Learning from Mistakes: Iterative Prompt Relabeling for Text-to-Image Diffusion Model Training

Xinyan Chen[1,2,*]   **Jiaxin Ge**[1,*]   **Tianjun Zhang**[3,*]
Jiaming Liu[1]   **Shanghang Zhang**[1,†,‡]

[*]Equal contributions   [†]Corresponding author
[1]State Key Laboratory of Multimedia Information Processing,
School of Computer Science, Peking University
[2] University of Science and Technology of China   [3]UC Berkeley
[‡]shanghang@pku.edu.cn

Abstract

Diffusion models have shown impressive performance in many domains. However, the model's capability to follow natural language instructions (e.g., spatial relationships between objects, generating complex scenes) is still unsatisfactory. In this work, we propose **I**terative **P**rompt **R**elabeling (IPR), a novel algorithm that aligns images to text through iterative image sampling and prompt relabeling *with feedback*. IPR first samples a batch of images conditioned on the text, then relabels the text prompts of unmatched text-image pairs with classifier feedback. We conduct thorough experiments on SDv2 and SDXL, testing their capability to follow instructions on spatial relations. With IPR, we improved up to 15.22% (absolute improvement) on the challenging spatial relation VISOR benchmark, demonstrating superior performance compared to previous RL methods. Our code is publicly available at https://github.com/cxy000000/IPR-RLDF.

1 Introduction

Recent advancements in the field of image generation have been notably driven by diffusion models, especially in the area of text-to-image conversion (Ho et al., 2020). However, a significant challenge arises when these models are tasked with interpreting and executing complex instructions, particularly those involving spatial relationships (Gal et al., 2022; Saharia et al., 2022). One simple example is that a prompt like "a dog left to a car" often results in images where the spatial relationship "left to" is not accurately depicted. This limitation underscores a crucial gap in the current models' ability to understand and render intricate spatial relationships.

In response to these challenges, we investigate how models can self-improve without using extensive training data. Some studies have found that self-supervised learning is useful for training diffusion models (Hu et al., 2023). However, these methods still lack an effective self-correction mechanism, making the self-generated dataset either noisy or being trimmed off too much. Recent advances in large language models (LLM) have found that learning from mistakes and receiving language feedback can effectively enhance the model's reasoning capacity (An et al., 2023). Inspired by this, in our research, we study whether we can effectively use the originally incorrect text-image pairs to train the diffusion model.

In particular, we propose **I**terative **P**rompt **R**elabeling (IPR), a novel technique designed to enhance the alignment of images with text through an iterative process of image sampling and prompt relabeling. Our approach uses rich language feedback for all the images and a simple reward design. A high-level overview of our approach is shown in Figure 1. We begin by establishing a reward function based on an external detection model that automatically classifies the correctness of an image and assigns rewards based on the detection results. This enables a straightforward reward design and neglects the complicated training of reward models. Then, we use prompt relabeling to relabel the input prompt of the mismatched image-text pairs based on the results of the detection model. This allows the models to use the mis-generated images with the correct version. Finally, we adopt iterative training that continuously trains the model with its self-generated images. Iterative training allows dynamically scaling up the dataset and receiving additional feedback through multiple training rounds, which can further enrich the feedback and progressively refine the model's performance.

We demonstrate the efficacy of IPR across SDv2 (Rombach et al., 2022) and SDXL (Podell et al., 2023), trained with and w/o LoRA, and we test the performance on the challenging spatial relation task, where we observe a substantial improvement of up to 15.22% (absolute improvement) on the VISOR (Gokhale et al., 2022) benchmark. This performance underlines the potential of IPR in pushing
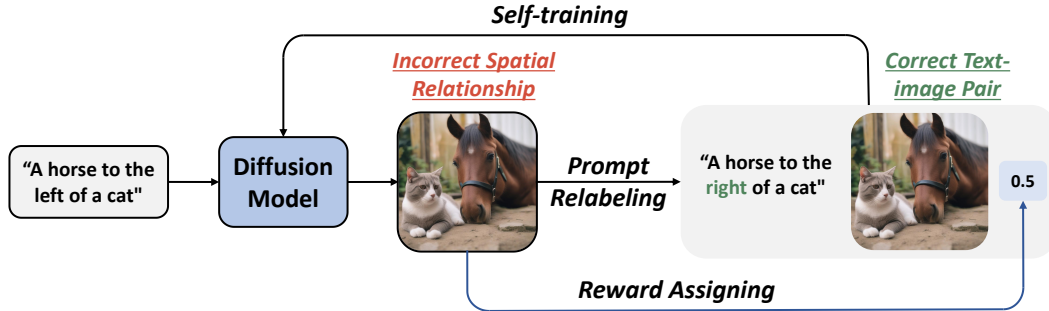
Figure 1: **A high-level overview of our approach.** We enhance the alignment of images with text through an iterative process of image sampling and prompt relabeling.

the boundaries of text-to-image generation models, especially in terms of understanding and rendering complex spatial relationships.

2 Related Work

Text-to-image models have facilitated the generation of high-resolution, multi-styled images (Saharia et al., 2022; Ramesh et al., 2021; Zhang et al., 2023b). However, training diffusion models from scratch demands a substantial amount of data and time. Therefore, various fine-tuning strategies have been explored. These approaches include associating a unique identifier with a particular subject (Ruiz et al., 2023), introducing new embeddings to represent user-provided concepts (Gal et al., 2022), adapting compositional generation (Liu et al., 2022), and implementing Low-Rank Adaptation (Hu et al., 2021). Our work utilizes LoRA to fine-tune the diffusion model and follow the reinforcement learning line of work to rescale loss for effective training.

Recent research has demonstrated the efficacy of prompt relabeling techniques. TEMPERA (Zhang et al., 2022b) offers interpretable prompts tailored to various queries through the creation of an innovative action space, enabling flexible adjustments to the initial prompts. Furthermore, the Hindsight Instruction Relabeling (HIR) (Zhang et al., 2023a) approach conceptualizes the instruction alignment problem as a goal-reaching problem within the context of decision-making. It entails the conversion of feedback into instructions by re-labeling the original instructions. Our approach represents a significant advancement as it employs prompt relabeling techniques with diffusion models for the first time.

3 Method

3.1 Method Overview

Our approach adopts four different stages: diffusion model sampling, reward-based loss rescal-

Table 1: Combined results of spatial accuracy and CLIP score on four RLDF settings. (1) Direct: The original diffusion models. (2) RLDF: Only applying RLDF on diffusion models. (3) IPR-RLDF: Our method (with RLDF). The results show that IPR outperforms other methods in generating images with correct spatial relationships.

Score Type	Method	SDv2(1)	SDv2 (LoRA)	SDXL (LoRA)	SDv2(2)
Spatial Accuracy (%)	Direct	18.75	18.75	27.00	17.00
	RLDF	21.50	22.00	29.75	22.44
	IPR-RLDF	**28.50**	**25.25**	**31.25**	**32.22**
CLIP Score	Direct	25.75	25.75	27.41	-
	RLDF	**26.67**	26.09	28.68	-
	IPR-RLDF	25.87	**26.15**	**28.74**	-

ing, prompt relabeling, and iterative training. Our pipeline is demonstrated in Figure 2. First, IPR samples a batch of images from the diffusion model using the text prompts. Then, IPR uses a detection model to check the correctness of the text-image pair and relabels the original text prompt to attain a correct text-image pair. Next, IRP assigns a reward to the text-image pair based on the correctness of the original text-image pair. Finally, IRP trains the diffusion model on the new relabeled text-image pairs in an iterative manner. We will introduce the details of each stage in the following sections.

3.2 Prompt Relabeling

The sampled images from the diffusion model consist of the ones that are aligned with the text prompt and the ones that are not. We then relabel the inconsistent image-text pairs to ensure that the textual description accurately reflects the content of the generated image. The algorithm consists of the following:

1. Detect objects in the generated image using a detection model, yielding bounding boxes (bbox) for each object.
2. Compare the number of detected objects with the object count specified in the original prompt.
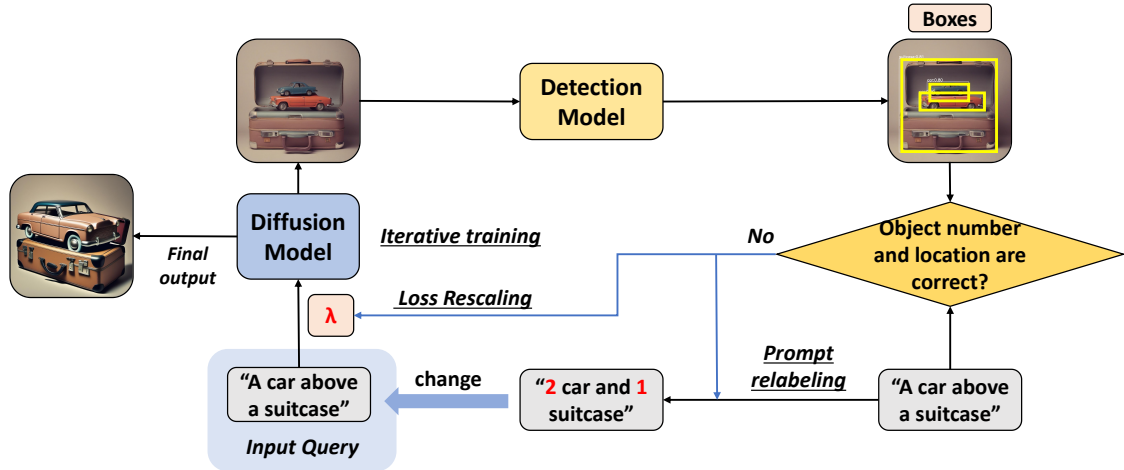3. If the object count matches, analyze the bound-

Figure 2: **The general pipeline of IPR.** Our approach adopts four different stages: (1) diffusion model sampling, (2) reward-based loss rescaling, (3) prompt relabeling, and (4) iterative training.

ing box centers to determine the actual spatial relationship between objects. Modify the original prompt to accurately describe this relationship.

4. If the object count does not match, revise the prompt to reflect the actual number and type of objects in the image (e.g., "2 cats and 1 dog").

3.3 Detection-Based Loss Rescaling

In our method, when attaining the relabeled text-image pairs, we assign a reward-rescaling factor to each pair. This reward is used to rescale the loss when training the diffusion model. The loss function is rescaled based on the following $L_{\text{rescaled}} = \mathbb{E}_{x_i,c_i} \left[\lambda_{x_i,c_i} L_\theta(x_i, c_i) \right]$, where L_θ is the standard diffusion model loss as defined in (Ho et al., 2020). The modulatory factor λ is defined as: 1, when the model correctly generates the text-image pair; 0.5, when the text-image pair is incorrect and needs relabeling.

The rescaling factor λ balances the model's capability to follow text prompts with the model's generated image distribution not too far away from the original one. A larger λ pushes the model to learn from the mis-generated text-image pairs while a smaller λ keeps the model closer to its original distribution. After the rescaling weight is assigned to each text-image pair, we obtain a dataset of the model's self-generated text-image pairs. We use this dataset to train the diffusion model.

3.4 Iterative Training

The self-generated dataset is often not large enough. So we train the diffusion model by repeating this process iteratively. Specifically, after training the model for an iteration, we use the updated model to repeat the previous process and get a

Table 2: Combined results of spatial accuracy and CLIP score on RLHF settings.

Score Type	Method	SDv2	SDv2(LoRA)
Spatial Accuracy(%)	Direct	18.75	18.75
	RLHF	22.25	24.00
	IPR-RLHF	**27.00**	**26.00**
CLIP Score	Direct	25.75	25.75
	RLHF	26.00	25.72
	IPR-RLHF	**26.06**	**26.10**

new dataset that contains text-image pairs and their corresponding rescaling weight. Then we use the new dataset to train on the new model to get the updated model for the next iteration. The training procedure is formulated as follows: $\theta_{\text{new}} = \text{Train}(\theta_{\text{old}}, \{(x_0^{(i)}, c_{\text{new}}^{(i)}, L_{\text{rescaled}}^{(i)})\}_{i=1}^N)$ where N denotes the number of samples in each iteration. This iterative process leads to continual improvement in the model's ability to produce spatially coherent images aligned with textual descriptions.

4 Experiments

In this section, we assess IPR's effectiveness through multiple settings. We first introduce our experiment settings. Then, we present the quantitative results and qualitative results of IRP. Finally, we present ablation studies to fully analyze our method.

4.1 Experimental Settings

Settings and baselines Our setting is mainly based on SDv2 (Rombach et al., 2022) and SDXL (Podell et al., 2023). Specifically, we use: (1,2) Two variants of standard SDv2 fine-tuning. (3) SDv2 LoRA fine-tuning, by freezing the model but training the rank-decomposition matrices injected into the UNet and text-encoder (Radford

Table 3: Ablation Study on the effect of three parts of our method (metric: spatial accuracy). (1) PR: Only applying prompt relabeling on diffusion models. (2) RLDF: Only applying RLDF on diffusion models. (3) PR-RLDF: Fine-tuning diffusion models with RLDF and prompt relabeling for one iteration. (4) IPR-RLDF: Iterative training. IPR-RLDF consistently achieves higher spatial accuracy in different settings.

Score Type	Method	SDv2(1)	SDv2 (LoRA)	SDXL (LoRA)	SDv2(2)
Spatial Accuracy (%)	PR	24.50	21.50	28.25	-
	RLDF	21.50	22.00	29.75	22.44
	PR-RLDF	25.75	24.25	30.00	25.22
	IPR-RLDF	**28.50**	**25.25**	**31.25**	**32.22**

et al., 2021). (4) SDXL LoRA fine-tuning. In each of the settings, we train 3 epochs for each iteration. Other experimental details are provided in Appendix A. We use GLIPv2 (Zhang et al., 2022a) as the detection model to assign a rescaling reward for the text-image pair. We refer to this as Reinforcement Learning with Detection Feedback (RLDF) in our later experiments.

We compare with the existing RL-trained diffusion baseline: Reinforcement Learning with Human Feedback (RLHF). We use ImageReward (Xu et al., 2024) as the human feedback reward model. This model was pre-trained on large paired datasets that have been annotated by humans. We use the Reward Feedback Learning (ReFL) method for fine-tuning diffusion models.

Dataset We use 100 self-training prompts from the VISOR benchmark (Gokhale et al., 2022), a challenging dataset focusing on spatial relations.

Metrics We use GLIPv2 for evaluation. This model detects objects and represents them as (object name, bounding box) pairs. It then evaluates accuracy by verifying both the correct count of detected objects and their spatial relationships, determined by comparing the bounding box centers. Images are classified as correct if both object count and spatial relations are accurate; otherwise, they are deemed incorrect. Besides the spatial accuracy, we also use the standard evaluation metric CLIP score (Radford et al., 2021).

4.2 Quantitative Results

Spatial accuracy The result shows that IPR-RLDF significantly outperforms the baseline across all settings and suggests that IPR is an effective algorithm in training diffusion models to accurately depict spatial relationships, outperforming conventional training without language feedback or iterative processes. The consistent improvement of

Table 4: Different kinds of spatial accuracy (left-right, above-below, and object number) of our model and baselines in RLDF settings, show that IPR-RLDF outperforms other baselines in recognizing a variety of spatial positional relationships, even though there are minor fluctuations in the above-below spatial relationship of fine-tuned SDXL model.

Spatial Accuracy (%)	Method	SDv2(1)	SDv2 (LoRA)	SDXL (LoRA)
Left-right	Direct	20.63	20.63	25.00
	PR	20.63	16.88	30.63
	RLDF	18.13	18.75	28.13
	PR-RLDF	20.00	22.50	29.38
	IPR-RLDF	**26.25**	**23.13**	**36.25**
Above-below	Direct	17.50	17.50	28.33
	PR	27.08	24.58	26.67
	RLDF	23.75	24.17	**30.83**
	PR-RLDF	29.58	25.42	30.42
	IPR-RLDF	**30.00**	**26.67**	27.92
Object number	Direct	46.75	46.75	51.00
	PR	53.00	48.75	54.75
	RLDF	46.75	48.00	56.50
	PR-RLDF	52.75	47.75	51.50
	IPR-RLDF	**58.75**	**49.75**	**58.25**

overall settings also verifies the method's robustness in different model training and techniques.

Comparison between IPR-RLDF and IPR-RLHF Table 2 demonstrates that IPR-RLHF outperforms RLHF method in both spatial accuracy and CLIP score across two settings, illustrating the effectiveness of the iterative prompt relabeling method with diverse feedback. Although IPR-RLDF achieves a similar increase compared to the IPR-RLHF method (Tables 1 and 2), it can get feedback easily from a general detection model, rather than a human feedback reward model specifically trained for the text-image alignment task.

4.3 Qualitative Results

Figure 3 provides a visual comparison of the original SDXL model with fine-tuned versions using RLDF, PR-RLDF, and IPR-RLDF, across four different prompts. The original SDXL model frequently misinterprets the number and placement of objects, a challenge also observed in RLDF. In contrast, it is evident that our algorithm outperforms both RLDF and the original SDXL model in these aspects, showcasing enhanced spatial awareness and accuracy in depicting the specified objects, while sacrificing some details.

4.4 Ablation Study

Method breakdown In Table 3, we conduct an ablation study to assess the impact of three components of our method, focusing on SDv2 and SDXL settings. The results clearly show that each component is essential to our algorithm's performance. Both prompt relabeling and iterative training significantly improve RLDF outcomes.

Original SDXL model	RLDF	P-RLDF	IP-RLDF(for 2 iters)
number ✗	location ✗	location ✗	number ✓ location ✓

(a) A horse to the left of a cat

Original SDXL model	RLDF	P-RLDF	IP-RLDF(for 2 iters)
number ✗	number ✗	number ✓ location ✓	number ✓ location ✓

(b) A book below a sports

number ✗	number ✗	number ✓ location ✓	number ✓ location ✓

(c) A stop sign above a book

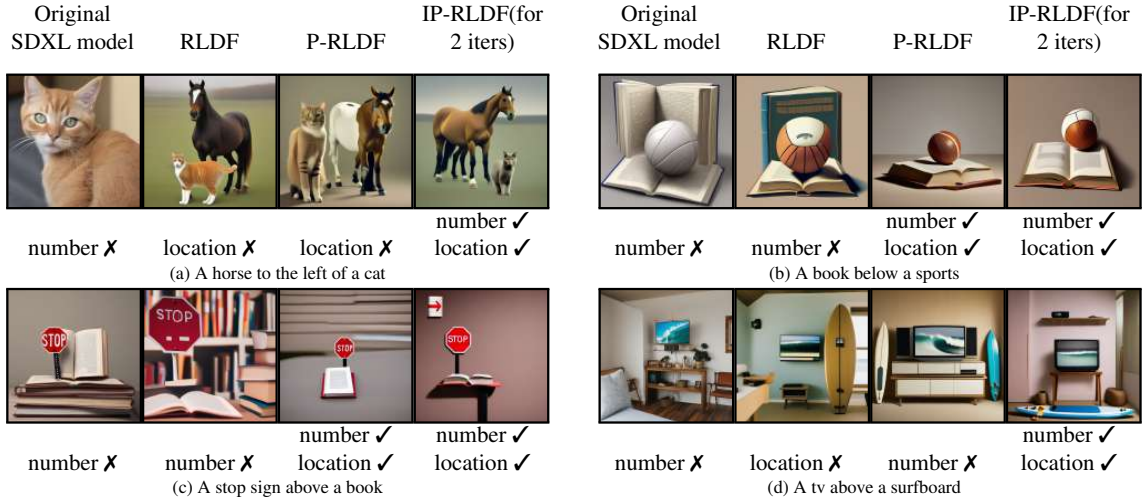| number ✗	location ✗	number ✗	number ✓ location ✓

(d) A tv above a surfboard

Figure 3: Visual comparison of the original SDXL model with fine-tuned versions using RLDF, PR-RLDF, and IPR-RLDF, across four different prompts. Our algorithm demonstrates superior spatial awareness and accuracy in object depiction, while sacrificing some details.

Table 5: Comparison of spatial accuracy at varying GLIP score thresholds for prompt relabeling and evaluation. Our algorithm consistently improves the spacial accuracy over different detection thresholds.

Score Type	Thresholds for Prompt Relabeling	Thresholds for Evaluation	Direct	PR-RLDF	IPR-RLDF
Spacial Acc (%)	0.45	0.45	18.75	25.75	**28.5**
Spacial Acc (%)	0.6	0.6	16.50	20.00	*20.75*

Table 6: Spatial accuracy of models in RLHF settings, showing that our method outperforms other baselines in recognizing a variety of spatial positional relationships.

Spatial Accuracy (%)	Method	SDv2	SDv2 (LoRA)
Left-right	Direct	20.63	20.63
	RLHF	**21.88**	18.13
	IPR-RLHF	21.88	**21.25**
Above-below	Direct	17.50	17.50
	RLHF	22.50	27.92
	IPR-RLHF	**30.42**	**29.17**
Object number	Direct	46.75	46.75
	RLHF	47.75	48.75
	IPR-RLHF	**52.25**	**51.25**

Spatial relationship type study To further investigate how our algorithm performs across a variety of spatial relationships, we apply an ablation study to different spatial relationships across RLDF and RLHF settings, with a focus on the SDv2 and SDXL models. The findings, detailed in Tables 4 and 6, reveal that our model consistently outperforms the baseline across diverse spatial relations.

GLIP score thresholds To investigate the impact of the detection model on the final results, we conduct experiments on the SDv2 model, using different thresholds for the GLIP score during detection. For prompt relabeling, we change the thresholds of the glip score from 0.45 to 0.6. We present the results in Table 5. We found that when the threshold for the GLIP score is set to 0.6, the performance is not as good as 0.45. This is likely because when

GLIPv2 only detects objects with a GLIP score above 0.6, the alignment between the image and the relabeled prompt decreases. Nevertheless, our algorithm consistently improves the spacial accuracy over different detection thresholds.

Additional ablation studies on reward assignments, unseen prompts, color tasks and rephrased prompts can be found in Appendix A.2.

5 Conclusion

In this work, we present IPR in response to the challenging task of spatial location generation. IPR is a novel algorithm that designs rich language feedback and incorporates it with the detection model rewards. Then it trains the diffusion model by iteratively receiving rewards and language feedback. This algorithm is a plug-and-play method that is applicable to a range of diffusion models. Extensive results show the model's effectiveness on the challenging spatial relationship benchmark across various settings.

Limitations

Our results demonstrate impressive out-of-distribution generation abilities. However, we haven't fully studied the mechanism of how self-training induces such ability. We leave this for future experiments.

References

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. 2023. Learning from mistakes makes llm better reasoner. *arXiv preprint arXiv:2310.20689*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *International Conference on Computer Vision (ICCV)*, pages 2425–2433.

Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. 2019. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems (NeurIPS)*, 32.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. 2018. Coco-stuff: Thing and stuff classes in context. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1209–1218.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, pages 213–229. Springer.

Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. 2019. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations (ICLR)*.

Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. 2022. Benchmarking spatial relationships in text-to-image generation. *arXiv preprint arXiv:2212.10015*.

Agrim Gupta, Piotr Dollar, and Ross Girshick. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5356–5364.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:6840–6851.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Vincent Tao Hu, David W Zhang, Yuki M Asano, Gertjan J Burghouts, and Cees GM Snoek. 2023. Self-guided diffusion models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18413–18422.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:78723–78747.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *International Conference on Computer Vision (ICCV)*, pages 1780–1790.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022. Grounded language-image pre-training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10965–10975.

Yiting Li, Haiyue Zhu, Yu Cheng, Wenxin Wang, Chek Sing Teo, Cheng Xiang, Prahlad Vadakkepat, and Tong Heng Lee. 2021. Few-shot object detection via classification refinement and distractor retreatment. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15395–15403.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *International Conference on Computer Vision (ICCV)*, pages 2980–2988.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. 2022. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision (ECCV)*, pages 423–439. Springer.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning (ICML)*, pages 8821–8831. Pmlr.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510.

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. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:36479–36494.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:25278–25294.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2024. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 36.

Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. 2022a. Glipv2: Unifying localization and vision-language understanding. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:36067–36080.

Tianjun Zhang, Fangchen Liu, Justin Wong, Pieter Abbeel, and Joseph E Gonzalez. 2023a. The wisdom of hindsight makes language models better instruction followers. In *International Conference on Machine Learning (ICML)*, pages 41414–41428. PMLR.

Tianjun Zhang, Xuezhi Wang, Denny Zhou, Dale Schuurmans, and Joseph E Gonzalez. 2022b. Tempera: Test-time prompt editing via reinforcement learning. In *International Conference on Learning Representations (ICLR)*.

Tianjun Zhang, Yi Zhang, Vibhav Vineet, Neel Joshi, and Xin Wang. 2023b. Controllable text-to-image generation with gpt-4. *arXiv preprint arXiv:2305.18583*.

A Supplementary Experimental Details

A.1 Method Details

We provide a breakdown of our method Figure 4 and pseudocode in Algorithm 1 to better illustrate our method.

Algorithm 1 Iterative Prompt Relabeling (IPR)

Require: Text prompts C, pre-trained model θ, detection model D, iterations T, data X
Ensure: Refined parameters θ'
1: Initialize $\theta' \leftarrow \theta$
2: **for** iteration $= 1$ **to** T **do**
3:   $X_0 \leftarrow$ Sample_Images(Diffusion_Model, C, θ')
4:   **for** each x_0 in X_0 **do**
5:     Objects, Bounding_Box $\leftarrow D(x_0)$
6:     Count, Relation $\leftarrow$ Analyze(D)
7:     Reward $\leftarrow$ CalcReward(Count, Relation, C)
8:     $\lambda \leftarrow$ (Count, Relation match C)?1 : 0.5
9:     $L_{\text{rescaled}} \leftarrow L_{\text{DDPM}}(x_0, \theta') \times \lambda$
10:    $c_{\text{new}} \leftarrow$ Relabel(C, D)
11:    Update X with $(x_0, c_{\text{new}}, L_{\text{rescaled}})$
12:   **end for**
13:   $\theta' \leftarrow$ TrainModel(X, θ')
14: **end for**
       return θ'

A.2 Additional Ablation Studies

Reward assignment To examine the impact of rescaling factor λ, we conduct ablation experiments on the SDv2 model (fine-tuned with LoRA), using different values of λ (0.1, 0.5, and 0.7). In Table 7, we find that 0.5 yields the best performance, yet, the model is not very sensitive towards different λ values.

Unseen prompts We assess the performance of our fine-tuned models on prompts that are unseen during self-training, we randomly select 100 new prompts from the VISOR benchmark different from the 100 in the self-training process and test on our trained LoRA fine-tuned SDv2 model. As shown in Table 8, our model can also outperform the original pre-trained model in spatial accuracy on the unseen prompts. This demonstrates that such training can be generalized to unseen prompts.

Color accuracy In this work, we primarily concentrate on spatial location generation. We study whether IRP can be applicable to other tasks as well. Therefore, we provide some experiments on the color category of T2I-CompBench (Huang et al., 2023), a comprehensive benchmark designed for open-world compositional text-to-image generation tasks. Instead of a detection model, we use the LLaVA (Liu et al., 2023) model, a large-scale multimodal model that combines a vision encoder

Table 7: Ablation study of λ on SDv2 model (fine-tuned with LoRA).

Score Type	λ	Direct	RLDF	PR-RLDF
	0.1	18.75	**21.75**	20.00
Spatial Acc(%)	0.5	18.75	22.00	**24.25**
	0.7	18.75	21.25	**24.00**

Table 8: Spatial accuracy of unseen prompts on SDv2 model (fine-tuned with LoRA).

Score Type	Direct	IPR-RLDF (iter2)	IPR-RLDF (iter3)
Spatial Acc(%)	17.00	**19.25**	18.75

with an LLM for visual and language understanding, to relabel and evaluate. Figure 6 shows an example of input prompts used for evaluation and prompt relabeling. The results presented in Table 9 demonstrate that our relabeling and iterative training method can generalize to new tasks like color accuracy.

Generalization abilities on other text inputs To evaluate the generalization capabilities of our approach across diverse text inputs, we use GPT-3.5-turbo to rephrase the 100 prompts into other formats. (e.g. "a car above a suitcase" is changed to "positioned above a suitcase is a car"; "an airplane to the right of a clock" is changed to "a clock with an airplane on its right side") Then, we evaluate our fine-tuned SDv2 models on the 100 rephrased prompts. Table 10 shows that our method can be generalized to other formats of text inputs as well.

Training on rephrased prompts We also conduct experiments on finetuning SDv2 with the 100 rephrased prompts. At each iteration, we relabel the prompts using our original method. Then we use GPT-3.5-turbo to rephrase the relabeled prompts into other formats. Then, we finetune the model using these rephrased prompts. After the model has been trained, we evaluate the finetuned models using the original 100 prompts. In Table 11, we find that training on rephrased prompts can also enhance the model's spatial accuracy on original prompts.

A.3 Additional Quantitative Results

Text-to-image alignment IPR-RLDF also demonstrates remarkable gains in CLIP score over both the RLDF algorithm and the original text-to-image models in SDv2 (LoRA) and SDXL (LoRA) settings, indicating that it keeps the image's quality and achieves remarkable text-to-image alignment while enhancing the spatial correctness. One exception occurs during
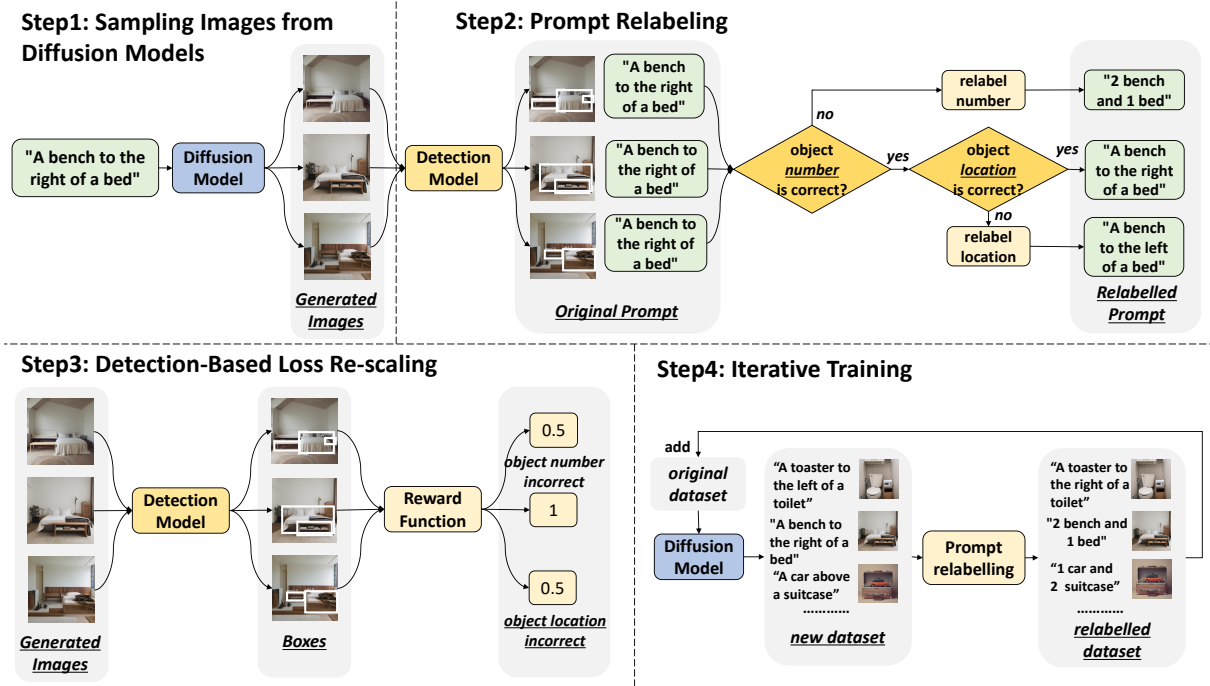
Figure 4: The process of our IPR algorithm. (1) Sampling Images from Diffusion Models: sample images from a diffusion model conditioned on textual prompts. (2) Prompt Relabeling: detect the generated image to yield a bounding box; analyze the box to modify original prompts. (3) Detection-Based Loss Re-scaling: apply a detection model to rescale the loss function. (4) Iterative Training: retrain the model with the updated dataset iteratively.

| SDv2(1) | SDv2(LoRA) | SDXL(LoRA) | | SDv2(1) | SDv2(LoRA) | SDXL(LoRA) |

(a) A laptop to the left of a cat

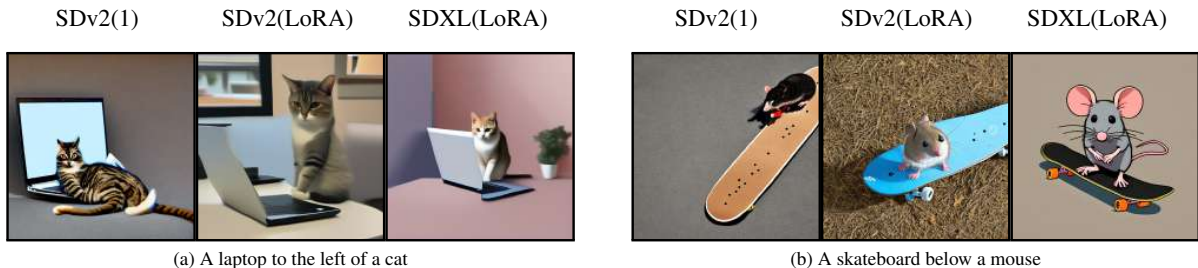

(b) A skateboard below a mouse

Figure 5: Samples from different models fine-tuned with IPR-RLDF, generated by two distinct prompts. (1) Left column: unfrozen fine-tuning on SDv2. (2)Mid column: using LoRA to fine-tune SDv2. (3) Right column: using LoRA to fine-tune SDXL. The LoRA training exhibits more notable image fidelity than unfrozen training.

Table 9: Application of IPR on color accuracy, using LLaVa model to relabel and evaluate, fine-tuning SDv2 with LoRA.

Score Type	Direct	IPR-RLDF (iter2)
LLaVA Evaluation Acc(%)	64.00	**66.5**

Table 10: The evaluation results of our approach on rephrased prompts.

Score Type	Direct	PR-RLDF
Spatial Acc (%)	16.25	**17.50**

SDv2(1), where the RLDF outperforms IPR-RLDF on the CLIP score. This is likely due to a trade-off between spatial accuracy and overall image quality, with IPR-RLDF achieving higher spatial accuracy at the expense of a slight decrement in CLIP score. As shown in Figure 7, IPR-RLDF blurs and RLDF

Table 11: Comparison between training on original and rephrased Prompts.

Score Type	Training Prompts	Direct	PR-RLDF	IPR-RLDF
Spatial Acc (%)	Original Prompts	18.75	25.75	**28.50**
Spatial Acc (%)	Rephrased Prompts	18.75	21.50	**23.00**

kept more details, while IPR-RLDF shows higher spatial accuracy.

LLaVA Evaluation Results Besides GLIP, we add another evaluation method. Since recent VLMs have shown comprehensive visual understanding abilities, we use the most recent VLM–the LLaVA model to evaluate spatial accuracy. Specifically, the model is asked to determine the spatial correctness of each input image. The results are shown in Table 12. Our method exhibits significant improvement over previous methods using LLaVA to

2945

> \# **Evaluate_input:** "Do '*a green bench and a red car*' matches the image? If any color or number of objects is wrong, please answer No. Answer 'Yes' or 'No':"
>
> -
>
> \# **Relabel_input:** "Separately describe the color of '*bench*' or '*car*' in the image. Answer in the form of: '<number> <color> *bench* and <number> <color> *car*' or '<number> <color> *bench*' or '<number> <color> *car*'. You can change the <color> and <number> in these example. You can choose <color> between: green, red, brown, blue, gold, pink, yellow."

Figure 6: An example prompt for evaluation and prompt relabeling. The evaluating input is directly passed to LLaVA. If the answer is "No", we will pass the relabeling input to LLaVA.

Table 12: **LLaVA and GPT-4-turbo Evaluation Results.** Our method exhibits significant improvement over previous methods using LLaVA and GPT-4-turbo to evaluate.

Metric	Direct	PR-RLDF	IPR-RLDF
LLaVA Evaluation Acc (%)	49.50	52.00	**53.50**
GPT-4-turbo Evaluation Acc (%)	9	13	**19**

evaluate.

GPT Models Evaluation Results We also use GPT-4-turbo to evaluate the spatial accuracy of the SDv2 model. As we did in the LLaVA evaluation experiments, the model is asked to determine the spatial correctness of each input image. The results are shown in Table 12. We found that our method also has an improvement in spatial accuracy.

Training method analysis Table 1 shows a performance difference between different training techniques. In the LoRA setting, where the weight of pre-trained models is frozen and only the rank-decomposition matrices injected are trained, the performance gain is less compared with fine-tuning all parameters. However, the LoRA training exhibits notable image fidelity compared with fine-tuning all parameters, as shown in Figure 5. This suggests a trade-off between enhancing spatial accuracy and keeping overall image quality, indicating possible directions for improvement in future research.

A.4 Model Details

A.4.1 Stable Diffusion v2 (SDv2)

Similar to Imagen (Ho et al., 2022), Stable Diffusion (Rombach et al., 2022) is a latent text-to-image model, with a frozen CLIP ViT-L/14 text encoder (Radford et al., 2021) and an 860M UNet (Ronneberger et al., 2015) constructure. This model was pretrained on 256*256 images followed by fine-tuning on 512*512 images sourced from the LAION 5B dataset (Schuhmann et al., 2022). It excelled in text-to-image tasks, while concurrently supporting image-to-image tasks. Stable Diffusion

2.0, however, employs OpenCLIP-ViT/H as its text encoder, which is trained from scratch. In our experiments, we use Stabilityai/stable-diffusion-2-1-base (512*512 resolution) and Stabilityai/stable-diffusion-2-1 (768*768 resolution) as base models, both fine-tuned on Stable Diffusion 2.0.

A.4.2 Stable Diffusion XL (SDXL)

Compared to the previous stable diffusion models, Stable Diffusion XL (Podell et al., 2023) features a UNet that is three times larger and integrates Open-CLIP ViT-bigG/14 with the original text encoder. It also introduces crop-conditioning and a two-stage model process to significantly enhance the quality of generated images. Stable Diffusion XL demonstrates improved support for shorter prompts and glyphs in images.

A.5 Training Details

A.5.1 Low-Rank Adaptation (LoRA)

Due to the considerable time and computational resources required for full fine-tuning of large models, Low-Rank Adaptation (Hu et al., 2021) was introduced for fine-tuning in particular tasks. It involves injecting rank decomposition matrices into transformer layers while freezing all other model weights. LoRA attains training quality comparable to full fine-tuning but in less time and with fewer computational resources. Initially deployed in large language models, LoRA can also be extended to cross-attention layers in text-to-image models such as Stable Diffusion, where it has demonstrated outstanding. In our experiments, we use LoRA to fine-tune SDv2 and SDXL, validating that our IPR algorithm serves as an additional plug-and-play algorithm capable of integration across various diffusion model settings.

Dataset We use 100 self-training prompts from the VISOR benchmark (Gokhale et al., 2022), a challenging dataset focusing on spatial relations. In each iteration, we sample 400 data in total to do

SDv2(1) RLDF SDv2(1) IP-RLDF   SDv2(1) RLDF SDv2(1) IP-RLDF

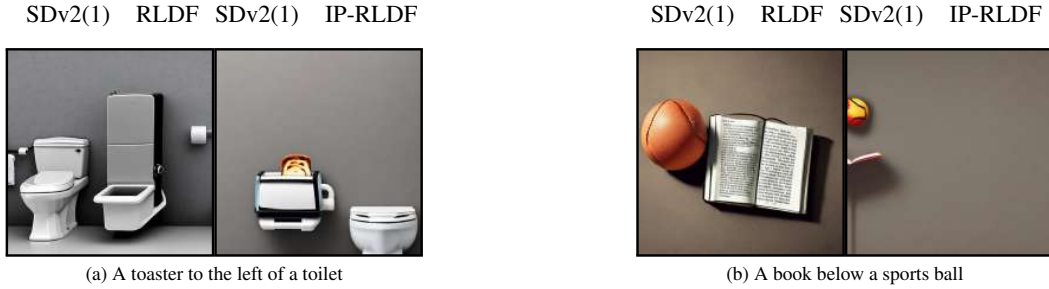(a) A toaster to the left of a toilet   (b) A book below a sports ball

Figure 7: Samples from different fine-tuned models. (1) Left column: unfrozen fine-tuning on SDv2 with RLDF. (2) Right column: unfrozen fine-tuning on SDv2 with IPR-RLDF. Compared with RLDF, the SDv2 model fine-tuned with IPR-RLDF exhibits notable image fidelity but lower spatial accuracy.

Table 13: Comparison between different settings of RLDF and RLHF, from five different aspects: (1) pretrained models. (2) Unfrozen parts of each model during fine-tuning. (3) Resolution of models. (4) The initial learning rate of iterative training. (5) Spacial accuracy improvement of different settings.

	Settings	Pretrained models	Unfrozen parts	Resolution	Initial lr	Spatial acc(%)
RLDF	SDv2(1)	Stabilityai/stable-diffusion-2-1	Full weights	768	1e-6	28.50/18.75
	SDv2(LoRA)	Stabilityai/stable-diffusion-2-1	Low-Rank Adaptation	768	1e-6	25.25/18.75
	SDXL(LoRA)	Stabilityai/stable-diffusion-xl-base-1.0	Low-Rank Adaptation	1024	1e-5	31.25/27.00
	SDv2(2)	Stabilityai/stable-diffusion-2-1-base	Full weights	512	1e-5	**32.22/17.00**
RLHF	SDv2	Stabilityai/stable-diffusion-2-1	Full weights	768	1e-6	27.00/18.75
	SDv2(LoRA)	Stabilityai/stable-diffusion-2-1	Low-Rank Adaptation	768	1e-5	26.00/18.75

the training. If overfitting, only the base dataset will increase by 2000 prompts in each iteration.

A.5.2 Training Cost

In our IPR-RLDF method, when fine-tuning SDv2 with LoRA on four NVIDIA Tesla V100 GPUs, the training time for each iteration is approximately 15 minutes. While in the IPR-RLHF baseline, fine-tuning SDv2 with LoRA on four V100s costs around 60 minutes for each iteration.

A.6 Difference between Fine-tuning Settings

In different RLDF and RLHF fine-tuning settings, we employ the previously mentioned models and fine-tuning methods. Detailed differences among settings are presented in Table 13.

A.7 Detection Model Details

GLIPv2 (Zhang et al., 2022a), a grounded Vision-Language (VL) understanding model, integrates localization (Lin et al., 2014; Caesar et al., 2018) and VL understanding (Chen et al., 2015; Antol et al., 2015; Kiros et al., 2014) to establish grounded VL understanding. It achieves this by innovatively transforming localization tasks into a concentration of category names within VL understanding (Li et al., 2022). This innovative approach effectively resolves the conflicting output format requirements between localization and VL understanding, allowing them to help each other mutually. Consequently, it attains outstanding performance in

both localization and understanding tasks. Leveraging its outstanding zero-shot detection capability, we employ the pretrained GLIPv2 model as the detection component in our framework, yielding excellent results.

A.8 Human Preference Reward Model

We use ImageReward (Xu et al., 2024) as the reward model in the RLHF baseline. It is the first general-purpose text-to-image human preference reward model, trained with 137k expert comparisons. It proficiently encodes human preferences. To fine-tune diffusion models with ImageReward, we use Reward Feedback Learning (ReFL) (Xu et al., 2024), an algorithm that fine-tunes LDMs directly. It treats the scores from the reward model as human preference losses, which are then back-propagated to a randomly selected step in the denoising process.

A.9 Additional Visualizations

We provide qualitative samples from original models and models fine-tuned with various methods (RLDF, PR-RLDF, and IPR-RLDF). In comparison to the original models, RLDF, and PR-RLDF methods, the IPR-RLDF method notably demonstrates superior spatial accuracy. Figure 10 shows samples from SDv2 LoRA fine-tuning models; Figure 11 shows samples from SDv2 unfrozen fine-tuning models; Figure 12 shows additional samples from

Original SDv2 model	SDv2 IPR-RLDF	SDv2 IPR-RLHF
location ✘	location ✔ number ✔	location ✔ number ✔

(a) A horse to the left of a cat

Original SDv2 model	SDv2 IPR-RLDF	SDv2 IPR-RLHF
number ✘	location ✔ number ✔	location ✔ number ✔

(b) A suitcase above a keyboard

number ✘	location ✔ number ✔	location ✔ number ✔

(c) A toothbrush above a chair

location ✘	location ✔ number ✔	location ✔ number ✔

(d) A car above a suitcase

Figure 8: Qualitative examples of different models, generated by four distinct prompts. Models are as follows: (1) Using SDv2 as the original model. (2) IPR-RLDF (LoRA fine-tuning). (3) IPR-RLHF (LoRA fine-tuning). Both IPR-RLDF and IPR-RLHF exhibit commendable spatial accuracy.

SDXL LoRA fine-tuning models. Images in each figure are generated by ten distinct prompts. Also, there are qualitative examples of IPR-RLDF and IPR-RLDF in Figures 8 and 9. As is presented, both IPR-RLDF and IPR-RLHF exhibit commendable spatial accuracy.

B Additional Related Works

B.1 Detection Model

Many studies have concentrated on object detection tasks (Redmon et al., 2016; Lin et al., 2017; Chen et al., 2019; Carion et al., 2020), while the ability to detect certain rare objects (Gupta et al., 2019) still lacks proficiency. Recent work aims to address this issue, adopting novel approaches such as zero-shot (Bucher et al., 2019), few-shot (Li et al., 2021), or weakly-supervised (Bucher et al., 2019) methods. Notably, MEDTER (Kamath et al., 2021), GLIP (Li et al., 2022), and GLIPv2 (Zhang et al., 2022a) introduce an innovative perspective by transforming object detection tasks into grounded Vision-Language tasks. This integration of Vision-Language aspects into object detection yields remarkable effects in tasks like few-shot object detection. Thus, in our approach, we employ GLIPv2 as our chosen detection model within the pipeline.

Original SDv2 model	SDv2(LoRA) IPR-RLDF	SDv2(LoRA) IPR-RLHF
location ✘	location ✔ number ✔	location ✔ number ✔

(a) A mouse to the left of a couch

Original SDv2 model	SDv2(LoRA) IPR-RLDF	SDv2(LoRA) IPR-RLHF
number ✘	location ✔ number ✔	location ✔ number ✔

(b) A cell phone above an apple

number ✘	location ✔ number ✔	location ✔ number ✔

(c) A skateboard below a potted plant

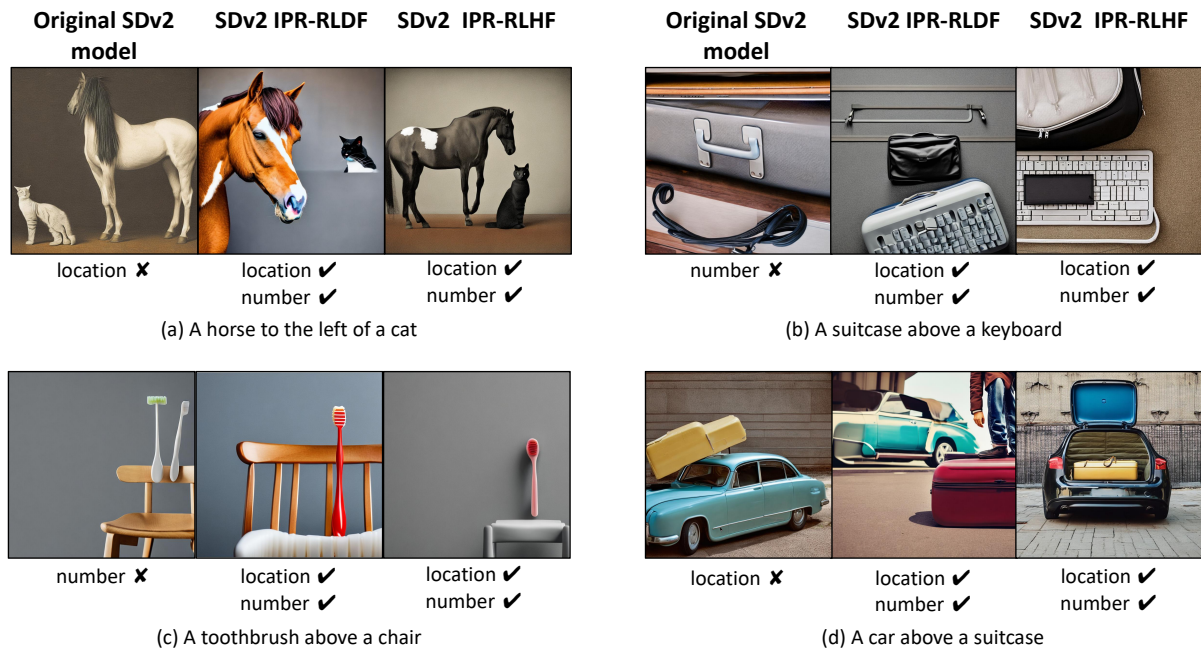| number ✘	location ✔ number ✔	location ✔ number ✔
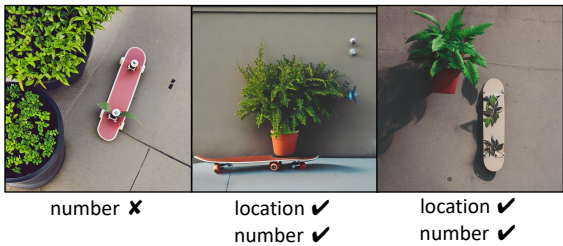
(d) A sports ball above a refrigerator

Figure 9: Qualitative examples of different models, generated by four distinct prompts. Models are as follows: (1) Using SDv2 as the original model. (2) IPR-RLDF (unfrozen fine-tuning). (3) IPR-RLHF (unfrozen fine-tuning). Both IPR-RLDF and IPR-RLHF exhibit commendable spatial accuracy.
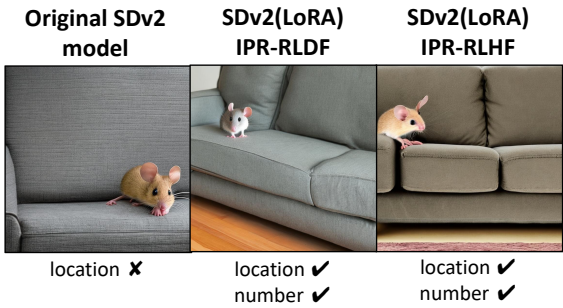
Figure 10: Samples from different models, generated by ten distinct prompts. Models are as follows: (1) Using SDv2 as the original model. (2) RLDF(LoRA fine-tuning). (3) PR-RLDF(LoRA fine-tuning). (4) IPR-RLDF(trained for two iters, LoRA fine-tuning). Compared to the original SDv2 model, RLDF, and PR-RLDF methods, our method showcases enhanced spatial awareness and accuracy in generating images of specified objects.

| Original SDv2 model | RLDF (unfrozen) | PR-RLDF (unfrozen) | IPR-RLDF (unfrozen) | Original SDv2 model | RLDF (unfrozen) | PR-RLDF (unfrozen) | IPR-RLDF (unfrozen) |

number ✗   location ✗   number ✓ location ✓   number ✓ location ✓

(a) A tv above a surfboard

location ✗   number ✗   number ✓ location ✓   number ✓ location ✓

(b) A hot dog to the right of an apple

number ✗   location ✗   number ✗   number ✓ location ✓

(c) A sink to the right of a tie

number ✗   number ✗   location ✗   number ✓ location ✓

(d) A carrot to the right of a tennis racket

location ✗   number ✗   number ✓ location ✓   number ✓ location ✓

(e) A bicycle to the left of an airplane

location ✗   number ✗   number ✗   number ✓ location ✓

(f) A cow to the right of an umbrella

number ✗   number ✗   number ✓ location ✓   number ✓ location ✓

(g) A skateboard below a potted plant

number ✗   location ✗   number ✗   number ✓ location ✓

(h) A bus above a banana

location ✗   number ✗   number ✓ location ✓   number ✓ location ✓

(i) A bear to the right of a toasterbear to the right of a toaster

number ✗   number ✗   number ✗ ✓   number ✓ location ✓
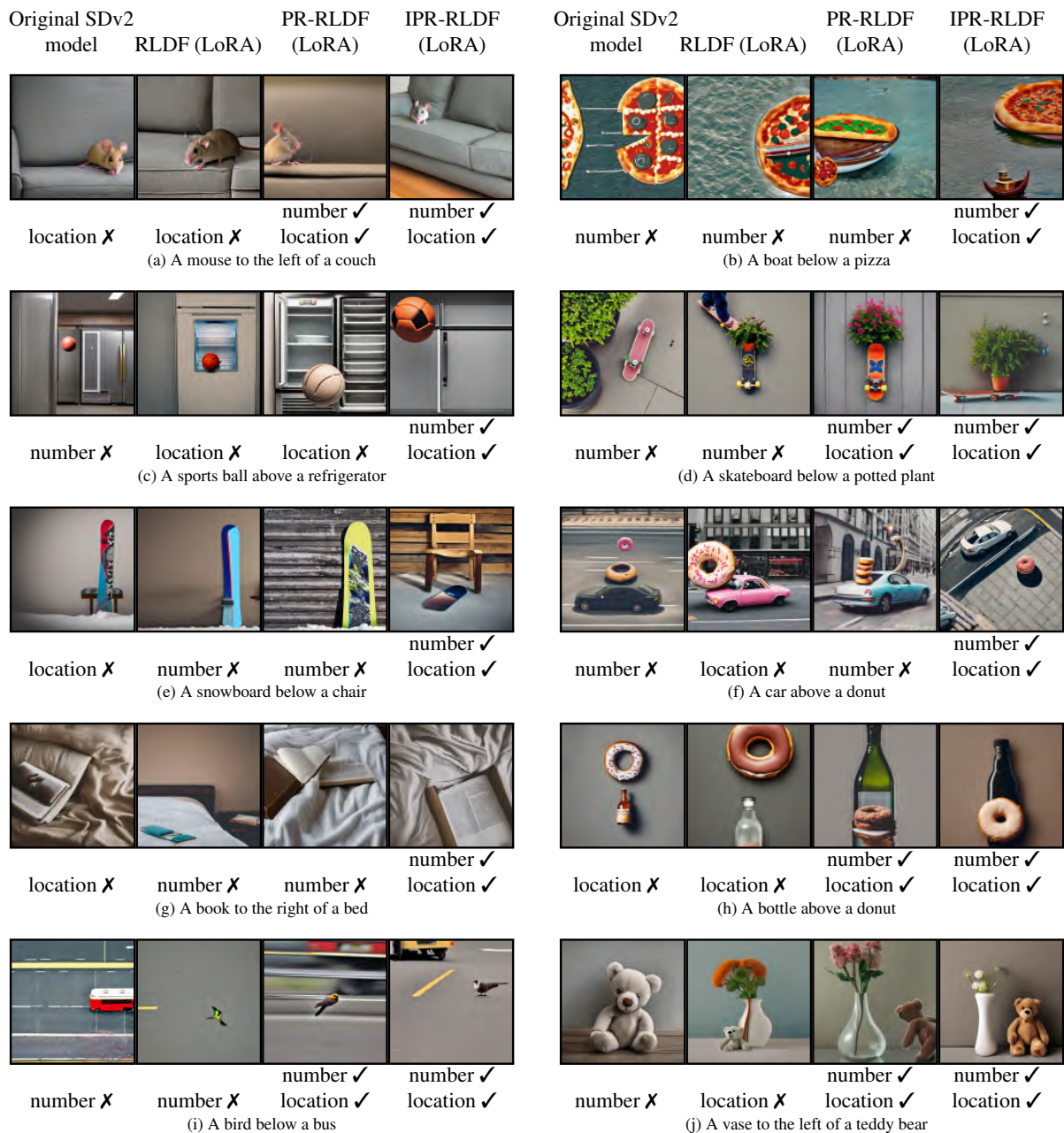
(j) A suitcase above a keyboard

Figure 11: Samples from different models, generated by ten distinct prompts. Models are as follows: (1) Using SDv2 as the original model. (2) RLDF(unfrozen fine-tuning). (3) PR-RLDF(unfrozen fine-tuning). (4) IPR-RLDF(trained for two iters, unfrozen fine-tuning). Compared to the original SDv2 model, RLDF, and PR-RLDF methods, our method showcases enhanced spatial awareness and accuracy in generating images of specified objects.
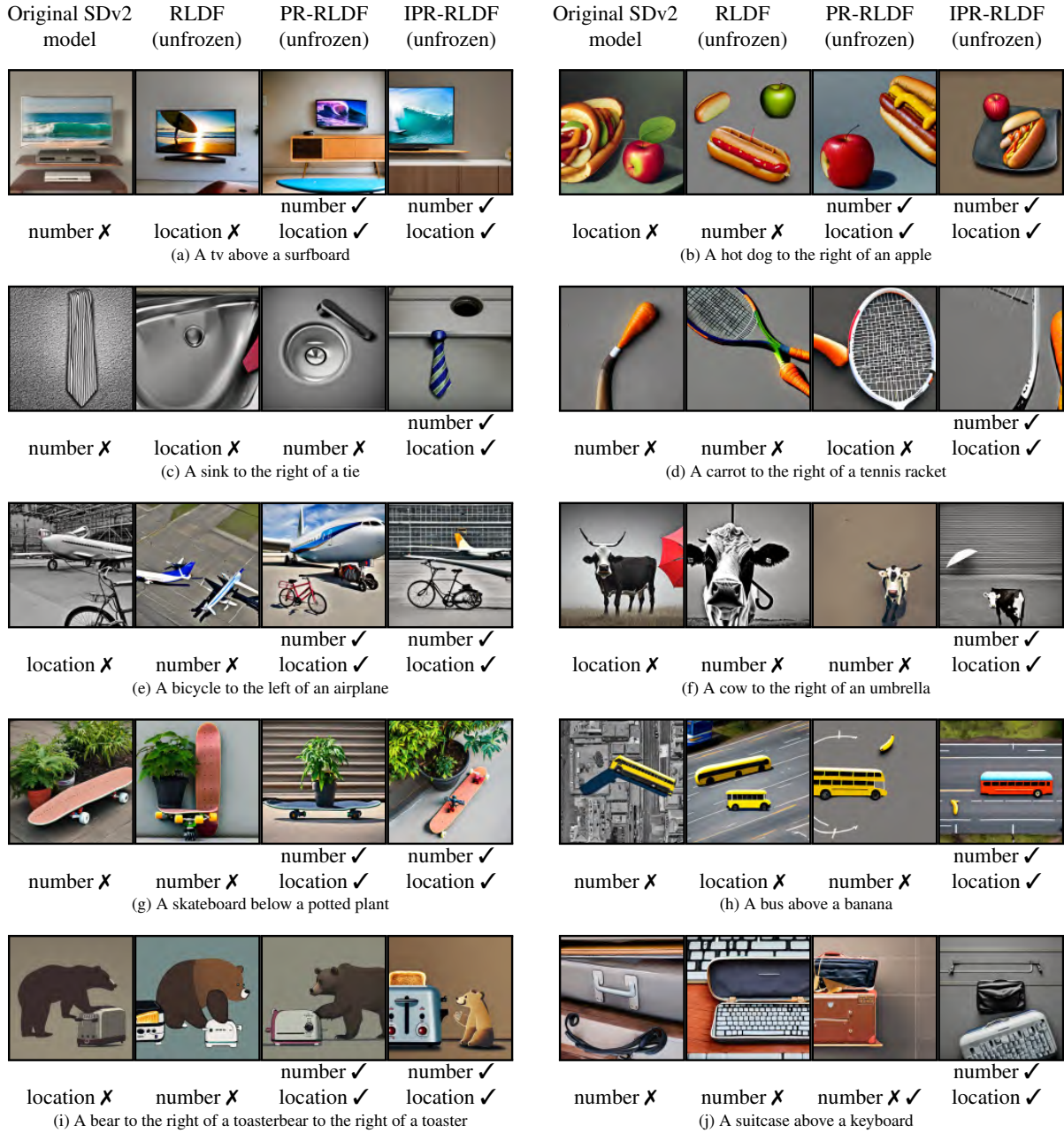
| | Original SDXL model | RLDF (LoRA) | PR-RLDF (LoRA) | IPR-RLDF (LoRA) | Original SDv2 model | RLDF (LoRA) | PR-RLDF (LoRA) | IPR-RLDF (LoRA) |

(a) A microwave to the right of a tennis racket

number ✗ · number ✗ · number ✗ · number ✓ location ✓

(b) A couch to the left of a laptop

number ✗ · number ✗ · number ✓ location ✓ · number ✓ location ✓

(c) A bicycle to the left of an airplane

location ✗ · location ✗ · location ✗ · number ✓ location ✓

(d) A airplane to the right of a clock

location ✗ · number ✗ · number ✓ location ✓ · number ✓ location ✓

(e) A keyboard to the left of a fire hydrant

location ✗ · number ✗ · number ✓ location ✓ · number ✓ location ✓

(f) A sandwich to the right of a horse

location ✗ · number ✗ · location ✗ · number ✓ location ✓

(g) A traffic light above a toilet

number ✗ · number ✗ · number ✓ location ✓ · number ✓ location ✓

(h) A traffic light to the right of a hot dog

number ✗ · number ✗ · number ✗ · number ✓ location ✓

(i) A bottle to the left of a hair drier

number ✗ · number ✗ · number ✗ · number ✓ location ✓

(j) A sandwich to the left of a boat

location ✗ · number ✗ · number ✗ · number ✓ location ✓

Figure 12: Samples from different models, generated by ten distinct prompts. Models are as follows: (1) Using SDXL as the original model. (2) RLDF(LoRA fine-tuning). (3) PR-RLDF(LoRA fine-tuning). (4) IPR-RLDF(trained for two iters, LoRA fine-tuning). Compared to the original SDXl model, RLDF, and PR-RLDF methods, our method showcases enhanced spatial awareness and accuracy in generating images of specified objects.