# LUMINA-MGPT: ILLUMINATE FLEXIBLE PHOTOREALISTIC TEXT-TO-IMAGE GENERATION WITH MULTIMODAL GENERATIVE PRETRAINING

**Anonymous authors**
Paper under double-blind review

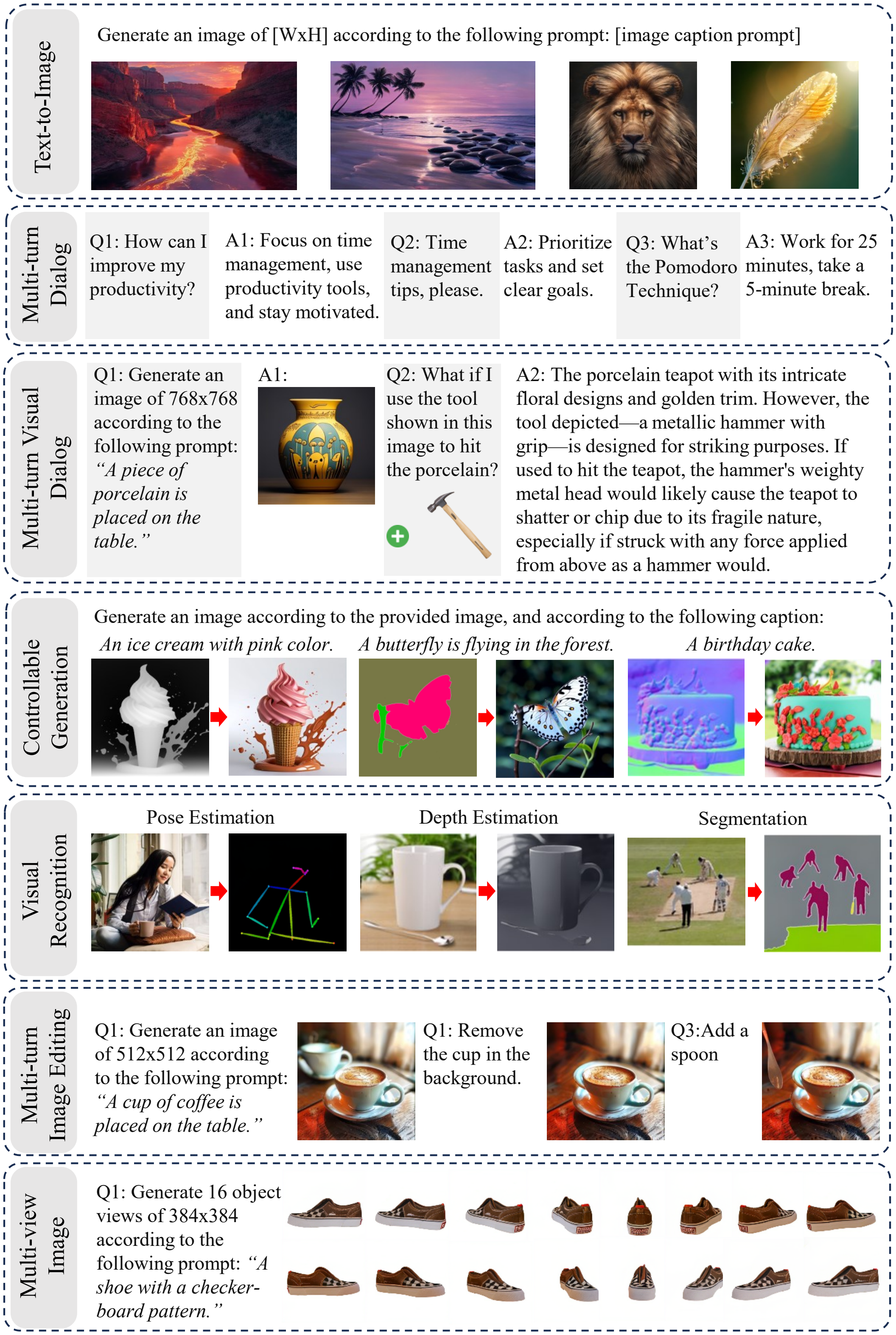
## ABSTRACT

We present Lumina-mGPT, a family of multimodal autoregressive models capable of various vision and language tasks, particularly excelling in generating flexible photorealistic images from text descriptions. By initializing from ***m**ultimodal* ***G**enerative* ***P**re**T**raining* (mGPT), Lumina-mGPT demonstrates that decoder-only Autoregressive (AR) model can achieve image generation performance comparable to modern diffusion model with high efficiency through *Flexible Progressive Supervised Finetuning* (FP-SFT). Equipped with our proposed ***Un**ambiguous* ***i**mage* ***Rep**resentation* (Uni-Rep), Lumina-mGPT can flexibly generate high-quality images of varying aspect ratios. Building on the strong image generation capabilities, we further explore *Ominiponent Supervised Finetuning* (Omni-SFT), an initial attempt to elevate Lumina-mGPT into a unified multi-modal generalist. The resulting model demonstrates versatile multimodal capabilities, including visual generation tasks like text-to-image/multiview generation and controllable generation, visual recognition tasks like segmentation and depth estimation, and vision-language tasks like multi-turn visual question answering, casting light on the rosy potential of this direction. We release all code and checkpoints, hoping to facilitate the progress toward building artificial general intelligence.

## 1 INTRODUCTION

Seminal models, including DALL-E 3 (Betker et al., 2023), Stable Diffusion 3 (Esser et al., 2024), and SoRA (Brooks et al., 2024), have demonstrated superior performance in photorealistic image and video generation using diffusion-based generative modeling over continuous latent image features. In contrast, autoregressive (AR) generative models, which rely on “next-token prediction,” have revolutionized text generation with groundbreaking reasoning abilities, as exemplified by models like GPT-4 (Achiam et al., 2023) and Gemini (Team et al., 2023), over discrete token representation. However, AR-based generative modeling over vector-quantized image features still lags far behind diffusion-based counterparts in terms of photorealistic and controllable image generation.

Although previous autoregressive efforts, such as DALL-E (Ramesh et al., 2021), CogView (Ding et al., 2021), Parti (Yu et al., 2022), OFA (Wang et al., 2022), Unified-IO (Lu et al., 2022; 2024a), LlamaGen (Sun et al., 2024), and Chameleon (Team, 2024a), have extensively explored generative modeling over vector-quantized image features following the paradigm of large language models (LLMs), their results on text-to-image generation have either been unsatisfactory or produced high-quality samples limited to academic benchmarks like ImageNet (Deng et al., 2009). We summarize the limitations of previous approaches: (1) AR-based image generation models often rely on randomly-initialized transformers, resulting in poor image quality and slow convergence due to the lack of pretrained representations; (2) the verbose encoder-decoder architectures used complicate the framework and hinder scalability and generalization to other tasks; (3) these models are limited in decoding resolution and flexibility, often producing low-quality images with fixed resolutions; (4) they lack task extensibility, restricting their applicability to a narrow range of scenarios without integration with other tasks such as dense labeling and controllable image generation. More detailed discussion of these challenges can be found in Appendix C.



Generate an image of [WxH] according to the following prompt: [image caption prompt]

Q1: How can I improve my productivity?

A1: Focus on time management, use productivity tools, and stay motivated.

Q2: Time management tips, please.

A2: Prioritize tasks and set clear goals.

Q3: What's the Pomodoro Technique?

A3: Work for 25 minutes, take a 5-minute break.

Q1: Generate an image of 768x768 according to the following prompt: *"A piece of porcelain is placed on the table."*

A1:

Q2: What if I use the tool shown in this image to hit the porcelain?

A2: The porcelain teapot with its intricate floral designs and golden trim. However, the tool depicted—a metallic hammer with grip—is designed for striking purposes. If used to hit the teapot, the hammer's weighty metal head would likely cause the teapot to shatter or chip due to its fragile nature, especially if struck with any force applied from above as a hammer would.

Generate an image according to the provided image, and according to the following caption:

*An ice cream with pink color.* *A butterfly is flying in the forest.* *A birthday cake.*

Q1: Generate an image of 512x512 according to the following prompt: *"A cup of coffee is placed on the table."*

Q1: Remove the cup in the background.

Q3:Add a spoon

Q1: Generate 16 object views of 384x384 according to the following prompt: *"A shoe with a checker-board pattern."*

Figure 1: Unified in the next-token prediction framework, Lumina-mGPT can perform a wide range of multi-modal tasks. See Figure 13 to Figure 18 in the Appendix for more demonstrations.

Table 1: Overview of the design choices and capabilities of multimodal autoregressive approaches. Lumina-mGPT is the only model capable of both flexible photorealistic image generation and multimodal task unification, due to its decoder-only transformer design and multimodal generative pretraining.

| Method | Model Architecture | Multimodal Pretraining | Flexible Photorealistic Image Generation | Task Extensibility |
|---|---|---|---|---|
| DALL-E (Ramesh et al., 2021) | Decoder-only | × | × | × |
| Cogview (Ding et al., 2021) | Decoder-only | × | × | × |
| Parti (Yu et al., 2022) | Encoder-Decoder | × | × | × |
| LlamaGen (Sun et al., 2024) | Encoder-Decoder | × | × | × |
| OFA (Wang et al., 2022) | Encoder-Decoder | ✓ | × | ✓ |
| Unified-IO (Lu et al., 2022) | Encoder-Decoder | × | × | ✓ |
| Unified-IO 2 (Lu et al., 2024a) | Encoder-Decoder | ✓ | × | ✓ |
| CM3Leon (Yu et al., 2023) | Decoder-only | ✓ | × | ✓ |
| Chameleon (Team, 2024a) | Decoder-only | ✓ | × | × |
| **Lumina-mGPT** | Decoder-only | ✓ | ✓ | ✓ |

To address the aforementioned challenges, we present Lumina-mGPT, a decoder-only transformer initiated with effective ***m**ultimodal **G**enerative **P**re**T**raining* (mGPT) and then supervised-finetuned over flexible, high-quality, high-resolution discrete image tokens in a progressive manner. This framework illuminates flexible high-resolution photorealistic image generation and can be easily extended to solve various component tasks in a unified manner. We provide a detailed comparison of the architecture design choices and model capabilities of existing multimodal autoregressive approaches in Table 1. The key features of Lumina-mGPT are outlined below:

① ***Effective Multimodal Generative Pretraining:*** Unlike commonly adopted approaches that utilize a randomly-initialized causal transformer to generate discrete image tokens in an autoregressive manner, Lumina-mGPT is initialized from an effective **m**ultimodal **G**enerative **P**ret**T**raining (mGPT) representation, which is a multimodal autoregressive transformer trained at large scale using a simple "next-token prediction" loss. We find that by leveraging the rich knowledge from large-scale multimodal pretraining, mGPT can learn broad and general multimodal representation. This serves as an excellent starting point, remarkably simplifying the task of text-to-image generation, accelerating training convergence on downstream tasks, and unifying various vision-language tasks. Due to the lack of training resources, this mGPT representation is directly adopted from the pretrained Chameleon 7B and 30B models released by Meta (Team, 2024a).

② ***Simple Decoder-Only Architecture:*** Lumina-mGPT adopts the simple decoder-only architecture. Compared to more complex designs like encoder-decoder architectures, decoder-only models have a significant innate advantage: they provide a simple, elegant, and extensible framework to unify various understanding and generation tasks across different modalities. This makes them a promising candidate for achieving true unification. Besides, by sharing the same architecture as the flourishing text-only LLMs, Lumina-mGPT can benefit from well-established theories and infrastructures in the LLM community, such as scaling properties (Chowdhery et al., 2022; Brown et al., 2020) and techniques to optimize the training and inference processes (Dao et al., 2022; Dao, 2024; Kwon et al., 2023). We train a family of Lumina-mGPTs with up to 30 billion parameters to provide a flexible trade-off between efficiency and performance and to explore the limit of its multimodal capabilities as model parameters scale.

③ ***Flexible High-quality Image Generation:*** Despite the aforementioned strengths of decoder-only architecture, the image generation capabilities of such models remain limited, creating a gap between the potential and the reality of this architecture. We thus propose Flexible Progressive Supervised Finetuning (FP-SFT) to fully fulfill the potential of high-quality text-to-image generation. This approach starts with low-resolution discrete tokens and progressively transitions to high-resolution discrete tokens. Combined with ***Un**ambiguous **i**mage **Rep**resentation* (Uni-Rep), this weak-to-strong SFT strategy effectively grants the model with the ability to generate high-quality photorealistic images with flexible aspect ratios.

④ ***Omnipotent Task Unification:*** The high-quality image generation capabilities achieved during the FP-SFT stage provide the prerequisites for further exploring the fundamental advantages of the decoder-only architecture. We thus propose Omnipotent Supervised Finetuning (Omni-SFT), a preliminary attempt to create an omnipotent generalist. Specifically, Omni-SFT treats various tasks– such as multi-turn dialog, visual multi-turn understanding, dense labeling, text-to-image generation, text-to-multiview generation, image editing, and spatial-conditional image generation–as a unified discrete modeling task, allowing the model to achieve omnipotent task unification via a natural language interface.

We summarize our contributions as follows: (1) We are the first (especially in the open-source domain) to demonstrate that a decoder-only AR model can achieve image generation performance comparable to modern diffusion models. Furthermore, by initializing from mGPT, the aforementioned capability can be achieved at low computational cost (32 A100 GPUs × 7 days for 7B model). (2) We propose UniRep, an image representation that empowers decoder-only AR models with the ability to flexibly generate images of varying aspect ratios. (3) Building on the strong image generation capabilities, we further explore Omni-SFT, an initial attempt to elevate the model into a unified generalist. Experiments underscore the promising potential of this direction. (4) We open-source the entire pipeline to encourage the community's further exploration of this topic.

## 2 METHODOLGY

Lumina-mGPT is a decoder-only transformer initialized with multimodal Generative PreTraining (mGPT) and finetuned over high-quality multimodal tokens derived from various tasks. Based on the robust mGPT representation and our proposed supervised finetuning strategies with unambiguous image representation, Lumina-mGPT achieves superior performance in photorealistic image generation and omnipotent task unification with high flexibility in image resolution and aspect ratio. In this section, we first introduce mGPT, followed by the training and inference details of Lumina-mGPT.

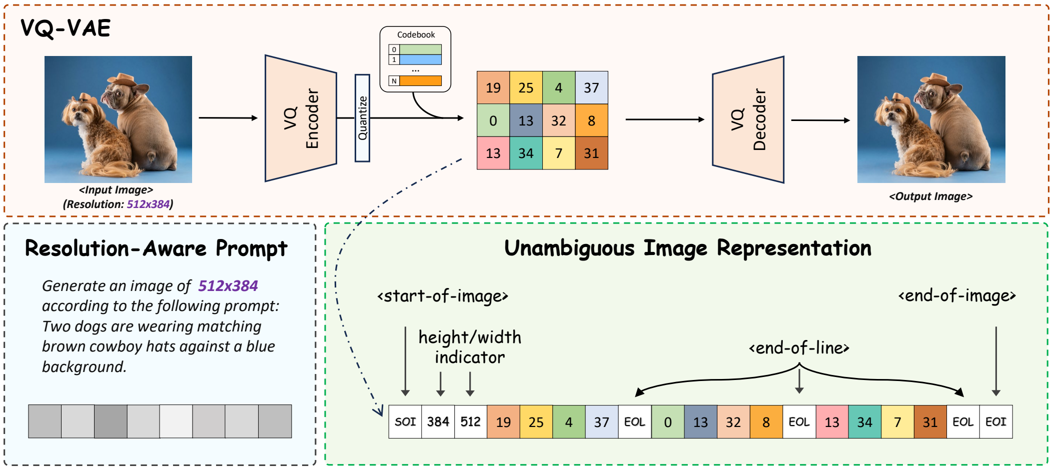
### 2.1 REVISITING MGPT WITH CHAMELEON

mGPT represents the family of models utilizing a decoder-only transformer architecture, pretrained on extensive multimodal token sequences. These models exhibit exceptional native multimodal understanding, generation, and reasoning capabilities, offering the potential for universal modeling across various modalities and tasks. We use the recent open-source model, Chameleon (Team, 2024a), as an example to illustrate the design choices and implementing details of mGPT.

**Multimodal Tokenization** To unify text and images into a multimodal token sequence, it is essential first to tokenize both text and image into discrete space. Especially for images, the choice of tokenizer is crucial as it determines the upper limit of generation quality. Specifically, Chameleon trains a new byte pair encoding tokenizer for text. For images, it adopts the quantization-based tokenization method following prior works (Esser et al., 2021; Yu et al., 2022; Ramesh et al., 2021), converting continuous image patches into discrete tokens from a fixed codebook while reducing spatial dimensions. The quantized image tokens are then flattened into a 1D sequence and concatenated with text tokens in various ways to form a multimodal token sequence for unified modeling.

**Decoder-Only Transformer** Unlike Unified-IO and Parti using an encoder-decoder architecture with pretrained encoders, mGPT trains a decoder-only autoregressive transformer from scratch after transforming text and image inputs into a unified sequence of discrete tokens $x = (x_1, x_2, ..., x_T)$, leading to a simpler and more unified approach for multimodal generative modeling. mGPT adopts a standard dense transformer architecture for scalability and generalizability, with minor adaptations like RoPE (Su et al., 2024) and SwiGLU (Shazeer, 2020) activation function, following the LLaMA series (Touvron et al., 2023). However, this standard transformer architecture exhibits training instability when scaling up the model size and context length of multimodal token sequences, as observed in existing works (Team, 2024a; Zhuo et al., 2024). It has been found that this instability is caused by the uncontrollable growth of network activations in transformer blocks. Therefore, Pre-Norm, Post-Norm, and QK-Norm (Henry et al., 2020) are added to each transformer block to preserve the magnitude of intermediate activation and further stabilize the training process.

**Training Objective** During training, mGPT models the conditional probability $p(x_t|x_1, ..., x_{t-1})$ of multimodal sequences using the standard next-token prediction objective. Additionally, Chameleon applies z-loss (Chowdhery et al., 2022) to stabilize the training of 7B and 30B models. Initially, we



*Generate an image of* ***512x384*** *according to the following prompt: Two dogs are wearing matching brown cowboy hats against a blue background.*

Figure 2: Illustration of Resolution-Aware Prompt (bottom left) and Unambiguous Image Representation (bottom right). These designs are used in all supervised finetuning stages to eliminate the ambiguity in image representation, enabling flexible resolution image modeling.

underestimated the importance of z-loss in our attempts because it is absent in most works relevant to the training of (multimodal) LLMs. However, we found that without this term, the magnitude of logits surges, causing the loss to diverge. On the other hand, with z-loss, we observe that for the 7B and 30B models, the optimal temperature for inference-time image generation is much lower, as the magnitude of logits is significantly reduced in the larger models.

**Limitations of Chameleon** Although Chameleon demonstrates potential for joint image and text understanding within one decoder-only transformer, its image generation ability remains inferior to state-of-the-art diffusion-based frameworks (Esser et al., 2024; Chen et al., 2024a; Li et al., 2024c; Team, 2024b; Zhuo et al., 2024) in both quality and resolution flexibility. Moreover, it is worth noting that the image generation ability is even absent in the open-source version of Chameleon. Additionally, the capabilities of Chameleon are confined to vision-language and text-only tasks, excluding a broader range of vision-centric tasks. These include classic visual recognition tasks such as segmentation and depth prediction, as well as creative visual generation tasks like controllable generation and image editing. Lumina-mGPT is built upon Chameleon to unlock its full potential for flexible photorealistic image generation and to become a versatile vision generalist.

## 2.2 LUMINA-MGPT

### 2.2.1 EFFECTIVE INITLIZATION

Large-scale pre-training and scalable model architecture have been widely verified as the golden path to advanced intelligence. As mGPTs like Chameleon are pretrained on large-scale image-text interleaved datasets and have developed effective and generalizable representations for both image and text, they can better serve as the starting point for flexible photorealistic image generation and beyond than random initialization or language-only models. Furthermore, the LLaMA (Touvron et al., 2023) architecture integrated with features like QK-Norm has demonstrated strength and scalability through extensive validation (Touvron et al., 2023; Meta, 2024; Sun et al., 2024; Gao et al., 2024). By initializing from the Chameleon mGPT, which adheres to the LLaMA architecture, we can leverage these architectural advantages. Consequently, initializing from mGPTs allowed us to efficiently train high-performing Lumina-mGPT models, with parameters ranging from 7B to 30B, using just 10M high-quality image-text data points.

### 2.2.2 SUPERVISED FINETUNING FOR LUMINA-MGPT

**Unambiguous Image Representation** Existing methods represent images as 1D flattened sequences of 2D discrete image codes. While adequate for fixed resolutions, this approach becomes ambiguous when supporting variable resolutions, as with Lumina-mGPT. For instance, images with resolutions of 512 × 512, 256 × 1024, and 1024 × 256 can all be encoded into the same number of tokens, making it impossible to infer the original shape without examining the token contents. This ambiguity poses significant challenges for both image perception and generation.

To address this problem, we propose **Un**ambiguous **i**mage **Rep**resentation (Uni-Rep), which augments the image representations by adding extra height/width indicator tokens immediately after the `<start-of-image>` token and inserting `<end-of-line>` tokens after image tokens belonging to the same row. As shown in Figure 2, this modification ensures that the original shape of the images can be accurately parsed from the 1D representation without additional context or delving into the contents of the image tokens. This enhancement provides the foundation for Lumina-mGPT's ability to perform image-related tasks at any resolution and aspect ratio.

Note that while either the height/width indicators or the `<end-of-line>` tokens can independently achieve disambiguation, we still use both simultaneously because they have distinct benefits. When generating images, the height/width indicators, generated before any image tokens, pre-determine the shape of the image, aiding Lumina-mGPT in composing the image contents. On the other hand, the `<end-of-line>` tokens can serve as anchors, offering the 1D tokens sequence with additional explicit spatial information. We detailed in Section 3.3 to illustrate the role of these indicators.

**Flexible Progressive Supervised Finetuning (FP-SFT)** The FP-SFT process equips the pretrained mGPT with the capability to generate high-resolution images with flexible aspect ratios in a progressive manner. The process is divided into three stages, where the product of width and height approximates $512^2$, $768^2$, and $1024^2$, respectively. In each stage, a set of candidate resolutions with similar areas but different height-width ratios are prepared, and each image is matched to the most suitable resolution. In the low-resolution stage, shorter sequence lengths and the resulting high training throughput allow the model to quickly traverse a large amount of data, learning the general composition of images and a broad spectrum of visual concepts. Conversely, in the high-resolution stage, the model is expected to focus on learning high-frequency fine-grained details unique to high-resolution images. Benefiting from the strong foundation built during the high-throughput pretraining and low-resolution finetuning stages, the low-throughput high-resolution finetuning stage is data-efficient, thereby enhancing the overall efficiency of the FP-SFT process.

A meticulously curated dataset of high-resolution photorealistic image-text pairs is used for FP-SFT. Moreover, the pure-text data from OpenHermess (Teknium, 2023) and the image-to-text data from Mini-Gemini (Li et al., 2024b) are also incorporated during training to prevent catastrophic forgetting. To provide users with a natural way to specify the desired resolution of generated images, we developed the resolution-aware prompt (Figure 2). For each image and its corresponding description, the prompt is structured as follows:

```
 Generate an image of {width}x{height} according to the following
prompt:  \n {description}
```

**Omnipotent Supervised Finetuning (Omni-SFT)** While flexible photorealistic image generation is the primary target of Lumina-mGPT, we find that the resulting model after FP-SFT can be efficiently transferred to a wide spectrum of image understanding and generation tasks. We thus present Omni-SFT, a preliminary exploration toward boosting Lumina-mGPT to a visual generalist. Training tasks and data for Omni-SFT consists of the following:

1. Single- and multi-turn language-guided image-editing with data from MagicBrush (Zhang et al., 2024) and SEED (Ge et al., 2024) (only involving the real-world and multi-turn subsets).

2. Dense prediction tasks, including surface norm estimation from NYUv2 Silberman et al. (2012) and ScanNet Dai et al. (2017), depth estimation from Kitti v2 Cabon et al. (2020) and Sintel Butler et al. (2012), pose estimation from MSCOCO Lin et al. (2014), semantic segmentation data annotated with OneFormer (Jain et al., 2023) on image from Laion (Schuhmann et al., 2022), and grounding data from RefCOCO (Kazemzadeh et al., 2014).

3. In-house spatial-conditional image generation following ControlNet (Zhang et al., 2023), with conditions including surface norm, depth, pose, and segmentation.

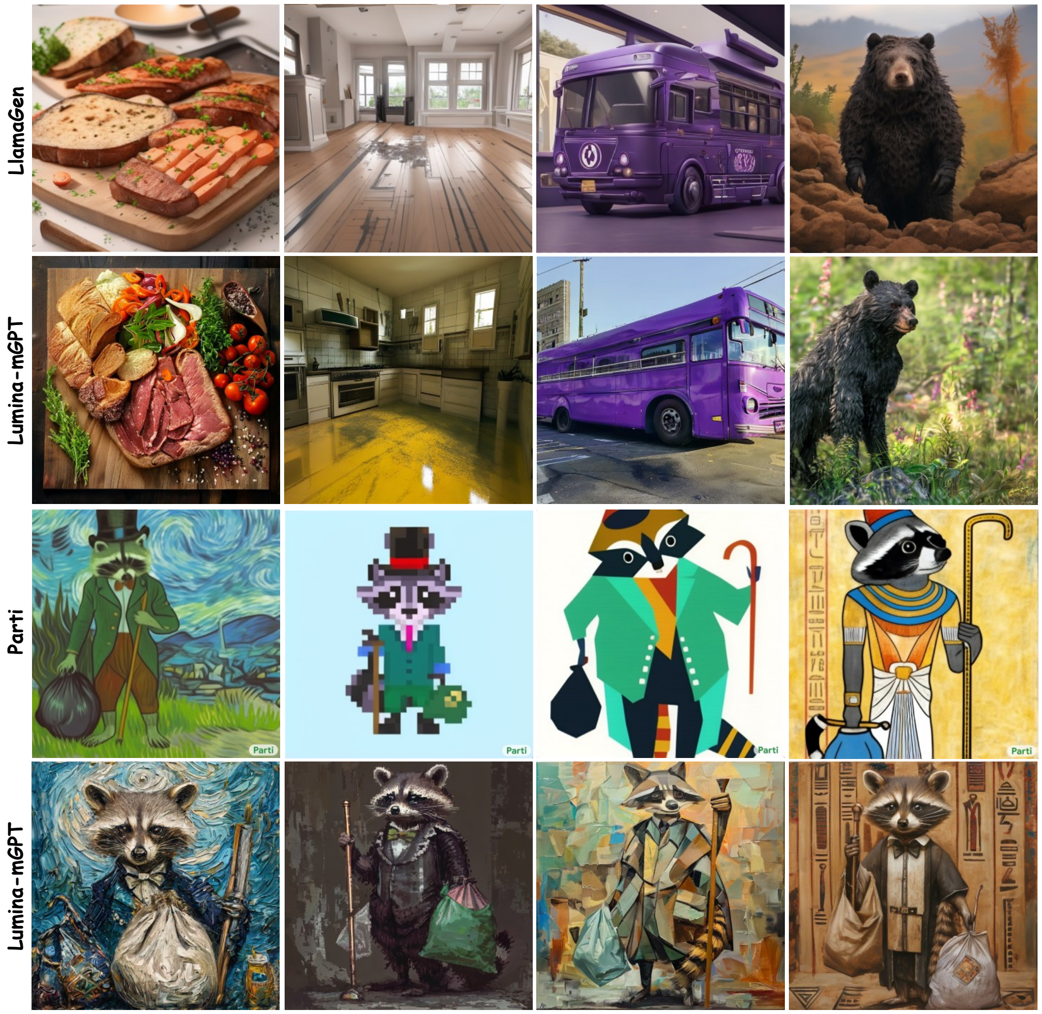


Figure 3: Qualitative comparison with LlamaGen and Parti. Lumina-mGPT can generate more photorealistic images with high aesthetics and fine-grained details.

4. Text-conditional multiview generation using an internal dataset consisting of 100k high-quality samples with rendered $384^2$ images from 16 viewpoints uniformly distributed in azimuth angles.

5. A small fraction of data sampled from those used in the previous FP-SFT process, including both text modeling and text-to-image generation to maintain its learned capabilities.

We tokenize all text and images into discrete tokens and formulate these tasks as a unified next-token prediction objective. Notably, we also incorporate tasks multiview generation, which requires generating a sequence of image frames, as a preliminary for video generation. As demonstrated in Section 3.2, after Omni-SFT, Lumina-mGPT exhibits a general capability for completing a wide range of tasks other than text-to-image generation, indicating the potential for building a multimodal generalist along this direction.

**Training Setup** Though multiple tasks are involved in the SFT process, a unified next-token-prediction loss is used for all of the tasks. As Lumina-mGPT is designed as a chat model, all data are organized into single or multi-turn dialogs, with the loss applied only to the response parts. For all experiments, the AdamW (Loshchilov & Hutter, 2017) optimizer with weight decay = $0.1$ and betas = $(0.9, 0.95)$ is used, and the learning rate is set to 2e-5. For stabilizing training, z-loss is applied with weight 1e-5 for both the 7B and the 30B model, and for the 7B model dropout with probability 0.05 is additionally applied. Inspired by the classifier-free guidance in diffusion models (Ho &

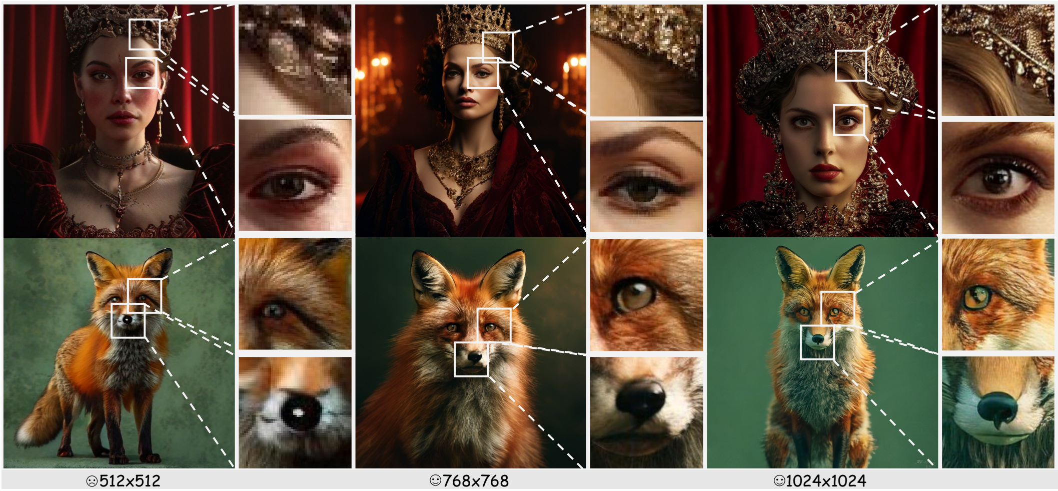


Figure 4: Samples with zoom-ins generated by Lumina-mGPT in different resolution finetuning stages. The visual details continuously improve along with the progressively increasing resolution.

Salimans, 2022), we randomly drop the context by a probability of 10% during training, as detailed in Appendix E. To accommodate the large model volume, PyTorch FSDP (Zhao et al., 2023) is employed with gradient checkpointing. To increase training throughput, all data are pre-tokenized before training and are clustered according to the number of tokens, ensuring that each global batch is composed of data with similar lengths.

## 3 EXPERIMENTS

### 3.1 FUNDAMENTAL PHOTOREALISTIC TEXT-TO-IMAGE GENERATION

We first demonstrate the fundamental text-to-image generation capabilities of Lumina-mGPT with FP-SFT. As shown in Figure 13, Lumina-mGPT can generate photorealistic images in a variety of resolutions, achieving the first native 1K autoregressive generation without cascaded models, a common technique in text-to-mage generation (Yu et al., 2022; Saharia et al., 2022; Chang et al., 2023; Pernias et al., 2024). These generated images exhibit strong semantic coherence with intricate visual details, despite being finetuned on limited computational resources and text-image pairs.

**Qualitative Comparison with SoTA AR-based Approaches** We compare the text-to-image synthesis ability of Lumina-mGPT with LlamaGen (Sun et al., 2024) and Parti (Yu et al., 2022). LlamaGen beats state-of-the-art diffusion models on ImageNet FID score. Compared with LlamaGen, Lumina-mGPT can achieve better visual quality as shown in Figure 3 in text-to-image generation. Note that Lumina-mGPT only requires 10M image-text pairs while LlamaGen is trained over 50M low-quality image-text pairs accompanied by 10M in-house aesthetic image-text pairs. Compared to Parti, an AR text-to-image model with 20 billion parameters, Lumina-mGPT also demonstrates better visual quality and aesthetics. However, due to significant differences in computational costs and training datasets, Lumina-mGPT demonstrates inferior text instruction following ability compared to Parti. In addition, neither LlamaGen nor Parti supports the end-to-end generation of 1K resolution images with arbitrary aspect ratios, as achieved by Lumina-mGPT. LlamaGen only supports a fixed resolution of 512 × 512, while Parti generates 1024 × 1024 images using an additional super-resolution upsampler. Beyond AR-based approaches, we also provide a side-by-side comparison with diffusion-based counterparts by training on the same dataset, as detailed in Appendix F.

**Quantitative Comparison on text-to-image benchmarks** We evaluate Lumina-mGPT on the popular text-to-image benchmarks T2I-CompBench (Huang et al., 2023), GenEval (Ghosh et al., 2023), and DPG-Bench (Hu et al., 2024) to objectively demonstrate its performance. Results are shown in Tab. 3.1. Lumina-mGPT shows clear improvement over Chameleon. Furthermore, Lumina-mGPT outperforms Lumina-Next, a modern diffusion transformer trained using the same text-to-image data as Lumina-mGPT, and we guess that generative pre-training may be the reason behind this performance gap. Additionally, Lumina-mGPT shows competitive performance compared to SDXL, while maintaining a stable advantage over SDv2.1, providing an intuitive picture of Lumina-mGPT's position in the field.

Table 2: Quatitative results of Lumina-mGPT-7B on text-to-image benchmarks.

| | **T2I-CompBench** | | | **GenEval** | **DPG-Bench** |
|---|---|---|---|---|---|
| | Color | Shape | Texture | Overall | Average |
| Lumina-Next (Zhuo et al., 2024) | 0.5088 | 0.3386 | 0.4239 | 0.46 | 75.66 |
| SDv2.1 (Rombach et al., 2022) | 0.5694 | 0.4495 | 0.4982 | 0.50 | - |
| SDXL (Podell et al., 2023) | 0.6369 | 0.5408 | 0.5637 | 0.55 | 74.65 |
| Chameleon (Team, 2024a) | - | - | - | 0.39 | - |
| Lumina-mGPT (ours) | 0.6371 | 0.4727 | 0.6034 | 0.56 | 79.68 |

**On the Effectiveness of FP-SFT** To further validate the effectiveness of FP-SFT, we visualize the images generated at different finetuning stages in Figure 4. With increasing image resolution, we observe a progressive decrease in visual artifacts introduced by VQ-VAE and the emergence of diverse fine-grained visual details. From these illustrations, we can conclude that our FP-SFT can unleash the potential of generating high-quality images from mGPT in a progressive manner.

**Decoding Configuration Matters** In Appendix E, we highlight that optimal hyperparameters differ between text and image generation, with a status-aware control mechanism switching settings based on the type of content being generated. We then explore how different inference configurations, such as temperature, top-k, and classifier-free guidance scale, affect the quality of generated images. For example, lower temperatures and top-k values often result in over-smoothed images, while higher values enhance detail but may introduce artifacts. Inspired by these observations, we leverage min-p sampling (Nguyen et al., 2024) for visual generation and find it works significantly better compared to standard top-k sampling by striking the balance between coherence and diversity.

Table 3: Quatitative results of Lumina-mGPT-7B on comprehensive VQA benchmarks.

| | MMBench | MME-p | MME-r | SEEDBench-I | MMMU-val | POPE |
|---|---|---|---|---|---|---|
| Chameleon | 19.80 | 153.10 | 49.60 | 30.50 | 22.40 | 19.40 |
| Lumina-mGPT | **32.20** | **976.85** | **290.36** | **50.93** | **27.11** | **70.43** |

### 3.2 OMNIPOTENT TASK UNIFICATION WITH LUMINA-MGPT

By applying Omni-SFT over FP-SFT, Lumina-mGPT demonstrates a multitude of capabilities, which can be categorized into text-only multi-turn dialog, visual multi-turn dialog, multi-turn image editing, dense labeling, spatial-conditional image synthesis, and multiview generation. To intuitively illustrate these capabilities, we qualitatively visualize how various types of downstream tasks can be seamlessly integrated into Lumina-mGPT from Figure 14 to Figure 18 in the Appendix.

First, Lumina-mGPT effectively handles general text-only tasks in LLMs, such as solving math problems, coding, and commonsense reasoning, thanks to the extensive pertaining in Chameleon and our multi-task finetuning to mitigate catastrophe forgetting. As shown in Figure 14, Lumina-mGPT correctly answers "`Which is bigger? 9.9 or 9.11`", a question that has confused almost all existing LLMs including GPT-4 (Achiam et al., 2023) and Gemini (Team et al., 2023).

As depicted in Figure 15, Lumina-mGPT is also capable of handling various vision-language tasks including image captioning, visual question answering, and general multi-turn multi-image dialog. Quantitative results on multimodal understanding benchmarks (Liu et al., 2023b; Yue et al., 2024; Li et al., 2023; 2024a; Fu et al., 2023) presented in Table 3 illustrate that Omni-SFT significantly enhances the visual perception capabilities over the original Chameleon model.

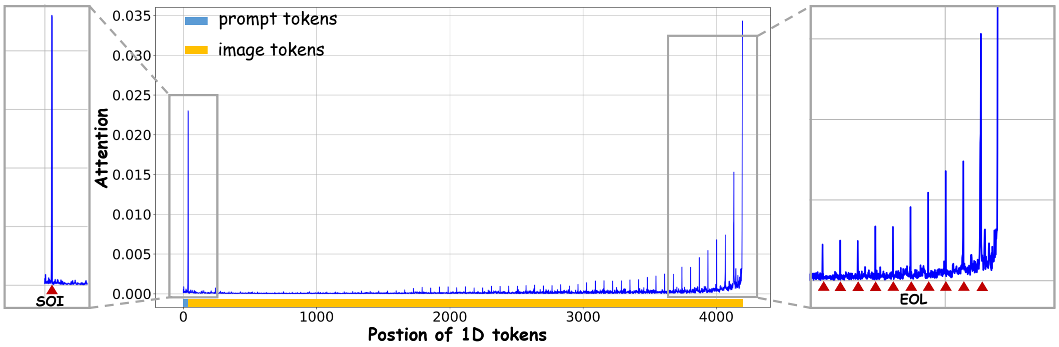


Figure 5: Visualization of averaged attention logits from the last image token. We observe “attention sinks” (Xiao et al., 2024), where added indicator tokens allocate a large proportion of attention score.

As a visual generalist, Lumina-mGPT incorporates classic visual recognition tasks. Using natural language as a unified interface, Lumina-mGPT can perform multiple high-level computer vision tasks including image segmentation, pose estimation, depth estimation, surface normal estimation, and referring object detection. See Figure 16 and 17 for examples.

For image generation, Lumina-mGPT supports both text and versatile spatial conditions, such as depth maps, segmentation maps, normal maps, and human poses, to guide the generation of target images, demonstrated in Figure 17 and 18. Beyond single-image generation, Lumina-mGPT can produce a sequence of consistent images from multiview, as shown in Figure 18. This task marks a preliminary step toward video generation by rendering the multiview images into a coherent video.

Given the above examples, though preliminary, they showcase that Lumina-mGPT can effectively follow diverse instructions, highlighting its promising potential as a unification of various challenging tasks in one framework.

### 3.3 ATTENTION VISUALIZATION

To better understand the sampling behavior of Lumina-mGPT, we visualize the average attention logits of the last image token during text-to-image generation, as shown in Figure 5. The results reveal that the attention score decreases for distant tokens, indicating that the model pays more attention to local tokens compared to distant ones. This behavior aligns with the long-term decay property designed in RoPE. Besides, we observe a similar pattern known as “attention sinks” in LLMs (Xiao et al., 2024), where a large proportion of attention score is allocated to a small number of tokens. After looking into details, these sink tokens are identified to be the indicators, including the `<start-of-image>` and `<end-of-line>` tokens. Notably, the attention scores of text tokens are significantly smaller than those of the `<start-of-image>` token. This suggests that most of the semantic information in text tokens may have been encapsulated into the `<start-of-image>` token. These intriguing findings indicate that Lumina-mGPT aggregates more information from these sink tokens, demonstrating the effectiveness of these indicators in our proposed Uni-Rep and opening up the potential for accelerating the sampling process with these sink tokens.

## 4 CONCOLUSION

In this work, we introduce Lumina-mGPT, a decoder-only transformer that can produce diverse, photorealistic images at any resolution from text prompts. Instead of random initialization, Lumina-mGPT features initializing from an autoregressive transformer with multimodal Generative PreTraining (mGPT). Leveraging the general multimodal representation learned from massive interleaved data, we design two efficient finetuning strategies named FP-SFT and Omni-SFT to unleash the potential of mGPT on text-to-image generation and omnipotent task unification, respectively. We demonstrate Lumina-mGPT’s broad multimodal capabilities across a wide range of tasks, showcasing its potential as a general vision-language assistant.

# REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 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*, 2023.

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

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL `https://openai.com/research/video-generation-models-as-world-simulators`.

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. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, pp. 611–625. Springer, 2012.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020.

Huiwen Chang, Han Zhang, Jarred Barber, Aaron Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Patrick Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. In *International Conference on Machine Learning*, pp. 4055–4075. PMLR, 2023.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-\sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *arXiv preprint arXiv:2403.04692*, 2024a.

Junsong Chen, Jincheng YU, Chongjian GE, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-$\alpha$: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024b. URL `https://openreview.net/forum?id=eAKmQPe3m1`.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017.

Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in neural information processing systems*, 34:19822–19835, 2021.

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. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=y01KGvd9Bw`.

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

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

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Peng Gao, Le Zhuo, Ziyi Lin, Chris Liu, Junsong Chen, Ruoyi Du, Enze Xie, Xu Luo, Longtian Qiu, Yuhang Zhang, et al. Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. *arXiv preprint arXiv:2405.05945*, 2024.

Yuying Ge, Sijie Zhao, Chen Li, Yixiao Ge, and Ying Shan. Seed-data-edit technical report: A hybrid dataset for instructional image editing. *arXiv preprint arXiv:2405.04007*, 2024.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems*, volume 36, pp. 52132–52152, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 630–645. Springer, 2016.

Alex Henry, Prudhvi Raj Dachapally, Shubham Pawar, and Yuxuan Chen. Query-key normalization for transformers. *arXiv preprint arXiv:2010.04245*, 2020.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=rygGQyrFvH`.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment, 2024.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *arXiv preprint arXiv:2307.06350*, 2023.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *International Conference on Machine Learning*, 2024.

Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2989–2998, 2023.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 787–798, 2014.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9492–9502, 2024.

Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13299–13308, 2024a.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024b.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024c.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023a.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023a.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. In *The Eleventh International Conference on Learning Representations*, 2022.

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

Zeyu Lu, ZiDong Wang, Di Huang, Chengyue Wu, Xihui Liu, Wanli Ouyang, and LEI BAI. Fit: Flexible vision transformer for diffusion model. In *Forty-first International Conference on Machine Learning*, 2024b. URL `https://openreview.net/forum?id=jZVen2JguY`.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. *arXiv preprint arXiv:2309.15505*, 2023.

AI Meta. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*, 2024.

Minh Nguyen, Andrew Baker, Andreas Kirsch, and Clement Neo. Min p sampling: Balancing creativity and coherence at high temperature. *arXiv preprint arXiv:2407.01082*, 2024.

Pablo Pernias, Dominic Rampas, Mats Leon Richter, Christopher Pal, and Marc Aubreville. Würstchen: An efficient architecture for large-scale text-to-image diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=gU58d5QeGv`.

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, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

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*, pp. 8748–8763. PMLR, 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(140):1–67, 2020.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pp. 746–760. Springer, 2012.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. In *The Twelfth International Conference on Learning Representations*, 2023.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024a.

Gemini Team, Rohan 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.

Kolors Team. Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis. *arXiv preprint*, 2024b.

Teknium. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants, 2023. URL `https://huggingface.co/datasets/teknium/OpenHermes-2.5`.

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.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning*, pp. 23318–23340. PMLR, 2022.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=NG7sS51zVF`.

Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2955–2966, 2023.

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

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*, 2(3), 2023.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36, 2024.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023.

Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Lirui Zhao, Fu-Yun Wang, Zhanyu Ma, et al. Lumina-next: Making lumina-t2x stronger and faster with next-dit. *arXiv preprint arXiv:2406.18583*, 2024.

# A RELATED WORK

**Multimodal Large Language Models** Recent large language models (LLMs) (Chowdhery et al., 2022; Touvron et al., 2023; Brown et al., 2020; Achiam et al., 2023; Team et al., 2023) demonstrate strong instruction-following and reasoning capabilities, coupled with extensive world knowledge. To extend LLMs' expertise from the text domain to multimodal domains such as images and videos, prior works (Liu et al., 2023a; Lin et al., 2023a; Maaz et al., 2023; Lin et al., 2023b) have aligned pretrained encoders from various modalities with LLMs by curating multimodal instruction tuning datasets. Although these multimodal large language models (MLLMs) have exhibited powerful visual understanding capabilities, their multimodal functionality is primarily limited to perceiving the visual world, exemplified by tasks such as visual question answering and image captioning, rather than generating visual outputs. Another line of research Koh et al. (2024); Dong et al. (2024); Sun et al. (2023); Wu et al. has proposed augmenting MLLMs with the ability to generate images, videos, and audio from text instructions. These approaches introduce additional visual tokens for generation and align these generative tokens as conditional information with a pretrained generator, such as Stable Diffusion (Esser et al., 2024; Podell et al., 2023) for text-to-image generation. Consequently, the generation capabilities heavily rely on the external expert generator rather than MLLMs themselves, resulting in inconsistent and inferior generation results. To combine the strength of both approaches, our model aims to learn both understanding and generation of images using an MLLM with native multimodal capabilities, drawing inspiration from Chameleon (Team, 2024a), a mixed-modal early-fusion foundation model.

**Text-to-Image Generation** The task of text-to-image generation seeks to synthesize photorealistic and diverse images based on textual descriptions. Nowadays, diffusion models, whether in pixel space (Saharia et al., 2022) or in latent space (Podell et al., 2023; Esser et al., 2024; Pernias et al., 2024), have become the de-facto approaches due to their superior performance, particularly in producing extremely high-aesthetic images. Among these models, the recent trend of scaling diffusion transformers (DiTs) (Chen et al., 2024b;a; Gao et al., 2024; Esser et al., 2024; Li et al., 2024c; Team, 2024b) suggests a unified architecture for both text and image modeling. However, existing DiTs still leverage separate language models, such as CLIP (Radford et al., 2021) or T5 (Raffel et al., 2020), as text encoders. This modality gap between text and image representations not only leads to inaccurate generation but also hinders the development of a unified multimodal foundational generative model. Compared to the dominance of diffusion models, the progress of autoregressive image generation has received less attention in recent years. Early works (Ramesh et al., 2021; Ding et al., 2021) proposed a two-stage generation approach: first, training a VQ-VAE Van Den Oord et al. (2017); Esser et al. (2021) for image tokenization and de-tokenization, and then using an autoregressive transformer to model discrete image token sequences, akin to language modeling. Parti (Yu et al., 2022) scaled up the autoregressive transformer to 20 billion parameters, demonstrating promising high-fidelity image generation results. LlamaGen (Sun et al., 2024) further improved the image tokenizer and integrated advanced techniques in LLMs, bridging the performance gap with diffusion counterparts. Unlike Parti and LlamaGen, Lumina-mGPT proposes multimodal generative pertaining on unified text-image sequences, followed by supervised finetuning on high-quality text-to-image pairs, achieving flexible high-aesthetic image generation with autoregressive models.

# B DISCUSSIONS ON RECONSTRUCTION QUALITY

VQ-VAEs (Van Den Oord et al., 2017; Razavi et al., 2019; Esser et al., 2021) compress images at the cost of information loss, which introduces quality degradation at reconstruction, especially for high-frequency details such as edges, hair, and text. As Generative models such as Lumina-mGPT only has access to the VQ-VAE latents during training and cannot access original images, intuitively the VQ-VAE reconstruction quality should somehow build an upper bound for such models' image generation quality.

However, we observe an interesting and counter-intuitive phenomenon. Given two data flows:

1. Image $\xrightarrow{\text{VQVAE Encoder}}$ latent $\xrightarrow{\text{VQVAE Decoder}}$ Recon1

2. Image $\xrightarrow{\text{VQVAE Encoder}}$ latent $\xrightarrow{\text{Lumina-mGPT using editing system prompt with instruction "no edit"}}$ latent2 $\xrightarrow{\text{VQVAE Decoder}}$ Recon2

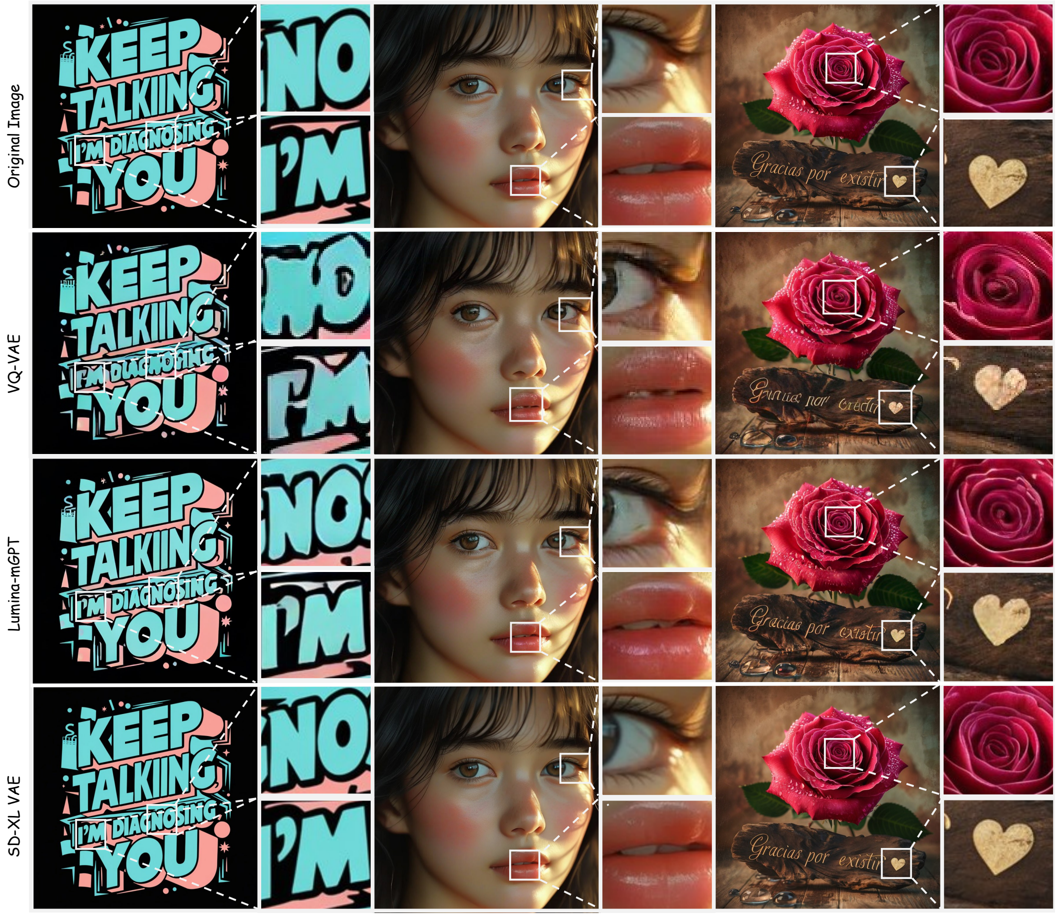


Figure 6: Reconstruction quality of different methods. Lumina-mGPT means first encode using VQ-VAE encoder, then send the latent to Lumina-mGPT using image editing system prompt with the instruction "no edit", and finally decode the newly generated latents using VQ-VAE decoder.

We surprisingly find that the quality of Recon2 sometimes surpasses that of Recon1, and we show such cases in Fig. 6. For reference, the reconstruction results using SDXL VAE (Podell et al., 2023) are also presented. Note that while Lumina-mGPT has been trained on the image editing task, it is not trained with the "no edit" instruction. This intriguing observation may suggest some meaningful insights. For example, it may possibly indicate that the latents encoded by VQ-VAE encoder could contain noises of certain patterns that can be learned and even corrected. It is worth noting that we believe this phenomenon should mainly be attributed to the intrinsic properties of VQ-VAE's discrete representation. We hypothesize that similar effects might also be observed in various generative models (e.g., autoregressive models, discrete diffusion models, etc.) and not due to any specific advantage or uniqueness of the decoder-only autoregressive architecture or other components of Lumina-mGPT, especially those claimed as contributions. We leave the further exploration of this phenomenon for future work.

## C LIMITATIONS OF EXISTING APPROACHES

**Randomly-Initialized Transformer** While transfer learning has revolutionized key fields such as visual recognition (He et al., 2016; Radford et al., 2021; Lu et al., 2019) and language generation (Raffel et al., 2020; Radford et al., 2018; 2019; Brown et al., 2020), popular autoregressive image generation approaches such as DALL-E, Parti, and LlamaGen all adopt a randomly-initialized causal transformer, which fails to utilize pretrained transferable representation and large-scale datasets. As a result, AR-based approaches often lead to poor image generation quality and slow convergence without leveraging proper large-scale pretraining.

**Verbose Encoder-Decoder Architeture** DALL-E and CogView initially propose using a decoder-only transformer for image generation with discrete representation, where a single transformer acts as both a text encoder and an image token decoder. However, subsequent approaches, such as Parti and LlamaGen, adopt a verbose

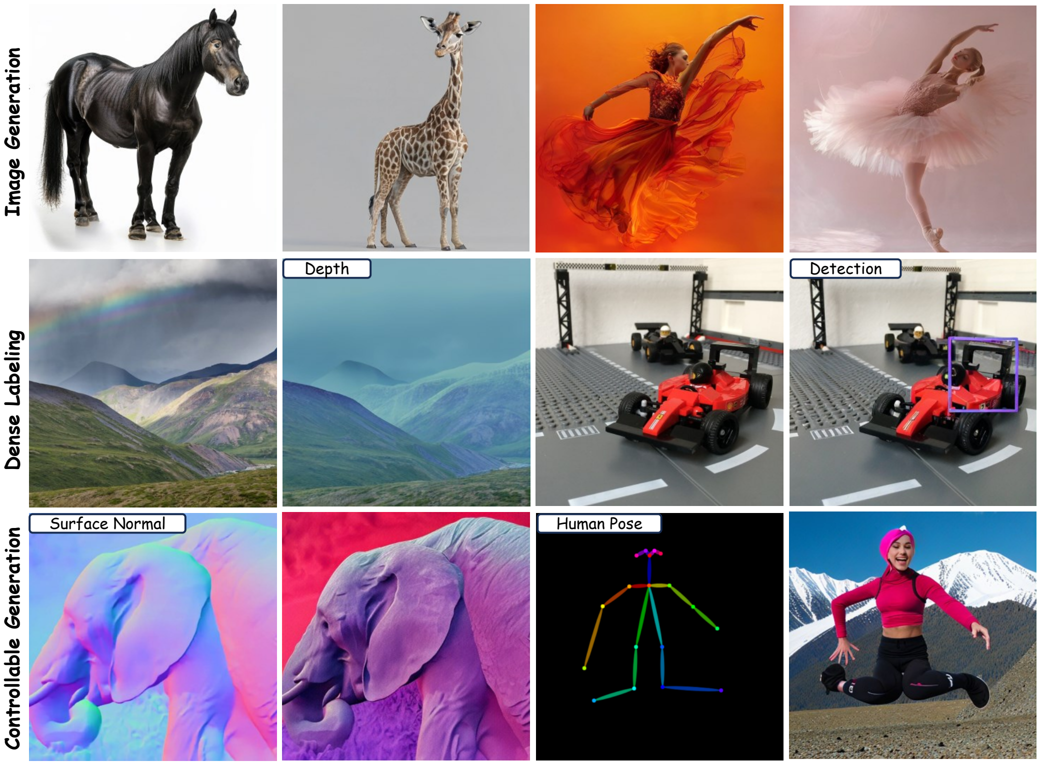


Figure 7: Failure cases of current Lumina-mGPT. Due to inadequate training and limited data size, Lumina-mGPT sometimes struggles to understand input conditions and produce visual artifacts.

encoder-decoder architecture that injects frozen T5 text features (Raffel et al., 2020) using cross-attention or prefix-filling approaches, motivated by the findings of Imagen (Saharia et al., 2022). Compared to the trend in scaling LLMs (Touvron et al., 2023; Bai et al., 2023; Bi et al., 2024), such encoder-decoder architecture is cumbersome due to the decoupling of text encoding and image token modeling. This design significantly complicates the autoregressive-decoding framework, limits the scalability of image generation, and hinders the generalization to additional modalities and tasks.

**Limited Decoding Resolution and Flexibility** Natural images exsit in various resolutions and aspect ratios. Advanced diffusion models (Chen et al., 2024a;b; Esser et al., 2024; Team, 2024b; Li et al., 2024c; Zhuo et al., 2024; Lu et al., 2024b) can successfully generate diverse photorealistic images at arbitrary resolution with skewed ratios. In contrast, current AR-based approaches (Yu et al., 2022; Team, 2024a; Sun et al., 2024) rely on central-cropping a low-resolution 512 × 512 image and transforming the cropped low-resolution image into a fixed-length sequence of discrete tokens using a pretrained Vector-Quantized Variational Autoencoder (VQ-VAE) (Van Den Oord et al., 2017; Razavi et al., 2019; Esser et al., 2021). This approach simplifies autoregressive training but at the cost of deteriorated image quality and generation flexibility.

**Poor Task Extensability** Autoregressive modeling excels at unified generative modeling for various tasks and modalities using discrete tokens (Wang et al., 2022; Lu et al., 2022; 2024a). However, previous AR-based image generation approaches (Ramesh et al., 2021; Ding et al., 2021; Yu et al., 2022; Sun et al., 2024) have been limited to text-to-image generation without exploring the unification with other tasks, such as dense labeling and controllable image generation. This lack of task extensibility significantly constrains the applicability of AR-based image generation models to a broader range of scenarios.

## D LIMITATIONS OF LUMINA-MGPT

**Failure Cases** Despite Lumina-mGPT can generate photorealistic images, it sometimes produces images with noticeable visual artifacts. For example, Lumina-mGPT may generate people and animals with unreasonable limbs, as shown in the first row of Figure 7. Besides, compared to SoTA text-to-image generation approaches including SD3 (Esser et al., 2024), Kolors (Team, 2024b), and HunyuanDiT (Li et al., 2024c), all pretrained over

1B image-text pairs, Lumina-mGPT's prompt-following ability is inferior due to the limited training resources and data size, which are many times smaller than these SoTA methods. Regarding dense labeling and controllable generation, Lumina-mGPT currently showcases preliminary results with a limited training budget. Hence, the second row in Figure 7 provides such an example where Lumina-mGPT produces inaccurate predictions or semantically inconsistent images, failing to understand the given image conditions. Therefore, we expect by scaling data size with more computational resources, Lumina-mGPT can effectively address the above failure cases such as inadequate instruction-following ability and visual artifacts.

**Generation Speed** Autoregressive models require numerous network evaluations during inference due to the nature of next-token prediction, similar to the iterative denoising process in diffusion models. This becomes worse when generating high-resolution images, which often require minutes to generate a full sequence of image tokens, significantly slower than current diffusion models with advanced samplers. However, there have been plenty of techniques to optimize the inference speed designed for autoregressive models, such as vLLM (Kwon et al., 2023) and FlashAttention (Dao et al., 2022; Dao, 2024). We believe that by integrating these approaches in the future, Lumina-mGPT can achieve a remarkable speed up during inference.

**VQ-VAE Reconstruction Quality** VQ-VAE is employed as the image tokenizer to convert continuous images into discrete token representations. Meanwhile, it also introduces information bottlenecks by compressing the spatial dimensions of images. As a result, the reconstruction quality of VQ-VAE largely determines the upper limit of generation quality. We discover that the VQ-VAE proposed in Chameleon sometimes struggles to reconstruct high-frequency details, especially when text and human faces are present in images. Incorporating further improvements on VQ-VAE, such as FSQ (Mentzer et al., 2023), may also enhance the generation quality of Lumina-mGPT.

## E INFERENCE CONFIGURATION OF LUMINA-MGPT

In autoregressive models, various configuration parameters during Lumina-mGPT's decoding stage significantly affect sample quality (Holtzman et al., 2020; Radford et al., 2018; 2019). Hyperparameters such as temperature (T), top-k, and classifier-free guidance scale (CFG) have not been extensively investigated in the visual domain. In this section, we explore how these hyperparameters influence the generated image in terms of quality, texture, and style.

**Different Decoding Hyperparameters for Image and Text** The sampling strategy of autoregressive models involves numerous hyperparameters that significantly influence the sampling results. We find that the optimal decoding hyperparameters differ greatly between text decoding and discrete image code decoding. For example, the top-k=5 setting performs well in generating text. However, when generating images, the value of top-k should be much larger (e.g. 2000) to avoid repetitive and meaningless patterns. Therefore, we implement a status-aware control mechanism for inference. Specifically, a set of default hyperparameters is used for text decoding; once a `<start-of-image>` token is generated, the hyperparameters switch to those optimized for image generation. After the `<end-of-image>` token is generated, the parameters revert to the initial settings.

**Classfier-Free Guidance** Classifier-Free Guidance (CFG) (Ho & Salimans, 2022) is originally proposed to enhance the quality and text alignment of generated samples in text-to-image diffusion models. We incorporate this technique into autoregressive models during inference. When generating an image token, the CFG-processed logits $l_{\text{cfg}}$ are formulated as $l_{\text{cfg}} = l + s(l - l')$, where $l$ represents the original logits conditioned on the complete context; $l'$ represents the context-independent logits, which are conditioned solely on the tokens following the `<start-of-image>` token of the currently generating image, and are independent of any prior context; $s$ denotes the guidance scale of Classifier-Free Guidance. To make CFG work, during training, the context before <start-of-image> is randomly dropped by a probability of 10%. In practice, KV cache can be used for accelerating the computation of both $l$ and $l'$. As shown in Fig. 8, similar to the trend of diffusion models, increasing CFG initially raises the quality and stability of generation, but increasing it further would make the quality deteriorate.

**Influence of Temperature** To evaluate the effect of these decoding parameters, we first set a standard decoding configuration: T=1.0, top-k=2000, CFG=4.0, which serves as a good-to-use setting. From this baseline, we gradually shift T from 0.7 to 1.0 to generate corresponding images at different temperatures. As shown in Figure 9, it is evident that when setting the temperature low, visual details diminish and objects tend to be over-smoothed. Conversely, when setting the temperature high, the generated images contain rich visual content but are prone to contain more artifacts.

**Influence of Top-k** Based on the standard decoding setting, we vary the top-k value, from 50 to 8192, where 8192 is equal to the size of the VQ-VAE codebook used usedin Lumina-mGPT. The results, visualized in Figure 9, indicate a similar trend with increasing temperature. When top-k is low, the image content and texture

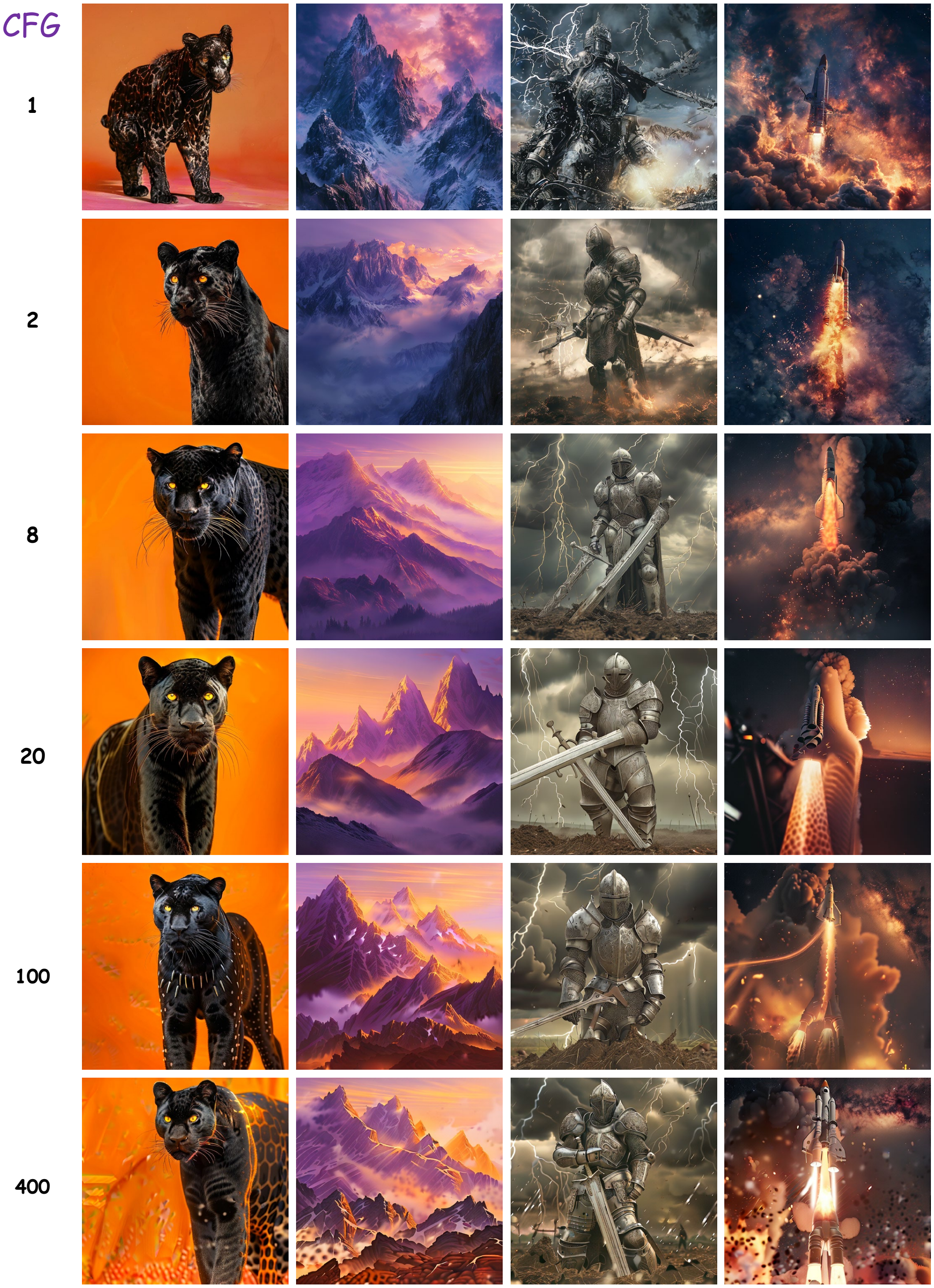


Figure 8: Samples generated by Lumina-mGPT using different CFG; T=1.0, Top-k=2000.

are relatively simple, exhibiting the over-smoothed problem as well. When top-k is set high, the image detail and texture are diverse, making it more aesthetically appealing, while increasing the potential of artifacts.

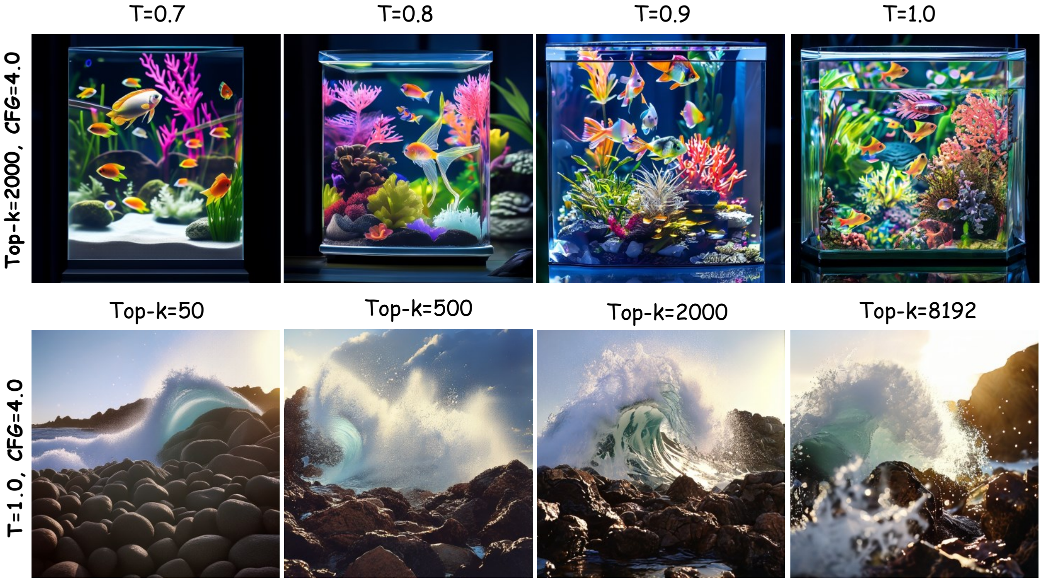


Figure 9: Samples generated by Lumina-mGPT using different Top-k and Temperature.

## F COMPARISON WITH DIFFUSION-BASED APPRAOCHES

For a long period of time, diffusion models have dominated the field of text-to-image generation compared to autoregressive models. Although LlamaGen claims to beat diffusion models, their results are limited to the ImageNet benchmark and there has been no direct comparison between these two architectures so far. In this section, we aim to provide a detailed comparison of autoregressive and diffusion-based methods trained on these same text-image datasets, focusing on image quality, diversity, text-rendering, and multilingual capabilities. Specifically, we adopt Lumina-mGPT and Lumina-Next-SFT (Zhuo et al., 2024) as representatives of autoregressive and diffusion-based methods, respectively. A direct visual comparison between Lumina-Next-SFT and Lumina-mGPT reveals both the similarities and differences between autoregressive and diffusion-based generative modeling approaches.

**On the Similarity between Diffusion- and AR-based Generation** Given the same set of text prompts, both diffusion- and AR-based approaches generate photorealistic images with similar aesthetic style and fine-grained details, illustrated in Figure 10. This reveals the fact that both architectures can achieve satisfactory text-to-image generation performance when provided with the same training data, training budget, and comparable model sizes. The AR-based methods display remarkable visual aesthetics on par with their diffusion counterparts, challenging the notion that diffusion models are more effective and promising architecture in generative modeling. This finding also aligns with the platonic representation hypothesis (Huh et al., 2024) that neural networks are converging to learn a shared representation space despite being trained with different architectures and objectives. Therefore, this hypothesis highlights the importance of collecting more high-quality data and optimizing training infrastructure as directions for data and model scaling, to boost the overall model performance that is agnostic to any specific architecture.

**On the Differences between Diffusion- and AR-based Generation** As shown in Figure 10, Lumina-mGPT exhibits more diversity using different random seeds, while Lumina-Next-SFT generates similar images with identical layouts and textures. This can be partly attributed to the use of high temperature and top-k values in Lumina-mGPT. However, excessive diversity also causes our model to be less stable and more prone to producing visual artifacts, which is discussed in Section D.

We also compare the text rendering and multilingual understanding capabilities between Lumina-mGPT and Lumina-Next-SFT. As illustrated in Figure 11, Lumina-mGPT exhibits significantly better text synthesizing results, while Lumina-Next-SFT struggles to generate any complete character. We argue that this underscores the importance of mGPT, where the model learns a seamless multimodal representation between text and images using massive interleaved data during the pertaining stage. However, when it comes to multilingual understanding, Lumina-mGPT performs worse than Lumina-Next-SFT in terms of emoji and Chinese prompts shown in Figure 12. The reason is that although Lumina-mGPT learns better text-image alignment, the lack of multilingual text corpus used in pertaining limits its performance. In contrast, the text encoder used in Lumina-Next-SFT showcases significantly stronger multilingual capabilities than Chameleon. Hence, we hope

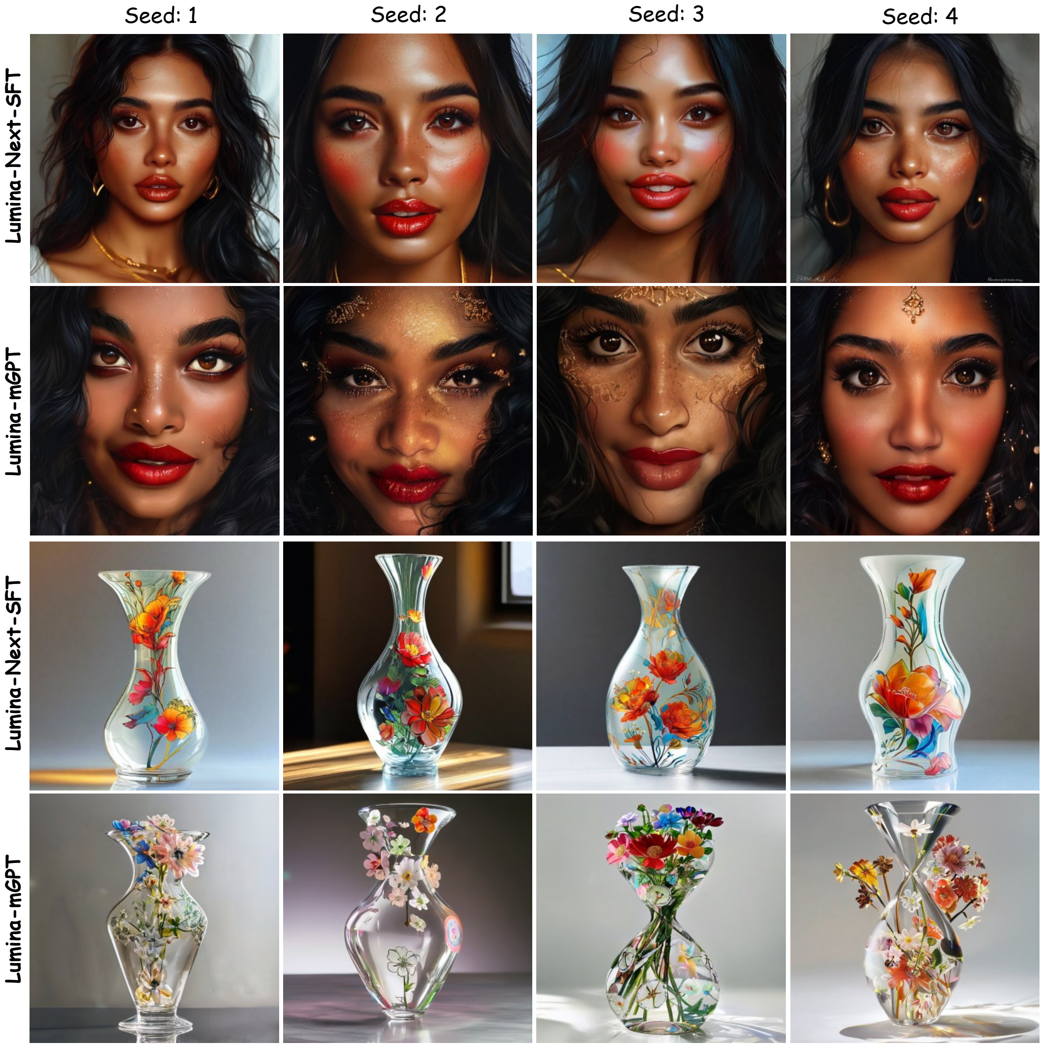


Figure 10: Comparison between Lumina-Next-SFT and Lumina-mGPT using different random seeds. Images generated by Lumina-mGPT exhibit comparable aesthetics with greater diversity.

that by comprehensively enhancing the capabilities of the base mGPT model, such as adding more multilingual data, Lumina-mGPT can benefit in all downstream tasks.

In addition to text-to-image generation, Lumina-mGPT supports various vision and language tasks within a unified framework. However, the design of diffusion models limits their compatibility and performance across multiple modalities and tasks. They often require specific architecture designs and additional training for each unseen task (Ke et al., 2024; Xu et al., 2023). In contrast, Lumina-mGPT treats input from all modalities as multimodal token sequences and leverages natural language as the interface to unify diverse tasks with next-token prediction.

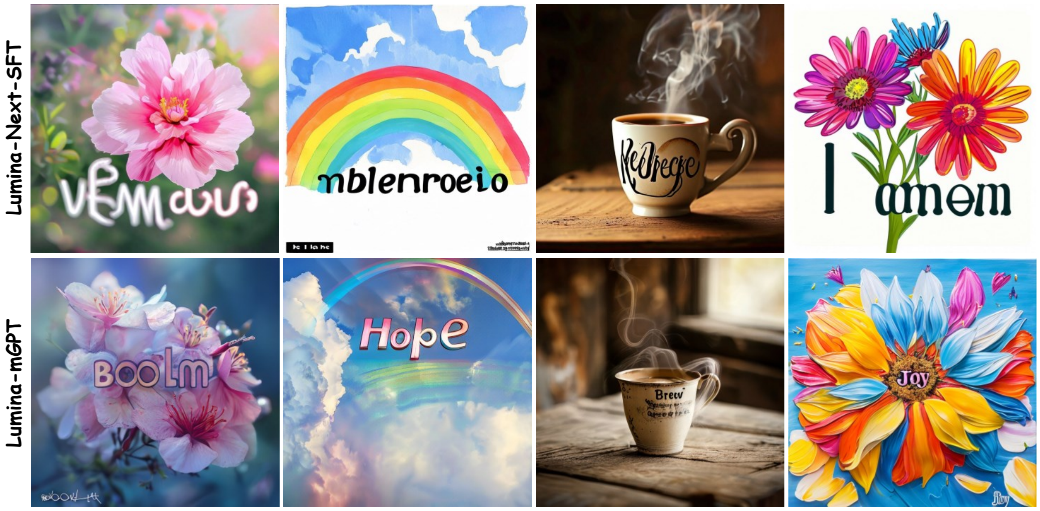
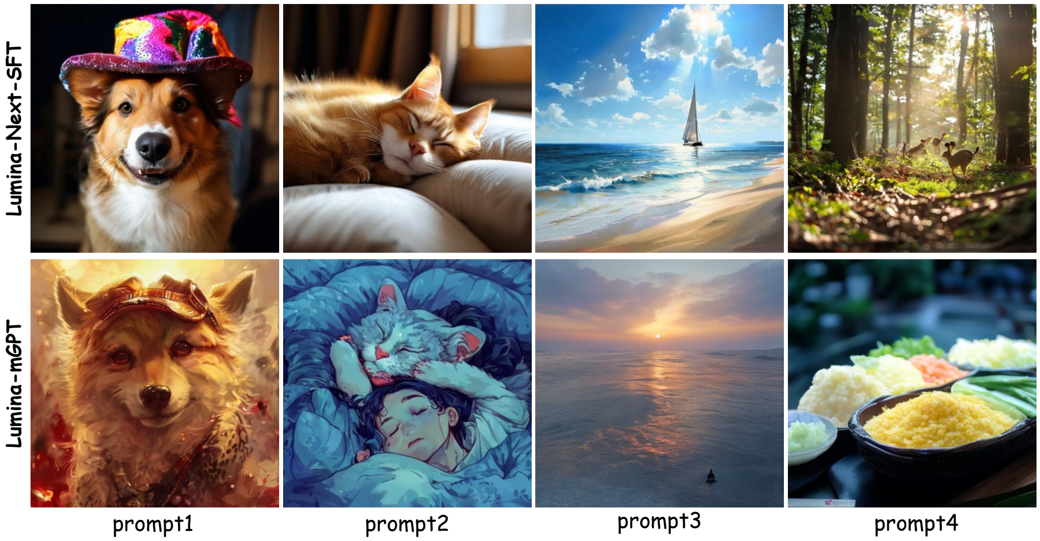


Figure 11: Text rendering comparison between Lumina-Next-SFT and Lumina-mGPT. From left to right, the correct texts to be rendered on the image are: "Bloom", "Hope", "Brew", and "Joy".



prompt1: 🐶 is wearing 🎩.
Prompt2: 🐱 is sleeping on 🛋.
prompt3: 日落时分的海滩，阳光透过云层在海面上撒下一片金色的光芒，海浪轻轻拍打着沙滩，远处的帆船在余晖中若隐若现。
prompt4: 森林中的清晨，阳光穿过茂密的树冠，形成斑驳的光影，地上覆盖着厚厚的落叶，几只小鹿在林间悠闲地觅食。

Figure 12: Emoji and multilingual instruction understanding ability comparison between Lumina-Next-SFT and Lumina-mGPT. Lumina-mGPT struggle to understand emojis and multilingual prompts.

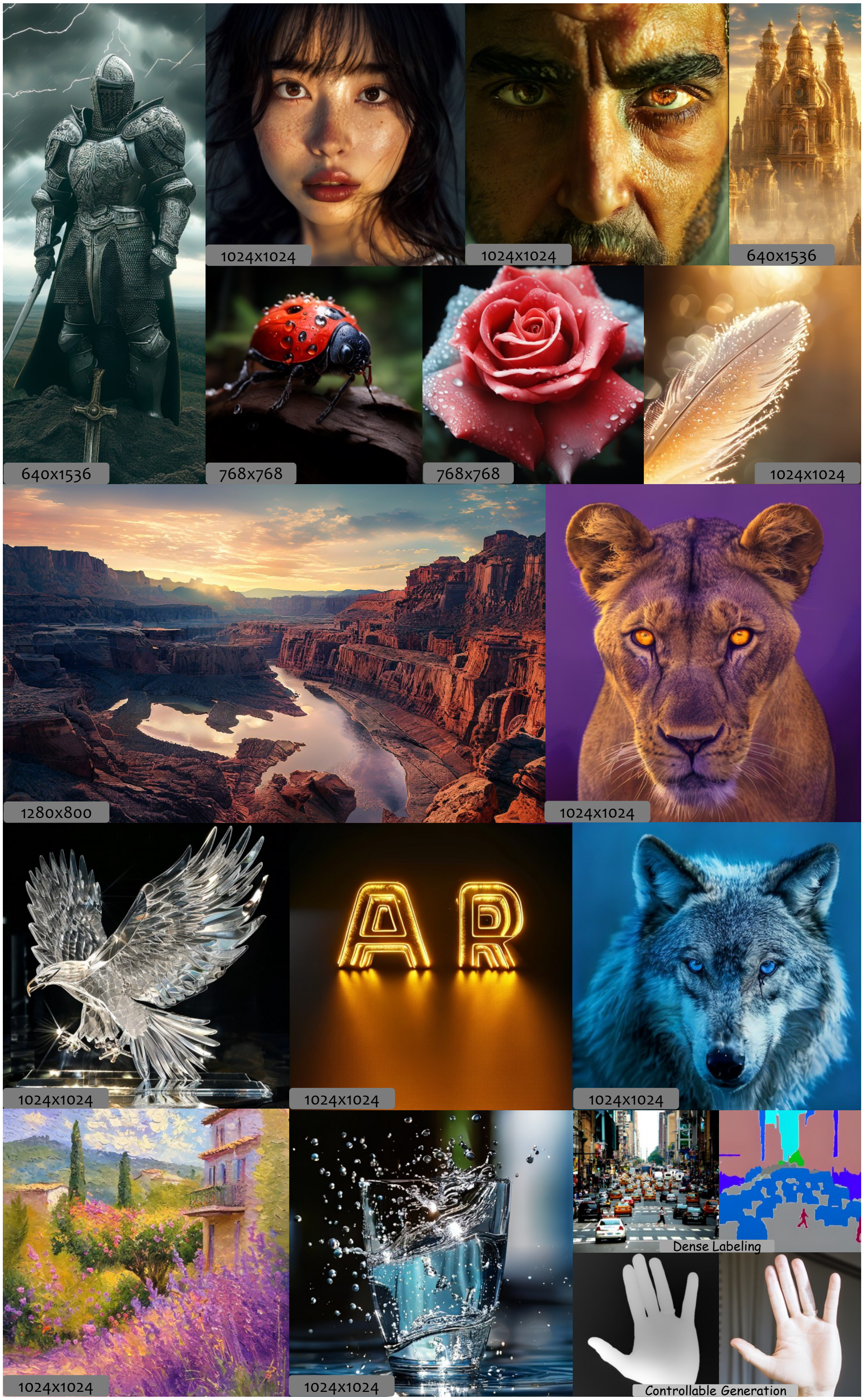
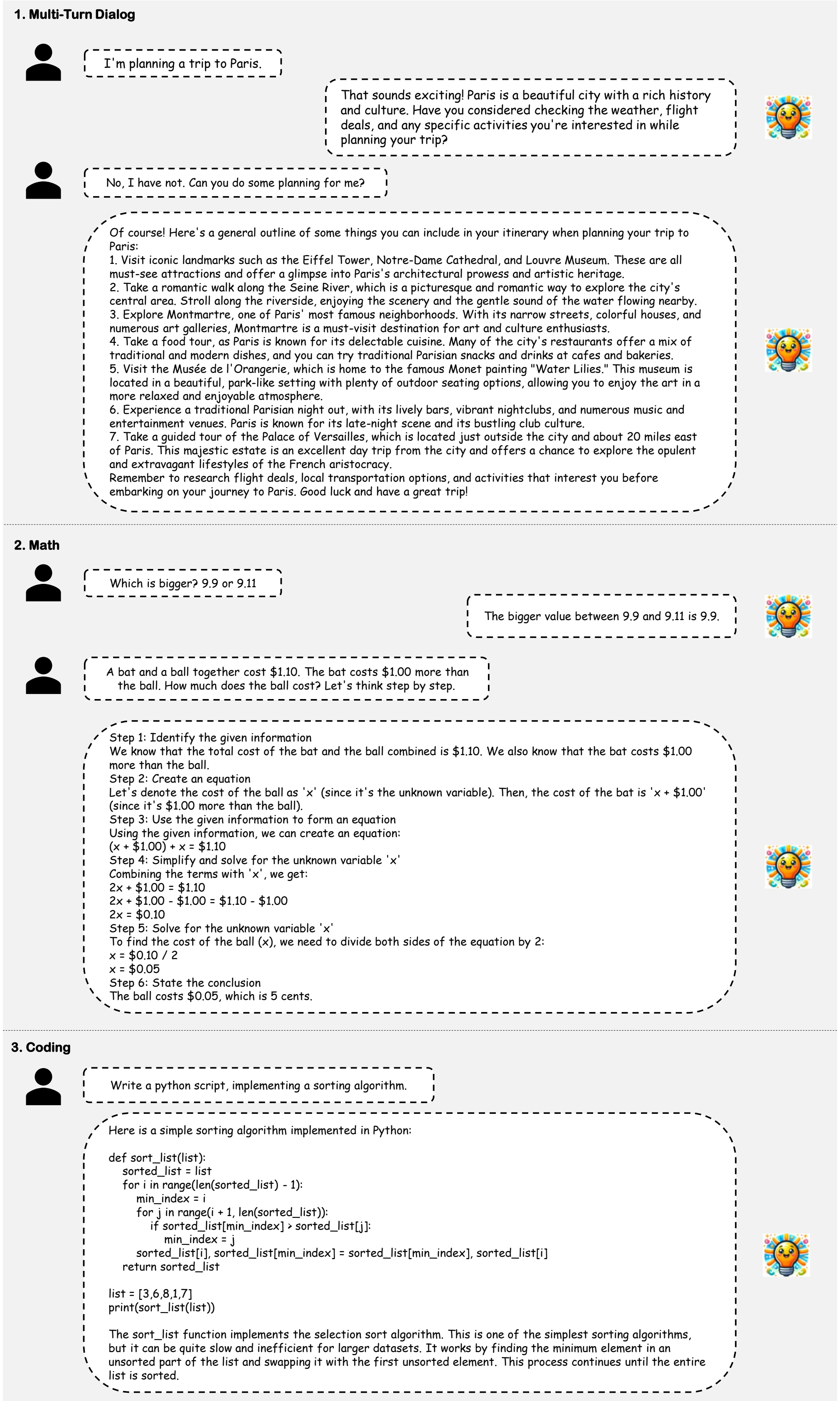


Figure 13: Selected photorealistic images of various resolutions generated by Lumina-mGPT. At the bottom, we include two examples to demonstrate the omnipotent task unification in Lumina-mGPT.



**1. Multi-Turn Dialog**

I'm planning a trip to Paris.

That sounds exciting! Paris is a beautiful city with a rich history and culture. Have you considered checking the weather, flight deals, and any specific activities you're interested in while planning your trip?

No, I have not. Can you do some planning for me?

Of course! Here's a general outline of some things you can include in your itinerary when planning your trip to Paris:
1. Visit iconic landmarks such as the Eiffel Tower, Notre-Dame Cathedral, and Louvre Museum. These are all must-see attractions and offer a glimpse into Paris's architectural prowess and artistic heritage.
2. Take a romantic walk along the Seine River, which is a picturesque and romantic way to explore the city's central area. Stroll along the riverside, enjoying the scenery and the gentle sound of the water flowing nearby.
3. Explore Montmartre, one of Paris' most famous neighborhoods. With its narrow streets, colorful houses, and numerous art galleries, Montmartre is a must-visit destination for art and culture enthusiasts.
4. Take a food tour, as Paris is known for its delectable cuisine. Many of the city's restaurants offer a mix of traditional and modern dishes, and you can try traditional Parisian snacks and drinks at cafes and bakeries.
5. Visit the Musée de l'Orangerie, which is home to the famous Monet painting "Water Lilies." This museum is located in a beautiful, park-like setting with plenty of outdoor seating options, allowing you to enjoy the art in a more relaxed and enjoyable atmosphere.
6. Experience a traditional Parisian night out, with its lively bars, vibrant nightclubs, and numerous music and entertainment venues. Paris is known for its late-night scene and its bustling club culture.
7. Take a guided tour of the Palace of Versailles, which is located just outside the city and about 20 miles east of Paris. This majestic estate is an excellent day trip from the city and offers a chance to explore the opulent and extravagant lifestyles of the French aristocracy.
Remember to research flight deals, local transportation options, and activities that interest you before embarking on your journey to Paris. Good luck and have a great trip!

**2. Math**

Which is bigger? 9.9 or 9.11

The bigger value between 9.9 and 9.11 is 9.9.

A bat and a ball together cost $1.10. The bat costs $1.00 more than the ball. How much does the ball cost? Let's think step by step.

Step 1: Identify the given information
We know that the total cost of the bat and the ball combined is $1.10. We also know that the bat costs $1.00 more than the ball.
Step 2: Create an equation
Let's denote the cost of the ball as 'x' (since it's the unknown variable). Then, the cost of the bat is 'x + $1.00' (since it's $1.00 more than the ball).
Step 3: Use the given information to form an equation
Using the given information, we can create an equation:
(x + $1.00) + x = $1.10
Step 4: Simplify and solve for the unknown variable 'x'
Combining the terms with 'x', we get:
2x + $1.00 = $1.10
2x + $1.00 - $1.00 = $1.10 - $1.00
2x = $0.10
Step 5: Solve for the unknown variable 'x'
To find the cost of the ball (x), we need to divide both sides of the equation by 2:
x = $0.10 / 2
x = $0.05
Step 6: State the conclusion
The ball costs $0.05, which is 5 cents.

**3. Coding**

Write a python script, implementing a sorting algorithm.

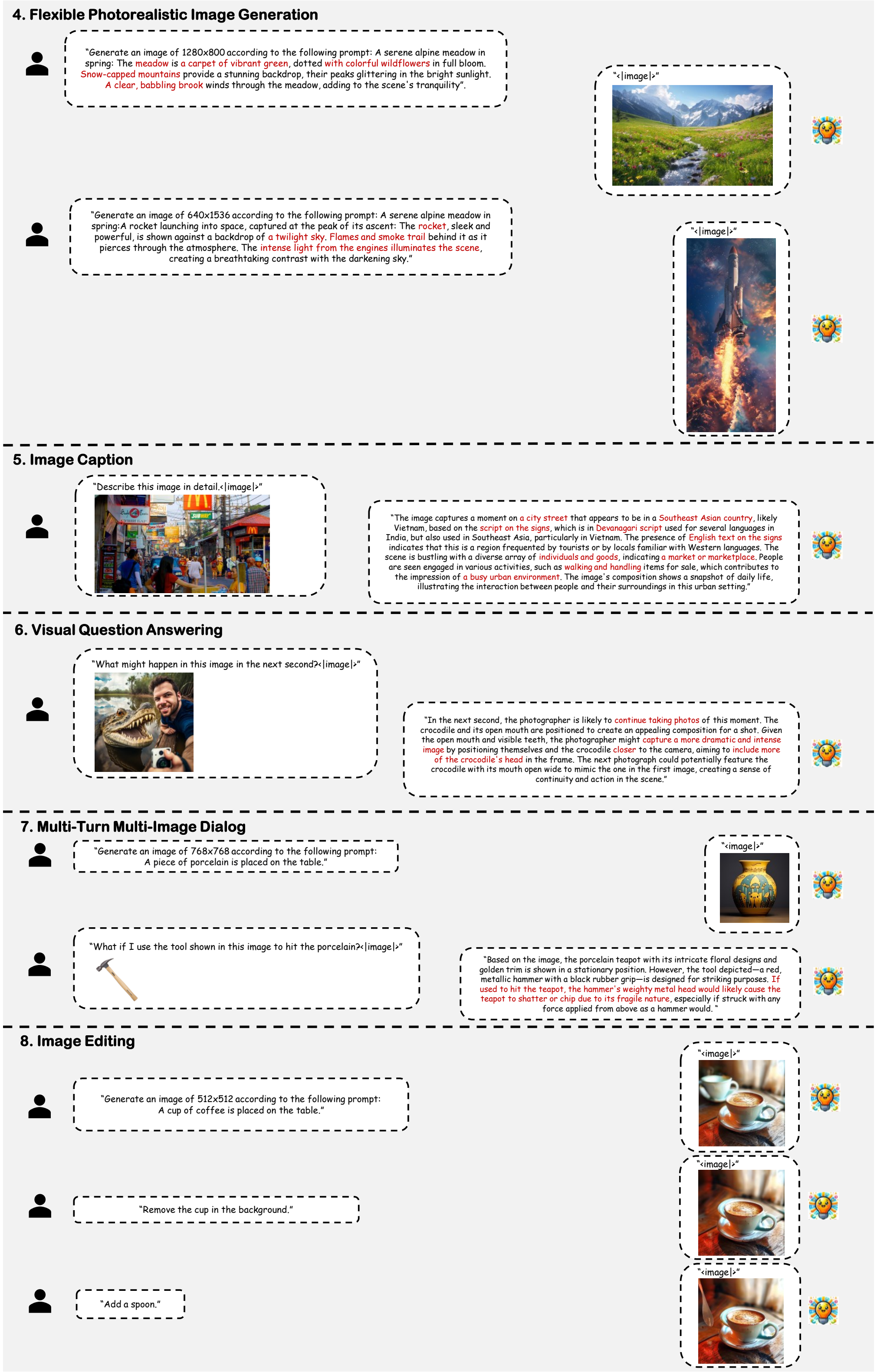
Here is a simple sorting algorithm implemented in Python:

```
def sort_list(list):
    sorted_list = list
    for i in range(len(sorted_list) - 1):
        min_index = i
        for j in range(i + 1, len(sorted_list)):
            if sorted_list[min_index] > sorted_list[j]:
                min_index = j
        sorted_list[i], sorted_list[min_index] = sorted_list[min_index], sorted_list[i]
    return sorted_list

list = [3,6,8,1,7]
print(sort_list(list))
```

The sort_list function implements the selection sort algorithm. This is one of the simplest sorting algorithms, but it can be quite slow and inefficient for larger datasets. It works by finding the minimum element in an unsorted part of the list and swapping it with the first unsorted element. This process continues until the entire list is sorted.

Figure 14: Lumina-mGPT as a generalist for various multimodal tasks.



**4. Flexible Photorealistic Image Generation**

“Generate an image of 1280x800 according to the following prompt: A serene alpine meadow in spring: The meadow is a carpet of vibrant green, dotted with colorful wildflowers in full bloom. Snow-capped mountains provide a stunning backdrop, their peaks glittering in the bright sunlight. A clear, babbling brook winds through the meadow, adding to the scene's tranquility”.

“<|image|>”

“Generate an image of 640x1536 according to the following prompt: A serene alpine meadow in spring:A rocket launching into space, captured at the peak of its ascent: The rocket, sleek and powerful, is shown against a backdrop of a twilight sky. Flames and smoke trail behind it as it pierces through the atmosphere. The intense light from the engines illuminates the scene, creating a breathtaking contrast with the darkening sky.”

“<|image|>”

**5. Image Caption**

“Describe this image in detail.<|image|>”

“The image captures a moment on a city street that appears to be in a Southeast Asian country, likely Vietnam, based on the script on the signs, which is in Devanagari script used for several languages in India, but also used in Southeast Asia, particularly in Vietnam. The presence of English text on the signs indicates that this is a region frequented by tourists or by locals familiar with Western languages. The scene is bustling with a diverse array of individuals and goods, indicating a market or marketplace. People are seen engaged in various activities, such as walking and handling items for sale, which contributes to the impression of a busy urban environment. The image's composition shows a snapshot of daily life, illustrating the interaction between people and their surroundings in this urban setting.”

**6. Visual Question Answering**

“What might happen in this image in the next second?<|image|>”

“In the next second, the photographer is likely to continue taking photos of this moment. The crocodile and its open mouth are positioned to create an appealing composition for a shot. Given the open mouth and visible teeth, the photographer might capture a more dramatic and intense image by positioning themselves and the crocodile closer to the camera, aiming to include more of the crocodile's head in the frame. The next photograph could potentially feature the crocodile with its mouth open wide to mimic the one in the first image, creating a sense of continuity and action in the scene.”

**7. Multi-Turn Multi-Image Dialog**

“Generate an image of 768x768 according to the following prompt: A piece of porcelain is placed on the table.”

“<image|>”

“What if I use the tool shown in this image to hit the porcelain?<|image|>”

“Based on the image, the porcelain teapot with its intricate floral designs and golden trim is shown in a stationary position. However, the tool depicted—a red, metallic hammer with a black rubber grip—is designed for striking purposes. If used to hit the teapot, the hammer's weighty metal head would likely cause the teapot to shatter or chip due to its fragile nature, especially if struck with any force applied from above as a hammer would. “

**8. Image Editing**

“Generate an image of 512x512 according to the following prompt: A cup of coffee is placed on the table.”

“<image|>”

“Remove the cup in the background.”

“<image|>”

“Add a spoon.”

“<image|>”

Figure 15: Lumina-mGPT as a generalist for various multimodal tasks.

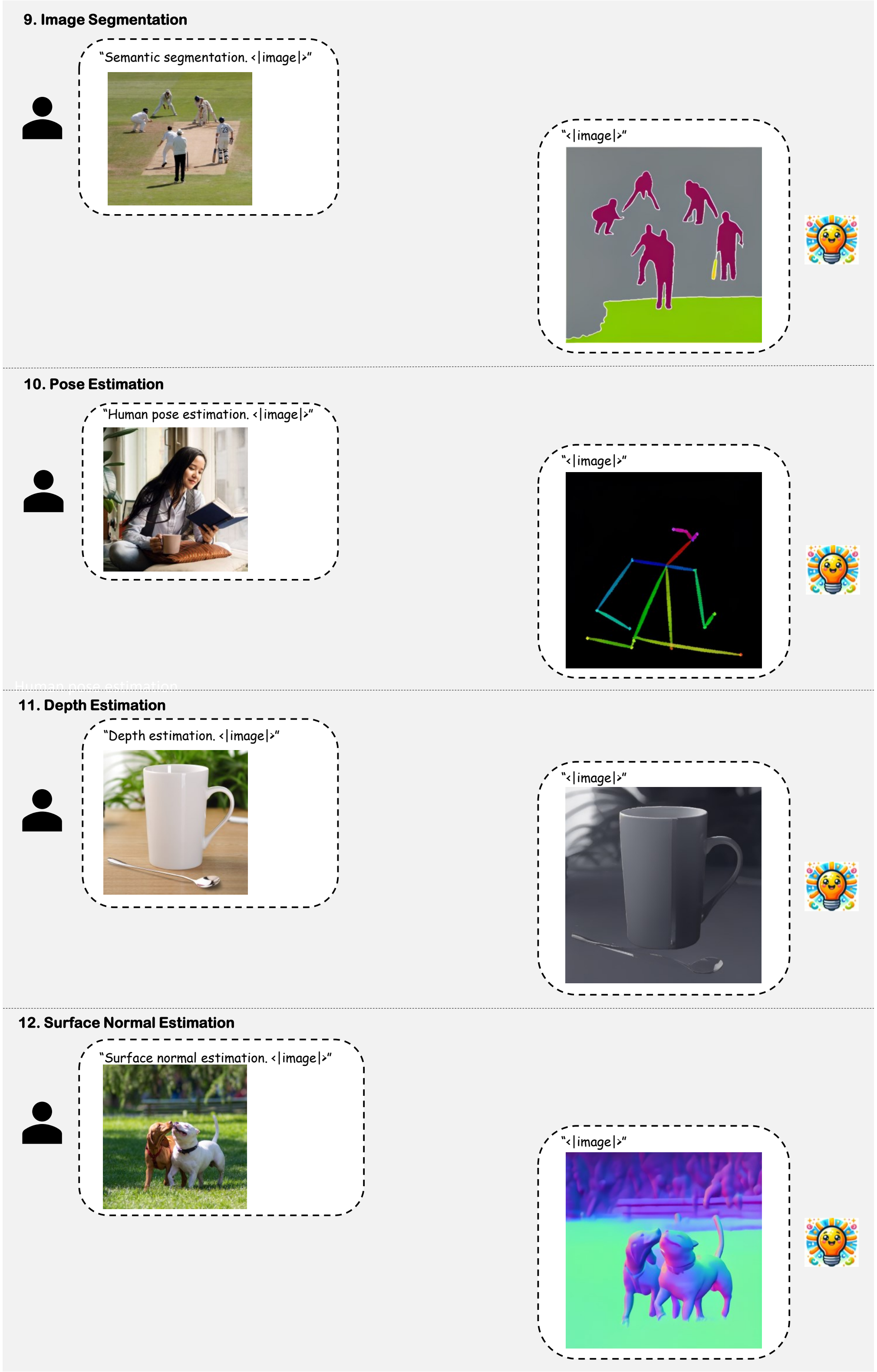
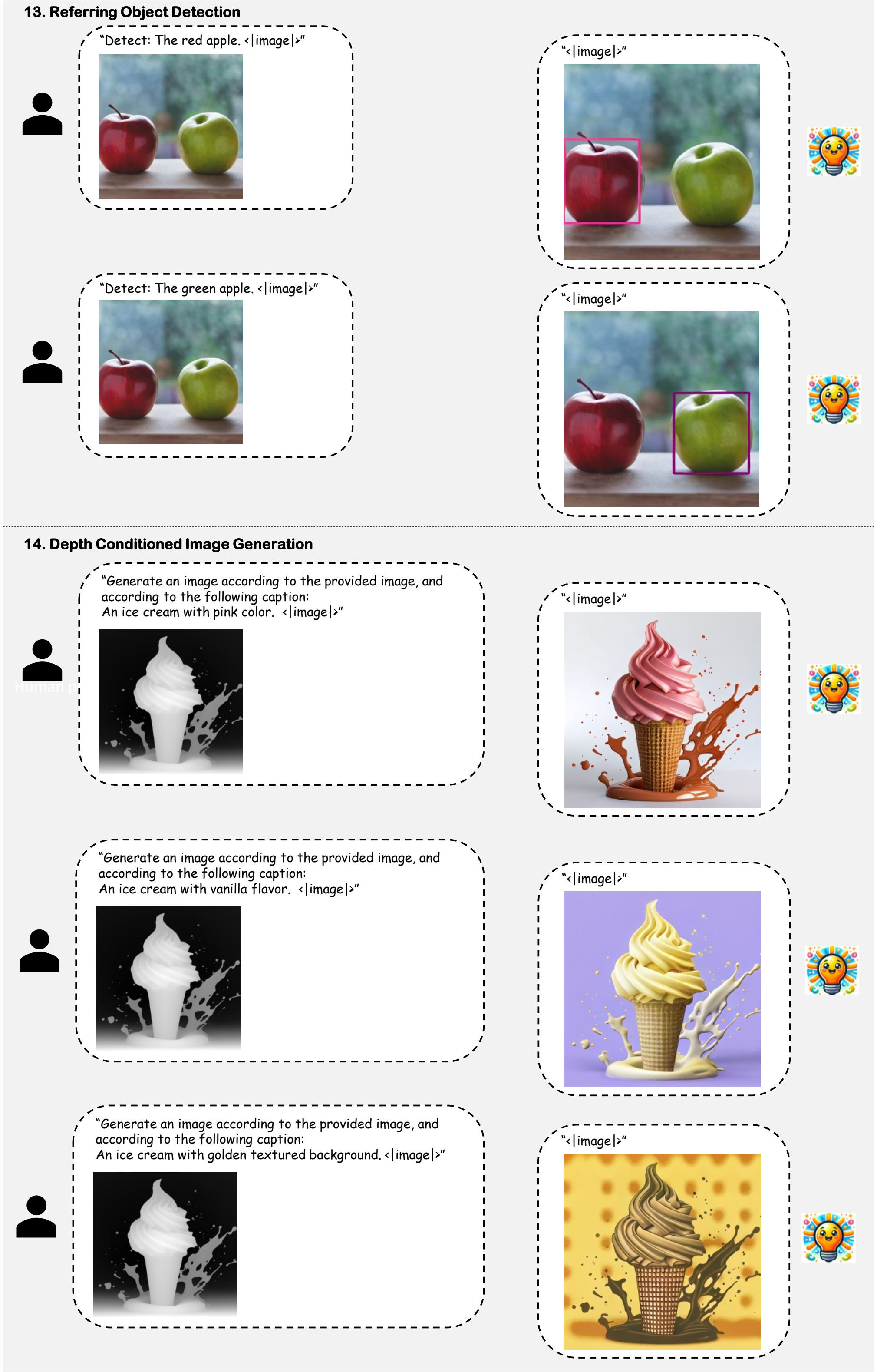

9. Image Segmentation
"Semantic segmentation. <|image|>"
"<|image|>"
10. Pose Estimation
"Human pose estimation. <|image|>"
"<|image|>"
11. Depth Estimation
"Depth estimation. <|image|>"
"<|image|>"
12. Surface Normal Estimation
"Surface normal estimation. <|image|>"
"<|image|>"

Figure 16: Lumina-mGPT as a generalist for various multimodal tasks.



**13. Referring Object Detection**

"Detect: The red apple. <|image|>"

"<|image|>"

"Detect: The green apple. <|image|>"

"<|image|>"

**14. Depth Conditioned Image Generation**

"Generate an image according to the provided image, and according to the following caption:
An ice cream with pink color. <|image|>"

"<|image|>"

"Generate an image according to the provided image, and according to the following caption:
An ice cream with vanilla flavor. <|image|>"

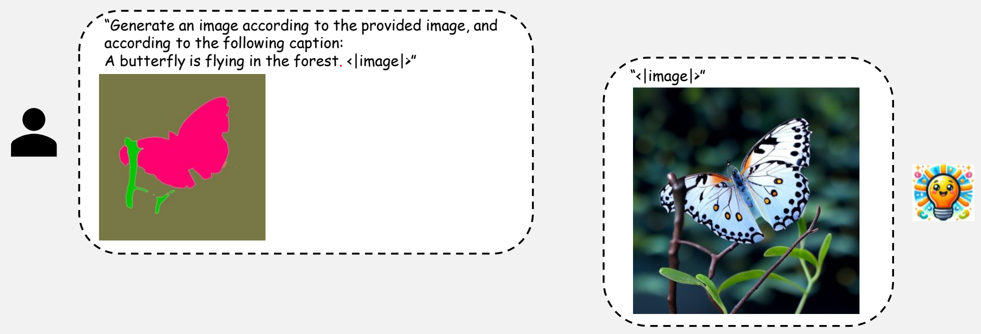
"<|image|>"

"Generate an image according to the provided image, and according to the following caption:
An ice cream with golden textured background. <|image|>"

"<|image|>"

Figure 17: Lumina-mGPT as a generalist for various multimodal tasks.

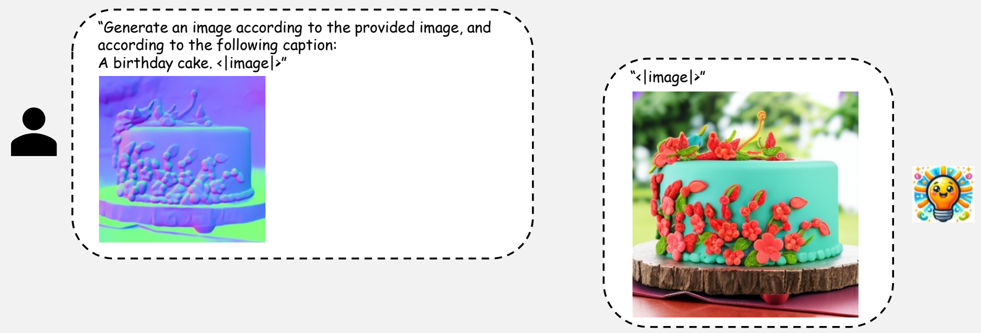
**15. Segmentation Conditioned Image Generation**



"Generate an image according to the provided image, and according to the following caption:
A butterfly is flying in the forest. <|image|>"

"<|image|>"

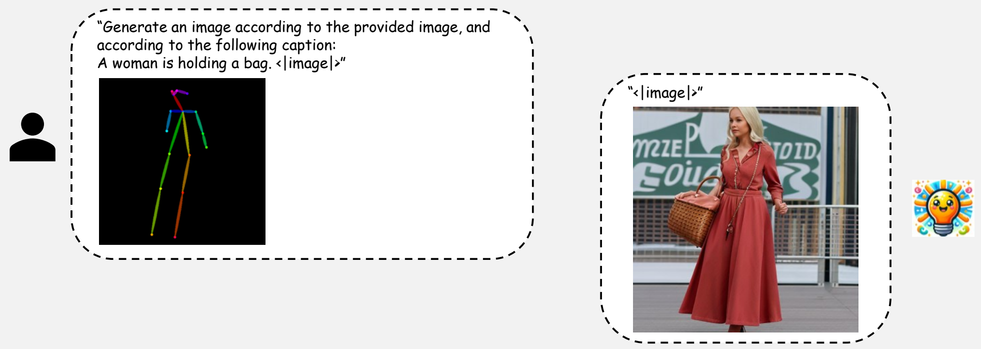
**16. Surface Normal Conditioned Image Generation**



"Generate an image according to the provided image, and according to the following caption:
A birthday cake. <|image|>"

"<|image|>"

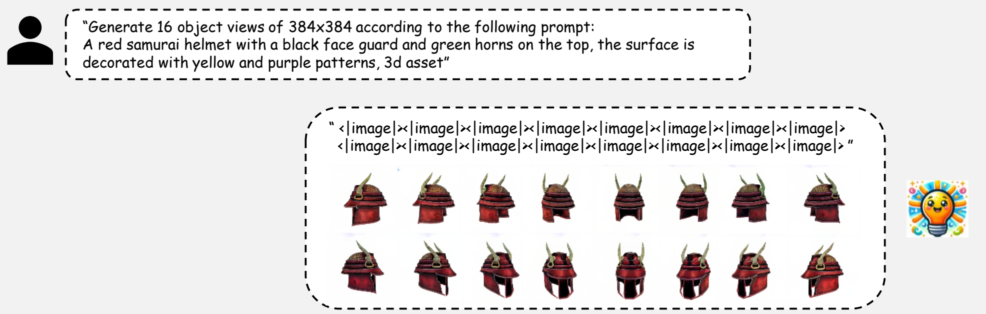
**17. Pose conditioned Image Generation**



"Generate an image according to the provided image, and according to the following caption:
A woman is holding a bag. <|image|>"

"<|image|>"

**18. Multi-view Image Generation**



"Generate 16 object views of 384x384 according to the following prompt:
A red samurai helmet with a black face guard and green horns on the top, the surface is decorated with yellow and purple patterns, 3d asset"

" <|image|><|image|><|image|><|image|><|image|><|image|><|image|><|image|>
<|image|><|image|><|image|><|image|><|image|><|image|><|image|><|image|> "

Figure 18: Lumina-mGPT as a generalist for various multimodal tasks.