VerbDiff: Text-Only Diffusion Models with Enhanced Interaction Awareness

SeungJu Cha Kwanyoung Lee Ye-Chan Kim Hyunwoo Oh Dong-Jin Kim
Hanyang University, South Korea
{sju9020, mobled37, dpcksdl78, komji, djdkim}@hanyang.ac.kr

Abstract

Recent large-scale text-to-image diffusion models generate photorealistic images but often struggle to accurately depict interactions between humans and objects due to their limited ability to differentiate various interaction words. In this work, we address the challenge of capturing nuanced interactions within text-to-image diffusion models. We propose a novel text-to-image generation model that weakens the bias between interaction words and objects, enhancing the understanding of interactions. Specifically, we disentangle various interaction words from frequency-based anchor words and leverage localized interaction regions from generated images to help the model better capture semantics in distinctive words without extra conditions. Our approach enables the model to accurately understand the intended interaction between humans and objects, producing high-quality images with accurate interactions aligned with specified verbs. Extensive experiments on the HICO-DET dataset validate the effectiveness of our method compared to previous approaches.

1. Introduction

Recently, text-to-image diffusion models have demonstrated the ability to generate photorealistic images from natural text [16], yet they often fail to accurately capture interactions between humans and objects [6]. For example, they often fail to differentiate between semantically distinct prompts, such as "A person *walking* a bicycle" and "A person *riding* a bicycle," particularly in capturing the semantics in verbs that are crucial for accurately depicting the intended interaction. To enhance prompt interpretation ability of diffusion models, previous methods have leveraged additional modules such as large language models (LLMs) [4, 5, 12] or extra bounding boxes [6, 11, 21, 24] to provide explicit relational visual information.

For example, InteractDiffusion [6] introduces controllable image generation using Human-Object Interaction (HOI) information with bounding boxes for human, relation, and object <h, r, o> triplets. Nevertheless, as shown in Fig. 1, the model still struggles with accurate interactions even with these additional conditions. For instance, given the prompt "walking, bicycle," the model fails to depict the intended interaction when semantically distinctive interaction words share a similar bounding box (*e.g.*, *walking* and *riding*). This suggests that the model lacks an understanding of interaction semantics and relies heavily on precise bounding boxes, which are labor-intensive to provide.

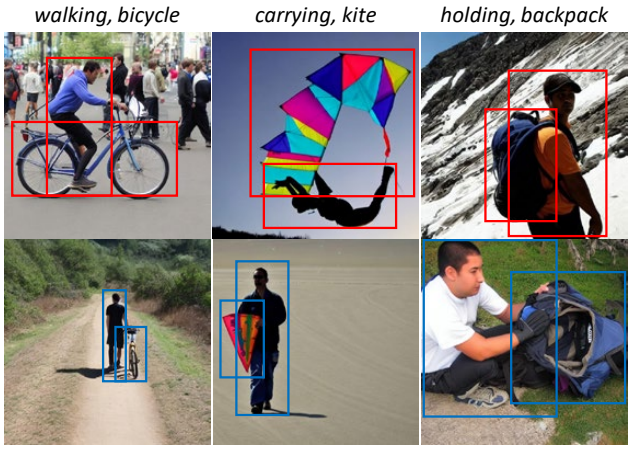

Figure 1. **Examples of interactions from InteractDiffusion [6], showing a lack of understanding of interaction words.** The model relies on precise bounding boxes rather than understanding interaction words to exhibit accurate interactions, as shown in the results for the "red" and "blue" boxes.

To improve the model's interaction understandability without additional conditions, we propose VerbDiff, a novel text-to-image generation method that better distinguishes interaction words via several objectives. First, we apply Relation Disentanglement Guidance (RDG) using frequency-based anchor texts from the dataset for each human-object pair to reduce interaction bias in generated images. We observe that the exhibited interactions in generated images have bias, which follows frequent verbs corresponding to each human-object pair in the data distribution. To address this, we align interaction features between real and generated images while separating those of generated images

from anchor text features, enabling a more effective distinction of interaction words.

Secondly, to enforce the model to better focus on the interaction regions between humans and objects, we introduce Interaction Direction Guidance (IDG). This approach emphasizes fine-grained semantic distinctions in localized interactions. We design an Interaction Region (IR) module to capture specific interaction areas in generated images, leveraging cross-attention maps for region extraction in text-only diffusion models without bounding box conditions. The IR module extracts interaction regions using the centroids of cross-attention maps associated with <h, r, o> tokens. We obtain biased interaction features from these regions and apply interaction direction guidance. We validate our method's effectiveness on the HICO-DET dataset [1]. Our approach accurately captures semantic differences between interaction words and generates high-quality images with accurate interactions compared to previous methods.

2. Relation Generation

Relation generation between objects has been explored leveraging LLMs [4, 5, 12] and scene graph [14, 17, 19, 20, 22, 23]. While LLM-based methods can generate accurate object layouts, they primarily focus on compositional relations rather than interactions [4, 5, 12]. Scene graph-based approaches have addressed these limitations by combining scene graph encoders to enhance prompt understandability [14, 17, 19, 20, 22, 23]. However, they require additional training to extract scene graphs from prompts and struggle to capture subtle semantic differences in interactions.

To depict accurate interactions during generation, researchers have focused on human-object interactions (HOI) [6, 8–10]. Some methods use inversion-based frameworks to capture interaction semantics but require optimization for each interaction word [8, 9]. More generalized approaches leverage additional conditions such as human poses [10] or bounding boxes [6]. Specifically, InteractDiffusion [6] leverages the HOI information with bounding boxes to generate interactions. However, this method still relies on additional information rather than semantic differences between prompts. Our work is similar to InteractDiffusion [6] by leveraging HOI information but differs from focusing on semantic differences between interaction words to enhance interaction understanding in SD without additional conditions.

3. Method

3.1. Relation Disentanglement Guidance

To improve interaction understanding in text-to-image diffusion models, we propose a method that reduces the bias toward frequent interaction verbs. Given a ground-truth text T^{gt} = "*A photo of a* $\{h\}$ $\{r^{gt}\}$ *a/an* $\{o\}$", SD generates images I^{gen} during training, which often reflect biased verbs frequently observed in the dataset. For example, when the text describes "*holding* a backpack," I^{gen} often depicts "*wearing* a backpack," a more frequent verb for the human-object pair. We define such a frequent verb as the anchor verb $r^{anc} = \arg\max_{r \in R_o} \mathcal{C}(r|o)$, where R_o denotes the set of verb in each human-object pair and $\mathcal{C}(\cdot)$ is the verb count. Then, we construct an anchor sentence T^{anc} = "*A photo of a* $\{h\}$ $\{r^{anc}\}$ *a/an* $\{o\}$," and encourage the model to distinguish between T^{gt} and T^{anc} using a triplet loss [7] with margin m. We compute the loss using CLIP [15] features f^{gen}, e^{gt}, and e^{anc} from I^{gen}, T^{gt}, and T^{anc}:

$$\mathcal{L}_{\text{triple}} = \max\left(0, m + \text{sim}\left(f^{gen}, e^{gt}\right) - \text{sim}\left(f^{gen}, e^{anc}\right)\right). \tag{1}$$

To further align the generated image with the ground-truth interaction at the image level, we apply an image alignment loss. Since I^{gt} often includes multiple interactions, we extract the mask $\mathcal{M}$ from the human-object bounding boxes and isolate the relevant region $I^{gt}_{\mathcal{M}} = I^{gt} \odot \mathcal{M}$. The image encoder then extracts $f^{gt}_{\mathcal{M}}$, and the alignment loss becomes:

$$\mathcal{L}_{\text{align}} = 1 - \frac{f^{gt}_{\mathcal{M}} \cdot f^{gen}}{|f^{gt}_{\mathcal{M}}||f^{gen}|}. \tag{2}$$

This guides the model to focus on interaction-specific regions. However, as each interaction verb has widely varying sample counts, we apply effective number α to balance the modification extent across each words. Following [2], each class k is weighted by $\alpha(k) = \frac{1-\beta^{n_k}}{1-\beta}$, where $\beta = (N-1)/N$, n_k is the number of samples, and N is the total number of samples in the dataset. Finally, we multiply α and define Relation Disentanglement Guidance as:

$$\mathcal{L}_{\text{RDG}} = \alpha \cdot (\mathcal{L}_{\text{triple}} + \mathcal{L}_{\text{align}}). \tag{3}$$

3.2. IR Module & Interaction Direction Guidance

While Relation Disentanglement Guidance (RDG) helps the model distinguish interaction words, it often fails to reflect human expectations in fine-grained interaction regions. To address this, we hypothesize that focusing on more localized regions can better capture detailed interactions. We apply Interaction Direction Guidance (IDG), which leverages cross-attention maps of <h, r, o> tokens to guide the model toward more precise region-level interaction depiction.

Interaction Region Module. To achieve a more detailed interaction expression, we first extract interaction regions from both real and generated images. Although real images contain explicit human and object bounding boxes in the dataset, it is challenging to extract specific interaction regions without additional conditions from generated images. We leverage the cross-attention maps $\mathcal{A}_h$, $\mathcal{A}_r$ and $\mathcal{A}_o$ corresponding to <h,r,o> token to extract interaction region between human and objects effectively.

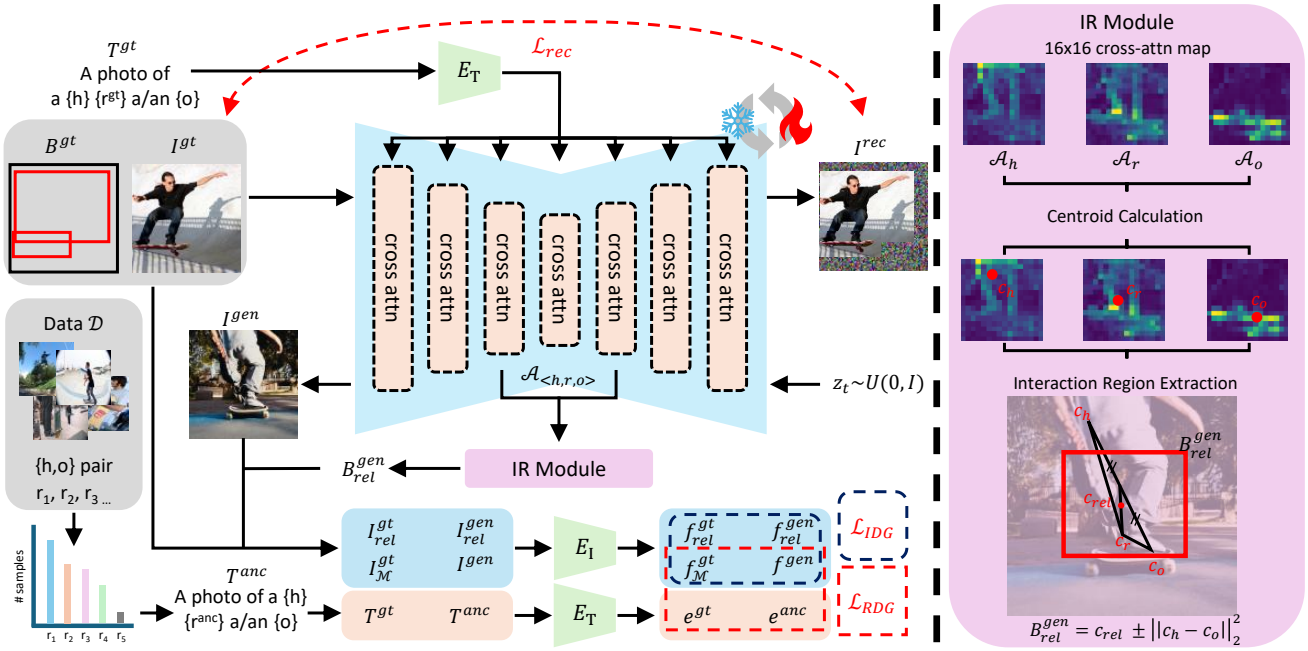

Figure 2. **Pipeline of VerbDiff.** VerbDiff uses Relation Disentanglment Guidance (Sec. 3.1) that separates the interaction features from the anchor text for each human-object pair. Additionally, it contains the IR module (Sec. 3.2) (right) which extracts localized interaction regions from generated images without explicit bounding boxes and Interaction Direction Guidance (Sec. 3.2) that guides the model to focus more on the fine-grained interaction regions.

We apply the centroid extraction mechanism proposed in [3] to compute specific points within attention maps. Specifically, we calculate c_h, c_r and c_o from each cross-attention maps as follows:

$$c = \frac{1}{\sum_{h,w} \mathcal{A}} \begin{bmatrix} \sum_{h,w} w \cdot \mathcal{A} \\ \sum_{h,w} h \cdot \mathcal{A} \end{bmatrix}, \tag{4}$$

where h and w denote the height and width in the attention map $\mathcal{A}$. Then, we define the interaction center c_{rel} as the centroid of a triangle defined by c_h, c_r, and c_o. Finally, we extract interaction region B^{gen}_{rel} by leveraging the distance between human and object center with $B^{gen}_{rel} = c_{rel} \pm \|c_h - c_o\|^2_2$. Additionally, we extract B^{gt}_{rel} in the same manner, leveraging the human and object bounding boxes in the dataset. We treat c_{rel} as the midpoint between the human and object centers within the given boxes B^{gt}.

Interaction Direction Guidance. To generate images with more realistic interactions, we design guidance that aligns feature differences at the image level with those at the interaction region. With interaction region B^{gt}_{rel} and B^{gen}_{rel}, we obtain interaction region image by masking the real and generated images: $I^g_{\text{rel}} = B^g_{\text{rel}} \odot I^g, \quad \text{where } g \in \{\text{gt}, \text{gen}\}$. We then encode both interaction region images through a CLIP image encoder and extract interaction region features f^{gt}_{rel} and f^{gen}_{rel}. Then, we calculate the direction between f^{gt}_{rel} and f^{gen}_{rel}, referred to as biased relation feature $f^{bias}_{rel} = f^{gt}_{rel} - f^{gen}_{rel}$, to align the direction of biased interaction features with that of the real images. The interaction direction guidance is as follows:

$$\mathcal{L}_{\text{IDG}} = 1 - \frac{(f^{gt}_{\mathcal{M}} - f^{gen}) \cdot (f^{bias}_{rel})}{|f^{gt}_{\mathcal{M}} - f^{gen}||f^{bias}_{rel}|}. \tag{5}$$

3.3. Training Phase

We train only the cross-attention layer in SD to capture the semantically distinct relation words, as the cross-attention reflects the existence of each word token [13]. We adopt the reconstruction loss to train T2I models in addition to relation disentanglement guidance and interaction direction guidance. However, when an image contains multiple humans and objects, it can lead to multiple interactions that do not match a single target interaction verb. To accurately separate the human and object corresponding to interaction words, we apply mask $\mathcal{M}$ when calculating the reconstruction loss as follows:

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{z,\epsilon\sim\mathcal{N}(0,1),t}\left[\|(\epsilon \odot \mathcal{M} - \epsilon_\theta(z_t, t, T) \odot \mathcal{M})\|^2_2\right]. \tag{6}$$

Finally, we combine reconstruction loss with disentanglement guidance and interaction direction guidance leveraging λ_1, λ_2, and λ_3 as below:

$$\mathcal{L}_{\text{total}} = \lambda_1 \cdot \mathcal{L}_{\text{rec}} + \lambda_2 \cdot \mathcal{L}_{\text{RDG}} + \lambda_3 \cdot \mathcal{L}_{\text{IDG}}. \tag{7}$$

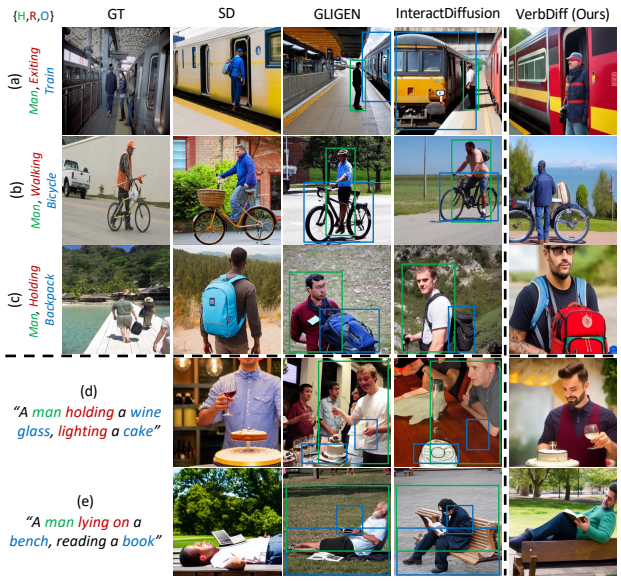

Figure 3. **Interaction comparison of generated images across models.** Images are generated using a fixed template: "A {H} {R} a/an {O}". The top three rows show single interactions, while the bottom rows illustrate multiple interactions. Green and blue boxes indicate grounding boxes used for humans and objects, respectively, during image generation.

4. Experiments

We train our model on SD v1.4 at a resolution of 512×512. Using the Adam optimizer with a learning rate of 4×10^{-6}, we trained for a single epoch over 17 hours. For image generation during training, we utilize the DDIM scheduler [18] with 30 sampling steps, and during inference, we increase sampling steps to 50. Finally, λ_1, λ_2, and λ_3 are set to 1.0, 10, and 0.8, respectively, and the margin m is set to 0.2.

Dataset. To train our model, we realign the training set according to the 600 HOI text descriptions and modify annotations to match each image with its corresponding text descriptions. This realignment yields 69,404 images across 600 prompts. We exclude text classes containing "and" (*e.g.* "A photo of a person and an airplane") to avoid ambiguity from various interactions within the same class, resulting in a final training set of 61,114 images across 501 prompts.

4.1. Qualitative Results

Fig. 3 shows a comparison of generated images focusing on the interactions between humans and objects across our method and other models. Compared to previous approaches, our results exhibit more detailed interactions, closely resembling ground-truth images. For example, in Fig. 3 (a)-(c), GLIGEN [11] mainly focuses on generating humans and objects, failing to depict the accurate interactions related to the interaction words. While InteractDiffusion [6] improves interaction representation, its results show inaccurate or ambiguous interactions that lack fine-grained interaction details (a)-(c).

We further evaluate our model using complex prompts containing multiple interactions to assess its ability to distinguish between interaction words. As shown in Fig. 3 (d) and (e), our model successfully captures each specified interaction, whereas other models often struggle. While GLIGEN and InteractDiffusion depict some interactions, they frequently miss key elements such as a wine glass or a bench. In contrast, our model consistently generates accurate images that faithfully reflect the intended interactions.

34emModels	SOV-STG-S (Acc ↑)				SOV-STG-Swin-L (Acc ↑)			
	Def.		KO.		Def.		KO.	
	Full	Rare	Full	Rare	Full	Rare	Full	Rare
HICO-DET	26.52	6.78	28.68	7.29	29.98	12.66	31.16	13.43
SD [16]	16.09	4.59	18.22	4.85	20.08	8.07	21.69	8.66
GLIGEN [11]	15.88	4.85	17.91	5.24	17.83	7.00	19.35	7.57
InteractDiffusion [6]	19.67	7.00	21.31	7.69	23.53	10.27	24.86	11.18
VerbDiff (Ours)	**22.59**	**7.62**	**24.79**	**7.83**	**27.05**	**12.60**	**28.43**	**13.18**

Table 1. **HOI accuracy comparison between VerbDiff and previous methods.** Def. and KO. refer to Default and Known Object.

4.2. Quantitative Results

Tab. 1 presents the HOI accuracy scores, where VerbDiff achieves the highest accuracy across all settings compared to previous methods. In particular, although different backbones are used for computing accuracy, VerbDiff consistently shows the highest accuracy across all settings. This demonstrates that our method has robust interaction word understanding and produces images with precise interactions. Overall, VerbDiff consistently outperforms other models across nearly all evaluation metrics, demonstrating its strong ability to comprehend interaction words and generate high-quality images with accurate human-object interactions, even without extra conditions.

5. Conclusion

We propose VerbDiff, a novel text-to-image (T2I) diffusion model that addresses the interaction misunderstanding problem in SD in a simple yet effective manner, without additional conditions. Although previous methods leverage additional conditions to help the model understand the interaction between humans and objects, they still rely on precise instructions and lack an understanding of the semantic differences between interaction verbs. Our model successfully captures the semantic meanings inherent in interaction words and generates high-quality images with accurate interactions. Extensive experiments demonstrate the effectiveness of our method in enhancing the ability to understand interaction words of T2I models, achieving better interaction comprehension compared to previous state-of-the-art methods.

References

[1] 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. 2

[2] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019. 2

[3] Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36:16222–16239, 2023. 3

[4] Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2

[5] Hanan Gani, Shariq Farooq Bhat, Muzammal Naseer, Salman Khan, and Peter Wonka. Llm blueprint: Enabling text-to-image generation with complex and detailed prompts. *arXiv preprint arXiv:2310.10640*, 2023. 1, 2

[6] 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, 4

[7] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *Similarity-based pattern recognition: third international workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3*, pages 84–92. Springer, 2015. 2

[8] 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

[9] Xu Jia, Takashi Isobe, Xiaomin Li, Qinghe Wang, Jing Mu, Dong Zhou, Huchuan Lu, Lu Tian, Ashish Sirasao, Emad Barsoum, et al. Customizing text-to-image generation with inverted interaction. In *ACM Multimedia 2024*, 2024. 2

[10] 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. *arXiv preprint arXiv:2407.17911*, 2024. 2

[11] 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, 4

[12] 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. 1, 2

[13] Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7817–7826, 2024. 3

[14] Jinxiu Liu and Qi Liu. R3cd: Scene graph to image generation with relation-aware compositional contrastive control diffusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3657–3665, 2024. 2

[15] 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. 2

[16] 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, 4

[17] Guibao Shen, Luozhou Wang, Jiantao Lin, Wenhang Ge, Chaozhe Zhang, Xin Tao, Yuan Zhang, Pengfei Wan, Zhongyuan Wang, Guangyong Chen, et al. Sg-adapter: Enhancing text-to-image generation with scene graph guidance. *arXiv preprint arXiv:2405.15321*, 2024. 2

[18] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 4

[19] Shengqiong Wu, Hao Fei, Hanwang Zhang, and Tat-Seng Chua. Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[20] Yinwei Wu, Xingyi Yang, and Xinchao Wang. Relation rectification in diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7685–7694, 2024. 2

[21] 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. 1

[22] Bicheng Xu, Qi Yan, Renjie Liao, Lele Wang, and Leonid Sigal. Joint generative modeling of scene graphs and images via diffusion models. *arXiv preprint arXiv:2401.01130*, 2024. 2

[23] Ling Yang, Zhilin Huang, Yang Song, Shenda Hong, Guohao Li, Wentao Zhang, Bin Cui, Bernard Ghanem, and Ming-Hsuan Yang. Diffusion-based scene graph to image generation with masked contrastive pre-training. *arXiv preprint arXiv:2211.11138*, 2022. 2

[24] 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