# GR-Diffusion: Graph-guided Relational-aware Diffusion via Attention Alignment

Xiaochen Liu[1] Xiaoting Xi[1] Chao Yin[1] Xiaoqiang Li[1,*] Daoguo Dong[2,*]

[1]School of Computer Engineering and Science, Shanghai University

[2] Institute of Trustworthy Embodied AI, Fudan University

{liuxc, xixiaoting, yincaho, xqli}@shu.edu.cn dgdong@fudan.edu.cn

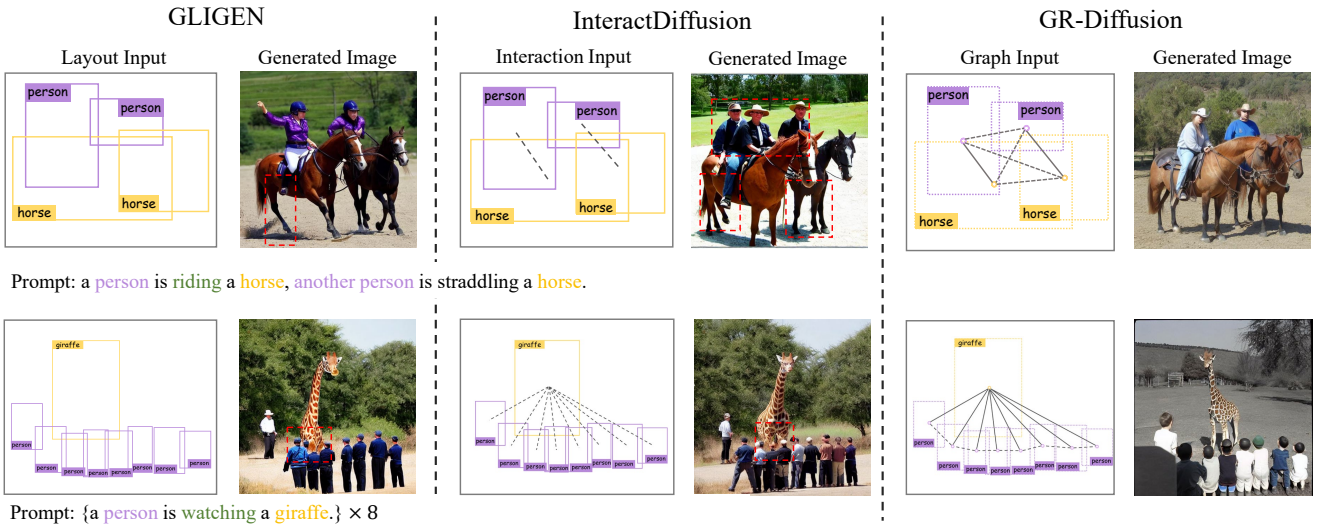


Figure 1. Generated 512x512 samples. GLIGEN [15] conditions on layout inputs. InteractDiffusion [11] introduces interaction labels and corresponding locations. Our proposed GR-Diffusion operates as a training-free guidance mechanism by reframing the interaction directives as a Target Scene Graph to enforce precise structural and relational alignment.

## Abstract

*Large-scale text-to-image diffusion models excel at generating high-fidelity images but struggle with control over complex human-object interaction (HOI), due to guidance conflicts between layout and interaction constraints. In this work, we introduce **G**raph-guided **R**elational-aware Diffusion (GR-Diffusion), a training-free framework for precise control over complex HOI in diffusion models. GR-Diffusion leverages a Target Scene Graph (TSG) as a structural guidance to steer the internal attention at each denoising step via two plug-and-play modules. First, to control the spatial layout, the Node Alignment Guidance (NAG) module guides the cross-attention maps by reducing the structural deviation between the TSG and a dynamic attention graph. Subsequently, to reinforce the semantic interactions, the Edge Enhancement Guidance (EEG) module constructs a relational mask from the corrected cross-attention maps and injects the mask into the self-attention layers. Our GR-Diffusion achieves state-of-the-art control over both spatial layout and semantic interactions on the HICO-DET benchmark, and significantly outperforms existing baselines in both the HOI detection score and image fidelity measured by FID and KID.*

## 1. Introduction

Text-to-Image (T2I) diffusion models [21, 23] excel at generating high-fidelity images, yet they inherently lack precise control over the generated content. To address this, existing approaches have investigated various conditioning mechanisms, including class[7, 34], text [21–23, 25], spatial layouts [1, 5, 15, 29, 35] and scene graph [21, 24], images cues

*Corresponding author.

(*e.g.*, edges and skeletons) [1, 12, 33]. While these methods have enhanced controllability, they often fall short in scenarios that demand a deeper level of semantic understanding, particularly in modeling the complex human-object interaction (HOI).

**(a) Guidance Conflict:** Naively enforcing complex interaction conditions simultaneously with layout constraints often induces attention conflicts, leading to attribute drift and incorrect number of entities. As shown in Fig. 1, the complex interactions between *"person"* and *"giraffe"* cause attribute drift, resulting in blended entities. Furthermore, *"person"* and *"horse"* within a specific layout results in incorrect number of legs and people.

**(b) Implicit Interaction Modeling:** Existing methods, whether based on triplets [11] or layout boxes [26], primarily rely on a weak, implicit signal for the interaction relationship. For instance, existing methods fail to model a coherent relational structure of the *'person riding horse'* pairs in Fig. 1. The lack of an explicit semantic link results in spatially adjacent but semantically non-interacting entities.

**(c) Generalization and Efficiency Trade-offs:** Existing paradigms force a dilemma between two types of computational cost. Training-based methods suffer from high training costs and architectural rigidity, requiring costly retraining to generalize. Conversely, training-free methods introduce high inference costs by relying on external, large-scale LLMs or VLMs for guidance.

To address these issues, we introduce **GR-Diffusion**, a plug-and-play training-free framework for high-fidelity interaction control. The core of GR-Diffusion is the Target Scene Graph (TSG), an explicit representation of the desired scene where nodes define the spatial layout of entities, while edges define the semantic interactions. The TSG reframes the HOI control problem as a graph structural alignment task, which serves as a structural guidance to decouple and steer the spatial layout and semantic interactions at each denoising step.

GR-Diffusion deploys the TSG via two key guidance modules operating at different stages of the U-Net: First, to control the spatial layout, the NAG module constructs a dynamic attention graph from the current cross-attention layers of the UNet, compares it to the TSG to quantify the layout deviation, and then applies an adaptive warping operation to correct the cross-attention maps. The NAG module ensures entities are precisely positioned prior to modeling of their interactions. Subsequently, to enforce semantic interactions defined by the graph's edges, the Edge Enhancement Guidance (EEG) module leverages the corrected entity locations provided by NAG. The EEG module constructs a relational mask corresponding to the TSG's edges, which is injected into self-attention layers. The EEG module reinforces the mutual focus between interacting entities, constructing implicit interactions. Through the dynamic graph structural alignment mechanism, GR-Diffusion addresses the challenges of guidance conflict and implicit interaction modeling.

Our main contributions can be summarized as follows:

- We propose **GR-Diffusion**, a training-free framework for high-fidelity Human-Object Interaction (HOI) control. The core of GR-Diffusion is the Target Scene Graph (TSG), which reframes the HOI control as a graph alignment task, enabling an adaptive guidance mechanism that generalizes across different diffusion models.
- To address the guidance conflict and implicit relational modeling challenges, GR-Diffusion introduces an adaptive mechanism to align the model's internal attention with the TSG. First, the Node Alignment Guidance (NAG) module enforces the TSG node layout via cross-attention maps. Subsequently, the Edge Enhancement Guidance (EEG) module injects an edge-derived relational mask to reinforce semantic interactions.
- Through comprehensive experiments, we demonstrate that GR-Diffusion achieves **state-of-the-art** performance on the HICO-DET benchmark. It significantly outperforms existing baselines in both interaction controllability and image fidelity.

## 2. Related Work

**Controllable Image Generation.** Research in controllable image generation aims to provide explicit control beyond a single text prompt. Training-based approaches involve integrating auxiliary modules into pre-trained diffusion models [23], such as lightweight encoders for spatial conditions [20, 33]. A significant line of this research has focused on precise layout and instance control. GLIGEN [15] proposed trainable gated self-attention layers to condition generation on bounding boxes. This was followed by methods focusing on instance-level control [27, 35], and more advanced architectures for complex layout generation [31, 32]. Other methods target compositional control [12] or use LLMs for enhanced guidance [6, 14].

Training-free approaches achieves control purely at inference time by manipulating the model's attention maps. This was pioneered by Prompt-to-Prompt [9] for style editing and was later adapted for layout control in methods like Layout-Guidance [5] and FreeControl [19].

**Human-Object Interactions Generation.** A more challenging task is the explicit control of Human-Object Interactions (HOI) [4, 8]. This requires moving beyond simple layout to model the specific semantic relationships between entities. Research on this specific problem has also split into two paths.

Training-based methods aim to bake interaction control directly into the model weights. InteractDiffusion [11] proposed a pluggable module trained on HOI datasets to integrate ⟨subject, action, object⟩ triplets. This line of work

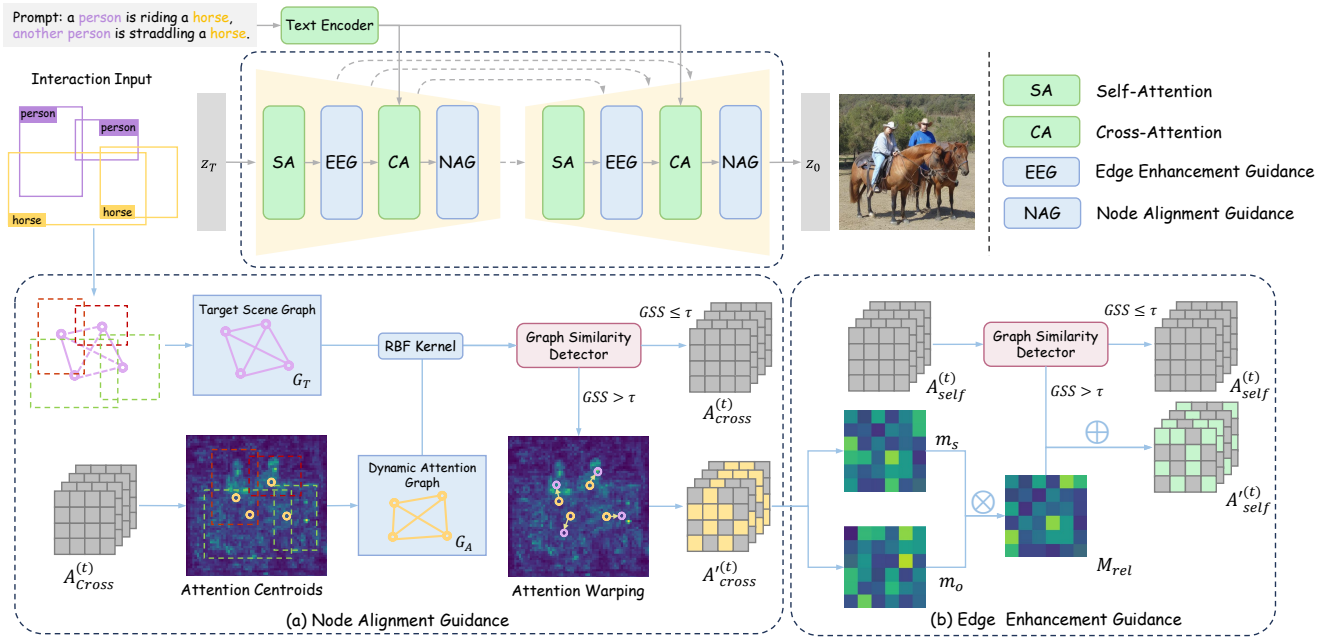


Figure 2. Overview of GR-Diffusion. GR-Diffusion aligns attention of a pre-trained U-Net with a Target Scene Graph (TSG): (a) NAG corrects the spatial layout corresponding to the TSG's nodes, computing a layout deviation by comparing the cross-attention map to the TSG and applying an attention warping operation to produce a corrected map. (b) EEG reinforces semantic interactions corresponding to the TSG's edges, leveraging corrected spatial priors from (a) to construct a relational mask, which is injected into the self-attention layers.

has been extended by methods like VerbDiff [3], which fine-tunes the model to disambiguate verb semantics, EmIT [30], which introduces modules to control the emotional context of interactions. While effective, these methods rely on extensive, task-specific training, limiting their flexibility.

Training-free methods achieve control at inference stage, which often rely on large external models. One category leverages Large Language Models (LLMs). CEIDM [28] uses LLMs to mine relationships for generation refinement. B2B [26] utilizes LLMs to generate pre-defined layout boxes. Another category employs external Visual Language Models (VLMs), such as ReCorD [13], which utilizes a multi-step reasoning and correcting loop.

This leaves a critical gap for a unified HOI generation framework that is simultaneously training-free, efficient (*i.e.*, without external LLMs or VLMs), and explicitly models both layout and semantic relationships in a unified manner. Therefore, we propose a novel training-free approach GR-Diffusion, which is the first to fill this gap by reframing the control problem as a graph structural alignment task at inference stage.

## 3. Method

We propose GR-Diffusion, a training-free guidance framework to generate images with complex human-object interactions, as illustrated in Fig. 2. GR-Diffusion first models the interaction directive as a Target Scene Graph (TSG), which provides a structural guidance for the denoising process. GR-Diffusion comprises two key guidance modules: (a) **Node Alignment Guidance**, which corrects object layout by aligning the model's cross-attention maps with the TSG's nodes, and (b) **Edge Enhancement Guidance**, which reinforces semantic interactions by injecting a relational mask, corresponding to the TSG's edges, into the self-attention layers.

### 3.1. Target Scene Graph Formulation

Our training-free guidance mechanism operates on Stable Diffusion [23], a pre-trained Latent Diffusion Model (LDM). LDMs utilize a U-Net $\epsilon_\theta$ to perform the diffusion process in a computationally efficient latent space. During inference, the model iteratively denoises a random latent $z_T$ to produce $z_0$. For our method, text conditions $c$ are injected into the U-Net via cross-attention layers.

To address Human-Object Interaction (HOI) generation, our plug-in guidance steers this pre-trained LDM by interpreting a set $\mathcal{D}$ of $N$ HOI directives. Unlike frameworks that require task-specific training [11], our guidance operates directly on the pre-trained model. We define each directive $d_i \in \mathcal{D}$ by its core semantic triplet and the spatial layout of its participating entities:

$$d_i = (\langle s_i, a_i, o_i \rangle, b_{s_i}, b_{o_i}), \tag{1}$$

where $\langle s_i, a_i, o_i \rangle$ are the labels for the (subject, action, object) triplet, and $(b_{s_i}, b_{o_i})$ are the respective bounding boxes for the subject and object.

While this set of directives $\mathcal{D}$ provides structured information, it represents a flat list of constraints. To better model the holistic scene structure, our core idea is to reframe this set of directives into a unified representation. We introduce the **Target Scene Graph (TSG)** $\mathcal{G}_T = (V, E_T)$ as a guidance. The TSG is constructed directly from the directives $\mathcal{D}$ as follows:

- **Nodes ($V$):** The $M$ unique entities (subjects and objects) from $\mathcal{D}$ form the vertex set. This set $V$ is defined as the collection of $M$ target position vectors, $V = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_M\}$, where the node $\boldsymbol{v}_j \in \mathbb{R}^2$ *is* its target center coordinate, derived from its bounding box $b_j$. Each node $\boldsymbol{v}_j$ is also associated with its original class label.
- **Edges ($E_T$):** The interactions from $\mathcal{D}$ form the edges. An edge $e_{jk} \in E_T$ exists if the entities corresponding to nodes $\boldsymbol{v}_j$ and $\boldsymbol{v}_k$ are linked by an action, and is attributed with the corresponding action label.

This graph-based formulation explicitly decouples the spatial layout (nodes $V$) from the semantic interactions (edges $E_T$). Our guidance mechanism, detailed next, is designed to align the model's internal attention with this structured representation.

### 3.2. Node Alignment Guidance

The Node Alignment Guidance (NAG) module is designed to enforce the spatial layout defined by the Target Scene Graph (TSG). The NAG module operates by constructing a Dynamic Attention Graph (DAG), $\mathcal{G}_A^{(t)} = (U^{(t)}, E_A^{(t)})$, at each guided timestep $t$ and aligning its structure to $\mathcal{G}_T$. The vertex set of the DAG, $U^{(t)} = \{\boldsymbol{u}_1^{(t)}, \ldots, \boldsymbol{u}_M^{(t)}\}$, is composed of the corresponding centroids [16, 26] $\boldsymbol{u}_j^{(t)}$ calculated from the cross-attention maps at timestep $t$. The module functions in a three-step process: (1) calculating these centroids $\boldsymbol{u}_j^{(t)}$ to define $\mathcal{G}_A^{(t)}$, (2) computing the similarity between $\mathcal{G}_A^{(t)}$ and $\mathcal{G}_T$, and (3) applying a region-confined attention warp to correct the layout.

**Dynamic Attention Graph Construction.** The NAG module constructs the DAG $\mathcal{G}_A^{(t)}$, representing the model's current spatial beliefs. Its vertices $U^{(t)}$ are computed by identifying the spatial centroid $\boldsymbol{u}_j^{(t)}$ for each entity. The process begins by extracting the cross-attention map $\boldsymbol{A}_{cross}^{(t)} \in \mathbb{R}^{H \times W \times K}$ from the U-Net. As shown in Fig. 3, the $j$-th entity correspond to a set of text tokens $K_j$, its average attention map $\boldsymbol{A}_{cross,j}^{(t)}$ is computed by averaging the maps for all tokens $k \in K_j$:

$$\boldsymbol{A}_{cross,j}^{(t)} = \frac{1}{|K_j|} \sum_{k \in K_j} \boldsymbol{A}_{cross,(:,:,k)}^{(t)}. \tag{2}$$

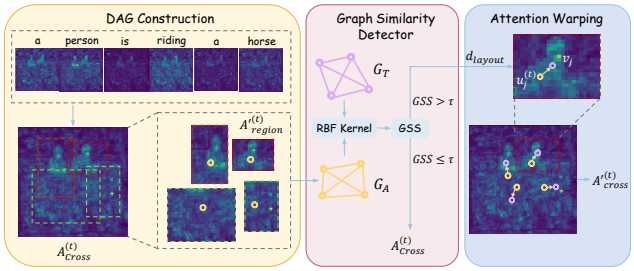


Figure 3. Overview of the Node Alignment Guidance (NAG), which corrects the spatial layout based on the deviation between the Dynamic Attention Graph and the Target Scene Graph.

To accurately reflect the attention scores inside the target bounding box $b_j$, a binary mask $\boldsymbol{M}_j \in \{0,1\}^{H \times W}$ is applied to its spatial region. As illustrated in Fig. 3, the centroid $\boldsymbol{u}_j^{(t)} \in \mathbb{R}^2$ is then computed as the center of scores only within this masked region. This calculation, shown in Eq. (3), first isolates the regional attention $\boldsymbol{A}_{region,j}^{(t)}$ and then calculates the weighted average of coordinates, normalized by the total regional attention scores $Z_j^{(t)}$.

$$\begin{aligned} \boldsymbol{A}_{region,j}^{(t)} &= \boldsymbol{A}_{cross,j}^{(t)} \odot \boldsymbol{M}_j, \\ \boldsymbol{u}_j^{(t)} &= \frac{1}{Z_j^{(t)}} \sum_{x=1}^{W} \sum_{y=1}^{H} (\boldsymbol{A}_{region,j}^{(t)})_{x,y} \begin{pmatrix} x \\ y \end{pmatrix}, \end{aligned} \tag{3}$$

as shown in Eq. (3), where $Z_j^{(t)} = \sum_{x,y} (\boldsymbol{A}_{region,j}^{(t)})_{x,y}$ is the normalization factor representing the total attention scores within the region. This localized calculation ensures the centroid is not skewed by spurious attention elsewhere in the map. The set of $M$ calculated centroids $U^{(t)}$ thus forms the vertices of $\mathcal{G}_A^{(t)}$.

**Graph Similarity Detector.** To quantify the discrepancy, The Graph Similarity Detector (GSD) compare the structure of $\mathcal{G}_A^{(t)}$ with $\mathcal{G}_T$. We represent both graphs using weighted adjacency matrices, $W_A^{(t)} \in \mathbb{R}^{M \times M}$ and $W_T \in \mathbb{R}^{M \times M}$. The weight $w_{jk}$ between any two nodes $j$ and $k$ in $W_T$ is computed based on their pairwise Euclidean distance $\|\boldsymbol{v}_j - \boldsymbol{v}_k\|_2$ (similarly for $W_A^{(t)}$), transformed by an RBF kernel:

$$W_{T,(j,k)} = \exp\left(-\frac{\|\boldsymbol{v}_j - \boldsymbol{v}_k\|_2^2}{2\sigma^2}\right) \cdot S_{jk}, \tag{4}$$

where $\sigma$ is the kernel bandwidth and $S_{jk}$ is a semantic prior. We then compute the **Graph Structural Similarity (GSS)** as the normalized cosine similarity of their adjacency matrices:

$$\mathrm{GSS}(\mathcal{G}_A^{(t)}, \mathcal{G}_T) = \frac{\mathrm{vec}(W_A^{(t)}) \cdot \mathrm{vec}(W_T)}{\|\mathrm{vec}(W_A^{(t)})\|_2 \|\mathrm{vec}(W_T)\|_2}. \tag{5}$$

The GSS score serves as an adaptive detector. We define $\tau$ as a hyperparameter representing the similarity threshold for activating the guidance. The guidance is only applied if $GSS < \tau$. When guidance is active, the layout deviation is defined as $d_{layout} = 1 - GSS$.

**Region-Confined Attention Warping.** The magnitude of the guidance, $f_{warp}$, is first made proportional to the current layout deviation:

$$f_{warp} = 1 + \gamma \cdot d_{layout}. \tag{6}$$

This force magnitude is then translated into a concrete spatial adjustment for each entity $j$. We compute a displacement vector $\Delta \boldsymbol{p}_j$ that points from the entity's current centroid $\boldsymbol{u}_j^{(t)}$ towards its target position $\boldsymbol{v}_j$. The direction of $\Delta \boldsymbol{p}_j$ is given by the unit vector from $\boldsymbol{u}_j^{(t)}$ to $\boldsymbol{v}_j$, while its magnitude is determined by scaling the original distance between them by the adaptive factor $f_{warp}$:

$$\Delta \boldsymbol{p}_j = f_{warp} \cdot (\boldsymbol{v}_j - \boldsymbol{u}_j^{(t)}).$$

The vector $\Delta \boldsymbol{p}_j$ is used to define an affine transformation for the entity's corresponding attention region. This affine transformation generates a sampling grid, which is applied to warp the original cross-attention map $\boldsymbol{A}_{cross,j}^{(t)}$ via grid sampling, producing the spatially corrected map $\boldsymbol{A}_{cross}^{\prime(t)}$. Through the region-confined attention warping, the attention map of each entity is effectively shifted, pulling its centroid $\boldsymbol{u}_j^{(t)}$ closer to the target $\boldsymbol{v}_j$.

### 3.3. Edge Enhancement Guidance

While the NAG module corrects the spatial layout (nodes), the Edge Enhancement Guidance (EEG) module is responsible for enforcing the semantic interactions defined by the graph's edges ($E_T$). Correctly generating an interaction requires establishing a strong contextual link between the corresponding pixel regions. This link is primarily modeled by the self-attention mechanism. The EEG module therefore directly manipulates the self-attention maps to structurally reinforce these crucial relationships defined by all edges in $E_T$, the module performs the following two steps:

**Edge Relation Mask Construction.** While the EEG module executes before NAG in the network block (as shown in Fig. 2), it utilizes the corrected cross-attention maps provided by the NAG module which is cached from its previous execution. Using the cached cross-attention maps, the Edge Enhancement Guidance (EEG) module is responsible for enforcing the semantic interactions defined by the graph's edges ($E_T$). First, all entity nodes in the Target Scene Graph $\mathcal{G}_T$ that participate in an interaction are conceptually divided into two sets: the set of subject nodes and the set of object nodes. The EEG module then creates subject map and object map by averaging the corrected cross-attention maps for all text tokens corresponding to them:

These aggregated maps ($m_s, m_o \in \mathbb{R}^{HW}$) represent the model's spatial belief for the subject nodes and object nodes as a whole. An Edge Relation Mask, $M_{rel} \in \mathbb{R}^{HW \times HW}$, is then constructed via an outer product to encourage mutual attention between these two regions:

$$M_{rel} = m_s m_o^T + m_o m_s^T,$$

where $M_{rel}$ effectively represents the desired relational structure for all interactions defined in $E_T$.

**Adaptive Self-Attention Injection.** The final step is to inject aggregate mask $M_{rel}$ into the self-attention layers. Let the original self-attention map at step $t$ be $\boldsymbol{A}_{self}^{(t)} = \text{softmax}(\boldsymbol{S}_{self}^{(t)})$, where $\boldsymbol{S}_{self}^{(t)}$ are the raw attention scores. The guided self-attention map $A_{self}^{\prime(t)}$ is then computed as:

$$\boldsymbol{A}_{self}^{\prime(t)} = \text{softmax}(\boldsymbol{S}_{self}^{(t)} + \beta \cdot M_{rel}),$$

where the guidance strength $\beta$ is made adaptive based on the GSS score computed and cached during the NAG phase: $\beta = \beta_{base} \cdot d_{layout}$. This formulation ensures that poor layout alignment triggers a stronger self-attention correction.

## 4. Experiments

This section presents a comprehensive evaluation of the proposed GR-Diffusion framework. Experiments are conducted on top of InteractDiffusion (v1.2) and InteractDiffusion-XL (v1.0) [11]. All images are generated at a 512x512 resolution. Consistent with the baseline setup [11], The diffusion process utilizes the PLMS [17] sampler with 50 inference steps and a guidance scale of 7.5. The key hyperparameters for the guidance mechanism, including the the GSS scaling factor $\gamma$, GSS threshold $\tau$ and the base self-attention strength $\beta_{base}$, are held constant across all experiments to ensure a fair comparison.

### 4.1. Datasets

Our experiments are conducted on the widely-used HICO-DET dataset [4]. This large-scale benchmark is the standard for evaluating human-object interaction understanding and generation. It contains 47,776 images in total, split into a training set of 38,118 images and a test set of 9,658 images. The dataset is annotated with 117,871 HOI instances in the training set and 33,405 in the test set. It encompasses 600 distinct HOI categories, which are combinations of 117 verb classes and 80 object classes.

Following the standard protocol established by prior work [11], we use the ground-truth annotations (HOI triplets and bounding boxes) from the HICO-DET test set as the conditional inputs for our generation task. We then evaluate the generated images against these ground-truth directives. We report our results on the Full subset (all 600 HOI classes) and the more challenging Rare subset (138 classes

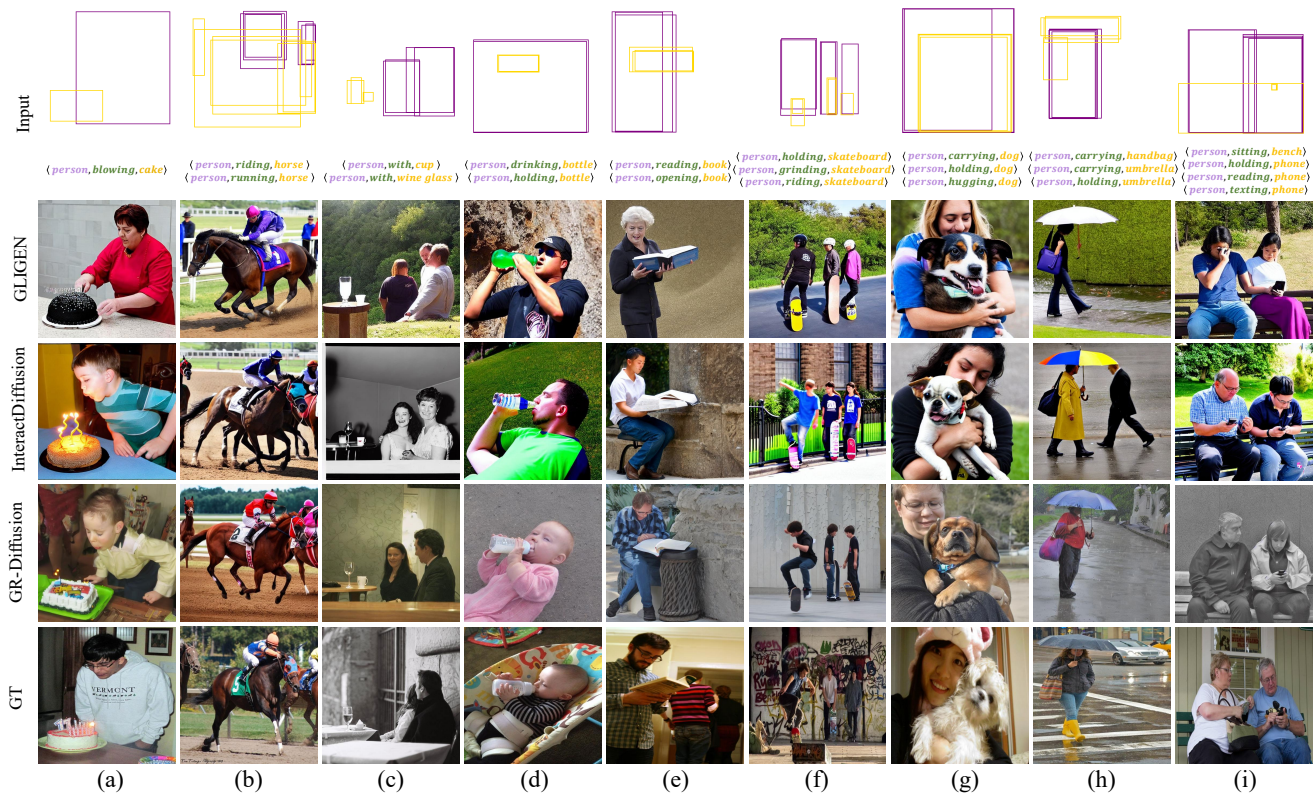


Figure 4. Visual comparison with existing baselines. In all methods, we use the text caption format of *"a person action a object"*. Input rows display the layout input and interaction labels. Purple and yellow boxes represent bounding boxes for each subject and object. GR-Diffusion generates images with superior interaction control and image quality under both simple and complex interaction conditions.

with fewer than 10 samples), under both the Default and Known Object evaluation settings as defined by the official HICO-DET benchmark.

### 4.2. Evaluation Metrics

We evaluate our method from two primary aspects: overall image quality and the controllability of the interactions.

**Image Quality.** We use standard metrics to assess the perceptual realism and diversity of the generated images.

- Fréchet Inception Distance (FID) [10] measures the distributional similarity between generated images and the real data distribution in the Inception feature space.
- Kernel Inception Distance (KID) [2] computes the squared Maximum Mean Discrepancy (MMD) between Inception features, which is less sensitive to sample size,

**Interaction Controllability.** Standard quality metrics like FID assess overall realism but fail to guarantee specific semantic relationships (*e.g.*, *"riding"* in *"a person riding a horse"*) are correctly rendered. Therefore, the HOI Detection Score (mAP) is adopted as the primary controllability metric. This score is obtained by employing the pre-trained FGAHOI [18] detector to identify ⟨subject, action, object⟩ triplets in generated images. The detected triplets are then compared against the ground-truth input directives. This metric specifically evaluates the model's ability to render the intended interactions, not just the co-location of entities. Scores are reported following the FGAHOI protocol under both Default and Known Object settings, using Swin-Tiny and Swin-Large detector backbones for evaluation.

### 4.3. Qualitative Analysis

To visually illustrate the superiority of our method, we provide qualitative comparisons against baselines in Fig. 4. The results demonstrate that GR-Diffusion generates images with significantly higher semantic accuracy and visual coherence, particularly in complex scenes that challenge baseline models.

The improved control of GR-Diffusion is evident in both complex and simple interaction scenarios. In simple interactions, GR-Diffusion enforces stronger semantic fidelity. For example, In Fig 4 (a), (d) and (e), GR-Diffusion render a depicting a natural *'blowing'*,*'drinking'* and *'reading'* pose, which is more accurate and closer to the ground truth. Furthermore, as seen in Fig 4(b), GR-Diffusion successfully generates the correct relationship while preserving quantitative integrity, whereas other methods suf-

Table 1. Quantitative comparison of GR-Diffusion against baselines. Our training-free guidance significantly improves both image quality (FID/KID ↓) and interaction controllability (HOI Score ↑) on both Stable Diffusion v1.x and SDXL backbones. Best results are in **bold**.

| Model | Quality ↓ | | Swin-Tiny HOI Score ↑ | | | | Swin-Large HOI Score↑ | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | FID | KID | Default | | Known Object | | Default | | Known Object | |
| | | | Full | Rare | Full | Rare | Full | Rare | Full | Rare |
| *Backbone: Stable Diffusion v1.x* | | | | | | | | | | |
| StableDiffusion [23] | 35.85 | 0.01297 | 0.63 | 0.68 | 0.66 | 0.70 | 0.64 | 0.83 | 0.65 | 0.84 |
| GLIGEN [15] | 29.35 | 0.01275 | 21.73 | 15.35 | 23.31 | 17.24 | 23.99 | 19.56 | 24.99 | 20.37 |
| InteractDiffusion [11] | 18.69 | 0.00676 | 29.53 | 23.02 | 30.99 | 24.93 | 31.56 | 26.09 | 32.52 | 27.04 |
| B2B [26] | 18.35 | 0.00639 | 31.12 | 23.85 | 31.67 | 24.98 | 32.88 | 27.13 | 33.96 | 28.22 |
| EmIT [30] | 17.07 | 0.00459 | 29.88 | 22.31 | 32.46 | 24.42 | - | - | - | - |
| **GR-Diffusion** | **15.55** | **0.00451** | **32.16** | **25.32** | **33.33** | **26.29** | **34.90** | **28.98** | **35.79** | **29.89** |
| *Backbone: Stable Diffusion XL (SDXL)* | | | | | | | | | | |
| GLIGEN | 28.01 | 0.00820 | 20.73 | 18.55 | 21.73 | 19.82 | 22.97 | 22.09 | 23.42 | 22.55 |
| InteractDiffusion | 17.87 | **0.00491** | 28.98 | 23.70 | 29.92 | 25.10 | 31.37 | 27.17 | 32.21 | 28.09 |
| **GR-Diffusion** | **17.07** | 0.00493 | **29.03** | **25.63** | **30.00** | **27.02** | **31.89** | **29.43** | **32.65** | **30.43** |
| *Ground Truth* | | | | | | | | | | |
| HICO-DET GT | - | - | 29.94 | 22.24 | 32.48 | 24.16 | 37.18 | 30.71 | 38.93 | 31.93 |

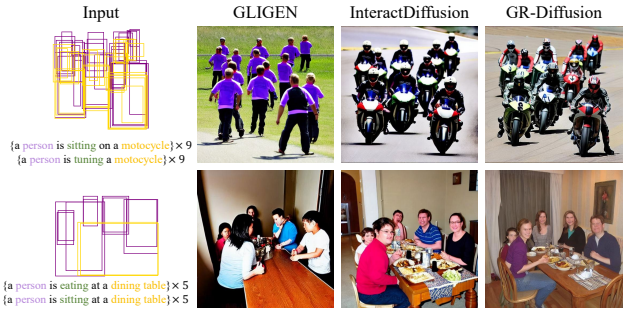


Figure 5. Robustness comparison in extremely conditions. GR-Diffusion is tested against baselines with highly complex prompts.

fer from artifacts such as generating superfluous horse legs. In complex scenes involving multiple concurrent interactions, baselines often struggle to disentangle or render all specified directives. For instance, In Fig 4 (f)-(i), baselines fail to model distinct actions (*e.g.*, missing *'carrying'* when paired with *'holding'*) or generate an incorrect number of subjects for different actions (*e.g.*, *'sitting'* and *'texting'*). GR-Diffusion, successfully resolves these conflicts and accurately models all specified interactions.

The framework's robustness is further tested under extremely complex conditions in Fig. 5. When confronted with a high density of overlapping directives, such as numerous interacting motorcycles (*e.g.*, *"sitting"*, *"turning"*) or a crowded dining scene (*e.g.*, *"sitting"*, *"eating"*), baseline methods exhibit significant failures. For instance, GLIGEN may fail to generate the specified objects entirely or omit key actions, while InteractDiffusion struggles to maintain coherence, producing semantically loose results. In contrast, GR-Diffusion successfully generates these complex scenes with precision and naturalism. These results demonstrate that the graph-guided approach effectively resolves severe guidance conflicts and the challenge of modeling implicit interactions. As a result, it maintains coherence even in the most demanding scenarios.

### 4.4. Quantitative Analysis

As presented in Tab. 1, our GR-Diffusion framework demonstrates a significant and consistent performance improvement over all baseline methods across every metric and model backbone.

**Image Quality.** The proposed guidance mechanism not only enhances controllability but also substantially improves overall image quality. This is demonstrated by the application to the InteractDiffusion (v1.2) model, where GR-Diffusion achieves a new state-of-the-art FID score of **15.55** (down from 18.69) and reduces the KID score from 0.00676 to **0.00451**. This indicates that by resolving internal attention conflicts and enforcing a coherent scene structure, the method effectively mitigates artifacts and generates images that are perceptually more realistic and closer to the real data distribution. A similar trend of improved fidelity is also observed when applied to the SDXL backbone, confirming the method's broad applicability.

**Interaction Controllability.** The core advantage of GR-Diffusion lies in its superior ability to control complex interactions. On the SD 1.x backbone, the framework sets a

Table 2. Plug-and-Play Generalizability of GR-Diffusion (GRD). Applying GRD as a plug-and-play module to baselines (GLIGEN, InteractDiffusion) improves both quality and controllability.

| Model | Quality ↓ | | Default | | Known Object | |
|---|---|---|---|---|---|---|
| | FID | KID | Full | Rare | Full | Rare |
| GLIGEN | 29.35 | 0.01275 | 21.73 | 15.35 | 23.31 | 17.24 |
| GLIGEN+GRD | 28.06 | 0.00914 | 23.65 | 16.89 | 23.65 | 19.02 |
| Interact | 18.69 | 0.00676 | 29.53 | 23.02 | 30.99 | 24.93 |
| Interact+GRD | **15.55** | **0.00451** | **32.16** | **25.32** | **33.33** | **26.29** |

Table 3. Efficiency and Computational Cost Comparison. All evaluations are performed at 512x512 resolution.

| Model | Time (s) | Trainable Parameters (M) |
|---|---|---|
| GLIGEN | 2.67 | 209M |
| InteractDiffusion | 4.01 | 210M |
| B2B | 7.50 | 0 |
| GR-Diffusion | 5.78 | 0 |

new state-of-the-art, boosting the HOI detection score from 32.52 to **35.79** in the (Swin-Large, Known Object, Full) setting. This substantial gain underscores the effectiveness of the graph-based alignment, ensuring that subjects and objects are not merely co-located but are rendered in a semantically correct, interactive relationship. The improvements are notable in the challenging Rare subset, highlighting the model's robustness. The efficacy of GR-Diffusion on more powerful foundational models is further confirmed by the HOI detection score on the SDXL backbone as well.

**Plug-and-Play Generalizability.** To validate its model-agnostic nature, GR-Diffusion (GRD) was tested as a plug-and-play module on GLIGEN, with results evaluated by the FGAHOI Swin-Tiny detector. As shown in Tab. 2, GRD provides consistent improvements in both fidelity and controllability. The improvement is particularly notable in the challenging Known Object (Rare) subset, where the HOI Detection Score was raised from 17.24 to 19.02, which confirms GRD functions as a generalizable guidance mechanism for diverse diffusion models.

**Efficiency and Cost Analysis.** The framework's computational costs are detailed in Tab. 3. GR-Diffusion introduces no trainable parameters that bypasses the training-based methods like GLIGEN and InteractDiffusion. During inference, GR-Diffusion demonstrates lower latency than B2B. Therefore GR-Diffusion achieves a balance, offering plug-and-play control while maintaining lower computational overhead at both training and inference stages compared to respective baselines.

Table 4. Ablation study on the components of GR-Diffusion. All components contribute to the final performance, with their combination yielding the best balance of fidelity and controllability.

| Model | NAG | EEG | GSD | Quality ↓ | | HOI Score | |
|---|---|---|---|---|---|---|---|
| | | | | FID | KID | Def. | KO . |
| Interact | | | | 18.69 | 0.00676 | 29.53 | 30.99 |
| Ours | ✓ | | | 17.63 | 0.00588 | 30.75 | 32.22 |
| | ✓ | ✓ | | **14.99** | **0.00384** | 30.31 | 31.66 |
| | ✓ | ✓ | ✓ | 15.55 | 0.00451 | **32.16** | **33.33** |

### 4.5. Ablation Study

The proposed GR-Diffusion framework incorporates three key modules: the Node Alignment Guidance (NAG), the Edge Enhancement Guidance (EEG), and the Graph Similarity Detector (GSD). The ablation study in Tab. 4 analyzes each component's contribution, starting from the InteractDiffusion baseline. Adding only NAG improves both metrics, confirming the benefit of spatial layout correction. Further adding EEG achieves the best image quality, but this combination reveals a fidelity-control trade-off, as the strong fidelity boost unexpectedly compromises controllability, causing the HOI Detection Score to fall. This trade-off motivates the adaptive guidance mechanism. The full model, which uses the GSD to adaptively apply both guidance modules, successfully resolves this conflict. This final configuration recovers and improved detection score from 31.66 to 33.33 while maintaining excellent fidelity.

## 5. Conclusion

This paper introduces GR-Diffusion, a training-free framework for complex human-object interaction (HOI) control in pre-trained diffusion models. The core contribution of GR-Diffusion is reframing the complex control challenge as a graph structural alignment task, that allows the framework to decouple and guide both the spatial layout of entities as graph nodes and the semantic interactions as graph edges during the denoising process. GR-Diffusion is realized as a plug-and-play framework that steers the internal cross-attention and self-attention layers. Both quantitative and qualitative evaluations confirm that GR-Diffusion establishes a new state-of-the-art, significantly enhancing the interaction controllability and image fidelity over existing methods.

## Acknowledgments

This work is supported in part by Science and Technology Innovation Plan of Shanghai Science and Technology Commission under Grant No.25511106200.

## References


[1] 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. 1, 2

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

[3] SeungJu Cha, Kwanyoung Lee, Ye-Chan Kim, Hyunwoo Oh, and Dong-Jin Kim. VerbDiff: Text-only diffusion models with enhanced interaction awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8041–8050, 2025. 3

[4] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 381–389. IEEE, 2018. 2, 5

[5] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5343–5353, 2024. 1, 2

[6] Yushi Chen, Haoran Li, and Jian Yang. Rare-to-Frequent: Unlocking compositional generation power of diffusion models on rare concepts with LLM guidance. In *Proceedings of the International Conference on Learning Representations*, 2025. 2

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 1

[8] 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 *Proceedings of the ACM International Conference on Multimedia*, pages 165–173, 2020. 2

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

[10] 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

[11] Jiun Tian Hoe, Xudong Jiang, Chee Seng Chan, Yap-Peng Tan, and Weipeng Hu. InteractDiffusion: Interaction control in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6180–6189, 2024. 1, 2, 3, 5, 7

[12] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. In *Proceedings of the International Conference on Machine Learning*, pages 1–21, 2023. 2

[13] Jian-Yu Jiang-Lin, Kang-Yang Huang, Ling Lo, Yi-Ning Huang, Terence Lin, Jhih-Ciang Wu, Hong-Han Shuai, and Wen-Huang Cheng. RECORD: Reasoning and correcting diffusion for HOI generation. In *Proceedings of the ACM International Conference on Multimedia*, pages 9465–9474, 2024. 3

[14] Long Li, Weimian Zhang, Hong Yin, Jialu Wang, Jiatong Cheng, Jue Wu, Xin Wang, Wen-Jun Wang, Yueting Zhuang, and Han Zhang. LLM-Grounded Diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. In *Proceedings of the International Conference on Learning Representations*, 2024. 2

[15] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. GLIGEN: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 1, 2, 7

[16] Tianyi Liang, Jiangqi Liu, Yifei Huang, Shiqi Jiang, Jianshen Shi, Changbo Wang, and Chenhui Li. TextCenGen: Attention-guided text-centric background adaptation for text-to-image generation. In *Proceedings of the International Conference on Machine Learning*, 2025. 4

[17] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In *Proceedings of the International Conference on Learning Representations*, 2022. 5

[18] Shuailei Ma, Yuefeng Wang, Shanze Wang, and Ying Wei. FGAHOI: Fine-grained anchors for human-object interaction detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2415–2429, 2023. 6

[19] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7465–7475, 2024. 2

[20] Chong Mou, Xintao Wang, Liangbin Zuo, Jian Zhang, Yujiu Qiao, Ying Shan, and Wenping He. T2I-Adapter: Learning adapters to inject spatial control in text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2

[21] 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. 1

[22] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022.

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2, 3, 7

[24] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 1

[25] 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. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1

[26] Ashkan Taghipour, Morteza Ghahremani, Mohammed Bennamoun, Aref Miri Rekavandi, Hamid Laga, and Farid Boussaid. Box it to bind it: Unified layout control and attribute binding in text-to-image diffusion models. *IEEE Transactions on Multimedia*, pages 1–15, 2025. 2, 3, 4, 7

[27] Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. InstanceDiffusion: Instance-level control for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6232–6242, 2024. 2

[28] Mingyue Yang, Dianxi Shi, Jialu Zhou, Xinyu Wei, Leqian Li, Shaowu Yang, and Chunping Qiu. CEIDM: A controlled entity and interaction diffusion model for enhanced text-to-image generation. *arXiv preprint arXiv:2508.17760*, 2025. 3

[29] 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. 1

[30] Haofan Zhang and Shangfei Wang. EmIT: Emotional interaction control in text-to-image diffusion models. In *Proceedings of the ACM International Conference on Multimedia*, pages 9950–9958. Association for Computing Machinery, 2025. 3, 7

[31] Hui Zhang, Dexiang Hong, Yitong Wang, Jie Shao, Xinglong Wu, Zuxuan Wu, and Yu-Gang Jiang. CreatiLayout: Siamese multimodal diffusion transformer for creative layout-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18487–18497, 2025. 2

[32] Kai Zhang, Zeqiang Zhao, Fang Liu, and Zhe Wang. HiCo: Hierarchical controllable diffusion model for layout-to-image generation. In *Advances in Neural Information Processing Systems*, 2024. 2

[33] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2

[34] 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 *Proceedings of the European Conference on Computer Vision*, pages 754–769. Springer, 2022. 1

[35] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. LayoutDiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22490–22499, 2023. 1, 2