

# Can MLLMs Perform Text-to-Image In-Context Learning?

Yuchen Zeng<sup>\*1</sup>, Wonjun Kang<sup>\*2,3</sup>, Yicong Chen<sup>1</sup>, Hyung Il Koo<sup>2,4</sup>, and Kangwook Lee<sup>1</sup>

<sup>1</sup>University of Wisconsin-Madison    <sup>2</sup>FuriosaAI

<sup>3</sup>Seoul National University    <sup>4</sup>Ajou University

## Abstract

The evolution from Large Language Models (LLMs) to Multimodal Large Language Models (MLLMs) has spurred research into extending In-Context Learning (ICL) to its multimodal counterpart. Existing such studies have primarily concentrated on image-to-text ICL. However, the Text-to-Image ICL (T2I-ICL), with its unique characteristics and potential applications, remains underexplored. To address this gap, we formally define the task of T2I-ICL and present **CoBSAT**, the first T2I-ICL benchmark dataset, encompassing ten tasks. Utilizing our dataset to benchmark six state-of-the-art MLLMs, we uncover considerable difficulties MLLMs encounter in solving T2I-ICL. We identify the primary challenges as the inherent complexity of multimodality and image generation, and show that strategies such as fine-tuning and Chain-of-Thought prompting help to mitigate these difficulties, leading to notable improvements in performance. Our code and dataset are available at <https://github.com/UW-Madison-Lee-Lab/CoBSAT>.

## 1 Introduction

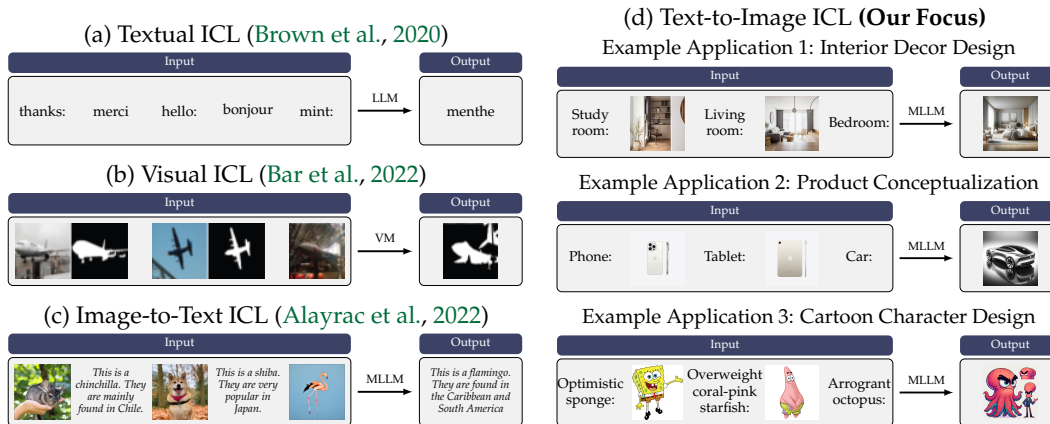


Figure 1: Comparison of various In-Context Learning (ICL) settings. (a) Textual ICL, where both the input and output in each example are textual. (b) Visual ICL, where both input and output in each demonstration are presented as images. (c) Image-to-Text ICL (I2T-ICL), featuring images as input and texts as output in each demonstration. (d) Text-to-Image ICL (T2I-ICL, our focus), which involves text input and image output in each demonstration. T2I-ICL introduces greater complexities and presents different potential applications. The examples in (d) provide three potential applications of T2I-ICL, with the output generated using ChatGPT-4 (OpenAI, 2023) with DALL-E 3 (Betker et al., 2023) capabilities.

\*Equal contribution. Emails: yzeng58@wisc.edu, kangwj1995@furiosa.ai

Object-Inference Tasks							Attribute-Inference Tasks								
		Latent Var	Prompt				Output			Latent Var	Prompt				Output
Colors-I	Car	White:		Blue:		Red:		Color-II	Black	Chair:		Cup:		Box:	
	Cup	Green:		Purple:		Orange:			Pink	Leaf:		Car:		Book:	
Background-I	Pig	Beach:		Desert:		Glacier:		Background-II	Volcano	Lion:		Zebra:		Tiger:	
	Zebra	Cave:		Space:		Waterfall:			Seafloor	Bird:		Pig:		Monkey:	
Style-I	Apple	Icon:		Lego:		Origami:		Style-II	Icon	Hat:		Chair:		Car:	
	Hat	Pixel:		Sketch:		Graffiti:			Sketch	Apple:		Ball:		Leaf:	
Action-I	Cat	Sing:		Read:		Swim:		Action-II	Sleep	Cat:		Cow:		Lion:	
	Dog	Run:		Sleep:		Fly:			Cry	Sheep:		Dog:		Bird:	
Texture-I	Ball	Metal:		Leather:		Wood:		Texture-II	Denim	Box:		Cup:		Apple:	
	Box	Wicker:		Plastic:		Paper:			Wood	Chair:		Ball:		Hat:	

Figure 2: **Overview of example prompts in the CoBSAT benchmark.** CoBSAT covers five themes: color, background, style, action, and texture, each with two different emphases: object-inference and attribute-inference. In object-inference tasks, the attribute (e.g., color) is directly provided in the textual input, and the model is required to infer the object (e.g., car) from the images. In other words, the latent variable (denoted as “Latent Var.” in the figure) of object-inference tasks is the object. Conversely, in attribute-inference tasks, the object is specified in the text. The model is tasked with inferring the attribute from the images in the demonstrations, i.e., the attribute serves as the latent variable in attribute-inference tasks.

In the rapidly evolving landscape of artificial intelligence, Multimodal Large Language Models (MLLMs) (Ge et al., 2023b; Koh et al., 2023; Sun et al., 2023c; OpenAI, 2023; Liu et al., 2023a; Bai et al., 2023b; Gemini Team Google: Anil et al., 2023; Li et al., 2023; Anthropic, 2024) extend the frontier of Large Language Models (LLMs) (Devlin et al., 2019; Radford et al., 2019; Brown et al., 2020; OpenAI, 2023; Touvron et al., 2023) by handling not only text but also images, videos, and audio. This multimodal capability enables MLLMs to undertake complex tasks, integrating visual, auditory, and textual cues. The versatility of MLLMs makes them powerful tools in AI, offering context-rich interpretations across various domains.

In-Context Learning (ICL) (see Figure 1(a)) is a prevalent technique that enables predictions based on context through a sequence of input-output pairs, termed *demonstrations*, without requiring any model parameter updates. This capability was initially identified and applied by Brown et al. (2020) and has since become a widely used standard prompt engineering method to enhance LLM inference performance for various downstream tasks. This method has been applied in computer vision to produce output images contextually aligned with

provided image-image pair examples, termed Visual ICL (V-ICL) (see Figure 1(b)) (Bar et al., 2022; Wang et al., 2023a). In another development, Tsimpoukelli et al. (2021) introduced Multimodal ICL (M-ICL) for the first time for image-to-text generation tasks, including applications such as visual question answering and image captioning. Unlike ICL, which is exclusively text-focused, and V-ICL, which is solely image-oriented, M-ICL uses demonstrations that incorporate samples from two modalities.

The majority of existing M-ICL work (Tsimpoukelli et al., 2021; Alayrac et al., 2022; Monajatipoor et al., 2023; Chen et al., 2023b; Zhao et al., 2023) has mainly centered on the performance of image-to-text tasks, the goal of which is transforming high-dimensional, image-based input into low-dimensional, text-based output. However, when the roles are reversed, models can exhibit significantly different performance characteristics. To distinguish between these two tasks, we refer to M-ICL for image-to-text generation as *Image-to-Text ICL* (I2T-ICL) (see Figure 1(c)), and M-ICL for text-to-image generation as *Text-to-Image ICL* (T2I-ICL) (see Figure 1(d)), with the latter being the focus of our work. It is important to note that potential applications of T2I-ICL are completely different from I2T-ICL, which include areas like product design and personalized content creation.

**Our Contributions.** We summarize our main contributions as follows.

- **Identifying an Important Problem: T2I-ICL.** Our work first identifies the important yet underexplored ICL setting on text-to-image generation, termed T2I-ICL.
- **Introducing the CoBSAT Benchmark.** To systematically assess the T2I-ICL capability of MLLMs, we introduce a comprehensive benchmark featuring ten tasks across five different themes — Color, Background, Style, Action, and Texture, which is named as CoBSAT (see Figure 2).
- **Benchmarking MLLMs in T2I-ICL.** We utilize our dataset to evaluate the T2I-ICL capabilities of ten state-of-the-art MLLMs. This includes Emu (Sun et al., 2023c), GILL (Koh et al., 2023), SEED-LLaMA (Ge et al., 2023b), Qwen-VL (Bai et al., 2023b), Gemini (Gemini Team Google: Anil et al., 2023), Claude (Anthropic, 2024), and GPT-4V (OpenAI, 2023), which are elaborated upon in the main paper, alongside Emu2 (Sun et al., 2023a), LLaVA-1.5 (Liu et al., 2023a), and LLaVA-NeXT (Liu et al., 2024), detailed in the appendix. We observe that the T2I-ICL performance of these models is significantly influenced by their respective training paradigms. Among them, SEED-LLaMA, Qwen-VL, Gemini, Claude, and GPT-4V demonstrate the capability to perform T2I-ICL. Yet, except for Gemini, their accuracy rates hover around or fall below 60% in most scenarios.
- **Understanding Challenges in T2I-ICL.** We then investigate the key factors contributing to the underperformance of MLLMs in T2I-ICL. Our findings point to two principal challenges: (i) the intrinsic complexity involved in processing multimodal data, and (ii) the inherent difficulties associated with the task of image generation.
- **Enhancing MLLMs’ T2I-ICL Capabilities.** To augment MLLMs’ T2I-ICL capabilities, we delve into various potential techniques. Our study demonstrates that fine-tuning and Chain-of-Thought (CoT) (Wei et al., 2022) significantly boost T2I-ICL performance.

## 2 Related Works

**Unimodal ICL.** Ever since Brown et al. (2020) demonstrated that language models are in-context learners (see Figure 1(a)), there has been substantial interest in comprehending this capability, both empirically (Liu et al., 2022; Min et al., 2022b; Chen et al., 2022; Mishra et al., 2022; Lampinen et al., 2022; Garg et al., 2022; Hendel et al., 2023) and theoretically (Xie et al., 2022; Wies et al., 2023; Akyürek et al., 2023; Von Oswald et al., 2023; Bai et al., 2023c; Ahn et al., 2023; Zhang et al., 2023b). Textual ICL (T-ICL) enables the adaptation of LLMs to downstream tasks simply by providing a few illustrative examples, bypassing any need for updating model parameters. The concept of V-ICL is then employed in computer vision, starting with the introduction of visual prompts (see Figure 1(b)). The pioneering works by Bar et al. (2022); Wang et al. (2023a) propose to automatically generate output images that are contextually aligned with provided examples. Specifically, Bar et al. (2022) developed

a method that combines three images - an example input, its corresponding output, and a query - into a single composite image. In this layout, the example input is placed in the upper left, the example output in the upper right, the query image in the bottom left, and the bottom right patch is left blank for output construction via an image inpainting model. Bar et al. (2022) demonstrated the effectiveness of V-ICL in tasks like edge detection, colorization, inpainting, etc. Unlike T-ICL and V-ICL which are limited to handling unimodal inputs, M-ICL integrates demonstrations encompassing both text and images.

**Image-to-Text ICL.** Most existing work on M-ICL focuses on image-to-text generation, i.e., I2T-ICL (Tsimpoukelli et al., 2021; Alayrac et al., 2022; Monajatipoor et al., 2023; Chen et al., 2023b; Zhao et al., 2023). In particular, Tsimpoukelli et al. (2021) were the first to extend ICL from the text domain to the multimodal domain, focusing on image-to-text generation such as visual question-answering (see Figure 1(c)). Alayrac et al. (2022) introduced Flamingo, an MLLM that achieves good performance in a variety of image and video understanding tasks using I2T-ICL with 32 demonstrations, implying the efficacy of I2T-ICL in performance enhancement in their model. Concurrently, efforts have been made to develop datasets specifically designed for evaluating the I2T-ICL capability of MLLMs (Zhao et al., 2023).

**Text-to-Image ICL.** There are limited attempts to evaluate MLLMs based on their T2I-ICL capabilities. A notable exception is concurrent research by Sun et al. (2023a). They evaluated the performance of their model on T2I-ICL with DreamBooth dataset (Ruiz et al., 2023). However, it is important to note that the DreamBooth dataset, primarily developed for fine-tuning models to modify image contexts, was not specifically designed for T2I-ICL applications, making it more challenging and mostly focusing on background altering. The complexity, as seen in style transfer examples that emulate artists like Vincent van Gogh or Michelangelo, can pose challenges even for human interpretation.

**MLLMs.** Recently, there has been a surge in the release of MLLMs, which are designed to address more challenging multimodal tasks, thereby enabling the perception of images, videos, and audios (Li et al., 2022; Alayrac et al., 2022; Hao et al., 2022; Laurençon et al., 2023; Huang et al., 2023b; Peng et al., 2023b; Li et al., 2023; Ge et al., 2023b; Koh et al., 2023; Zhu et al., 2023a; Sun et al., 2023c; Zheng et al., 2023a; OpenAI, 2023; Liu et al., 2023b;a; Bai et al., 2023b; Sun et al., 2023a; Driess et al., 2023; Gemini Team Google: Anil et al., 2023; Borsos et al., 2023; Huang et al., 2023a; Chen et al., 2023a; Zhang et al., 2023a; Anthropic, 2024).

Since our main focus is T2I-ICL, we only consider models capable of processing both text and multiple images. We consider two types of MLLMs: (i) proficient in generating both text and images, including Emu (Sun et al., 2023c), Emu2 (Sun et al., 2023a), GILL (Koh et al., 2023), and SEED-LLaMA (Ge et al., 2023b), and (ii) those limited to text generation, including GPT-4V (OpenAI, 2023), LLaVA-1.5 (Liu et al., 2023b), LLaVA-NeXT (Liu et al., 2024), Gemini (Gemini Team Google: Anil et al., 2023), Claude (Anthropic, 2024) and Qwen-VL (Bai et al., 2023b). For text-only MLLMs, we evaluate their capacity to infer visual outputs by prompting them to describe the anticipated image. Conversely, for MLLMs capable of image generation, we not only elicit image outputs but also ask for descriptive text, ensuring an apple-to-apple comparison with text-only models.

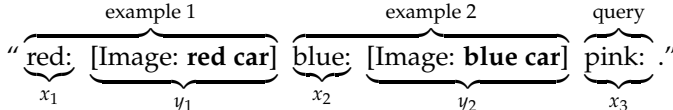
Owing to page constraints, we provide a more detailed overview of related works in Sec. B.

### 3 Dataset: CoBSAT

We start by describing the definition of in-context learning. Consider a task with data  $(x, y)$ , where input  $x \in \mathcal{X}$ , output  $y \sim f_\theta(x)$ , where distribution  $f_\theta$  is parameterized by latent variable  $\theta \in \Theta$ . We denote the model by  $M$ . For in-context demonstrations, we are given  $N$  input-output pairs  $\{(x_n, y_n)\}_{n=1}^N$  and one test query  $x_{N+1}$ . In-context learning make the prediction by incorporating these demonstrations  $\{(x_n, y_n)\}_{n=1}^N$  and the test query  $x_{N+1}$  in the prompt. The prediction made by model  $M$  is formulated as  $\hat{y}_{N+1} = M(x_1, y_1, x_2, y_2, \dots, x_N, y_N, x_{N+1})$ . In this work, we mainly focus on scenarios where the input  $x$  is textual data and output  $y$  corresponds to an image. We use notation

[Image: **description**] to denote an image corresponding to the text description. For instance, [Image: **red car**] refers to an image depicting a red car.

**Dataset Structure.** We begin by outlining the structure of our dataset, which evaluates whether models are capable of learning the mapping from textual input to visual output, based on the given in-context demonstrations. For instance, task Color-I in our experiment involves generating an image of an object of a particular color, where the object to be drawn is not explicitly stated in the text query  $x_{N+1}$ . The information of the object is instead implicitly contained in  $\theta$  (and hence in  $y_i$ 's since  $y_i \sim f_\theta(x_i)$  for all  $i = 1, \dots, N$ ), which can be learned from the demonstrations. An example prompt when  $\theta = \text{“car”}$  is



Ideally, MLLMs can learn the object  $\theta$  from the context, and generate an image of a pink car.

CoBSAT comprises ten tasks, divided into two categories: (i) *object-inference tasks*, which give the attributes (e.g., color, texture) in the text input and require identifying objects (e.g., car, cup) from images, and (ii) *attribute-inference tasks*, which provide the object to be drawn in the text input but require identifying the common attribute from the images (see Figure 2). Each task has predefined lists for text inputs and latent variables, denoted as  $\mathcal{X}$  and  $\Theta$ , each containing ten distinct items. For instance, in the Color-I task, the predefined list for the latent variable (i.e., the object) is  $\Theta = \{\text{leaf, hat, cup, chair, car, box, book, ball, bag, apple}\}$ , and the predefined list for the text input (i.e., the attribute) is  $\mathcal{X} = \{\text{yellow, white, red, purple, pink, orange, green, brown, blue, black}\}$ . The predefined lists for all tasks are provided in Sec. C. In our experiment, for each specified number of shots (i.e., 2, 4, 6, 8 in our experiments), we create 1,000 prompts per task. This is accomplished by randomly selecting a latent variable  $\theta$  from the predefined list  $\Theta$  and a sequence of textual inputs  $(x_n)_{n=1}^{N+1}$  from  $\mathcal{X}^{N+1}$ . Then, we pair each textual input  $x_n$  with the corresponding image  $y_n \sim f_\theta(x_n)$  to instruct in-context demonstrations.

**Data Collection.** For each task, we gather one image for every possible pairing of the textual input  $x \in \mathcal{X}$  and latent variable  $\theta \in \Theta$ , resulting in  $|\mathcal{X}| \times |\Theta| = 10 \times 10 = 100$  images for each task. For instance, for task Color-I, we collect an image of a red car to correspond to the case where  $x = \text{“red”}$  and  $\theta = \text{“car”}$ , and likewise for other images. It is noteworthy that the tasks with the same theme, such as Color-I (object-inference task) and Color-II (attribute-inference task), share the same images. In addition, all object lists and attribute lists, along with the images, are carefully selected so that LLaVA can correctly identify the specified objects and the corresponding attributes (i.e., color, background, texture, action, and style) of the images. This ensures an appropriate level of difficulty for T2I-ICL tasks and allows LLaVA to perform reliable evaluations on generated images. In total, we collect 500 images from the web and DALL-E 3 (Betker et al., 2023). We then construct in-context prompts for 2, 4, 6, and 8 shots as previously described, with each shot resulting in 10,000 prompts.

## 4 Methodology

**MLLMs.** In our study, we assess the performance of models in T2I-ICL, specifically Emu (Sun et al., 2023c), Emu2 (Sun et al., 2023a), SEED-LLaMA (Ge et al., 2023b), and GILL (Koh et al., 2023), which can generate images. In addition to image generation scenarios, we instruct the text-only generation models — Qwen-VL (Bai et al., 2023b), LLaVA-1.5 (Liu et al., 2023a), LLaVA-NeXT (Liu et al., 2024), Gemini (Gemini Team Google: Anil et al., 2023), Claude (Anthropic, 2024), and GPT-4V (OpenAI, 2023)), together with aforementioned models capable of generating images, to generate textual descriptions for expected images. This assesses if they learn the mapping from low-dimensional textual input to high-dimensional visual output based on the demonstrations. An extensive review of these

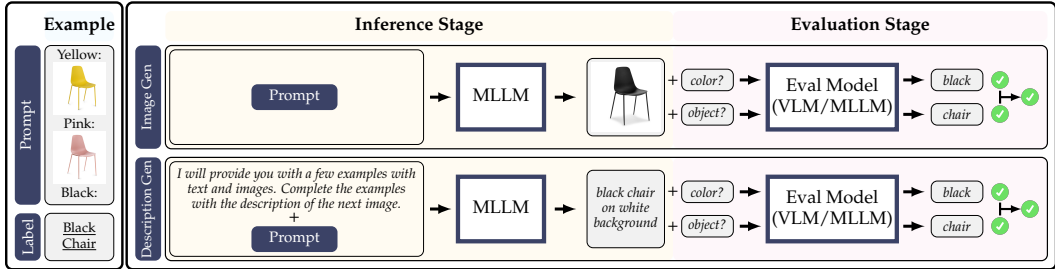


Figure 3: **Benchmarking pipeline for MLLMs in T2I-ICL with CoBSAT.** (i) For MLLMs with image generation capabilities, we feed prompts from our dataset into the MLLM under evaluation to prompt image generation. If the MLLM accurately interprets the text-image relationship in the provided demonstrations, it should produce an image of a “black chair.” To verify this alignment, we employ one evaluation model, it could be either a Vision-Language Model (VLM, e.g., CLIP) or an MLLM adept at visual question answering (e.g., LLaVA). This allows us to determine whether the generated image accurately corresponds to the target label. (ii) For MLLMs that do not generate images, we modify the process by instructing the MLLMs to describe the image textually, following the same evaluation criteria as in the image generation scenario.

MLLMs, and detailed information about the prompts used for each model, are provided in Sec. A and Sec. D.1, respectively.

In particular, since LLaVA models are primarily designed for visual question answering (Liu et al., 2023a; 2024) and are tailored to work with single-image inputs accompanied by questions, they do not perform well on T2I-ICL tasks as expected. Furthermore, Emu2 requires a significant amount of memory, especially for cases with a large number of demonstrations, which limits our ability to obtain comprehensive results due to resource constraints. Therefore, we defer the results of LLaVA models, as well as the partial results obtained for Emu2 in two-shot and four-shot cases, to Sec. F. In the main body of the paper, we primarily focus on discussing the other seven models.

**Evaluation.** Our evaluation pipeline is depicted in Figure 3, where we leverage both VLM and MLLM to assess whether the generated images or descriptions accurately represent the intended objects (e.g., “car” in the first example in Figure 2) and attributes (e.g., “red” in the same example). Specifically, we employ CLIP for its proficiency in vision-and-language tasks (Hessel et al., 2021; Ruiz et al., 2023), and MLLMs including LLaVA, Qwen-VL, and Gemini to determine the accuracy of the generated content. For CLIP’s evaluation, we identify the main object and attribute in the generated content by calculating the similarity between the embeddings of the generated content and the embeddings of all entries within our object and attribute lists. The items with the highest similarity are deemed the predicted labels. In the case of MLLMs, the generated content is embedded into the input, prompting MLLMs to identify the main object and attribute in the generated content, which are then assigned as the predicted labels. We then measure the accuracy of these predictions against the true labels to determine the correctness of the generated content.

In Sec. E, we compare these evaluation models in terms of alignment with human evaluation, and find Gemini > LLaVA-1.5 > CLIP > Qwen-VL in terms of alignment. Since Gemini is not open-sourced and there is a high correlation between the accuracies of LLaVA-1.5 and Gemini, we use free and open-sourced LLaVA-1.5 for all accuracy evaluations in our paper, unless otherwise stated. Additionally, we find that LLaVA-1.5 accurately identifies the correct object and attribute for all images in our dataset, ensuring the reliability of our evaluations. We provide more details such as prompts utilized for evaluation in Sec. D.2.

## 5 Benchmarking MLLMs in T2I-ICL

We visualize the T2I-ICL performance of the considered MLLMs in Figure 4.

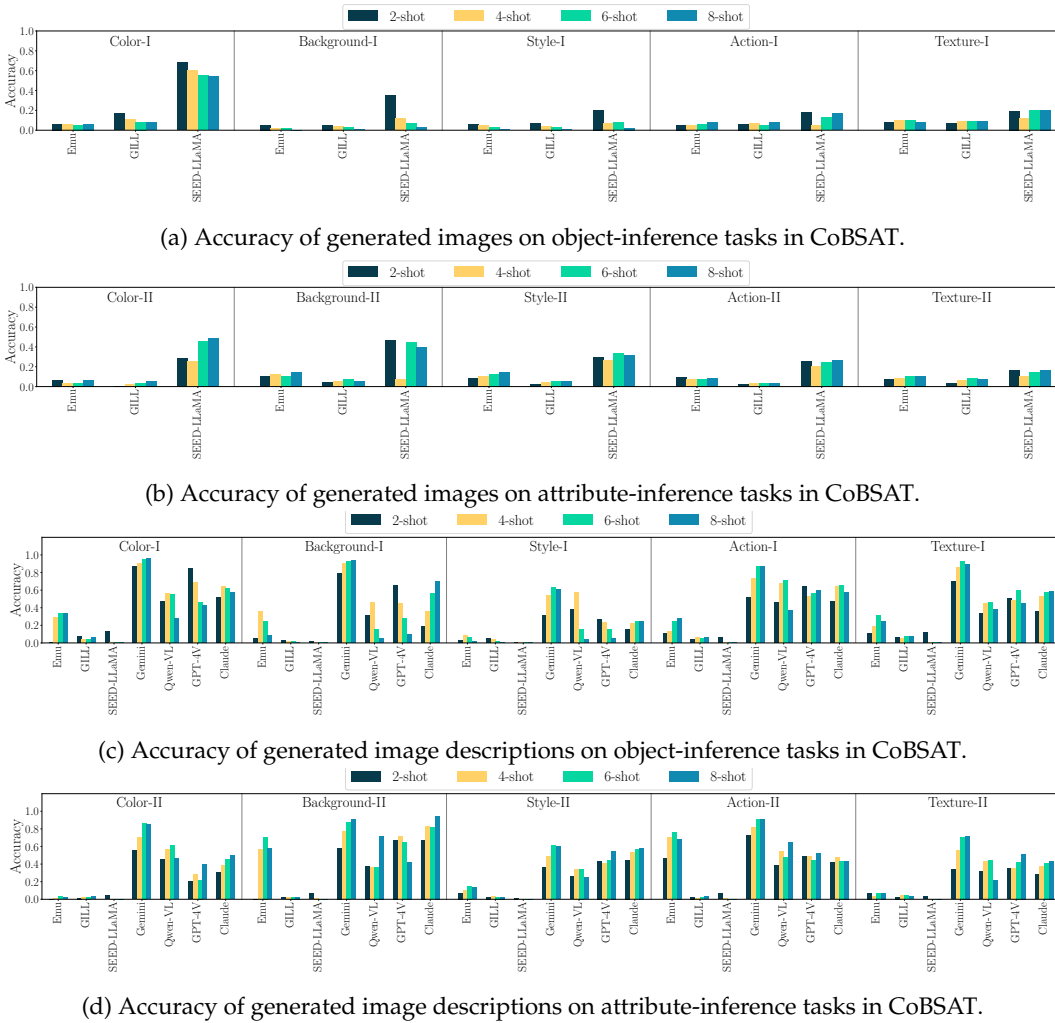


Figure 4: T2I-ICL performance of MLLMs on CoBSAT with 2,4,6,8 demonstrations.

**Assessing Generated Images.** In terms of image generation, we focus on the three MLLMs that have this capability: Emu, GILL, and SEED-LLaMA. Among these, SEED-LLaMA significantly outperforms the others, as evidenced by Figure 4(a) and (b), where it attains accuracies exceeding or nearing 20% across various tasks. Notably, on the Color-I task, SEED-LLaMA reaches an impressive 68% accuracy. In contrast, Emu and GILL exhibit low performance, achieving accuracies around or even below 10%.

GILL’s limited performance can be attributed to its training paradigm, which is not optimized for tasks requiring a unified understanding and generation of multimodal content (Ge et al., 2023b). Specifically, this limitation stems from its training that omits interleaved image-text data and the absence of an image generation model during its training process (Koh et al., 2023). In contrast, SEED-LLaMA benefits from instruction fine-tuning across a broad range of datasets, including both multimodal and text-to-image generation datasets such as Instructpix2pix (Brooks et al., 2023), MagicBrush (Zhang et al., 2024), JourneyDB (Sun et al., 2024), DiffusionDB (Wang et al., 2023c), LAION-Aesthetics (LAION, 2022), and VIST (Huang et al., 2016). Emu, on the other hand, has been fine-tuned exclusively on the LLaVA dataset (Liu et al., 2023b) in the context of image-text tasks. This expansive and varied instruction fine-tuning likely accounts for SEED-LLaMA’s enhanced performance in T2I-ICL tasks when compared to Emu.

Model	Shot	Method	Object-Inference Task					Attribute-Inference Task				
			Color-I	Background-I	Style-I	Action-I	Texture-I	Color-II	Background-II	Style-II	Action-II	Texture-II
Gemini	2	T2I-ICL	.865	.794	.315	.517	.704	.555	.583	.360	.725	.340
		T-ICL	<u>.979</u>	<u>.907</u>	<u>.692</u>	<u>.895</u>	<u>.764</u>	.150	.410	<u>.645</u>	<u>.468</u>	<u>.361</u>
	4	T2I-ICL	.904	.908	.540	.737	.861	.709	.773	.484	<u>.818</u>	.553
		T-ICL	<u>.988</u>	<u>.965</u>	<u>.888</u>	<u>.965</u>	<u>.927</u>	<u>.777</u>	<u>.780</u>	<u>.835</u>	.783	<u>.812</u>

Table 1: **Comparison of T2I-ICL v.s. T-ICL accuracy** (see Table 7 for the full version). To perform T-ICL on our dataset, we replace all images in the prompts with their corresponding descriptions. Underlined numbers indicate the highest accuracy achieved for each model and task across various shot numbers, while bold numbers indicate the highest accuracy for each specific combination of model, task, and shot count.

**Assessing Generated Image Descriptions.** Figures 4(c) and (d) reveal that Gemini, Qwen-VL, Claude, and GPT-4V stand out by significantly surpassing other MLLMs in most tasks. It is observed that MLLMs with image-generation capabilities often struggle with generating image descriptions. Among these leading models, Claude, Qwen-VL and GPT-4V show comparable results, whereas Gemini outperforms all of them. Given the lack of detailed information on the training datasets and paradigms for Gemini, Claude, and GPT-4V, our analysis can only extend to Qwen-VL. Notably, Qwen-VL benefits from pretraining on a broader dataset than Emu, GILL, and SEED-LLaMA, contributing to its enhanced performance (Bai et al., 2023b).

**Impact of Number of Demonstrations.** An interesting observation from Figure 4 is the lack of a consistent pattern in how performance is influenced by an increase in the number of demonstrations. For example, the accuracy in generating image descriptions for models such as Emu and Qwen-VL first increases and then decreases with an increasing number of demonstrations generally. Conversely, SEED-LLaMA’s accuracy first decreases and then increases. This non-monotonic performance trend with a growing number of demonstrations can potentially be attributed to two factors. Firstly, with a higher number of demonstrations, there may be an insufficient number of pertaining samples featuring the corresponding number of image inputs. Secondly, existing evidence indicates that an increase in demonstrations does not necessarily correlate with enhanced performance (Xie et al., 2022; Brown et al., 2020; Lin & Lee, 2024). Brown et al. (2020) demonstrate that for some datasets (e.g., LAMBADA, HellaSwag, PhysicalQA, RACE-m, CoQA/SAT analogies for smaller models), GPT-3’s zero-shot performance may surpass one-shot performance. Similarly, Xie et al. (2022) found that zero-shot scenarios can sometimes outperform few-shot ones, although performance tends to recover with the addition of more examples. Lin & Lee (2024) provided a theoretical explanation for this phenomenon by considering in-context learning as a process that involves both task retrieval and task learning.

We offer a more in-depth analysis in Sec. F.1, which delves further into the discussion above, and additionally (i) explores the impact of textual and visual information on predictions, (ii) investigates the performance of MLLMs in accurately generating the objects and attributes, respectively, and (iii) presents results for a more challenging variant of the CoBSAT benchmark.

## 6 Understanding Challenges in T2I-ICL

In Sec. 5, we observe that most MLLMs still face challenges in performing T2I-ICL effectively. Notably, SEED-LLaMA, Gemini, and Qwen-VL are notable free models, each capable of performing T2I-ICL tasks; SEED-LLaMA performs well for image generation scenarios, whereas Gemini and Qwen-VL specialize in image description generation scenarios. Therefore, unless otherwise stated, our subsequent investigations concentrate on these three models, specifically utilizing SEED-LLaMA for image generation scenarios and Gemini and Qwen-VL for image description generation.

In this section, our goal is to understand the main difficulties leading to this suboptimal performance in T2I-ICL. We hypothesize that the primary difficulties lie in (i) the complexity



Model	Shot	Precise Textual Inputs	Object-Inference Task					Attribute-Inference Task				
			Color-I	Background-I	Style-I	Action-I	Texture-I	Color-II	Background-II	Style-II	Action-II	Texture-II
SEED-LLaMA	0	✓	.730	<u>.456</u>	<u>.356</u>	<u>.264</u>	.275	.582	.314	.298	.207	<u>.286</u>
	2	✗ ✓	.680 <u>.801</u>	.348 <u>.409</u>	.203 <u>.241</u>	.182 <u>.192</u>	.196 <u>.326</u>	.287 <u>.385</u>	.467 <u>.485</u>	.297 <u>.393</u>	.261 <u>.317</u>	.163 <u>.268</u>
	4	✗ ✓	.482 <u>.669</u>	.211 <u>.318</u>	.141 <u>.284</u>	.053 <u>.161</u>	.122 <u>.286</u>	.252 <u>.608</u>	.076 <u>.441</u>	.268 <u>.299</u>	.207 <u>.278</u>	.105 <u>.248</u>

Table 2: **Accuracy comparison: with or without providing precise textual inputs** (see Table 8 for the full version). Bold numbers represent the highest accuracy for each task and shot count, comparing scenarios with and without descriptive textual inputs. Underlined numbers indicate the highest accuracy for each task across various shots.

Model	Shot	Fine-tuned	Object-Inference Task					Attribute-Inference Task				
			Color-I	Background-I	Style-I	Action-I	Texture-I	Color-II	Background-II	Style-II	Action-II	Texture-II
Qwen-VL	2	✗ ✓	.540 <u>.852</u>	.236 <u>.744</u>	.248 <u>.212</u>	.412 <u>.856</u>	.372 <u>.532</u>	.276 <u>.516</u>	.244 <u>.344</u>	.112 <u>.148</u>	.232 <u>.520</u>	.224 <u>.284</u>
	4	✗ ✓	.680 <u>.876</u>	.492 <u>.604</u>	<u>.448</u> <u>.216</u>	.228 <u>.812</u>	.556 <u>.588</u>	.512 <u>.696</u>	.448 <u>.308</u>	<u>.240</u> <u>.088</u>	.320 <u>.656</u>	.420 <u>.480</u>

Table 3: **T2I-ICL accuracy comparison of pretrained-only versus fine-tuned (FT) MLLM** (see Table 9 for the full version). Underlined numbers denote the highest performance achieved across different methods and shots for each task, while bold numbers indicate the top performance for each shot across various methods within their tasks.

inherent to multimodality, and (ii) the intrinsic challenges of the image generation task itself, which might be independent of the T2I-ICL process. We test these hypotheses as below.

**Is Multimodality a Primary Challenge in T2I-ICL?** The low performance of MLLMs in T2I-ICL is in contrast to the impressive results their underlying LLM demonstrated in T-ICL (Touvron et al., 2023; Bai et al., 2023a). To study whether multimodality is one primary challenge for T2I-ICL, we consider a textual version of our tasks by replacing every image in the prompts with corresponding detailed descriptions, which are initially created by LLaVA and ChatGPT and reviewed and updated by humans. Results in Table 1 show that T-ICL significantly improves the accuracy, especially in the 4-shot scenario. This improvement is also observed in the performance of Qwen-VL and SEED-LLaMA. For an in-depth exploration of the performance of Qwen-VL and SEED-LLaMA, detailed experimental settings, and comprehensive discussion, refer to Sec. F.2.1. These findings validate our hypothesis that multimodality is a principal challenge in T2I-ICL.

**Is the Image Generation a Primary Challenge in T2I-ICL?** We conduct an experiment with 0, 2, and 4-shot image generation tasks, with textual inputs updated as precise labels. For example, in the initial scenario from Figure 2, the terms “White,” “Blue,” and “Red” are updated to “White car,” “Blue car,” and “Red car,” respectively. The results, as shown in Table 2, reveal that even when precise textual inputs are provided, the accuracies of SEED-LLaMA remain below 50% in most scenarios, maintaining a similar relative performance across different tasks to scenarios without these inputs. This indicates that the task of image generation itself poses a significant challenge for current MLLMs, contributing to their underperformance on the CoBSAT dataset. Similar investigations with Emu and GILL yield consistent conclusions (see Sec. F.2.2).

## 7 Enhancing MLLMs’ T2I-ICL Capabilities

In the previous sections, we observed the suboptimal performance of MLLMs in executing T2I-ICL and investigated the primary challenges involved. This section delves into exploring techniques that could potentially enhance the performance of MLLMs in T2I-ICL. Additional details on our experiments, including choices of hyperparameters, prompt templates, results of other MLLMs, and other interesting technique explorations, are provided in Sec. F.3.

Model	Shot	CoT	Object-Inference Task				Attribute-Inference Task					
			Color-I	Background-I	Style-I	Action-I	Texture-I	Color-II	Background-II	Style-II	Action-II	Texture-II
SEED-LLaMA	2	✗	.680	<b>.348</b>	.203	<b>.182</b>	.196	<b>.287</b>	<b>.467</b>	<b>.297</b>	.261	<b>.163</b>
		✓	<b>.781</b>	.179	<b>.206</b>	.167	<u>.222</u>	.179	.389	.195	<b>.300</b>	.154
	4	✗	.482	.211	.141	.053	.122	.252	.076	.268	.207	.105
		✓	<b>.650</b>	<b>.353</b>	<b>.244</b>	<b>.242</b>	<b>.208</b>	<b>.303</b>	<b>.370</b>	<b>.335</b>	<b>.241</b>	<b>.171</b>

Table 4: **Accuracy comparison between T2I-ICL with vs. without CoT** (see Table 10 for the full version). Numbers in bold highlight the highest accuracy achieved for each model, number of shots, and task, and underlined numbers indicate the highest accuracy achieved for each model and task across different numbers of shots.

**Fine-tuning MLLMs on CoBSAT.** Building on the work of Min et al. (2022a), which demonstrates that tuning models on a collection of ICL tasks enables them to learn new tasks in context at test time, we fine-tune two instances of Qwen-VL, one on a 2-shot dataset and the other on a 4-shot dataset, and then compare their performances with their non-fine-tuned counterparts on the T2I-ICL test set. Note that all objects and attributes in the test set are not present in the training set. The results are summarized in Table 3. The results indicate a significant improvement in Qwen-VL’s T2I-ICL performance post fine-tuning. A similar trend is observed with SEED-LLaMA, as discussed in Sec. F.3.1. This suggests that fine-tuning MLLMs on a T2I-ICL dataset enhances T2I-ICL capability of MLLMs. Furthermore, a more challenging training-test dataset split is considered in Sec. F.3.1 to study the generalizability of the fine-tuned models in terms of T2I-ICL.

**Integrating Chain-of-Thought with T2I-ICL.** Another widely utilized method in prompt engineering is Chain-of-Thought (CoT) (Wei et al., 2022). This approach involves incorporating a simple instruction, such as “let’s think step by step,” prompting the model to sequentially generate concise sentences that outline the reasoning process, commonly referred to as reasoning chains or rationales. The chains are subsequently embedded into the subsequent prompt to obtain the final answer. In this experiment, we investigate the impact of integrating CoT on the T2I-ICL performance of MLLMs. The results are reported in Table 4. With the integration of CoT, SEED-LLaMA shows significant improvement in T2I-ICL performance across all ten tasks in the 4-shot scenario. Similar improvement is observed for Gemini, see Sec. F.3.2.

## 8 Conclusion and Future Works

In this work, we identify an important yet underexplored problem — T2I-ICL, and explore the capability of MLLMs to solve it. To facilitate this investigation, we introduce CoBSAT, a comprehensive benchmark dataset. Our experimental evaluation of MLLMs on this dataset reveals that many MLLMs have difficulty in effectively performing T2I-ICL. We identify two key challenges in T2I-ICL: (i) the integration and understanding of multimodal information; and (ii, particularly for image generation models) the actual process of image creation. To improve MLLMs’ performance in T2I-ICL, we carry out additional experimental studies, which suggest that fine-tuning and CoT can substantially enhance T2I-ICL capabilities.

As we identify T2I-ICL as an important problem for the first time, many interesting questions remain open. First, the impact of demonstration selection on T2I-ICL performance is yet to be fully understood. Furthermore, the application of other prevalent prompt engineering techniques to T2I-ICL remains open. While our dataset only covers basic themes, we identify expanding the themes of our dataset and extending it for image editing tasks as two interesting future directions. For a more in-depth discussion, please refer to Sec. G.

## Acknowledgement

The work of Kangwook Lee is supported in part by NSF CAREER Award CCF-2339978, Amazon Research Award, and a grant from FuriosaAI. We would like to express our

appreciation to Prof. Dimitris Papailiopoulos, Hanrong Ye, Changho Shin, Mu Cai, and anonymous reviewers for their insightful comments.

## References

- Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, et al. CM3: A causal masked multimodal model of the internet. *arXiv preprint arXiv:2201.07520*, 2022.
- Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. *arXiv preprint arXiv:2306.00297*, 2023.
- Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? Investigations with linear models. In *International Conference on Learning Representations*, 2023.
- 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. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022.
- Anthropic. The Claude 3 Model Family: Opus, Sonnet, Haiku. 2024. URL <https://api.semanticscholar.org/CorpusID:268232499>.
- Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. OpenFlamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023a.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 3(1), 2023b.
- Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *arXiv preprint arXiv:2306.04637*, 2023c.
- Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. *arXiv preprint arXiv:2312.00785*, 2023d.
- Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1728–1738, 2021.
- Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. In *Advances in Neural Information Processing Systems*, volume 35, pp. 25005–25017, 2022.
- 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>, 2023.
- Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. AudioLM: a language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

- Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.
- 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*, volume 33, pp. 1877–1901, 2020.
- Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Sae-hoon Kim. COYO-700M: Image-text pair dataset. <https://github.com/kakaobrain/coyo-dataset>, 2022.
- Mu Cai, Zeyi Huang, Yuheng Li, Haohan Wang, and Yong Jae Lee. Leveraging large language models for scalable vector graphics-driven image understanding. *arXiv preprint arXiv:2306.06094*, 2023.
- Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3558–3568, 2021.
- Qian Chen, Yunfei Chu, Zhifu Gao, Zerui Li, Kai Hu, Xiaohuan Zhou, Jin Xu, Ziyang Ma, Wen Wang, Siqi Zheng, et al. LauraGPT: Listen, attend, understand, and regenerate audio with GPT. *arXiv preprint arXiv:2310.04673*, 2023a.
- Shuo Chen, Zhen Han, Bailan He, Mark Buckley, Philip Torr, Volker Tresp, and Jindong Gu. Understanding and improving in-context learning on vision-language models. *arXiv preprint arXiv:2311.18021*, 2023b.
- Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.
- Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. Meta-learning via language model in-context tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 719–730, 2022.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.
- Tuan Dinh, Yuchen Zeng, Ruisu Zhang, Ziqian Lin, Michael Gira, Shashank Rajput, Jy-yong Sohn, Dimitris Papailiopoulos, and Kangwook Lee. LIFT: Language-interfaced fine-tuning for non-language machine learning tasks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 11763–11784, 2022.
- Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi. DreamLLM: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.
- Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An embodied multimodal language model. In *International Conference on Machine Learning*, volume 202, pp. 8469–8488, 2023.

- Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA: exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19358–19369, 2023.
- Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruva Ghosh, Jieyu Zhang, et al. DataComp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.
- Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? A case study of simple function classes. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30583–30598, 2022.
- Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. Planting a SEED of vision in large language model. *arXiv preprint arXiv:2307.08041*, 2023a.
- 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*, 2023b.
- Rohan Gemini Team Google: Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. *Int. J. Comput. Vis.*, 127:398–414, 2019.
- Yaru Hao, Haoyu Song, Li Dong, Shaohan Huang, Zewen Chi, Wenhui Wang, Shuming Ma, and Furu Wei. Language models are general-purpose interfaces. *arXiv preprint arXiv:2206.06336*, 2022.
- Roe Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 9318–9333, 2023.
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.
- 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. In *International Conference on Learning Representations*, 2022.
- Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *arXiv preprint arXiv:2303.11897*, 2023.
- Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. AudioGPT: Understanding and generating speech, music, sound, and talking head. *arXiv preprint arXiv:2304.12995*, 2023a.
- Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023b.

- Ting-Hao K. Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Aishwarya Agrawal, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *15th Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2016.
- Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6700–6709, 2019.
- Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. URL <https://doi.org/10.5281/zenodo.5143773>.
- Kushal Kafle, Brian L. Price, Scott Cohen, and Christopher Kanan. DVQA: understanding data visualizations via question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5648–5656, 2018.
- Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 787–798, 2014.
- Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Generating images with multimodal language models. *arXiv preprint arXiv:2305.17216*, 2023.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123:32–73, 2017.
- LAION. LAION-Aesthetics. <https://laion.ai/blog/laion-aesthetics>, 2022.
- Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. Can language models learn from explanations in context? In *Findings of the Association for Computational Linguistics (EMNLP)*, pp. 537–563, 2022.
- Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M Rush, Douwe Kiela, et al. OBELISC: An open web-scale filtered dataset of interleaved image-text documents. *arXiv preprint arXiv:2306.16527*, 2023.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, volume 162, pp. 12888–12900, 2022.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, volume 202, pp. 19730–19742, 2023.
- Ziqian Lin and Kangwook Lee. Dual operating modes of in-context learning. *arXiv preprint arXiv:2402.18819*, 2024.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023a.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023b.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. <https://llava-vl.github.io/blog/2024-01-30-llava-next/>, 2024.

- Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 100–114, 2022.
- Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11–20, 2016.
- Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for VQA on document images. In *IEEE Winter Conference on Applications of Computer Vision, WACV*, pp. 2199–2208, 2021.
- Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2791–2809, 2022a.
- Sewon Min, Xinxu Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, 2022b.
- Suvir Mirchandani, Fei Xia, Pete Florence, brian ichter, Danny Driess, Montserrat Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, and Andy Zeng. Large language models as general pattern machines. In *7th Annual Conference on Robot Learning*, 2023.
- Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. OCR-VQA: visual question answering by reading text in images. In *2019 International Conference on Document Analysis and Recognition, ICDAR*, pp. 947–952, 2019.
- Swaroop Mishra, Daniel Khoshdel, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. Reframing instructional prompts to GPT’s language. In *Findings of the Association for Computational Linguistics*, pp. 589–612, 2022.
- Masoud Monajatipoor, Liunian Harold Li, Mozhdeh Rouhsedaghat, Lin Yang, and Kai-Wei Chang. MetaVL: Transferring in-context learning ability from language models to vision-language models. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 495–508, 2023.
- Thao Nguyen, Yuheng Li, Utkarsh Ojha, and Yong Jae Lee. Visual instruction inversion: Image editing via visual prompting. *arXiv preprint arXiv:2307.14331*, 2023.
- OpenAI. Can i sell images i create with DALL·E? <https://help.openai.com/en/articles/6425277-can-i-sell-images-i-create-with-dall-e>, 2023.
- OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2Text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, volume 24, pp. 1143–1151, 2011.
- Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhui Chen, and Furu Wei. Kosmos-G: Generating images in context with multimodal large language models. *arXiv preprint arXiv:2310.02992*, 2023.
- Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11410–11420, 2022.

- Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023a.
- Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023b.
- 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.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, 2015.
- Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2655–2671, 2022.
- 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*, pp. 22500–22510, 2023.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: an open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 25278–25294, 2022a.
- Christoph Schuhmann, Andreas Köpf, Richard Vencu, Theo Coombes, and Romain Beaumont. LAION COCO: 600M synthetic captions from laion2b-en. <https://laion.ai/blog/laion-coco/>, 2022b.
- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2556–2565, 2018a.
- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018b.
- Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. Selective annotation makes language models better few-shot learners. In *International Conference on Learning Representations*, 2023.



- Keqiang Sun, Juntong Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. JourneyDB: A benchmark for generative image understanding. *Advances in Neural Information Processing Systems*, 36, 2024.
- Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiyong Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multi-modal models are in-context learners, 2023a.
- Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: Improved training techniques for CLIP at scale. *arXiv preprint arXiv:2303.15389*, 2023b.
- Quan Sun, Qiyong Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multi-modality. *arXiv preprint arXiv:2307.05222*, 2023c.
- Yanpeng Sun, Qiang Chen, Jian Wang, Jingdong Wang, and Zechao Li. Exploring effective factors for improving visual in-context learning. *arXiv preprint arXiv:2304.04748*, 2023d.
- 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.
- Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In *Advances in Neural Information Processing Systems*, volume 34, pp. 200–212, 2021.
- Unsplash Team. Unsplash dataset. <https://unsplash.com/data>, 2023. Accessed: 2024-01-30.
- Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6306–6315, 2017.
- Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, volume 202, pp. 35151–35174, 2023.
- Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6830–6839, 2023a.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023b.
- Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13: 600–612, 2004.
- Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 893–911, 2023c.
- 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. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.
- Noam Wies, Yoav Levine, and Amnon Shashua. The learnability of in-context learning. *arXiv preprint arXiv:2303.07895*, 2023.

- Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. Tree of Thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, 2023.
- Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*, 2023a.
- Qiyang Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Xinlong Wang, and Jingjing Liu. CapsFusion: Rethinking image-text data at scale. *arXiv preprint arXiv:2310.20550*, 2023b.
- Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. MERLOT RESERVE: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16375–16387, 2022.
- Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 15757–15773, 2023a.
- Kai Zhang, Lingbo Mo, Wenhui Chen, Huan Sun, and Yu Su. MagicBrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36, 2024.
- Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *arXiv preprint arXiv:2306.09927*, 2023b.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022a.
- Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. GPT-4V (ision) as a generalist evaluator for vision-language tasks. *arXiv preprint arXiv:2311.01361*, 2023c.
- Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. LlaVAR: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023d.
- Yiming Zhang, Shi Feng, and Chenhao Tan. Active example selection for in-context learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9134–9148, 2022b.
- Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. What makes good examples for visual in-context learning? *arXiv preprint arXiv:2301.13670*, 2023e.
- Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. MMICL: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915*, 2023.
- Kaizhi Zheng, Xuehai He, and Xin Eric Wang. MiniGPT-5: Interleaved vision-and-language generation via generative vokens. *arXiv preprint arXiv:2310.02239*, 2023a.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-judge with MT-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023b.

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*, 2023a.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal C4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*, 2023b.

# Appendix

<b>A In-Depth Overview of MLLMs</b>	<b>21</b>
<b>B Extended Related Works</b>	<b>23</b>
<b>C More Details of CoBSAT Dataset</b>	<b>25</b>
<b>D Detailed Experiment Setup</b>	<b>25</b>
D.1 Prompt Templates for Model Inference . . . . .	25
D.1.1 Instructing MLLMs for Generating Image Descriptions . . . . .	26
D.1.2 Articulating the Text-to-Image Relationship in Prompts . . . . .	27
D.2 Prompt Templates for Model Evaluation . . . . .	28
<b>E Comparison of T2I-ICL Evaluation Metrics</b>	<b>29</b>
<b>F Detailed and Extended Results of Experiments</b>	<b>29</b>
F.1 Benchmarking MLLMs in T2I-ICL (Detailed Version of Sec. 5) . . . . .	30
F.1.1 Assessing Generated Images . . . . .	30
F.1.2 Assessing Generated Image Descriptions . . . . .	32
F.1.3 Impact of Number of Demonstrations . . . . .	32
F.1.4 Textual Information v.s. Visual Information . . . . .	33
F.1.5 Object Generation v.s. Attribute Generation . . . . .	33
F.1.6 A Challenging Version of CoBSAT . . . . .	34
F.2 Understanding Challenges in T2I-ICL (Detailed Version of Sec. 6) . . . . .	35
F.2.1 Is Multimodality a Primary Challenge in T2I-ICL? . . . . .	35
F.2.2 Is the Image Generation a Primary Challenge in T2I-ICL? . . . . .	35
F.3 Enhancing MLLMs’ T2I-ICL Capabilities (Detailed Version of Sec. 7) . . . . .	36
F.3.1 Fine-tuning MLLMs on CoBSAT . . . . .	36
F.3.2 Integrating Chain-of-Thought with T2I-ICL . . . . .	37
F.3.3 Articulating the Text-to-Image Relationship in Prompts . . . . .	39
<b>G Extended Discussion</b>	<b>40</b>
G.1 Conclusion . . . . .	40
G.2 Limitations and Future Works . . . . .	40
<b>H Sample Outputs Generated by MLLMs</b>	<b>41</b>
H.1 Sample Prompts and Corresponding Outputs . . . . .	41
H.2 Sample Outputs from Fine-tuning SEED-LLaMA on CoBSAT . . . . .	42
H.3 Sample Outputs from Integrating CoT with T2I-ICL . . . . .	42

## A In-Depth Overview of MLLMs

In this section, we provide a detailed overview of the MLLMs used in our experiments, including (i) four MLLMs with image generation capabilities: Emu (Sun et al., 2023c), Emu2 (Sun et al., 2023a), SEED-LLaMA (Ge et al., 2023a;b), and GILL (Koh et al., 2023), and (ii) five state-of-the-art MLLMs that can only generate text: Qwen-VL (Bai et al., 2023b), LLaVA models (LLaVA-1.5 (Liu et al., 2023a) and LLaVA-NeXT (Liu et al., 2024)), Gemini (Gemini Team Google: Anil et al., 2023), and GPT-4V (OpenAI, 2023).

**Emu (Sun et al., 2023c).** Emu integrates EVA-CLIP (Fang et al., 2023) as the Visual Encoder, the Causal Transformer, LLaMA-13B (Touvron et al., 2023), and Stable Diffusion v1.5 as the Visual Decoder. Given any sequence including images and texts, the images are encoded into dense visual features via EVA-CLIP (Fang et al., 2023). These features are then transformed into visual causal embeddings via a Causal Transformer, which converts 2D spatial visual signals into 1D causal sequences. Two special image tokens, [IMG] and [/IMG], are prepended and appended to the visual causal embeddings of each image. The visual causal embeddings are then combined with the text tokens and fed into the LLaMA. In the output generated by LLaMA, the visual embeddings in-between image tokens [IMG] and [/IMG] are decoded using the fine-tuned Stable Diffusion 1.5. All components of Emu are further trained from their initial state using image-text pairs from LAION-2B (Schuhmann et al., 2022a) and LAION-COCO (Schuhmann et al., 2022b), video-text pairs from WebVid-10M (Bain et al., 2021), interleaved image and text from MMC4 (Zhu et al., 2023b), an expanded version of the text-only C4 (Raffel et al., 2020), and interleaved video and text from YT-Storyboard-1B (Zellers et al., 2022; Sun et al., 2023c). Furthermore, Emu can also process videos by treating various frames as a sequence interspersed with text and images.

**Emu2 (Sun et al., 2023a).** Emu2 represents a upscaled version of its predecessor, Emu, featuring significant upgrades in its component architecture. Unlike Emu, which utilized EVA-CLIP, LLaMA-13B, and Stable Diffusion v1.5 for its Visual Encoder, Multimodal Modeling, and Visual Decoder, respectively, Emu2 employs larger versions: EVA-02-CLIP-E-plus (Sun et al., 2023b) for the Visual Encoder, LLaMA-33B for Multimodal Modeling, and SDXL (Podell et al., 2023) as the Visual Decoder. Moreover, Emu2 replaced Emu’s C-Former with mean pooling followed by a linear projection for connecting Visual Encoder and Multimodal modeling. Its pretraining regime also differs, utilizing datasets that includes image-text pairs from LAION-2B (Schuhmann et al., 2022a) and CapsFusion-120M (Yu et al., 2023b), video-text pairs from WebVid-10M (Bain et al., 2021), interleaved image-text data from MMC4 (Zhu et al., 2023b), interleaved video-text data from YT-Storyboard-1B (Zellers et al., 2022; Sun et al., 2023c), grounded image-text pairs from GRIT-20M (Peng et al., 2023b) and CapsFusion-grounded-100M (Yu et al., 2023b), and language-focused data from Pile (Gao et al., 2020).

**SEED-LLaMA (Ge et al., 2023b).** SEED-LLaMA introduces a tokenizer named SEED, which consists of a ViT encoder (Dosovitskiy et al., 2021) derived from the pretrained BLIP-2 (Li et al., 2023), a Causal Q-Former, a VQ Codebook (van den Oord et al., 2017), a multi-layer perception, and a UNet decoder (Ronneberger et al., 2015) derived from the Stable Diffusion model. When given an input that includes both text and images, the images are first transformed into 2D raster-ordered features by the ViT encoder. These features are then converted into a sequence of causal semantic embeddings via the Causal Q-Former, discretized by the VQ Codebook, and projected by a multi-layer perceptron. The resulting embeddings are integrated with the text embeddings and fed into the LLaMA. The generated image embeddings are subsequently inputted into the Stable Diffusion model to generate realistic images. All components, except for the embedding layer, have been further trained on datasets including COCO Caption (Chen et al., 2015), CC3M (Sharma et al., 2018b), Unsplash (Unsplash Team, 2023), LAION-COCO (Schuhmann et al., 2022b), MMC4 (Zhu et al., 2023b), OBELISC (Laurençon et al., 2023), and WebVid (Bain et al., 2021). Additionally, 26 datasets are employed for supervised instruction tuning of SEED-LLaMA to align it with human instructions.

**GILL (Koh et al., 2023).** GILL employs a pretrained visual backbone and linear projection mapping to process image input, while a tokenizer is used for text input. These inputs are concatenated and fed into OPT-6.7B (Zhang et al., 2022a). The output image embeddings are then processed by a decision model to determine whether to retrieve real images or generate realistic fake ones. For generating realistic images, GILL proposes a GILLMapper, which encompasses a Transformer Encoder that receives image embeddings, and a Transformer Decoder that processes the Encoder’s outputs along with certain learned queries. The sequences produced by the Decoder are transformed through a linear layer to generate the predicted embeddings, which are then provided to the Stable Diffusion v1.5 model to create realistic images. For image retrieval, GILL projects the image embeddings via a linear layer and then measures the similarity between these embeddings and those of potential image candidates obtained through the CLIP ViT-L model (Radford et al., 2021). The image exhibiting the highest similarity score is then selected for output. GILL is pretrained on the CC3M dataset (Sharma et al., 2018b).

The three models previously mentioned are MLLMs capable of generating images. Next, we will describe MLLMs that can only generate text.

**Qwen-VL (Bai et al., 2023b).** Qwen-VL is an extension of the Qwen-7B language model (Bai et al., 2023a), equipped with visual capabilities. To achieve this, Qwen-VL incorporates a Vision Transformer (ViT) (Dosovitskiy et al., 2021) with weights initialized from OpenCLIP’s ViT-bigG (Ilharco et al., 2021), and a single-layer cross-attention module to convert images into a feature sequence that can be directly fed into Qwen-7B. Qwen-VL is pre-trained using (i) a variety of web-crawled image-text datasets, including LAION-5B, LAION-COCO (Schuhmann et al., 2022a), DataComp (Gadre et al., 2023), Coyo (Byeon et al., 2022), CC12M (Changpinyo et al., 2021), CC3M (Sharma et al., 2018a), SBU (Ordonez et al., 2011), COCO Caption (Chen et al., 2015), and in-house data (Bai et al., 2023b); and (ii) other visual question-answering datasets and visual reasoning datasets, including GQA (Hudson & Manning, 2019), VGQA (Krishna et al., 2017), VQAv2 (Goyal et al., 2019), DVQA (Kafle et al., 2018), OCR-VQA (Mishra et al., 2019), DocVQA (Mathew et al., 2021), GRIT (Peng et al., 2023a), Visual Genome (Krishna et al., 2017), RefCOCO (Kazemzadeh et al., 2014), RefCOCO+, and RefCOCOg (Mao et al., 2016).

**LLaVA (Liu et al., 2023a).** LLaVA is built upon the Vicuna-v1.5-13B LLM (Zheng et al., 2023b). To enable the visual perceiving capability, it incorporates a vision encoder, specifically the CLIP-ViT-L-336px (Radford et al., 2021), along with an MLP projection to encode visual features into image embeddings. These image embeddings, along with text embeddings encoded by tokenization, are then concatenated and fed into the LLM to generate the textual output. Its training follows a two-stage protocol. First, during the vision-language alignment pretraining stage, the model leverages the image-text pairs dataset CC3M (Sharma et al., 2018a) to align the visual features with the language model’s word embedding space. Second, the visual instruction tuning stage involves tuning the model on visual instructions to enable it to follow users’ diverse requests involving visual content. For this stage, LLaVA utilizes GPT-4V (OpenAI, 2023) to expand the existing COCO (Chen et al., 2015) bounding box and caption dataset into a multimodal instruction-following dataset, which includes three types of instruction-following data: conversational-style QA, detailed description, and complex reasoning. LLaVA-NeXT (Liu et al., 2024) is an improved version of LLaVA, particularly in reasoning, OCR, and world knowledge. It achieves this by increasing the input image resolution to capture more visual details and utilizing Mistral-7B and Nous-Hermes-2-Yi-34B as the additional backbones. Moreover, LLaVA-NeXT utilizes a better mixture of visual instruction tuning data, comprising high-quality user instructions and multimodal document/chart data.

**Claude (Anthropic, 2024).** Claude series is one of the leading LLMs developed by Anthropic. Anthropic recently introduced Claude 3, a family of MLLMs: Claude 3 Opus, Claude 3 Sonnet, and Claude 3 Haiku. Claude 3 can understand multimodal inputs such as photos, tables, and graphs. Besides multimodality, Claude 3 shows better fluency, especially for non-English languages. We chose Claude 3 Haiku for our experiment due to its speed and cost-effectiveness.

**Gemini (Gemini Team Google: Anil et al., 2023).** Gemini, a family of MLLMs developed by Google, is built on Transformer decoders and trained on extensive images, audio, video, and text datasets (including natural images, charts, screenshots, and PDFs). With a 32k context length support, it provides three variants: Ultra, Pro, and Nano, with Ultra offering the highest capabilities and Nano excelling in efficiency. We employ Gemini-pro in our paper.

**GPT-4V (OpenAI, 2023).** GPT-4V has emerged as one of the most proficient MLLMs, demonstrating exceptional performance and achieving human-level results on a majority of professional and academic examinations. Despite being a closed-source MLLM, with undisclosed details about its architecture and dataset construction, GPT-4V is included in our evaluation due to its superior performance compared to other MLLMs (Bai et al., 2023b).

## B Extended Related Works

This section provides detailed related works.

**Textual ICL.** Ever since Brown et al. (2020) demonstrated that language models are in-context learners (see Figure 1(a)), there has been substantial interest in comprehending this capability, both empirically (Liu et al., 2022; Min et al., 2022b; Chen et al., 2022; Mishra et al., 2022; Lampinen et al., 2022; Garg et al., 2022; Hendel et al., 2023) and theoretically (Xie et al., 2022; Wies et al., 2023; Akyürek et al., 2023; Von Oswald et al., 2023; Bai et al., 2023c; Ahn et al., 2023; Zhang et al., 2023b). Textual ICL (T-ICL) enables the adaptation of LLMs to downstream tasks simply by providing a few illustrative examples, bypassing any need for updating model parameters. The existing works indicate that LLMs possess the capability to comprehend context and perform reasoning through T-ICL (Brown et al., 2020).

**Visual ICL.** The concept of V-ICL is then employed in computer vision, starting with the introduction of visual prompts (see Figure 1(b)). The pioneering works by Bar et al. (2022); Wang et al. (2023a) propose to automatically generate output images that are contextually aligned with provided examples. Specifically, Bar et al. (2022) developed a method that combines three images - an example input, its corresponding output, and a query - into a single composite image. In this layout, the example input is placed in the upper left, the example output in the upper right, the query image in the bottom left, and the bottom right patch is left blank for output construction via an image inpainting model. Bar et al. (2022) demonstrated the effectiveness of V-ICL in tasks like edge detection, colorization, inpainting, segmentation, and style transfer. Wang et al. (2023a) introduced a similar approach and trained a generalist model named “Painter,” which exclusively uses visual prompts without any textual data for V-ICL. Experiments on standard computer vision benchmarks revealed competitive performance against task-specific models. Nguyen et al. (2023) further applied visual prompts to image editing by inverting visual prompts into text-based editing directions, leveraging the pre-trained capabilities of diffusion models.

A subsequent empirical study by Zhang et al. (2023e) highlighted that the success of V-ICL significantly depends on the choice of in-context demonstrations. The aspect of demonstration selection was further explored by Sun et al. (2023d), who also examined the impact of prompt fusion on performance. Their findings indicate a high sensitivity of performance to the arrangement of sub-images in in-context learning. Moreover, innovative approaches to structuring V-ICL, such as the concept of “visual sentences,” have been introduced in recent studies, notably by Bai et al. (2023d). Unlike V-ICL which only handles images, M-ICL integrates demonstrations encompassing both text and images.

**MLLMs.** In light of the significant success of LLMs, there has been an increase in the release of MLLMs. These models are designed to address more challenging multimodal tasks, thereby enabling the perception of images (Li et al., 2022; Alayrac et al., 2022; Hao et al., 2022; Laurençon et al., 2023; Huang et al., 2023b; Peng et al., 2023b; Li et al., 2023; Ge et al., 2023b; Koh et al., 2023; Zhu et al., 2023a; Sun et al., 2023c; Zheng et al., 2023a; OpenAI, 2023; Liu et al., 2023b;a; Bai et al., 2023b; Sun et al., 2023a; Gemini Team Google: Anil et al.,

2023; Driess et al., 2023; Anthropic, 2024), videos (Li et al., 2022; Alayrac et al., 2022; Li et al., 2023; Sun et al., 2023c; Gemini Team Google: Anil et al., 2023), and audio (Hao et al., 2022; Borsos et al., 2023; Huang et al., 2023a; Chen et al., 2023a; Zhang et al., 2023a; Gemini Team Google: Anil et al., 2023). Existing models capable of handling images can be categorized as follows: (i) those that use language as a general interface and directly employ LLMs without altering the model architectures (Dinh et al., 2022; Cai et al., 2023; Aghajanyan et al., 2022; Yu et al., 2023a; Huang et al., 2023b; Mirchandani et al., 2023; Cai et al., 2023); (ii) those that add one or more modules before feeding the input sequence into the LLM to perceive multimodal inputs (Tsimpoukelli et al., 2021; Alayrac et al., 2022; Awadalla et al., 2023; Laurençon et al., 2023; Li et al., 2023; Hao et al., 2022; Liu et al., 2023b;a; Zhu et al., 2023a; Gemini Team Google: Anil et al., 2023; Liu et al., 2024); (iii) those that add one or more modules after the LLM processing for generating multimodal outputs (Pan et al., 2023); (iv) those that add modules to both inputs and outputs of the LLMs to process the multimodal input and generate multimodal outputs (Dong et al., 2023; Sun et al., 2023c; Koh et al., 2023; Ge et al., 2023b; Zheng et al., 2023a; Sun et al., 2023a).

In this paper, our main focus is T2I-ICL. We aim to investigate whether MLLMs can learn to transform low-dimensional textual input into high-dimensional visual output based on demonstrations, and to accurately generate images from new textual queries. Consequently, we focus on models capable of processing both text and multiple images. We consider two types of MLLMs: (i) proficient in generating both text and images, including Emu (Sun et al., 2023c), Emu2 (Sun et al., 2023a), GILL (Koh et al., 2023), and SEED-LLaMA (Ge et al., 2023b), and (ii) those limited to text generation, including GPT-4V (OpenAI, 2023), Gemini (Gemini Team Google: Anil et al., 2023), Claude (Anthropic, 2024), LLaVA-1.5 (Liu et al., 2023a), LLaVA-NeXT (Liu et al., 2024), and Qwen-VL (Bai et al., 2023b). For text-only MLLMs, we evaluate their capacity to infer visual outputs by prompting them to describe the anticipated image. Conversely, for MLLMs capable of image generation, we not only elicit image outputs but also ask for descriptive text, ensuring an apple-to-apple comparison with text-only models.

**Image-to-Text ICL in MLLMs.** Most existing work on M-ICL focuses on the image-to-text generation, i.e., I2T-ICL, which involves mapping from high-dimensional input (i.e., images) to low-dimensional output (i.e., text). In particular, Tsimpoukelli et al. (2021) were the first to extend ICL from the text domain to the multimodal domain, focusing on image-to-text generation such as visual question-answering (see Figure 1(c)). Alayrac et al. (2022) introduced Flamingo, an MLLM that achieves state-of-the-art performance in a variety of image and video understanding tasks using I2T-ICL with 32 demonstrations, implying the efficacy of I2T-ICL in performance enhancement in their model. In contrast, Monajatipoor et al. (2023) explores whether the in-context capabilities of LLMs can be seamlessly extended to I2T-ICL by incorporating a visual encoder. Chen et al. (2023b) conducted a systematic study on the importance of visual and textual information in I2T-ICL. Concurrently, efforts have been made to develop datasets specifically designed for evaluating I2T-ICL capability of MLLMs (Zhao et al., 2023). In contrast, there have been only a few attempts (Sun et al., 2023a) to evaluate the T2I-ICL capability of MLLMs, a domain that remains relatively unexplored compared to its image-to-text counterpart.

**Zero-Shot Image Generation in MLLMs.** A relatively small number of MLLMs are capable of image generation (Yu et al., 2023a; Dong et al., 2023; Zheng et al., 2023a; Sun et al., 2023c; Ge et al., 2023b; Koh et al., 2023; Pan et al., 2023; Sun et al., 2023a). Zero-shot text-to-image generation typically generates images directly from textual descriptions without relying on any examples. This does not require the model to integrate a combination of textual and visual inputs. Another common task for MLLMs in image generation is context modifications. In this more complex scenario, the model receives visual inputs (e.g., an image of a dog) along with associated textual instructions (e.g., “swimming underwater”). This task requires a nuanced understanding and manipulation of the image, guided by the textual instructions, thereby blending image comprehension with contextual transformation based on text. Unlike zero-shot image generation, our focus is on studying whether MLLMs can learn the implicit relationship between the input and output from multiple in-context demonstrations.



**Text-to-Image ICL in MLLMs.** There are limited attempts to evaluate MLLMs based on their T2I-ICL capabilities. A notable exception is concurrent research by Sun et al. (2023a). They evaluated the performance of their model on T2I-ICL with DreamBooth dataset (Ruiz et al., 2023). However, it is important to note that the DreamBooth dataset, primarily developed for fine-tuning models to modify image contexts, was not specifically designed for T2I-ICL applications. This leads to certain constraints, such as its concentrated emphasis on altering backgrounds only and a level of complexity that may not align well with T2I-ICL. In contrast, our dataset spans five themes and provides well-designed prompts to assess whether models can understand both visual and textual information, learn mappings from demonstrations, and make inferences.

**Image Evaluation Metrics.** A variety of metrics exist for assessing the quality of generated images. Classical ones like Peak Signal-to-Noise Ratio (PSNR) (Wang et al., 2004) evaluate the quality of reconstructed images or videos by measuring pixel-level errors compared to the target images. Fréchet Inception Distance (FID) (Parmar et al., 2022) gauges the quality of images produced by generative models, such as Generative Adversarial Networks, by calculating the similarity between the distributions of generated and real images. However, these metrics are not entirely suitable for our purpose, where no single definitive ground-truth target image exists but rather a textual label (e.g., “red car” in the first example of Figure 2).

In the realm of text-to-image generation, the CLIP similarity (Radford et al., 2021) metric has gained popularity (Ruiz et al., 2023). It measures the cosine similarity between the CLIP embeddings of the textual ground truth and the visual output. Meanwhile, there is a growing trend of utilizing MLLMs for evaluation (Zhang et al., 2023c; Hu et al., 2023), showing promising results in text-to-image tasks. Our study both approaches, utilizing CLIP (Radford et al., 2021) and MLLMs including LLaVA-1.5 (Liu et al., 2023a), Gemini (Gemini Team Google: Anil et al., 2023), and Qwen-VL (Bai et al., 2023b) to assess the accuracy of generated images. To be more specific, we utilize CLIP and MLLMs to identify the object (e.g., “car”) and attribute (e.g., “red”) in the image generated by MLLMs and then compare these identifications with the actual label (e.g., “red car” for the first example in Figure 2). The details are provided in Sec. 4. Unless specified otherwise, the accuracy reported in our studies is primarily estimated using LLaVA-1.5, whose effectiveness has been validated by its ability to accurately recognize objects and attributes, achieving a 100% accuracy rate within our dataset, and closely aligning with human evaluation, as detailed in our analysis in Sec. E.

## C More Details of CoBSAT Dataset

**Detailed Structure.** The detailed structure of all tasks in our dataset is provided in Table 5.

**Copyright Considerations.** It is important to note that the images generated using DALL-E 3 for our dataset are not subject to copyright restrictions. As per the content policy and terms of the DALL-E 3 service, users retain ownership rights over the images they create, including the rights to reprint, sell, and merchandise, irrespective of whether the images were generated using free or paid credits (OpenAI, 2023).

## D Detailed Experiment Setup

In this section, we provide the details of our experiment setup, including prompt template design for model inference (Sec. D.1) and prompt design for model evaluation (Sec. D.2).

### D.1 Prompt Templates for Model Inference

For generating images based on in-context input-output pairs, we employ the prompt template depicted in Figure 3 for SEED-LLaMA and Emu. This template simply includes

Category	Task	Text Input $x \in \mathcal{X}$	Latent Variable $\theta \in \Theta$	Image Output $y \sim f_\theta(x)$
Object-Inference	Color-I	[Text: <b>color</b> $\in$ {yellow, white, red, purple, pink, orange, green, brown, blue, black}]	object $\in$ {leaf, hat, cup, chair, car, box, book, ball, bag, apple}	[Image: <b>object</b> $\theta$ of color $x$ ]
	Background-I	[Text: <b>background</b> $\in$ {beach, desert, glacier, volcano, park, gym, waterfall, space, cave, seafloor}]	animal $\in$ {zebra, tiger, sheep, pig, monkey, lion, dog, cow, cat, bird}	[Image: <b>animal</b> $\theta$ in background $x$ ]
	Style-I	[Text: <b>style</b> $\in$ {watercolor, sketch, pixel, origami, lego, icon, graffiti, futuristic, wireframe, old}]	object $\in$ {leaf, hat, cup, chair, car, box, book, ball, bag, apple}	[Image: <b>object</b> $\theta$ in style $x$ ]
	Action-I	[Text: <b>action</b> $\in$ {swim, sleep, sing, run, read, fly, eat, drink, cry, angry}]	animal $\in$ {zebra, tiger, sheep, pig, monkey, lion, dog, cow, cat, bird}	[Image: <b>animal</b> $\theta$ doing $x$ ]
	Texture-I	[Text: <b>texture</b> $\in$ {wood, wicker, sequined, plastic, paper, metal, leather, lace, denim, ceramic}]	object $\in$ {leaf, hat, cup, chair, car, box, book, ball, bag, apple}	[Image: <b>object</b> $\theta$ in texture $x$ ]
Attribute-Inference	Color-II	[Text: <b>object</b> $\in$ {leaf, hat, cup, chair, car, box, book, ball, bag, apple}]	color $\in$ {yellow, white, red, purple, pink, orange, green, brown, blue, black}	[Image: <b>object</b> $x$ of color $\theta$ ]
	Background-II	[Text: <b>animal</b> $\in$ {zebra, tiger, sheep, pig, monkey, lion, dog, cow, cat, bird}]	background $\in$ {beach, desert, glacier, volcano, park, gym, waterfall, space, cave, seafloor}	[Image: <b>animal</b> $x$ in background $\theta$ ]
	Style-II	[Text: <b>object</b> $\in$ {leaf, hat, cup, chair, car, box, book, ball, bag, apple}]	style $\in$ {watercolor, sketch, pixel, origami, lego, icon, graffiti, futuristic, wireframe, old}	[Image: <b>object</b> $x$ in style $\theta$ ]
	Action-II	[Text: <b>animal</b> $\in$ {zebra, tiger, sheep, pig, monkey, lion, dog, cow, cat, bird}]	action $\in$ {swim, sleep, sing, run, read, fly, eat, drink, cry, angry}	[Image: <b>animal</b> $x$ doing $\theta$ ]
	Texture-II	[Text: <b>object</b> $\in$ {leaf, hat, cup, chair, car, box, book, ball, bag, apple}]	texture $\in$ {wood, wicker, sequined, plastic, paper, metal, leather, lace, denim, ceramic}	[Image: <b>object</b> $x$ in texture $\theta$ ]

Table 5: **Task summary of CoBSAT.** We use [Text: **description**] to denote the text providing the corresponding description. For instance, [Text: **color**] could refer to terms such as “red” and “black.” Each task is characterized by the input space  $\mathcal{X}$ , and the latent variable space  $\Theta$ . For  $N$ -shot inference, we generate 1,000 prompts. Each prompt is obtained by randomly sampling  $\theta \in \Theta$  and  $(x_n)_{n=1}^{N+1} \in \mathcal{X}^{N+1}$ , followed by collecting the corresponding images  $(y_n)_{n=1}^N$ , where  $y_n \sim f_\theta(x_n)$ .

the in-context samples and the text query, without any additional instructions. For GILL, we add an additional system message: “You are a professional assistant who can generate a new image based on the sequence.”

In the subsequent subsections, we present our prompts for instructing MLLMs to generate image descriptions, continuing from the discussion in Sec. 4, and prompts for articulating the text-to-image relationship, continuing from the discussion in Sec. 7.

### D.1.1 Instructing MLLMs for Generating Image Descriptions

In this part, we provide the prompt templates used for instructing all considered models to generate image descriptions:

- **Emu:** We add the instruction as a system message: “Based on the sequence, describe the next image clearly, including attributes such as the main object, color, texture, background, action, style, if applicable.”
- **Emu2:** We append “Based on the sequence, describe the next image clearly, including details such as the main object, color, texture, background, action, style, if applicable.” to the end of the input.
- **GILL:** We insert “You are a professional assistant and always answer my question directly and perfectly without any excuses.” at the beginning of the prompt and append “Based on the sequence, describe what the next image should be clearly, including attributes such as the main object, color, texture, background, action, style, if applicable. Your response should only contain a description of the image, and any additional information can cause significant loss.” at the end of the input.

- **SEED-LLaMA:** We insert *“I will provide you a few examples with text and image. Complete the example with the description of next image. Tell me only the text prompt and I’ll use your entire answer as a direct input to A Dalle-3. Never say other explanations.”* at the beginning of the prompt.
- **LLaVA-1.5 & LLaVA-NeXT:** We add *“Based on the sequence, describe the next image to be generated clearly, including attributes such as the main object, color, texture, background, action, style, if applicable.”* at the end of the prompt.
- **Qwen-VL:** We insert *“You are a professional assistant and always answer my question directly and perfectly without any excuses.”* to the start of the prompt and append *“Based on the sequence, describe what the next image should be clearly, including attributes such as the main object, color, texture, background, action, style, if applicable. Your response should only contain a description of the image, and all other information can cause huge loss.”* to the end of the input.
- **Gemini:** We append *“Based on the sequence, describe the next image clearly, including details such as the main object, color, texture, background, action, style, if applicable.”* at the end of the prompt.
- **Claude:** We prepend *“I will provide you a few examples with text and image. Complete the example with the description of next image. Never say other explanations.”* to the beginning of the prompt, and append *“Give me the description of the your predicted next image.”* at the end of the prompt.
- **GPT-4V:** We add *“I will provide you with a few examples with text and images. Complete the example with the description of the next image. The description should be clear with main object, and include attributes such as color, texture, background, style, and action, if applicable. Tell me only the text prompt and I’ll use your entire answer as a direct input to A Dalle-3. Never say other explanations.”* at the start of the input.

### D.1.2 Articulating the Text-to-Image Relationship in Prompts

We now present the instructions for articulating the text-to-image relationship for the experiment presented in Sec. 7.

For image generation, we add the following sentences to the start of the prompts for each task.

- **Color-I:** *“Please identify the common main object in the images, and generate another image of this object of the requested color.”*
- **Color-II:** *“Please identify the common color in the images, and generate another image of the requested object in the same color.”*
- **Background-I:** *“Please identify the common animal in the images, and generate another image of this animal walking in the requested background.”*
- **Background-II:** *“Please identify the common background in the images, and generate another image of the requested animal walking in the same background.”*
- **Style-I:** *“Please identify the common object in the images, and generate another image of this object in the requested style.”*
- **Style-II:** *“Please identify the common style in the images, and generate another image of the requested object in the same style.”*
- **Action-I:** *“Please identify the common animal in the images, and generate another image of this animal doing the requested action.”*
- **Action-II:** *“Please identify the common action/mood the animal is doing in the images, and generate another image of the requested animal doing the same action/mood.”*
- **Texture-I:** *“Please identify the common main object in the images, and generate another image of this object of the requested texture.”*
- **Texture-II:** *“Please identify the common texture of the objects in the images, and generate another image of the requested object in the same texture.”*

For image description, we add the following sentences to the start of the prompts for each task.

- **Color-I:** *“Please identify the common main object in the images, and describe the next image to be generated based on the sequence below. Your description of the image should contain the description of the common main object and the requested color.”*
- **Color-II:** *“Please identify the common main color in the images, and describe the next image to be generated based on the sequence below. Your description of the image should contain the description of the requested object and the common color.”*
- **Background-I:** *“Please identify the common animal in the images, and describe the next image to be generated based on the sequence below. Your description of the image should contain the description of the common animal and the requested background.”*
- **Background-II:** *“Please identify the common background in the images, and describe the next image to be generated based on the sequence below. Your description of the image should contain the description of the requested animal and the common background.”*
- **Style-I:** *“Please identify the common object in the images, and describe the next image to be generated based on the sequence below. Your description of the image should contain the description of the common object and the requested style.”*
- **Style-II:** *“Please identify the common style in the images, and describe the next image to be generated based on the sequence below. Your description of the image should contain the description of the requested object and the common style.”*
- **Action-I:** *“Please identify the common animal in the images, and describe the next image to be generated based on the sequence below. Your description of the image should contain the description of the common animal and the requested action.”*
- **Action-II:** *“Please identify the common action/mood the animal is doing in the images, and describe the next image to be generated based on the sequence below. Your description of the image should contain the description of the requested animal and the common action/mood.”*
- **Texture-I:** *“Please identify the common main object in the images, and describe the next image to be generated based on the sequence below. Your description of the image should contain the description of the common main object and the requested texture.”*
- **Texture-II:** *“Please identify the common texture of the objects in the images, and describe the next image to be generated based on the sequence below. Your description of the image should contain the description of the requested object and the common texture.”*

## D.2 Prompt Templates for Model Evaluation

In this section, we present our prompt templates for model evaluation. The evaluation encompasses two scenarios: (i) assessing the generated images, and (ii) assessing the generated image descriptions.

**Assessing Generated Images.** Unless otherwise stated, we employ LLaVA-1.5 to evaluate the generated images in terms of whether they generated the right object (e.g., “car” in the first example in Figure 2) and attribute (e.g., “red” in the first example in Figure 2). To facilitate this evaluation, we design specific prompts for LLaVA. Here are the prompts designed for tasks Color-I and II:

- **Object Identification:** *“[Image: **generated image**] What is the main object in this image? Answer from the following options: (1)leaf (2)hat (3)cup (4)chair (5)car (6)box (7)book (8)ball (9)bag (10)apple. Answer the number only and do not include any other texts (e.g., 1).”*
- **Attribute Identification:** *“[Image: **generated image**] What is the color (of the main object) in this image? Answer from the following options: (1)yellow (2)white (3)red (4)purple (5)pink (6)orange (7)green (8)brown (9)blue (10)black. Answer the number only and do not include any other texts (e.g., 1).”*

For other tasks involving different themes, the options and the attribute category (e.g., replace “color” in the attribute inference prompt with “style” for tasks Style-I and II) are updated correspondingly.

**Assessing Generated Image Descriptions.** We also use LLaVA-1.5 to evaluate the generated image descriptions. However, in this case, we modify the prompts used for assessing generated images by replacing “[Image: **generated image**]” with “Image caption: [Text: **generated description**].”

## E Comparison of T2I-ICL Evaluation Metrics

In our experiments, we leverage LLaVA-1.5 for estimating the accuracy of the output of T2I-ICL. However, there are also many other alternatives such as CLIP, Gemini, and Qwen-VL. In this experiment, we study and compare the effectiveness of different models in terms of evaluating the performance of T2I-ICL.

**Evaluation Metrics.** This comparison focuses on the accuracy metrics derived from CLIP and MLLMs including Gemini, LLaVA-1.5, and Qwen-VL, with results gathered from SEED-LLaMA’s 2-shot T2I-ICL on CoBSAT. *MLLM accuracy* is determined by using MLLM to identify the main object and specific attribute (e.g., color) in the generated images or descriptions leveraging prompts provided in Sec. D.2, which are then matched against the true labels. *CLIP accuracy* is computed based on CLIP similarity. CLIP similarity measures the cosine similarity between the true label’s CLIP embedding and that of the generated content. CLIP accuracy involves selecting the most similar object and attribute from the predefined list based on their CLIP embedding’s cosine similarity with the generated image or description. These selections are then compared with the true labels to determine accuracy.

**Alignment of T2I-ICL Evaluation Metrics with Human Evaluation.** We first investigate their alignments with human evaluation. We manually labeled 100 images generated by SEED-LLaMA through T2I-ICL, selecting ten random images from each task to serve as a baseline. It is important to note that some images were of suboptimal quality, presenting ambiguities that could be interpreted both as correct or incorrect. Despite these difficulties, our evaluations using the LLaVA-1.5 show strong alignment with human assessments, achieving a consistency rate of 89% (computed by the ratio of agreement between the two methods). Notably, other MLLMs, especially Gemini, also exhibited commendable performance, as shown in Table 6.

Model	CLIP	LLaVA-1.5	Qwen-VL	Gemini
Consistency Rate to Human Evaluation	.85	.89	.78	.92

Table 6: Alignment between human evaluations and automatic evaluations performed by CLIP, LLaVA-1.5, Qwen-VL, and Gemini.

**Comparison among Evaluation Metrics.** We further conducted a scaled statistical study with 20,000 images to compare the performance of these automatic metrics, particularly focusing on how other metrics relate to Gemini’s results, given its closest alignment with human evaluations.

Figure 6, 7, and 8 depict the alignment between the accuracy estimates of Gemini and those provided by CLIP, Qwen-VL, and LLaVA-1.5, respectively. The analyses demonstrate a robust correlation between the accuracy estimates of LLaVA-1.5 and Gemini, highlighted by the narrow confidence interval represented by the purple shadow in the figures. This correlation strengthens our confidence in LLaVA-1.5 as a reliable and accessible open-source evaluation alternative to closed-source models in evaluating MLLMs’ T2I-ICL performance.

## F Detailed and Extended Results of Experiments

In this section, we supplement the experimental details, extended experiments, and discussions that could not be addressed in the main body due to space limitations. Specifically,

Sec F.1, F.2, and F.3 provide additional experiment results and discussions for Sec 5, 6, and 7, respectively.

## F.1 Benchmarking MLLMs in T2I-ICL (Detailed Version of Sec. 5)

This is an extended discussion of Section 5.

In this section, we present and analyze our experimental results on the evaluation of all the considered MLLMs’ performance on T2I-ICL, including the MLLMs that are not discussed in the main paper, i.e., LLaVA-1.5, LLaVa-NeXT, and Emu2. The full evaluation results are visualized in Figure 9. In addition to supplementing more detailed information on top of the main body, we also present a comparison of textual and visual information in Sec. F.1.4, and a comparison of object and attribute generation in Sec. F.1.5. Furthermore, we explore a more complex variation of our dataset, with detailed descriptions of the experiments and results presented in Section F.1.6.

### F.1.1 Assessing Generated Images

In terms of image generation, we focus on the four MLLMs that have this capability: Emu, Emu2, GILL, and SEED-LLaMA. Among these, SEED-LLaMA significantly outperforms the others, as evidenced by Figure 9(a) and (b), achieving a score of 68% on Color-I tasks. In contrast, Emu, Emu2, and GILL exhibit low performance, achieving accuracies around or even below 10%. For a more tangible understanding, we present specific prompts alongside the images generated using Emu, Emu2, GILL, and SEED-LLaMA in Figure 14, 15, 16, 17, and 18. We observe that while Emu, Emu2, and GILL exhibit low performance, GILL does manage to generate images that either align with the textual query (e.g., “pink” in the fourth example of Figure 14(a) or adhere to common visual patterns (e.g., “monkey” in the fourth example of Figure 15(a)). Conversely, Emu occasionally generates random images, as seen in the fourth example of Figure 14(a). On the other hand, Emu2’s generated images more closely resemble a blend of the input images in the prompt, such as the fifth example of Figure 16(b).

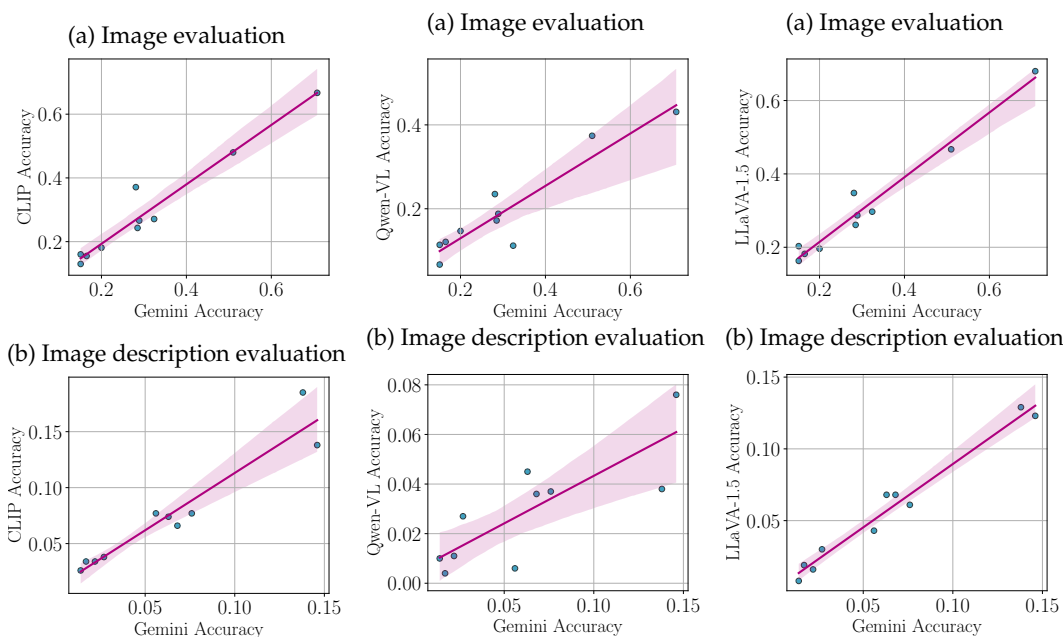
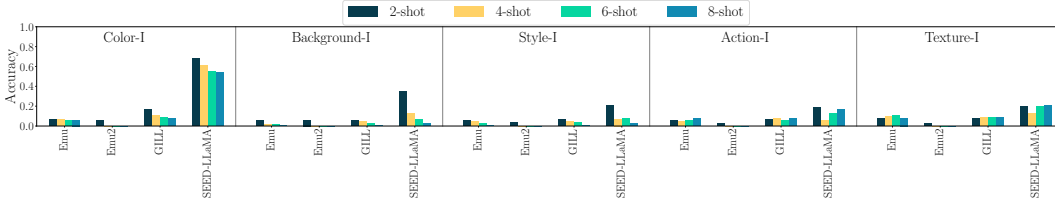


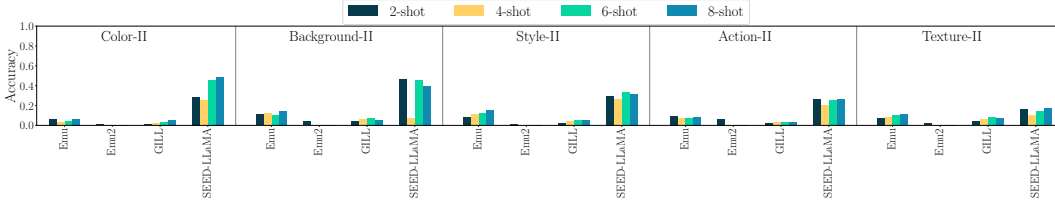
Figure 6: Accuracy estimated by CLIP versus accuracy estimated by Gemini.

Figure 7: Accuracy estimated by Qwen-VL versus accuracy estimated by Gemini.

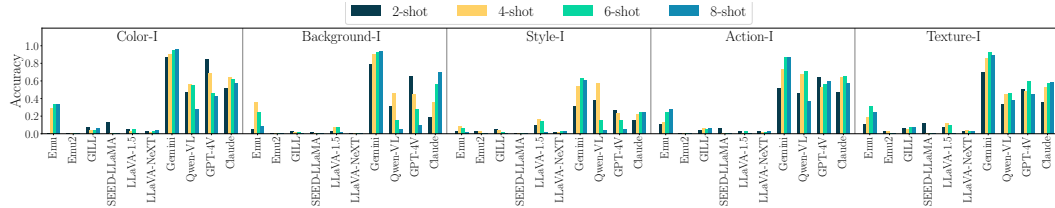
Figure 8: Accuracy estimated by LLaVA-1.5 versus accuracy estimated by Gemini.



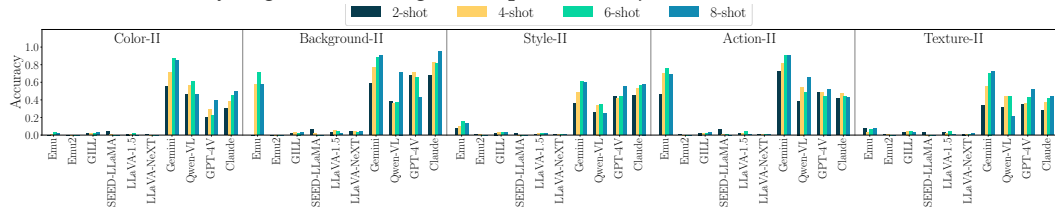
(a) Accuracy of generated images on object-inference tasks in CoBSAT.



(b) Accuracy of generated images on attribute-inference tasks in CoBSAT.



(c) Accuracy of generated image descriptions on object-inference tasks in CoBSAT.



(d) Accuracy of generated image descriptions on attribute-inference tasks in CoBSAT.

Figure 9: T2I-ICL performance of all evaluated MLLMs on the CoBSAT benchmark with 2,4,6,8 in-context demonstrations.<sup>1</sup>

GILL’s limited performance can be attributed to its training paradigm, which is not optimized for tasks requiring a unified understanding and generation of multimodal content (Ge et al., 2023b). Specifically, this limitation stems from its training that omits interleaved image-text data and the absence of an image generation model during its training process (Koh et al., 2023). Meanwhile, both Emu and Emu2 update all components in their model, and there is empirical evidence showing that they can better understand multimodal prompts. During instruction fine-tuning, Emu has been fine-tuned exclusively on the LLaVA dataset (Liu et al., 2023b) in the context of image-text tasks, while Emu2 is fine-tuned on more image-text pair data, including LLaVA and LLaVAR (Zhang et al., 2023d). In contrast, SEED-LLaMA benefits from instruction fine-tuning across a broad range of datasets, including both multimodal and text-to-image generation datasets such as Instructpix2pix (Brooks et al., 2023), MagicBrush (Zhang et al., 2024), JourneyDB (Sun et al., 2024), DiffusionDB (Wang et al., 2023c), LAION-Aesthetics (LAION, 2022), and VIST (Huang et al., 2016). This specific text-to-image generation dataset for instruction fine-tuning likely accounts for SEED-LLaMA’s enhanced performance in T2I-ICL tasks when compared to Emu and Emu2.

<sup>1</sup>Warning: it should be noted that Emu2 has results for 2 and 4-shot scenarios, but results for 6 and 8-shot scenarios are unavailable due to resource constraints, as Emu2 demands excessive memory.

### F1.2 Assessing Generated Image Descriptions

For image description generation, Figure 9(c) and (d) illustrate the performance of MLLMs in performing T2I-ICL for object-inference and attribute-inference tasks, respectively. We observe that Gemini, Qwen-VL, Claude, and GPT-4V stand out by significantly surpassing other MLLMs in most tasks. Among these leading models, Qwen-VL, Claude, and GPT-4V show comparable results, whereas Gemini outperforms them all.

To further investigate the performance of each model, we offer examples of prompts and their corresponding image descriptions, generated by MLLMs in Figure 19, 20, 21, 22, and 23. We observe that SEED-LLaMA and GILL often struggle to produce relevant textual output. GILL, tends to produce disjointed sentences like “person - bird on the beach - watercolor painting - watercolor,” as exemplified in Figure 20(a). SEED-LLaMA, on the other hand, predominantly generates images, defaulting to the text “I have generated an image,” regardless of varying instructions. Emu, Emu2, LLaVA-1.5, and LLaVA-NeXT all tend to describe the images contained in the prompt instead of making predictions. This is expected for LLaVA models since they are mostly trained for single image inputs with related questions and answers. Their primary function is to describe and answer questions related to the single image inputs rather than making predictions. In terms of image-text datasets, Emu and Emu2 are also instruction fine-tuned on LLaVA and LLaVAR datasets, thus sharing the same property as LLaVA models. However, they perform slightly better than LLaVA models. For instance, Emu makes the correct prediction in the second example in Figure 19(a). This improvement can be attributed to their pretraining on many other datasets, including interleaved image and text datasets such as Multimodal-C4 (Zhu et al., 2023b). For the leading models, which include Gemini, Claude, Qwen-VL, and GPT-4V, Qwen-VL is the only one that includes detailed information on the training datasets and paradigms. Notably, Qwen-VL benefits from pretraining on a broader dataset than Emu, GILL, SEED-LLaMA, LLaVA-1.5, and LLaVA-NeXT, contributing to its enhanced performance (Bai et al., 2023b).

### F1.3 Impact of Number of Demonstrations

In this part, we analyze how the number of demonstrations affects the performance of T2I-ICL. An interesting observation from Figure 4 is the lack of a consistent pattern in how performance is influenced by an increase in the number of demonstrations. For example, the accuracy in generating image descriptions for models such as Emu, Qwen-VL, and LLaVA initially increases and then decreases with an increasing number of demonstrations generally. Conversely, SEED-LLaMA’s accuracy first decreases and then increases.

This non-monotonic performance trend with a growing number of demonstrations can potentially be attributed to two factors. Firstly, with a higher number of demonstrations, there may be an insufficient number of pertaining samples featuring the corresponding number of image inputs. For example, LLaVA encounters a context length limitation when presented with eight image inputs, resulting in the model generating only empty strings in 8-shot cases. Secondly, existing evidence indicates that an increase in demonstrations does not necessarily correlate with enhanced performance (Xie et al., 2022; Brown et al., 2020). Brown et al. (2020) demonstrate that for some datasets (e.g., LAMBADA, HellaSwag, PhysicalQA, RACE-m, CoQA/SAT analogies for smaller models), GPT-3’s zero-shot performance may surpass one-shot performance. Similarly, Xie et al. (2022) found that zero-shot scenarios can sometimes outperform few-shot ones, although performance tends to recover with the addition of more examples. Xie et al. (2022) posit that an initial decrease in accuracy may be due to the distracting structure of prompts in such settings. Theoretical insights from Lin & Lee (2024) shed light on this phenomenon, suggesting that models initially rely on task retrieval and prior knowledge for predictions with a low number of demonstrations, shifting towards task learning as the number of demonstrations increases. The presence of a limited number of initial demonstrations might result in the retrieval of an incorrect task, potentially causing a decline in ICL performance. As more demonstrations are added, performance is anticipated to improve, as the model increasingly depends on task learning, which is improved by a greater number of demonstrations. However, in the MLLM scenario,



due to the scarcity of prompts with multiple images in the pretrained dataset, we do not anticipate observing this phenomenon.

#### F.1.4 Textual Information v.s. Visual Information

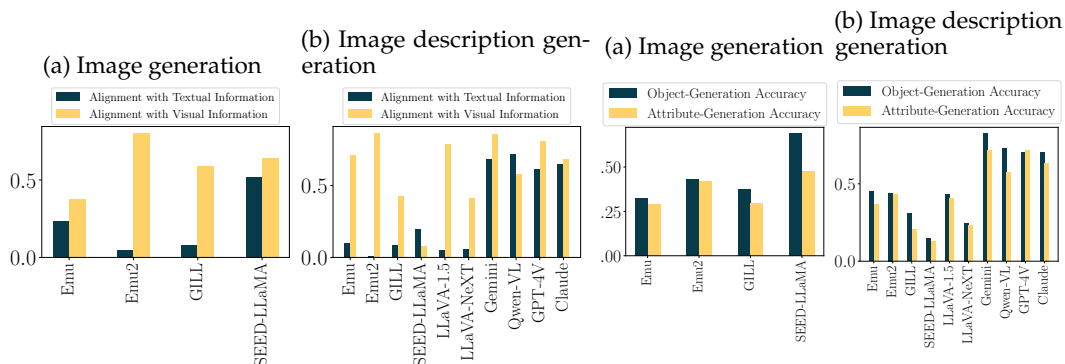


Figure 11: Comparison of alignment with textual information versus visual information when MLLMs perform two-shot T2I-ICL on the CoBSAT tasks.

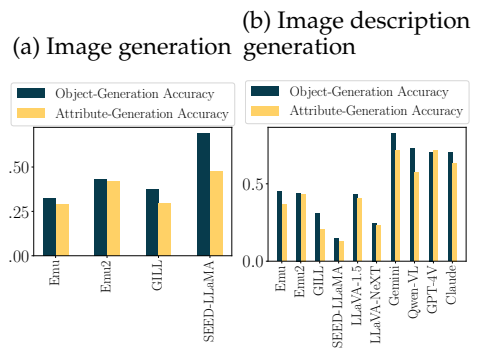


Figure 12: Comparison of object generation accuracy and attribute generation accuracy when MLLMs perform two-shot T2I-ICL on the CoBSAT tasks.

In this part, we investigate whether textual or visual information contributes more to the prediction of MLLMs. We assess this by evaluating how well the output of MLLMs aligns with both types of information. For textual alignment, we concentrate on how accurately the models generate images or image descriptions that match the given textual query. As an example, consider the scenario in Figure 2, where the text instructs “red.” In this context, we consider the output textually aligned if it features a red object. We employ a similar approach to measure the visual alignment of the outputs. Similarly, for visual alignment, we examine whether the generated images or their descriptions accurately incorporate elements from the images presented in the prompts. Taking the same example, an output is visually aligned if it correctly represents aspects like “car,” which is the common feature in the demonstration images. Employing these criteria allows us to determine which models are more influenced by textual queries and which lean toward visual cues.

Figure 11 reveals distinct patterns in how MLLMs respond to these inputs. Models such as Emu, Emu2, GILL, LLaVA-1.5, and LLaVA-NeXT demonstrate a marked reliance on visual information in their inputs. This aligns with our findings discussed in Sec. F.1.1 and F.1.2. As we discussed in Sec. F.1.1, GILL’s training exclusively on the CC3M dataset (Sharma et al., 2018b), an image-caption corpus, limits its predictive capabilities. For Emu, Emu2, LLaVA-1.5, and LLaVA-Next, they consistently generate descriptions of the images present in the prompt rather than predicting the next image based on the sequence, thus ignoring the textual query in the prompt. In contrast, the models that perform well, such as SEED-LLaMA for image generation and Qwen-VL, GPT-4V, Claude, and Gemini for image description generation, demonstrate a more balanced use of both textual and visual information compared to the other models.

#### F.1.5 Object Generation v.s. Attribute Generation

We are also interested in evaluating the proficiency of different MLLMs in inferring objects (e.g., car, chair) and attributes (e.g., color, style). As such, we report the accuracy of MLLMs in generating the correct objects and attributes. These results are depicted in Figure 12. Our observations reveal that all MLLMs perform better in generating correct objects, indicating that the task of generating accurate attributes presents a greater challenge compared to generating correct objects.

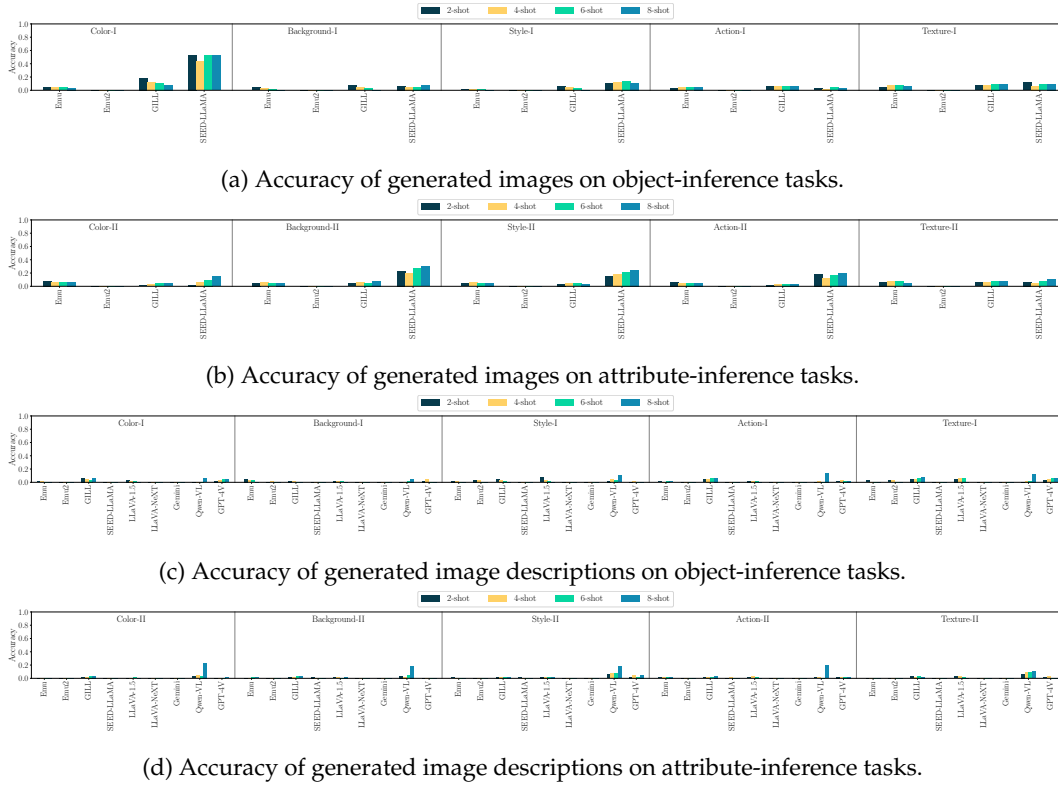


Figure 13: **Performance of considered MLLMs on the challenging version of the CoBSAT dataset with misleading information in the textual inputs.** We evaluate the T2I-ICL performance of various MLLMs with 2,4,6,8 in-context demonstrations. Low performance is observed across almost all evaluated MLLMs on this variant of our dataset, indicating a limited capacity of existing MLLMs in filtering out misleading information from prompts. <sup>2</sup>

### F.1.6 A Challenging Version of CoBSAT

We also investigate a more challenging task type that introduces misleading information into the textual input. This aims to evaluate whether MLLMs can accurately identify and ignore irrelevant information. Note that Claude is not considered in this experiment.

**Prompt Design.** To this end, we consider inputs of the form  $(x, \tilde{\theta})$  instead of just  $x$ , where  $\tilde{\theta} \in \Theta$  represents the misleading information that does not affect the output  $y$ . Here is an example with 4-shot input when  $\theta = \text{“car:”}$

“ $\underbrace{\text{red box: [Image: red car]}}_{x_1 \quad \tilde{\theta}_1 \quad y_1}$   $\underbrace{\text{blue chair: [Image: blue car]}}_{x_2 \quad \tilde{\theta}_2 \quad y_2}$   $\underbrace{\text{yellow leaf: [Image: yellow car]}}_{x_3 \quad \tilde{\theta}_3 \quad y_3}$   $\underbrace{\text{black book: [Image: black car]}}_{x_4 \quad \tilde{\theta}_4 \quad y_4}$   $\underbrace{\text{green bag:}}_{x_2 \quad \tilde{\theta}_2}$ ”

For prompt generation, we base it on the original prompt design, but with an added twist: for each prompt created with a sampled latent variable  $\theta$ , we introduce a misleading instruction for each example within the prompt. This is done by sampling the misleading information  $\tilde{\theta} \in \Theta / \{\theta\}$  without replacement, thereby adding an extra layer of complexity to the task. In the case of prompts with misleading textual inputs, for each prompt with a sampled  $\theta$ , we further sample  $\tilde{\theta} \in \Theta / \{\theta\}$  without replacement for each example.

<sup>2</sup>Warning: it should be noted that Emu2 has results for 2 and 4-shot scenarios, but results for 6 and 8-shot scenarios are unavailable due to resource constraints, as Emu2 demands excessive memory.

**Results.** Figure 13 illustrates the performance of MLLMs on this challenging version of the CoBSAT dataset. We note a significantly poor performance across all MLLMs, with the exception of SEED-LLaMA in image generation. However, even SEED-LLaMA’s performance shows a decline in most tasks compared to those in the CoBSAT dataset, which is visualized in Figure 4. These results suggest that current MLLMs also heavily rely on textual instructions and struggle to filter out misleading information. We anticipate this to be a challenging task for future MLLMs to overcome.

## F.2 Understanding Challenges in T2I-ICL (Detailed Version of Sec. 6)

In this section, we add more comprehensive details related to the experiments in Sec 6, to better understand the challenges in T2I-ICL. Our further experiments will mainly explore SEED-LLaMA, Gemini, and Qwen-VL. Previously, we identified SEED-LLaMA as the leading free model for image generation, whereas Gemini and Qwen-VL excel as the top free models for image description generation scenarios.

### F.2.1 Is Multimodality a Primary Challenge in T2I-ICL?

In Sec. 5, we find that SEED-LLaMA and Qwen-VL achieve only around or less than 50% accuracy on most tasks. This is in contrast to the impressive results their underlying LLM demonstrates in Textual ICL (T-ICL) (Touvron et al., 2023; Bai et al., 2023a). In this experiment, our objective is to determine whether multimodality is the primary cause of this reduced performance, or whether MLLMs intrinsically struggle with these tasks even for T-ICL.

**Prompt Design.** In this part, we evaluate SEED-LLaMA’s capability in image generation and Qwen-VL and Gemini’s proficiency in image description generation by modifying the prompts to be entirely textual. We achieve this by replacing every image in the prompts with corresponding detailed descriptions, which are initially created by LLaVA and ChatGPT. These descriptions are then reviewed and corrected by humans to ensure their accuracy. For example, in the first example depicted in Figure 2, the image [Image: red car] is replaced with a descriptive text: *“The image portrays a red Volkswagen Golf R, a compact sports car, stationed on a wet road under a dark sky, with its vivid red color prominently contrasting the background.”* Furthermore, to guide SEED-LLaMA in generating images rather than their descriptions, we append the following instruction at the beginning of the prompt: *“We provide a few examples, each with an input, and an output of the image description. Based on the examples, predict the next image description and visualize it.”* Similarly, for Qwen-VL and Gemini, we include an instruction: *“We provide a few examples, each with an input, and an output of the image description. Based on the examples, predict the next image description,”* with the focus on predicting the next image description without the visualization component. This distinction aims to direct SEED-LLaMA towards image generation, whereas Qwen-VL and Gemini are instructed to generate image descriptions.

**Results.** The results are reported in Table 7. For 2-shot cases, SEED-LLaMA exhibits similar accuracies in both T2I-ICL and T-ICL, but in 4-shot instances, T-ICL surpasses T2I-ICL in eight out of ten tasks. The disparity is even more evident for Qwen-VL and Gemini; under T-ICL, it significantly outperforms T2I-ICL, especially in 4-shot situations. These findings confirm our first hypothesis, indicating that multimodality is the primary challenge in T2I-ICL.

### F.2.2 Is the Image Generation a Primary Challenge in T2I-ICL?

To verify the second hypothesis, that image generation itself presents a primary challenge in T2I-ICL, we conduct an experiment with 0, 2, and 4-shot image generation tasks, with textual inputs updated as precise labels. For example, in the initial scenario from Figure 2, the terms “White,” “Blue,” and “Red” are updated to “White car,” “Blue car,” and “Red car,” respectively. For this experiment, we exclude MLLMs that do not generate images. Instead, we focus on MLLMs that are capable of generating images, including Emu, GILL, and SEED-LLaMA.

Model	Shot	Method	Object-Inference Task					Attribute-Inference Task				
			Color-I	Background-I	Style-I	Action-I	Texture-I	Color-II	Background-II	Style-II	Action-II	Texture-II
SEED-LLaMA	2	T2I-ICL	<b>.680</b>	.348	.203	.182	.196	.287	<b>.467</b>	<b>.297</b>	<b>.261</b>	.163
		T-ICL	<u>.614</u>	<b>.380</b>	<b>.246</b>	<b>.279</b>	<b>.265</b>	<b>.531</b>	<u>.315</u>	<u>.206</u>	<u>.184</u>	<b>.192</b>
	4	T2I-ICL	.482	.211	.141	.053	.122	.252	.076	<b>.268</b>	<b>.207</b>	.105
		T-ICL	<b>.584</b>	<b>.404</b>	<b>.289</b>	<b>.317</b>	<b>.276</b>	<b>.667</b>	<b>.343</b>	<u>.266</u>	<u>.195</u>	<b>.228</b>
Qwen-VL	2	T2I-ICL	.475	.313	.378	.464	.338	.457	.379	.258	.388	.316
		T-ICL	<b>.854</b>	<b>.822</b>	<b>.692</b>	<b>.892</b>	<b>.679</b>	<b>.272</b>	<b>.409</b>	<b>.559</b>	<b>.428</b>	<b>.431</b>
	4	T2I-ICL	.560	.459	.571	.679	.454	.568	.364	.341	.546	.434
		T-ICL	<b>.973</b>	<b>.851</b>	<b>.857</b>	<b>.972</b>	<b>.890</b>	<b>.740</b>	<b>.805</b>	<b>.793</b>	<b>.719</b>	<b>.827</b>
Gemini	2	T2I-ICL	.865	.794	.315	.517	.704	.555	.583	.360	.725	.340
		T-ICL	<b>.979</b>	<b>.907</b>	<b>.692</b>	<b>.895</b>	<b>.764</b>	.150	.410	<b>.645</b>	<b>.468</b>	<b>.361</b>
	4	T2I-ICL	.904	.908	.540	.737	.861	.709	.773	.484	<b>.818</b>	.553
		T-ICL	<b>.988</b>	<b>.965</b>	<b>.888</b>	<b>.965</b>	<b>.927</b>	<b>.777</b>	<b>.780</b>	<b>.835</b>	<b>.783</b>	<b>.812</b>

Table 7: **Comparison of Text-to-Image ICL (T2I-ICL) versus Textual ICL (T-ICL) accuracy on our dataset.** To perform T-ICL on our dataset, we replace all images in the prompts with their corresponding detailed descriptions. Underlined numbers indicate the highest accuracy achieved for each model and task across various shot numbers, while bold numbers indicate the highest accuracy for each specific combination of model, task, and shot count. In this experiment, we focus on three MLLMs: SEED-LLaMA, which is used for image generation; Qwen-VL and Gemini, utilized for generating image descriptions. MLLMs demonstrate notably superior performance in T-ICL compared to T2I-ICL, particularly in the 4-shot scenario.

Model	Shot	Precise Textual Inputs	Object-Inference Task					Attribute-Inference Task				
			Color-I	Background-I	Style-I	Action-I	Texture-I	Color-II	Background-II	Style-II	Action-II	Texture-II
SEED-LLaMA	0	✓	.730	<b>.456</b>	<b>.356</b>	<b>.264</b>	.275	.582	.314	.298	.207	<b>.286</b>
	2	✗	.680	.348	.203	.182	.196	.287	.467	.297	.261	.163
		✓	<b>.801</b>	<b>.409</b>	<b>.241</b>	<b>.192</b>	<b>.326</b>	<b>.385</b>	<b>.485</b>	<b>.393</b>	<b>.317</b>	<b>.268</b>
	4	✗	.482	.211	.141	.053	.122	.252	.076	.268	.207	.105
✓		<b>.669</b>	<b>.318</b>	<b>.284</b>	<b>.161</b>	<b>.286</b>	<b>.608</b>	<b>.441</b>	<b>.299</b>	<b>.278</b>	<b>.248</b>	
Emu	0	✓	<b>.094</b>	<b>.102</b>	.052	.064	.047	.054	.075	.069	<b>.160</b>	.028
	2	✗	.065	.051	.057	.052	.078	.062	<b>.109</b>	<b>.081</b>	<b>.092</b>	<b>.074</b>
		✓	.050	<b>.086</b>	<b>.101</b>	<b>.070</b>	<b>.116</b>	<b>.122</b>	.087	.074	.079	.060
	4	✗	.063	.018	.045	.048	<b>.097</b>	.037	<b>.122</b>	<b>.109</b>	<b>.077</b>	<b>.088</b>
✓		.061	<b>.069</b>	<b>.136</b>	<b>.056</b>	.091	<b>.136</b>	.083	.076	.072	.081	
GILL	0	✓	<b>.341</b>	<b>.286</b>	<b>.244</b>	<b>.135</b>	<b>.237</b>	<b>.297</b>	<b>.223</b>	<b>.178</b>	<b>.176</b>	<b>.226</b>
	2	✗	.171	.054	.069	.063	.074	.010	.043	.024	.022	.040
		✓	<b>.245</b>	<b>.112</b>	<b>.100</b>	<b>.066</b>	<b>.108</b>	<b>.023</b>	<b>.092</b>	<b>.054</b>	<b>.021</b>	<b>.075</b>
	4	✗	.106	.044	.041	.073	.087	.022	.059	.044	.032	.067
✓		<b>.178</b>	<b>.084</b>	<b>.125</b>	<b>.064</b>	<b>.133</b>	<b>.072</b>	<b>.092</b>	<b>.055</b>	<b>.037</b>	<b>.095</b>	

Table 8: **Accuracy comparison on SEED-LLaMA, Emu, and GILL: with or without providing precise textual inputs.** Bold numbers represent the highest accuracy for each task and shot count, comparing scenarios with and without descriptive textual inputs. Underlined numbers indicate the highest accuracy for each task across various shots.

**Results.** Table 8 presents a comparative analysis of three considered MLLMs’s performance in T2I-ICL, both with and without the inclusion of precise textual inputs. We observe that in scenarios with both 2 and 4 shots, the presence of precise textual inputs leads to significantly higher accuracy in image generation compared to when these inputs are absent for SEED-LLaMA and GILL, whereas Emu’s performance does not follow a discernible trend. Crucially, the analysis shows that, even with precise inputs, all models sustain a comparable level of performance across different tasks, with accuracies remaining under 50% in most cases. This suggests that the task of image generation remains a considerable challenge for contemporary MLLMs, affecting their efficacy on the CoBSAT dataset.

### F3 Enhancing MLLMs’ T2I-ICL Capabilities (Detailed Version of Sec. 7)

In this section, we supplement Sec 7 with additional experimental details, discussion, and expanded experiment results, exploring techniques that could potentially enhance the performance of MLLMs in T2I-ICL.

#### F3.1 Fine-tuning MLLMs on CoBSAT

In this experiment, we investigate the impact of fine-tuning MLLMs on our dataset in improving its T2I-ICL capabilities. We focus on SEED-LLaMA Qwen-VL for this investiga-

tion. Consequently, we compare the T2I-ICL performance of the pretrained-only version of Qwen-VL nad SEED-LLaMA with their corresponding variant that is fine-tuned on our dataset.

**Training Setup.** We fine-tune two instances of both SEED-LLaMA and Qwen-VL, one on a 2-shot dataset and the other on a 4-shot dataset, and then compare their performances with their non-fine-tuned counterparts on the T2I-ICL test set. For both models, we fine-tune their LLM backbone only using LoRA (Hu et al., 2022) with a rank of 64, a weight decay of 0.1, and a warm-up ratio of 0.01 for 5 epochs.

**Training and Test Sets.** We employ two distinct strategies for splitting the training and test datasets. In the first strategy, the training set comprises prompts from all ten themes, ensuring that attributes and objects in the test set are not previously exposed in any training prompts. In the second strategy, the training set excludes the themes that are included in the test set, enabling us to assess whether a model fine-tuned on specific tasks can generalize to other tasks. Note that the tasks configured by the second approach are inherently more challenging.

(Data Split A) The training and test sets are constructed by splitting the predefined lists of text inputs and latent variables (from Table 5, denoted as  $\mathcal{X}$  and  $\Theta$ ) into training ( $\mathcal{X}_{\text{train}}, \Theta_{\text{train}}$ ) and testing ( $\mathcal{X}_{\text{test}}, \Theta_{\text{test}}$ ) subsets for each task, in a 1:1 ratio. Therefore, all the in-context demonstrations and textual queries in the test sets are unseen from the training set. For the training set, we create the dataset by considering all possible combinations and sequences of text inputs from  $\mathcal{X}_{\text{train}}$  and latent variables  $\Theta_{\text{train}}$  across all tasks. We fine-tune Qwen-VL on this unified training set containing the prompts from all tasks. For the testing set, we generate 250 prompts for each shot across various tasks. Each prompt is obtained by randomly sampling  $\theta \in \Theta_{\text{test}}$  and  $(x_n)_{n=1}^{N+1} \in \mathcal{X}_{\text{test}}^{N+1}$ , which are then paired with the corresponding collected images  $(y_n)_{n=1}^N$ , where  $y_n \sim f_{\theta}(x_n)$ . This process results in  $N$  in-context demonstrations  $(x_n, y_n)_{n=1}^N$  and a single textual query  $x_{N+1}$ .

(Data Split B) In this data split, we intensify the challenge by increasing the disparity between the training and testing distributions. Instead of merely including unseen objects and attributes from the same themes in the testing dataset, this split introduces unseen themes. For example, the results shown in Table 9 for split B on color-themed tasks (i.e., Color-I and Color-II) are derived from a model fine-tuned on the other four themes (i.e., eight tasks). Thus, the model is not fine-tuned on color-themed tasks but is evaluated on them. This method is uniformly applied across all themes: each theme is evaluated using a model fine-tuned on tasks from the other four themes. Thus, the training set includes four themes, while the test set comprises a different, fifth theme. Consequently, the results in Table 9 for split B reflect different models, each fine-tuned on a distinct training set.

**Results.** The results are summarized in Table 9. For data split A, both models demonstrate significant improvements in T2I-ICL performance following fine-tuning. In the more challenging tasks defined by data split B, SEED-LLaMA demonstrates strong generalization to unseen tasks after fine-tuning, while Qwen-VL exhibits more difficulty in generalizing. Overall, these results suggest that fine-tuning MLLMs on a T2I-ICL dataset generally enhances their overall T2I-ICL capabilities. Example output images from the pre-trained and fine-tuned versions of SEED-LLaMA on split A are provided in Sec. H.2.

### F.3.2 Intergrating Chain-of-Thought with T2I-ICL

Another widely utilized method in prompt engineering is Chain-of-Thought (CoT) (Wei et al., 2022). This approach involves incorporating a simple instruction, such as “let’s think step by step,” prompting the model to sequentially generate concise sentences that outline the reasoning process, commonly referred to as reasoning chains or rationales. The chains are subsequently embedded into the subsequent prompt to obtain the final answer. CoT has been particularly effective in enhancing performance, especially for complex reasoning tasks,

Model	Shot	Fine-tuned	Split	Object-Inference Task					Attribute-Inference Task				
				Color-I	Background-I	Style-I	Action-I	Texture-I	Color-II	Background-II	Style-II	Action-II	Texture-II
SEED-LLaMA	2	X	-	.636	.292	.088	.196	.108	.360	.536	.164	.196	.080
			A	<u>.776</u>	<b>.540</b>	.164	<b>.284</b>	<b>.208</b>	.468	<b>.588</b>	.108	.192	<b>.140</b>
			B	<u>.752</u>	<b>.484</b>	<b>.208</b>	<u>.272</u>	<b>.200</b>	<b>.568</b>	.376	<b>.240</b>	.180	<u>.104</u>
	4	X	-	.612	.360	.092	.044	.048	.380	.532	.140	.196	.148
			A	<b>.784</b>	<b>.516</b>	<b>.152</b>	<b>.160</b>	<b>.172</b>	<b>.504</b>	<b>.564</b>	.104	.192	<b>.200</b>
			B	<u>.748</u>	<b>.556</b>	<b>.208</b>	<b>.256</b>	<b>.244</b>	<b>.616</b>	.488	.112	.132	<b>.216</b>
Qwen-VL	2	X	-	.540	.236	.248	.412	.372	.276	.244	.112	.232	.224
			A	<b>.852</b>	<b>.744</b>	.212	<b>.856</b>	<b>.532</b>	.516	<b>.344</b>	<b>.148</b>	<b>.520</b>	<b>.284</b>
			B	<u>.708</u>	<b>.552</b>	<b>.376</b>	<b>.308</b>	<b>.328</b>	<b>.592</b>	.272	<b>.224</b>	.212	.172
	4	X	-	.680	.492	.448	.228	.556	.512	<b>.448</b>	<b>.240</b>	.320	.420
			A	<b>.876</b>	<b>.604</b>	.216	<b>.812</b>	<b>.588</b>	<b>.696</b>	.308	.088	<b>.656</b>	<b>.480</b>
			B	<u>.812</u>	<b>.728</b>	.300	<b>.352</b>	.464	<b>.740</b>	.380	<b>.240</b>	.212	.308

Table 9: **T2I-ICL accuracy comparison of pretrained-only versus fine-tuned (FT) MLLMs.** Underlined numbers signify instances where the fine-tuned model surpasses the pretrained model in the same scenario, while bold numbers indicate the top performance for each shot across various methods within their tasks.

when applied to large-scale models (Wei et al., 2022). In our experiment, we investigate the impact of integrating CoT on the T2I-ICL performance of MLLMs.

**Prompt Design.** We employ a two-step inference process utilizing two distinct prompts. The initial prompt builds upon the default examples showcased in Figure 2. To this, we prepend the statement, “We provide a few examples, each of which is an input-output pair where the output is a description of the image associated with the input. Based on the examples, the task is to predict the next image description. \n\n\n” This is placed at the beginning of the prompt. Additionally, we append, “\n\n\n Before predicting the next image, let’s first think step by step and analyze the relationship between the text input and image output in each example. \n\n\n” to the end of the prompt.

Following the MLLMs’ responses, the second prompt comes into play. It includes the first prompt and the MLLM’s response as part of the new prompt and extends it with, “\n\n\n Based on the analysis, please generate the next image for the request ‘red: ’ ” for the image generation scenario, and “\n\n\n Based on the analysis, please describe what the next image should look like for the request ‘red: ’ ” for the image description generation scenario when the textual query is ‘red.’ In each case, ‘red’ is replaced with the respective textual query according to different prompts.

Model	Shot	Method	Object-Inference Task					Attribute-Inference Task				
			Color-I	Background-I	Style-I	Action-I	Texture-I	Color-II	Background-II	Style-II	Action-II	Texture-II
SEED-LLaMA	2	T2I-ICL	.680	<b>.348</b>	.203	<b>.182</b>	.196	.287	<b>.467</b>	.297	.261	<b>.163</b>
		CoT + T2I-ICL	<b>.781</b>	.179	<b>.206</b>	.167	<b>.222</b>	.179	.389	.195	<b>.300</b>	.154
	4	T2I-ICL	.482	.211	.141	.053	.122	.252	.076	.268	.207	.105
		CoT + T2I-ICL	<b>.650</b>	<b>.353</b>	<b>.244</b>	<b>.242</b>	<b>.208</b>	<b>.303</b>	<b>.370</b>	<b>.335</b>	<b>.241</b>	<b>.171</b>
Qwen-VL	2	T2I-ICL	.475	.313	.378	<b>.464</b>	.338	<b>.457</b>	<b>.379</b>	.258	<b>.388</b>	<b>.316</b>
		CoT + T2I-ICL	.281	<b>.494</b>	<b>.387</b>	.217	<b>.363</b>	.150	.349	<b>.260</b>	.176	.181
	4	T2I-ICL	<b>.560</b>	<b>.459</b>	<b>.571</b>	<b>.679</b>	.454	<b>.568</b>	.364	<b>.341</b>	<b>.546</b>	<b>.434</b>
		CoT + T2I-ICL	.548	.379	.274	.404	<b>.573</b>	.207	<b>.690</b>	<b>.409</b>	.424	.340
Gemini	2	T2I-ICL	.865	.794	.315	.517	.704	.555	.583	.360	<b>.725</b>	.340
		CoT + T2I-ICL	<b>.938</b>	<b>.861</b>	<b>.647</b>	<b>.882</b>	<b>.731</b>	<b>.655</b>	<b>.908</b>	<b>.672</b>	.701	<b>.445</b>
	4	T2I-ICL	.904	.908	.540	.737	.861	.709	.773	<b>.484</b>	.818	.553
		CoT + T2I-ICL	<b>.986</b>	<b>.957</b>	<b>.799</b>	<b>.916</b>	<b>.945</b>	<b>.917</b>	<b>.977</b>	.293	<b>.897</b>	<b>.755</b>

Table 10: **Assessing the impact of Chain-of-Thought (CoT) prompting on T2I-ICL.** The evaluation metric is accuracy, with the numbers in bold highlighting the highest accuracy achieved for each model, number of shots, and task, and underlined numbers indicate the highest accuracy achieved for each model and task across different numbers of shots. In this experiment, we evaluate three MLLMs: SEED-LLaMA for image generation, and Qwen-VL and Gemini for image description generation. Our findings reveal that CoT significantly improves Gemini’s performance. For SEED-LLaMA and Qwen-VL, the enhancement offered by CoT is ambiguous in 2-shot scenarios. However, in 4-shot instances, CoT markedly enhances the performance of SEED-LLaMA, while it still shows no benefit for Qwen-VL.

**Results.** The results are reported in Table 10. With the integration of CoT, Gemini shows better performance across the most of tasks in both 2-shot and 4-shot scenarios. Similarly, SEED-LLaMA shows significant improvement in T2I-ICL performance across all ten tasks

in the 4-shot scenario. Conversely, for Qwen-VL, no concrete evidence suggests that CoT enhances its T2I-ICL performance. In fact, we find that Qwen-VL often avoids providing definitive answers in the second step of making predictions, and responds with general statements like “Given the request ‘black:’, we can infer that the image output should be related to a black object or theme. However, without more specific information, it’s difficult to determine the exact relationship between the text input and image output. Without additional context, it’s impossible to accurately predict the next image.” Therefore, in many instances, standard T2I-ICL without CoT appears to outperform the version integrated with CoT for Qwen-VL. Exploring additional prompt engineering methods such as self-consistency sampling (Wang et al., 2023b) or Tree-of-Thought (Yao et al., 2023) to elicit more concrete responses from Qwen-VL is a possibility. Specifically, self-consistency sampling involves generating multiple outputs at a non-zero temperature setting and selecting the most appropriate one from these options. On the other hand, Tree-of-Thought expands upon CoT by considering multiple lines of reasoning at each step. However, such investigations fall outside the scope of this paper, and we identify it as one of the interesting future directions. We provide example conversations of integrating CoT and T2I-ICL in Sec. H.3.

### F.3.3 Articulating the Text-to-Image Relationship in Prompts

In our dataset, the goal is to check if the MLLMs are able to learn the mapping from the textual input and the visual output based on the in-context demonstrations. In this experiment, we investigate the performance of MLLMs on T2I-ICL if we explicitly write down this relationship in the text prompt. For instance, for the Color-I task, we directly add the following sentence to the beginning of the prompt: “Please identify the common main object in the images, and generate another image of this object in the requested color.” The detailed prompts for all tasks are provided in Sec. D.1.2.

Model	Shot	Explicit Instruction	Object-Inference Task					Attribute-Inference Task				
			Color-I	Background-I	Style-I	Action-I	Texture-I	Color-II	Background-II	Style-II	Action-II	Texture-II
SEED-LLaMA	2	$\times$ $\checkmark$	.680 <u>.779</u>	.348 <u>.391</u>	.203 <u>.231</u>	.182 <u>.301</u>	.196 <u>.270</u>	<b>.287</b> <u>.257</u>	<b>.467</b> <u>.446</u>	.297 <u>.350</u>	<b>.261</b> <u>.249</u>	.163 <u>.185</u>
	4	$\times$ $\checkmark$	.482 <b>.832</b>	.211 <u>.408</u>	.141 <u>.281</u>	.053 <u>.318</u>	.122 <b>.322</b>	.252 <u>.388</u>	.076 <b>.483</b>	.268 <u>.406</u>	.207 <b>.268</b>	.105 <u>.228</u>
Qwen-VL	2	$\times$ $\checkmark$	.475 <u>.407</u>	.313 <u>.496</u>	.378 <u>.240</u>	.464 <u>.516</u>	.338 <u>.300</u>	<b>.457</b> <u>.240</u>	.379 <b>.697</b>	.258 <u>.317</u>	.388 <b>.600</b>	.316 <u>.373</u>
	4	$\times$ $\checkmark$	<b>.560</b> <u>.315</u>	<b>.459</b> <u>.291</u>	<b>.571</b> <u>.341</u>	<b>.679</b> <u>.475</u>	<b>.454</b> <u>.473</u>	<b>.568</b> <u>.277</u>	.364 <b>.591</b>	<b>.341</b> <u>.317</u>	<b>.546</b> <u>.527</u>	<b>.434</b> <u>.404</u>
Gemini	2	$\times$ $\checkmark$	.865 <u>.119</u>	.794 <u>.624</u>	.315 <u>.553</u>	.517 <u>.620</u>	.704 <u>.176</u>	.555 <u>.128</u>	.583 <u>.735</u>	.360 <u>.155</u>	.725 <u>.373</u>	.340 <u>.118</u>
	4	$\times$ $\checkmark$	<b>.904</b> <u>.198</u>	<b>.908</b> <u>.655</u>	.540 <b>.564</b>	<b>.737</b> <u>.675</u>	<b>.861</b> <u>.356</u>	<b>.709</b> <u>.125</u>	.773 <b>.921</b>	<b>.484</b> <u>.199</u>	<b>.818</b> <u>.520</u>	<b>.553</b> <u>.125</u>

Table 11: Effect of explicit instruction on T2I-ICL performance of MLLMs: articulating the text-to-image relationship in prompts. The evaluation metric is accuracy, where underlined numbers denote the highest accuracy achieved by each model and task across varying shot numbers, and bold numbers represent the top accuracy for each specific combination of model, task, and shot count. This evaluation focuses on SEED-LLaMA for image generation, and Qwen-VL and Gemini for image description generation. We find that explicit instructions significantly enhance the T2I-ICL capability of SEED-LLaMA, especially in the 4-shot scenario. However, for Qwen-VL and Gemini, explicit instructions do not show similar performance gains.

**Results.** We present the experiment results in Table 11. Results show that SEED-LLaMA’s performance significantly improves with explicit instructions, surpassing its performance in T2I-ICL without instructions for seven out of ten tasks in the 2-shot cases and all tasks in the 4-shot cases. Notably, in the 4-shot case for the Color-I task, SEED-LLaMA achieves a high accuracy of 83.2% with explicit instructions, compared to only 48.2% without them. Furthermore, the performance of T2I-ICL with explicit instructions improves when moving from 2-shot to 4-shot scenarios, in contrast to the situation without explicit instructions where SEED-LLaMA’s T2I-ICL performance declines as the number of demonstrations increases from 2 to 4. In contrast, Qwen-VL does not show comparable improvements, owing to reasons similar to those discussed in Sec. F.3.2, including the generation of irrelevant responses like “Received.”. Similarly, Gemini also fails to demonstrate improvements. To be

more specific, we find that Gemini consistently ignores the textual query after articulating the text-to-image relationship in prompts. To handle these issues, more careful prompt engineering could be applied, although it is beyond the scope of this paper.

## G Extended Discussion

This section is an expanded version of Sec 8, discussing the conclusion, limitations, and future works in greater detail.

### G.1 Conclusion

In this work, we identify an important yet underexplored problem — T2I-ICL, and explore the capability of MLLMs to solve it. To facilitate this investigation, we introduce CoBSAT, a comprehensive benchmark dataset encompassing ten tasks. Our experimental evaluation of MLLMs on this dataset reveals that while many MLLMs have difficulty in effectively learning from in-context demonstrations during text-to-image generation, a few MLLMs, such as GPT-4V, Qwen-VL, Gemini, Claude, and SEED-LaMA, show comparatively reasonable performance. Through further studies on free top-performing models SEED-LLaMA, Gemini, and Qwen-VL for both image and image description generation, we identify two key challenges in T2I-ICL: (i) the integration and understanding of multimodal information, evidenced by superior results achieved with textual ICL for the same tasks; and (ii, particularly for image generation models) the actual process of image creation, as even straightforward image requests with clear descriptions often yield suboptimal performances.

To improve MLLMs’ performance in T2I-ICL, we carry out additional experimental studies. These studies suggest that fine-tuning and CoT can substantially enhance T2I-ICL capabilities. Meanwhile, it is worth noting that in our dataset, we intentionally exclude explicit task descriptions to assess whether MLLMs can autonomously adapt to tasks based solely on in-context demonstrations alone. In the ablation studies, we find that providing clearer task instructions might be a promising strategy for enhancing T2I-ICL performance. However, these prompting engineering strategies might need to be combined with others to achieve consistent improvements.

### G.2 Limitations and Future Works

While our study is an early attempt to explore the T2I-ICL benchmark dataset, many interesting questions remain open.

**Impact of Demonstration Selection on T2I-ICL Performance.** Existing research in textual ICL has consistently demonstrated that the choice of demonstrations significantly influences ICL performances (Liu et al., 2022; Su et al., 2023; Rubin et al., 2022; Zhang et al., 2022b). In our study, we only employ random sampling to select in-context demonstrations. This opens an interesting question: to what extent does the selection of demonstrations affect T2I-ICL performance? Moreover, our evaluation primarily assesses whether MLLMs can accurately generate images with the current content, without a specific focus on the quality of these images. Another natural question arises: how significantly does the quality of images used in demonstrations influence the overall quality of the generated image output?

**Prompt Engineering Techniques for MLLMs.** As discussed in Sec. 7, CoT demonstrates significant improvements in T2I-ICL performance for SEED-LLaMA and Gemini. However, the prompt sensitivity of models like Qwen-VL poses a challenge, as they tend to provide non-committal responses such as, “Without additional context, it’s impossible to accurately predict the next image.” This issue underscores the necessity of implementing more advanced prompt engineering techniques, including methods like Tree-of-Thought and self-consistency sampling, to address these limitations.

Once such issues are resolved, another interesting question arises: Is it feasible to enhance existing prompt engineering techniques with multimodal capabilities? For instance, while



Sec. 7 focuses on prompting MLLM to perform CoT through textual analysis (as further exemplified in Sec. F.3.2), expanding this approach to a multimodal CoT that integrates both textual analysis and image grounding could potentially yield better performance. These open questions are identified as interesting future directions.

**T2I-ICL for Image Editing.** One notable absence in our dataset is tasks related to image editing. For instance, in the Color-I task, the goal is to generate an image of a car in a color specified by the text query. In our evaluation, the car type and background in both the example images from the prompt and the newly generated image may differ. However, there is a growing need for image editing applications where the task is to alter specific attributes (e.g., the color of a car) in an otherwise unchanged image. For such tasks, selecting images that strictly adhere to the given criteria (identical images with only specific attributes or objects altered), coupled with the development of sophisticated metrics, is critical to assess these more complex challenges effectively.

**Exploring a Wider Range of Themes.** Our dataset primarily assesses MLLMs on elementary themes, incorporating a specific range of objects and attributes within narrowly defined categories. For instance, in the style task, we consider styles such as watercolor, sketch, pixel art, origami, and others. Nonetheless, the realm of styles in real-world applications is far more intricate and varied, extending to include oil painting, rococo, steampunk, and beyond. Additionally, our dataset encompasses a limited set of themes. There are also many other interesting themes such as counting. While it serves to test basic capabilities in T2I-ICL, a more comprehensive dataset, covering a broader spectrum of themes and a finer list of objects and attributes, will be crucial for evaluating more advanced model capabilities. In this scenario, the evaluation methodology may require refinement to more accurately identify the more fine-grained attributes and objects.

## H Sample Outputs Generated by MLLMs

In this section, we provide sample outputs generated by the models under different scenarios. Specifically, Sec. H.1 contains a selection of sample prompts along with the corresponding responses generated by MLLMs across all ten tasks. Sec. H.2 displays selected sample prompts and the images produced by both the pretrained and fine-tuned SEED-LLaMA for all ten tasks. Furthermore, in Sec. H.3, we illustrate sample dialogues between users and MLLMs (including SEED-LLaMA and Qwen-VL) from our experiments that combine the CoT with T2I-ICL.

### H.1 Sample Prompts and Corresponding Outputs




































Here, we showcase examples of outputs generated by various MLLMs, accompanied by their respective prompts.

**Image Generation** For image generation, five examples are provided for each themed task, as depicted in Figures 14, 15, 16, 17, and 18 for color, background, style, action, and texture themes, respectively. Observing these figures, it is evident that SEED-LLaMA excels in image quality among the three MLLMs capable of image generation. Notably, SEED-LLaMA produces images that not only align with the true labels but also closely resemble the images in the prompts. For instance, in Figure 14, images with plain backgrounds in the prompts lead to similarly styled outputs.

**Image Description Generation** In Figures 19, 20, 21, 22, and 23, we showcase two examples for each task, covering color, background, style, action, and texture themes. For brevity, some lengthy responses from MLLMs have been truncated in these figures, retaining only the key parts of the responses.

Latent Var.	Input			Emu	Emu2	GILL	SEED-LLaMA
Bag	Brown: 	Blue: 	Purple: 				
Car	Brown: 	Purple: 	Orange: 				
Hat	Green: 	Yellow: 	White: 				
Book	Purple: 	White: 	Pink: 				
Box	White: 	Yellow: 	Green: 				

(a) Color-I

Latent Var.	Input			Emu	Emu2	GILL	SEED-LLaMA
Purple	Cup: 	Box: 	Bag: 				
Orange	Ball: 	Book: 	Car: 				
White	Book: 	Leaf: 	Hat: 				
Pink	Box: 	Hat: 	Book: 				
Green	Ball: 	Bag: 	Box: 				

(b) Color-II

Figure 14: Examples of prompts and corresponding images generated by MLLMs for tasks within the Color theme.

## H.2 Sample Outputs from Fine-tuning SEED-LLaMA on CoBSAT

In Figure 24, we provide sample outputs generated by pretrained-only and fine-tuned SEED-LLaMA from the experiments described in Sec. 7 and F.3.1. We observe that the images generated by fine-tuned SEED-LLaMA generally align better with the expected output.

## H.3 Sample Outputs from Integrating CoT with T2I-ICL

In this section, we provide sample outputs generated by Qwen-VL and SEED-LLaMA from the experiment of integrating CoT with T2I-ICL, as detailed in Sec. 7. Specifically, Figures 25, 26, and 27 illustrate the example prompts and corresponding outputs generated by Qwen-VL for the tasks Color-I, Action-II, and Texture-II. Similarly, Figures 28, 29, and 30 display the example prompts and outputs produced by SEED-LLaMA for the tasks Color-I, Background-II, and Style-I.

Latent Var.	Input			Emu	Emu2	GILL	SEED-LLaMA
Bird	Seafloor:	Beach:	Waterfall:				
Lion	Gym:	Seafloor:	Beach:				
Tiger	Desert:	Park:	Glacier:				
Monkey	Waterfall:	Cave:	Desert:				
Zebra	Beach:	Waterfall:	Cave:				

(a) Background-I

Latent Var.	Input			Emu	Emu2	GILL	SEED-LLaMA
Waterfall	Tiger:	Monkey:	Bird:				
Beach	Monkey:	Dog:	Lion:				
Glacier	Monkey:	Lion:	Tiger:				
Desert	Lion:	Dog:	Monkey:				
Cave	Cow:	Cat:	Zebra:				

(b) Background-II

Figure 15: Examples of prompts and corresponding images generated by MLLMs for tasks within the Background theme.

Latent Var.	Input			Emu	Emu2	GILL	SEED-LLaMA
Cup	Lego: 	Origami: 	Sketch: 				
Chair	Icon: 	Pixel: 	Watercolor: 				
Book	Sketch: 	Lego: 	Old: 				
Ball	Pixel: 	Sketch: 	Graffiti: 				
Apple	Futuristic: 	Origami: 	Icon: 				

(a) Style-I

Latent Var.	Input			Emu	Emu2	GILL	SEED-LLaMA
Sketch	Car: 	Apple: 	Cup: 				
Watercolor	Cup: 	Apple: 	Chair: 				
Old	Cup: 	Ball: 	Book: 				
Graffiti	Bag: 	Hat: 	Ball: 				
Icon	Leaf: 	Car: 	Apple: 				

(b) Style-II

Figure 16: Examples of prompts and corresponding images generated by MLLMs for tasks within the Style theme.

Latent Var.	Input			Emu	Emu2	GILL	SEED-LLaMA
Dog	Run: 	Drink: 	Swim: 				
Zebra	Fly: 	Swim: 	Drink: 				
Cat	Run: 	Cry: 	Eat: 				
Pig	Eat: 	Drink: 	Angry: 				
Sheep	Eat: 	Angry: 	Sleep: 				

(a) Action-I

Latent Var.	Input			Emu	Emu2	GILL	SEED-LLaMA
Swim	Monkey: 	Sheep: 	Dog: 				
Drink	Cat: 	Sheep: 	Zebra: 				
Eat	Bird: 	Pig: 	Cat: 				
Angry	Bird: 	Cow: 	Pig: 				
Sleep	Pig: 	Dog: 	Sheep: 				

(b) Action-II

Figure 17: Examples of prompts and corresponding images generated by MLLMs for tasks within the Action theme.

Latent Var.	Input			Emu	Emu2	GILL	SEED-LLaMA
Ball	Leather: 	Wicker: 	Metal: 				
Box	Metal: 	Sequined: 	Wood: 				
Hat	Ceramic: 	Paper: 	Wicker: 				
Bag	Ceramic: 	Sequined: 	Denim: 				
Cup	Plastic: 	Denim: 	Ceramic: 				

(a) Texture-I

Latent Var.	Input			Emu	Emu2	GILL	SEED-LLaMA
Metal	Bag: 	Hat: 	Ball: 				
Wood	Leaf: 	Book: 	Box: 				
Wicker	Ball: 	Cup: 	Hat: 				
Denim	Ball: 	Box: 	Bag: 				
Ceramic	Chair: 	Box: 	Cup: 				

(b) Texture-II



Figure 18: Examples of prompts and corresponding images generated by MLLMs for tasks within the Texture theme.

Latent Var.	Input		
Bag	Brown: 	Blue: 	Purple:
Emu	Emu2	GILL	
<i>A brown leather backpack with two zippers.</i>	<i>A blue backpack with two zippers on it.</i>	<i>image of the backpack in the image</i>	
SEED-LLaMA	LLaVA-1.5	LLaVA-NeXT	
<i>I have generated an image.</i>	<i>The next image should be a close-up of a blue backpack with a brown strap.</i>	<i>brown leather backpack.</i>	
Qwen-VL	GPT-4V		
<i>The next image should be a close-up of a blue backpack with brown leather details, showcasing the main object clearly.</i>	<i>A purple backpack with a similar style to the previous images, featuring a vibrant purple fabric, sturdy zippers</i>		
Gemini	Claude		
<i>A purple backpack with a black zipper and a small black tag on the front.</i>	<i>The next image appears to be a purple backpack. The backpack is likely made of a durable, high-quality material and features multiple compartments and pockets for storage and organization.</i>		



  

Latent Var.	Input		
Car	Brown: 	Purple: 	Orange:
Emu	Emu2	GILL	
<i>An orange sports car driving down a road.</i>	<i>A purple car is driving down the road.</i>	<i>automobile model is a car that is a sports car</i>	
SEED-LLaMA	LLaVA-1.5	LLaVA-NeXT	
<i>I have generated an image.</i>	<i>The next image should be a close-up of the front of the purple sports car, showcasing its grill and headlights. The car should be parked on a street,</i>	<i>The next image to be generated should be a close-up of the front grille of the purple sports car.</i>	
Qwen-VL	GPT-4V		
<i>The next image should be an orange sports car driving on a road with a blurry background.</i>	<i>A sleek race car in bright orange, with black detailing, glossy finish, and racing stripes, speeding on a racetrack with motion blur in the background.</i>		
Gemini	Claude		
<i>A bright orange Lamborghini Aventador SVJ Roadster</i>	<i>The image appears to depict a sleek and sporty sports car in a vibrant orange color.</i>		

(a) Color-I

Latent Var.	Input		
Purple	Cup: 	Box: 	Bag:
Emu	Emu2	GILL	
<i>A purple coffee mug sitting on top of a box.</i>	<i>A purple box sitting on a white background.</i>	<i>product - a box of the product.</i>	
SEED-LLaMA	LLaVA-1.5	LLaVA-NeXT	
<i>The bag is blue and has a strap.</i>	<i>The next image should be a close-up of a purple cupcake with a purple frosting, placed on a white surface.</i>	<i>3D rendering of a purple coffee mug.</i>	
Qwen-VL	GPT-4V		
<i>The next image should be a close-up of a purple gift bag on a white background.</i>	<i>A purple box with a simple, clean design on a white background.</i>		
Gemini	Claude		
<i>A purple gift bag with a white ribbon.</i>	<i>The next image appears to be a purple square-shaped box or container. It has a solid, uniform purple color and a simple, minimalist design.</i>		

Latent Var.	Input		
Orange	Ball: 	Book: 	Car:
Emu	Emu2	GILL	
<i>An orange car is parked on a white surface.</i>	<i>An orange book sitting on a white background.</i>	<i>personality and the way you move your body.</i>	
SEED-LLaMA	LLaVA-1.5	LLaVA-NeXT	
<i>I have generated an image.</i>	<i>The next image should be a close-up of a person's hand holding the orange book.</i>	<i>o range of colors.</i>	
Qwen-VL	GPT-4V		
<i>The next image should be of a car, possibly a sports car or a classic car. The color of the car should be a bright and bold color, such as orange or red.</i>	<i>A red book with a plain cover, lying flat with pages visible from the side, isolated on a white background.</i>		
Gemini	Claude		
<i>In the style of pixar, render a close up of the front of a red 1950s car, with a chrome bumper and white wall tires.</i>	<i>The next image appears to be a book. The image shows an orange-colored book with a thick cover and pages.</i>		

(b) Color-II



Figure 19: Examples of prompts and corresponding images description generated by MLLMs for tasks within the Color theme. For brevity, some lengthy responses from MLLMs have been truncated, retaining only the key parts of the responses.

Latent Var.	Input		
Bird	Seafloor: 	Beach: 	Waterfall:
Emu	Emu2	GILL	
The next image is a close-up of a bird standing on the edge of a cliff overlooking a waterfall.	A blue bird walking on a beach next to the ocean.	person - bird on the beach - watercolor painting - watercolor	
SEED-LLaMA	LLaVA-1.5	LLaVA-NeXT	
unanswerable.	The next image should be a blue bird, possibly a blue jay. The background should be a blend of natural elements, such as leaves, branches, and possibly some sky.	hills in the distance.	
Qwen-VL	GPT-4V		
The next image should be of a waterfall, with the main object being a large stream of water falling from a height into a pool below.	a bluebird standing in front of a cascading waterfall surrounded by green moss-covered rocks		
Gemini	Claude		
A blue heron standing in a shallow part of a river. The water is crystal clear. The heron is looking at a fish. The background is a lush green forest.	The next image appears to depict a stunning waterfall scene. The image is likely to showcase a powerful, cascading waterfall surrounded by lush, verdant foliage and a serene natural setting.		



  

Latent Var.	Input		
Lion	Gym: 	Seafloor: 	Beach:
Emu	Emu2	GILL	
In the next image, a lion is walking through a gym, surrounded by weights and exercise equipment.	An image of a lion that is walking underwater	person - person - person - person - person - person	
SEED-LLaMA	LLaVA-1.5	LLaVA-NeXT	
I have generated an image.	a lion's face. The background should be a blend of the beach and the underwater environment, with sand, water, and possibly some seaweed or coral elements.	200 feet below the surface of the water.	
Qwen-VL	GPT-4V		
The next image should be a close-up of a lion standing on a sandy beach with a blue sky and the ocean in the background.	A majestic lion walking on a sandy beach under a clear sky.		
Gemini	Claude		
A lion walking on the beach	The next image appears to depict a serene beach scene.		

(a) Background-I

Latent Var.	Input		
Waterfall	Tiger: 	Monkey: 	Bird:
Emu	Emu2	GILL	
The next image is a photograph of a tiger standing on a rock in front of a waterfall.	A monkey that is standing on a rock.	person is a professional photographer and digital artist from the United States.	
SEED-LLaMA	LLaVA-1.5	LLaVA-NeXT	
I have generated an image.	The next image should be a close-up of the monkey's face, focusing on its eyes and facial features. The background should be the waterfall.	s.	
Qwen-VL	GPT-4V		
The next image should feature a bird perched on a branch in front of a waterfall.	A majestic bird with vibrant plumage perched on a mossy branch in the foreground, with a serene waterfall cascading in the background		
Gemini	Claude		
A bird standing on a branch in front of a waterfall. The bird is blue and the waterfall is in the background.	The image appears to depict a majestic waterfall surrounded by lush, verdant foliage. The cascading waters flow over rocky cliffs, creating a breathtaking natural spectacle.		




  

Latent Var.	Input		
Beach	Monkey: 	Dog: 	Lion:
Emu	Emu2	GILL	
The image shows a lion walking on the beach with the sun setting in the background.	A dog running on a beach next to the ocean.	person - dog - beach - dog - dog - dog -	
SEED-LLaMA	LLaVA-1.5	LLaVA-NeXT	
The lion is walking on the beach near the water.	The next image should be a close-up of the dog's face. The background should be the beach, with the ocean in the distance.	es of water.	
Qwen-VL	GPT-4V		
A lion standing on a rocky outcropping in the savannah, with the sun setting in the background.	A lion sitting on a savanna under a clear blue sky, with dry grass around and an acacia tree in the background.		
Gemini	Claude		
A lion walking on the beach with a lighthouse in the background.	The image appears to depict a lion on a beach. The lion is standing on the sandy shore, with waves crashing in the background and a lighthouse or other coastal structure visible in the distance.		




(b) Background-II

Figure 20: Examples of prompts and corresponding image descriptions generated by MLLMs for tasks within the Background theme. For brevity, some lengthy responses from MLLMs have been truncated, retaining only the key parts of the responses.






Latent Var.	Input		
Cup	Lego: 	Origami: 	Sketch: 
Emu	Emu2	GILL	
<i>The next image is a red coffee cup made out of Legos.</i>	<i>A paper mug and saucer sitting on a table.</i>	<i>step 2</i>	
SEED-LLaMA	LLaVA-1.5	LLaVA-NeXT	
<i>Enjoy this picture.</i>	<i>The next image should be a close-up of a cup of coffee, with the cup placed on a table. The style should be a realistic representation of a coffee cup</i>	<i>The next image to be generated should be a photograph of a red Lego brick structure.</i>	
Qwen-VL	GPT-4V		
<i>The next image should be a close-up of the Lego mug</i>	<i>A complex pencil sketch of a teacup and saucer with detailed shading and texture on white paper.</i>		
Gemini	Claude		
<i>A pencil drawing of a coffee mug</i>	<i>The next image appears to be an origami creation. It shows a white, folded paper structure that resembles a cup or container shape.</i>		




  

Latent Var.	Input		
Chair	Icon: 	Pixel: 	Watercolor: 
Emu	Emu2	GILL	
<i>The next image is a watercolor painting of a wooden chair.</i>	<i>A pixelated image of a wooden chair on a white background</i>	<i>The next image is a chair.</i>	
SEED-LLaMA	LLaVA-1.5	LLaVA-NeXT	
<i>I have generated an image.</i>	<i>The next image to be generated should be a close-up of a wooden chair with a purple background. The image should be in a watercolor style</i>	<i>The next image to be generated should be a pixel art style illustration of a wooden chair. The chair should have a brown color with a visible wood grain texture.</i>	
Qwen-VL	GPT-4V		
<i>The next image should be a watercolor painting of a chair</i>	<i>A watercolor painting of a chair with a blending of colors giving it a soft and fluid appearance</i>		
Gemini	Claude		
<i>A watercolor painting of a chair. The chair is made of wood and has a brown finish.</i>	<i>The next image appears to be a pixelated, retro-style rendering of a wooden chair. The chair has a simple, geometric design with a backrest and seat composed of rectangular pixels in a striped pattern.</i>		

(a) Style-I

Latent Var.	Input		
Sketch	Car: 	Apple: 	Cup: 
Emu	Emu2	GILL	
<i>A pencil drawing of an apple.</i>	<i>A drawing of an apple with a leaf on it.</i>	<i>personality.</i>	
SEED-LLaMA	LLaVA-1.5	LLaVA-NeXT	
<i>I'm sorry, but there is no description provided for the cup image.</i>	<i>The next image to be generated should be a black and white drawing of a car. The drawing should be done in a minimalist style</i>	<i>0</i>	
Qwen-VL	GPT-4V		
<i>The next image should be a cup with a handle, filled with coffee or tea, and surrounded by coffee beans or tea leaves. The style should be realistic</i>	<i>A pencil drawing of a cup with a handle on the right side, shaded to give a three-dimensional effect, on a plain background without any other objects.</i>		
Gemini	Claude		
<i>A pencil drawing of a coffee cup</i>	<i>The next image appears to be a pencil sketch or drawing of a cup or mug.</i>		

Latent Var.	Input		
Watercolor	Cup: 	Apple: 	Chair: 
Emu	Emu2	GILL	
<i>A watercolor painting of a red apple with green leaves on a white background.</i>	<i>A painting of an apple with green leaves on it.</i>	<i>watercolor painting of an apple on a white background</i>	
SEED-LLaMA	LLaVA-1.5	LLaVA-NeXT	
<i>I have generated an image.</i>	<i>The next image to be generated should be a painting of a cup of coffee with a spoon in it. The style of the painting should be impressionistic</i>	<i>ps: I'm not sure if this is a painting or a drawing. Can you tell me?</i>	
Qwen-VL	GPT-4V		
<i>A watercolor painting of a red chair with a green leaf on the back.</i>	<i>A red apple with a watercolor texture, featuring green leaves and a painterly style with visible brushstrokes on a white background.</i>		
Gemini	Claude		
<i>A watercolor painting of a wooden chair with a cushion on it. The chair is painted in a light brown color, and the cushion is painted in a dark red color.</i>	<i>Flower</i>		

(b) Style-II




Figure 21: Examples of prompts and corresponding images description generated by MLLMs for tasks within the Style theme. For brevity, some lengthy responses from MLLMs have been truncated, retaining only the key parts of the responses.

Latent Var.	Input		
Dog	Run: 	Drink: 	Swim: 
Emu	Emu2	GILL	
The next image is a cartoon dog drinking water.	A cartoon dog is licking a puddle of water.	vector illustration of a dog drinking water from a bowl vector illustration	
SEED-LLaMA	LLaVA-1.5	LLaVA-NeXT	
unanswerable.	The next image to be generated should be a close-up of the dog's tongue licking the water. The action should be the dog's tongue making contact with the water	and white dog.	
Qwen-VL	GPT-4V		
The next image should be of a dog swimming in a pool, with its tongue hanging out and its eyes closed in enjoyment.	A cartoon brown and white dog with a wagging tail, puddling in clear blue water, splashes around it, and a joyful expression on its face.		
Gemini	Claude		
A cartoon dog swimming in a lake.	The next image appears to show a cartoon dog happily playing in a puddle of water. The dog has a friendly, playful expression on its face as it splashes around, enjoying the water.		




  

Latent Var.	Input		
Zebra	Fly: 	Swim: 	Drink: 
Emu	Emu2	GILL	
In the next image, a zebra with wings is seen swimming in the ocean on an inner tube.	A zebra with sunglasses is riding on an inner tube.	person on a white background with a blue background vector illustration	
SEED-LLaMA	LLaVA-1.5	LLaVA-NeXT	
unanswerable.	The next image to be generated should be a cartoon zebra sitting on a colorful, rainbow-striped float in a body of water.	2 clouds in the sky.	
Qwen-VL	GPT-4V		
The next image should be of a zebra drinking from a water bottle, with its head tilted back and the bottle placed on a table.	A cute zebra cartoon character drinking water from a clear blue lake with ripples around the snout, reflecting the warm sunlight.		
Gemini	Claude		
A zebra drinking from a waterhole.	The next image appears to depict a colorful and whimsical scene of a zebra floating on a colorful inner tube in a body of water, surrounded by a bright, sunny sky.		

(a) Action-I

Latent Var.	Input		
Swim	Monkey: 	Sheep: 	Dog: 
Emu	Emu2	GILL	
The next image is a cartoon dog wearing a life preserver.	A cartoon sheep on a life preserver in water.	illustration of a cute cartoon dog with a rope on a	
SEED-LLaMA	LLaVA-1.5	LLaVA-NeXT	
I'm sorry, but I cannot provide the text prompt as it is not relevant to the image.	The next image to be generated should be a cartoon-style illustration of a dog wearing a life jacket and riding a surfboard on a wave.	tie around the monkey's waist.	
Qwen-VL	GPT-4V		
The next image should be of a dog swimming in a pool with a life preserver on.	A cartoon sheep with a fluffy white coat jumping into a red and white lifebuoy against a plain light blue background.		
Gemini	Claude		
A dog wearing a red life jacket is swimming in a lake. The dog is brown and white, and it is smiling. The water is blue and clear.	The next image appears to be a cartoon-style illustration of a dog. It may be depicted in a swimming or floating pose, surrounded by water and bubbles, suggesting a playful or aquatic scene.		

Latent Var.	Input		
Drink	Cat: 	Sheep: 	Zebra: 
Emu	Emu2	GILL	
A zebra is drinking water from a pond.	A sheep drinking water from a pond with a straw.	You can use the text box below to add more details.	
SEED-LLaMA	LLaVA-1.5	LLaVA-NeXT	
unanswerable.	The next image to be generated should be a black and white cartoon of a cat sitting on a table.	ish color.	
Qwen-VL	GPT-4V		
The next image should be a black and white striped zebra drinking water from a pond. The action would be of the zebra taking a sip of water	A zebra standing on a grassy plain with a clear blue sky in the background. The zebra's pose should suggest movement or grazing.		
Gemini	Claude		
A cartoon zebra wearing a pink tutu is drinking from a martini glass. The zebra is standing on its hind legs and has a straw in the glass.	The image shows a zebra standing in a body of water, with trees in the background. The zebra appears to be drinking from the water, which is reflected in the image.		

(b) Action-II

Figure 22: Examples of prompts and corresponding images description generated by MLLMs for tasks within the Action theme. For brevity, some lengthy responses from MLLMs have been truncated, retaining only the key parts of the responses.























































	Latent Var.	Input			Pretrained	Fine-tuned
Color-I	Ball	Brown: 	Blue: 	Green: 		
Color-II	Green	Bag: 	Box: 	Ball: 		
BG-I	Bird	Cave: 	Seafloor: 	Gym: 		
BG-II	Gym	Cat: 	Dog: 	Bird: 		
Style-I	Box	Old: 	3d: 	Graffiti: 		
Style-II	Graffiti	Bag: 	Book: 	Box: 		
Action-I	Cat	Cry: 	Eat: 	Fly: 		
Action-II	Fly	Cow: 	Dog: 	Cat: 		
Texture-I	Apple	Denim: 	Lace: 	Metal: 		
Texture-II	Metal	Bag: 	Book: 	Apple: 		


Figure 24: Comparison between images generated by the pretrained-only SEED-LLaMA and its fine-tuned counterpart. The shorthand 'BG' represents 'background'.


User

We provide a few examples, each of which is an input-output pair where the output is a description of the image associated with the input. Based on the examples, the task is to predict the next image description.

Blue: 

Brown: 

Orange: 

Black: 

Red:

Before predicting the next image, let's first think step by step and analyze the relationship between the text input and image output in each example.

Qwen-VL

Sure, let's go through each example and analyze the relationship between the text input and image output.

1. blue: Picture 1: This image is a blue picture of a leaf. The text input is "blue" and the image output is a blue leaf. The relationship between the text input and image output is that the text input matches the color of the image output. 2. brown: Picture 2: This image is a brown picture of a leaf. The text input is "brown" and the image output is a brown leaf. The relationship between the text input and image output is that the text input matches the color of the image output. 3. orange: Picture 3: This image is an orange picture of a leaf. The text input is "orange" and the image output is an orange leaf. The relationship between the text input and image output is that the text input matches the color of the image output. 4. black: Picture 4: This image is a black picture of a leaf. The text input is "black" and the image output is a black leaf. The relationship between the text input and image output is that the text input matches the color of the image output.

Therefore, based on the examples, the task is to predict the next image description. Since we have seen four examples, we can make a prediction for the next image description by considering the color of the leaf in each example. For example, if the next image description is "a leaf with a white background", we can choose the image with the white background as the next image description.

User

Based on the analysis, please describe what the next image should be look like given the request 'red: .'





Qwen-VL

Given the request "red: ", we can make a prediction by considering the color of the leaf in each example. Since we have seen four examples, we can choose the image with the red color as the next image description. Therefore, the next image description should be "a red leaf with a white background".

Figure 25: A four-shot example of outputs generated by Qwen-VL using CoT prompting for task Color-I.

User

We provide a few examples, each of which is an input-output pair where the output is a description of the image associated with the input. Based on the examples, the task is to predict the next image description.

Cow:  Monkey:  Pig:  Dog:  Sheep:

Before predicting the next image, let's first think step by step and analyze the relationship between the text input and image output in each example.

Qwen-VL

Sure, I'll do my best to analyze the relationship between the text input and image output in each example. Please provide me with the examples.

User

Based on the analysis, please describe what the next image should look like given the request 'sheep: '.



Qwen-VL

I'm sorry, but I don't have any examples to base my analysis on. Could you please provide me with the text inputs and image outputs for each example so I can analyze the relationship between the text input and image output?

Figure 26: A four-shot example of outputs generated by Qwen-VL using CoT prompting for task Action-II.

User

We provide a few examples, each of which is an input-output pair where the output is a description of the image associated with the input. Based on the examples, the task is to predict the next image description.

Book:  Chair:  Ball:  Leaf:  Cup:

Before predicting the next image, let's first think step by step and analyze the relationship between the text input and image output in each example.

Qwen-VL

Sure, let's go through each example and analyze the relationship between the text input and image output.

1. book: Picture 1: This example is about a book with a lace cover. The cover is described as being embossed with flowers and leaves. The image output is a book with a lace cover, which is consistent with the text input.
2. chair: Picture 2: This example is about a chair with a lace cover. The cover is described as being white and made of lace. The image output is a chair with a lace cover, which is consistent with the text input.
3. ball: Picture 3: This example is about a ball with a lace pattern. The ball is described as being covered in a floral lace pattern. The image output is a ball covered in a floral lace pattern, which is consistent with the text input.
4. leaf: Picture 4: This example is about a leaf with a lace pattern. The leaf is described as being embossed with a delicate lace pattern. The image output is a leaf with a delicate lace pattern, which is consistent with the text input.

In each example, the image output is a visual representation of the text input. The text inputs describe objects or patterns, and the image outputs are images that match the described objects or patterns.

User

Based on the analysis, please describe what the next image should look like given the request 'cup: '.

Qwen-VL

Based on the analysis, the next image should be a cup with a lace pattern. The cup could be any type of cup, such as a coffee cup, a tea cup, or a decorative cup. The lace pattern could be any type of lace, such as floral lace, geometric lace, or abstract lace. The color of the cup could be any color, such as white, blue, pink, or green. The image should be a clear and detailed representation of a cup with a lace pattern.

Figure 27: A four-shot example of outputs generated by Qwen-VL using CoT prompting for task Texture-II.

User

We provide a few examples, each of which is an input-output pair where the output is a description of the image associated with the input. Based on the examples, the task is to predict the next image description.

Blue:  Brown:  Orange:  Black:  Red:

Before predicting the next image, let's first think step by step and analyze the relationship between the text input and image output in each example.

SEED-LLaMA

Yes, let's analyze the relationship between the text input and the image output in each example. In the first example, the text input is "blue" and the image output is a picture of a leaf. The second example has the text input "brown" and the image output is a picture of a leaf. The third example has the text input "orange" and the image output is a picture of a leaf. The fourth example has the text input "brown" and the image output is a picture of a leaf. The fifth example has the text input "orange" and the image output is a picture of a leaf. The sixth example has the text input "red" and the image output is a picture of a leaf. The seventh example has the text input "black" and the image output is a picture of a leaf. The eighth example has the text input "orange" and the image output is a picture of a leaf. The ninth example has the text input "red" and the image output is a picture of a leaf. The tenth example has the text input "brown" and the image output is a picture of a leaf.

From these examples, we can see that the text input is related to the color of the leaf in the image output. The input color is used to describe the color of the leaf in the output image.

User

Based on the analysis, please describe what the next image should be look like given the request 'red: '

SEED-LLaMA

I have generated the next image. 


Figure 28: A four-shot example of outputs generated by SEED-LLaMA using CoT prompting for task Color-I.

We provide a few examples, each of which is an input-output pair where the output is a description of the image associated with the input. Based on the examples, the task is to predict the next image description.


Cat:




Cow:



Lion:



Bird:



Sheep:

Before predicting the next image, let's first think step by step and analyze the relationship between the text input and image output in each example.

User

SEED-LLaMA

Sure, let's take a closer look at the relationship between the text input and the image output in each example.

1. In the first example, the text input is "cat" and the image output is a cat walking on the beach. The image shows a cat walking on the sandy beach near the water. 2. In the second example, the text input is "cow" and the image output is a cow walking on the beach. The image shows a cow walking on the sandy beach near the water. 3. In the third example, the text input is "lion" and the image output is a lion walking on the beach. The image shows a lion walking on the sandy beach near the water. 4. In the fourth example, the text input is "bird" and the image output is a bird walking on the beach. The image shows a bird walking on the sandy beach near the water. 5. In the fifth example, the text input is "sheep" and the image output is a sheep walking on the beach. The image shows a sheep walking on the sandy beach near the water.

From these examples, we can see that the text input and the image output are related in that they both describe an animal walking on the beach. The animals in the image are all different, but they all share the common trait of being on the beach.

Based on the analysis, please describe what the next image should look like given the request 'sheep: '

User

SEED-LLaMA

I have generated an image.









Figure 29: A four-shot example of outputs generated by SEED-LLaMA using CoT prompting for task Background-II.



User

We provide a few examples, each of which is an input-output pair where the output is a description of the image associated with the input. Based on the examples, the task is to predict the next image description.

Wireframe:  Futuristic:  Icon:  Old:  Pixel: 

Before predicting the next image, let's first think step by step and analyze the relationship between the text input and image output in each example.

SEED-LLaMA

Yes, let's analyze the relationship between the text input and the image output in each example.  
 1. In the first example, the text input is "wireframe" and the image output is a wireframe of a leaf. This means that the image is a wireframe representation of a leaf. 2. In the second example, the text input is "futuristic" and the image output is a futuristic leaf. This means that the image is a futuristic representation of a leaf. 3. In the third example, the text input is "icon" and the image output is a leaf icon. This means that the image is a representation of a leaf as an icon. 4. In the fourth example, the text input is "old" and the image output is a leaf with age. This means that the image is a representation of a leaf that has aged. 5. In the fifth example, the text input is "pixel" and the image output is a pixelated leaf. This means that the image is a pixelated representation of a leaf. 6. In the sixth example, the text input is "futuristic" and the image output is a futuristic leaf. This means that the image is a futuristic representation of a leaf. 7. In the seventh example, the text input is "pixel" and the image output is a pixelated leaf. This means that the image is a pixelated representation of a leaf. 8. In the eighth example, the text input is "wireframe" and the image output is a wireframe of a leaf. This means that the image is a wireframe representation of a leaf. 9. In the ninth example, the text input is "old" and the image output is a leaf with age. This means that the image is a representation of a leaf that has aged. 10. In the tenth example, the text input is "pixel" and the image output is a pixelated leaf. This means that the image is a pixelated representation of a leaf.  
 In summary, the relationship between the text input and the image output in each example is that the text input is used to describe the type of representation of the leaf in the image output.

User

Based on the analysis, please describe what the next image should be look like given the request 'pixel: '

SEED-LLaMA

Here is the next image based on the request 'pixel: '




Figure 30: A four-shot example of outputs generated by SEED-LLaMA using CoT prompting for task Style-I.