DraCo: Draft as CoT for Text-to-Image Preview and Rare Concept Generation

Supplementary Material

Overview

1. Related Work

Unified Multimodal Large Language Models Many prior works have focused on empowering a single model with both understanding and generation capabilities. Although most methods are built upon MLLMs [11, 12, 25], current mainstream approaches have diverged into different streams. One direction employs an external diffusion model for image generation [4, 22, 23, 29]. For example, OmniGen2 [27] uses a diffusion transformer to accept high-level conditions from the MLLM to produce images. Another direction adopts the image generation process into the auto-regressive paradigm [1, 26, 28]. [26] employs CLIP for vision understanding and VQVAE to discretize images for auto-regressive generation. Recently, some works have attempted to merge auto-regressive generation and diffusion into a single framework [3, 36]. Transfusion employs a single transformer model to generate text in an auto-regressive manner and images through iterative denoising. Extending this method further, Bagel employs a mixture of transformer architectures to separately process text tokens and VAE tokens of images.

Reasoning with Images With the introduction of OpenAI's o1 [15], reasoning with images has been widely explored for image understanding tasks [7, 10, 14, 31, 32]. These methods have attempted to incorporate more information from images into the reasoning process, including the use of external tools [21, 35], the reuse of image tokens from the vision encoder [2, 5], and the generation of auxiliary images [20, 30]. More recently, the field of visual generation has also adopted this paradigm [6, 8, 9, 17, 34]. Image-Gen-CoT [8] evaluates the effectiveness of direct preference optimization [19] for image generation and proposes PARM as a reward model for autoregressive generation in test-time scaling. Later, T2I-R1 [9] proposes generating a semantic-level CoT to analyze the given prompt and design the image. The semantic-level CoT is then fed back into the model to generate the image. Recently, several methods have proposed reflecting on generated images for improved image generation [18, 33, 37]. [37] utilizes a text-to-image diffusion model to generate the image, then a verifier produces reflection and a refined prompt instead of a correction method, and then the generator generates a new image. This method does not sufficiently connect the first generated image with the subsequently generated one, and therefore does not adequately address the difficulty of generating an image in one pass. Concurrently, [33] and [18] propose editing the first generated image instead of creating an entirely new one. However, there are several key differences between *DraCo* and these methods: (1) These methods follow a post-reflection strategy, rather than the pre-planning strategy adopted by *DraCo*. Therefore, they need to generate the image at the same resolution as the final output and then edit it, incurring much extra cost. (2) Additionally, these methods strictly utilize the image editing setting to edit the first generated image, such as restricting the background to remain exactly the same, thereby limiting the editing content. (3) In our pilot study, we find that the inherent editing capability is insufficient to handle various correction scenarios. Therefore, these methods that rely solely on editing capability cannot sufficiently meet the requirements for correcting errors.

2. More Details of *DraCo*-CFG

2.1. Original CFG of Bagel

Bagel [3] does not directly support *DraCo*, so no CFG can be directly applied to *DraCo*. However, the task that most closely resembles *DraCo* is "thinking before editing." In this scenario, Bagel takes as input: the image to edit (with both ViT features vit and VAE features vae) and an edit instruction $edit$. Bagel then generates a textual chain-of-thought $think$ to analyze how to perform the editing, followed by generating the edited image. Bagel employs three forms of multimodal context in CFG: $m(\phi, \phi, edit, \phi)$, $m(vit, vae, \phi, \phi)$, and the full condition $m(vit, vae, edit, think)$. The final output is computed through two sequential steps:

$$\hat{m}(vit, vae, edit, think)' = m(vit, vae, edit, think) + s_{text} \cdot (m(vit, vae, edit, think) - m(vit, vae, \phi, \phi)), \tag{1}$$

$$\hat{m}(vit, vae, edit, think) = \hat{m}(vit, vae, edit, think)' + s_{\text{img}} \cdot (\hat{m}(vit, vae, edit, think)' - m(\phi, \phi, edit, \phi)). \tag{2}$$

This formulation has two key issues:

1. **Incomplete decoupling of conditions.** The three input conditions, image, edit instruction, and text CoT, are not

fully decoupled in the two multimodal contexts used for CFG. Instead, two conditions are always dropped simultaneously. In $m(\phi, \phi, edit, \phi)$, both the image and CoT are absent, while in $m(vit, vae, \phi, \phi)$, both the edit instruction and CoT are absent. This means the CoT is emphasized twice across these two CFG types. Moreover, emphasizing two different conditions simultaneously may produce unexpected results.

2. **Sequential computation couples the scales.** Unlike *DraCo*-CFG, which computes CFG in one round, this approach uses two sequential rounds. This couples the scales s_{text} and s_{img}. Formally, let $M \triangleq m(vit, vae, edit, think)$, $M_{\text{text}} \triangleq m(vit, vae, \phi, \phi)$, and $M_{\text{img}} \triangleq m(\phi, \phi, edit, \phi)$ denote the latent vectors from the full model, the text-only model, and the image-only model, respectively. By substituting $\hat{m}(vit, vae, edit, think)'$ from Equation 1 into Equation 2, we obtain:

$$\begin{aligned}\hat{m} = &(1 + s_{text})(1 + s_{\text{img}})M \\ &- s_{text}(1 + s_{\text{img}})M_{text} - s_{\text{img}}M_{img}\end{aligned} \tag{3}$$

As shown, M_{text} is controlled by both s_{text} and s_{img}, meaning the condition from M_{text} may be overemphasized.

2.2. System Prompt

In *DraCo*-CFG, we adopt different system prompts for $m(\phi, \phi, \phi)$, $m(\phi, vit, \phi)$, and $m(p, vit, v)$ to ensure more emphasis on the specific condition. We omit the details of the system prompt in the full submission for brevity. In practice, we do not adopt system prompt for $m(\phi, \phi, \phi)$. For $m(\phi, vit, \phi)$, we want the model to fully construct the draft image, so the prompt is:

Draft-Only System Prompt

You should generate a larger image with the same content as the input image, with better details and clarity.

For full condition $m(p, vit, v)$, the system prompt is:

Full-Condition System Prompt

You should first generate a scratch image. Then you should analyze whether the scratch image is aligned with the prompt. If not, you should think about how to modify the scratch image to make it aligned with the prompt and then modify the scratch image. The analysis process is enclosed within `<think>` and `</think>` tags, i.e., `<think>` analysis process here modification process here (optional) `</think>` modified image here (optional).

3. More Dataset Details

We provide additional examples of our proposed *DraCo*-240K in Figure 1. As shown, the subset for general correction includes various correction scenarios such as object replacement and background modification. For instance manipulation, the focus is on handling objects of the same class. The training data includes either adding, removing, or changing attributes for one or multiple instances in the image. Regarding layout reorganization, we include examples that require swapping the positions of objects, or adding or removing an object at a specific position relative to existing objects. To guarantee the accuracy of the generated prompts, we conduct cross-validation for prompts in instance manipulation and layout reorganization. The reason is that prompts in instance manipulation contain numeracy information about objects, and prompts in layout reorganization contain spatial information, which is universally acknowledged that current MLLMs, even powerful ones like Qwen3-VL [24], still struggle to accurately capture. Therefore, we also use GroundingDINO [13] to detect all the objects present in the prompt and record their positions and numbers. We only retain samples where the detection results align with the prompt. Besides, the draft images in the dataset are roughly the same size as the final images. We downsample the draft images to a size of 384 $\times$ 384 during training. For the verification, we deliberately add a conclusion about what needs to be changed so that the model can better follow the correction instruction.

4. More Experiment Details

To facilitate Bagel to generate valid small resolution images as a valid draft, we first conduct a text-to-image fine-tuning stage before training for *DraCo*. We input various forms of prompts to Bagel to generate 1024 $\times$ 1024 images. Then we resize these images to 384 $\times$ 384 to train the model. We also adopt images generated from GPT-4o [16]. During training, half of the images in trained in 384 $\times$ 384, and half is trained in 1024 $\times$ 1024 to ensure the model's generation capability of both sizes of images. We finetune the model for 14K steps and adopt the EMA weight. We continue training on this weight for *DraCo*.

5. Limitations and Future Work

While DraCo demonstrates significant improvements in text-to-image generation through interleaved reasoning, several limitations remain. The application of this paradigm to other fields has not been extensively studied. Specifically, the low-resolution draft designed in *DraCo* cannot be directly or optimally used in other scenarios, such as videos, 3D assets, or scenes. Clearly, using a low-resolution draft still seems computationally expensive for video generation. The draft for different modalities requires capturing the ma-

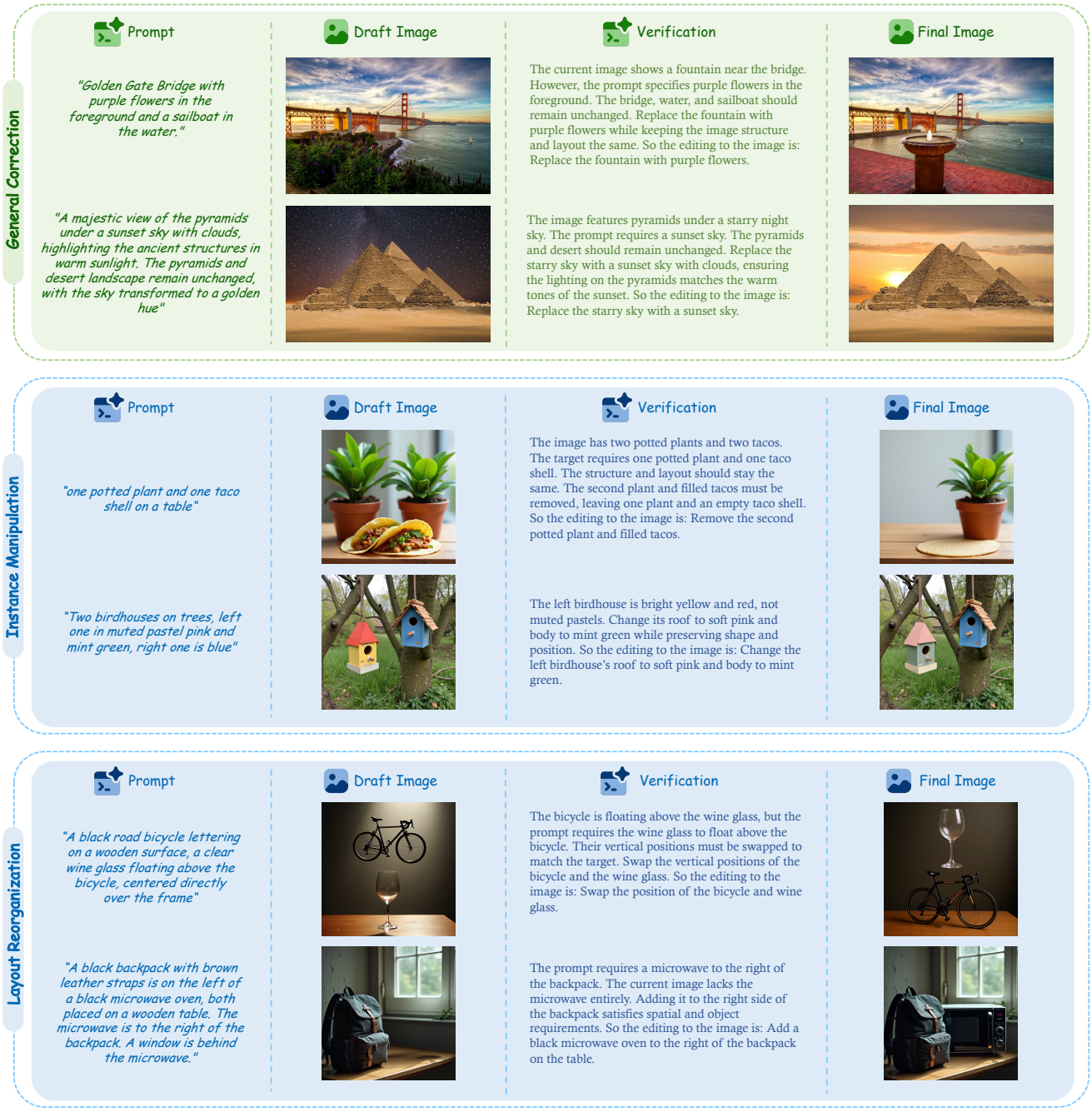

Figure 1. **More Examples of *DraCo*-240K.**

jor difficulty during generation, such as consistency in the video, and then constructing the form of the draft to give a second chance for the most challenging part. These are not clearly studied in this work. Moreover, the role of humans in improving data curation and the training loop is not well studied. Incorporating humans in this process could help better align the generation and correction methods.

6. Qualitative Examples

We show more qualitative examples and comparison in Fig. 2. In Fig. 2 (a), we showcase the draft image, final image, and its corresponding prompt. The examples include various correction scenarios, including: position correction, rare attribute generation, precise object editing, and numeracy correction. We also show more detailed exam-

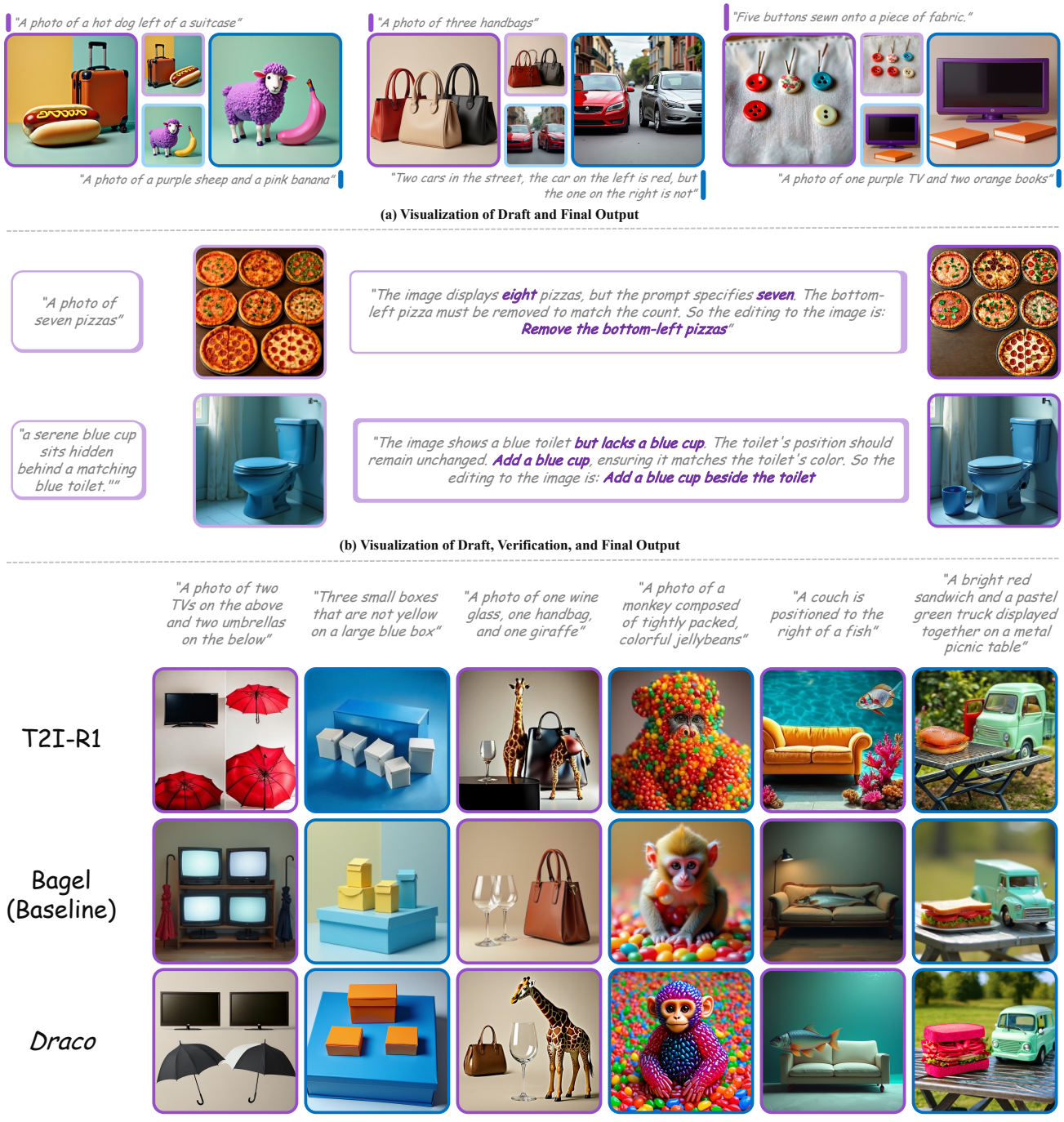

Figure 2. **More Qualitative Examples of *DraCo*.**

ples including the verification in Fig. 2 (b). During verification, the model first analyzes the inconsistency between the prompt and the image, then proposes a correction method, and also highlights what should not be changed. Ultimately, the model summarizes the correction in its final outcome. In Fig. 2 (c), we compare our method with the CoT-powered method, T2I-R1 [9], and our baseline model, Bagel [3]. As shown, T2I-R1 tends to generate artifacts, e.g., half of the giraffe and the monkey, or over-saturated images. While Bagel cannot accurately follow the prompt. In contrast,

DraCo produces satisfying results with both high quality and precise alignment of the prompt.

References

[1] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 1

[2] Xinyan Chen, Renrui Zhang, Dongzhi Jiang, Aojun Zhou, Shilin Yan, Weifeng Lin, and Hongsheng Li. Mint-cot: Enabling interleaved visual tokens in mathematical chain-of-thought reasoning. *arXiv preprint arXiv:2506.05331*, 2025. 1

[3] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 1, 4

[4] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023. 1

[5] Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. Interleaved-modal chain-of-thought. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19520–19529, 2025. 1

[6] Zeqi Gu, Markos Georgopoulos, Xiaoliang Dai, Marjan Ghazvininejad, Chu Wang, Felix Juefei-Xu, Kunpeng Li, Yujun Shi, Zecheng He, Zijian He, et al. Improving chain-of-thought efficiency for autoregressive image generation. *arXiv preprint arXiv:2510.05593*, 2025. 1

[7] Ziyu Guo, Ray Zhang, Hao Chen, Jialin Gao, Dongzhi Jiang, Jiaze Wang, and Pheng-Ann Heng. Sciverse: Unveiling the knowledge comprehension and visual reasoning of lmms on multi-modal scientific problems. *arXiv preprint arXiv:2503.10627*, 2025. 1

[8] Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. Can we generate images with cot? let's verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*, 2025. 1

[9] 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. 1, 4

[10] Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, et al. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*, 2025. 1

[11] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 1

[12] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 1

[13] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2

[14] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chun yue Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *ArXiv*, abs/2310.02255, 2023. 1

[15] OpenAI. Introducing openai o1, 2024., 2024. 1

[16] OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. 2

[17] Kaihang Pan, Wendong Bu, Yuruo Wu, Yang Wu, Kai Shen, Yunfei Li, Hang Zhao, Juncheng Li, Siliang Tang, and Yueting Zhuang. Focusdiff: Advancing fine-grained text-image alignment for autoregressive visual generation through rl. *arXiv preprint arXiv:2506.05501*, 2025. 1

[18] Luozheng Qin, Jia Gong, Yuqing Sun, Tianjiao Li, Mengping Yang, Xiaomeng Yang, Chao Qu, Zhiyu Tan, and Hao Li. Uni-cot: Towards unified chain-of-thought reasoning across text and vision. *arXiv preprint arXiv:2508.05606*, 2025. 1

[19] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023. 1

[20] Weikang Shi, Aldrich Yu, Rongyao Fang, Houxing Ren, Ke Wang, Aojun Zhou, Changyao Tian, Xinyu Fu, Yuxuan Hu, Zimu Lu, et al. Mathcanvas: Intrinsic visual chain-of-thought for multimodal mathematical reasoning. *arXiv preprint arXiv:2510.14958*, 2025. 1

[21] Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, et al. Openthinkimg: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*, 2025. 1

[22] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. *arXiv: 2312.13286*, 2023. 1

[23] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *arXiv: 2307.05222*, 2023. 1

[24] Qwen Team. Qwen3 technical report, 2025. 2

[25] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1

[26] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024. 1

[27] 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

[28] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024. 1

[29] 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

[30] Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. Visual planning: Let's think only with images. *arXiv preprint arXiv:2505.11409*, 2025. 1

[31] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023. 1

[32] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *ECCV 2024*, 2024. 1

[33] Xinchen Zhang, Xiaoying Zhang, Youbin Wu, Yanbin Cao, Renrui Zhang, Ruihang Chu, Ling Yang, and Yujiu Yang. Generative universal verifier as multimodal meta-reasoner. *arXiv preprint arXiv:2510.13804*, 2025. 1

[34] Yu Zhang, Yunqi Li, Yifan Yang, Rui Wang, Yuqing Yang, Dai Qi, Jianmin Bao, Dongdong Chen, Chong Luo, and Lili Qiu. Reasongen-r1: Cot for autoregressive image generation models through sft and rl. *arXiv preprint arXiv:2505.24875*, 2025. 1

[35] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025. 1

[36] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024. 1

[37] Le Zhuo, Liangbing Zhao, Sayak Paul, Yue Liao, Renrui Zhang, Yi Xin, Peng Gao, Mohamed Elhoseiny, and Hongsheng Li. From reflection to perfection: Scaling inference-time optimization for text-to-image diffusion models via reflection tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15329–15339, 2025. 1