# PSR: Scaling Multi-Subject Personalized Image Generation with Pairwise Subject-Consistency Rewards

## Supplementary Material

Table 5. Quantitative Comparison on DreamBench

| Method | CLIP-T | CLIP-I | DINO |
|---|---|---|---|
| UNO [47] | 0.319 | 0.713 | 0.492 |
| OmniGen2 [46] | 0.329 | **0.720** | 0.506 |
| XVerse [2] | 0.327 | 0.715 | 0.479 |
| Qwen-Image-Edit-2509 [45] | **0.337** | 0.716 | 0.494 |
| **Ours (PSR)** | 0.335 | 0.717 | **0.529** |

## 7. More Implementation Details

We adopt LoRA [12] in both training stages, using a rank of 512 in the first stage and 64 in the second stage. During the GRPO training phase, we set the group size to 6, which yields stable convergence in practice. Since the aesthetic preference reward cannot be naturally normalized to the $[0, 1]$ range and is often larger than the other rewards in our experiments, we adjust the weighting scheme to stabilize multi-reward training. Specifically, we assign weights $w_1 = 0.4, w_2 = 0.4$ and $w_3 = 0.2$ to the respective reward components.

## 8. Experimental Results on DreamBench

To further validate the effectiveness of our approach, we also conduct evaluations on DreamBench [30, 47]. Specifically, we randomly sample 100 test cases from the DreamBench [30, 47] multi-ip subset and compare our method against recent state-of-the-art approaches, including UNO [47], OmniGen2 [46], XVerse [2], and Qwen-Image-Edit-2509 [45]. The evaluation follows the official testing protocol provided by UNO [47]. As shown in Table 5, our method achieves state-of-the-art performance on the DINO metric, surpassing the second-best method on DINO by a margin of 0.23. In addition, our model attains competitive results with Qwen-Image-Edit-2509 [45] on the CLIP-T metric. These findings collectively demonstrate the superiority of our approach. Although our method outperforms existing approaches on these metrics, it is important to acknowledge that there are still limitations in this evaluation method. For example, CLIP's semantic evaluation is not always accurate, and the global DINO score for assessing subject consistency is susceptible to interference from background information. We will discuss these issues in further sections.

In addition, Figure 7 presents the qualitative results of our model on the DreamBench [30]. Consistent with the observations on PSRBench, our method demonstrates superior subject consistency compared to other approaches. For instance, in the fourth row, existing methods fail to preserve the appearance of the robot toy in the generated images. In contrast, our method performs exceptionally well in maintaining both the dog's and the robot's appearances. Furthermore, as shown in the fifth row, when the input image features a candle that is not similar to a conventional candle, existing methods either disregard this 'non-standard' candle or generate one that is completely dissimilar to the original. Our method, on the other hand, handles such challenging cases effectively.

## 9. Quantitive Comparison with More Baseline

To further validate the effectiveness of our approach, we benchmark it against MS-Diffusion [42] and the state-of-the-art MOSAIC [33]. As shown in Table 6, our method consistently outperforms these baselines across all metrics. Specifically, SC, HPS, and SA denote Subject Consistency, Aesthetic Score, and Semantic Alignment, respectively.

## 10. User Study

Tab. 7 reports the user study results. Specifically, we randomly sample 100 test cases and ask five participants to rank images generated by these four methods for each case, where the ranked images are assigned scores of 1, 0.8, 0.6, and 0.4 from highest to lowest. The results indicate that our method achieves the best performance.

## 11. Additional Visualization Results and Analysis

As shown in Fig. 5-A, combining multiple rewards effectively mitigates reward hacking, whereas PSR alone tends to induce copy-and-paste behavior. Fig. 5-B presents cases where our model successfully captures inter-subject interactions, demonstrating that the results go beyond simple subject stacking and exhibit overall coherence. Fig. 5-C exemplifies a notable failure case of our method, specifically illustrating that identity preservation struggles to be maintained for small subjects.

## 12. More Infomation about Datasets and PSR-Bench

In the construction of the dataset, we use Qwen3-32B [52] to generate instructions for T2I and single subject personalization, with the prompt template for Qwen3-32B shown

Table 6. Quantitive Comparison with More Baseline Models

| Method | Attribute | Background | Action | Position | Comlex | Three | Four | Overal |
|---|---|---|---|---|---|---|---|---|
| MS-Diffusion (SC) [42] | 0.352 | 0.366 | 0.401 | 0.329 | 0.322 | 0.303 | 0.300 | 0.339 |
| MOSAIC(SC) [33] | 0.569 | 0.505 | 0.580 | 0.555 | 0.513 | 0.419 | 0.366 | 0.501 |
| **Ours (SC)** | **0.626** | **0.721** | **0.741** | **0.727** | **0.713** | **0.615** | **0.571** | **0.673** |
| MS-Diffusion (HPS) [42] | 0.659 | 0.930 | 0.704 | 0.865 | 0.774 | 0.977 | 0.974 | 0.840 |
| MOSAIC (HPS) [33] | 0.922 | 1.106 | 0.933 | 1.116 | 0.964 | 1.107 | 0.997 | 1.021 |
| **Ours (HPS)** | **1.040** | **1.150** | **0.881** | **1.200** | **1.050** | **1.260** | **1.290** | **1.124** |
| MS-Diffusion (SA) [42] | 0.587 | 0.849 | 0.479 | 0.254 | 0.706 | 0.692 | 0.684 | 0.607 |
| MOSAIC (SA) [33] | 0.776 | 0.921 | 0.693 | 0.274 | **0.818** | 0.827 | 0.786 | 0.728 |
| **Ours (SA)** | **0.908** | **0.926** | **0.739** | **0.468** | 0.692 | **0.884** | **0.866** | **0.783** |

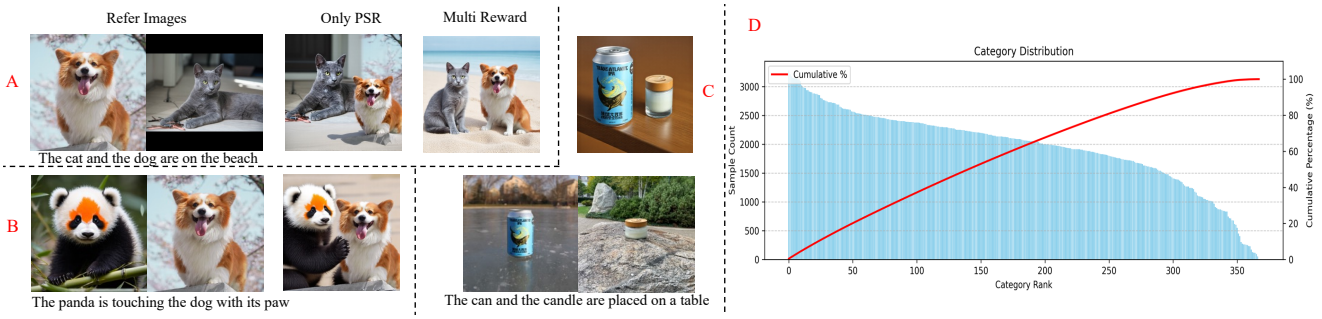


Figure 5. (A) Hacking hacking. (B) Interaction results. (C) Failure case. (D) Data statistics.

Table 7. User Study

| Method | SC | SA | HPS |
|---|---|---|---|
| Omnigen2 [46] | 0.50 | 0.58 | 0.74 |
| XVerse [2] | 0.64 | 0.66 | 0.60 |
| Qwen-Image-Edit-2509 [45] | 0.74 | 0.76 | 0.62 |
| **Ours (PSR)** | **0.92** | **0.80** | **0.82** |

in Figure 8. We utilize the Text-to-Image model FLUX.1-schnell [39] to generate $I_{out}$. During the construction process, since Qwen-Image [45] performs better when generating images containing four subjects, we use Qwen-Image [45] for generating data involving four subjects. Fig. 5-D visualizes the number of samples per category in our dataset. The results show that our dataset exhibits a diverse and relatively balanced category distribution.

Our benchmark consists of seven subsets, as illustrated in Figure 6, with each subset containing 50 evaluation samples. Our input images are generated by Qwen-Image [45] and are manually filtered to select images with distinctive subject appearances, ensuring the high quality of the benchmark. The basic information for each subset is as follows:

- **Action:** Each input image in this subset features a different animal, with the prompt explicitly requiring each animal to perform a distinct action, such as "sit", "run", and so on.

- **Attribute:** This subset also consists of animal images, with the prompt specifying that each animal must possess a unique attribute, such as "wearing a hat", "wearing a crown", etc. This is considered a more challenging subset.

- **Background:** The input images in this subset can include both objects and animals, with the prompt requiring the images to feature a specific background.

- **Position:** The inputs in this subset are arbitrary, with the prompt instructing that the subjects be placed in fixed orientations within the image, such as "on the right", "on the left", etc.

- **Complex:** The prompts in this subset are more intricate, potentially including detailed backgrounds, specific animal actions, or subjects placed in precise locations within the image.

- **Three:** This subset consists of three input subjects, which must be placed within a specific background.

- **Four:** This subset contains four input subjects, each requiring a specific background.

For evaluation, we consider three complementary dimensions. First, for subject consistency, we detect and crop the corresponding subjects in both the input references and the generated images, compute pairwise similarities, and average the results. Because subject consistency is defined as

an averaged similarity score, its value naturally falls within the range of 0 to 1. For semantic alignment, we employ Qwen2.5-VL-32B-Instruct [1] to assess whether the generated image semantically aligns with the given prompt. The MLLM outputs a score between 0 and 10, where 10 indicates perfect alignment and 0 indicates complete misalignment. We then normalize this score to the range of 0–1. For each specific subset, the corresponding prompt template used for evaluation by the MLLM is shown in Figure 9. For aesthetic preference, we evaluate image quality using HPSv3 [22], an uncertainty-aware ranking model that provides a non-normalized aesthetic score. To make it compatible with other reward components during training, we normalize the score by dividing it by 10.

## 13. Data Cleaning

To obtain high-quality paired data, we further design a stringent data-filtering pipeline. Specifically, we filter samples based on both subject consistency and semantic alignment. Similar to our evaluation protocol, we compute a paired DINO [24] score to assess the consistency between the input subjects and the generated outputs, and we employ Qwen2.5-VL [1] to evaluate semantic alignment with the prompt. After this filtering stage, we obtain a curated dataset of approximately 350K high-quality multi-subject personalization samples.

## 14. Metric Comparison

In this section, we compare several metrics used to evaluate subject consistency in previous works.

UNO [47] and MIP-Adapter [14] assess subject consistency by directly comparing the DINO [24] score of the input image with that of the output image. However, the output image often involves other subjects and background information, which can significantly affect the evaluation of subject consistency. As shown in Figure 10, the results of the traditional Global DINO Score evaluation are presented. It is evident that the bag in 'Output Image 1' is more similar to the bag in the 'Ref Image'. However, the Global DINO Score assigns a higher score to 'Output Image 2'. This is because the background of the reference image and 'Output Image 2' are more similar, thus distorting the subject consistency evaluation. In contrast, our method first grounds the subject and then performs cropping, effectively eliminating the irrelevant background information that could interfere with the evaluation. As demonstrated in Figure 10, our evaluation metric accurately identifies which image is more similar to the input subject, highlighting the precision of our approach.

Some existing methods use multimodal large language models to evaluate subject consistency. Therefore, we also compare our method with MLLM-based evaluations. We utilize the powerful multimodal large language model GPT-5 and employ the evaluation prompt proposed in OmniContext [46]. As shown in Figure 11, even the most advanced GPT-5 model fails to be highly sensitive to changes in the subject's appearance during subject consistency evaluation, often producing hallucinations.

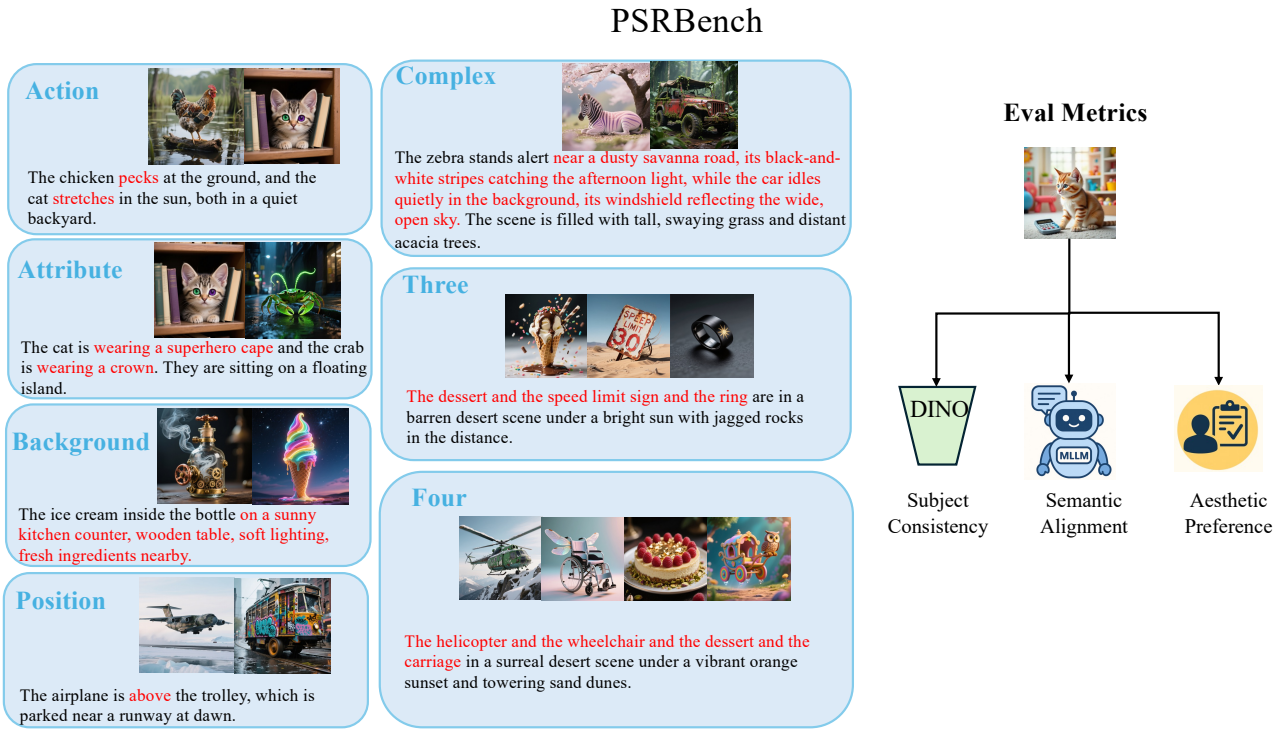


Figure 6. Overview of PSRBench, with a case from each subset shown on the left and the three evaluation dimensions for each subset displayed on the right.

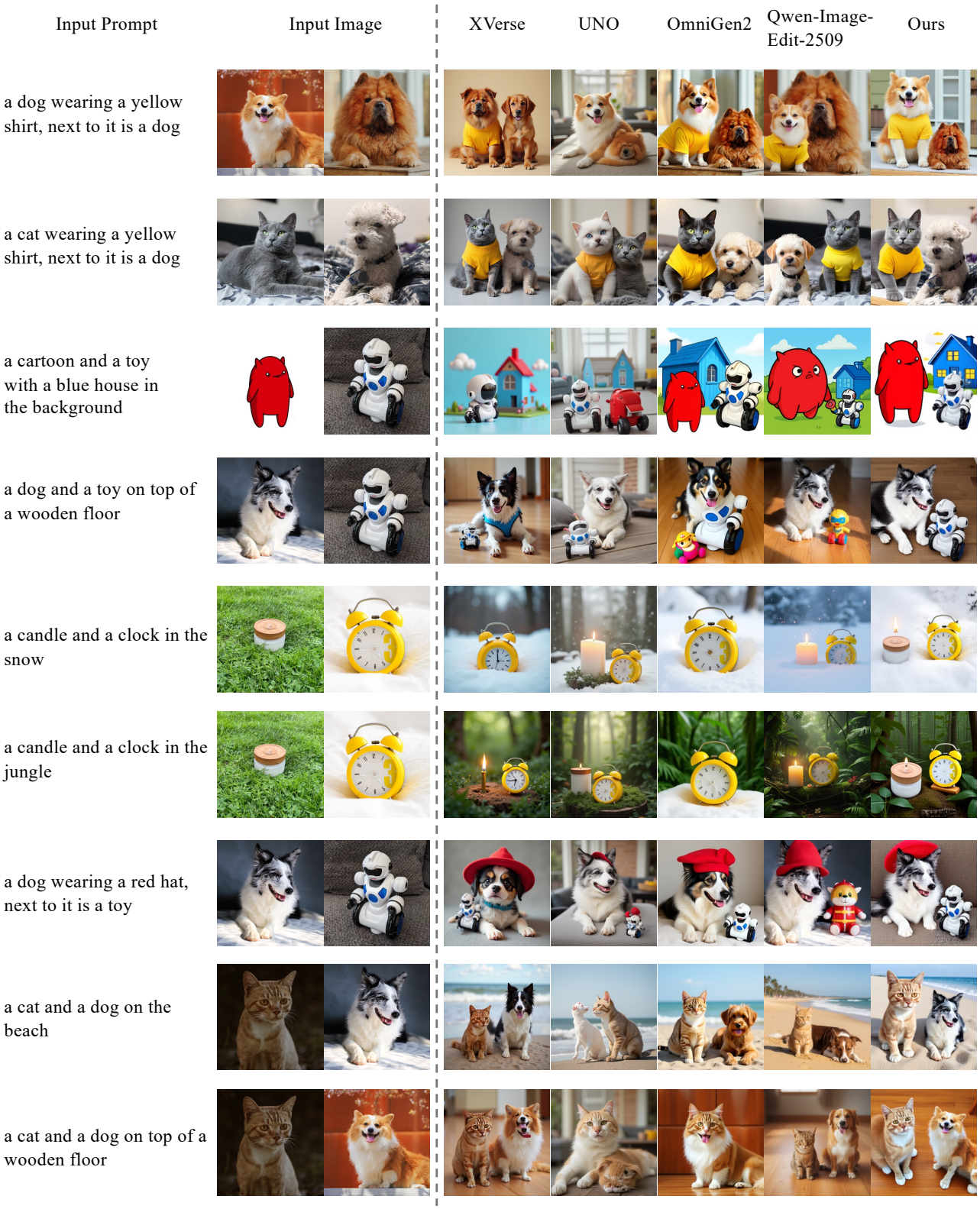


Figure 7. Qualitative Analysis on DreamBench

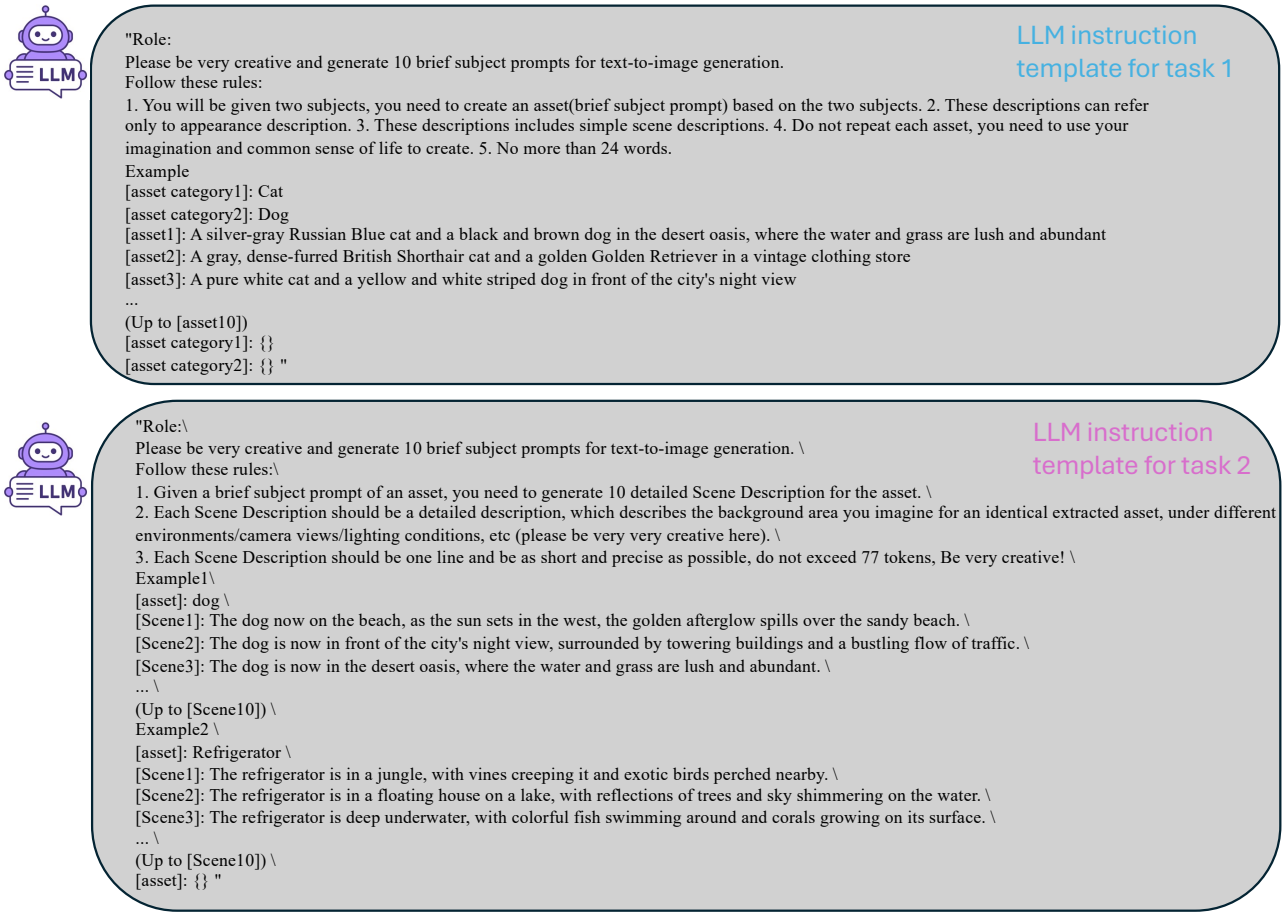


Figure 8. Instruction template used for providing to Qwen3 to construct the dataset.

## Action

Your role is to evaluate whether the action of the objects in the image match the given action descriptions.
Role1: The score range is 0-10, where 10 indicates a perfect match between the action in the image and the action description, with higher scores indicating a better match, and 0 indicating no match at all.
Rule2: You only need to focus on the degree of matching between the action in the image an the action description.
Rule3: The criteria for judging action should be relatively stringent.
Rule4: Please first provide a detailed analysis of the evaluation process, including the criteria for judging action alignment, then give a final Score from 0 to 10. The output text must follow the format [Thought]: … [Score]: …
[action description]: {}
[Thought]:
[Score]:

## Attribute

Your role is to evaluate whether the object attribute binding and attribute description in the image are consistent.
Role1: The score range is 0-10, where 10 indicates that all attributes of each object are completely consistent with the attribute description. A lower score indicates low consistency.
Rule2: You only need to focus on the degree of matching between the attributes of subjects of image and the attribute description.
Rule3: The criteria for judging attributes should be relatively stringent.
Rule4: Please first provide a detailed analysis of the evaluation process, including the criteria for judging attribute alignment, then give a final Score from 0 to 10. The output text must follow the format [Thought]: … [Score]: …
[attribute description]: {}
[Thought]:
[Score]:

## Background

Your role is to evaluate whether the background in the image and background description are consistent.
Role1: The score range is 0-10, where 10 indicates a perfect match between the background in the image and the background description, with higher scores indicating a better match, and 0 indicating no match at all.
Rule2: You only need to focus on the degree of matching between the background in the image and the background description.
Rule3: The criteria for judging attributes should be relatively stringent.
Rule4: Please first provide a detailed analysis of the evaluation process, including the criteria for judging background alignment, then give a final Score from 0 to 10. The output text must follow the format [Thought]: … [Score]: …
[background description]: {}
[Thought]:
[Score]:

## Complex/Three/Four

Your role is to evaluate whether the content in the image match the description in the prompt.
Role1: The score range is 0-10, where 10 indicates that the content in the image is completely consistent with the description, and lower scores indicate low consistency.
Rule2: You need to evaluate the alignment of all content in the image and the prompt.
Rule3: The criteria for judging alignment should be relatively stringent.
Rule4: Please first provide a detailed analysis of the evaluation process, including the criteria for judging alignment, then give a final Score from 0 to 10. The output text must follow the format [Thought]: … [Score]: …
[description]: {}
[Thought]:
[Score]:

Figure 9. Instruction template used for providing to Qwen2.5-VL to evaluate the semantic alignment scores of different subsets.

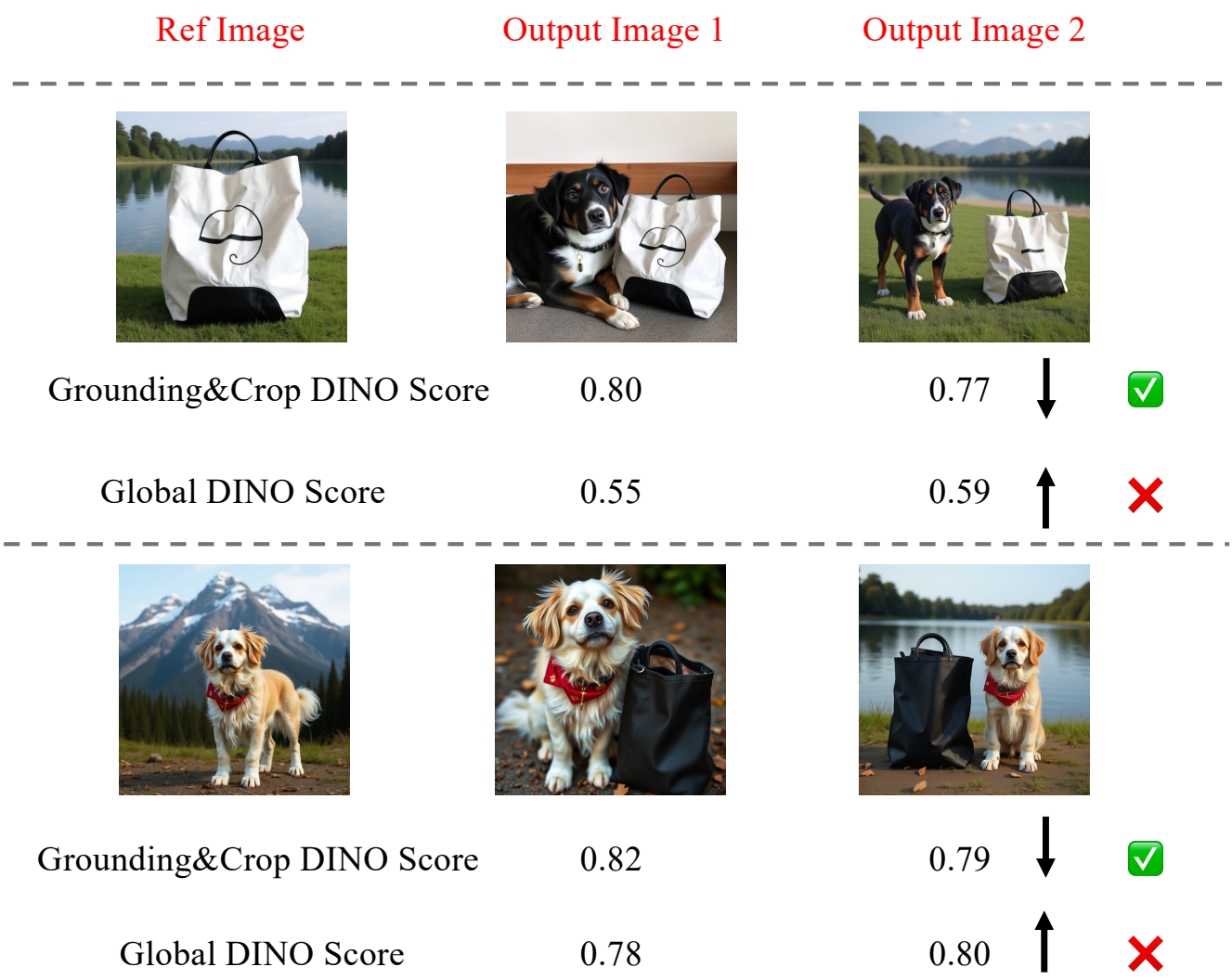


Figure 10. Comparison of different metrics.

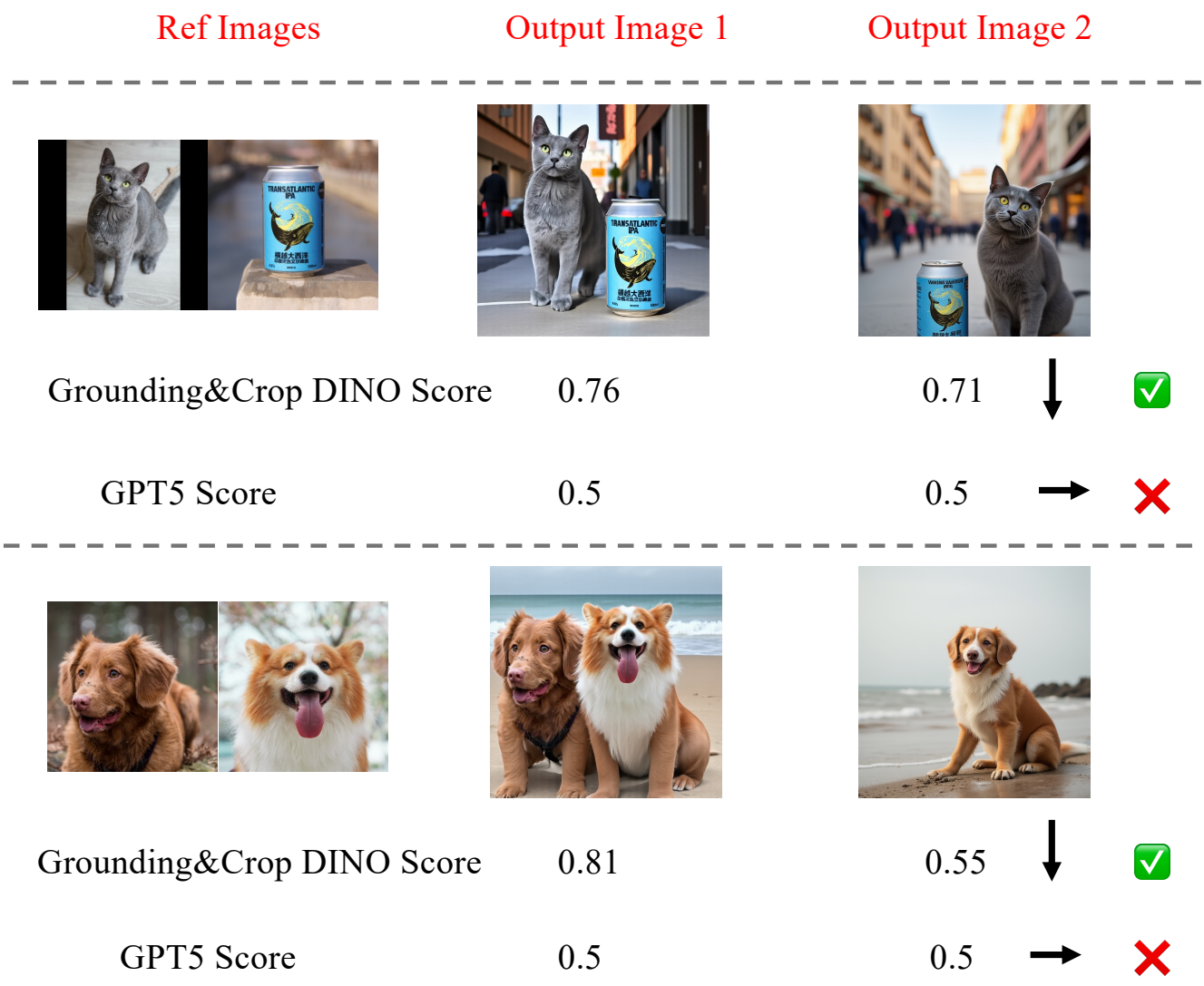


Figure 11. Comparison of different metrics.

## References


[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 4, 6, 3

[2] Bowen Chen, Mengyi Zhao, Haomiao Sun, Li Chen, Xu Wang, Kang Du, and Xinglong Wu. Xverse: Consistent multi-subject control of identity and semantic attributes via dit modulation. *arXiv preprint arXiv:2506.21416*, 2025. 2, 4, 6, 7, 1

[3] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023. 3

[4] Chengqi Duan, Rongyao Fang, Yuqing Wang, Kun Wang, Linjiang Huang, Xingyu Zeng, Hongsheng Li, and Xihui Liu. Got-r1: Unleashing reasoning capability of mllm for visual generation with reinforcement learning. *arXiv preprint arXiv:2505.17022*, 2025. 3

[5] Xie Fan, Zeng Dan, Shen Qiaomu, and Tang Bo. A comprehensive survey on text-to-video generation. *Chinese Journal of Electronics*, 34(4):1009–1036, 2025. 1

[6] Ying Fan and Kangwook Lee. Optimizing ddpm sampling with shortcut fine-tuning. *arXiv preprint arXiv:2301.13362*, 2023. 3

[7] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:79858–79885, 2023. 3

[8] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[9] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 3

[10] Zinan Guo, Yanze Wu, Chen Zhuowei, Peng Zhang, Qian He, et al. Pulid: Pure and lightning id customization via contrastive alignment. *Advances in neural information processing systems*, 37:36777–36804, 2024. 2

[11] Wenkang Han, Wang Lin, Yiyun Zhou, Qi Liu, Shulei Wang, Chang Yao, and Jingyuan Chen. Show and polish: reference-guided identity preservation in face video restoration. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10315–10324, 2025. 2

[12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 6, 1

[13] Mengqi Huang, Zhendong Mao, Mingcong Liu, Qian He, and Yongdong Zhang. Realcustom: Narrowing real text word for real-time open-domain text-to-image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7476–7485, 2024. 2

[14] Qihan Huang, Siming Fu, Jinlong Liu, Hao Jiang, Yipeng Yu, and Jie Song. Resolving multi-condition confusion for finetuning-free personalized image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3707–3714, 2025. 3

[15] Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. *arXiv preprint arXiv:2505.00703*, 2025. 3

[16] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2023. 2

[17] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. 1, 3, 4, 5, 6, 7

[18] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8640–8650, 2024. 1, 2

[19] Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, et al. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19401–19411, 2024. 3

[20] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025. 3, 5

[21] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 1, 4

[22] Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. *arXiv preprint arXiv:2508.03789*, 2025. 4, 6, 3

[23] Zhendong Mao, Mengqi Huang, Fei Ding, Mingcong Liu, Qian He, and Yongdong Zhang. Realcustom++: Representing images as real-word for real-time customization. *arXiv preprint arXiv:2408.09744*, 2024. 2

[24] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 4, 6, 3

[25] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini

Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 3

[26] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 1

[27] Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023. 3

[28] Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion alignment via reward gradients. *arXiv preprint arXiv:2407.08737*, 2024. 3

[29] 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. 4

[30] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 2, 4, 1

[31] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 3

[32] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 3

[33] Dong She, Siming Fu, Mushui Liu, Qiaoqiao Jin, Hualiang Wang, Mu Liu, and Jidong Jiang. Mosaic: Multi-subject personalized generation via correspondence-aware alignment and disentanglement. *arXiv preprint arXiv:2509.01977*, 2025. 1, 2

[34] Fei Shen and Jinhui Tang. Imagpose: A unified conditional framework for pose-guided person generation. *Advances in neural information processing systems*, 37:6246–6266, 2024. 2

[35] Fei Shen, Xin Jiang, Xin He, Hu Ye, Cong Wang, Xiaoyu Du, Zechao Li, and Jinhui Tang. Imagdressing-v1: Customizable virtual dressing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6795–6804, 2025.

[36] Fei Shen, Cong Wang, Junyao Gao, Qin Guo, Jisheng Dang, Jinhui Tang, and Tat-Seng Chua. Long-term talkingface generation via motion-prior conditional diffusion model. *arXiv preprint arXiv:2502.09533*, 2025. 2

[37] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 2024. 2, 3, 5

[38] Jiale Tao, Yanbing Zhang, Qixun Wang, Yiji Cheng, Haofan Wang, Xu Bai, Zhengguang Zhou, Ruihuang Li, Linqing Wang, Chunyu Wang, et al. Instantcharacter: Personalize any characters with a scalable diffusion transformer framework. *arXiv preprint arXiv:2504.12395*, 2025. 2

[39] XLabs AI team. x-flux, 2025. Accessed: 2025-02-07. 1, 4, 2

[40] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 1, 2

[41] Shulei Wang, Wang Lin, Hai Huang, Hanting Wang, Sihang Cai, WenKang Han, Tao Jin, Jingyuan Chen, Jiacheng Sun, Jieming Zhu, et al. Towards transformer-based aligned generation with self-coherence guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18455–18464, 2025. 1

[42] Xierui 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, 5, 1, 2

[43] Yibin Wang, Zhimin Li, Yuhang Zang, Yujie Zhou, Jiazi Bu, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Pref-grpo: Pairwise preference reward-based grpo for stable text-to-image reinforcement learning. *arXiv preprint arXiv:2508.20751*, 2025. 3

[44] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15943–15953, 2023. 2

[45] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 1, 3, 4, 6, 7, 2

[46] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025. 1, 3, 4, 6, 7, 2

[47] Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. *arXiv preprint arXiv:2504.02160*, 2025. 1, 2, 3, 4, 5, 6, 7

[48] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13294–13304, 2025. 1

[49] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023. 3

[50] Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. Visionreward: Fine-grained multi-dimensional

human preference learning for image and video generation. *arXiv preprint arXiv:2412.21059*, 2024.

[51] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025. 3, 5

[52] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 3, 4, 1

[53] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2