ROICtrl: Boosting Instance Control for Visual Generation

Supplementary Material

Contents

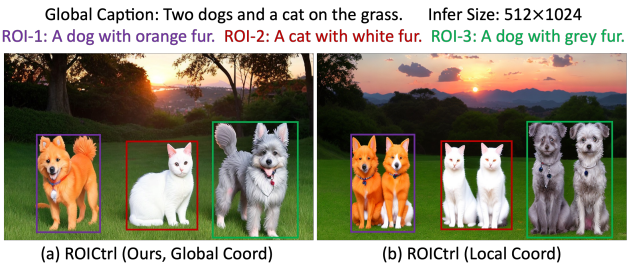


Figure 1. Comparison of regional and global coordinate conditioning. Regional coordinate conditioning leads to repetition issues when the inference size is doubled relative to the training size.

Table 1. Ablation study of design choices in ROICtrl.

Models	ROICtrl-Bench		MIG-Bench		Instdiff-Bench		
	mIoU	Acc	mIoU	Acc	AP	Color Acc	Texture Acc
ROICtrl (Ours)	0.652	48.7	0.66	0.73	41.0	62.3	29.3
multi-scale roi → single-scale roi	0.639	49.6	0.65	0.73	40.0	62.5	29.9

1. Detailed Evaluation Settings

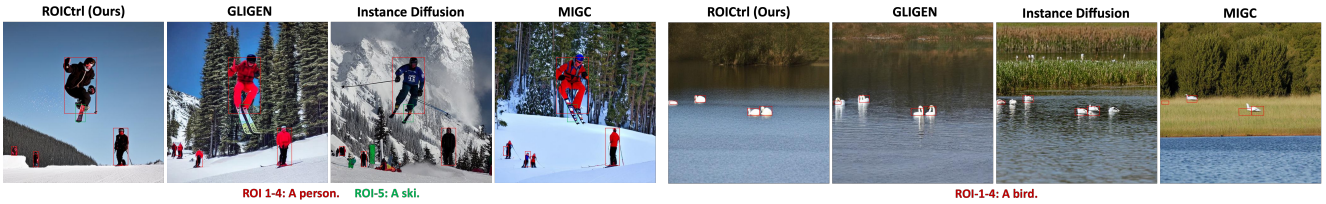
1.1. ROICtrl-Bench

ROICtrl-Bench contains 200 samples, divided into groups $\{1, 2, 3, 4, 5, 6\text{-}10, 11\text{-}15, 16\text{-}20, 21\text{-}25, 26\text{-}30\}$ based on instance counts. Each group includes 20 examples randomly selected from the MS-COCO 2017 evaluation set [4]. Half of the evaluation examples contain small-sized ROIs with spatial size smaller than 32×32.

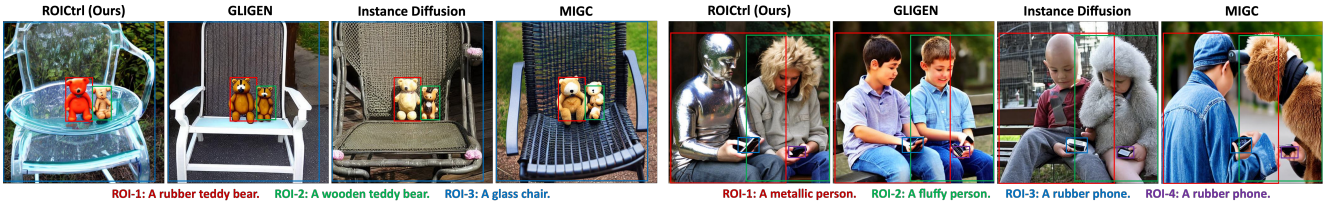
As discussed in Sec. 4.2, we create four types of instance captions for each example, corresponding to four tracks, resulting in a total of 800 evaluation examples. For template-based captions in tracks 1 and 2, we follow the GLIGEN [3] evaluation protocol, using only category labels as instance captions. For free-form instance captions, we leverage a multi-modal large language model to provide instance captions. We report the spatial alignment (mIoU) and regional text alignment (Acc) metrics for each track.

1.2. InstDiff-Bench

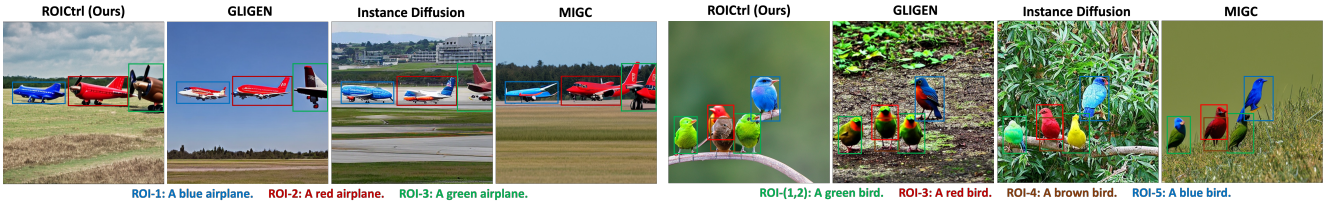
InstDiff-Bench [8] uses the entire MS-COCO 2017 evaluation set [4] as its benchmark. For spatial alignment evaluation, it calculates YOLOv8 detection metrics (AP) based on in-distribution instance captions (*i.e.*, object categories). To assess the model's ability to generate out-of-distribution instance captions, it defines 8 common colors: black, white, red, green, yellow, blue, pink, purple, and 8 common textures: rubber, fluffy, metallic, wooden, plastic, fabric, leather, glass. For each instance, a texture or color adjective is randomly selected from that predefined adjective pool, and the caption is constructed using the template `[adj.]-[noun.]`. InstDiff-Bench inputs the cropped



(a) Qualitative comparison of ROICtrl and previous methods on ***small-sized ROIs*** in InstDiff-Bench [8]. (***Zoom in*** for details.)



(b) Qualitative comparison of ROICtrl and previous methods on ***out-of-distribution instance captions*** in InstDiff-Bench [8].



(c) Qualitative comparison of ROICtrl and previous methods on ***out-of-distribution instance captions*** in MIG-Bench [9].

Figure 2. Qualitative comparison of ROICtrl and previous methods on InstDiff-Bench [8] and MIG-Bench [9]. We provide examples for small-sized ROIs and out-of-distribution instance captions.

box into the CLIP model to predict attributes (colors and textures) and evaluates the accuracy of the predicted adjectives (*i.e.*, $\text{Acc}_{\text{color}}$ or $\text{Acc}_{\text{texture}}$). Additionally, it reports the regional CLIP score for each instance caption.

1.3. MIG-Bench

MIG-bench [9] mainly evaluates spatial alignment and regional text alignment on out-of-distribution instance captions. It selects 800 layouts from COCO, randomly assigns a color to each instance, and constructs the caption based on the template `[adj.]-colored-[noun]`. In their evaluation, they filter out small-sized ROIs and dense ROIs with more than 6 instances. MIG-bench primarily reports spatial alignment (mIoU) and regional text alignment (instance success rate).

2. Additional Experiments

2.1. Qualitative Comparison

We have demonstrated the qualitative comparison on ROICtrl-Bench in Sec. 4.3 of the main paper. Therefore, in this section, we primarily present the qualitative comparison on InstDiff-Bench [8] and MIG-Bench [9].

Small-Sized ROIs. As shown in Fig. 2(a), previous instance diffusion [8] tends to generate redundant instances beyond the box, while GLIGEN [3] and MIGC [9] do not accurately follow the box. In comparison, ROICtrl can accurately generate small-sized ROIs.

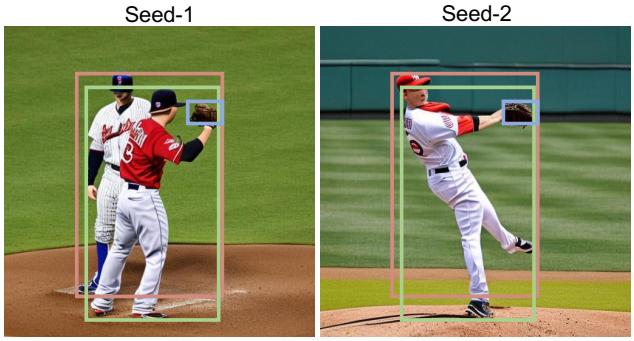


Figure 3. Limitation of ROICtrl. ROICtrl prioritizes the use of instance captions to solve attribute binding but performs unstably when instance boxes with similar captions are heavily overlapped.

Out-of-Distribution Instance Caption. As shown in Fig. 2(b, c), previous methods do not accurately follow the instance caption when generating out-of-distribution attributes and exhibit attribute leakage. In comparison, ROICtrl follows the instance caption accurately.

Figure 4. Applications of ROICtrl on Video Instance Control. We encourage readers to click and play the video clips in this figure using Adobe Acrobat.

2.2. Ablation Study

Regional *vs.* Global Coordinate Conditioning. ROICtrl employs global coordinate conditioning, following GLIGEN [3], whereas recent works such as MIGC [9] and BlobGEN [5] utilize regional coordinate conditioning. Tab. **??** shows that local coordinate conditioning achieves slightly better quantitative performance. However, in real-world applications, we find that local coordinate conditioning does not generalize well to varying resolutions. As shown in Fig. 1, with an inference size of 512×1024 (double the training size of 512×512), regional coordinate conditioning suffers from subject repetition issues.

Multi-Scale ROIs. In ROICtrl, we set multi-scale ROIs to adapt to the multi-scale feature maps of U-Net. Assume the spatial feature size of U-Net is $H = W = R$. The ROI size $r \times r$ is defined by the relation $r = 6 \times \log_2 R - 11$. For example, if the diffusion model operates at a resolution of 512×512, where its feature resolutions are $R = \{64, 32, 16, 8\}$, the corresponding ROI sizes are $r = \{25, 19, 13, 7\}$. We compare these multi-scale ROIs with single-scale ROIs, where $r = \{7, 7, 7, 7\}$, as shown in Tab. 1. Multi-scale ROIs achieve better spatial alignment while maintaining similar text alignment.

3. Limitation and Future Works

3.1. Limitation Analysis

The attribution leakage problem is largely addressed in ROICtrl, as we prioritize using instance captions in the learnable blending process. However, generating the same instance for highly overlapping bounding boxes remains a challenge. As illustrated in Fig. 3, when the boxes exhibit significant overlap, the model need to rely on the global caption for additional information. However, ROICtrl tends to favor instance captions instead, making it unstable to solve this case. We believe that further improving the learnable blending strategy to dynamically reweight the global and instance captions could solve this issue.

3.2. Future Works

Apply ROICtrl to Video Instance Control. In our preliminary experiments on VideoCrafter2 [1], we find that with slight fine-tuning of the pretrained ROICtrl on a video dataset (about 2K iterations), ROICtrl can be used to control video instances, as shown in Fig. 4. However, improving the temporal consistency of video instances remains a challenge, presenting a potential direction for future development of ROICtrl.

Apply ROI-Unpool to Diffusion Transformers. ROICtrl is primarily designed for UNet-based diffusion models. Another future direction is to explore combining ROI-Unpool with transformer-based diffusion models [2, 6] to explicitly separate instance features and inject instance control.

3.3. Potential Negative Social Impact

This project aims to provide the community with an effective method for performing multi-instance control. However, a risk exists wherein malicious entities could exploit this framework, in combination with image customization, to generate deceptive images of multiple public figures, potentially misleading the public. This concern is not owing to our approach but rather a shared consideration in concept customization. One potential solution to mitigate such risks involves adopting methods similar to anti-dreambooth [7], which introduce subtle noise perturbations to the published images to mislead the customization process. Additionally, applying unseen watermarking to the generated image could deter misuse and prevent them from being used without proper recognition.

References

[1] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7310–7320, 2024. 3

[2] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 3

[3] 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, 3

[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[5] Weili Nie, Sifei Liu, Morteza Mardani, Chao Liu, Benjamin Eckart, and Arash Vahdat. Compositional text-to-image generation with dense blob representations. *arXiv preprint arXiv:2405.08246*, 2024. 3

[6] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 3

[7] Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N Tran, and Anh Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2116–2127, 2023. 3

[8] 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. 1, 2

[9] Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. Migc: Multi-instance generation controller for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6818–6828, 2024. 2, 3