

DreamRelation: Bridging Customization and Relation Generation

Supplementary

Supplementary Material

Overview. The supplementary includes these sections:

- **Sec. 1.** Introduction video.
- **Sec. 2.** More experimental results.
- **Sec. 3.** Implementation details of the experiments.
- **Sec. 4.** Failure cases and limitations of our model.
- **Sec. 5.** Details of the RelationBench.
- **Sec. 6.** Incorporate with CogVideoX-I2V.
- **Sec. 7.** Social impacts.

1. Introduction Video

We provide a video introduction to our work. Please refer to “**introduction_video.mp4**” in the supplementary file.

2. Additional Experimental Results

Stable Diffusion 1.5 as base model. We implement DreamRelation on the SSR-Encoder (SD 1.5-based). As shown in Fig. 3 and Tab. 1, DreamRelation outperforms other SD 1.5-based models in our task.

Keypoint Matching Loss. We use X-Pose [8] as our keypoint detector due to its open-vocabulary detection capabilities that are compatible with a wide range of objects, instead of humans only. The keypoint matching loss (KML) facilitates relation generation by explicitly guiding the model’s pose manipulation, resulting in more accurate pose generation. As shown in Fig. 2, the cat’s arm crosses the dog’s body, accurately depicting the “hug” relation. The visualization in Fig. 4 shows the cosine similarity between the latent representation of the image prompt and model prediction. Implementation with KML shows a better alignment of specific parts during training and inference.

Local Tokens Injection. To understand why local features enhance relation-aware generation, we employ Principal Component Analysis (PCA) to compactly project the dense features. As shown in Fig. 5, these dense features provide more fine-grained information than CLIP image tokens, offering detailed insights about each object to construct interactions between them. This detailed representation aids in distinguishing different objects during the generation process and helps avoid object confusion, particularly in cases of heavy overlap, while also facilitating object appearance alignment.

Regarding the injection method, we observe that simply concatenating local tokens with image-level tokens yields the best performance, as demonstrated in Fig. 3. Additional ablation studies on the architecture of the Local Image Encoder, presented in Tab. 2, reveal that CLIP ViT bigG pro-

Method	CLIP-T	CLIP-R	CLIP-I	DINO
Custom Diffusion	20.1	<u>15.4</u>	64.7	55.3
SSR-Encoder	24.2	14.6	<u>72.1</u>	56.2
DreamRelation(SD 1.5)	26.1	19.1	72.4	58.9

Table 1. Quantitative results using SD 1.5 as the base model.

Model	Multi-object			
	CLIP-T	CLIP-R	CLIP-I	DINO
EVA-CLIP-L14	23.6	15.7	56.4	54.8
CLIP-ViT-L14	22.9	14.8	<u>58.3</u>	52.7
CLIP-ViT-bigG	28.9	20.4	75.4	62.1

Table 2. Ablation study on Local Image Encoder’s architecture.

Injection method	Multi-object			
	CLIP-T	CLIP-R	CLIP-I	DINO
Add	25.4	<u>18.5</u>	<u>71.0</u>	<u>56.9</u>
Linear Projection	<u>25.8</u>	18.3	68.2	54.4
Concatenate	28.9	20.4	75.4	62.1

Table 3. Ablation study on local token injection methods.

duces the best results across all metrics. We attribute this success to the compatibility between the local image encoder and the CLIP image encoder. Moreover, our investigation into the injection method for local tokens in the generation process confirms that concatenation, despite its simplicity, is highly effective and consistently achieves superior results across all metrics.

Relation-Aware Customization Data Engine. Our relation-aware data are generated using DALL-E-3 [1], which effectively preserves the identity of certain categories through its multi-turn dialogue capability. Specifically, we maintain object identity by appending “the photo of the same” to the text prompt. Although our relation-aware data primarily consist of animal categories, they effectively capture relational information in images and generalize well to a wide range of objects, as demonstrated in Fig. 1.

Relation-Aware Customized Image Generation. We present additional qualitative results in Fig. 9, Fig. 14, and Fig. 1, comparing our method with both training-based and tuning-based approaches. Our method demonstrates a clear advantage in pose manipulation to achieve the desired relations. Further results in Fig. 8 highlight the differences between DreamRelation and our base model, MS-Diffusion.

Relation Inversion Task. As shown in Fig. 10, our DreamRelation substantially enhances SDXL’s ability to generate



Figure 1. Additional results of relation-aware generation across a wide range of objects.

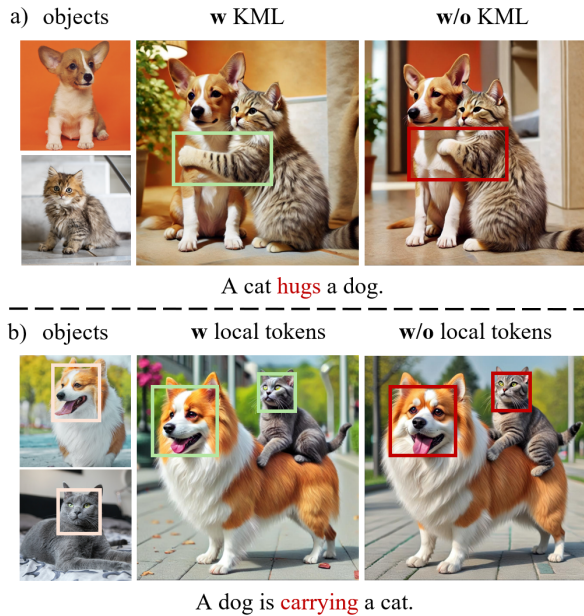


Figure 2. Additional ablation studies on KML and Local Tokens.



Figure 3. Results of SD 1.5 as base model



Figure 4. KML enhances the alignment of specific parts of the object in the image prompt and model prediction.

images that strictly adhere to specific relations. Compared to ReVersion [4], our method produces more accurate relations without any object confusion or omissions, highlighting DreamRelation’s robust performance in the Relation Generation task.

3. Additional Implementation Details

Local Image Encoder’s Implementation Details. To enhance region-language alignment of dense features, we employ self-distillation on CLIP-ViT-bigG. The training is conducted on the train2017 split of the COCO dataset for 6 epochs, using 8 A100 GPUs with a batch size of 2 per GPU. We apply the Adam optimizer with a learning rate of $1e-5$ and a weight decay of 0.1. The Local Image Encoder, containing 1.8B parameters, extracts local tokens that are

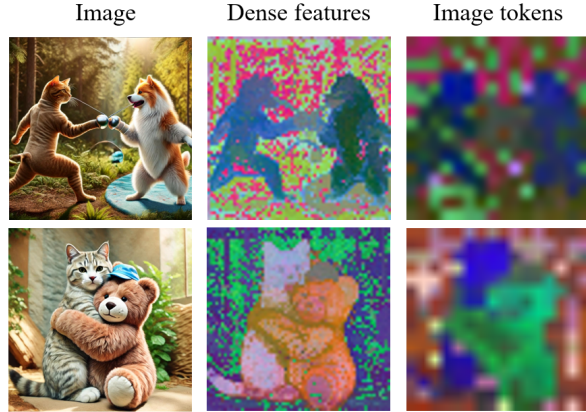


Figure 5. Visualization of Dense feature and Image tokens by Principal Component Analysis (PCA).



Figure 6. Multi-object relation-aware image customization results of pet, toy, plushie, and person.

concatenated with image tokens during fine-tuning and inference to mitigate the confusion problem between objects.

Baselines’ Implementation Details. We incorporate ReVersion [4] with MS-Diffusion [7] by its official implementation. We fine-tune a learnable text embedding on a set of relation-specific images. We inject the text embedding into the text prompt embedding during inference. We implement ReVersion [4] on DreamBooth [6] by a similar procedure. For tuning-based methods, we implement Textual Inversion [2], DreamBooth [6], and Custom Diffusion [5] using their respective diffuser versions, with learning rates and tuning steps aligned to those reported in the original papers. We implement Mix-of-Show [3] from their official repository. We utilize the official implementations and checkpoints for training-based methods, adjusting hyperparameters as needed during evaluation. Specifically, we set the

scale to 0.6 in MS-Diffusion [7] and sample 30 steps using the EulerDiscreteScheduler. For the SSR Encoder [9], we employ the UniPCMultistepScheduler, sampling 30 steps and adjusting the scale for each object to accommodate different cases. For λ -Eclipse, we apply the default settings of the official implementation without modification.

4. Failure cases

As illustrated in Fig. 11, we present three typical failure cases from our experiments. First, when given unreasonable relation generation requests—such as asking a plushie octopus, which inherently lacks limbs, to “shake hands”—our model compensates by generating additional arms, resulting in a mismatched appearance. Second, some generated relations appear unnatural, exemplified by a duck that fails to make contact with a cat as intended. Lastly, object confusion at the interaction point remains a common challenge across all multi-object generation models.

5. RelationBench

In this section, we show the objects and text prompts contained in our RelationBench in Fig. 12, Fig. 13, and Tab. 4. The objects are selected from well-known benchmarks, DreamBench and CustomConcept101, covering the commonly seen categories in the real world. The relations in text prompts have covered the most common relations in the real world.

6. Incorporate with CogVideoX-5b-I2V

Additionally, we use our generated relation-aware customized images as the first frame to generate videos by the CogVideoX-I2V model, the generated results are shown in Fig. 15. CogVideoX-I2V can handle simple relations such as “shaking hands” but struggles with complex interactions such as “hugging”.

7. Social Impact

Positive societal impacts. Relation-aware image customization enables users to generate images that not only contain customized objects but also capture their meaningful relationships. This opens up new opportunities for creative professionals, such as designers, advertisers, and educators, to communicate complex ideas visually with greater precision and flexibility. It has the potential to streamline content creation in diverse fields, from personalized marketing to educational tools, making high-quality, contextually rich imagery accessible without the need for extensive resources.

Potential negative societal impacts. The ability to generate customized images that involve specific relationships between objects could be misused to fabricate misleading or

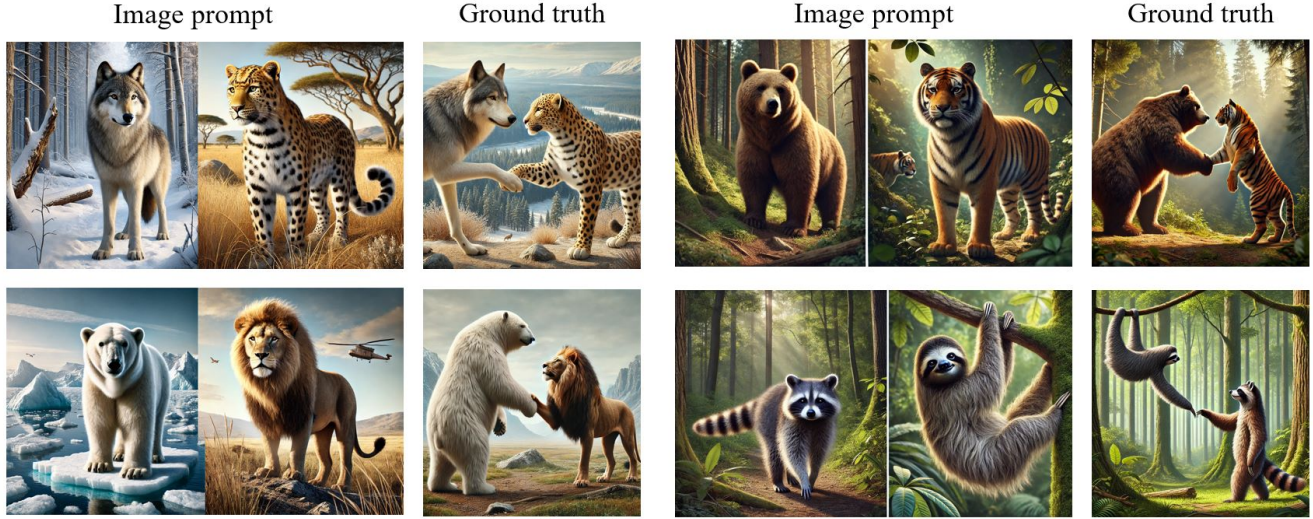


Figure 7. Our fine-tuning dataset as an example.



Figure 8. Single-object comparison with our base model MS-Diffusion: The results demonstrate that our method generates more accurate and natural relation-aware images.

harmful visual narratives, including false representations of events or manipulative visual content in political or social contexts. Additionally, if the models are trained on biased data, they may reinforce existing societal biases, marginalizing certain groups, or distorting the accuracy of represented relationships.

Mitigation strategies. To reduce misuse, ethical guidelines should be established to govern the responsible development and application of this technology. Promoting transparency about generated content and integrating fairness and diversity considerations into dataset selection are key strategies for mitigating potential harms.

References

[1] 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. 1

[2] 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

[3] 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. 3

[4] 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

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

[6] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch,

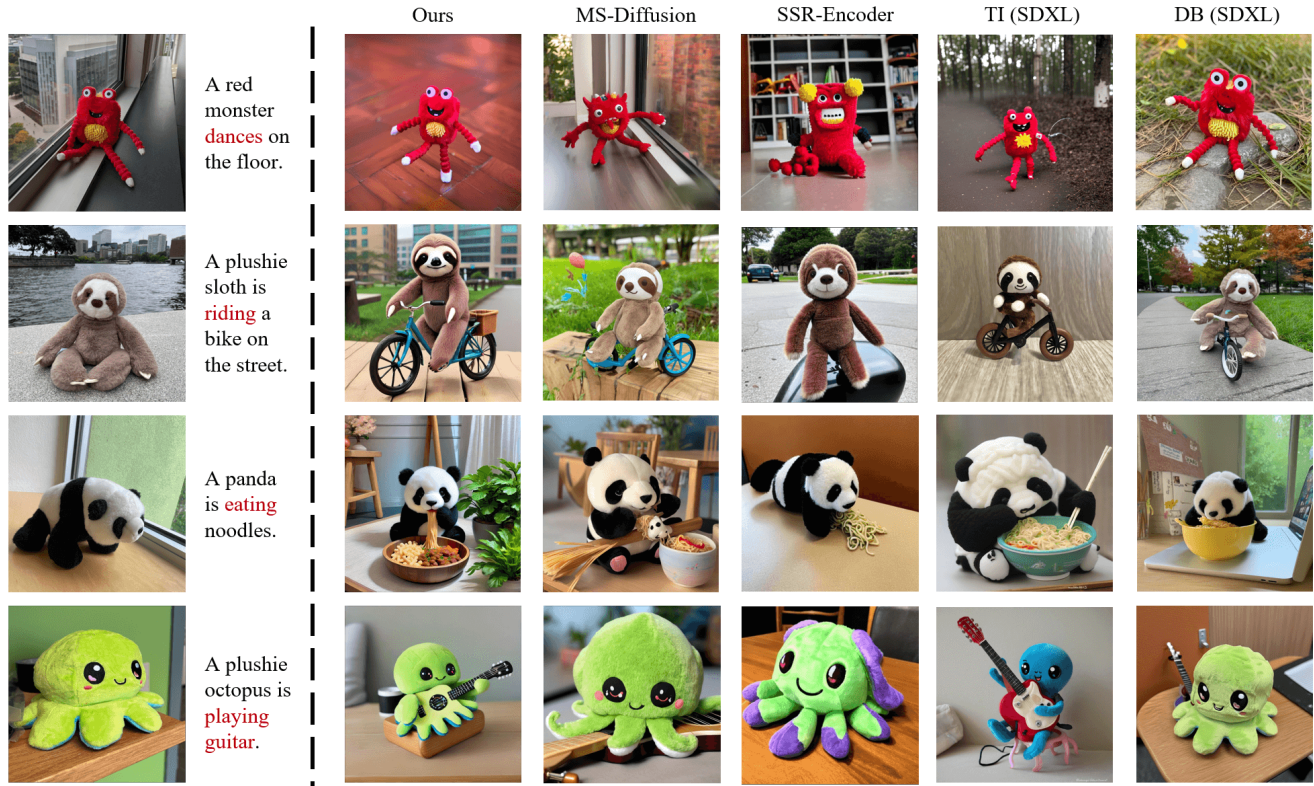


Figure 9. Single-object comparison. TI and DB indicate Textual Inversion and DreamBooth, respectively. Our methods achieve the best balance between relation generation and identity preservation.

Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 3

- [7] 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. 3
- [8] Jie Yang, Ailing Zeng, Ruimao Zhang, and Lei Zhang. Uni-pose: Detecting any keypoints. *ECCV*, 2024. 1
- [9] 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. 3



Figure 10. Our DreamRelation is compatible with SDXL to address the Relation Inversion task.

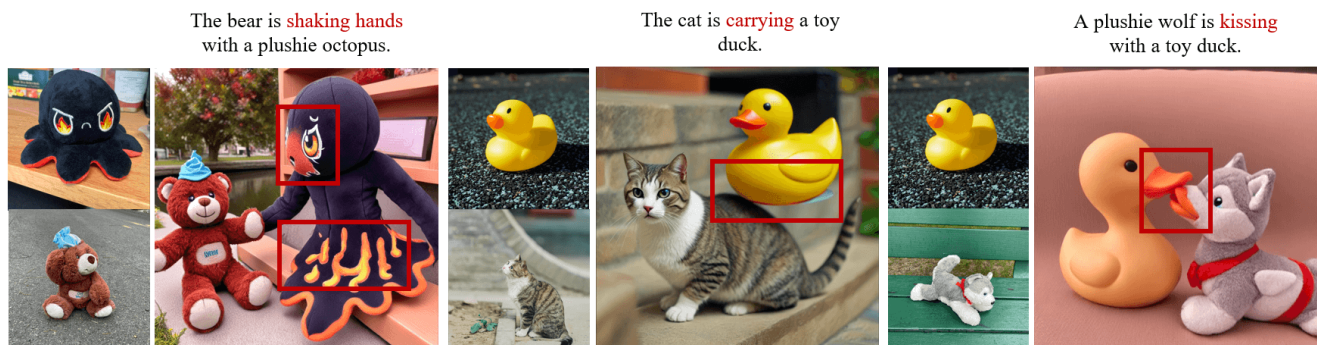


Figure 11. Failure cases of our DreamRelation.

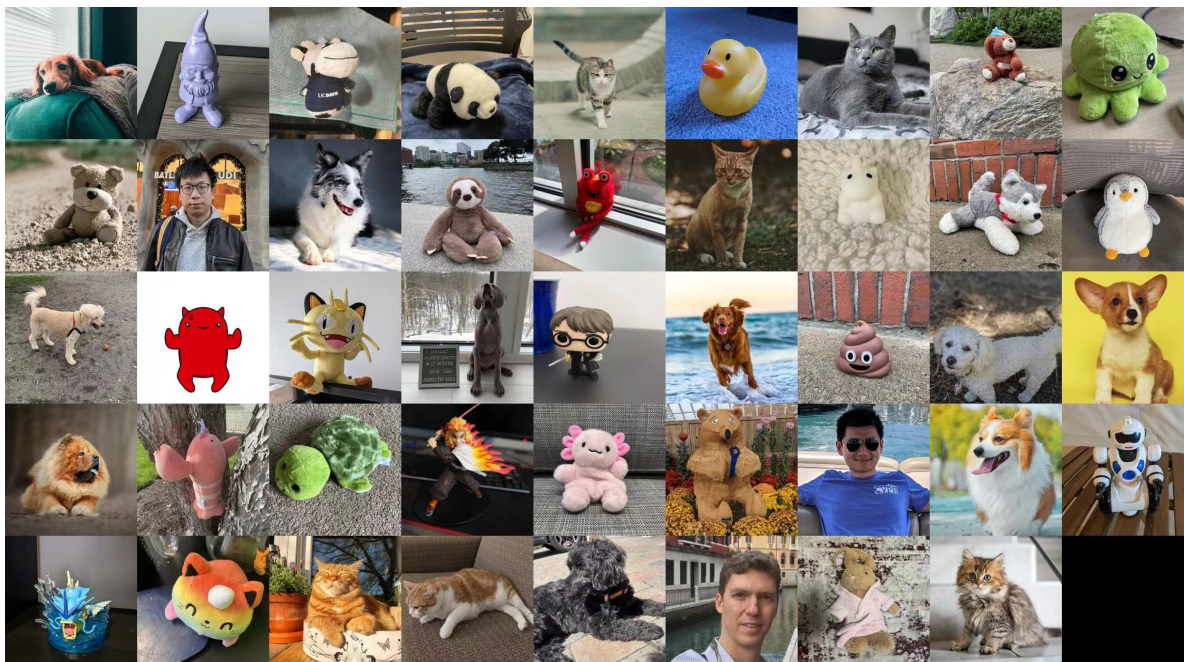


Figure 12. Objects in our proposed RelationBench

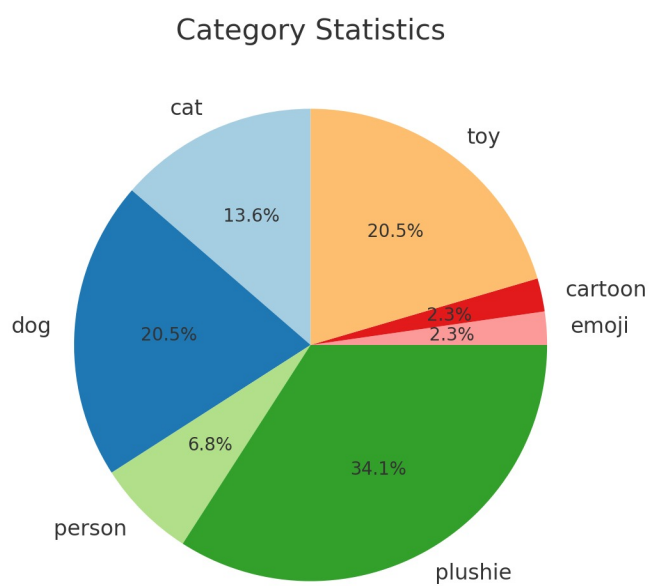


Figure 13. Object category in RelationBench

Table 4. Text prompt in RelationBench.

No.	Prompt
1	A { } is playing guitar on a park bench.
2	A { } is playing piano in a grand hall.
3	A { } is eating dinner in a bustling restaurant.
4	A { } is dancing in the moonlight.
5	A { } is lifting weights in a modern gym.
6	A { } is reading a book by the fireplace.
7	A { } is skiing down a steep slope in the Alps, with snowflakes falling gently.
8	A { } is sleeping peacefully in a hammock under the shade of a palm tree.
9	A { } is cooking lunch in a kitchen.
10	A { } is singing on stage during a vibrant music festival.
11	A { } is riding a bike along the scenic countryside road.
12	A { } is riding a horse on the grassland.
13	A { } is riding a motorbike on the street.
14	A { } is playing soccer on football playground.
15	A { } is playing chess with a { } under a tree.
16	A { } is partner dancing with a { } in a vintage ballroom.
17	A { } is carrying a { } on the diving room.
18	A { } is fencing with a { } in an elegant arena.
19	A { } shakes hands with a { } in the forest.
20	A { } is kissing a { }.
21	A { } is playing basketball with a { } on a street court.
22	A { } is wrestling with a { } in a championship ring.
23	A { } is hugging a { } in front of the mountain.
24	A { } is fighting with a { } in a garden.
25	A { } is sitting back to back with a { } on a hilltop.



Figure 14. Single-object comparison with our base model MS-Diffusion



Figure 15. Incorporate with CogVideoX-5b-I2V.