MM-R1: Unleashing the Power of Unified Multimodal Large Language Models for Personalized Image Generation

Qian Liang[1*], Yujia Wu[1*], Kuncheng Li[1], Jiwei Wei[1†], Shiyuan He[1], Jinyu Guo[2], Ning Xie[1]

[1] School of Computer Science and Engineering, University of Electronic Science and Technology of China
[2] School of Information and Software Engineering, University of Electronic Science and Technology of China
{202422080310, 202322080314}@std.uestc.edu.cn, asyouwishasyousee@gmail.com, mathematic6@gmail.com, shiyuanhe.david@gmail.com, guojinyu@uestc.edu.cn, seanxiening@gmail.com

Abstract

Multimodal Large Language Models (MLLMs) with unified architectures excel across a wide range of vision-language tasks, yet aligning them with personalized image generation remains a significant challenge. Existing methods for MLLMs are frequently subject-specific, demanding a data-intensive fine-tuning process for every new subject, which limits their scalability. In this paper, we introduce **MM-R1**, a framework that integrates a cross-modal Chain-of-Thought (X-CoT) reasoning strategy to unlock the inherent potential of unified MLLMs for personalized image generation. Specifically, we structure personalization as an integrated visual reasoning and generation process: (1) grounding subject concepts by interpreting and understanding user-provided images and contextual cues, and (2) generating personalized images conditioned on both the extracted subject representations and user prompts. To further enhance the reasoning capability, we adopt Grouped Reward Proximal Policy Optimization (GRPO) to explicitly align the generation. Experiments demonstrate that MM-R1 unleashes the personalization capability of unified MLLMs to generate images with high subject fidelity and strong text alignment in a zero-shot manner.

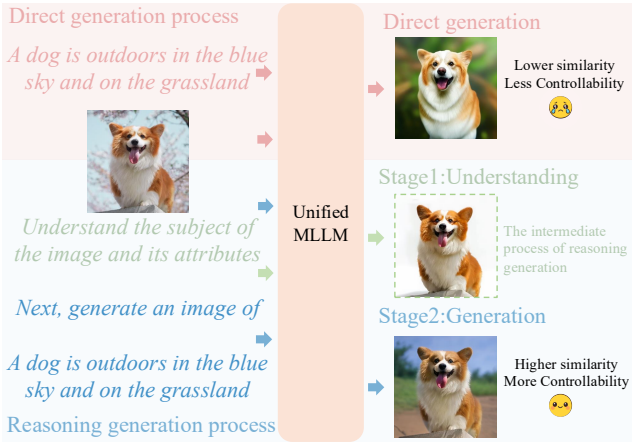

Figure 1: **Comparison between reasoning generation and ordinary generation.** Reasoning generation first understands the subject and its attributes in the image, and then injects these attributes into the generation process.

Introduction

With the rapid advancement of Multimodal Large Language Models (MLLMs), the capabilities of unified architectures for vision–language understanding and generation have increasingly converged and mutually reinforced one another. Recent advances (Team 2025; Liu et al. 2025; Xie, Yang, and Shou 2025; Zou et al. 2025) have shown that these architectures can effectively handle a variety of vision-language tasks, such as visual question answering and text-to-image generation, demonstrating strong multimodal understanding and generation capabilities. However, their potential in more fine-grained tasks, particularly personalized image synthesis, remains insufficiently explored.

Existing studies on personalizing unified MLLMs for image generation often involve training subject-specific tokens using the specific dataset. For instance, (Sun et al. 2025) performed subject-specific fine-tuning by creating trainable parameters for each subject. Yo'chameleon (Nguyen et al. 2025) introduces soft prompts as learnable tokens to represent user concepts during both understanding and generation. UniCTokens (An et al. 2025) proposed unified concept tokens refined through progressive staged optimization. Despite methodological differences, these approaches all rely on external token mechanisms and subject-specific optimization, which limit the scalability and generalization of personalization heavily.

Recently, some works (Sun et al. 2024a,b; Wu et al. 2025) established personalization as a two-fold task: it first requires the model to comprehend the user-specified subject, and then to generate images that are faithful to both this subject and the accompanying prompt. This principle, the tight coupling of understanding and generation, mirrors the core design philosophy of unified MLLMs, which are engineered to integrate the understanding and generating within a single, coherent framework (An et al. 2025). This inherent alignment suggests a promising yet underexplored path: achieving personalization by directly enhancing the model's intrinsic reasoning abilities, unlocking significant potential for future development and applications.

[*]These authors contributed equally.
[†]Corresponding author.

To validate this, we conducted a preliminary trial of a two-step personalization process as illustrated in Fig. 1. By prompting an MLLM (Liu et al. 2025) to first analyze the attributes of a subject in an image and then generate, it synthesizes significantly better results even without any additional training. This observation supports for our hypothesis and logically leads to our proposed solution: a unified framework that achieves personalization by deeply integrating reasoning.

In this paper, we propose **MM-R1**, a reasoning-enhanced framework for personalized image synthesis based on unified MLLMs. MM-R1 integrates a cross-modal reasoning mechanism with reward-guided optimization to better align the model's understanding and generation capabilities with user-defined content. Inspired by how Chain-of-Thought (CoT) improves reasoning in LLMs (Kojima et al. 2022; Wei et al. 2022a; Wang et al. 2023), we begin by introducing a fundamental cross-modal Chain-of-Thought (**X-CoT**) strategy, which explicitly defines the personalization as visual understanding and generation process. In the understanding stage, the model deconstructs the user-provided image and contextual cues to distill two key outputs: an explicit textual description of the subject, and an intermediate "focus image" that visually isolates the identified concept. Subsequently, in the generation stage, the model synthesizes the final scene by seamlessly weaving together its understanding of these extracted representations with the user-specified prompt, thereby formulating a coherent layout and rendering a personalized image faithful to both the subject's identity and the prompt's semantics.

To this, we design an X-CoT Data Engine. This automated pipeline generates high-quality, cross-modal Chain-of-Thought annotations that mirror our desired reasoning path. Thus, the model receives structured supervision, learning to explicitly link the initial subject concept with the final conditioned generation. This foundational training helps learn the core reasoning patterns necessary for enhancing subject fidelity and text alignment. Considering X-CoT is used as cold-start, we further employ Reinforcement Learning (RL). Specifically, we adopt the Group Relative Policy Optimization (GRPO) strategy (DeepSeek-AI 2025). GRPO enables fine-grained, reward-guided optimization by using multifaceted signals, such as image similarity for subject fidelity and text-image consistency for prompt adherence, to directly evaluate and enhance the generated outputs. In essence, our approach combines supervised pre-training with reinforcement learning fine-tuning, creating a principle-based, end-to-end framework that fully leverages the reasoning and generation capabilities of a unified MLLM for effective and controllable personalized image synthesis.

In summary, our main contributions are as follows:

- We propose MM-R1, a framework that equips unified MLLMs with personalized image synthesis capabilities through a cross-modal reasoning strategy (X-CoT) combining subject grounding and conditioned generation.
- We develop an X-CoT Data Engine, an automated pipeline that generates structured cross-modal reasoning annotations for models to complete personalized tasks, thereby achieving training without manual labeling.
- We extend GRPO to personalized image synthesis with multi-aspect rewards, improving subject fidelity, prompt alignment, and controllability.
- Our experiments demonstrate strong zero-shot personalization and superior performance over existing methods in both fidelity and controllability.

Related Work

Personalized Image Synthesis

Personalized image synthesis focuses on generating images of specific subjects in diverse contexts while maintaining subject fidelity and prompt alignment. Diffusion-based approaches have advanced this task substantially, with optimization-based methods such as DreamBooth (Ruiz et al. 2023), Custom Diffusion (Kumari et al. 2023), and Textual Inversion (Gal et al. 2023) fine-tuning model weights or token embeddings for each subject, achieving high fidelity but requiring costly per-subject optimization. To improve efficiency, tuning-free diffusion methods extract subject features and understand subject attributes with pre-trained encoders, enabling zero-shot or few-shot generation and inspiring more scalable personalization techniques (Ye et al. 2023; Labs et al. 2025). More recently, unified MLLMs have been explored for personalized image synthesis, including autoregressive models (Team 2025) that adopt a two-stage training strategy combining text embedding optimization and transformer fine-tuning, as well as methods such as Yo'Chameleon (Nguyen et al. 2025) and UniCTokens (An et al. 2025), which inject subject information into unified MLLMs via soft prompts or concept tokens to support both understanding and generation. Despite these advances, existing approaches still rely on separate per-subject representations, increasing complexity and limiting generalization. In contrast, our work focuses on personalized image synthesis within unified MLLMs, achieving efficient zero-shot personalization without disjoint subject-specific components or an additional encoder for understanding.

Reasoning in Multimodal Large Language Models

Chain-of-Thought (CoT) reasoning (Wei et al. 2022b; Zhang et al. 2023; Kojima et al. 2022) has emerged as a powerful mechanism to enhance the reasoning capabilities of LLMs by breaking complex problems into intermediate steps, improving both transparency and performance. Building on these advances, recent work has extended CoT to the multimodal setting, giving rise to MLLMs (Liu et al. 2023; Team et al. 2024; Thawakar et al. 2025) capable of processing both textual and visual inputs. For instance, T2I-R1 (Jiang et al. 2025) and GoT-R1 (Duan et al. 2025) respectively enable the model to generate images that better meet the user's requirements and those that better conform to the spatial positional relationship requirements through CoT. Multimodal CoT explores strategies such as grounding visual inputs into symbolic representations (Gao et al. 2025), generating visualizations to accompany reasoning (Zhao et al. 2025), or producing visual rationales through cropping and zooming (Wang et al. 2025). Visual Planning (Xu et al. 2025) further

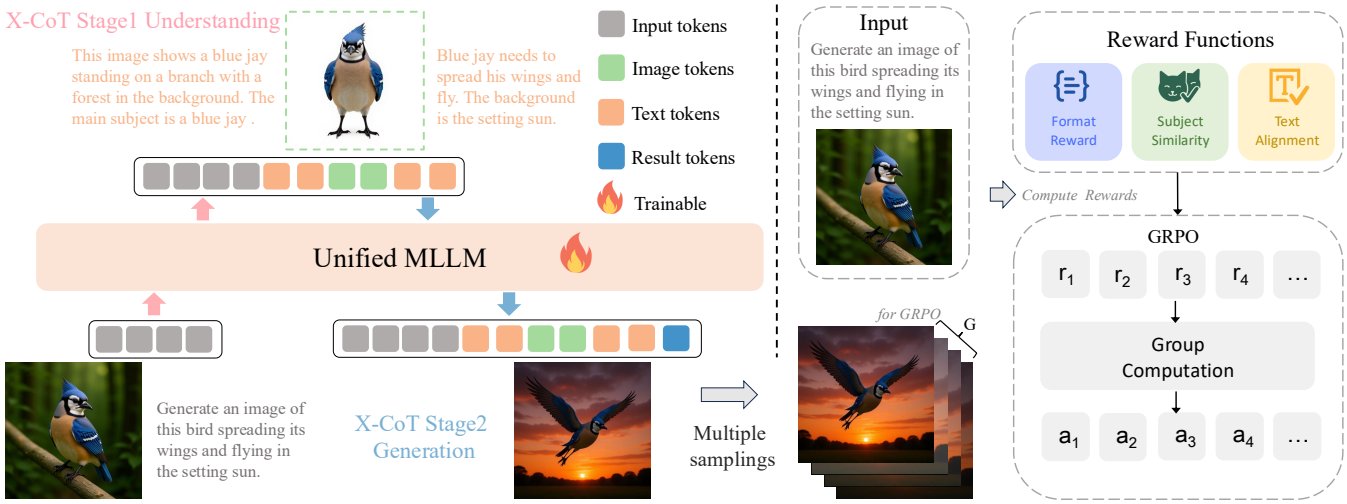

Figure 2: **Schematic diagram of the method proposed in this paper.** The left part is our X-CoT process. The model first understands the user input to obtain the subject image and then generates the image. After that, for each sample, multiple outputs are generated and trained together with the user input for GRPO. In the reinforcement learning process, three reward functions are used to calculate the rewards for different generated results, and finally these rewards are used to adjust the model.

investigates whether reasoning can emerge purely within the visual modality by structuring the reasoning process as a sequence of images without textual mediation. In contrast to the aforementioned approaches, our work explicitly incorporates cross-modal reasoning into unified MLLMs for personalized image synthesis, leveraging multi modalities in a unified and integrated manner that aligns with the pretraining paradigm and the foundational design principles of unified MLLMs.

Method

We present the MM-R1 framework, which enables unified MLLMs (e.g., Chameleon (Team 2025), Lumina-mGPT (Liu et al. 2025)) to perform zero-shot personalized image generation by leveraging a novel cross-modal Chain-of-Thought (X-CoT) reasoning strategy and reinforcement learning methods GRPO. MM-R1 extends the capabilities of these models for personalized vision-language tasks, allowing them to generate personalized images without the need for subject-specific fine-tuning.

Overview of MM-R1 Framework

As depicted in Fig. 2, our proposed MM-R1 framework is architected around a central cross-modal Chain-of-Thought (X-CoT) reasoning pipeline, which is fine-tuned using GRPO-based reinforcement learning. It enables zero-shot personalized image generation through a structured, two-step inference process.

First, in the understanding and planning phase, the X-CoT strategy processes multimodal inputs (including user images and text prompts) to generate a high-level "generation blueprint". This blueprint embodies a deep semantic understanding of the subject and a coherent plan for the final image generation. This initial step provides a robust, semantically-grounded foundation for the subsequent synthesis. Next, in the generation phase, the model translates this abstract blueprint into a sequence of detailed visual tokens. This ensures the final rendered image is not only visually coherent but also strictly adheres to both the subject's identity and the prompt's semantics.

Then, we use GRPO to teach the model to think better to improve the generation results. For each prompt, the model generates a batch of candidate images. These candidates then undergo a evaluation and ranking process based on a suite of predefined reward functions, which assess critical attributes like subject fidelity, text alignment, and format consistency. Ultimately, this process iteratively teaches the model to favor policies that produce images with higher subject consistency and stricter textual adherence.

X-CoT Data Engine

Training our two-stage reasoning model requires a specialized dataset that goes beyond simple image-text pairs, but thinking content of the X-CoT. To this end, we engineered an automated data construction pipeline. We chose the Subjects200K dataset (Tan et al. 2025) as our foundation due to its rich diversity of subject categories. We reconstructed this dataset by regenerating images and texts using the Flux-Kontext (Labs et al. 2025) and Qwen2.5-VL-7B-Instruct (Bai et al. 2025) model respectively. Fig. 3 shows an example creation process of our dataset.

Specifically, the Subjects200K dataset consists of samples from multiple categories, each of which contains two images with the same subject and the descriptive text of each image. We use Flux-Kontext to first extract the subject image from each sample and then use this image and the description text of the dataset to generate the two images with the same subject we need while retaining the text.

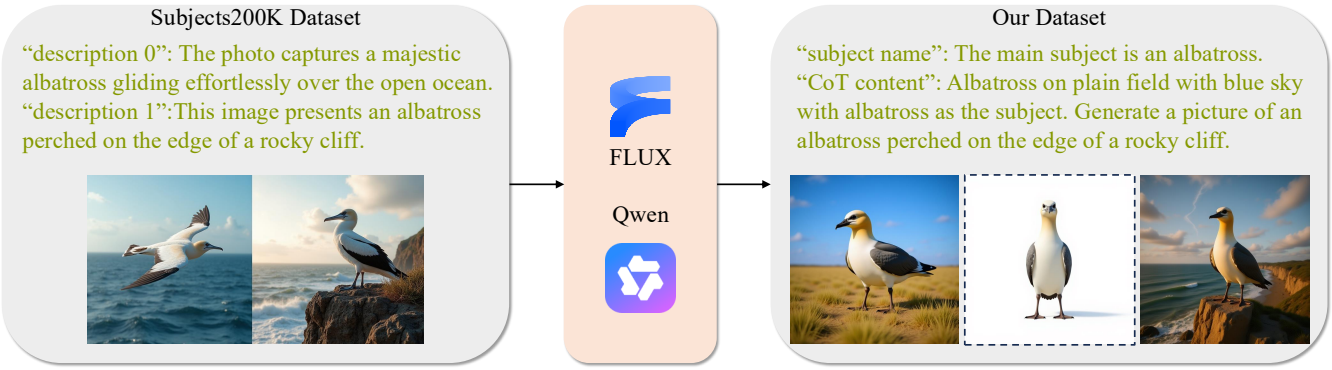

Figure 3: **An example of our data construction pipeline.** We reconstructed each sample in the Subjects200K dataset using Flux-Kontext and Qwen2.5-VL-7B-Instruct respectively, and obtained three images and thinking content.

Thinking contents consist of three parts. The first part is to understand the subject and its attributes of the reference image. We use the Qwen2.5-VL-7B-Instruct model to understand the reference image and generate the reference image title that meets the personalized generation task. The second part is the extraction of the subject by the model. Here, we directly use the subject image previously extracted. The third part of the text thinking contents is the integration and understanding of the reference image and the target text, which helps the model understand the paradigm of the target image. The Qwen2.5-VL-7B-Instruct model is also used to understand the content of the target image and obtain the corresponding generation prompt.

Cold-Start Training with X-CoT

By default, the model will directly integrate all the information and generate the result image when using unified MLLMs for personalized generation tasks. But this ability of the model is insufficient to achieve satisfactory generation results for personalized generation tasks (Wang et al. 2024). To address these issues, our approach is to have the model conduct reasoning before generating the results. The reasoning process includes understanding the content of the reference image and extracting the subject. To ensure the consistency of the final generated results with the reference images. During the reasoning process, we add a subject image to help the model better extract and understand the subject of the reference image, and at the same time the model uses the subject image rather than just user input to generate subsequent result images. The thinking contents include three parts, In the first part, the model conducts the necessary understanding of the reference image information. In the second part, for the personalized generation task, the model extracts the subject name and thinks obtaining the corresponding subject image. In the third part, based on the text prompt and all reference images, the model plans the content, layout, subject action, scene and other information of the resulting image. Finally, the model generates the result based on all the input information and the thinking content.

In order to enable the model to learn the ability to extract and utilize the subject information from the reference image through reasoning, we adopt a cold-start strategy based on supervised fine-tuning, dividing the generation process of the unified MLLMs into two stages. The first stage, the reasoning process, is used for the extraction and integration of information. The second stage utilizes this information to generate the resulting image. We utilized the constructed dataset to guide the model in generating reasoning contents in the expected format and semantics, strengthening the inherent information extraction and generation capabilities of the model to complete personalized generation tasks provides a stable and effective foundation for subsequent Reinforcement Learning.

Reinforcement Learning

The model has already acquired a certain ability to generate personalized images after supervised fine-tuning. To further enhance the reasoning and generation capabilities of the models and ensure the format of the reasoning contents and the quality of the generation results, we treat the personalized generation task as an RL problem and adopt the GRPO strategy (DeepSeek-AI 2025) to optimize the generated images. Unlike RL methods such as PPO (Schulman et al. 2017) that require value networks or DPO with paired preference data (Rafailov et al. 2023), GRPO does not require explicit reward functions or human-labeled preferences. It utilizes intra-group ranking feedback to optimize the generation strategy by encouraging better samples within each group, effectively reducing computational consumption and resource requirements without the need to train additional evaluation models.

Tailored specifically for personalized generation tasks, we introduce the multi-reward mechanism and design multiple rewards which is used to enhance the consistency between the generated results and reference informations to better guide policy learning. These reward signals reflect the generation effect through groupwise preference sorting in GRPO rather than being directly used in scalar form.

Format Reward. To ensure that the reasoning contents generated by the model have a parsable format, we introduce the Format Reward (Ouyang et al. 2025) R_{format}. This reward detects whether the result generated strictly adheres

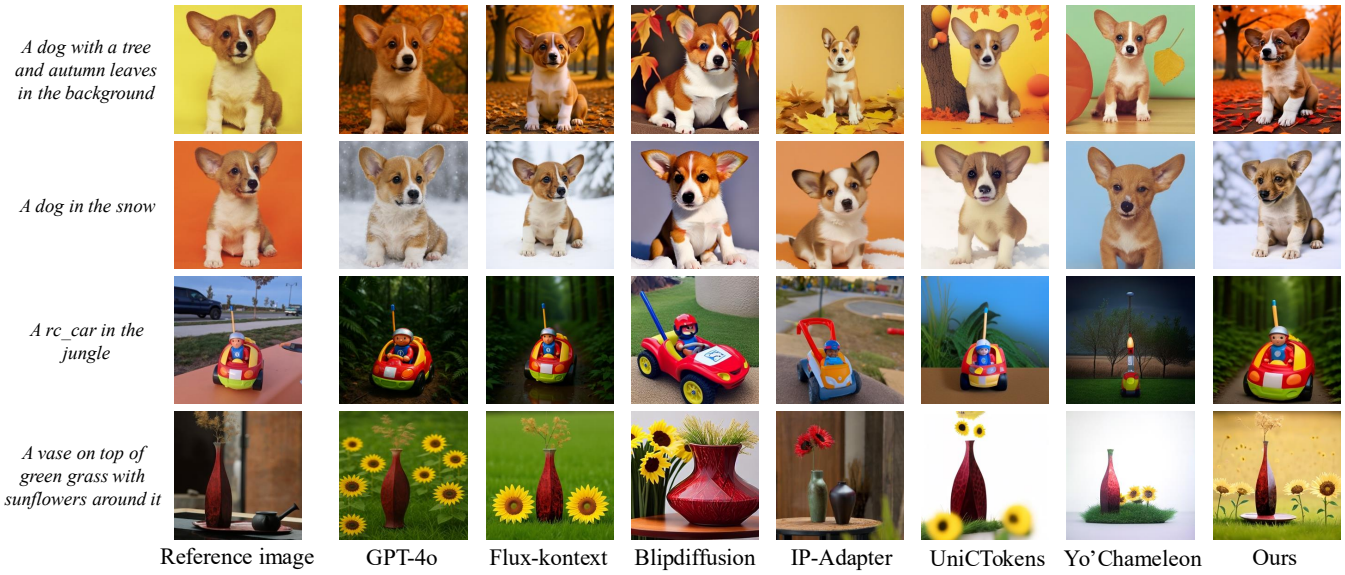

Figure 4: **Visualization of the qualitative results on DreamBench.** Our method can ensure extremely high subject similarity (the last column of the third row) and the aesthetic degree of the image (the last column of the fourth row). Meanwhile, our method can also generate diverse poses on the same subject.

to the predefined format through a regularized expression: {<text><|image|><text>}<|image|>. Here, the curly braces and the contents they contain represent the reasoning content, Specifically, the generated text thinking content and the corresponding image tokens that are used to help the model to understand the subject and to generate. The remaining contents are the generated result.

Text Alignment Reward. To ensure that the generated images are consistent with the input prompt in semantic content, we use the PickScore (Kirstain et al. 2023) metric as the standard to evaluate the quality of the generated results. It mimics the preference judgments of human users, directly optimizes the model to predict the images that users prefer, and scores them based on text conditions, making it more suitable for evaluating and optimizing the semantic consistency and user satisfaction of text-generated images. We calculate the score between the input text prompt and the generated result as R_t.

Subject Similarity Reward. To enable the model to learn the ability to generate personalized images from the original image to the target image, we use DreamSim (Fu et al. 2023) to calculate the similarity between the reference image and the result generated as R_i. It is a metric for evaluating visual similarity based on synthetic image triples and human judgment, focusing on high-level features such as foreground and semantic information, while also taking into account low-level features like color and layout.

Experiments

Implementation Details

We construct the X-CoT dataset using Qwen2.5-VL-7B-Instruct (Bai et al. 2025) and Flux-Kontext (Labs et al. 2025), which generate step-by-step cross-modal reasoning contents from user-provided images and prompts. These contents serve as semantic guidance for personalized image generation. We adopt Lumina-mGPT (Liu et al. 2025) as the unified backbone model and train it in two stages: 16K steps of cold-start training on the constructed X-CoT dataset, followed by 500 steps of GRPO-based reinforcement learning using subject fidelity and text consistency as reward signals. The training process is carried out on NVIDIA A6000 GPUs. All tests are conducted on NVIDIA A100 GPUs.

Experimental Settings

Evaluation Dataset. We evaluate our method on two benchmarks: DreamBench (Ruiz et al. 2023) and Kontext-Bench (Labs et al. 2025). DreamBench consists of 30 subjects, each with 4–6 reference images and 25 prompts covering diverse scenes and subject variations. The subject categories span pets, household objects, and artistic sculptures, providing a broad testbed for evaluating personalization and content controllability. For Kontext-Bench, we focus on the Character Reference subset and sample approximately 200 test cases, each containing one reference image and a prompt requiring identity-consistent image generation in novel contexts. For the results of the zero-shot methods on DreamBench, we generated each image sample in the same category and took the average.

Evaluation Metrics. Following standard protocols for personalized generation (Ruiz et al. 2023), we adopt DINO (Caron et al. 2021) and CLIP-I (Radford et al. 2021) to assess subject fidelity, and CLIP-T (Radford et al. 2021) to assess the consistency between generated images and input text prompts. These metrics embed images and text into a shared semantic space and compute similarity scores to reflect subject fidelity and text-image alignment. For each sub-

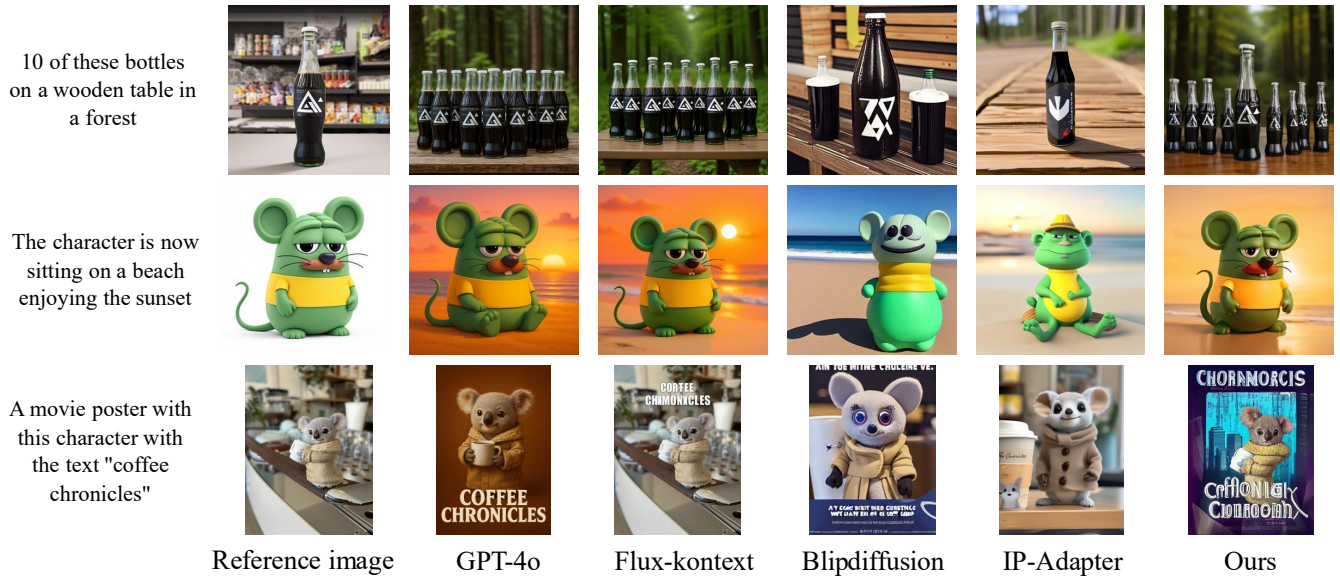

Figure 5: **Visualization of the qualitative results on Kontext-Bench.** Since this benchmark does not support subject-specific fine-tuning, we compare only with zero-shot methods on it.

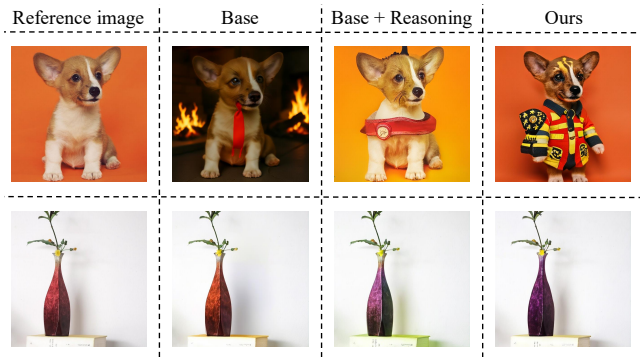

Figure 6: **Comparison of results in the ablation experiment.** *Base* means that two images are generated using the base model, and *Base + Reasoning* represents generating images using the Base model and reasoning generation strategy. Prompt is *"a dog in a fireman outfit."*.

ject, we average the scores across all prompts to obtain the final evaluation results. The specific models are ViT-B/32 and facebook/dinov2-large (Oquab et al. 2024) respectively.

Quantitative Evaluation

We evaluate the performance of MM-R1 on the DreamBench and Kontext-Bench, comparing it with state-of-the-art methods across both diffusion-based and unified MLLM-based approaches (Tab. 1). For diffusion models, we include DreamBooth (Ruiz et al. 2023), Blip-Diffusion (Li, Li, and Hoi 2023), IP-Adapter (Ye et al. 2023), and Flux-Kontext (Labs et al. 2025). For unified MLLMs, we compare against Yo'Chameleon (Nguyen et al. 2025), UniCTokens (An et al. 2025), and GPT-4o (Hurst et al. 2024). Among them, although Yo'Chameleon and UniCTokens can perform both understanding and generation tasks, here we only test the generation task and do not focus on their performance in the understanding task.

On DreamBench, MM-R1 achieves the best DINO score (0.786) and CLIP-T score (0.313), indicating strong subject fidelity and prompt alignment. On Kontext-Bench, it also ranks first in DINO (0.562) and performs competitively in CLIP-I (0.750) and CLIP-T (0.296), validating its effectiveness in single-image reference generation. Overall, MM-R1 establishes a new state-of-the-art among unified MLLMs and achieves performance that is competitive with the strongest diffusion-based approaches, all without relying on subject-specific fine-tuning.

Qualitative Results

Fig. 4 and 5 show the comparison of qualitative results between our method and some other methods. It can be seen that our method can maintain a high level of image fidelity and text controllability. As shown in our picture (the last row of the third column), the rc_car is highly similar not only overall but also in detail to the rc_car in the reference image. This indicates that, through X-CoT, the model can extract more attributes of the reference subject to generate consistent images. Furthermore, our method can generate more diverse images, such as the last column of the second row in Fig. 4. This is because our method performs information integration and subject extraction before generating images, which can help the model decouple subject from input and reduce the interference of irrelevant information.

Ablation Study

Ablation Study on MM-R1 Modules. To evaluate the contribution of each component in our framework, we con-

Type	Methods	DreamBench			Kontext-Bench			Zero-Shot
		DINO↑	CLIP-I↑	CLIP-T↑	DINO↑	CLIP-I↑	CLIP-T↑	
Diffusion Model	Dreambooth	0.631	0.803	0.305	-	-	-	No
	Blip-Diffusion	0.660	0.818	0.283	0.500	0.742	0.246	Yes
	IP-Adapter	0.671	0.836	0.291	0.516	**0.762**	0.249	Yes
	Flux-Kontext	0.682	**0.846**	0.310	0.554	0.750	**0.308**	Yes
Unified MLLM	Yo'Chameleon	0.542	0.795	0.225	-	-	-	No
	UniCTokens	0.599	0.782	0.304	-	-	-	No
	GPT-4o	0.722	0.803	0.274	0.501	0.725	0.303	Yes
	Ours	**0.786**	0.842	**0.313**	**0.562**	0.750	0.296	Yes

Table 1: **Quantitative comparison on DreamBench and Kontext-Bench.** Our MM-R1 achieves the best results across multiple metrics, notably scoring highest on DreamBench (DINO: 0.786, CLIP-T: 0.313) and Kontext-Bench (DINO: 0.562), while also supporting zero-shot capabilities. For each test sample we conduct four tests and take the average as the result.

Method	DreamBench			Kontext-Bench		
	DINO ↑	CLIP-I ↑	CLIP-T ↑	DINO ↑	CLIP-I ↑	CLIP-T ↑
R_f	0.729	0.768	0.298	0.512	0.712	0.284
$R_f + R_i$	0.782	**0.845**	0.284	0.559	0.749	0.269
$R_f + R_t$	0.737	0.801	0.311	0.525	0.714	**0.301**
Ours	**0.786**	0.842	**0.313**	**0.562**	**0.750**	0.296

Table 2: **Ablation results of different reward functions in GRPO.** Each line represents the result of GRPO reinforcement learning using different reward functions after the model has undergone X-CoT cold-start. Among them, R_f represents Format Reward, R_i represents Subject Similarity Reward, R_t represents Text Alignment Reward.

duct an ablation study as shown in Tab. 3. We begin by evaluating our reasoning generation strategy on the base model without any additional training. In this setting, the understanding stage guides the model to interpret the reference image and generate an intermediate image that isolates the subject and corresponding text thinking contents. This result is then provided to the generation stage, together with the user prompt, to guide the final personalized image generation. As shown in Tab. 3 and Fig. 6, this structurally guided inference pipeline already yields noticeable improvements, highlighting the effectiveness of our X-CoT formulation.

We further evaluate the impact of supervised X-CoT training without applying GRPO. This variant enhances the model's ability to extract subject-specific concepts and align them with user prompts, leading to clear improvements in fidelity and consistency. Conversely, we assess the effect of applying GRPO reinforcement learning without X-CoT supervision, which also results in notable gains, particularly in text alignment. When all components are combined, the model achieves the highest scores across all evaluation metrics, confirming the complementary strengths of structured reasoning and reward-based optimization.

Ablation Study on Reward Design. To evaluate the contribution of each reward component in GRPO, we conduct an ablation study as summarized in Tab. 2. The Format reward (R_f) is used by default to enforce structural constraints

Method	DreamBench			Kontext-Bench		
	DINO ↑	CLIP-I ↑	CLIP-T ↑	DINO ↑	CLIP-I ↑	CLIP-T ↑
Base	0.631	0.728	0.270	0.372	0.663	0.216
Reasoning	0.650	0.742	0.278	0.380	0.683	0.221
w/o GRPO	0.727	0.761	0.297	0.505	0.702	0.287
w/o X-CoT	0.732	0.801	0.288	0.525	0.714	0.282
Ours	**0.786**	**0.842**	**0.313**	**0.562**	**0.750**	**0.296**

Table 3: **Ablation results of our proposed components.** The quantitative results of each behavioral model after different steps of training. Among them, *Base* refers to the generation using only Lumina-mGPT, *Reasoning* indicates that the base model is used for reasoning understanding first and then for generation, *w/o GRPO* indicates that the model has been trained by our X-CoT, and *w/o X-CoT* indicates that the base model adjusted by the GRPO.

in the X-CoT process, and also brings slight improvements in output stability. Adding the Subject Similarity Reward (R_i) yields notable gains in identity preservation, as it directly encourages alignment between the generated and reference images. In contrast, the Text Alignment Reward (R_t) enhances semantic consistency with the input prompt. When both R_i and R_t are incorporated, the model achieves further improvements, indicating that these two rewards offer complementary supervision signals during training.

Conclusion

In this paper, we introduce MM-R1 which achieving personalization by directly enhancing the MLLMs' intrinsic reasoning abilities. Inspired by the relevance of understanding and generating tasks in MLLMs, we divide the personalization task into understanding the attributes of subject and image generation, and using this strategy to propose the X-CoT framework for image generation. To better unleash the model's capabilities of understanding and generation, we train the model that has been fine-tuned with the X-CoT strategy using GRPO. The experiments show that this strategy can effectively help the model better complete the personalized generation task.

Acknowledgments

This research is partially supported by the National Natural Science Foundation of China under Grant 62306067 and Grant 62220106008, the Sichuan Science and Technology Program under Grant 2024NSFSC1463, and the Guangdong Basic and Applied Basic Research Foundation under Grant 2025A1515010108. Additional support is provided by the Postdoctoral Fellowship Program (Grade C) of the China Postdoctoral Science Foundation (Grant No. GZC20251053) and Huawei Funding (Project ID: H04W241592).

References

An, R.; Yang, S.; Zhang, R.; Shen, Z.; Lu, M.; Dai, G.; Liang, H.; Guo, Z.; Yan, S.; Luo, Y.; Zou, B.; Yang, C.; and Zhang, W. 2025. UniCTokens: Boosting Personalized Understanding and Generation via Unified Concept Tokens. arXiv:2505.14671.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Duan, C.; Fang, R.; Wang, Y.; Wang, K.; Huang, L.; Zeng, X.; Li, H.; and Liu, X. 2025. Got-r1: Unleashing reasoning capability of mllm for visual generation with reinforcement learning.

Fu, S.; Tamir, N.; Sundaram, S.; Chai, L.; Zhang, R.; Dekel, T.; and Isola, P. 2023. DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data. *Advances in Neural Information Processing Systems*, 36: 50742–50768.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-or, D. 2023. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *The Eleventh International Conference on Learning Representations*.

Gao, J.; Li, Y.; Cao, Z.; and Li, W. 2025. Interleaved-modal chain-of-thought. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19520–19529.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. GPT-4o System Card. arXiv:2410.21276.

Jiang, D.; Guo, Z.; Zhang, R.; Zong, Z.; Li, H.; Zhuo, L.; Yan, S.; Heng, P.-A.; and Li, H. 2025. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot.

Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; and Levy, O. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 36652–36663.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35: 22199–22213.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1931–1941.

Labs, B. F.; Batifol, S.; Blattmann, A.; Boesel, F.; Consul, S.; Diagne, C.; Dockhorn, T.; English, J.; English, Z.; Esser, P.; Kulal, S.; Lacey, K.; Levi, Y.; Li, C.; Lorenz, D.; Müller, J.; Podell, D.; Rombach, R.; Saini, H.; Sauer, A.; and Smith, L. 2025. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. arXiv:2506.15742.

Li, D.; Li, J.; and Hoi, S. 2023. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36: 30146–30166.

Liu, D.; Zhao, S.; Zhuo, L.; Lin, W.; Xin, Y.; Li, X.; Qin, Q.; Qiao, Y.; Li, H.; and Gao, P. 2025. Lumina-mGPT: Illuminate Flexible Photorealistic Text-to-Image Generation with Multimodal Generative Pretraining. arXiv:2408.02657.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36: 34892–34916.

Nguyen, T.; Singh, K. K.; Shi, J.; Bui, T.; Lee, Y. J.; and Li, Y. 2025. Yo'Chameleon: Personalized Vision and Language Generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14438–14448.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Assran, M.; Ballas, N.; Galuba, W.; Howes, R.; Huang, P.-Y.; Li, S.-W.; Misra, I.; Rabbat, M.; Sharma, V.; Synnaeve, G.; Xu, H.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2024. DINOv2: Learning Robust Visual Features without Supervision. arXiv:2304.07193.

Ouyang, R.; Li, H.; Zhang, Z.; Wang, X.; Zhu, Z.; Huang, G.; and Wang, X. 2025. Motion-R1: Chain-of-Thought Reasoning and Reinforcement Learning for Human Motion Generation. arXiv:2506.10353.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PmLR.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36: 53728–53741.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. 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*, 22500–22510.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. arXiv:1707.06347.

Sun, K.; Liu, X.; Teng, Y.; and Liu, X. 2025. Personalized Text-to-Image Generation with Auto-Regressive Models. arXiv:2504.13162.

Sun, Q.; Cui, Y.; Zhang, X.; Zhang, F.; Yu, Q.; Wang, Y.; Rao, Y.; Liu, J.; Huang, T.; and Wang, X. 2024a. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14398–14409.

Sun, Z.; Chu, Z.; Zhang, P.; Wu, T.; Dong, X.; Zang, Y.; Xiong, Y.; Lin, D.; and Wang, J. 2024b. X-Prompt: Towards Universal In-Context Image Generation in Auto-Regressive Vision Language Foundation Models. arXiv:2412.01824.

Tan, Z.; Liu, S.; Yang, X.; Xue, Q.; and Wang, X. 2025. OminiControl: Minimal and Universal Control for Diffusion Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Team, C. 2025. Chameleon: Mixed-Modal Early-Fusion Foundation Models. arXiv:2405.09818.

Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv:2403.05530.

Thawakar, O.; Dissanayake, D.; More, K.; Thawkar, R.; Heakl, A.; Ahsan, N.; Li, Y.; Zumri, M.; Lahoud, J.; Anwer, R. M.; et al. 2025. LlamaV-o1: Rethinking Step-by-step Visual Reasoning in LLMs. *CoRR*.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv:2203.11171.

Wang, X.; Zhang, X.; Luo, Z.; Sun, Q.; Cui, Y.; Wang, J.; Zhang, F.; Wang, Y.; Li, Z.; Yu, Q.; Zhao, Y.; Ao, Y.; Min, X.; Li, T.; Wu, B.; Zhao, B.; Zhang, B.; Wang, L.; Liu, G.; He, Z.; Yang, X.; Liu, J.; Lin, Y.; Huang, T.; and Wang, Z. 2024. Emu3: Next-Token Prediction is All You Need. arXiv:2409.18869.

Wang, Y.; Wu, S.; Zhang, Y.; Yan, S.; Liu, Z.; Luo, J.; and Fei, H. 2025. Multimodal Chain-of-Thought Reasoning: A Comprehensive Survey. arXiv:2503.12605.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; Chi, E. H.; Hashimoto, T.; Vinyals, O.; Liang, P.; Dean, J.; and Fedus, W. 2022a. Emergent Abilities of Large Language Models. arXiv:2206.07682.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Wu, Y.; Zhu, L.; Liu, L.; Qiao, W.; Li, Z.; Yu, L.; and Li, B. 2025. Proxy-Tuning: Tailoring Multimodal Autoregressive Models for Subject-Driven Image Generation. arXiv:2503.10125.

Xie, J.; Yang, Z.; and Shou, M. Z. 2025. Show-o2: Improved Native Unified Multimodal Models. arXiv:2506.15564.

Xu, Y.; Li, C.; Zhou, H.; Wan, X.; Zhang, C.; Korhonen, A.; and Vulić, I. 2025. Visual Planning: Let's Think Only with Images. arXiv:2505.11409.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. arXiv:2308.06721.

Zhang, Z.; Zhang, A.; Li, M.; and Smola, A. 2023. Automatic Chain of Thought Prompting in Large Language Models. In *The Eleventh International Conference on Learning Representations*.

Zhao, Q.; Lu, Y.; Kim, M. J.; Fu, Z.; Zhang, Z.; Wu, Y.; Li, Z.; Ma, Q.; Han, S.; Finn, C.; et al. 2025. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1702–1713.

Zou, J.; Liao, B.; Zhang, Q.; Liu, W.; and Wang, X. 2025. OmniMamba: Efficient and Unified Multimodal Understanding and Generation via State Space Models. arXiv:2503.08686.