This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Let's Verify and Reinforce Image Generation Step by Step

Renrui Zhang[*†1], Chengzhuo Tong[*4], Zhizheng Zhao[*3], Ziyu Guo[*2], Haoquan Zhang[4]
Manyuan Zhang[1], Jiaming Liu[3], Peng Gao[4,5], Hongsheng Li[‡1,6]

CUHK [1]MMLab & [2]MiuLar Lab [3]Peking University
[4]Shanghai AI Lab [5]SIAT [6]CPII under InnoHK

Abstract

*Chain-of-Thought (CoT) reasoning has been extensively explored in large models to tackle complex understanding tasks. However, it remains an open question whether such strategies can be applied to verifying and reinforcing image generation scenarios. In this paper, we provide the first comprehensive investigation in the potential of CoT reasoning to enhance autoregressive image generation. We focus on three techniques: scaling test-time computation for verification, aligning model preferences with Direct Preference Optimization (DPO), and integrating these techniques for complementary effects. Our results demonstrate that these approaches can be effectively adapted and combined to significantly improve image generation performance. Furthermore, given the pivotal role of reward models in our findings, we propose the **P**otential **A**ssessment **R**eward **M**odel (**PARM**) specialized for autoregressive image generation. PARM adaptively assesses each generation step through a potential assessment mechanism, merging the strengths of existing reward models. Using our investigated reasoning strategies, we enhance a baseline model, Show-o, to achieve superior results, with a significant +24% improvement on the GenEval benchmark, surpassing Stable Diffusion 3 by +15%. We hope our study provides unique insights and paves a new path for integrating CoT reasoning with autoregressive image generation. Code is released at* https://github.com/ZiyuGuo99/Image-Generation-CoT.

1. Introduction

Large Language Models (LLMs) [2, 47, 48] and Large Multimodal Models (LMMs) [10, 33, 57] have gained remarkable achievements across language [5, 31], 2D image [8, 27], video [9, 17], and 3D [14, 16]. Building upon general understanding skills, recent efforts have been made toward enhancing LLMs and LMMs with Chain-of-Thought (CoT) reasoning capabilities [19, 50, 52, 60], e.g., OpenAI o1 [34], contributing to superior performance in mathematics [28, 58], science [15, 42], and coding [13, 62].

Despite the success in multimodal understanding, it remains under-explored *whether multi-step reasoning strategies can be effectively applied to image generation*. Considering the discrepancy between two tasks, we observe that, autoregressive image generation [3, 45, 51, 53] shares a similar output manner to the nature of LLMs and LMMs. Specifically, they both quantize the target data (language and image) into discrete tokens, and iteratively predict partial content conditioned on previously generated tokens.

As illustrated in Figure 1, LMMs leverage CoT to break down complex mathematical problems into manageable steps, which enables scaling test-time computation with reward models [24, 30, 44, 49] and reinforcement learning for preference alignment [18, 20, 29, 59]. Likewise, autoregressive image generation through step-by-step decoding can produce intermediate images, potentially allowing for similar verification and reinforcement techniques. This raises the question: *Can we verify and reinforce image generation step-by-step with strategies revealed by OpenAI o1?*

To this end, we conduct a systematic investigation into the potential of CoT reasoning for autoregressive image generation. We adopt Show-o [53], a latest discrete generative model, as our baseline, and evaluate on a challenging text-to-image generation benchmark: GenEval [12]. Specifically, we focus on examining two key perspectives: *1) Scaling test-time computation with Outcome/Process Reward Model (ORM/PRM) as verifiers*; and *2) Reinforced preference alignment with Direct Preference Optimization (DPO)*. The specifics are as follows:

- **ORM *vs* PRM as Test-time Verifiers.** As the top-1 result may not always be reliable, reward models are employed to score sampled candidates and perform outcome selection, where ORM is instance-level and PRM is process-level. In our settings, the score assesses whether each

[*]Equal Contribution [†]Project Lead [‡]Corresponding Author

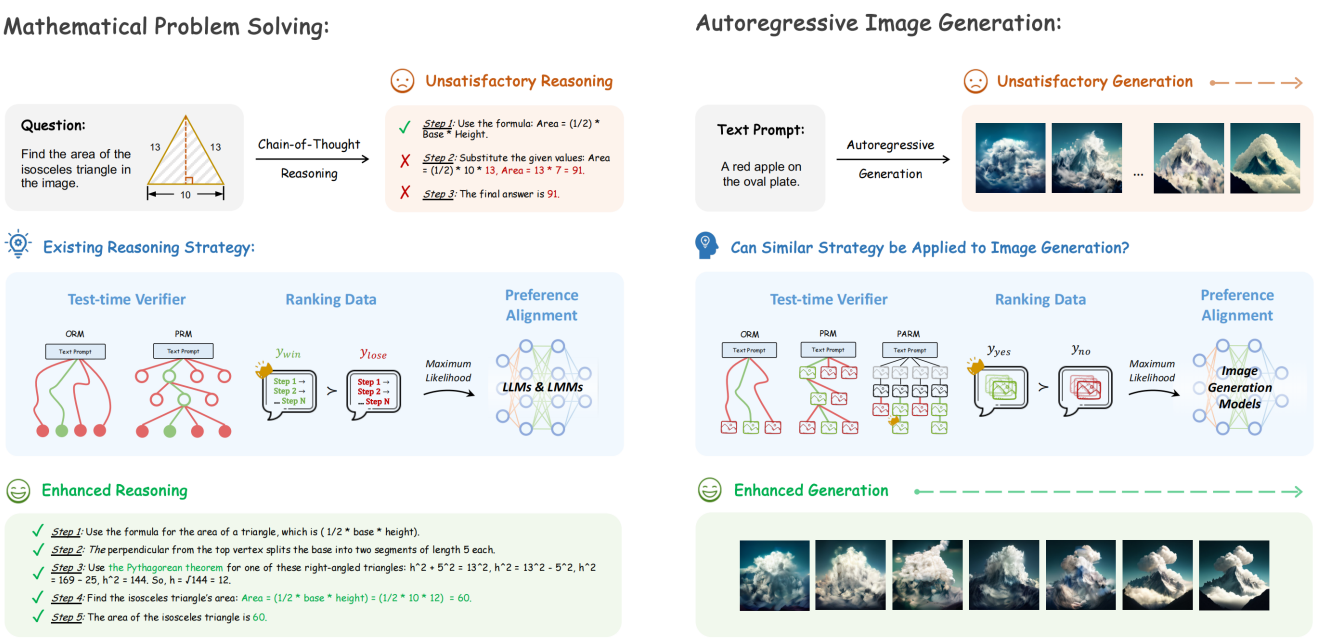

Figure 1. ***Can We Verify and Reinforce Image Generation with Chain-of-Thought (CoT) Reasoning Strategies?*** Given the success of CoT reasoning in LLMs and LMMs (Left), we provide *the first* investigation to comprehensively explore the potential of applying current reasoning techniques to autoregressive image generation (Right), including test-time verification and preference alignment.

candidate image is inherently reasonable and aligns with the given textual prompt. We prompt LLaVA-OneVision (7B) [22] as a zero-shot reward model, and then curate text-to-image ranking data for reward fine-tuning. We apply a best-of-N selection approach in the comparison of zero-shot and fine-tuned reward models.

Observation: *ORM demonstrates significant improvement, while PRM offers minimal benefit.*

- **Test-time Verifiers *vs* Preference Alignment.** Exploring the trade-off between inference-time and post-training offers valuable insights into the model's attainable performance. Preference alignment are adopted to elicit the implicit reasoning capabilities from their widely learned knowledge. In this study, we construct ranking preference data and apply DPO alignment with iterative training [36] to optimize the generation decoding process, comparing its effectiveness to test-time verification.

  Observation: *DPO alignment with iterative training attains stronger results to the fine-tuned ORM verifier.*

- **Preference Alignment *plus* Test-time Verifiers.** After investigating the individual impact, we integrate the two techniques to highlight their complementary potential in autoregressive image generation. We consider three approaches: *1) DPO with reward model guidance*, i.e., integrating DPO's policy with reward models' objectives for alignment; *2) Verification after DPO alignment*, i.e., applying reward models for best-of-N selection on DPO-aligned models; and *3) Verification after DPO with reward model guidance*, i.e., a combination of *1* and *2*.

  Observation: *All integration methods lead to greater improvements, indicating complementary characteristics.*

Through our experiments, we demonstrate the promsing potential of applying CoT reasoning strategies to image generation scenarios, uncovering their adaptation methods and compatibility. Furthermore, we identify significant room for improvement in reward models tailored for autoregressive image generation. For ORM, the global-level assessments are unable to capture the nuanced step-wise information, leading to inaccurate reward judgments. For PRM, the early-stage images tend to appear blurry, while later-stage images across different paths often converge to visually similar outputs, limiting its discrimination ability.

To alleviate these issues, we propose a specialized reward model for autoregressive image generation, termed **P**otential **A**ssessment **R**eward **M**odel (**PARM**). PARM adaptively verifies the generation process step by step with three delicately designed tasks: *1) Judge which step is clear and convincing enough to be evaluated*, given that most early-stage images are typically blurry; *2) Assess whether the current step has the potential to yield a high-quality final image*, since later-stage images generally do not change too much; and *3) Score the remaining final paths for selecting the best one*, similar to an ORM. In this way, PARM can adaptively conduct assessment at appropriate steps (overcoming PRM's scoring challenges), while effectively cap-

turing fine-grained step-by-step cues (complementing the coarse evaluation of ORM). Our experiments showcase that PARM significantly outperforms both ORM and PRM, which improves the baseline model by +24% on GenEval, surpassing the advanced Stable Diffusion 3 [7] by +15%.

Our main contributions are summarized as follows:

- We present *the first* comprehensive empirical study of applying CoT reasoning strategies to autoregressive image generation domains, providing unique insights into the future advancement of this field.
- We investigate specific adaption methods of techniques, including test-time verification, preference alignment, and ranking data curation, to autoregressive image generation, indicating their performance and complementarity.
- We further introduce PARM, a new reward model tailored for image generation scenarios, which adaptively performs step-wise potential assessment and path selection, significantly enhancing text-to-image generation quality.

2. Our Investigation

Chain-of-Thought (CoT) reasoning has been widely exploited to solve complex problems for language and multimodal understanding. In this study, we conduct a systematic investigation aiming to find out, whether we can verify and reinforce image generation step-by-step.

2.1. Overview

Task Formulation. To enable the applicability of current reasoning techniques, we focus on autoregressive image generation tasks, demonstrated by models such as MaskGiT [3], LlamaGen [45], and Show-o [53]. This task employs a data representation and output paradigm akin to those used in LLMs and LMMs, while achieving comparable performance to continuous diffusion models [21, 23, 26, 37, 39, 40, 56, 61]. Specifically, it leverages quantized autoencoders [6] to transform images into discrete tokens, allowing for the classification loss of Direct Preference Optimization (DPO) [38] in post-training. Additionally, it iteratively predicts one or more tokens at each step, conditioned on prior outputs, thereby creating reasoning paths suitable for step-wise verification.

Experimental Settings. We select Show-o as our baseline model for this investigation, a latest autoregressive image generation model with advanced capabilities. To comprehensively evaluate the effectiveness of different strategies, we assess the text-to-image generation performance on a rigorous benchmark: GenEval [12]. This scenario challenges the model to produce images with not only high visual quality and image-text alignment, but also accurate object attribute and co-occurrence. In the subsequent sections, we explore three strategies to improve the step-by-step decoding of autoregressive image generation: test-time verification (Section 2.2), preference alignment (Section 2.3), and their combination (Section 2.4).

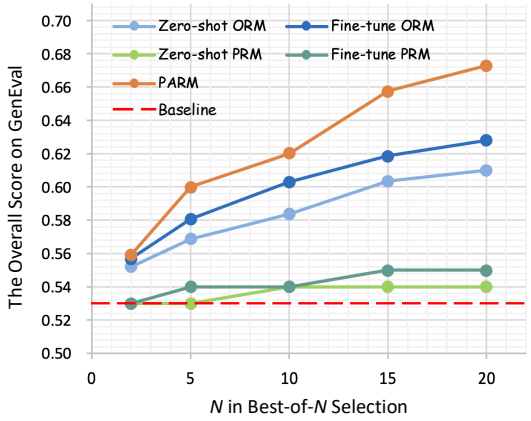

Figure 2. **Comparison of Reward Models as Test-time Verifiers.** We adopt Show-o [53] as the 'Baseline' and evaluate Best-of-N selection on the GenEval [12] benchmark.

2.2. ORM *vs* PRM as Test-time Verifiers

Scaling test-time computation [24, 30, 44, 49] to enhance reasoning capabilities has emerged as an effective alternative to scaling training costs. Current approaches often employ reward models as test-time verifiers within CoT reasoning paths, typically using two main categories: Outcome Reward Model (ORM) and Process Reward Model (PRM). Drawing inspiration from these methods, we respectively implement and evaluate them within the context of autoregressive image generation, as illustrated in Figure 3.

ORM. Based on multiple complete reasoning outputs, ORM assigns each candidate a reword score and select the most confident one using a best-of-N strategy. In our study, we adopt ORM solely to evaluate the generated image at the final step, rather than the entire CoT process in mathematical reasoning tasks. Specifically, we begin with a zero-shot ORM, followed by curating a text-to-image ranking dataset to fine-tune the ORM for enhancement, as outlined below:

- ***Zero-shot ORM:*** We leverage a pre-trained LLaVA-OneVision (7B) [22], an LMM with superior generalization, as our zero-shot ORM. We input the text prompt along with the generated image into LLaVA-OneVision, and devise a prompt template (detailed in the Supplementary Material) to activate its visual understanding capabilities. This model assesses the quality of candidate images, providing binary responses, 'yes' (good quality) or

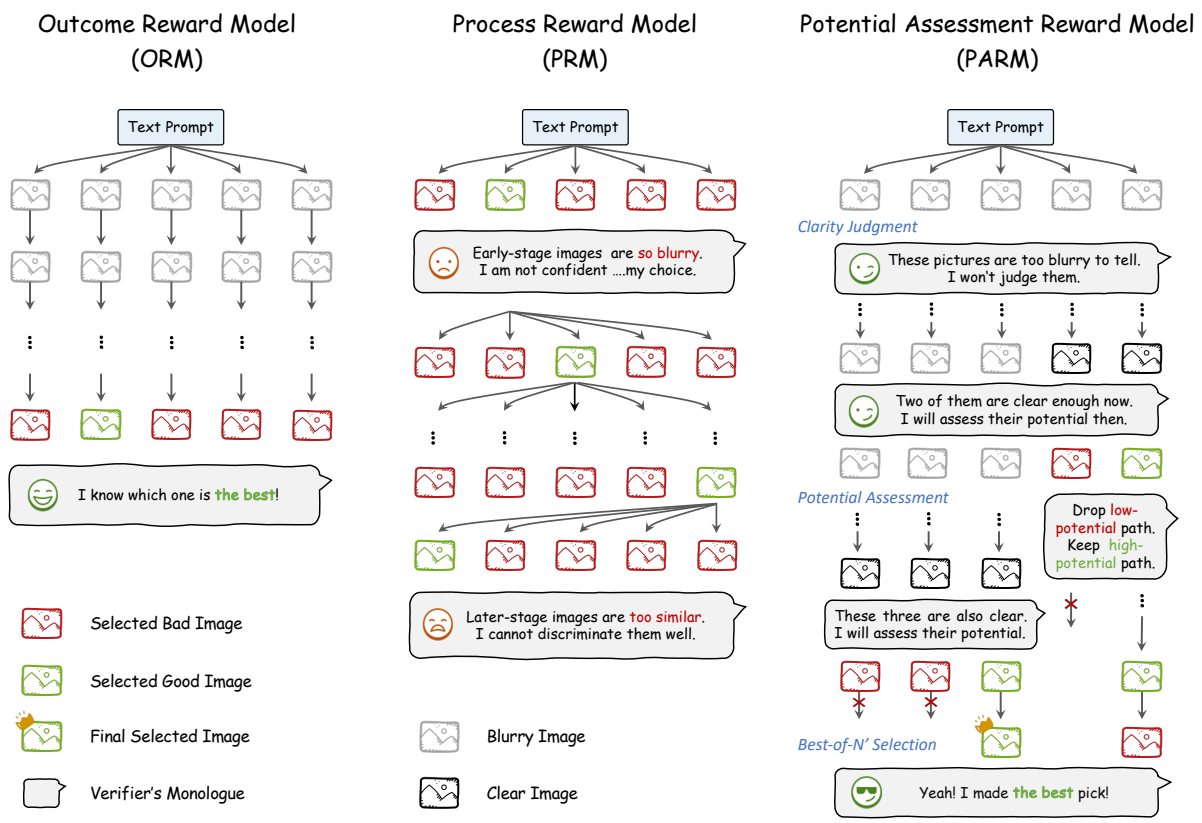

Figure 3. **Investigation of Reward Models in Autoregressive Image Generation.** For test-time verification, we implement Outcome Reward Model (ORM) and Process Reward Model (PRM), and introduce a new Potential Assessment Reward Model customized for image generation scenarios, which progressively performs three tasks (highlighted in blue) to enhance the reasoning of generation process.

'no' (low quality). The candidate image with the highest probability of 'yes' is then selected as the final output.

- ***ORM Ranking Data Curation:*** To enhance the accuracy of outcome rewards, we curate a dataset of 288K text-to-image ranking examples for fine-tuning ORM. First, we prompt GPT-4 [32] to generate a list of 200 countable daily object names with specific colors. Using these objects, we apply the six object-centered prompt templates from GenEval, constructing a diverse set of 13K text prompts. We perform a strict filtering to ensure that these prompts do not overlap with the GenEval test samples. Then, using our baseline model, Show-o, we synthesize around 50 images per prompt at a high temperature. After that, we label each image with a binary annotation of 'yes' or 'no' using the evaluation metric in GenEval.
- ***Fine-tuned ORM:*** Using the curated ranking dataset, we fine-tune LLaVA-OneVision to enhance its capability for assessing image quality and cross-modal alignment. The training data format is consistent with the prompt template used in the zero-shot ORM, incorporated with our constructed 288K text prompts and associated images. The model is fine-tuned for one epoch, using a batch size of 8 and a learning rate of $1e^{-5}$. This fine-tuning process enables the ORM to capture more intricate aspects of object composition and nuanced visual-text relationships, resulting in more reliable reward scoring.

PRM. Different from ORM that evaluates only the final output, we utilize PRM to provide a reward score to each candidate with different steps throughout the generation process. Similar to our previous investigation, we start from a zero-shot PRM, LLaVA-OneVision, and then curate 10K step-wise text-to-image ranking data to obtain a fine-tuned PRM. We refer to the Supplementary Material for detailed impelemtation of PRM.

Experiments and Insights. As showcased in the middle of Table 1, we compare the test-time verification results between ORM and PRM with a best-of-20 strategy. The observations are summarized below:

- ***Test-time verification can significantly boost generation performance.*** Compared to the baseline scores of 53% on GenEval, the fine-tuned ORM as a test-time verifier achieves the highest gains of +10%. This finding suggests that current autoregressive image generation models, similar to LLMs and LMMs, face challenges with in-

Table 1. **Test-time Verifiers (ORM *vs* PRM) *vs* Preference Alignment.** We evaluate text-to-image generation on the GenEval [12] benchmark and adopt Show-o [53] as the autoregressive baseline model. 'ORM/PRM' and 'DPO' denote Outcome/Process Reward Model and Direct Preference Optimization [38], respectively. We adopt the best-of-N selection for test-time verifiers, setting $N = 20$, and highlight the better-performed variant of each reasoning strategy in green.

Reasoning Strategy	Method	Setting	Single object	Two object	Counting	Colors	Position	Attribute binding	**Overall**
Baseline	-	-	0.95	0.52	0.49	0.82	0.11	0.28	0.53
Test-time Verifier	ORM	Zero-Shot	0.99	0.63	0.63	0.84	0.19	0.39	0.61
		Fine-tuned	0.99	0.72	0.65	0.84	0.25	0.33	0.63
	PRM	Zero-Shot	0.98	0.51	0.54	0.82	0.11	0.23	0.53
		Fine-tuned	0.98	0.55	0.54	0.83	0.13	0.29	0.55
Preference Alignment	DPO	-	0.96	0.70	0.50	0.82	0.30	0.43	0.62
		Iterative	0.98	0.72	0.53	0.84	0.40	0.46	0.65

consistent and unstable decoding paths. Consequently, a test-time verification strategy is essential to identify and follow the most reliable reasoning path.

- ***ORM exhibits stronger enhancement capabilities than PRM.*** In contrast to ORM providing a clear benefit, PRM yields only marginal improvements, e.g., +2% on GenEval after fine-tuning. This discrepancy arises from the unique characteristics of the autoregressive image generation task in two key ways: *1)* Images at early steps are too blurry for PRM to effectively interpret their visual features and image-text alignment. *2)* Images at later steps tend to exhibit minimal differences, making it challenging for PRM to discriminate. Whereas, ORM evaluates images at the final step, which provides sufficient visual and semantic information for accurate judgment.
- ***Fine-tuning by ranking data enhances verification results and demonstrates improved scaling performance.*** As illustrated in Figure 2, both fine-tuned ORM and PRM outperform their zero-shot counterparts, achieving higher scores with larger N values in the best-of-N selection. Additionally, as N increases, fine-tuned reward models show greater improvements, indicated by steeper curves, reflecting a better scaling response to test-time computation. This highlights the effectiveness of our curated ranking dataset in refining reward accuracy and benefiting scalability across a broader range of candidates.

2.3. Test-time Verifiers *vs* Preference Alignment

Post-training has been widely utilized in existing LLMs and LMMs to align model outputs with human preferences. Common techniques include reinforcement learning with reward models, e.g., Proximal Policy Optimization (PPO) [43], and its streamlined counterpart with classification objectives, e.g., Direct Preference Optimization (DPO) [38]. Given that most autoregressive image generation models inherently function within a classification framework, we leverage the simplicity of DPO alignment to enhance the quality of generated images.

DPO Ranking Data Curation. To bypass reinforcement learning, DPO leverages an implicit rewarding mechanism by training on a ranking dataset of paired preferred and dispreferred responses, corresponding to well-generated and poor-quality images in our case. Fortunately, we have already constructed a substantial amount of ranking data for training ORM, annotated with 'yes' and 'no' labels to indicate the generation quality. Building on this, we utilize the 13K unique text prompts from the ORM training dataset and, for each prompt, randomly pair two generated images, one labeled 'yes' and the other 'no', yielding a total of 10K paired data for DPO alignment.

DPO for Autoregressive Image Generation. Since autoregressive image generation models are also trained using a cross-entropy loss, we can directly apply the maximum likelihood objective in DPO to our setting. In detail, the parameterized policy is initialized from Show-o and optimized during training, while the reference policy is also initialized from Show-o but kept frozen. The objective encourages the model to assign a higher likelihood to preferred images over dispreferred images, aligning with the curated preference structure. The DPO training is conducted over one epoch with a batch size of 10 and a learning rate of $1e^{-5}$.

DPO with Iterative Training. Following the initial stage of DPO alignment, the model has learned to generate images that better align with the preferred responses. Inspired by iterative DPO [35], we further refine this alignment by applying the newly aligned model to generate updated ranking data based on the text prompts in $\mathcal{D}$. We annotate these new images with 'yes' or 'no' labels using the same method in Section 2.2. For each prompt, we collect paired images

Table 2. **Test-time Verifiers *plus* Preference Alignment.** We evaluate text-to-image generation on the GenEval [12] benchmark and adopt Show-o [53] as the autoregressive baseline model. 'Ft. ORM' and 'It. DPO' denote the fine-tuned ORM and iterative DPO [38]. We explore three combination approaches ('1*st*, 2*nd*, and 3*rd* Integration') of reward models and preference alignment, comparing 'individual' results. We adopt the best-of-N selection for test-time verifiers, setting $N = 20$, and highlight the best-performing integration in green.

Reasoning Strategy	Test-time Verifier	Preference Alignment	Reward Guidance	Single object	Two object	Counting	Colors	Position	Attribute binding	**Overall**
Baseline	-	-	-	0.95	0.52	0.49	0.82	0.11	0.28	0.53
Individual	Ft. ORM	-	-	0.99	0.72	0.65	0.84	0.25	0.33	0.63
	-	It. DPO	-	0.98	0.72	0.53	0.84	0.40	0.46	0.65
1*st* Integration	-	It. DPO	Ft. ORM	0.98	0.78	0.44	0.81	0.50	0.48	0.67
2*nd* Integration	Ft. ORM	It. DPO	-	0.98	0.80	0.62	0.83	0.59	0.54	0.72
3*rd* Integration	Ft. ORM	It. DPO	Ft. ORM	0.98	0.84	0.64	0.85	0.66	0.52	0.75

labeled y_{yes} and y_{no}, and exclude samples where all images receive the same label, resulting in a refined DPO ranking dataset of 7K samples. By conducting another round of DPO, the model can be further improved by learning from more informative preference relations. We iterate the DPO training process once with the same training configurations.

Experiments and Insights. In the bottom of Table 1, we present the evaluation results of DPO alignment and compare it with the performance with test-time verification. The observations are summarized below:

- ***DPO alignment can effectively reinforce the generation performance.*** On GenEval, initial DPO alignment improves the baseline model's performance by +9%. With iterative training, these gains are further extended as +11%, highlighting the effectiveness of a refined preference dataset in strengthening model alignment with desired outputs. This demonstrates that DPO alignment can serve as a powerful method for enhancing autoregressive image generation models, especially in scenarios where explicit preference data is available to guide training.
- ***Initial DPO matches test-time verification, while iterative DPO surpasses.*** After the initial alignment, the model achieves performance comparable to that of the fine-tuned ORM, the top-performing variant for verification. However, with iterative alignment on refined ranking data, the model outperforms all test-time verifiers, i.e., +2% over the fine-tuned ORM. This demonstrates the potential of iterative DPO to progressively reinforce image generation capabilities through updated ranking data.

2.4. DPO Alignment *plus* Test-time Verifiers

The investigations above have demonstrated the individual effectiveness of test-time verification and DPO alignment. Next, we explore three approaches to integrate these two techniques to assess their potential for complementary enhancement in image generation, leveraging both the adaptability of verifiers and the reinforcement of DPO.

DPO with Reward Model Guidance. As discussed in previous works [54], DPO can struggle with out-of-distribution responses due to distribution shifts from the ranking dataset. A potential solution [1, 55] is to incorporate prompt-only datasets during post-training and leverage a reward model to provide online preference guidance. Following this approach, we adopt our fine-tuned ORM as the explicit reward model to offer more generalized preference feedback, and add the online objectives with the original DPO loss We maintain the same training data and configurations as in the initial DPO alignment stage.

Verification after DPO Alignment. We observe that verification and DPO techniques may naturally complement each other in two key ways: *1)* They operate independently at different stages of implementation, i.e., post-training and test-time; and *2)* DPO refines the internal knowledge distribution within the model to enhance reasoning, while verification focuses on selecting the optimal reasoning path within this refined distribution. Therefore, we apply the fine-tuned ORM for best-of-N selection directly on the model after DPO alignment.

Verification after DPO with Reward Model Guidance. In this approach, we combine the strengths of both DPO with reward model guidance and test-time verification. Our goal is to achieve optimal alignment, enhancing the model's generalization capabilities during training, while also ensuring reliable image generation paths at inference time.

Experiments and Insights. Table 2 reports the text-to-image generation scores with different integration methods. The observations are summarized below:

- ***Verification and alignment perform expected complementary characteristics.*** Across all three approaches, verification and alignment complement each other effectively. For instance, the third integration method outperforms DPO alignment alone by +10%, and surpasses the

Table 3. **Performance Comparison on the GenEval [12] Benchmark.** Compared to existing diffusion and autoregressive models, we investigate the potential of Chain-of-Thought (CoT) reasoning strategies in text-to-image generation. 'Ft. ORM' and 'It. DPO' denote the fine-tuned ORM and iterative DPO [38]. **PARM** refers to our Potential Assessment Reward Model specialized for autoregressive image generation. We adopt the best-of-20 selection for test-time verifiers, and highlight the best and second-best overall scores in green and red.

Model	Test-time Verifier	Preference Alignment	Reward Guidance	Single object	Two object	Counting	Colors	Position	Attribute binding	**Overall**
PixArt-α [4]	-	-	-	0.98	0.50	0.44	0.80	0.08	0.07	0.48
SD v2.1 [41]	-	-	-	0.98	0.51	0.44	0.85	0.07	0.17	0.50
DALL-E 2 [39]	-	-	-	0.94	0.66	0.49	0.77	0.10	0.19	0.52
SDXL [37]	-	-	-	0.98	0.74	0.39	0.85	0.15	0.23	0.55
SD 3 (d=24) [7]	-	-	-	0.98	0.74	0.63	0.67	0.34	0.36	0.62
LlamaGen [45]	-	-	-	0.71	0.34	0.21	0.58	0.07	0.04	0.32
Chameleon [46]	-	-	-	-	-	-	-	-	-	0.39
LWM [25]	-	-	-	0.93	0.41	0.46	0.79	0.09	0.15	0.47
SEED-X [11]	-	-	-	0.97	0.58	0.26	0.80	0.19	0.14	0.49
Show-o [53]	-	-	-	0.95	0.52	0.49	0.82	0.11	0.28	0.53
	Ft. ORM	-	-	0.99	0.72	0.65	0.84	0.25	0.33	0.63
	-	It. DPO	-	0.98	0.72	0.53	0.84	0.40	0.46	0.65
	Ft. ORM	It. DPO	Ft. ORM	0.98	0.84	0.64	0.85	0.66	0.52	0.75
	PARM	-	-	0.99	0.77	0.68	0.86	0.29	0.45	0.67
	-	It. DPO	**PARM**	0.97	0.75	0.60	0.83	0.54	0.53	0.69
	PARM	It. DPO	-	0.98	0.83	0.64	0.84	0.59	0.62	0.74
	PARM	It. DPO	**PARM**	0.99	0.86	0.67	0.84	0.66	0.64	0.77

verification alone by +12%. These results highlight the potential of combining verification and alignment techniques in future autoregressive image generation tasks, enabling the production of high-quality outputs that are both preference-aligned and test-time reliable.

- ***Applying verifiers to both training and test time yields maximum enhancement.*** We observe the third combination approach delivers the most significant gains, outperforming the first approach by +8%, and the second by +3%. This suggests that, even with a model already aligned to preferences, the fine-tuned ORM can play complementary roles in the test-time decoding. These dual functions reinforce each other, highlighting the versatility of reward models in autoregressive image generation.

3. Potential Assessment Reward Model

From our comprehensive investigation, reward models prove valuable by enabling both decoding path selection and preference reward guidance. However, we still observe considerable room for enhancing reward models.

Limitation of ORM and PRM. *1)* ORM showcases strong performance by selecting optimal final outputs, yet it lacks the capacity to provide fine-grained, step-wise evaluation at each generation step. *2)* While PRM has demonstrated effectiveness in understanding tasks such as mathematics, it is less suitable for autoregressive image generation. As analyzed in Section 2.2, PRM struggles with early-stage images that are too blurry for reliable evaluation, given that only a few regions are decoded. In later stages, images derived from similar previous steps lack sufficient distinction, challenging for PRM to discriminate.

PARM. Motivated by these observations, we propose the Potential Assessment Reward Model (PARM), a specialized reward model tailored for autoregressive image generation, as illustrated in Figure 3. PARM combines the best of both worlds: *1)* it operates adaptively in a step-wise manner, using a potential assessment mechanism to overcome PRM's evaluation challenges; and *2)* it performs a best-of-N' selection across N' ($N' \leq N$) high-potential reasoning paths, thus inheriting ORM's advantage. Specifically, the methodology of PARM contains three progressive tasks:

1. **Clarity Judgment.** In the best-of-N setting, we first sample N different reasoning paths for image generation. Then, at each intermediate step, PARM evaluates whether the partially generated image contains enough visual clarity to be meaningfully assessed, assigning a binary label If labeled 'no', the model skips to the next step. If labeled 'yes', the model proceeds to the next task for potential assessment. This pre-judgment prevents scoring on early, blurry images that lack informative content (as seen in PRM), ensuring only sufficiently clear steps are considered for rewarding.

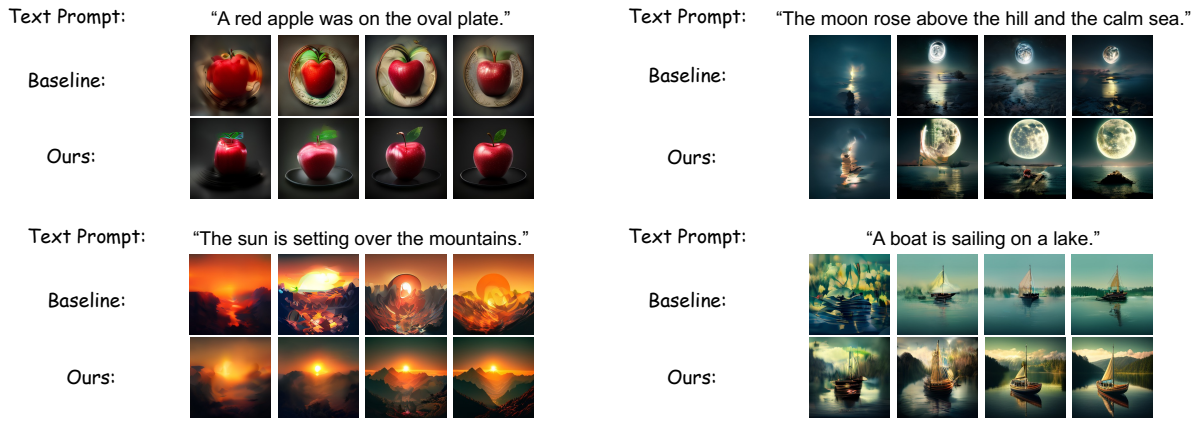

Figure 4. **Qualitative Results using Our Reasoning Strategies.** Show-o [53] is adopted as the baseline model, and compared to our best-performing reasoning strategy: integrating PARM with iterative DPO for both reward model guidance and test-time verification.

2. **Potential Assessment.** For each clear step that passes the clarity judgment, PARM assesses the potential of the current step to determine whether it can lead to a high-quality final image, again using a binary label. If labeled 'no', the generation path is truncated immediately. If labeled 'yes', the path is preserved to produce the final image. This approach is based on the observation that, once an image at a given step is clear enough to evaluate, its overall layout and structure are unlikely to change significantly in subsequent steps, making it a reliable candidate for potential assessment. This task helps identify promising intermediate steps accurately, effectively pruning low-potential candidates during inference.
3. **Best-of-N' Selection.** After completing the above two tasks, suppose there are N' high-potential paths remaining to produce the final images ($N' \leq N$). PARM then performs a best-of-N' selection to identify the most promising image candidates as the output. If $N' = 0$, the model defaults to selecting the reasoning path with the lowest probability of a 'no' label as the output. This final task leverages ORM's global selection capabilities to ensure a high-quality generated image.

PARM Ranking Data Curation. To empower PARM with robust capabilities, we curate a new ranking dataset with 400K instances by re-annotating the 13K text prompts from ORM ranking data. The dataset is structured into three subsets corresponding to the three evaluation tasks, containing 120K, 80K, and 200K instances, respectively. Please refer to the Supplementary Material for detailed data formats.

Experiments and Insights. With the new reward model, we revisit our previous investigation by applying PARM to different approaches enhancing autoregressive image generation. The observations are summarized below:

- ***PARM demonstrates the best-performing reward model across different strategies.*** Table 3 and Figure 2 present the effectiveness of PARM as test-time verifiers, significantly outperforming other reward models, e.g., +6% to the fine-tuned ORM. Additionally, PARM scales effectively with increasing N, highlighting its potential for further improvement with larger test-time computation. PARM also outperforms iterative DPO, the enhanced preference alignment strategy with refined data. Furthermore, PARM better harnesses the complementary strengths with post-training, consistently attaining higher integration scores than the fine-tuned ORM. These results underscore PARM's capability as a versatile and robust reward model for autoregressive image generation.
- ***With PARM, our baseline model (Show-o) is enhanced to achieve leading generation performance.*** Compared to other image generation models in Table 3, our best-performing configuration, i.e., integrating PARM with iterative DPO in both post-training and test-time, achieves a score of 77%, improving the baseline by +24% and surpassing the advanced Stable Diffusion 3 [7] by +15%. Qualitative results are showcased in Figure 4. This result demonstrates that, using appropriate reasoning strategies, autoregressive image generation can be effectively verified and reinforced for superior performance.

4. Conclusion

In this work, we investigate the adaption and potential of CoT reasoning strategies in autoregressive image generation. Through a systematic investigation, we demonstrate that different reasoning strategies can effectively improve image generation. We further introduce the tailored Potential Assessment Reward Model (PARM), which evaluates the step-wise potential of image generation for adaptive reward scoring with superior results. Our experiments underscore the promise of CoT reasoning in autoregressive image generation, advancing this field in new directions.

Acknowledgment

This project is funded in part by National Key R&D Program of China Project 2022ZD0161100, by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Commission (ITC)'s InnoHK, by General Research Fund of Hong Kong RGC Project 14204021, by NSFC-RGC Joint Fund Project N_CUHK498/24, by the National Natural Science Foundation of China (No. 62206272). Hongsheng Li is a PI of CPII under the InnoHK.

References

[1] AI@Meta. Llama 3 model card, 2024.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in neural information processing systems*, pages 1877–1901, 2020.

[3] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325, 2022.

[4] Junsong Chen, Jincheng YU, Chongjian GE, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024.

[5] OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass, 2023.

[6] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024.

[8] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[9] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

[10] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

[11] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation, 2024.

[12] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

[13] Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yu Wu, YK Li, et al. Deepseek-coder: When the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*, 2024.

[14] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*, 2023.

[15] Ziyu Guo, Renrui Zhang, Hao Chen, Jialin Gao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. Sciverse. *https://sciverse-cuhk.github.io*, 2024.

[16] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Chengzhuo Tong, Peng Gao, Chunyuan Li, and Pheng-Ann Heng. Sam2point: Segment any 3d as videos in zero-shot and promptable manners. *arXiv preprint arXiv:2408.16768*, 2024.

[17] Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*, 2023.

[18] Alex Havrilla, Yuqing Du, Sharath Chandra Raparthy, Christoforos Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravinskyi, Eric Hambro, Sainbayar Sukhbaatar, and Roberta Raileanu. Teaching large language models to reason with reinforcement learning. *arXiv preprint arXiv:2403.04642*, 2024.

[19] 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.

[20] Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. *arXiv preprint arXiv:2406.18629*, 2024.

[21] Jiayi Lei, Renrui Zhang, Xiangfei Hu, Weifeng Lin, Zhen Li, Wenjian Sun, Ruoyi Du, Le Zhuo, Zhongyu Li, Xinyue Li, et al. Imagine-e: Image generation intelligence evaluation of state-of-the-art text-to-image models. *arXiv preprint arXiv:2501.13920*, 2025.

[22] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[23] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and

Kaiming He. Autoregressive image generation without vector quantization. *arXiv preprint arXiv:2406.11838*, 2024.

[24] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

[25] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention, 2024.

[26] Jiaming Liu, Hao Chen, Pengju An, Zhuoyang Liu, Renrui Zhang, Chenyang Gu, Xiaoqi Li, Ziyu Guo, Sixiang Chen, Mengzhen Liu, et al. Hybridvla: Collaborative diffusion and autoregression in a unified vision-language-action model. *arXiv preprint arXiv:2503.10631*, 2025.

[27] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

[28] 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.

[29] Zimu Lu, Aojun Zhou, Ke Wang, Houxing Ren, Weikang Shi, Junting Pan, and Mingjie Zhan. Step-controlled dpo: Leveraging stepwise error for enhanced mathematical reasoning. *arXiv preprint arXiv:2407.00782*, 2024.

[30] Qianli Ma, Haotian Zhou, Tingkai Liu, Jianbo Yuan, Pengfei Liu, Yang You, and Hongxia Yang. Let's reward step by step: Step-level reward model as the navigators for reasoning. *arXiv preprint arXiv:2310.10080*, 2023.

[31] OpenAI. Chatgpt. https://chat.openai.com, 2023.

[32] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

[33] OpenAI. GPT-4V(ision) system card, 2023.

[34] OpenAI. Openai o1. [Online], 2024. https://openai.com/index/learning-to-reason-with-llms/.

[35] Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization, 2024. *URL https://arxiv.org/abs/2404.19733*.

[36] Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization. *arXiv preprint arXiv:2404.19733*, 2024.

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

[38] 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, 2024.

[39] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[42] Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301, 2022.

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

[44] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters, 2024.

[45] Linzhuang Sun, Hao Liang, and Wentao Zhang. Beats: Optimizing llm mathematical capabilities with backverify and adaptive disambiguate based efficient tree search. *arXiv preprint arXiv:2409.17972*, 2024.

[46] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2024.

[47] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[48] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[49] Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, 2024.

[50] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[51] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

[52] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[53] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o:

One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

[54] Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. Is dpo superior to ppo for llm alignment? a comprehensive study. *arXiv preprint arXiv:2404.10719*, 2024.

[55] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[56] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. *ICLR 2024*, 2023.

[57] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. LLaMA-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *The Twelfth International Conference on Learning Representations*, 2024.

[58] 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.

[59] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, et al. Mavis: Mathematical visual instruction tuning. *arXiv preprint arXiv:2407.08739*, 2024.

[60] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

[61] 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.

[62] Qihao Zhu, Daya Guo, Zhihong Shao, Dejian Yang, Peiyi Wang, Runxin Xu, Y Wu, Yukun Li, Huazuo Gao, Shirong Ma, et al. Deepseek-coder-v2: Breaking the barrier of closed-source models in code intelligence. *arXiv preprint arXiv:2406.11931*, 2024.