This ICCV 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.

ImageGen-CoT: Enhancing Text-to-Image In-context Learning with Chain-of-Thought Reasoning

Jiaqi Liao[1†], Zhengyuan Yang[1], Linjie Li[1], Dianqi Li, Kevin Lin[1], Yu Cheng[2], Lijuan Wang[1✉]
[1] Microsoft [2] The Chinese University of Hong Kong

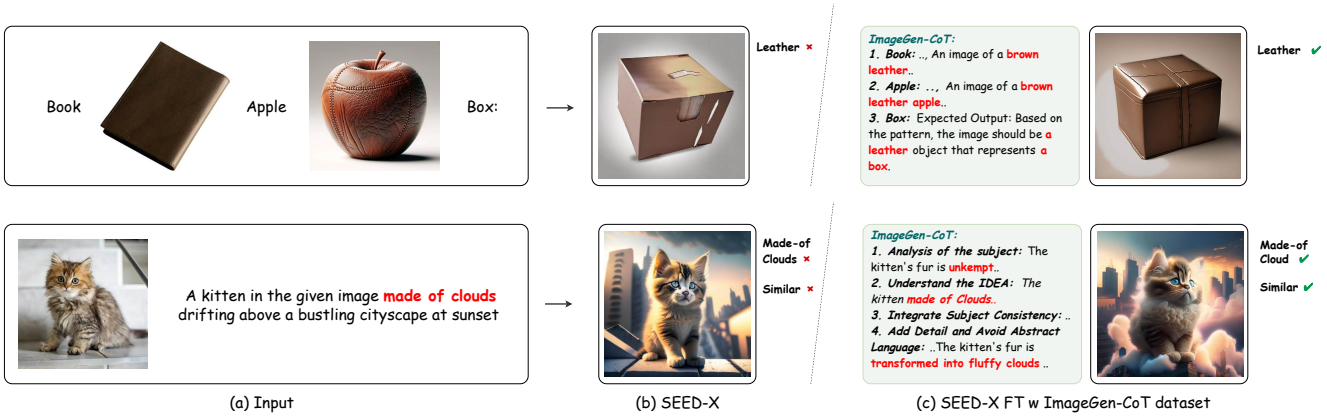


Figure 1. **Comparisons between SEED-X and SEED-X FT w ImageGen-CoT dataset.** Our method forces the model to generate a reasoning chain before the ImageGen. The top row shows SEED-X failing to infer 'Leather' style, generating only a box, while our approach enables SEED-X to recognize the leather style and produce the intended leather box. The bottom row shows SEED-X failing to capture the unkempt fur and cloud, while our method successfully recognizes these key elements.

Abstract

In this work, we study the problem of Text-to-Image In-Context Learning (T2I-ICL). While Unified Multimodal LLMs (MLLMs) have advanced rapidly in recent years, they struggle with contextual reasoning in T2I-ICL scenarios. To address this limitation, we propose a novel framework that incorporates a reasoning chain called ImageGen-CoT prior to image generation. To avoid generating ineffective reasoning steps, we develop an automatic pipeline to curate a high-quality ImageGen-CoT dataset. We then fine-tune MLLMs using this dataset to enhance their contextual reasoning capabilities. To further enhance performance, we explore test-time scale-up strategies and propose a novel hybrid scaling approach. This approach first generates multiple reasoning chains and then produces multiple images for each chain via sampling. Extensive experiments demonstrate the effectiveness of our proposed method. Notably, fine-tuning with the ImageGen-CoT dataset leads to a substantial 80% performance gain for SEED-X on T2I-ICL tasks. See our project page at https://ImageGen-CoT.github.io/. *Code will be open-sourced.*

† Interns at Microsoft.

1. Introduction

Human intelligence excels at learning novel concepts through contextual observation and adapting to new inputs. When presented with a series of interleaved text-image examples—such as "a leather-bound book", followed by "a leather apple"—and then asked to generate an image for the query "a box," humans intuitively infer the implicit pattern of "leather" and apply it to the new query, resulting in "a leather box". This reasoning ability to learn novel concepts from multimodal contexts underpins creative problem-solving. Existing unified Multimodal Large Language Models (unified MLLMs) [5, 8, 9, 14, 18, 28, 34] have demonstrated remarkable capabilities in multimodal understanding and generation within a single model architecture. Given their ability to process and generate across modalities similar to human cognition, it is natural to investigate whether these models can exhibit reasoning capabilities comparable to those of humans. To evaluate this, we adopt the Text-to-Image In-Context Learning (T2I-ICL)

task [40], which requires models to process interleaved text-image inputs and generate coherent outputs by learning from multimodal contexts (Figure 1). Despite the impressive capabilities of unified MLLMs, our experiments reveal that they struggle to replicate this reasoning capability, often failing to grasp contextual relationships or preserve compositional consistency in T2I-ICL tasks.

To overcome these challenges, building upon the demonstrated success of CoT prompting in enhancing complex task processing for LLMs, we propose a novel framework that involves a reasoning chain called **ImageGen-CoT** prior to image generation. Our key insight is that explicitly **generating reasoning steps before image synthesis** helps unified MLLMs better understand multimodal contexts and produce more coherent outputs. However, these models often produce disorganized and incoherent reasoning chains, leading to suboptimal performance. To address these limitations, we first propose an automated dataset construction pipeline to generate ImageGen-CoT datasets, where each sample consists of a pair of reasoning chain and a corresponding image. The pipeline comprises three main stages: 1) collecting T2I-ICL instructions, 2) using MLLMs to generate step-by-step reasoning chains (ImageGen-CoT), and 3) producing image descriptions via MLLMs for diffusion models to generate images. To further enhance the dataset quality, we employ an iterative refinement process: The model first generates multiple text prompts and corresponding images, selects the best one, critiques the generated image, and iteratively refines the prompt until the max round is reached. **Then, we fine-tune the model using this dataset** which significantly enhances the image generation capabilities of unified-MLLMs in T2I-ICL tasks.

Despite the strong performance, T2I-ICL tasks' complexity leaves room for improvement. Inspired by NLP's Best-of-N paradigm, we explore **three test-time scaling strategies**: 1. Multi-Chain: Generate multiple reasoning chains, each producing one image; 2. Single-Chain: Create multiple image variants from one reasoning chain; 3. **Hybrid:** Combine both methods - multiple reasoning chains with multiple image variants per chain. Our empirical studies reveal two critical insights: (1) Instead of changing seeds, generating multiple ImageGen-CoTs via high-temperature LLM decoding achieves a similar performance to scaling. (2) **ImageGen-CoT enables bidirectional expansion**—either generating multiple instances of reasoning chains or modifying seeds to create diverse images—outperforming single-dimension scaling, opening new pathways for performance optimization in complex multimodal tasks.

To evaluate the effectiveness of our method, we experiment with leading Unified MLLMs. These models can be categorized into two types based on their visual representations: discrete visual tokens [8, 18, 22, 29, 34] and continuous visual embeddings [5, 15, 30, 35]. We select SEED-LLaMA [8] as a representative of the discrete approach and SEED-X [9] for the continuous approach, considering their open-source availability and support for interleaved text-image input. Extensive experiments demonstrate the effectiveness of our method. Specifically, as shown in Figure 2, SEED-X FT with ImageGen-CoT improves by 89% and 114% on CoBSAT and DreamBench++. With scaling strategy, it further achieves 0.909 and 0.543 respectively.

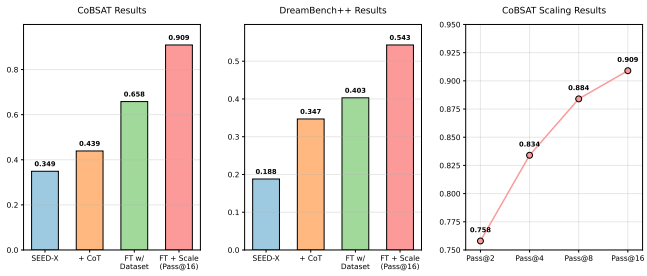


Figure 2. **Performance comparison on CoBSAT and DreamBench++ benchmarks.** Our method significantly improves SEED-X's performance through progressive enhancements: adding ImageGen-CoT, fine-tuning with the ImageGen-CoT dataset, and applying test-time scaling strategies.

Our contributions can be summarized as follows:

1. We propose a novel framework that generates a reasoning chain (called ImageGen-CoT) to enhance the performance of unified MLLMs in T2I-ICL tasks.

2. We construct high-quality ImageGen-CoT datasets for fine-tuning unified MLLMs through an automatic dataset construction pipeline.

3. We explore Best-of-N test-time scaling up paradigms and propose a hybrid scaling approach that first generates multiple ImageGen-CoT chains and then generates multiple image variations per chain.

2. Related Work

2.1. In-Context Learning

Large language models (LLMs) [3] have exhibited exceptional capabilities for text in-context learning (T-ICL). This ability allows LLMs to adapt to new tasks by observing a few illustrative examples provided as context, without requiring parameter updates. These models demonstrate remarkable performance on various tasks, with extension to multimodal models [1, 31, 37]. With the development of image generation, recent studies have proposed text-to-image in-context learning (T2I-ICL). For instance, CoBSAT [40] is the first benchmark designed to evaluate a model's T2I-ICL (Text-to-Image In-Context Learning) gen-

eration capabilities. This includes assessing the model's ability to rapidly adapt to tasks given the in-context demonstrations, which are key aspects of T2I-ICL. Emu2 [28] also evaluates models' T2I-ICL capabilities through subject customization in DreamBench [25], where the model needs to bind visual concepts from reference images to generate customized outputs. In this work, following previous studies, we validate our approach's improvement on the T2I-ICL task using CoBSAT and DreamBench++ [20].

2.2. Text-to-Image Generation

Text-to-Image (T2I) generation [21, 23, 24, 26, 39] aims to generate images based on a user's textual description. With the development of T2I diffusion models, such as DALL-E 3 [2], SD3 [6], and FLUX.1-Schnell [13], users can now generate high-quality and vivid images directly from text descriptions. Building on this success, there is an increasing demand for models to generate customized content, such as specific subjects, styles, or attributes tailored to individual user needs. Consequently, a variety of methods have emerged to address the challenge of subject-customized generation. These methods [7, 12, 19, 25, 27] typically rely on fine-tuning techniques, such as LoRA [11] or contrastive learning [10], to specialize a general T2I model for subject customization. However, these methods require the collection of subject-specific datasets and involve time-consuming retraining for each new user request. This makes them resource-intensive, limiting their ability to generalize quickly to new needs. To address these challenges, researchers [28] train EMU2 on multimodal sequences, leveraging its inherent ICL ability to quickly bind visual concepts from the context. Despite these advancements, their performances remain limited. In this work, we explore how introducing a reasoning chain prior to image generation, called ImageGen-CoT, can significantly enhance their performance on the T2I-ICL task.

2.3. Unified Multimodal Language Models

Recent years have witnessed remarkable progress in multimodal AI across two key domains: understanding and generation. In understanding, Large Vision-Language Models (LVLMs) [1, 4, 14, 16, 17, 32, 36, 38, 41] have achieved impressive capabilities in complex visual-textual reasoning tasks. Meanwhile, in generation, Text-to-Image diffusion models [2, 6, 13] have advanced to produce photorealistic images that can rival professional artists' work. Given these developments, researchers have been exploring ways to unify multimodal understanding and generation capabilities within a single model architecture. These models can be categorized into two approaches based on their visual representations: discrete visual tokens [8, 18, 22, 29, 34] and continuous visual embeddings [5, 15, 30, 35]. Discrete approaches leverage VQ-VAE to tokenize images into discrete tokens, enabling training and inference similar to language processing. In contrast, continuous approaches generate latent embeddings that are subsequently processed through diffusion models for image synthesis.

3. Method

In this section, we present our ImageGen-CoT framework in detail. First, we introduce the formulation of ImageGen-CoT. (Sec.3.1). Second, we describe our automated pipeline for collecting high-quality ImageGen-CoT datasets (Sec.3.2). Third, we provide a detailed formulation of the dataset and the loss function used to fine-tune the model with the collected dataset (Sec.3.3). Finally, we explore various strategies to enhance model performance during inference, culminating in a novel hybrid scaling approach that addresses both contextual comprehension and generation challenges (Sec.3.4).

3.1. Formulation of ImageGen-CoT

As described above, T2I-ICL tasks require models to have a high level of comprehension. To enhance the model's capacity, we propose a new framework that generates a reasoning chain, which we call ImageGen-CoT, before performing ImageGen. While we initially expected models to simultaneously output both ImageGen-CoT reasoning chains and corresponding images in a single forward pass.

However, during our practice, we observe that models frequently fail to generate images even when explicitly prompted to first generate ImageGen-CoT followed by image output. As illustrated in Figure 3, to ensure reliable image generation, we develop a two-stage inference protocol. The first stage involves prompting the model to generate the ImageGen-CoT reasoning chain R. In the second stage, we combine the original input X with the generated ImageGen-CoT R, along with a mandatory image generation token $\langle\text{image}\rangle$, to guarantee the production of the target image I. This process can be formally expressed as:

$$\begin{aligned} &\text{Stage 1: } R = \mathcal{M}(X \oplus \text{instruction}) \\ &\text{Stage 2: } I = \mathcal{M}(X \oplus R \oplus \langle\text{image}\rangle) \end{aligned} \tag{1}$$

where $\mathcal{M}$ denotes the unified MLLMs, and $\oplus$ represents the concatenation operation.

3.2. Dataset Construction

Due to the limitations of some unified MLLMs in generating well-structured ImageGen-CoT, which leads to suboptimal performance, we propose an automated pipeline to collect an ImageGen-CoT dataset and fine-tune the model using this dataset.

To collect high-quality ImageGen-CoT datasets, we first establish an instruction pool by collecting instructions from

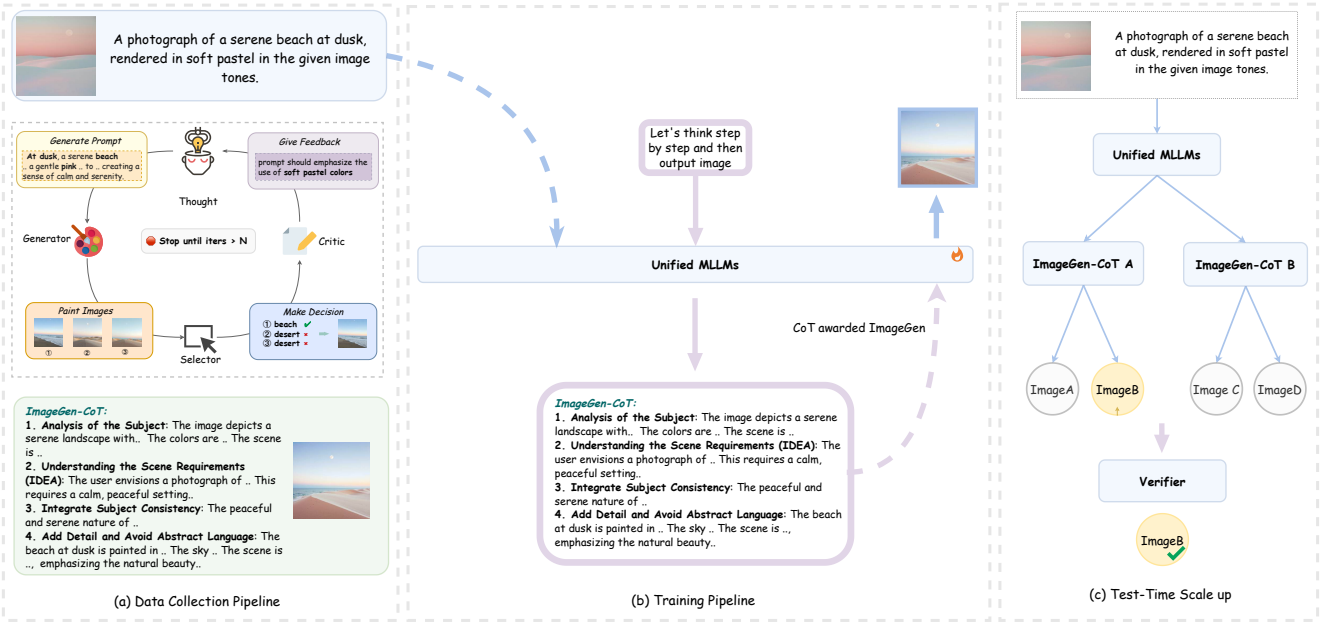


Figure 3. **Main Pipeline.** (a) **Data Collection Pipeline**: An automated iterative process where the MLLM acts as Generator, Selector, Critic, and Refiner to produce high-quality ImageGen-CoT (reasoning chains) and aligned images. (b) **Training Pipeline**: Fine-tuning unified MLLMs on the collected ImageGen-CoT dataset to enhance contextual reasoning and image generation. (c) **Test-Time Scaling**: Strategies for performance improvement via hybrid scaling during inference.

existing training datasets in T2I-ICL tasks. Second, we propose an automatic dataset construction pipeline. As illustrated in Figure 3, our pipeline proceeds as follows: In the initial stage, we let MLLM act as a **Generator** to generate N outputs, each consisting of an ImageGen-CoT and a prompt for the next image, which are then used by T2I-Model [13] to generate N images. Then, MLLM acts as a **Selector** to select the best image from the N candidates. After that, if the selected image meets our quality threshold or reaches the maximum iteration limit, the pipeline terminates and outputs the corresponding ImageGen-CoT and image pair. Otherwise, we let MLLM act as a **Critic** to write a critique of the selected image, assessing how well it matches the T2I-ICL prompt. Finally, MLLM acts as a **Refiner** to refine the prompt based on the critique, and the process iterates until meeting the termination. Finally, based on the collected responses, we construct our ImageGen-CoT dataset as follows:

$$\mathcal{D}_{\text{ImageGen-CoT}} = \{(T_i^*, I_i^*)\}_{i=1}^{n} \tag{2}$$

where (T_i^*, I_i^*) represents the i-th high-quality pair selected by our pipeline. T_i^* is the ImageGen-CoT that successfully guided the generation. I_i^* is the corresponding generated image that meets our quality standards. n is the total number of collected pairs in the dataset.

3.3. Training Pipeline

After constructing the dataset, we explore the training paradigm to fine-tune Unified MLLMs using our collected dataset. In this section, we detail the process of training Unified MLLMs with the ImageGen-CoT dataset, focusing on data formulation and training objectives.

To maintain consistency between the training and inference stages, we divide the ImageGen-CoT dataset into two splits:

(1) $[X, p_{\text{cot}}] \rightarrow$ ImageGen-CoT, which generates the ImageGen-CoT

(2) $[X, \text{ImageGen-CoT}, p_{\text{image}}] \rightarrow$ image, which generates the final image.

When training with the training dataset split 1, since model only generate the ImageGen-CoT text, we apply the normal lm_loss, formulated as follows:

$$lm_loss = -\frac{1}{N}\sum_{i=1}^{N} \log P(y_i \mid y_{<i}, X) \tag{3}$$

where y_i is the i-th token in the ImageGen-CoT text, $y_{<i}$ represents the preceding tokens, X is the input, and N is the total number of tokens in the ImageGen-CoT sequence.

For training with dataset split 2, there is no uniform training loss, as different Unified MLLMs utilize varying visual representations (e.g., discrete visual tokens [8, 34]

or continuous visual embeddings [9]). For models using discrete visual tokens, the same loss as language modeling (lm_loss) is applied. For models using continuous visual embeddings, we apply the mse_loss between the generated and target visual embeddings, formulated as:

$$mse_loss = \|\hat{z} - z\|^2 \tag{4}$$

where $\hat{z}$ is the generated visual embedding and z is the corresponding target visual embedding. In this study, our primary objective is to enhance the model's capability to generate accurate ImageGen-CoT. So, by default, we utilize data split 1 for fine-tuning Unified MLLMs, with more results presented in the Appendix.

3.4. Test time scale up

Though fine-tuning with the ImageGen-CoT dataset significantly improves model performance in T2I-ICL tasks, substantial room for improvement remains. Inspired by test-time scaling methods in NLP, we explore whether increasing computational investment during inference can further enhance performance. We first investigate a conventional paradigm: using SEED-X as the base model, generating multiple images by varying the seed value, and outputs are filtered via a ground-truth verifier aligned with the Pass@N metric. However, we observe that even with N=16, this approach underperforms compared to SEED-X fine-tuned with ImageGen-CoT Dataset.

This observation motivates our exploration of test-time scaling in the context of ImageGen-CoT, which we approach through three distinct strategies: 1. **Single-Chain Scaling:** This approach generates one reasoning chain and produces multiple image variants by varying the seed values. 2. **Multi-Chain Scaling:** Similar to NLP's "Best-of-N" sampling, we generate multiple reasoning chains through high-temperature LLM decoding. Each chain produces a unique image, potentially capturing different aspects of the contextual requirements. 3.**Hybrid Scaling:** Regarding the dual challenges of contextual comprehension and generation in T2I-ICL tasks, we propose a hybrid approach that combines the strengths of both strategies. As illustrated in Figure 3, this method first generates multiple reasoning chains and then creates multiple image variations for each chain. Our extensive experiments further reveal the effectiveness of this hybrid scaling strategy: the integration of ImageGen-CoT enables effective bidirectional scaling across both comprehension and generation dimensions. This dual-axis scalability opens new pathways for optimizing MLLM performance in complex multimodal tasks.

4. Experiments

4.1. Implementation details

To validate the effectiveness of our ImageGen-CoT framework and dataset, we conduct experiments on two T2I-ICL benchmarks: CoBSAT [40] and DreamBench++ [20]. We employ SEED-LLaMA [8] and SEED-X [9] as our base Unified MLLMs for both ImageGen-CoT reasoning and image generation. For the dataset construction pipeline, we utilize different configurations: on DreamBench++, InternVL2.5-78B-MPO-AWQ [33] serves as the Generator, Selector, Critic, and Refiner. When adapting to CoBSAT, we tested two selectors: self-select and self-consistency. The self-consistency selector performed better, so we chose it, while keeping the other components the same. However, the self-select method still achieved high scores on CoBSAT. This choice does not affect the generalization ability of our pipeline. FLUX.1-schnell [13] is selected as the base T2I model for both benchmarks. We maintain CoBSAT's original split strategy, while implementing an image-level split for DreamBench++ to ensure no subject overlap. During data construction, we generate 3 outputs per query using a sampling temperature of 0.7 and top-p of 0.8, with a maximum of 2 iterations. Additional details regarding computational cost, dataset splits, training procedures, and ablation studies are provided in the Appendix.

4.2. Main Results

In this section, we seek to answer the following questions: a) How much the ImageGen-CoT improves model's performance (via prompting)? b) To what extent does the performance of the model improve after fine-tuning with the ImageGen-CoT dataset? c) Can we invest more time in inference time to improve the performance? Finally, to better demonstrate the effectiveness of our method, we present visible comparison results.

Question 1: How much the ImageGen-CoT (via prompt) improves model's performance.

To verify the effectiveness of ImageGen-CoT (via prompt), we compare the model's performance with and without generating ImageGen-CoT before ImageGen via prompt on CoBSAT [40] and Dreambench++ [20]. Since CoBSAT includes 10 tasks, we calculate the average score to represent overall performance. Similarly, for Dreambench++, we compute the average score across its tasks.

Results As shown in Tables 1 and 2, integrating ImageGen-CoT through prompting yields consistent improvements across benchmarks. On CoBSAT, SEED-LLaMA's average score improves from 0.254 to 0.283 (+11.4% relative gain), while SEED-X shows a more substantial improvement from 0.349 to 0.439 (+25.8%). The trend persists on Dreambench++, where SEED-X achieves a 84.6% rel-

Table 1. **Main results on CoBSAT benchmark.** "FT w/ GT Image" denotes fine-tuning with ground truth images, while "FT w/ ImageGen-CoT" represents fine-tuning with our ImageGen-CoT dataset. The results demonstrate that ImageGen-CoT significantly improves model performance, with relative improvements over baseline model shown in red.

Method	Object-Inference Task					Attribute-Inference Task					Avg.↑
	Color-I	Bkg-I	Style-I	Action-I	Texture-I	Color-II	Bkg-II	Style-II	Action-II	Texture-II	
SEED-LLaMA	.616	.216	.272	.592	.112	.088	.168	.192	.220	.056	.254
+ ImageGen-CoT (via Prompt)	.700	.276	.300	.408	.084	.176	.292	.272	.192	.132	.283
+ FT w/ GT Image	.632	.272	.352	.540	.128	.164	.200	.256	.172	.112	.283
+ FT w/ ImageGen-CoT Dataset	.620	.368	.384	.424	.060	.192	.288	.208	.216	.148	**.291** ↑14.6%
SEED-X	.796	.412	.316	.596	.240	.176	.344	.260	.252	.104	.349
+ ImageGen-CoT (via Prompt)	.724	.440	.660	.784	.216	.312	.472	.228	.320	.240	.439
+ FT w/ GT Image	.936	.712	.896	.860	.468	.280	.324	.388	.636	.424	.592
+ FT w/ ImageGen-CoT Dataset	.884	.692	.928	.936	.420	.504	.612	.660	.524	.424	**.658** ↑88.5%

Table 2. **Evaluation results on Dreambench++ benchmark.** CP refers to concept preservation and PF refers to prompt following metrics. "FT" stands for fine-tuning. The relative gains over baseline model are shown in red.

Method	Concept Preservation					Prompt Following				CP·PF↑
	Animal	Human	Object	Style	Overall	Photorealistic	Style	Imaginative	Overall	
SEED-LLaMA	.436	.315	.288	.381	.358	.306	.202	.154	.218	.078
+ ImageGen-CoT (via Prompt)	.390	.241	.262	.346	.317	.291	.211	.170	.222	.078
+ FT w/ ImageGen-CoT Dataset	.399	.290	.271	.318	.325	.348	.355	.210	.310	**.101** ↑29.5%
SEED-X	.647	.420	.526	.571	.559	.346	.342	.303	.337	.188
+ ImageGen-CoT (via Prompt)	.547	.293	.369	.424	.427	.862	.775	.737	.817	.347
+ FT w/ ImageGen-CoT Dataset	.549	.410	.403	.432	.458	.922	.851	.846	.881	**.403** ↑114.4%

ative improvement(0.188 → 0.347) compared to its baseline. These results highlight the effectiveness of incorporating ImageGen-CoT in enhancing model performance. However, the SEED performance on Dreambench remains unchanged. This is attributed to its limited comprehension capabilities, which result in unreasonable and disorganized ImageGen-CoT outputs. To address this, we fine-tune the model using our collected ImageGen-CoT datasets, enabling higher-quality generation. More details are provided below.

Question 2: To what extent does the performance of the model improve after fine-tuning with the ImageGen-CoT dataset?

To further enhance model performance, we fine-tuned both SEED-LLaMA [8] and SEED-X [9] on the ImageGen-CoT dataset, which was collected using an automatic dataset construction pipeline. The ImageGen-CoT dataset consists of two components: the first part focuses on teaching the model how to generate ImageGen-CoT text, while the second part trains the model to generate images based on the generated ImageGen-CoT text. As described in sec.3.3, our primary goal is to improve the model's capabilities in generating high-quality ImageGen-CoT. To this end, we fine-tune the models using Part I of the ImageGen-CoT Dataset by default. We compare the performance of our fine-tuned model with its version using ImageGen-CoT (via prompt) and the standard version.

Results As shown in Table 1, SEED-LLaMA and SEED-X fine-tuned with ImageGen-CoT Dataset achieve improvements of +2.8% (0.283 → 0.291) and +49.9% (0.439 → 0.658), compared to generating ImageGen-CoT via prompting, respectively. What's more, they even outperform themselves fine-tuned with GT Images by +2.8% (0.283 → 0.291) and +11.1% (0.592 → 0.658). Additionally, on the Dreambench++ benchmark, SEED-LLaMA fine-tuned with ImageGen-CoT Dataset shows an improvement of +29.5% (0.078 → 0.101) in CP·PF score, while SEED-X achieves a +16.1% gain (0.347 → 0.403). These strong results on COBSAT and Dreambench++ underscore the effectiveness and generalizability of our collected ImageGen-CoT dataset in enhancing model reasoning and understanding abilities.

Question 3: Can we invest more time in inference time to improve the performance?

To further enhance model performance, we explore various test-time scaling strategies. we implement a Best-of-N approach where the model generates multiple image variations, with ground-truth metric evaluation (pass@N). As a baseline approach, we first experiment with the vanilla SEED-X model, generating multiple images by varying the seed values. We then investigate three advanced scaling strategies using SEED-X fine-tuned with ImageGen-CoT dataset: (1) Multi-Chain Scaling, which generates multiple distinct reasoning chains, with each chain producing an image; (2) Single-Chain Scaling, which produces mul-

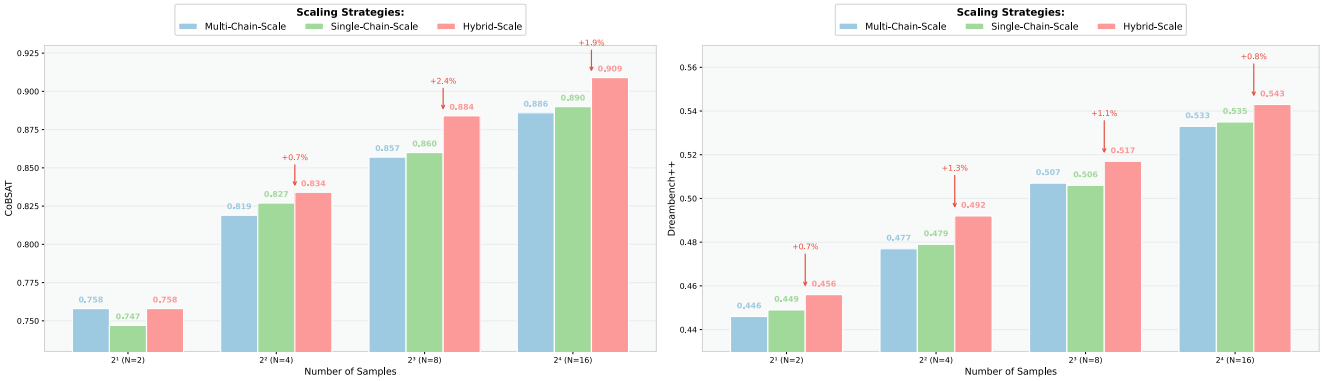


Figure 4. **Test-time scaling strategies comparison.** We conducted a comprehensive evaluation of three distinct scaling strategies: Multi-Chain Scaling, Single-Chain Scaling, and Hybrid Scaling, examining their performance across varying numbers of generated outputs (N=2,4,8,16). The experimental results are presented in two figures, with the left figure showing results on CoBSAT and the right figure displaying results on Dreambench++. The red numbers indicate the performance improvements achieved by Hybrid Scaling compared to Single-Chain Scaling.

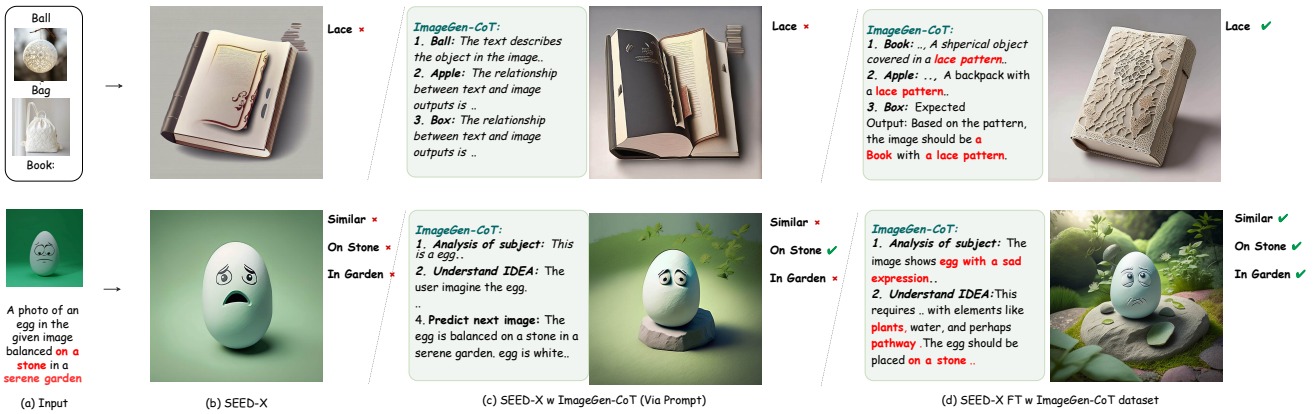


Figure 5. **Qualitative Results.** Comparison of generation results on COBSAT (top) and Dreambench+ (bottom) using baseline SEED-X, SEED-X with ImageGen-CoT prompting, and SEED-X fine-tuned with ImageGen-CoT dataset.

tiple image variations from a single reasoning chain; and (3) Hybrid Scaling, a novel approach that combines the strengths of both strategies by first generating multiple reasoning chains and then producing multiple image variations for each chain. For each paradigm, we systematically evaluate scalability by generating 2, 4, 8, and 16 outputs. For Hybrid Scaling, we implement specific configurations: Hybrid@16 uses 4 reasoning chains with 4 images per chain; Hybrid@8 explores two alternatives (2 chains × 4 images or 4 chains × 2 images); Hybrid@4 employs 2 chains × 2 images; and Hybrid@2 tests either 2 chains × 1 image or 1 chain × 2 images. Due to the significant scale difference, we visualize the latter strategy here.

Results

As shown in Figure 4, our experiments reveal three key insights. First, the Vanilla SEED-X@16 baseline (0.67 on CobSAT, 0.312 on Dreambench++) underperforms even the simplest scaling strategies (e.g., 0.747 on CobSAT@2), highlighting the necessity of ImageGen-CoT integration. Second, Multi-Chain Scaling matches Single-Chain Scaling in performance, proving that generating diverse reasoning paths is as effective as varying outputs from a single chain. Finally, Hybrid Scaling consistently achieves the highest scores across benchmarks. At N=16, Hybrid Scaling improves CobSAT performance to 0.909 (1.9% over Single-Chain) and Dreambench++ to 0.543 (0.8% higher than Single-Chain). The integration of ImageGen-CoT enables effective bidirectional scaling across both comprehension and generation dimensions. This dual-axis scalability suggests new pathways for optimizing MLLM performance in complex multimodal tasks.

Qualitative Results We further validate the effectiveness of our proposed methods through visualization. Figures 5 showcase the generation results from SEED-X under different configurations: baseline SEED-X, SEED-X with ImageGen-CoT (via prompting), and SEED-X fine-tuned

Table 3. We presents a comprehensive analysis of model performance on the COBSAT benchmark. Each model is evaluated in two generation modes: image generation (Img) and text generation (Txt).

Method	Gen Mode		Object Inference Task					Attribute Inference Task					Avg.↑
	Img	Txt	Color	Bkg	Style	Action	Texture	Color	Bkg	Style	Action	Texture	
SEED-X	✓	–	.796	.412	.316	.596	.240	.176	.344	.260	.252	.104	.349
	–	✓	.440	.388	.096	.080	.060	.116	.080	.180	.164	.132	.174
+ ImageGen-CoT (via Prompt)	✓	–	.724	.440	.660	.784	.216	.312	.472	.228	.320	.240	.439
	–	✓	.744	.212	.648	.476	.388	.356	.780	.292	.540	.136	.457
+ FT w/ ImageGen-CoT Dataset	✓	–	.884	.692	.928	.936	.420	.504	.612	.660	.524	.424	**.658** ↑88.5%
	–	✓	.984	.568	.968	1.00	.640	.516	.984	.592	.712	.628	**.760** ↑117.8%

with the ImageGen-CoT dataset. As shown in the top of Figure 5, baseline SEED-X (b) generates a basic book shape but misses the implicit "lace" attribute. With ImageGen-CoT prompting (c), the model's weak comprehension leads to poor ImageGen-CoT quality and even degraded generation quality. After fine-tuning with ImageGen-CoT dataset (d), with help of ImageGen-CoT, the model first successfully infers the shared attribute "lace" in CoT text and then generates the correct image - a book made of lace. Similarly, as shown in the bottom of Figure 5, baseline SEED-X (b) only generates a simple egg with an open mouth, ignoring key requirements like "on stone", "in garden", and similar expression (sad with upturned closed mouth). With ImageGen-CoT prompting (c), while the egg is placed on stone, it lacks both the required facial expression and garden environment. After fine-tuning (d), the model successfully understands all task requirements and generates a complete scene with an egg properly placed on a stone in a garden setting, maintaining similar facial features to the input. These qualitative results visually demonstrate the effectiveness of ImageGen-CoT and its corresponding dataset in enhancing model comprehension and generation capability, particularly in handling complex tasks that require attention to detail and scene understanding.

5. Further Analysis

5.1. The principles behind ImageGen-CoT contribute to enhancing the model's performance.

As described above, our proposed method, ImageGen-CoT, significantly enhances model performance on T2I-ICL tasks. To better understand why ImageGen-CoT improves performance, we hypothesize that 'a better understanding leads to better generation.' Specifically, we believe that ImageGen-CoT enhances the comprehension capabilities of Unified-MLLMs. To quantitatively assess the model's comprehension ability, we have the model generate a text description for the next image, as indicated by the 'Gen_mode' label in Table 3. Then we conduct a series of experiments to validate this hypothesis.

Results: The results in Table 3 demonstrate that integrating ImageGen-CoT significantly enhances model comprehension capabilities. When applied via prompting, SEED-X's text generation mode (Txt) achieves substantial gains, with the average score improving from 0.174 to 0.457. Fine-tuning with the ImageGen-CoT dataset further amplifies this advantage, elevating the text mode to a remarkable average score of 0.760 (vs. SEED-X's baseline of 0.174). Notably, enhanced comprehension also improves image generation (Img): SEED-X with ImageGen-CoT via prompt raises the average score from 0.349 to 0.439, while fine-tuning further boosts it to 0.658. This aligns with our hypothesis: "a better understanding leads to better generation."

5.2. Main obstacles in T2I-ICL

In this section we further discuss the main obstacles in T2I-ICL. We identify two primary challenges: First, as shown in Table 3, SEED-X's text generation mode (Txt) demonstrates relatively low performance scores (0.174), highlighting its difficulties in comprehending complex T2I-ICL instructions. Second, the image generation capabilities remain a bottleneck - notably, SEED-X fine-tuned with ImageGen-CoT dataset shows lower performance in image generation mode compared to text mode, indicating that while understanding may improve, translating this understanding into accurate image generation remains challenging.

6. Conclusion

In this work, we propose a novel framework to enhance Unified MLLMs' performance on T2I-ICL tasks by incorporating CoT reasoning before ImageGen. We also develop an automatic pipeline to curate high-quality ImageGen-CoT datasets and fine-tune models. Our experiments show significant improvements, with SEED-X achieving up to 80% gains after fine-tuning. Additionally, we explore test-time scaling and propose a hybrid approach combining multiple reasoning chains with diverse image generation, establishing a new paradigm for complex multimodal generation tasks.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 2, 3

[2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 3

[3] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 2

[4] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024. 3

[5] 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, 2, 3

[6] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 3

[7] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 3

[8] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*, 2023. 1, 2, 3, 4, 5, 6

[9] 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. *arXiv preprint arXiv:2404.14396*, 2024. 1, 2, 5, 6

[10] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3

[12] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 3

[13] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2023. 3, 4, 5

[14] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, Jianfeng Gao, et al. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision*, 16(1-2):1–214, 2024. 1, 3

[15] Hao Li, Changyao Tian, Jie Shao, Xizhou Zhu, Zhaokai Wang, Jinguo Zhu, Wenhan Dou, Xiaogang Wang, Hongsheng Li, Lewei Lu, et al. Synergen-vl: Towards synergistic image understanding and generation with vision experts and token folding. *arXiv preprint arXiv:2412.09604*, 2024. 2, 3

[16] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 3

[17] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 3

[18] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26439–26455, 2024. 1, 2, 3

[19] Jae Wan Park, Sang Hyun Park, Jun Young Koh, Junha Lee, and Min Song. Cat: Contrastive adapter training for personalized image generation. *arXiv preprint arXiv:2404.07554*, 2024. 3

[20] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. *arXiv preprint arXiv:2406.16855*, 2024. 3, 5

[21] 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. *arXiv preprint arXiv:2307.01952*, 2023. 3

[22] Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. *arXiv preprint arXiv:2412.03069*, 2024. 2, 3

[23] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[24] 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. 3

[25] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 3

[26] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 3

[27] Kihyuk Sohn, Lu Jiang, Jarred Barber, Kimin Lee, Nataniel Ruiz, Dilip Krishnan, Huiwen Chang, Yuanzhen Li, Irfan Essa, Michael Rubinstein, et al. Styledrop: Text-to-image synthesis of any style. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[28] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024. 1, 3

[29] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 2, 3

[30] Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*, 2024. 2, 3

[31] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021. 2

[32] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 3

[33] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024. 5

[34] 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. 1, 2, 3, 4

[35] 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. 2, 3

[36] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 3

[37] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3081–3089, 2022. 2

[38] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 2023. 3

[39] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022. 3

[40] Yuchen Zeng, Wonjun Kang, Yicong Chen, Hyung Il Koo, and Kangwook Lee. Can mllms perform text-to-image in-context learning? *arXiv preprint arXiv:2402.01293*, 2024. 2, 5

[41] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3