the learning process. Fine-tuning solely on $\{x_k\}_{k=1}^n$ makes it challenging for the model to effectively capture the relationship. As shown in the second row of Fig. 4, where ReVersion [11] struggles to generate the ”hugging” relationship even in the absence of c_i. In the following subsections, we will introduce our data collection method and DreamRelation to bridge customization and relation generation.

3.3. Data Collection

To address the balance between c_i and c_t while facilitating relation learning from $\{x_k\}_{k=1}^n$, we first design a data engine for collecting high-quality tuning data. The ideal tuning data should be in the form of $\mathcal{D} = (x_k, c_i, c_t)$. The c_i contain the object in x_k while c_t describes the relationship between objects in x_k. The identity information in c_i helps decouple the relationship within x_k, enhancing relation learning while maintaining a balanced integration of c_i and c_t. Unfortunately, directly cropping the objects from x_k to obtain c_i results in a copy-and-paste effect [5], even when using data augmentation techniques like flipping and rotation. Therefore, we introduce our data engine below.

Relation-aware data engine. Fig. 3 illustrates the difference of our data engine. Both of x_k and c_i are generated by the recent text-to-image generation model, rather than through cropping and augmentation. Specifically, inspired by the performance of DALL-E 3 and its multi-turn dialog capability. We use DALL-E 3 to generate $\mathcal{D} = (x, c_i, c_t)$. Where x_k and c_i share the same object identity by leveraging the prompt “The photo of the same.” We observe that prompting in this manner allows DALL-E 3 to remember and preserve identity in common categories such as ”tiger”, which is enough for relation learning. Next, we employ X-Pose [41], SAM [13], and LLaVA [19] to annotate x, c_i with keypoints, masks, and captions for samples.

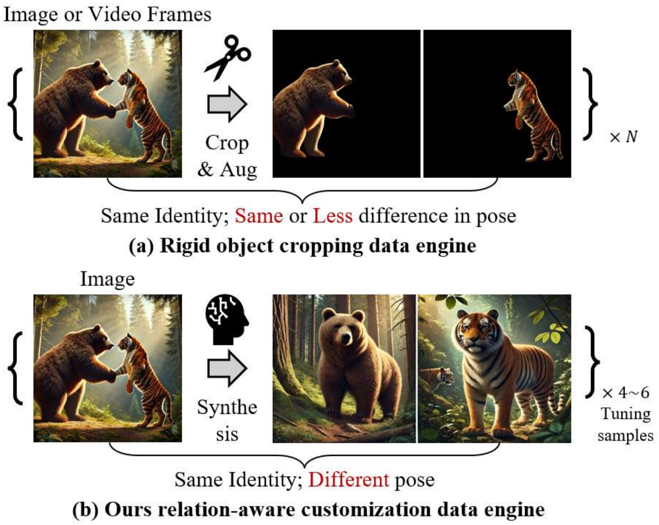


Image or Video Frames
Crop & Aug
× N
Same Identity; Same or Less difference in pose
Image
Synthesis
× 4~6 Tuning samples
Same Identity; Different pose

(a) Rigid object cropping data engine

(b) Ours relation-aware customization data engine

Figure 3. The rigid object cropping data engine, which results in minimal changes to object pose, the cropped image prompts also contain relation information, leading to a copy-and-paste effect, where the text prompt is neglected. Our relation-aware data engine, on the other hand, focuses on relation learning by decoupling identity information from images.

3.4. DreamRelation

3.4.1. Relation Learning

To balance c_i and c_t while enhancing relation learning, we decouple the identity information and relationships in x_k. Specifically, we leverage an off-the-shelf ID extractor from the state-of-the-art customized model [35]. The identity in-

Methods	Single-Object Cus	Multi-Object Cus	Non-Confusing	Relation-Aware
Textual Inversion [7]	✓	✗	✗	✗
DreamBooth [31]	✓	✗	✗	✗
Custom Diffusion [15]	✓	✓	✗	✗
Mix-of-Show [8]	✓	✓	✗	✗
CLIF [18]	✓	✓	✓	✗
MultiBooth [45]	✓	✓	✗	✗
BLIP-Diffusion [16]	✓	✗	✗	✗
ELITE [36]	✓	✗	✗	✗
AnyDoor [5]	✓	✗	✗	✗
SSR-Encoder [43]	✓	✓	✗	✗
λ-eclipse [24]	✓	✓	✗	✗
MS-Diffusion [35]	✓	✓	✗	✗
ReVersion [11]	✗	✗	✓	✓
ADI [10]	✗	✗	✗	✓
DreamRelation	✓	✓	✓	✓

Table 1. Setting Comparison for Different Models. We include several representative methods here. Our proposed DreamRelation can meet the demand of the relation-aware customization task.

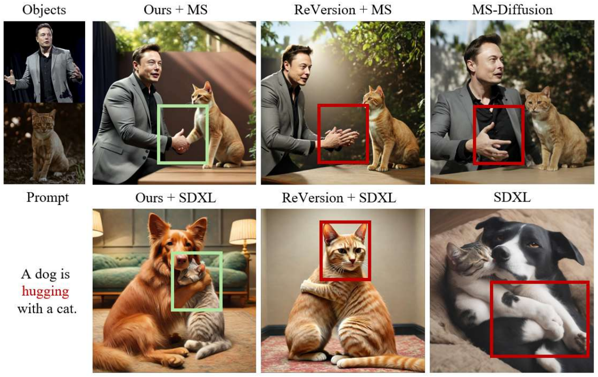


Figure 4. Comparing our method with ReVersion across different base models, our approach demonstrates superior performance in the relation-aware generation task.

formation in c_i enables the model to focus on the relationship information in x_k and c_t during relation learning. To achieve this, we employ parallel cross-attention layers to process c_i and c_t, which can be formulated as:

$$h = \gamma \cdot \text{Softmax}(\frac{QK_i}{\sqrt{d}} + \mathcal{M})V_i + \text{Softmax}(\frac{QK_t}{\sqrt{d}})V_t \quad (4)$$

where $Q = W_q h$, $K_i = W_{k_i} c_i$, $V_i = W_{v_i} c_i$, $K_t = W_{k_t} c_t$, $V_t = W_{v_t} c_t$, and γ is a hyperparameter for scaling control from c_i. For clarity, we omit the final linear layer W_{out}. The relation information mainly depends on text prompts. Therefore, we inject LoRA layers in all of the W_q, W_{k_t}, W_{v_t}, and W_{out} to encourage the model to pay more attention to the relationship in c_t. We freeze all other parameters during fine-tuning. As shown in Fig. 4, benefiting from the decoupled identity information in c_i, our method effectively captures the relationship in x_k and c_t, accurately generating the corresponding images. Notably, after fine-tuning, our method can also be directly integrated into SDXL [27], effectively enabling the model to address the relation-aware generation task. The extensive results are presented in the Supplementary Material (SM).

3.4.2. Keypoint Matching Loss

Most relationships have specific requirements for the pose, such as ”hugging” requiring the arms to cross over the object’s body, and ”riding a bicycle” requiring the feet of the object to be on the pedal plate. Therefore, it’s important to manipulate the pose of objects in the right status accurately. Intuitively, we introduce explicit supervision in the diffusion latent space during the tuning process, named Keypoint Matching Loss (KML).

Specifically, considering our task only involves common objects instead of humans, we employ X-Pose [41] as a keypoint detector because it can detect any keypoints in complex real-world scenarios. We detect 17 keypoints for each object in x_k and c_i, and denote the keypoint coordinates as $c_{kp}^{x_k}, c_{kp}^{c_i} \in \mathcal{R}^{17\times 2}$ respectively. To encourage the model to generate accurate poses in $\mathcal{D}(\hat{z}_0)$, where $\mathcal{D}$ is the VAE decoder. We use the U-Net’s output $\hat{\epsilon}$ to predict z_0 during fine-tuning:

$$\hat{z}_0 = \frac{z_t - \sqrt{1-\bar{\alpha}_t}\hat{\epsilon}}{\sqrt{\bar{\alpha}_t}} \quad (5)$$

We use the VAE encoder $\mathcal{E}$ to obtain the latent representation of image prompts $\mathcal{E}(c_i)$. Then we calculate the MSE loss on the corresponding keypoint locations between $\mathcal{E}(c_i)$ and $\hat{z}_0$:

$$\mathcal{L}_{KML} = \frac{1}{n_{kp}} \mathbb{E}_{\boldsymbol{z_t}, c_i} \left\| \mathcal{E}(c_i)[c_{kp}^{c_i}] - \hat{z}_0[c_{kp}^{x_k}] \right\|_2^2 \quad (6)$$

$$\mathcal{L} = \mathcal{L}_{denoise} + \lambda \cdot \mathcal{L}_{KML} \quad (7)$$

where n_{kp} denote the number of keypoint, $c_{kp}^{c_i}$ and $c_{kp}^{x_k}$ are keypoint’s coordinates in c_i and x_k respectively. We scale the coordinates to fit the size of $\mathcal{E}(c_i)$, $\hat{z}_0$ λ controls for the relative weight of KML. The part inside the dotted box in Fig. 2 illustrates the model fine-tuning with the KML. We find KML is effective in encouraging the model to manipulate the pose for relation-aware generation.

3.4.3. Local Token Injection

When generating relationships, local and detailed features are essential—for instance, the feature of arms to accurately construct relationships like “carrying”. However, the coarse image-level features extracted by the CLIP Image Encoder fail to provide such fine-grained details, leading to confusion between objects in Fig. 7. To address this limitation, we introduce dense features from the CLIP Image Encoder [44]. Specifically, we modify the last layer of the CLIP Image Encoder to obtain dense features: $h_{\text{dense}} = \text{ModifiedAttention}(h'_{\text{clip}})$, where h'_{clip} represents the hidden states of the second-to-last layer.

$$h_{\text{tmp}} = \text{Proj}_v(\text{norm}(h'_{\text{clip}})), \quad (8)$$

$$h_{\text{tmp}} = h'_{\text{clip}} + \text{Proj}_{out}(h_{\text{tmp}}), \quad (9)$$

$$h_{\text{dense}} = h_{\text{tmp}} + \text{FFN}(h_{\text{tmp}}) \quad (10)$$

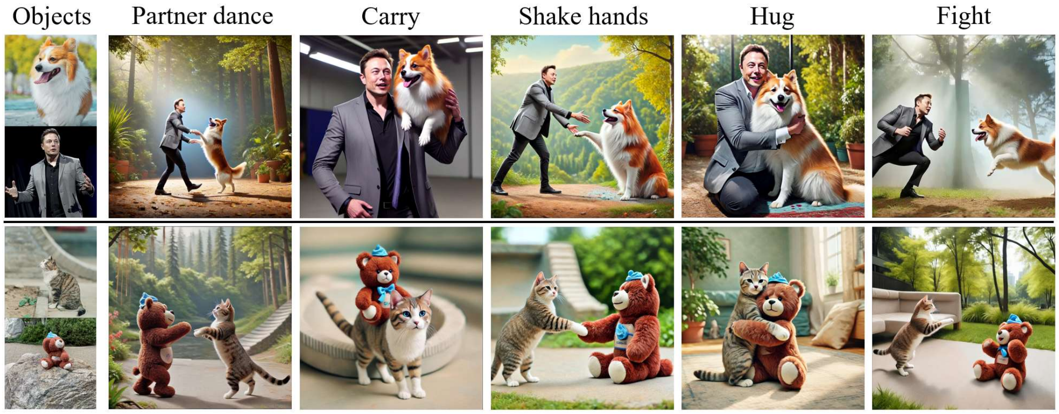


Figure 5. Additional results demonstrate the effectiveness of relation-aware generation. DreamRelation adapts the same pair of objects to different relationships in a natural and accurate manner.

To further enhance the compatibility between dense features and image-level features, we employ self-distillation on the CLIP Image Encoder [39] to derive our Local Image Encoder. Specifically, we align local regions of the dense features with their corresponding image-level features using cosine similarity.

During inference, we extract local tokens tok_{local} from the dense feature h_{dense} through partitioning and pooling. These local tokens are then concatenated with the image-level tokens tok_{image} and passed to the ID extractor. The ID extractor, which is a transformer-based architecture as illustrated in Fig. 2 (b), can be written as:

$$q = q + \text{Attention}(\text{concat}[q,\, tok_{\text{image}},\, tok_{\text{local}}]) \tag{11}$$

Regarding the compatibility between CLIP dense features and the original image-level features, experimental results demonstrate that our Local Image Encoder outperforms other dense representations [3]. For a detailed discussion, we refer readers to the Supplementary Material. Additionally, the injection of local tokens enhances identity preservation, as evidenced by improvements in evaluation metrics and visual results, shown in Fig. 7 and Tab. 3.

4. Experiments

Implementation Details. We generate a tuning set for relation learning with 4-6 samples per relationship. We incorporate LoRA layers into all text cross-attention layers of the U-Net. We set the LoRA rank to $r = 4$, the parallel cross-attention scaling factor to $\gamma = 0.6$, and the keypoint matching loss weight to $\lambda = 1e - 3$. The model is fine-tuned for 500 steps, using 2 A100 GPUs, with a total batch size of 8, completing the process in 10 minutes. We use the Adam optimizer [12] with a learning rate of $1e - 4$ and no weight decay, resulting in a total of 3.1M trainable parameters. Note that our learnable parameters are only in the text cross-attention layers, making it compatible with any SDXL-based models. Considering the strong identity preservation capabilities of MS-Diffusion, we implement our DreamRelation on MS-Diffusion. During fine-tuning and inference, we concatenate local tokens with image tokens. We put more self-distillation details of the Local Image Encoder in the Supplementary Material.

Evaluation. To evaluate relation-aware customized image generation, we propose RelationBench, consisting of 44 objects from DreamBench [31] and CustomConcept101 [15], along with 25 relationships. The object categories include pets, plushies, toys, people, and cartoons. Using GPT-4, we generate 100 cases for single-object and multi-object evaluations. Following previous works [31, 35], we evaluate our method on three metrics: (1) Identity Preservation, which assesses the alignment between the generated images and the image prompts, using the CLIP image score and DINO score that calculate the cosine similarity between the class embeddings of images, referred to as CLIP-I and DINO, respectively; (2) Text Alignment, which evaluates how well the generated images align with the text prompts, using the CLIP image-text score, denoted as CLIP-T; and (3) Relation Alignment. We observed that nouns in the prompts can inflate CLIP-T scores, making them less accurate for evaluating relation generation. To address this, we extract the relation word from the text prompt using spaCy and calculate the CLIP image-text score, denoted as CLIP-R.

4.1. Main Results

We evaluate our method through both quantitative and qualitative results in single- and multi-object cases. In single-object cases, we primarily focus on the alignment of the object's pose with the relationship in the text prompt and the model's ability to preserve identity. In multi-object cases, in addition to relation generation and identity preservation, we also need to ensure no confusion between the multiple objects.

Quantitative results. We compare our method with

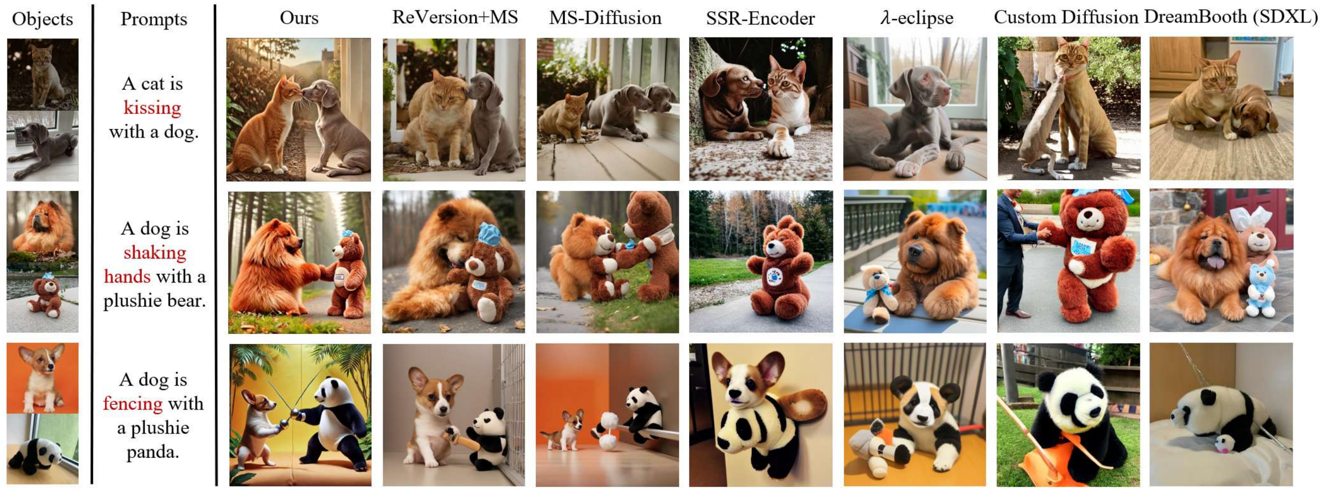

Objects
Prompts
Ours
ReVersion+MS
MS-Diffusion
SSR-Encoder
λ-eclipse
Custom Diffusion
DreamBooth (SDXL)
A cat is kissing with a dog.
A dog is shaking hands with a plushie bear.
A dog is fencing with a plushie panda.

Figure 6. We compare multi-object performance with training-based and tuning-based methods. Additionally, we integrate ReVersion [11] into MS-Diffusion, where the learnable text token is injected into the text prompt during MS-Diffusion's inference, resulting in ReVersion+MS. Our method shows a clear advantage in relation-aware generation and avoiding object confusion in overlapping scenarios.

Method	Single-object				Multi-object			
	CLIP-T	CLIP-R	CLIP-I	DINO	CLIP-T	CLIP-R	CLIP-I	DINO
DreamBooth (SDXL) [31]	27.8	18.2	74.2	62.6	24.3	16.2	67.8	57.2
Custom Diffusion [15]	26.5	15.2	73.2	58.8	20.1	15.4	64.7	55.3
Cones-V2 [20]	24.4	13.5	72.1	57.2	21.3	15.2	64.3	54.2
ELITE [36]	25.7	14.9	75.4	61.5	—	—	—	—
AnyDoor [5]	24.5	14.7	77.4	62.2	21.6	14.9	69.7	59.8
BLIP-Diffusion [16]	26.2	15.7	77.4	57.7	—	—	—	—
SSR-Encoder [43]	25.5	15.9	**80.4**	59.4	24.2	14.6	72.1	56.2
MS-Diffusion [35]	26.5	18.8	78.7	**64.5**	26.9	18.9	73.8	58.8
Reversion+MS [11]	27.8	19.3	77.8	63.1	27.2	19.2	74.2	59.7
Ours	**30.6**	**21.4**	77.9	63.4	**28.9**	**20.4**	**75.4**	**62.1**

Table 2. Quantitative comparison on RelationBench, **Bold** and underline represent the highest and second-highest metrics.

Method	Multi-object			
	CLIP-T	CLIP-R	CLIP-I	DINO
w/o Relation-aware Data	27.3	19.4	75.3	59.8
w/o Local Token Injection	28.5	19.5	75.1	59.9
w/o Keypoint Matching Loss	27.4	19.2	75.2	61.2
Full Model	**28.9**	**20.4**	**75.4**	**62.1**

Table 3. Ablation study on our proposed components

baseline models across three benchmarks: RelationBench, DreamBench, and multi-object cases in CustomConcept101. As shown in Tab. 2, our method demonstrates a clear advantage on RelationBench for single-object cases, particularly in the CLIP-T and CLIP-R metrics. However, its performance is slightly lower in the CLIP-I metric, likely due to significant pose variations in the objects, which may decrease the CLIP-I score. For multi-object generation, our method outperforms other approaches across all evaluation metrics, which we attribute to its ability to effectively avoid object confusion and significantly enhance CLIP-I performance. Furthermore, as shown in Tab. 4, we evaluate single-object generation performance on DreamBench. Our method outperforms others in the CLIP-T metric while delivering competitive results in CLIP-I and DINO. We conduct experiments on multi-object cases from CustomConcept101, which we denote as M-CustomConcept101. As shown in Fig. 5, our model achieves the highest performance in the DINO score and ranks second in both the CLIP-T and CLIP-I metrics.

Qualitative Comparison. We conduct qualitative comparison experiments on RelationBench to evaluate relation-aware customized image generation. As shown in Fig. 6, our method outperforms others in generating multi-object relationships, effectively avoiding object confusion while accurately capturing the intended relationships. Additional results can be found in the SM.

4.2. Ablation study and Analysis

Relation-Aware Data Engine. Instead of fine-tuning on our curated (x_k, c_i, c_t) pairs. We train a text embedding R^*

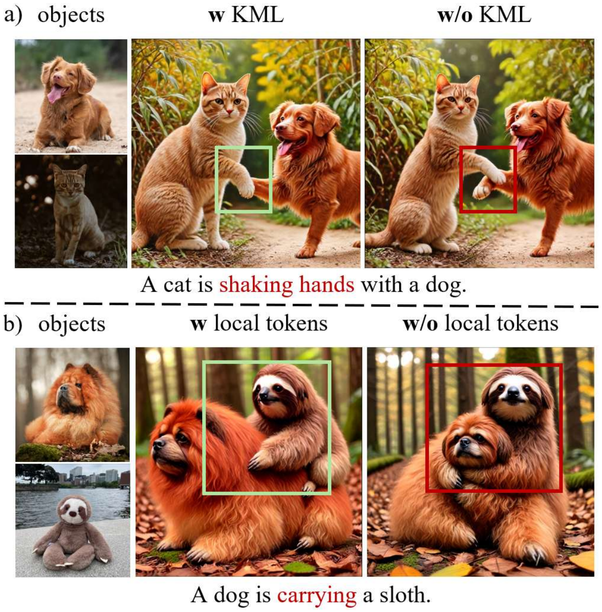


Figure 7. Ablation study of keypoint matching loss (KML) and local tokens: The images are generated using the same random seed. We use green and red boxes to highlight the main differences.

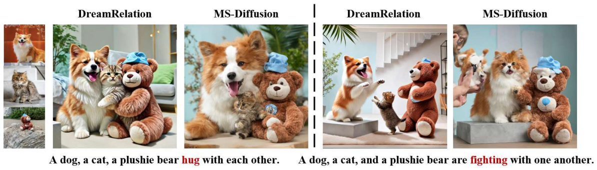

DreamRelation
MS-Diffusion
DreamRelation
MS-Diffusion
A dog, a cat, a plushie bear hug with each other.
A dog, a cat, and a plushie bear are fighting with one another.

Figure 8. DreamRelation for three objects. Compared to the baseline model, DreamRelation shows a clear advantage.

using a set of relation-specific data (x_k, c_i', c_t), where c_i' is cropped from x_k. The well-trained text embedding R^* is injected into the text embedding during inference. The quantitative results in Tab. 3 show that our relation-aware data engine outperforms in both CLIP-T and CLIP-R scores.

Keypoint Matching Loss (KML). We omit KML during fine-tuning. As shown in Fig. 7 a), including the KML leads to more accurate and natural relation-aware customized images using the same random seed. Quantitative results, presented in Tab. 3, show a decline across all evaluation metrics when the KML is removed. Additionally, we experimented with different values for λ and found that the best performance occurs at $\lambda = 1e - 3$.

Local Token Injection. For local token injection, the visual comparison in Fig. 7 b) shows severe object confusion when local tokens are omitted. The quantitative results in Tab. 3 demonstrate that local tokens improve relation-aware customization. We conducted an ablation study on the number of local tokens, as presented in Tab. 6, and determined the optimal number to be 16. Additional ablation studies are shown in the SM. Combining these methods ensures that DreamRelation achieves high-quality relation-aware customized generation.

Method	Single-object		
	CLIP-T	CLIP-I	DINO
DreamBooth (SDXL) [31]	31.2	81.5	**69.2**
Custom Diffusion [15]	28.4	77.2	66.8
Cones-V2 [20]	31.0	76.5	67.2
ELITE [36]	29.8	77.4	62.5
AnyDoor [5]	25.5	**82.1**	67.8
SSR-Encoder [43]	30.8	**82.1**	61.2
MS-Diffusion [35]	31.5	79.3	68.2
Ours	**31.6**	81.7	68.4

Table 4. Evaluation Results on DreamBench.

Method	Multi-object		
	CLIP-T	CLIP-I	DINO
DreamBooth (SDXL) [31]	29.5	67.4	49.2
Custom Diffusion [15]	28.1	66.2	48.8
Cones-V2 [20]	29.2	66.5	47.2
AnyDoor [5]	20.2	**72.1**	51.2
SSR-Encoder [43]	**30.6**	71.1	52.2
λ-eclipse [24]	29.2	68.2	48.2
MS-Diffusion [35]	28.0	70.2	51.2
Ours	29.7	71.4	**52.3**

Table 5. Evaluation Results on M-CustomConcept101.

Num Local Tokens	Multi-object			
	CLIP-T	CLIP-R	CLIP-I	DINO
N=2×2	28.3	19.4	75.3	60.1
N=4×4	**28.9**	**20.4**	**75.4**	**62.1**
N=8×8	27.9	18.1	74.1	59.4

Table 6. Ablation study on the number of local tokens.

Lambda	Multi-object			
	CLIP-T	CLIP-R	CLIP-I	DINO
λ=1e-2	27.5	19.3	72.6	59.6
λ=1e-3	**28.9**	**20.4**	**75.4**	**62.1**
λ=1e-4	26.8	18.1	73.9	60.4

Table 7. Ablation study on λ, controlling relative weight of KML.

5. Conclusion

In this work, we introduce a new challenging task: relation-aware customized image generation. This task aims to generate objects that adhere to the relationship specified in the text prompt and preserve the identities of the user-provided images. To support this, we propose a data engine for generating high-quality fine-tuning data. Our method, DreamRelation, incorporates Keypoint Matching Loss and Local Token Injection, effectively capturing relation information and generating natural relations between customized objects. We also present a fair comparison with other methods on a new benchmark, RelationBench. Extensive experiments demonstrate the effectiveness of our method, highlighting its potential for applications in interactive scenario generation, relation detection, and more. Our research can inspire the direction of relation-aware generation in the community.

Acknowledgement. This work is supported by the National Key Research and Development Program of China (No. 2023YFC3807600).

References

[1] Jinbin Bai, Tian Ye, Wei Chow, Enxin Song, Qing-Guo Chen, Xiangtai Li, Zhen Dong, Lei Zhu, and Shuicheng Yan. Meissonic: Revitalizing masked generative transformers for efficient high-resolution text-to-image synthesis. *arXiv preprint arXiv:2410.08261*, 2024. 2

[2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science.*, 2:3, 2023. 2

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Int. Conf. Comput. Vis.*, 2021. 6

[4] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*, 2024. 2

[5] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *CVPR*, 2024. 2, 3, 4, 5, 7, 8

[6] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 2

[7] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 3, 5

[8] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *NeurIPS*, 2024. 2, 3, 5

[9] Edward J Hu, Yelong Shen, Phillip Wallis abd Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 2

[10] Siteng Huang, Biao Gong, Yutong Feng, Xi Chen, Yuqian Fu, Yu Liu, and Donglin Wang. Learning disentangled identifiers for action-customized text-to-image generation. In *CVPR*, 2024. 5

[11] Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin CK Chan, and Ziwei Liu. Reversion: Diffusion-based relation inversion from images. *arXiv preprint arXiv:2303.13495*, 2023. 2, 3, 4, 5, 7

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[13] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. In *ICCV*, 2023. 4

[14] Lingjie Kong, Kai Wu, Xiaobin Hu, Wenhui Han, Jinlong Peng, Chengming Xu, Donghao Luo, Jiangning Zhang, Chengjie Wang, and Yanwei Fu. Anymaker: Zero-shot general object customization via decoupled dual-level id injection. *arXiv preprint arXiv:2406.11643*, 2024. 2

[15] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 2, 3, 5, 6, 7, 8

[16] Dongxu Li, Junnan Li, and Steven C. H. Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In *NeurIPS*, 2023. 2, 3, 5, 7

[17] Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*, 2023. 2

[18] Wang Lin, Jingyuan Chen, Jiaxin Shi, Yichen Zhu, Chen Liang, Junzhong Miao, Tao Jin, Zhou Zhao, Fei Wu, Shuicheng Yan, et al. Non-confusing generation of customized concepts in diffusion models. *ICML*, 2024. 2, 3, 5

[19] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2024. 4

[20] Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones 2: Customizable image synthesis with multiple subjects. In *NeurIPS*, 2023. 2, 3, 7, 8

[21] Joanna Materzynska, Josef Sivic, Eli Shechtman, Antonio Torralba, Richard Zhang, and Bryan Russell. Customizing motion in text-to-video diffusion models. *arXiv preprint arXiv:2312.04966*, 2023. 2

[22] mengmeng Ge, Xu Jia, Takashi Isobe, Xiaomin Li, Qinghe Wang, Jing Mu, Dong Zhou, liwang Amd, Huchuan Lu, Lu Tian, Ashish Sirasao, and Emad Barsoum. Customizing text-to-image generation with inverted interaction. In *ACM MM*, 2024. 2, 3

[23] Lianyu Pang, Jian Yin, Baoquan Zhao, Feize Wu, Fu Lee Wang, Qing Li, and Xudong Mao. Attndreambooth: Towards text-aligned personalized text-to-image generation. *arXiv preprint arXiv:2406.05000*, 2024. 2, 3

[24] Maitreya Patel, Sangmin Jung, Chitta Baral, and Yezhou Yang. λ-eclipse: Multi-concept personalized text-to-image diffusion models by leveraging clip latent space. *arXiv preprint arXiv:2402.05195*, 2024. 2, 3, 5, 8

[25] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 2

[26] Pablo Pernias, Dominic Rampas, Mats Leon Richter, Christopher Pal, and Marc Aubreville. Würstchen: An efficient architecture for large-scale text-to-image diffusion models. In *ICLR*, 2023. 2

[27] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 5

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual

models from natural language supervision. In *ICML*, 2021. 2, 3

[29] Jingjing Ren, Wenbo Li, Haoyu Chen, Renjing Pei, Bin Shao, Yong Guo, Long Peng, Fenglong Song, and Lei Zhu. Ultrapixel: Advancing ultra-high-resolution image synthesis to new peaks. *arXiv preprint arXiv:2407.02158*, 2024. 2

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[31] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 2, 3, 5, 6, 7, 8

[32] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 2

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2

[34] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. p+: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023. 3

[35] X Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. *arXiv preprint arXiv:2406.07209*, 2024. 2, 3, 4, 5, 6, 7, 8

[36] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. ELITE: encoding visual concepts into textual embeddings for customized text-to-image generation. In *ICCV*, 2023. 2, 3, 5, 7, 8

[37] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion. In *CVPR*, 2024. 3

[38] Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. Motionbooth: Motion-aware customized text-to-video generation. *arXiv preprint arXiv:2406.17758*, 2024. 2

[39] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. *ICLR*, 2024. 2, 6

[40] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *CVPR*, 2023. 2

[41] Jie Yang, Ailing Zeng, Ruimao Zhang, and Lei Zhang. Unipose: Detecting any keypoints. *ECCV*, 2024. 4, 5

[42] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arxiv:2308.06721*, 2023. 2

[43] Yuxuan Zhang, Yiren Song, Jiaming Liu, Rui Wang, Jinpeng Yu, Hao Tang, Huaxia Li, Xu Tang, Yao Hu, Han Pan, et al. Ssr-encoder: Encoding selective subject representation for subject-driven generation. In *CVPR*, 2024. 2, 3, 5, 7, 8

[44] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *ECCV*, 2022. 5

[45] Chenyang Zhu, Kai Li, Yue Ma, Chunming He, and Li Xiu. Multibooth: Towards generating all your concepts in an image from text. *arXiv preprint arXiv:2404.14239*, 2024. 5