

SHOW-O: ONE SINGLE TRANSFORMER TO UNIFY MULTIMODAL UNDERSTANDING AND GENERATION

Jinheng Xie^{1†} Weijia Mao^{1†} Zechen Bai^{1†} David Junhao Zhang^{1†} Weihao Wang²
 Kevin Qinghong Lin¹ Yuchao Gu¹ Zhijie Chen² Zhenheng Yang² Mike Zheng Shou^{1*}

¹ Show Lab, National University of Singapore ² ByteDance

ABSTRACT

We present a unified transformer, *i.e.*, Show-o, that unifies multimodal understanding and generation. Unlike fully autoregressive models, Show-o unifies autoregressive and (discrete) diffusion modeling to adaptively handle inputs and outputs of various and mixed modalities. The unified model flexibly supports a wide range of vision-language tasks including visual question-answering, text-to-image generation, text-guided inpainting/extrapolation, and mixed-modality generation. Across various benchmarks, it demonstrates comparable or superior performance to existing individual models with an equivalent or larger number of parameters tailored for understanding or generation. This significantly highlights its potential as a next-generation foundation model. Code and models are released at <https://github.com/showlab/Show-o>.

1 INTRODUCTION

“Alone we can do so little; together we can do so much.” – Helen Keller

Over the past few years, significant advancements have blossomed in the two key pillars of multimodal intelligence: understanding and generation (as depicted in Fig. 1(a) and (b)). For multimodal understanding, Multimodal Large Language Models (MLLMs) like LLaVA (Liu et al., 2024c) have demonstrated exceptional capabilities in vision-language tasks such as visual question-answering (VQA). For the other pillar of visual generation, denoising diffusion probabilistic models (DDPMs) (Sohl-Dickstein et al., 2015; Ho et al., 2020) have revolutionized the traditional generative paradigms (Kingma & Welling, 2013; Goodfellow et al., 2014), achieving unprecedented performance in text-to-image/video generation (Podell et al., 2023; Esser et al., 2024; Ho et al., 2022; Wu et al., 2023a).

Given these achievements in individual fields, it is only natural to explore the potential of connecting them. Recent works (Wu et al., 2023b; Ge et al., 2024; Ye et al., 2024a; Tang et al., 2024) have tried to assemble expert models from these two different domains to form a unified system that can handle both multimodal understanding and generation. However, existing attempts mainly treat each domain independently and often involve individual models responsible for understanding and generation separately (as shown on the left of Fig. 1(c)). For instance, NExT-GPT (Wu et al., 2023b) employs a base language model for multimodal understanding but requires an additional pre-trained diffusion model for image generation. Nonetheless, the mainstream understanding models like LLaVA are of transformer architecture (Vaswani et al., 2017b) while each leading generation models like Stable Diffusion 3 (SD3) (Esser et al., 2024) are just another transformer. This motivates a research question: **can one single transformer handle both multimodal understanding and generation?**

Very recently, Chameleon (Team, 2024) demonstrated this is possible. Specifically, Chameleon enables an early fusion of different modalities to generate both text and image tokens through the same manner of autoregressive modeling. While it is reasonable to model text tokens autoregressively (Touvron et al., 2023; Liu et al., 2024c), it is less clear whether it is better to model image

[†]Equal Contribution.

*Corresponding Author.

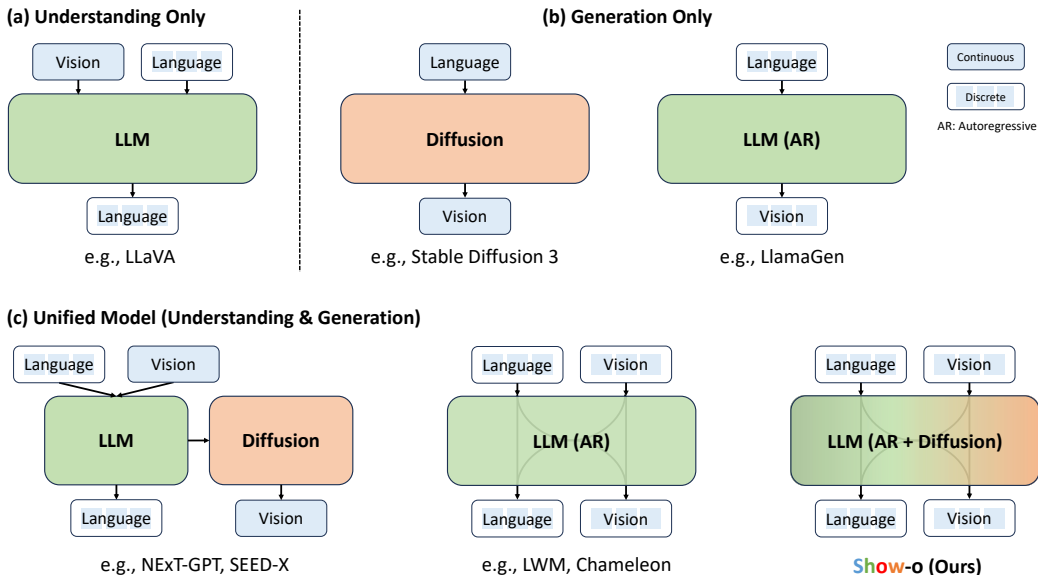


Figure 1: Characteristics comparison among understanding only, generation only, and unified (understanding & generation) models. “Vision” and “Language” indicate the representations from specific input modalities. In this context, “Diffusion” represents both continuous and discrete diffusion.

patches (or pixels) autoregressively as well. An apparent and significant bottleneck of autoregressively predicting an image is the large number of sampling steps required due to its causal attention, particularly when dealing with images/videos in higher resolution. Further, (continuous) diffusion models (Podell et al., 2023; Esser et al., 2024) have exhibited superior capabilities in visual generation than autoregressive ones and are in full attention.

This motivates us to ponder: **can such one single transformer involve both autoregressive and diffusion modeling?** Here we envision a new paradigm that text is represented as discrete tokens and modeled autoregressively same with large language models (LLMs), and continuous image pixels are modeled using denoising diffusion. However, it is non-trivial to integrate these two distinct techniques into one single network due to the significant differences between discrete text tokens and continuous image representations. Another challenge lies in the fact that existing state-of-the-art diffusion models typically rely on two distinct models, *i.e.*, a text encoder to encode text conditional information and a denoising network to predict noise. and diffusion modeling (as shown in Fig. 2). Specifically, Show-o is built upon a pre-trained LLM and inherits the autoregressive modeling capability for text-based reasoning. Inspired by Gu et al. (2022); Chang et al. (2022), we employ the discrete denoising diffusion to model discrete image tokens instead of continuous representations. Besides, Show-o inherently encodes text conditional information, eliminating additional text encoders. To accommodate diverse input data and variations of tasks, a text tokenizer and image tokenizer are employed to encode them into discrete tokens, and a unified prompting strategy is proposed to further process these tokens into structure sequences as input. Consequently, given an image accompanying questions, Show-o gives the answers autoregressively. When provided only text further to Show-o generates images in a style of discrete denoising diffusion.

Quantitatively, Show-o demonstrates comparable even better performance to individual models with an equivalent or larger number of parameters across benchmarks. In contrast to autoregressively generating an image, Show-o requires approximately 20 times fewer sampling steps, exhibiting inherent potential in acceleration. Besides, as shown in Fig. 2, Show-o naturally supports various downstream applications like text-based inpainting and extrapolation, without any fine-tuning. Moreover, we have demonstrated that Show-o has the potential for mixed-modality generation like interleaved video keyframe generation with text descriptions. This demonstrates the potential of the unified model as a feasible paradigm for long-form video generation. Beyond, we investigate the impact of different types of image representations (discrete or continuous) on the multimodal understanding performance, presenting systematic insights for the design of a unified model in the future.

In Fig. 1, we present a comparison of model characteristics between Show-o and existing representative methods across various domains. One can observe that Show-o is a unified model that flexibly involves existing advanced techniques to comprehensively address multimodal understanding and generation. Collectively, the main contributions of this paper can be summarized as:

- We present a unified model, *i.e.*, Show-o, which unifies multimodal understanding and generation using one single transformer.
- Show-o innovatively unifies autoregressive and (discrete) diffusion modeling within one single transformer, demonstrating versatility in handling both text and images distinctly.
- As a unified model, Show-o demonstrates comparable even better performance to individual baseline models with an equivalent or larger number of parameters in multimodal understanding and generation benchmarks.
- Show-o inherently supports various downstream applications like text-based inpainting and extrapolation, without necessitating any fine-tuning. Besides, it also demonstrates the potential for mixed-modality generation.
- We explore the impact of different types of representations (discrete or continuous) on multimodal understanding, providing valuable insights for improving multimodal understanding capabilities of a unified model.

2 RELATED WORK

2.1 MULTIMODAL UNDERSTANDING

Significant advancements in large language models (LLMs) (Touvron et al., 2023; Brown et al., 2020; Chowdhery et al., 2023) have inspired the development of multimodal large language models (MLLMs) (Li et al., 2024; Yin et al., 2023; Bai et al., 2024). Early MLLM efforts, such as LLaVA (Liu et al., 2024c), MiniGPT-4 (Zhu et al., 2023a), and InstructBLIP (Dai et al., 2023), demonstrate notable multimodal understanding capabilities. To integrate LLMs into multimodal domains, these studies explored projecting features from a pre-trained modal-specific encoder, such as CLIP (Radford et al., 2021), into the input space of LLMs, enabling multimodal understanding and reasoning within the transformer backbone. Although there are various design choices of MLLM (McKinzie et al., 2024; Tong et al., 2024), such as vision encoder, feature alignment adapter, and dataset, the training for most of these models adheres to the autoregressive generation paradigm, which is shown to be an effective approach of text-generation in LLMs. Despite their strong multimodal understanding capabilities, these models primarily focus on visual perception and lack the ability to generate multimodal outputs beyond text.

2.2 VISUAL GENERATION

Autoregressive models. Transformer models (Vaswani et al., 2017a; Raffel et al., 2020; Radford et al., 2018; Brown et al., 2020; Touvron et al., 2023) have demonstrated great success of autoregressive modeling in natural language processing. Inspired by such progress, previous studies (Parmar et al., 2018; Esser et al., 2021a; Ravuri & Vinyals, 2019; Chen et al., 2020; Kondratyuk et al., 2023) directly apply the same autoregressive modeling to learn the dependency of image pixels for image/video generation. For instance, VideoPoet (Kondratyuk et al., 2023) also employs the decoder-only transformer architecture for synthesizing high-quality videos from multimodal inputs. More recently, LlamaGen (Sun et al., 2024) has demonstrated that large language model architecture like Llama (Touvron et al., 2023) can also autoregressively model image tokens, which can accordingly obtain decent performance in class-conditional image generation.

Diffusion models. In recent years, numerous diffusion-based methods (Rombach et al., 2022; Ramesh et al., 2022b;a; Peebles & Xie, 2023; Podell et al., 2023; Chen et al., 2024; Nichol et al., 2021; Xue et al., 2024; Xie et al., 2023; Wu et al., 2023a) have demonstrated exceptional capabilities in text-to-image/video generation. Typically, the denoising diffusion process is operated on the continuous latent space, encoded by a VAE encoder. In this framework, the model is tasked with predicting Gaussian noise added to the continuous latent representations. In contrast, D3PM (Austin et al., 2021), Mask-predict (Ghazvininejad et al., 2019), ARDM (Hoogeboom et al., 2022) and

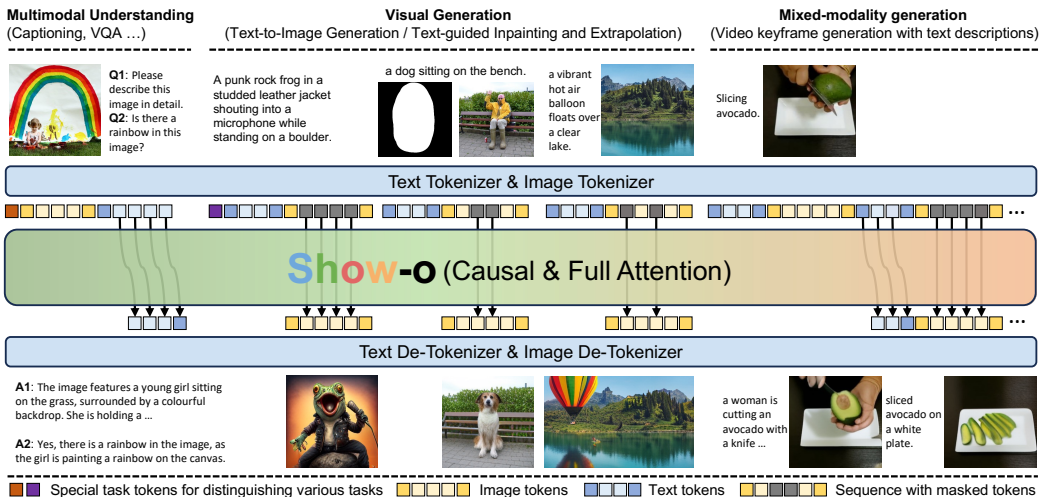


Figure 2: An overview of Show-o. The input data, regardless of its modalities, is tokenized and then prompted into a formatted input sequence. Show-o processes text tokens autoregressively with causal attention and image tokens in (discrete) denoising diffusion modeling via full attention, and then generates the desired output. Specifically, Show-o is capable of handling image captioning, visual question answering, text-to-image generation, text-guided inpainting/extrapolation, and mixed modality generation.

MaskGIT (Chang et al., 2022) formulate a discrete corruption process as an alternative to Gaussian diffusion. As summarized by Murphy (2023), specifically, an image is represented as a sequence of discrete tokens using image tokenizers (Esser et al., 2021b; Yu et al., 2023; Gu et al., 2024), and each token is associated with a categorical label. In this way, the token-wise distribution can be transformed into a uniform distribution through a stochastic sampling process. During training, a portion of these tokens is randomly masked, and the model is trained to predict the original values of these masked tokens. In this work, Show-o adopts the discrete diffusion modeling for visual generation.

2.3 UNIFIED VISION-LANGUAGE FOUNDATION MODEL

In recent years, an increasing number of studies (Ge et al., 2024; Wu et al., 2023b; Tang et al., 2024; Ye et al., 2024a) have focused on unified multimodal language models capable of both comprehension and generation. Some efforts (Zhu et al., 2023b; Sun et al., 2023c;b) use continuous representations interleaved with text tokens for autoregressive modeling to generate images. SEED-X (Ge et al., 2024) proposes a unified and versatile foundation system capable of handling both multimodal understanding and generation tasks. In this approach, continuous image representations from CLIP ViT encoder (Radford et al., 2021) are combined with text tokens and fed into a large language model (LLM) to perform next-word prediction and image representations regression. Chameleon (Team, 2024) introduces a family of token-based mixed-modal models capable of both comprehending and generating images. This approach represents all modalities as discrete tokens and utilizes a unified transformer-based architecture and trains the model from scratch in an end-to-end manner. Compared to this work, we also adopt discrete tokens to represent all modalities. In contrast, we utilize a discrete diffusion process instead of autoregressive modeling for visual generation.

3 METHODOLOGY

The objective is to develop a unified model that involves auto-regressive and diffusion modeling for jointly multimodal understanding and generation. Developing such a unified model poses non-trivial challenges, with core issues revolving around: **i)** How to define the model’s input/output space; **ii)** How to unify various kinds of input data from different modalities; **iii)** How to involve both auto-

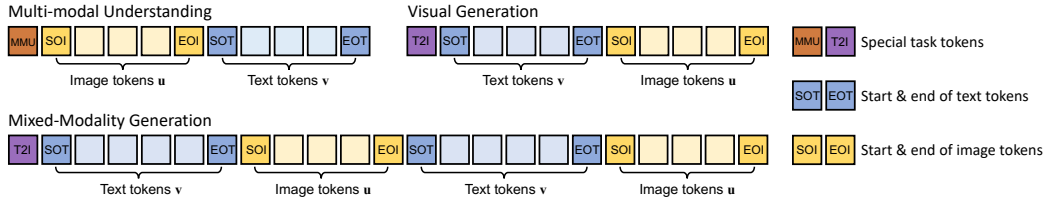


Figure 3: Illustration of the proposed unified prompting format.

regressive and diffusion modeling in one single transformer; **iv**) How to effectively train such a unified model.

Herein, we outline our packaged solutions – Show-o, to address the above challenges respectively. § 3.1: Initially, we elucidate the construction of input/output space by tokenizing text and image data into discrete tokens. § 3.2: We then introduce the default architecture of Show-o and the unified prompting strategy to structure the input data and modalities. § 3.2: Additionally, we also illustrate how to incorporate both auto-regressive and diffusion modeling within one single transformer. § 3.3: We finally present a three-stage training pipeline to effectively train the unified model.

3.1 TOKENIZATION

Given that the proposed Show-o is built upon pre-trained LLMs (Li et al., 2023; Touvron et al., 2023), it is natural to perform the unified learning on the discrete space. In this way, we maintain a unified vocabulary that includes discrete text and image tokens such that the unified model can be tasked with the same learning objective, *i.e.*, predicting discrete tokens.

Text Tokenization. Show-o is based on a pre-trained LLM and we utilize the same tokenizer for text data tokenization without any modifications.

Image Tokenization. Following MAGVIT-v2 (Yu et al., 2023), we train a lookup-free quantizer using around 35M image data. The quantizer maintains a codebook of size 8,192 and encodes images of 256×256 resolution into 16×16 discrete tokens (option (a) in Fig. 4). The reason for utilizing MAGVIT-v2 lies in its ease of fine-tuning to serve as a video tokenizer with temporal compression capability, a potential aspect that we intend to explore in the future.

An alternative approach is to use different tokenizers for understanding and generation, respectively. Inspired by existing studies (Liu et al., 2024c;b), we also extract the continuous image representations from the pre-trained MAGVIT-v2 and CLIP-ViT (Radford et al., 2021) encoder as input for exploring the improvement of multimodal understanding capabilities (options (b) and (c) in Fig. 4). We will present more details and discuss this exploration in Section 4.5. In the following sections, the default Show-o employs discrete image tokens as input for both multimodal understanding and generation (option (a) in Fig. 4). For simplicity, we only elaborate on the default Show-o in the methodology sections.

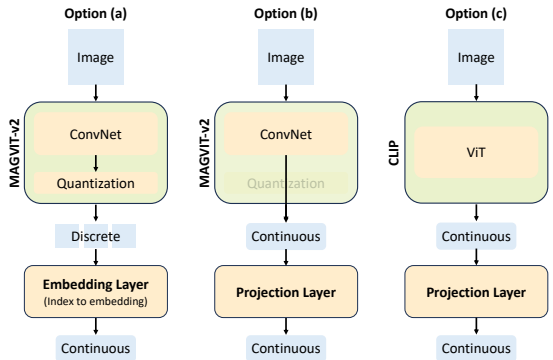


Figure 4: Optional inputs for multimodal understanding.

3.2 ARCHITECTURE

Show-o inherits the architecture of existing LLMs such as (Li et al., 2023; Touvron et al., 2023) without any architecture modifications except for prepending a QK-Norm operation (Dehghani et al., 2023; Wortsman et al., 2023; Team, 2024) to each attention layer. We initialize Show-o with the weights of a pre-trained LLM and expand the size of the embedding layer by incorporating 8,192 new learnable embeddings for discrete image tokens. Unlike state-of-the-art diffusion models that

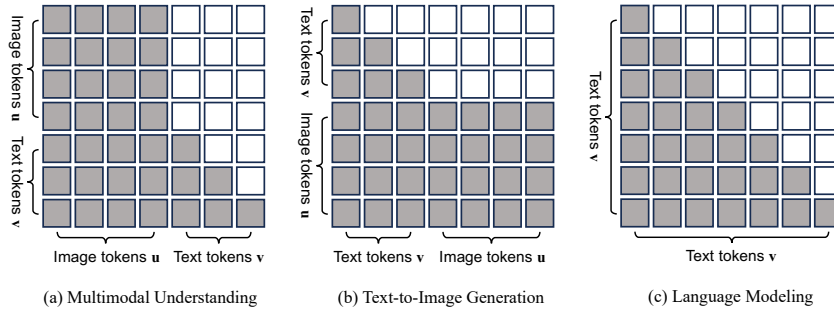


Figure 5: Omni-Attention Mechanism (The dark squares represent ‘allow to attend’, while the white squares indicate ‘prevent from attending’). It is a versatile attention mechanism with causal and full attention that adaptively mixes and changes according to the format of the input sequence. As illustrated in (a), (b), and (c), within a sequence containing both text and image tokens, the omni-attention mechanism distinctly processes text tokens using causal attention and image tokens using full attention. Besides, concerning the input sequence, (a) text tokens can attend to all preceding image tokens, (b) image tokens can access all preceding text tokens, and (c) in cases where only text tokens are provided, the attention degrades to causal attention.

require an additional text encoder, Show-o inherently encodes text conditional information by itself for text-to-image generation.

Unified Prompting. To perform unified learning on multimodal understanding and generation, we design a unified prompting strategy to format various kinds of input data. Given an image-text pair (\mathbf{x}, \mathbf{y}) , it is first tokenized into M image tokens $\mathbf{u} = \{u_i\}_{i=0}^M$ and N text tokens $\mathbf{v} = \{v_i\}_{i=0}^N$ by the image and text tokenizer, respectively. We form them into an input sequence according to the type of task in the format illustrated in Fig. 3. Specifically, [MMU] and [T2I] are pre-defined task tokens that indicate the learning task for the input sequence. [SOT] and [EOT] serve as special tokens denoting the start and end of text tokens, respectively. Similarly, [SOI] and [EOI] are pre-defined special tokens marking the start and end of image tokens.

By employing this prompt design, we can effectively encode various input data for multi-modal understanding, text-to-image generation, and mixed-modality generation as sequential data. This setup enables unified learning to operate seamlessly within sequences across these various tasks. Once trained, we can accordingly prompt Show-o to handle various vision-language tasks including visual question answering and text-to-image generation (as shown in Fig. 2).

Omni-Attention Mechanism. Different from existing works (Touvron et al., 2023; Team, 2024) that model sequence auto-regressively only, we propose an omni-attention mechanism to enable Show-o to model various types of signals in distinct ways. It is a comprehensive attention mechanism with causal and full attention that adaptively mixes and changes according to the format of the input sequence. We illustrate omni-attention examples for different input sequences in Fig. 5. Specifically, Show-o model text tokens \mathbf{v} within the sequence via causal attention. For image tokens \mathbf{u} , Show-o processes them via full attention, allowing each token to comprehensively interact with all others. Given a formatted input sequence, it is apparent that in multimodal understanding (Fig 5(a)), text tokens in a sequence can attend to all previous image tokens, and in text-to-image generation (Fig 5(b)), image tokens are able to interact with all preceding text tokens. Omni-attention maintains the text reasoning knowledge from the pre-trained LLM and enhances the efficiency of image generation by reducing sampling steps. Moreover, it naturally supports various downstream applications like inpainting and extrapolation without necessitating any fine-tuning. When given only text tokens, it degrades to causal attention (Fig. 5(c)).

Training Objectives. To perform both auto-regressive and (discrete) diffusion modeling, we employ two learning objectives: i) Next Token Prediction (NTP) and ii) Mask Token Prediction (MTP). Given a sequence with M image tokens $\mathbf{u} = \{u_1, u_2, \dots, u_M\}$ and N text tokens $\mathbf{v} = \{v_1, v_2, \dots, v_N\}$ for multimodal understanding, we maximize the likelihood of text tokens by

employing the standard language modeling objective:

$$\mathcal{L}_{\text{NTP}} = \sum_i \log p(v_i | v_1, \dots, v_{i-1}, u_1, \dots, u_M; \Theta), \tag{1}$$

where $p(\cdot|\cdot)$ indicates the conditional probability which is modeled by the weights Θ of Show-o and stochastic gradient descent is used to train the model. Note that, if the input sequence involves only text tokens, there are no conditional terms on image tokens $\mathbf{u} = \{u_1, u_2, \dots, u_M\}$.

As proven in [Murphy \(2023\)](#), generative masked models ([Ghazvininejad et al., 2019](#); [Wang & Cho, 2019](#); [Chang et al., 2022](#)) correspond to the discrete diffusion process. In this context, we refer to the generative paradigm of MaskGIT ([Chang et al., 2022](#)) for the seamless integration of (discrete) diffusion modeling within Show-o. Hence, for modeling image tokens $\mathbf{u} = \{u_1, u_2, \dots, u_M\}$ within the input sequence, we first randomly replace the image tokens with mask token u_* at a random ratio to create a masked sequence $\mathbf{u}_* = \{u_*, u_2, \dots, u_*, u_M\}$. Next, we aim to reconstruct the original image token from the masked tokens conditioning on unmasked tokens and preceding text tokens by maximizing the following likelihood:

$$\mathcal{L}_{\text{MTP}} = \sum_j \log p(u_j | u_*, u_2, \dots, u_*, u_M, v_1, \dots, v_N; \Theta). \tag{2}$$

Specifically, we follow the sampling strategy used in ([Chang et al., 2022](#); [2023](#)) to mask image tokens and reconstruct them via the information from all text and unmasked image tokens within the input sequence. Following the classifier-free guidance introduced by [Ho & Salimans \(2022\)](#), we randomly replace the conditioned text tokens with a null text “” with some probability.

Given a batch size of input sequences, the overall training loss is the combination of \mathcal{L}_{MTP} and \mathcal{L}_{NTP} :

$$\mathcal{L} = \mathcal{L}_{\text{MTP}} + \alpha \mathcal{L}_{\text{NTP}}, \tag{3}$$

where α is the hyper-parameter weighting the loss term \mathcal{L}_{NTP} .

3.3 TRAINING PIPELINE

Given that the embedding of image tokens is newly initialized, it necessitates large-scale pre-training to align for multimodal understanding and generation. Besides, Show-o eliminates the text encoder to extract text embeddings for text-to-image generation, which poses a significant challenge for achieving effective alignment between text and image content within one single transformer. To this end, we employ a three-stage approach to progressively and effectively train Show-o:

- i) **Image Token Embedding and Pixel Dependency Learning:** We employ RefinedWeb ([Penedo et al., 2023](#)) dataset to train Show-o to maintain the language modeling ability. Meanwhile, ImageNet-1K dataset ([Deng et al., 2009](#)) and 35M image-text pairs are adopted to train Show-o for class-conditional image generation and image captioning, respectively. Here, we directly leverage the class names from ImageNet-1K as textual inputs for learning class-conditional image generation. This stage primarily involves the learning of new learnable embeddings for discrete image tokens, pixel dependency for image generation, and alignment between image and text for image captioning.
- ii) **Image-Text Alignment for Multimodal Understanding and Generation:** Building upon the pre-trained weights, we proceed to involve training of text-to-image generation on the 35M image-text data instead of the ImageNet-1K. This stage mainly focuses on image and text alignment for both image captioning and text-to-image generation.
- iii) **High-Quality Data Fine-tuning:** Lastly, we further refine the pre-trained Show-o model by incorporating filtered high-quality image-text pairs for text-to-image generation and instructional data for multimodal understanding and mixed-modality generation.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. Three types of data are adopted for training Show-o: i) **Text-only Data:** We employ the publicly available RefinedWeb dataset ([Penedo et al., 2023](#)) to preserve the text reasoning capabilities of the pre-trained LLM. This dataset comprises approximately 1 billion instances (equivalent to 968 million individual web pages) and totals 2.8 terabytes of curated text data. ii) **Image**

Table 1: Evaluation on multimodal understanding benchmarks. Show-o is currently built upon Phi-1.5 (Li et al., 2023) and thus we implement LLaVA-v1.5-Phi-1.5 as our apple-to-apple baseline. Und. and Gen. denote “understanding” and “generation”, respectively. † denotes the improved Show-o that employs CLIP-ViT continuous representations. We highlight the model size of Show-o and LLaVA baseline in green, and we use blue to highlight the larger model size than ours.

Type	Model	# Params	POPE↑	MME↑	Flickr30k↑	VQAv2 _(test) ↑	GQA↑	MMMU↑
Und. Only	LLaVA-v1.5	7B	85.9	1510.7	-	78.5	62.0	35.4
	InstructBLIP	13B	78.9	1212.8	-	-	49.5	-
	Qwen-VL-Chat	7B	-	1487.5	-	78.2	57.5	-
	mPLUG-Owl2	7B	85.8	1450.2	-	79.4	56.1	-
	LLaVA-v1.5-Phi-1.5	1.3B	84.1	1128.0	69.6	75.3	56.5	30.7
Und. and Gen.	Gemini-Nano-1	1.8B	-	-	-	62.7	-	26.3
	CoDI	-	-	-	12.8	-	-	-
	Emu	13B	-	-	77.4	57.2	-	-
	NExT-GPT	13B	-	-	84.5	66.7	-	-
	SEED-X	17B	84.2	1435.7	52.3	-	47.9	35.6
	Chameleon	34B	-	-	74.7	66.0	-	-
	Show-o (Ours)	1.3B	73.8	948.4	36.2	59.3	48.7	25.1
	Show-o† (Ours)	1.3B	84.5	1182.7	64.3	73.9	57.5	27.4

Data with Class Names: Show-o learns pixel dependencies using 1.28M images sourced from the ImageNet-1K (Deng et al., 2009) dataset. iii) **Image-Text Data:** For pre-training tasks correspond to multimodal understanding and generation, we assemble roughly 35M image-text pairs from the publicly available datasets including CC12M (Changpinyo et al., 2021), SA1B (Kirillov et al., 2023), and LAION-aesthetics-12M*. Note that, we employ ShareGPT4V (Chen et al., 2023) to re-caption these datasets. Additionally, LAION-aesthetics-12M and JourneyDB (Sun et al., 2023a) serve as high-quality datasets for the final fine-tuning. Following LLaVA-v1.5 (Liu et al., 2024b), we incorporate LLaVA-Pretrain-558K and LLaVA-v1.5-mix-665K for instruction tuning. Moreover, the GenHowTo dataset (Souček et al., 2024) is utilized for mixed-modality generation.

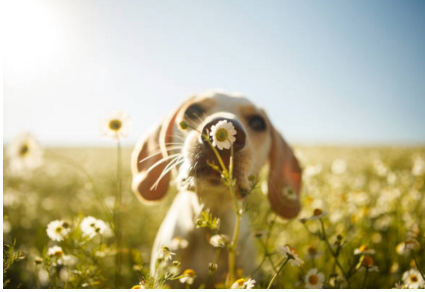
Evaluation Details. Following LLaVA (Liu et al., 2024b), we evaluate the multimodal understanding capabilities of Show-o on POPE, MME, Flickr30k, VQAv2, GQA, and MMMU benchmarks. Besides, following Stable Diffusion (Rombach et al., 2022), we adopt Fréchet Inception Distance (FID) on MSCOCO dataset to evaluate the potential generation fidelity of Show-o. Further, we follow SD3 (Esser et al., 2024) to evaluate the text-to-image generation capabilities of Show-o on the GenEval (Ghosh et al., 2023) benchmark. In experiments, we make comparisons with: i) understanding only models including LLaVA-v1.5 (Liu et al., 2024b), InstructBLIP (Dai et al., 2023), Qwen-VL-Chat (Bai et al., 2023), and mPLUG-Owl2 (Ye et al., 2024b); ii) generation only models including DALL-E (Ramesh et al., 2021), LDM (Rombach et al., 2022), DALL-E 2 (Ramesh et al., 2022a), SDv1.5 (Rombach et al., 2022), GigaGAN (Kang et al., 2023), PixArt (Chen et al., 2024), Imagen (Saharia et al., 2022), RAPHAEL Xue et al. (2024), LlamaGen (Sun et al., 2024), SDXL (Podell et al., 2023), and SD3 (Esser et al., 2024); unified models including Gemini (Anil et al., 2023), NExT-GPT (Wu et al., 2023b), Emu (Sun et al., 2023d), SEED-X (Ge et al., 2024), CoDI (Tang et al., 2024), LWM (Liu et al., 2024a), and Chameleon (Team, 2024).

Implementation Details. We initially conduct joint training of Show-o using the RefinedWeb, a collection of 35M image-text pairs, and the ImageNet-1K for language modeling, image captioning, and class-conditional image generation, respectively, over 500K steps. Subsequently, we replace the class-conditional generation with the training for text-to-image generation using the 35M image-text pairs for an additional 1,000K steps. The base model is trained on 48 A100 (80GB) GPUs with a total batch size of 1,152. We employ the AdamW optimizer with a weight decay of 0.01, 5,000 steps of warm-up, and an initial learning rate of 1e-4 with a cosine scheduling. Finally, we fine-tune Show-o with filtered high-quality image-text pairs and adhere to the configuration of LLaVA-v1.5 for instruction data tuning. Following the work Chang et al. (2023), in inference, we compute both the conditional logit l_c and unconditional logit l_u for masked tokens. The final logit l of each masked token is obtained by the following equation with a guidance scale t :

$$l = (1 + t)l_c - tl_u. \tag{4}$$

Note that, the current version of Show-o is based on Phi-1.5 (Li et al., 2023). In the following experiment sections, the default Show-o employs discrete image tokens as input for both multimodal

*<https://huggingface.co/datasets/dclure/laion-aesthetics-12m-umap>



User: Please describe this image in detail.

Show-o:

The image features a small dog standing in a field of flowers. The dog is positioned in the center of the scene, with its head tilted to the side as it sniffs the flowers. The flowers are scattered throughout the field, with some closer to the dog and others further away. The dog appears to be enjoying the beautiful surroundings, possibly smelling the flowers as a form of sensory exploration.

Chameleon:

This image is a close-up photograph of a dog's nose and face in a field of flowers. The dog's face is mostly out of focus, but the nose is in sharp focus. The dog is a golden retriever and is lying down on the ground in the middle of the image. There is a blurry yellow flower in the foreground of the image, with a clear yellow flower in the middle of the image. The flower in the middle has a few small yellow flowers.

SEED-X:

This image captures a charming scene of a small, brown and white dog in a field of daisies. The dog, with its tongue playfully sticking out, is in the center of the frame, surrounded by a sea of white daisies. The field extends to the horizon, suggesting a vast expanse of nature. The sky above is a clear blue, adding to the serene atmosphere of the scene. The dog appears to be enjoying its time in the field, perhaps taking a moment to smell the flowers ...



User: Do you think the image is unusual or not?

Show-o:

The image is unusual because it depicts a living room with a couch and a table in the middle of the ocean, surrounded by a large body of water. **This is not a typical setting for a living room, as living rooms are usually indoors and designed for relaxation and entertainment.** The presence of the couch and table in the ocean adds a unique and unexpected element to the scene, making it stand out from typical living room setups.

Chameleon:

The image is unusual because of the transparent and colorful fish and coral in the water. The image is also unusual because of the clear blue sky with white clouds. The unusual objects in the image are the fish and coral in the water. The unusual objects in the image are the fish and coral in the water. The unusual objects in the image are the fish and coral in the water.

SEED-X:

Yes, the image is unusual. It features a living room set, including a couch and two chairs, placed in the middle of the ocean. The room is surrounded by water, with coral reefs and fish visible in the background. **This is a surreal and unexpected scene, as living rooms are typically found on land, not in the middle of the ocean.** The presence of the couch and chairs in such an environment adds to the peculiarity of the image.

Figure 6: Comparisons of VQA capabilities among Chameleon, SEED-X, and Show-o. We emphasize distinctions in answers from different methods using a different color in the second example.

understanding and generation. Show-o[†] and Show-o[‡] indicate the use of continuous image representations from the pre-trained MAGViT-v2 and CLIP-ViT (corresponding to options (b) and (c) in Fig. 4), respectively, for multimodal understanding and we discuss this exploration in Section 4.5.

4.2 MULTIMODAL UNDERSTANDING

4.2.1 QUANTITATIVE EVALUATION

Table 1 presents the multimodal understanding capability of Show-o on public benchmarks, such as image captioning and visual question-answering tasks. i) The current version of Show-o is built upon Phi-1.5 (Li et al., 2023) and thus we follow LLaVA to train Show-o’s understanding only counterpart as our direct baseline, namely LLaVA-v1.5-Phi-1.5. The proposed Show-o exhibits comparable performance in all evaluation metrics to the baseline LLaVA-v1.5-Phi-1.5, which is dedicated and optimized to only multimodal understanding. This demonstrates the great potential of our framework to unify multimodal understanding and generation in one single transformer. ii) When comparing with

understanding only models including InstructBLIP (Dai et al., 2023), Qwen-VL-Chat (Bai et al., 2023), and mPLUG-Owl2 (Ye et al., 2024b) on multimodal understanding, our model with a much smaller model size also achieves competitive performance on POPE, MME, Flickr30k and VQAv2 benchmarks and performs better on GQA benchmark. iii) Compared with unified models with a much larger number of parameters, such as NExT-GPT-13B (Wu et al., 2023b) and Chameleon-34B (Team, 2024), our model also achieves decent performance on Flickr30k benchmark and perform much better on VQAv2 benchmark. Given such promising results, we envision Show-o as a potential next-generation foundation model for unifying understanding and generation. These results also demonstrate the potential of scaling Show-o to achieve state-of-the-art performance.

4.2.2 QUALITATIVE RESULTS

We present Show-o’s visual question-answering capability and make comparisons with Chameleon and SEED-X in Fig.6. It is evident that when presented with a query image, Show-o can describe the image in detail and respond to commonly asked questions, even addressing the unusual aspects within the image. In the example at the top of Fig.6, Chameleon, SEED-X, and Show-o all provide comprehensive descriptions of the image’s main content. However, when asked, “Do you think the image is unusual or not?” Chameleon fails to correctly identify the unusual aspect, and SEED-X’s response, while identifying the unusual, lacks precision, stating “as living rooms are typically found **on land.**” In contrast, Show-o’s response, “as living rooms are **usually indoors** and designed for relaxation and entertainment,” is more accurate.

4.3 VISUAL GENERATION

4.3.1 QUANTITATIVE EVALUATION

Results on MSCOCO 30K. We present zero-shot FID of Show-o on MSCOCO 30K in Table 2. It can be observed that, compared to generation models trained with larger numbers of parameters and training images such as GLIDE and DALL-E 2, Show-o achieves a better FID, *i.e.*, 9.24, with only 1.3B parameters and 35M training data. Though GigaGAN, Imagen, and RAPHAEL obtain a relatively better performance than Show-o, they are much larger in model size (3B v.s. 1.3B) and trained with much more data. In comparison

Table 2: MSCOCO zero-shot FID. Und. and Gen. denote “understanding” and “generation”, respectively.

Type	Method	# Params	# Images	FID-30K↓
Gen. Only	DALL-E	12B	250M	27.50
	GLIDE	5B	250M	12.24
	LDM	1.4B	400M	12.64
	DALL-E 2	6.5B	650M	10.39
	SDv1.5	0.9B	2000M	9.62
	GigaGAN	0.9B	2700M	9.09
	PixArt	0.6B	25M	7.32
	Imagen	3B	860M	7.27
	RAPHAEL	3B	5000M+	6.61
Und. and Gen.	CoDI	-	400M	22.26
	LWM	7B	-	12.68
	SEED-X	17B	-	14.99
	Show-o (Ours)	1.3B	35M	9.24

to unified models, Show-o also exhibits improvement. The above validates that Show-o, a unified transformer, can obtain competitive even better generation performance compared to individual models (generation only) with an equivalent or larger number of parameters and training data. However, it is worth noting that FID on MSCOCO 30K may not comprehensively accurate assessment of generation fidelity. The reason lies in the fact that existing generation models are commonly fine-tuned with high-quality and aesthetic images that do not align with the distribution of the MSCOCO dataset. Consequently, this mismatch leads to inaccurate measurement of generation fidelity.

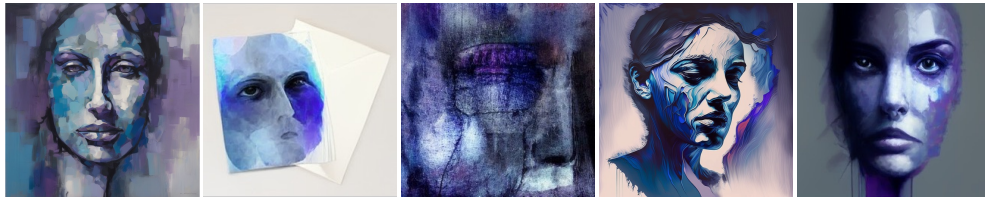
Results on GenEval. Following SD3 (Esser et al., 2024), we evaluate text-to-image generation capability of Show-o on the GenEval benchmark (Ghosh et al., 2023) across six dimensions including “Single Object”, “Two Objects”, “Counting”, “Colors”, “Position”, “Color Attribute” and present the quantitative comparison in Table 3. One can observe that when comparing to the model in a similar size such as LDM (1.4B), Show-o obtains significantly better performance in all six metrics, with an improvement of around 0.15 overall. Besides, Show-o achieves a relatively better performance than DALL-E 2, which is 5 times larger in model size. Further, Show-o, with only 1.3B parameters, achieves comparable performance to models with two times larger number of parameters such as SDXL (2.6B) and SD3 (2B). It indicates that our unified model’s generative capabilities are comparable to or even surpass those of specialized generation models. In comparison to unified models such as CoDI and SEED-X, Show-o also demonstrates significant improvements.



“A dynamic scene of a rally car race.”



“The breathtaking view of Santorini, a renowned landmark in Greece. The white-washed buildings with blue domes overlook the deep blue waters of the Aegean Sea, creating a stunning contrast against the vibrant sunset.”



“An abstract portrait of a pensive face, rendered in cool shades of blues, purples, and grays.”



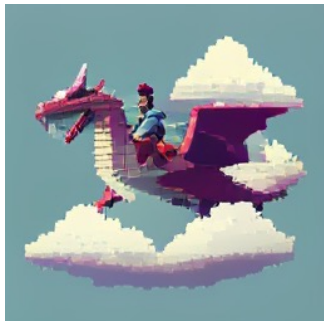
“Close-up view of a computer screen, with the screen displaying a webpage.”



“A tranquil scene of a lotus pond with koi fish swimming gracefully in a peaceful Chinese ink painting.”



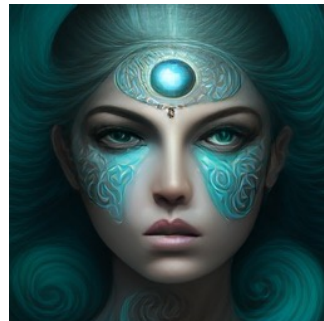
“Paper artwork, layered paper, colorful Chinese dragon surrounded by clouds.”



“Pixel art character riding a dragon through the clouds.”



“A peaceful village nestled at the foot of towering mountains in a tranquil East Asian watercolor scene.”



“A person with swirling patterns of teal paint on their face and a shimmering silver crescent moon placed above their eyebrow, symbolizing mystery and magic.”

Figure 7: (Top) Qualitative comparisons among SDXL, LlamaGen, LWM, SEED-X, and Show-o. (Bottom) Samples generated by Show-o.

Table 3: Evaluation of text-to-image generation on the GenEval (Ghosh et al., 2023) benchmark. Und. and Gen. denote “understanding” and “generation”, respectively. We highlight the model size of Show-o in green, and we use blue to highlight the larger model size than ours. Obj.: Object. Attri.: Attribute.

Type	Method	# Params	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall↑
Gen. Only	LlamaGen	0.8B	0.71	0.34	0.21	0.58	0.07	0.04	0.32
	LDM	1.4B	0.92	0.29	0.23	0.70	0.02	0.05	0.37
	SDv1.5	0.9B	0.97	0.38	0.35	0.76	0.04	0.06	0.43
	PixArt-alpha	0.6B	0.98	0.50	0.44	0.80	0.08	0.07	0.48
	SDv2.1	0.9B	0.98	0.51	0.44	0.85	0.07	0.17	0.50
	DALL-E 2	6.5B	0.94	0.66	0.49	0.77	0.10	0.19	0.52
	SDXL	2.6B	0.98	0.74	0.39	0.85	0.15	0.23	0.55
	SD3 (d=24)	2B	0.98	0.74	0.63	0.67	0.34	0.36	0.62
Und. and Gen.	CoDI	-	0.89	0.16	0.16	0.65	0.02	0.01	0.31
	LWM	7B	0.93	0.41	0.46	0.79	0.09	0.15	0.47
	SEED-X	17B	0.97	0.58	0.26	0.80	0.19	0.14	0.49
	Show-o (Ours)	1.3B	0.95	0.52	0.49	0.82	0.11	0.28	0.53

4.3.2 QUALITATIVE RESULTS

Qualitative Comparisons. We present qualitative comparisons with diffusion-based models, *e.g.*, SDv1.5, SDXL, auto-regressive based model, *i.e.*, LlamaGen, and unified models including LWM and SEED-X at the top of Fig. 7. One can observe that, given either short or long text prompts, Show-o can generate realistic images with consistent content described in the prompts. Compared to SDv1.5 and LlamaGen, Show-o demonstrates better visual quality and image-text alignment. For instance, as shown in the second column, both SDv1.5 and LlamaGen cannot fully comprehend the text prompt and miss some attributes such as sunset and blue domes in the generated images. In comparison to SDXL, Show-o exhibits comparable visual quality and alignment such as “a rally car race” and “stunning contrast against the vibrant sunset”. More samples generated by Show-o can be found at the bottom of Fig.7. One can observe that Show-o is capable of generating diverse and interesting visual content.

Text-guided Inpainting and Extrapolation. As mentioned, Show-o naturally supports text-based inpainting and extrapolation without requiring any fine-tuning. We illustrate examples in Fig. 8. As shown on the top of the figure, given an input image and inpainting mask, Show-o can inpaint the original red trolley car to a blue sports car with sleek curves and tinted windows based on the user-provided text prompt. Besides, Show-o is capable of extrapolating the original image horizontally/vertically based on the given text prompt. Further, we can flexibly extrapolate the original image with new objects or scenes such as “red wildflowers” (as illustrated in the second row). One can observe that the pixels in both inpainted regions and extrapolated ones are consistent with the original ones. These cases significantly demonstrate the inherent advantages of Show-o over those auto-regressive models for downstream applications.

4.4 MIXED-MODALITY GENERATION OF VIDEO KEYFRAMES AND CAPTIONS

Here, we explore the mixed-modality generation ability of Show-o based on the text descriptions and video keyframes in the GenHowTo dataset. Given a sequence of interleaved text descriptions and video keyframes (as shown at the bottom of Fig. 3), Show-o is trained to predict the next text tokens or keyframe tokens conditioning on all preceding tokens. Thus, Show-o can generate mixed-modality of text descriptions and video keyframes. Examining a single frame, these tokens are generated in a diffusion manner. When considering the modeling of long sequences, as subsequent keyframes are produced based on all preceding text and image information, this can also be viewed as a form of temporal auto-regressive modeling. Consequently, it becomes feasible to generate consistent video keyframes continuously. Substituting video keyframes with video clips could offer a viable strategy for long video generation, a direction we aim to explore in the future.

We present qualitative examples in Fig. 9. As shown on the top of Fig. 9, given text prompts, Show-o can generate consistent video keyframes. Besides, we have tried to train Show-o using instructional examples. For example, given a question “Can you guide me through making Avocado and Apple Juice”, Show-o exhibits the capability to generate video keyframes with text descriptions related to the question. It is apparent that the generated keyframes are temporally consistent. This exploration

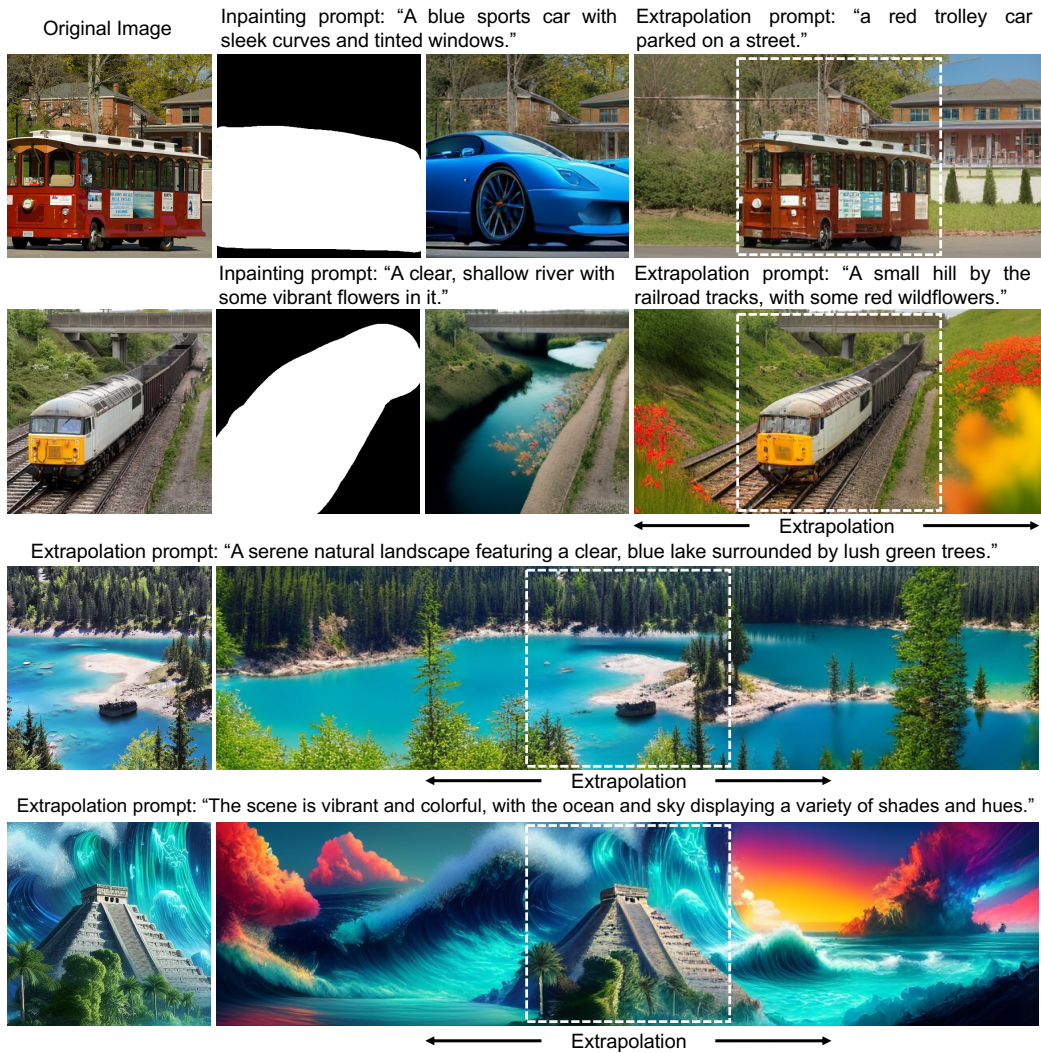


Figure 8: Examples of text-guided image inpainting/extrapolation.

reveals the potential of our model to be extended to the domain of long video generation, in which the model can produce a sequence of video clips. Specifically, the model continuously plans the next-scene textual prompt and proceeds to generate videos iteratively, allowing for the continuous generation of subsequent videos.

4.5 ABLATION STUDIES

As illustrated in Fig. 4(a), the default Show-o adopts the pre-trained MAGViT-v2 to tokenize input image to discrete tokens, which are then passed to the embedding layer to obtain embeddings as input for multimodal understanding. Beyond, we provide a systematic exploration of different design choices for the input of Show-o to enhance multimodal understanding. Specifically, as shown in Fig. 4(b) and (c), instead of discrete image tokens, we extract the continuous image representations from the pre-trained MAGViT-v2 and CLIP-ViT, respectively, as input for Show-o when dealing with multimodal understanding. The experimental results are illustrated in Table 4. Through this exploration, we unveil the following lessons and insights.

Impact of Vision Encoder for Multimodal Understanding. The default Show-o employs MAGViT-v2 to encode images into discrete tokens for both multimodal understanding and generation. Inspired by the literature (Liu et al., 2024b), we investigate the impact of the most popular design choice of vision encoder, *i.e.*, the pre-trained CLIP ViT (Radford et al., 2021), for mul-

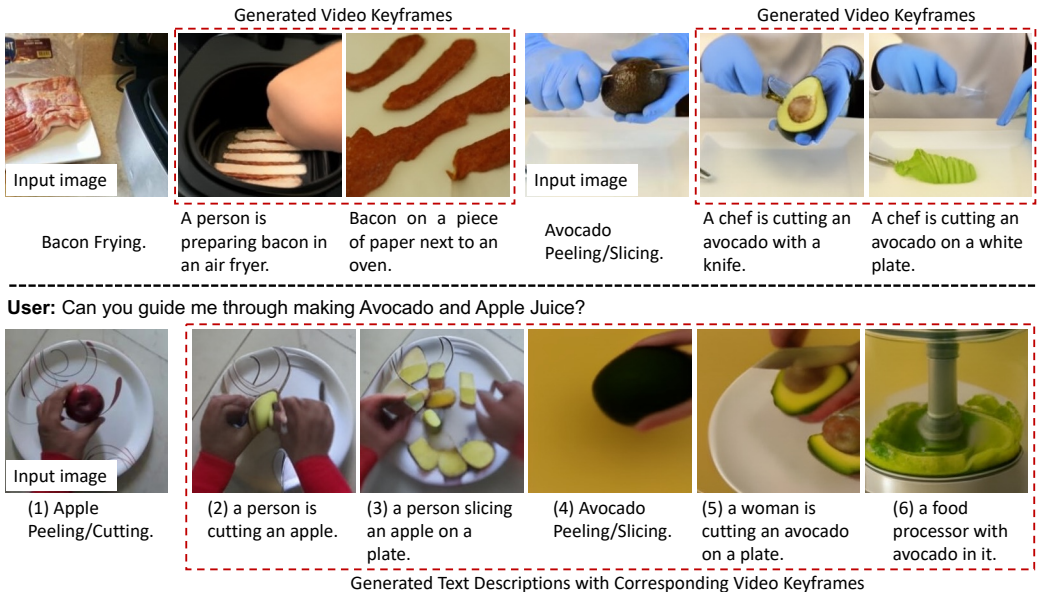


Figure 9: Mixed-modality generation. (Top) Given the first keyframe with text prompts for each frame, Show-o can correspondingly generate consistent video keyframes conditioning on the previous texts and keyframes. (Bottom) Given an instruction and the first frame with text descriptions, Show-o can alternately generate text descriptions and corresponding video keyframes.

timodal understanding. We first compare the two settings using our Show-o model. In Table 4, the comparison between Exp 2 and Exp 4, Exp 3 and Exp 5 clearly demonstrates that continuous representations from CLIP-ViT have significantly better performance on multimodal understanding than that of MAGVIT-v2. This mainly attributes to: i) The CLIP-ViT is pre-trained on a much larger dataset (400M) than that of our pre-trained MAGVIT-v2 (35M); ii) In contrast to image reconstruction learning objective in MAGVIT-v2, the discriminative loss, *i.e.*, image-text matching, in CLIP-ViT makes the extracted representations easier to be adapted for multimodal understanding.

Impact of Various Representations for Multimodal Understanding. In typical multimodal understanding models like LLaVA, the image representation extraction and cross-modal alignment usually happen in the continuous space. However, image tokenizers such as MAGVIT-v2 naturally yield discrete image tokens. As shown in Table 4, we compare the two types of input, *i.e.*, continuous representations and discrete tokens, in the multimodal understanding scenario. In Exp 6 and 7, we use the pre-trained MAGVIT-v2 to extract discrete tokens and train an embedding layer to embed the tokens into the continuous embedding space of the LLM. In Exp 4 and 5, we modify MAGVIT-v2 to output continuous representations without quantization. The cross-modal projection layer follows the setting of LLaVA. The comparison between Exp 5 and Exp 7 reveals that discrete tokens show much worse performance on most benchmarks. We attribute the performance gap to that popular multimodal understanding datasets, *e.g.*, LLaVA-Pretrain-558K, are not sufficient to align discrete image tokens into the language space, leading to an unsatisfactory cross-modal understanding. In contrast, continuous representations, already lying in a well-shaped embedding space, are much easier to align.

Impact of Unified Pre-training for Multimodal Understanding. Our training pipeline involves two-stage unified pre-training to learn image token embedding and image-text alignment for multimodal understanding and generation (as described in Section 3.3). Here we elaborate on the impact of the unified pre-training with different vision encoders and types of representations:

- CLIP-ViT with Continuous Representations. The comparison between Exp 2 and Exp 3 shows that the unified pre-training has a small negative effect on the CLIP ViT-based understanding, as the performance on most benchmarks has marginal degradations. We hypothesize that the MAGVIT-v2 token-based pre-training and the CLIP ViT-based tuning

Table 4: Ablation studies of various vision encoders and kinds input representations for multimodal understanding.

# Exp	Method	Vision Encoder	Unified Pretrain	Feature type	POPE	MME	Flickr30k	VQAv2 _(val)	GQA	MMMU
1	LLaVA	CLIP-ViT	✗	Continuous	84.1	1128.0	69.6	73.0	56.5	30.67
2	Show-o [‡]	CLIP-ViT	✓	Continuous	84.5	1182.7	64.3	71.9	57.5	27.4
3	Show-o [‡]	CLIP-ViT	✗	Continuous	84.5	1161.6	68.5	73.5	58.7	29.2
4	Show-o [†]	MAGVIT-v2	✓	Continuous	74.3	947.8	33.9	59.4	51.0	26.7
5	Show-o [†]	MAGVIT-v2	✗	Continuous	65.1	800.0	12.3	50.8	43.9	24.6
6	Show-o	MAGVIT-v2	✓	Discrete	73.8	948.4	36.2	57.8	48.7	25.1
7	Show-o	MAGVIT-v2	✗	Discrete	63.8	689.1	4.5	46.1	40.5	28.1

“A vibrant impressionist landscape painting depicting a serene rural scene with a loose, sketchy style, using a palette of bright blues, greens, and yellows.”



Figure 10: Illustration of generated samples using different sampling steps and classifier-free guidance scale t .

happen in nearly orthogonal dimensions, and the capability of the backbone has been spared to maintain the compatibility of the two tasks.

- **MAGVIT-v2 with Continuous Representations.** In the comparison between Exp 4 and Exp 5, we also notice a performance improvement brought by the unified pre-training, even though the pre-training uses discrete tokens while the experiments here use continuous features. This comparison further validates the hypothesis that unified pre-training enhances the multimodal understanding and reasoning capabilities of the backbone by diverse multimodal interactions during pre-training.
- **MAGVIT-v2 with Discrete Tokens.** The comparison between Exp 6 and Exp 7 shows that the unified pre-training has significantly boosted the multimodal understanding performance. This is intuitive since the pre-training also adopts MAGVIT-v2 discrete tokens as image representation. Specifically, we attribute the performance gain to that unified pre-training learns a better cross-modal alignment with large-scale data and enhances the multimodal understanding capabilities of the backbone.

Additionally, we present qualitative examples to illustrate the impact of sampling steps and classifier-free guidance for text-to-image generation in the following.

Impact of Sampling Steps. We present generated results at 256×256 resolution with varying sampling steps on the left of Fig. 10. With just two steps, Show-o can produce an image that is roughly related to the given prompt. Increasing the sampling steps to 8 allows the synthesis of an image that closely adheres to the prompt. When the sampling step is set as 16, the generated image becomes more detailed and realistic. In contrast, auto-regressive models Team (2024); Sun et al. (2024) require 256 sampling steps to generate an image of the same resolution when the downsampling rate is 16, which is 16 times more steps than our approach.

Impact of Classifier-free Guidance. The visual variations of generated images with different classifier-free guidance scales t are illustrated on the right of Fig. 10. It can be observed that the generated images lack detail and objects without classifier-free guidance. As the classifier-free guidance scale t is gradually increased to 1.25 and 1.75, the colors and contents become more diverse and consistent with the given text prompt.

4.6 FAILURE CASES

We provide failure cases of Show-o in multimodal understanding and generation in Fig. 11. The current version of Show-o exhibits challenges in text recognition/generation and object counting. For instance, Show-o struggles to identify the phrase “closing down” in the left of Fig. 11(a) and

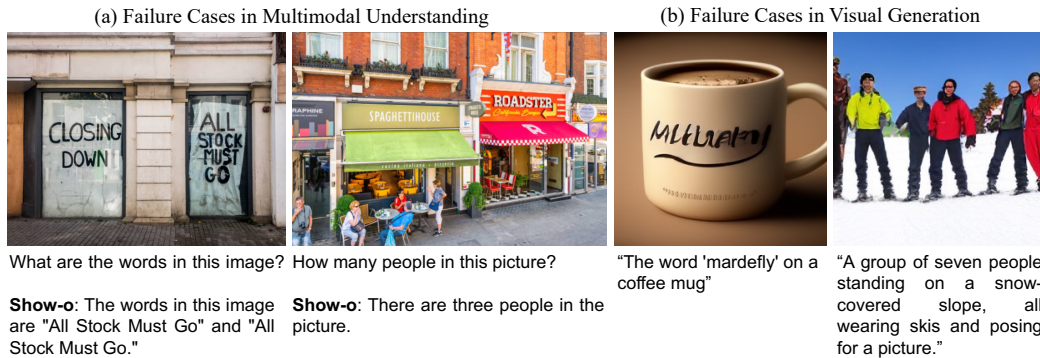


Figure 11: Illustration of failure cases of Show-o in multimodal understanding and generation.

is unable to generate the term “mardefly” (as shown left of Fig. 11(b)). This limitation is mainly attributed to the insufficiency of specific data tailored to these scenarios, as our model relies on a limited set of image-text pairs sourced from publicly available datasets and utilizes automatically generated captions. Enriching such kind of data holds promise for addressing these failure modes in Show-o, an aspect that will be explored in the future.

5 CONCLUSION

This paper proposed a unified transformer, *i.e.*, Show-o, to unify multimodal understanding and generation. Show-o for the first time unified auto-regressive and (discrete) diffusion modeling that can handle different modalities in distinct ways. Extensive experimental results demonstrated that Show-o is comparable even better than individual expert models across a wide range of vision-language tasks. This highlighted its potential as a next-generation foundation model.

6 ACKNOWLEDGMENTS

We would like to express our sincere gratitude to Henry Hengyuan Zhao for his valuable discussions and insightful feedback on multimodal understanding and Mingrui Wang for his patient assistance in helping us set up the development environment.

REFERENCES

- Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 1, 2023.
- Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *NeurIPS*, pp. 17981–17993, 2021.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *CoRR*, abs/2308.12966, 2023.
- Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Matusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020.

- Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *CVPR*, pp. 11315–11325, 2022.
- Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 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 *CVPR*, pp. 3558–3568, 2021.
- Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James T. Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart- α : Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*. OpenReview.net, 2024.
- Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.
- Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*, pp. 1691–1703, 2020.
- 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 Sankaranarayanan 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. *J. Mach. Learn. Res.*, 24:240:1–240:113, 2023.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.
- Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme Ruiz, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd van Steenkiste, Gamaleldin Fathy Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Collier, Alexey A. Gritsenko, Vignesh Birodkar, Cristina Nader Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetic, Dustin Tran, Thomas Kipf, Mario Lucic, Xiaohua Zhai, Daniel Keysers, Jeremiah J. Harmsen, and Neil Houlsby. Scaling vision transformers to 22 billion parameters. In *ICML*, pp. 7480–7512, 2023.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255, 2009.
- Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pp. 12873–12883, 2021a.
- Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pp. 12873–12883, 2021b.
- 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 *ICML*, 2024.

- Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024.
- Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-predict: Parallel decoding of conditional masked language models. In *EMNLP*, pp. 6111–6120, 2019.
- Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *NeurIPS*, 2023.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 2014.
- Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, pp. 10696–10706, 2022.
- Yuchao Gu, Xintao Wang, Yixiao Ge, Ying Shan, and Mike Zheng Shou. Rethinking the objectives of vector-quantized tokenizers for image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7631–7640, 2024.
- Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, pp. 6840–6851, 2020.
- Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *NeurIPS*, 2022.
- Emiel Hooeboom, Alexey A. Gritsenko, Jasmijn Bastings, Ben Poole, Rianne van den Berg, and Tim Salimans. Autoregressive diffusion models. In *ICLR*. OpenReview.net, 2022.
- Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *CVPR*, pp. 10124–10134. IEEE, 2023.
- Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pp. 4015–4026, 2023.
- Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vignesh Birodkar, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023.
- Chunyu Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, Jianfeng Gao, et al. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision*, 16(1-2):1–214, 2024.
- Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*, 2023.
- Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv preprint*, 2024a.
- Haotian Liu, Chunyu Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pp. 26296–26306, 2024b.
- Haotian Liu, Chunyu Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36, 2024c.

- Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruvi Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*, 2024.
- Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023. URL <http://probml.github.io/book2>.
- Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.
- Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *ICML*, pp. 4055–4064, 2018.
- William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pp. 4195–4205, 2023.
- Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refined-web dataset for falcon LLM: outperforming curated corpora with web data only. In *NeurIPS*, 2023.
- 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, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 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(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 *ICML*, pp. 8821–8831. Pmlr, 2021.
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125, 2022a.
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022b.
- Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. *NeurIPS*, 32, 2019.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 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. *NeurIPS*, 35:36479–36494, 2022.
- Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pp. 2256–2265, 2015.
- Tomáš Souček, Dima Damen, Michael Wray, Ivan Laptev, and Josef Sivic. Genhowto: Learning to generate actions and state transformations from instructional videos. In *CVPR*, pp. 6561–6571, 2024.

- Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, Jifeng Dai, Yu Qiao, Limin Wang, and Hongsheng Li. Journeydb: A benchmark for generative image understanding. In *NeurIPS*, 2023a.
- 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, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiyang Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. *CoRR*, abs/2312.13286, 2023b.
- Quan Sun, Qiyang Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *CoRR*, abs/2307.05222, 2023c.
- Quan Sun, Qiyang Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. In *ICLR*, 2023d.
- Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *NeurIPS*, 36, 2024.
- Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.
- Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017a.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017b.
- Alex Wang and Kyunghyun Cho. Bert has a mouth, and it must speak: Bert as a markov random field language model. *arXiv preprint arXiv:1902.04094*, 2019.
- Mitchell Wortsman, Peter J Liu, Lechao Xiao, Katie Everett, Alex Alemi, Ben Adlam, John D Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, et al. Small-scale proxies for large-scale transformer training instabilities. *arXiv preprint arXiv:2309.14322*, 2023.
- Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *ICCV*, 2023a.
- Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023b.
- Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *ICCV*, pp. 7452–7461, 2023.
- Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. Raphael: Text-to-image generation via large mixture of diffusion paths. *NeurIPS*, 36, 2024.
- Hanrong Ye, De-An Huang, Yao Lu, Zhiding Yu, Wei Ping, Andrew Tao, Jan Kautz, Song Han, Dan Xu, Pavlo Molchanov, et al. X-vila: Cross-modality alignment for large language model. *arXiv preprint arXiv:2405.19335*, 2024a.

- Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *CVPR*, pp. 13040–13051, 2024b.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.
- Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, et al. Language model beats diffusion-tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592, 2023a.
- Jinguo Zhu, Xiaohan Ding, Yixiao Ge, Yuying Ge, Sijie Zhao, Hengshuang Zhao, Xiaohua Wang, and Ying Shan. VL-GPT: A generative pre-trained transformer for vision and language understanding and generation. *CoRR*, abs/2312.09251, 2023b.