Muddit: Liberating Generation Beyond Text-to-Image with a Unified Discrete Diffusion Model

Qingyu Shi[1,2*], **Jinbin Bai**[2,3*†], **Zhuoran Zhao**[3], **Wenhao Chai**[4], **Kaidong Yu**[2], **Jianzong Wu**[1], **Shuangyong Song**[2], **Yunhai Tong**[1], **Xiangtai Li**[1‡], **Xuelong Li**[2‡], **Shuicheng Yan**[3‡]

[*]Equal Contribution, [†]Project Lead, [‡]Corresponding Authors

Abstract

Unified generation models aim to handle diverse tasks across modalities—such as text generation, image generation, and vision-language reasoning—within a single architecture and decoding paradigm. Autoregressive unified models suffer from slow inference due to sequential decoding, and non-autoregressive unified models suffer from weak generalization due to limited pretrained backbones. We introduce Muddit, a **u**nified **d**iscrete **di**ffusion **t**ransformer that enables fast and parallel generation across both text and image modalities. Unlike prior unified diffusion models trained from scratch, Muddit integrates strong **visual priors** from a pretrained text-to-image backbone with a lightweight text decoder, enabling flexible and high-quality multimodal generation under a unified architecture. Empirical results show that Muddit achieves competitive or superior performance compared to significantly larger autoregressive models in both quality and efficiency. The work highlights the potential of purely discrete diffusion, when equipped with strong visual priors, as a scalable and effective backbone for unified generation. The code and model are available at https://github.com/M-E-AGI-Lab/Muddit.

1 Introduction

Multimodal generative models capable of handling both text and images have rapidly advanced, typically relying on large autoregressive (AR) Transformers, also known as large language models (LLMs) [52]. These unified models represent both modalities as token sequences and generate outputs in a left-to-right autoregressive manner. However, this sequential decoding imposes a major inference bottleneck. For instance, in early unified transformers [46], as illustrated in Fig. 1(a), generating a single image requires sampling thousands of visual tokens one at a time. Despite strong correlation among adjacent image tokens, each token prediction triggers a full network forward, resulting in significant redundant computation. As a result, inference becomes **extremely slow and compute-intensive**. We refer to this as the first "dark cloud" over current unified generative models. Moreover, AR decoding enforces a rigid generation order. This prevents speed-quality trade-offs or flexible conditional generation like inpainting without fine-tuning, which severely limits practical applicability in interactive or real-time scenarios. To mitigate these limitations, some hybrid approaches [9, 11, 41], adopt AR language models paired with diffusion-based image synthesis heads (Fig. 1(b)). However, these "glue" architectures fall short of true unification, as they lack a shared generative modeling paradigm across modalities.

Recent work like Dual-Diffusion [29] (Fig. 1(c)) claims to unify modalities under discrete diffusion, but it ultimately relies on continuous diffusion for image generation via Stable Diffusion 3, a

[1]Peking University, [2]TeleAI, China Telecom, [3]National University of Singapore, [4]Princeton University
✉: jinbin.bai@u.nus.edu

Preprint. Under review.

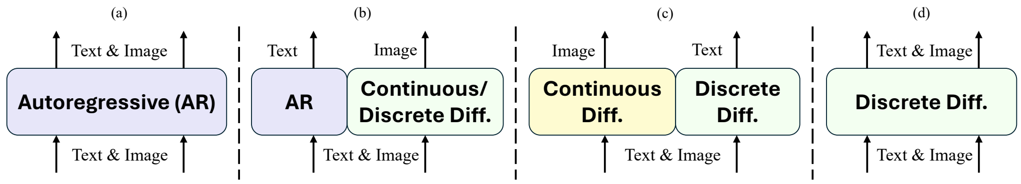


Figure 1: Four types of unified generative models. More details can be found in Sec. 2.

continuous diffusion paradigm. This fundamental mismatch in generative principles undermines its claim of true unification. UniDisc [48](Fig. 1(d)), takes a more promising step by applying discrete diffusion[1] over unified token spaces. This allows parallel refinement of text and image tokens, improving inference efficiency and enabling more flexible conditioning. However, the overall generation quality of UniDisc remains far from satisfactory. For example, it struggles to produce high-resolution 1024×1024 images, fails to match the fidelity of early diffusion models such as Stable Diffusion 1.5, and lacks support for vision-language reasoning tasks such as visual question answering (VQA). These limitations expose the second "dark cloud": **the absence of strong pre-trained discrete diffusion backbone models**: Unlike established unified autoregressive models that leverage powerful pretrained large language models, current unified discrete diffusion models are typically trained from scratch on mixed-modality tokens, which limits both their generative fidelity and transferability. Without modular components carrying rich pixel-level priors, these models face generalization and scalability bottlenecks.

Taken together, the two dark clouds: inefficient autoregressive sampling and the lack of strong pretrained foundations, highlight the need for a new generation of unified models. In this work, we present **Muddit**, a **M**askGIT-style **u**nified **d**iscrete **di**ffusion **t**ransformer equipped with a lightweight text decoder. By combining the strengths of parallel discrete diffusion and semantically rich image priors from a pre-trained Meissonic text-to-image backbone [5], Muddit enables scalable, efficient, and flexible sampling while significantly improving alignment and quality across modalities and various tasks such as high-resolution text-to-image synthesis, image-to-text synthesis, and visual question answering. We systematically detail the training objective of unified discrete diffusion models, the masking strategy, and the shared inference sampling strategy across three tasks. Finally, we conduct comprehensive evaluations with current popular unified models on several benchmarks, including GenEval, CIDEr, VQAv2, MME, and GQA, demonstrating Muddit's superior performance and efficiency, validating that the unexplored purely discrete diffusion approach can rival, or even surpass, much larger autoregressive-based unified models. While concurrent unified generation models [57] often build upon a language modeling prior—leveraging pretrained dLLMs as the backbone—we instead take a visual-first approach. Muddit is built upon an image generation prior, offering a new path toward unifying vision and language tasks within a discrete diffusion framework. We hope that this work inspires a new trend for unified generative modeling, grounded in discrete diffusion, beyond the boundaries of traditional text-to-image synthesis [5] and text synthesis [25, 39].

2 Related Work

2.1 Unified Models For Generation and Understanding

The success of LLMs in language modeling has inspired efforts to extend unified generation to multimodal domains. However, the divergence between autoregressive and diffusion-based paradigms presents fundamental architectural trade-offs. Autoregressive models naturally handle language, and several works [11, 16, 20, 47, 51, 53] extend this by connecting vision modules to LLMs via adapters or instruction tuning, with LLMs serving as planning modules that produce intermediate representations for image generation. While effective to some extent, these paradigms often exhibit limited interaction between text and image modalities and struggle with content consistency, particularly in image-to-image generation and complex instruction-based synthesis. To address these limitations,

[1]MaskGIT, MaskAR, RandomAR, and Discrete Diffusion share significant conceptual and practical overlaps, often differing only in decoding order or architectural nuances. We elaborate on their connections in the next section. While Meissonic [5] follows the naming convention of MaskGIT [8], we standardize terminology in this paper by referring to all such models under the umbrella of Discrete Diffusion.

recent research explores unified generation models that integrate understanding and generation within a single architecture, We categorize these into four major paradigms (see Fig. 1):

Fully Autoregressive: Both text and image are tokenized into discrete sequences and modeled with an AR Transformer [13, 22, 32, 34, 50, 54, 55, 59]. These models achieve strong cross-modal generation but suffer from high latency due to sequential decoding.

Text AR, Image Diffusion: LLMs generate text tokens while image synthesis is delegated to pretrained continuous diffusion backbones [38, 58, 60] or discrete diffusion [56]. Though visually strong, these models are not truly unified, as they rely on separate architectures and token spaces.

Image Diffusion, Text Discrete Diffusion: Emerging models experiment with discrete diffusion for text and images [29], though many, like Dual-Diffusion, still use continuous diffusion for image synthesis, failing to realize true modality symmetry.

Fully Discrete Diffusion: Recent work like UniDisc [48] pioneers full-token discrete diffusion over shared Transformer backbones. These models support parallel sampling and native integration, but currently lag behind in generation fidelity and scale.

Among these, the GPT-4o [40] model represents a significant advance as a unified multimodal generative system. However, its closed-source nature obscures critical architectural and training details, and its success may be largely attributable to scale rather than architectural novelty [12].

2.2 Masked Image Modeling

Masked Image Modeling (MIM) has emerged as a powerful self-supervised learning paradigm in computer vision, drawing inspiration from the success of Masked Language Modeling (MLM) in NLP, notably BERT [15]. The fundamental principle of MIM involves obscuring portions of an image, which could be raw pixels (MAE [23]), latent patches of pixels, or even discrete latent tokens (BEiT [6], MaskGIT [8]), and training a model, typically an autoencoder, to predict or reconstruct this missing information by leveraging the context provided by the visible parts.

MaskGIT [8] introduced parallel decoding via iterative token refinement, inspiring discrete diffusion models. Recent work such as RandomAR [18] and MAR [28] formalize this as random-order or masked autoregressive generation, blending AR and MIM principles. The major conceptual difference between RandomAR/MAR and MaskGIT is in the scanning order at inference time.

This class of techniques forms the conceptual foundation of discrete diffusion over tokenized spaces and plays a critical role in modern unified models. We will introduce discrete diffusion in the next section.

3 Method

3.1 Discrete Diffusion with Unified Image and Text Perspective

In discrete diffusion, a sample $x \in \mathcal{X}$ is treated as a one-hot vector $\mathbf{x}$, where $\mathcal{X} = \{1, \ldots, N\}$. For language models, N equals the vocabulary size. While for image models, N is the number of discrete image–token IDs obtained from a tokenizer or VQ-codebook. At each diffusion step, we stochastically corrupt the tokens, gradually transforming the data distribution into a maximally entropic categorical prior; the generative model then learns to invert this corruption. Following recent works [5, 36] that cast token corruption as a continuous–time Markov chain (CTMC) over the finite alphabet $\mathcal{X}$, we let

$$\frac{d\,p_t}{dt} = Q_t\,p_t, \tag{1}$$

where $p_t \in \mathbb{R}^{N+1}$ is the distribution of x_t, the time–dependent matrix Q_t transports the data distribution $p_0 \approx p_{\mathrm{data}}$ to the maximally entropic "noise" distribution $p_1 = p_{\mathrm{stationary}}$. We adopt the absorbing-state (masked) diffusion variant that has proved particularly effective in text modelling: every symbol can jump to a dedicated mask token $\mathtt{m} = (\underbrace{0, \ldots, 0}_{N}, 1)$ but never leaves it, i.e. $\mathtt{m}$ is an absorbing class.

Forward posterior. Marginalising $\mathbf{x}$ gives

$$q(x_t \mid \mathbf{x}) = \mathrm{Cat}\big(x_t \mid \alpha_t \mathbf{x} + (1-\alpha_t)\mathtt{m}\big). \tag{2}$$

$\mathrm{Cat}(\cdot)$ denotes a categorical distribution; it returns a one-hot token sampled from the probability vector inside the parentheses. $\alpha_t \in [0,1]$ is the *survival probability*, *i.e.* the probability that an individual token has not yet been masked by time t. Thus x_t equals the original clean token with probability α_t and equals the mask token $\mathtt{m}$ with probability $1-\alpha_t$.

Reverse process. For any $0 < s < t < 1$, the CTMC induces an analytic posterior

$$q(x_s \mid x_t, \mathbf{x}) = \begin{cases} \mathrm{Cat}(x_s \mid x_t), & x_t \neq \mathtt{m}, \\ \mathrm{Cat}\Big(x_s \mid \dfrac{(1-\alpha_s)\mathtt{m} + (\alpha_s - \alpha_t)\mathbf{x}}{1-\alpha_t}\Big), & x_t = \mathtt{m}, \end{cases} \tag{3}$$

x_t and x_s are the corrupted tokens at times t and s ($s < t$). If x_t is already a real vocabulary token ($x_t \neq \mathtt{m}$) it stays unchanged going backwards; otherwise, when $x_t = \mathtt{m}$, the distribution over x_s is a convex combination of the mask and the clean token $\mathbf{x}$, weighted by their respective survival probabilities α_s and α_t.

Training Objective. We employ a masked-token predictor $x_\theta(x_t, \alpha_t) \approx \mathbf{x}$, which leads to the continuous-time negative ELBO

$$\mathcal{L}_{\mathrm{NELBO}} = \mathbb{E}_{q(x_t|\mathbf{x})}\left[\int_0^1 \frac{\alpha'_t}{1-\alpha_t} \log\big(x_\theta(x_t, \alpha_t) \cdot \mathbf{x}\big)\, dt\right], \tag{4}$$

where $\alpha'_t = \frac{d\alpha_t}{dt}$ and $\mathbf{x}$ is the one-hot vector of ground truth. $x_\theta(x_t, \alpha_t) \in \mathbb{R}^{N+1}$ is the model's predicted categorical probability vector for the clean token given the corrupted input (x_t, α_t); $\mathbf{x}$ is the one-hot ground-truth clean token.

During generation, we start from an all-mask sequence ($t = 1$) and integrate the reverse CTMC towards $t = 0$, repeatedly replacing every masked position with the model's categorical prediction. Because the corruption schedule and objective are *identical* for any discrete alphabet $\mathcal{X}$, the same diffusion backbone unifies text and image generation. In the following section, we present Muddit, a unified framework that leverages discrete diffusion to model the generation tasks for both text and image jointly.

3.2 Muddit

3.2.1 Unified Architecture

As shown in Fig. 2, our architecture comprises a text encoder $\mathtt{E}_{\mathtt{txt}}$, image encoder $\mathtt{E}_{\mathtt{img}}$, transformer generator $\mathtt{G}$, sampler $\mathtt{S}$, text decoder $\mathtt{D}_{\mathtt{txt}}$, and image decoder $\mathtt{D}_{\mathtt{img}}$. The generator $\mathtt{G}$ is a single MM-DiT model, following the dual-/single-stream design of FLUX [26]. Importantly, the generator $\mathtt{G}$ is initialized from the Meissonic [5], which has been extensively trained for high-resolution text-to-image generation. This initialization brings in a strong pretrained image prior, capturing rich spatial structures and semantic correlations across image and text tokens, which significantly enhances sample quality and accelerates convergence in the multimodal setting. Consequently, the same MM-DiT predicts the masked tokens for both modalities, which produces a shared generator for text and image synthesis.

To reduce the computational cost of high-resolution imagery and lengthy captions, we quantize both modalities into a compact discrete space. A pre-trained VQ-VAE acts as the image encoder $\mathtt{E}_{\mathrm{img}}$, mapping pixels to codebook indices, while the CLIP text model, as $\mathtt{E}_{\mathrm{txt}}$, provides the text token embeddings. The MM-DiT predicts clean tokens in this shared space, which a lightweight linear head $\mathtt{D}_{\mathrm{txt}}$ converts back to text tokens.

3.2.2 Unified Training

Masking Strategy. We model the forward posterior in Eq. 2 of both modalities using time-dependent hyperparameters α_t, with the mask ratio defined as $\gamma_t = 1 - \alpha_t$. While BERT [15] employs a fixed mask ratio of 15%, this setting is suitable for token completion but insufficient for generation. To support generative tasks, the design of γ_t must satisfy the following criteria:

(i.e, they do not mask both text and image at the same time!)

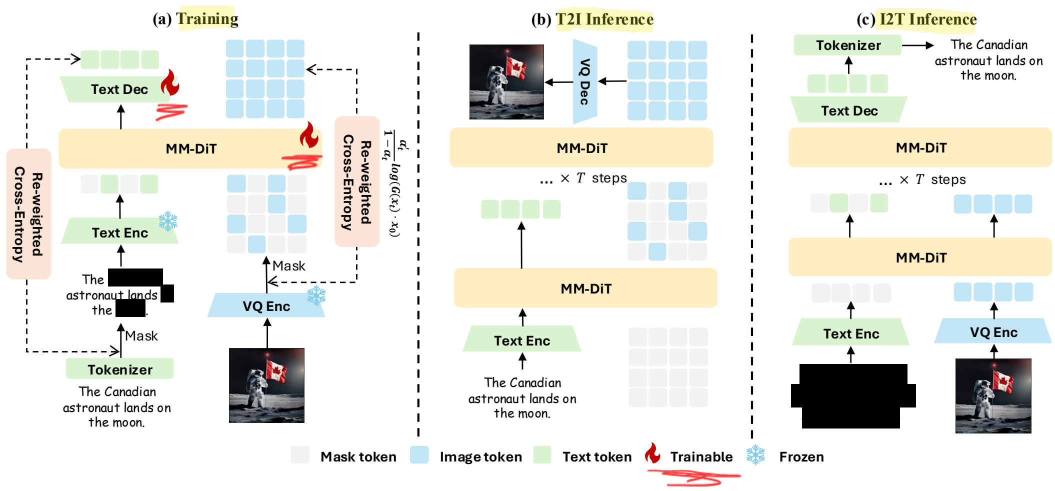


Figure 2: The training and inference architecture of Muddit. **(a)** During training, we randomly mask tokens from one of the two modalities. MM-DiT is trained to predict the masked tokens using a re-weighted cross-entropy loss, which jointly optimizes both the MM-DiT backbone and a lightweight text decoder. **(b)** In text-to-image inference, we initialize the image latent features using all-masked tokens and iteratively predict each latent token via MM-DiT. **(c)** In image-to-text inference, we similarly initialize all text tokens as masked and generate the text through the same iterative decoding process. Specifically for VQA tasks, we append mask token IDs to the end of the question and predict all masked token IDs as the final answer.

1. γ_t must be a continuous function, bounded between 0 and 1, for $t \in [0, 1]$.

2. γ_t should monotonically in decrease with respect to t, with boundary conditions $\gamma_0 \to 0$ (initially clean data) and $\gamma_1 \to 1$ (masking all tokens).

Several strategies for masking and sampling have been proposed to meet these criteria [8]. We adopt *cosine scheduling strategy*. During training, a timestep $t \in [0, 1]$ is sampled from a truncated arccos distribution, with the density function:

or rather, you sample t, and compute γ_t

$$\gamma_t = \frac{2}{\pi}(1 - (1-t)^2)^{-\frac{1}{2}}. \tag{5}$$

During training, a mask ratio $\gamma_t \in [0, 1)$ is randomly sampled for each modality x_0 (either image or text tokens), and the forward process (Eq. 2) is applied by randomly replacing clean tokens with mask tokens to obtain x_t.

Unified Training Objective. Let $\mathbf{c}$ denote the conditioning: the text embedding when synthesizing an image, or the image embedding when generating a caption. We randomly sample a mask ratio by Eq. 5. Then we corrupt the target sequence $\mathbf{x_0}$ (image or text tokens) with the CTMC described in Eq. 1 and train a single masked-token predictor $\mathtt{G}(x_t, \alpha_t, \mathbf{c})$ to reconstruct $\mathbf{x_0}$. Both directions—text $\to$ image and image$\to$ text—share the identical continuous-time negative ELBO

$$\mathcal{L}_{\text{unified}} = \mathbb{E}_{q(x_t|\mathbf{x})}\left[\int_0^1 \frac{\alpha'_t}{1-\alpha_t} \log\big(\mathtt{G}(x_t, \alpha_t, \mathbf{c})\cdot\mathbf{x}\big)\, dt\right], \tag{6}$$

where all symbols are as in Eq. 4 but the $\mathtt{G}$ now receives the cross-modal condition $\mathbf{c}$ as an additional input. **Key point:** switching from text$\to$ image to image$\to$ text merely changes the conditioning signal $\mathbf{c}$; the loss Eq. 6 itself is unchanged. This symmetry keeps optimization identical across tasks and allows us to train a single parameter set jointly for both generation directions. During inference we again start from an all-mask sequence (t=1) and integrate the reverse CTMC towards t=0, feeding in the desired condition $\mathbf{c}$ to obtain either an image or a sentence from the same diffusion backbone.

3.2.3 Unified Inference

Sampling Strategy. During inference, we apply the time-reversed posterior as defined in Eq. 3.

$$\mathtt{S}(\mathtt{G}, x_t, t) = p_\theta(x_s \mid x_t) = \begin{cases} \mathrm{Cat}(x_s \mid x_t), & x_t \neq \mathtt{m}, \\ \mathrm{Cat}\Big(x_s \mid \dfrac{(1-\alpha_s)\mathtt{m} + (\alpha_s - \alpha_t)\mathtt{G}(x_t, \alpha_t, \mathtt{c})}{1-\alpha_t}\Big), & x_t = \mathtt{m}, \end{cases} \tag{7}$$

where θ denotes the parameters of $\mathtt{G}$, $\mathtt{c}$ is the multimodal condition, and α_t in Eq. 5 is applied sequentially with t taking values $1, \frac{T-1}{T}, \ldots, \frac{1}{T}$, where T is the total number of reverse steps. At each timestep t, Muddit predicts a fraction $\gamma_{t+\frac{1}{T}} - \gamma_t$ of the masked tokens by $\mathtt{G}$ and update the masked tokens $\mathtt{x_t}$ by $\mathtt{S}$, continuing iteratively until all masked tokens are recovered. This dynamic approach offers several advantages over autoregressive methods, which require the model to learn conditional probabilities $P(x_i \mid x_{<i})$ based on a fixed token ordering. In contrast, random masking with a variable ratio enables the model to learn $P(x_i \mid x_\Lambda)$, where Λ denotes an arbitrary subset of observed tokens. This flexibility is essential for parallel sampling, allowing multiple tokens to be predicted simultaneously rather than sequentially.

Our Muddit supports three tasks with a single generator $\mathtt{G}$ and sampler $\mathtt{S}$: (i) text $\rightarrow$ image, (ii) image $\rightarrow$ text (captioning), and (iii) visual–question answering (VQA). The only change across tasks is the conditioning source $\mathbf{c}$ provided to $\mathtt{G}$; the diffusion process and guidance logic are shared.

(i) Text $\rightarrow$ image. Given a text prompt $\mathtt{tp} \in \mathcal{T}$, the text encoder $\mathtt{E_{txt}}$ produces a text token embedding $\mathtt{c_{txt}} = \mathtt{E_{txt}}(\mathtt{tp})$. Starting from a fully masked sequence x_1, the generator produces logits

$$l_t = \mathtt{G}(x_t, \alpha_t, \mathtt{c_{txt}}), \qquad x_{t-\frac{1}{T}} = \mathtt{S}(l_t, x_t, t), \tag{8}$$

for $k = 1, \frac{T-1}{T}, \ldots \frac{1}{T}$. After T steps we obtain visual tokens x_0, which the image decoder $\mathtt{D_{img}}$ converts to a pixel-space image $I = \mathtt{D_{img}}(x_0)$.

(ii) Image $\rightarrow$ text. For captioning, an input image $I \in \mathcal{I}$ is tokenized by the image encoder $\mathtt{E_{img}}$: $\mathtt{c_{img}} = \mathtt{E_{img}}(I)$. The generator now conditions on the *visual* tokens while progressively decoding text:

$$l_t = \mathtt{G}(x_t, \alpha_t, \mathtt{c_{img}}), \qquad t_{t-\frac{1}{T}} = \mathtt{S}(l_t, x_t, t), \tag{9}$$

yielding a text token sequence x_0, which $\mathtt{D_{txt}}$ maps to a caption $\mathtt{caption} = \mathtt{Detokenize}(\mathtt{D_{txt}}(x_0))$.

(iii) Image + question $\rightarrow$ answer (VQA). For visual–question answering we supply *both* an image and a question: $\mathtt{c_{img}} = \mathtt{E_{img}}(I)$ and $\mathtt{c_{txt}} = \mathtt{E_{txt}}(q)$. They are concatenated and fed to the generator, which outputs logits over answer tokens x_k:

$$l_t = \mathtt{G}(x_t, \alpha_t, [\mathtt{c_{img}}, \mathtt{c_{txt}}]), \qquad x_{t-\frac{1}{T}} = \mathtt{S}(l_t, x_t, t), \tag{10}$$

until the full answer a is produced and decoded by $a = \mathtt{Detokenize}(\mathtt{D_{txt}}(x_0))$.

Classifier-free guidance. At each decoding step, we apply the same guidance rule, independent of modality:

$$l_k \;\leftarrow\; \mathtt{G}(z_k, \alpha_k, \mathbf{c}) \;+\; \lambda\big[\mathtt{G}(z_k, \alpha_k, \mathbf{c}) - \mathtt{G}(z_k, \alpha_k, \mathbf{c}_{\mathrm{neg}})\big], \tag{11}$$

where z_k (image or text tokens) is the partial target sequence, $\mathbf{c}$ is the *positive* condition (prompt, image, or image +question), $\mathbf{c}_{\mathrm{neg}}$ is the corresponding negative condition, and λ is the guidance scale. Because the loss, decoding schedule, and guidance operator are *identical* in all three scenarios—only the conditioning signal changes—our framework realises a genuinely unified multimodal generator.

4 Experiment

4.1 Experimental Setup

Implementation details. We build Muddit on top of the open-sourced Meissonic models [5]. The MM-DiT backbone is initialized with pretrained weights, and a lightweight linear head is added as a text decoder. Following Meissonic, we adopt the CLIP tokenizer and encoder [43], as well as the VQ-VAE, keeping them entirely frozen throughout all experiments, including the mask token

Model	Text Gen Arch	Image Gen Arch	Params (B)	Overall ↑	Objects ↑		Counting ↑	Colors ↑	Position ↑	Color ↑ Attribution
					Single	Two				
PixArt-α [10]	-	Diffusion	0.6	0.48	0.98	0.50	0.44	0.80	0.08	0.07
SD 2.1 [45]	-	Diffusion	0.9	0.50	0.98	0.51	0.44	0.85	0.07	0.17
DALL-E 2 [44]	-	Diffusion	6.5	0.52	0.94	0.66	0.49	0.77	0.10	0.19
SDXL [42]	-	Diffusion	2.6	0.55	0.98	0.74	0.39	0.85	0.15	0.23
DALL-E 3 [7]	-	Diffusion	-	0.67	0.96	0.87	0.47	0.83	0.43	0.45
SD 3 [17]	-	Diffusion	2	0.62	0.98	0.74	0.63	0.67	0.34	0.36
LWM [35]	AR	AR	7	0.47	0.93	0.41	0.46	0.79	0.09	0.15
SEED-X [20]	AR	AR	17	0.49	0.97	0.58	0.26	0.80	0.19	0.14
Chameleon [50]	AR	AR	7	0.39	-	-	-	-	-	-
Show-O [56]	AR	Discrete Diff.	1.3	0.68	0.98	0.80	0.66	0.84	0.31	0.50
Transfusion [60]	AR	Diffusion	8	0.67	-	-	-	-	-	-
D-DiT [30]	Discrete Diff.	Diffusion	2	0.65	0.97	0.80	0.54	0.76	0.32	0.50
Monetico (512×512) [5]	-	Discrete Diff.	1	0.44	0.92	0.48	0.26	0.78	0.06	0.13
Meissonic (1024×1024) [5]	-	Discrete Diff.	1	0.54	0.99	0.66	0.42	0.86	0.10	0.22
UniDisc (512×512) [49]	Discrete Diff.	Discrete Diff.	1.4	0.42	0.92	0.47	0.15	0.67	0.13	0.19
Muddit (512×512)	Discrete Diff.	Discrete Diff.	1	0.61	0.98	0.72	0.54	0.82	0.19	0.41

Table 1: Evaluation of text-to-image generation performance on the GenEval [21].

embedding in CLIP. To support discrete denoising, we append a special <mask> token to CLIP's vocabulary for text masking, while the image mask token is inherited directly from Meissonic's initialization. During training, we use a constant learning rate of 1×10^{-4} and a weight decay of 1×10^{-2}. Gradient accumulation is applied in both pretraining and supervised fine-tuning, resulting in an effective batch size of 1024. During inference, we adopt the default Meissonic configuration, using cosine masking scheduling, 64 sampling steps, and a classifier-free guidance (CFG) scale of 9.0 for both text-to-image and image-to-text generation.

Training Data. We train Muddit in two stages using a mix of publicly available and internal datasets, comprising approximately 3.5 million image–text pairs. Both stages are optimized using the unified training objective defined in Eq. 6. Below, we detail the datasets and settings for each stage:

1. **Pretraining.** We pretrain Muddit for 70K steps with a batch size of 1024, using the unified objective across both modalities. Text inputs are truncated to a maximum of 77 tokens, and images are resized to 512×512. The pretraining corpus consists of 2 million image–text pairs, re-captioned using Qwen2.5-VL-3B for improved consistency. Each batch is evenly split between text-to-image and image-to-text samples to enable joint training in both directions.
2. **Supervised Fine-tuning.** After pretraining, we fine-tune the model on a combination of instruction-following datasets, including LLaVA-Instruct-150K and the MG-LLaVA tuning set. During this stage, only the answer portion of each prompt is masked. Additionally, we construct a curated dataset of 500K high-quality image–text pairs to support multi-task training on VQA and image generation. Following the task instructions embedded in each sample, Muddit learns to produce long-form answers, concise replies, and image captions via task-specific prompting.

We present both quantitative and qualitative results for T2I and I2T tasks in the following sections. Additional experiments and ablation studies are provided in the Appendix.

4.2 Text-to-Image Generation

Quantitive Results. Following prior work, we evaluate our 512×512 model on the GenEval [21] benchmark after supervised fine-tuning, measuring its ability to generate images aligned with textual prompts. As shown in Tab. 2, Muddit achieves a strong overall accuracy of 0.61, outperforming previous discrete diffusion models such as Monetico (0.44) and Meissonic (0.54), and approaching the performance of Stable Diffusion 3 (0.62), despite using only 1B parameters. Muddit exhibits strong compositional reasoning, scoring 0.72 on the "Two Objects" subset and 0.54 on "Counting". Notably, we observe that joint training across modalities significantly enhances the text-to-image generation capabilities of the Meissonic backbone. These results highlight the potential of Muddit as the first unified model to adopt a pure discrete diffusion framework for both text and image modalities, achieving competitive quality with a compact, scalable architecture.

Qualitative Results. We present diverse generations from our model conditioned on rich textual prompts in Fig. 3. The outputs exhibit strong text-image alignment, capturing fine details in both

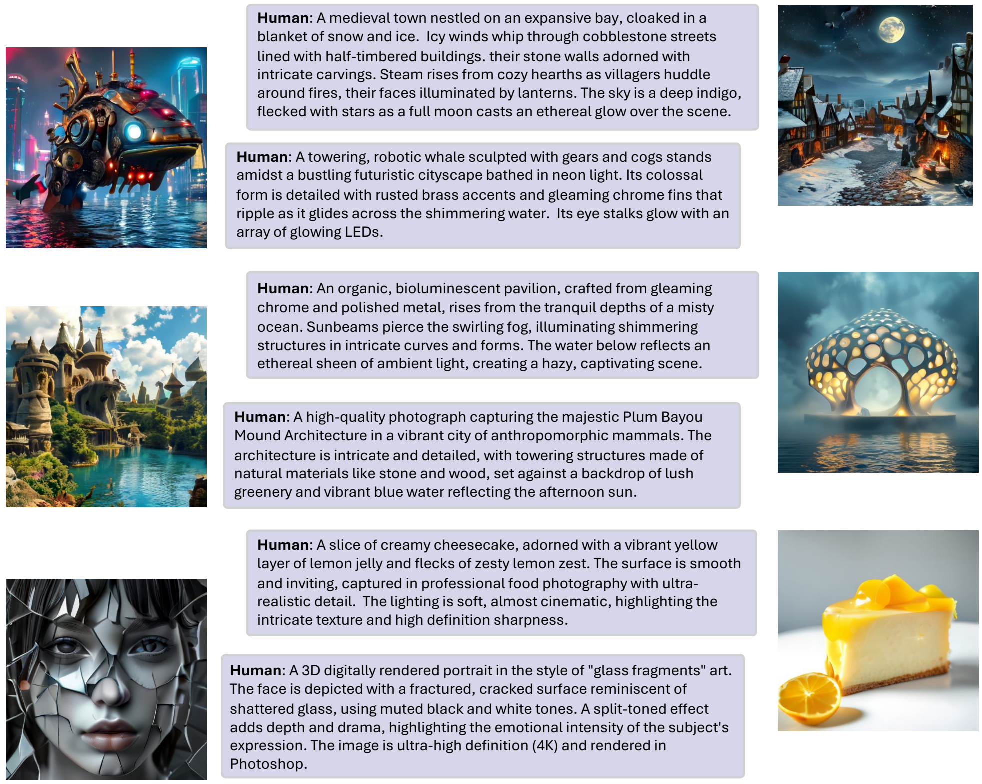


Figure 3: Samples of Text-to-Image Generation by Muddit.

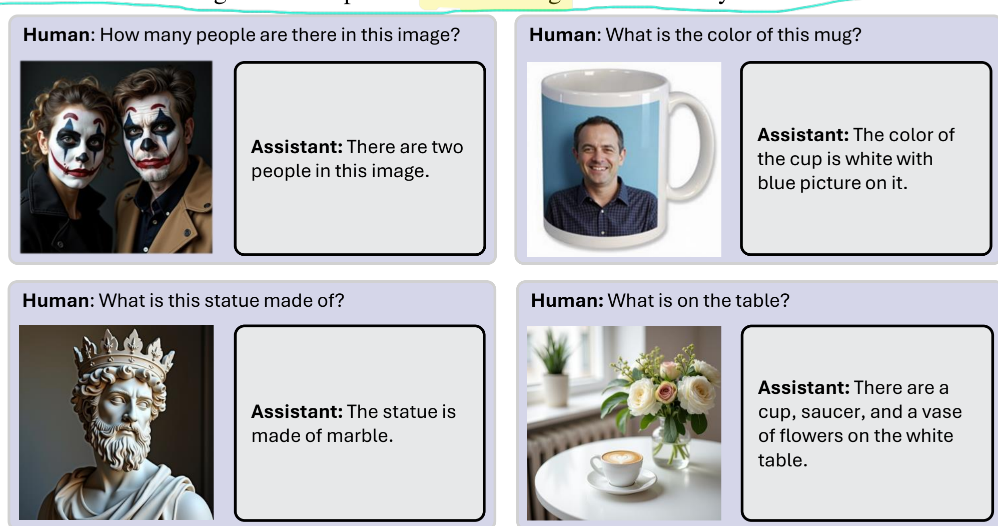


Figure 4: Samples of Visual Question Answering by Muddit.

realistic and imaginative scenes. Our model effectively renders complex structures, lighting, and textures across various domains.

4.3 Image-to-Text Generation

We present a comprehensive comparison of our model Muddit against other multimodal models across four benchmarks: MS-COCO (image captioning) [31], VQAv2 [2], MME [19], and GQA [24] in Tab. 4.3. Notably, Muddit is the first unified model to employ discrete diffusion for both image and text generation, demonstrating that this approach is not only viable but also highly competitive.

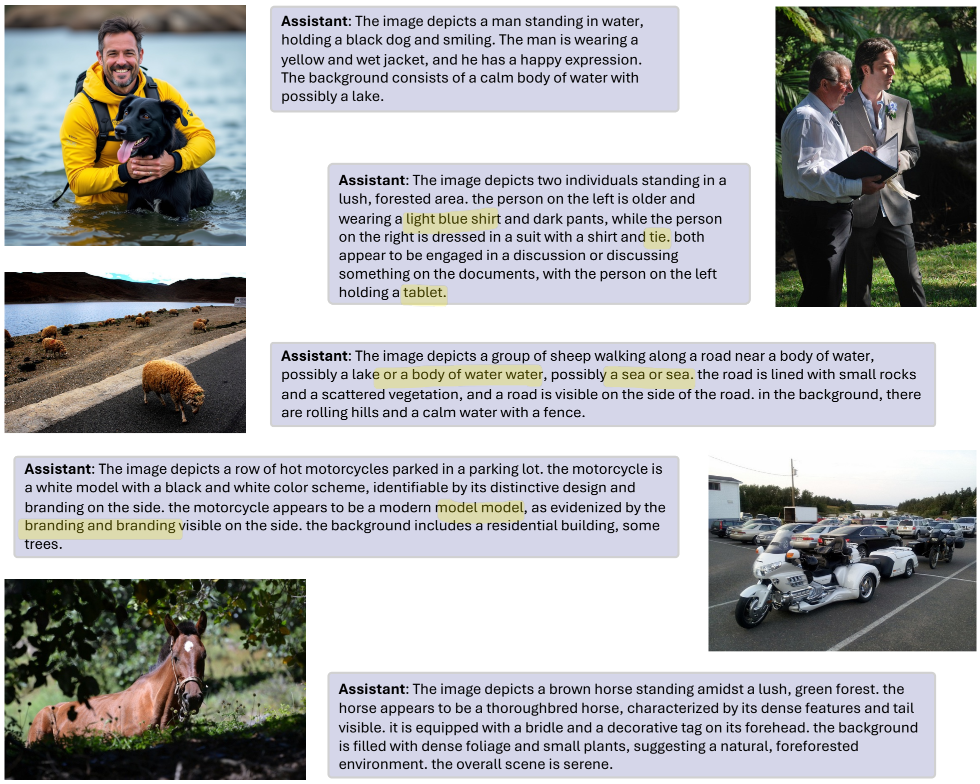


Figure 5: Samples of Image-to-Text Generation by Muddit.

Model	**Params** (B)	**Text Gen** Arch	**Image Gen** Arch	**MS-COCO** CIDEr ↑	**VQAv2** Acc. ↑	**MME** Acc. ↑	**GQA** Acc. ↑
InternVL-2.0 [14]	8	AR	-	-	-	1648.1	61.0
LLaVA-Next [33]	13	AR	-	-	82.8	1575.0	65.4
BLIP-2 [27]	13	AR	-	-	65.0	1293.8	41.0
QWEN-VL [4]	7	AR	-	-	78.2	1487.5	57.5
OpenFlamingo [3]	9	AR	-	65.5	43.5	-	-
Flamingo [1]	9	AR	-	79.4	51.8	-	-
Chameleon [50]	7	AR	AR	18.0	-	-	-
LWM [34]	7	AR	AR	-	55.8	-	44.8
Show-O (256×256) [56]	1.3	AR	Discrete Diff.	-	64.7	1014.9	54.2
Show-O (512×512) [56]	1.3	AR	Discrete Diff.	-	69.4	1097.2	58.0
Transfusion [60]	7	AR	Diffusion	29.0	-	-	-
D-DiT (256×256) [29]	2	Discrete Diff.	Diffusion	-	59.5	897.5	55.1
D-DiT (512×512) [29]	2	Discrete Diff.	Diffusion	56.2	60.1	1124.7	59.2
UniDisc [49]	0.33	Discrete Diff.	Discrete Diff.	46.8	-	-	-
Muddit (512×512)	1	Discrete Diff.	Discrete Diff.	59.7	67.7	1104.6	57.1

Table 2: Evaluation of image-to-text generation and visual question answering.

Quantitive Comparison. Despite having only 1B parameters—substantially fewer than most competing models—Muddit achieves strong performance across both image captioning and visual question answering tasks. It obtains a CIDEr score of 59.7 on MS-COCO, surpassing larger models such as Show-O and D-DiT. On the VQAv2 benchmark, it reaches 67.7%, outperforming other diffusion-based models like D-DiT (512×512) and approaching the performance of much larger autoregressive models such as LLaVA-Next (13B). Muddit also demonstrates competitive results on MME and GQA (1104.6 and 57.1 accuracy, respectively), highlighting its capability as a unified model without compromising task-specific quality.

Qualitative Results. We present example captions generated by our model across diverse scenarios in Fig. 5, including humans, animals, vehicles, and natural landscapes. The model demonstrates strong visual grounding and fine-grained descriptive ability, accurately capturing attributes such as

Metric	0.2	0.4	0.6	0.8
GenEval ↑	60.1	60.5	**61.6**	60.8
MS-COCO CIDEr ↑	50.2	51.2	**58.4**	58.3
VQAv2 ↑	62.1	65.8	67.8	**67.9**

Table 3: Impact of text loss weight.

Metric	w/o joint training	w/ joint training
GenEval ↑	28.3	**61.6**
MS-COCO CIDEr ↑	**59.4**	58.4
VQAv2 ↑	**69.2**	67.8

Table 4: Effect of joint training.

Timestep	T=8	T=16	T=24	T=32	T=40	T=50	T=64
GenEval ↑	51.4	58.1	59.2	**61.6**	61.5	61.4	60.8
MS-COCO CIDEr ↑	43.4	58.5	58.6	58.4	59.2	**60.0**	59.7
VQAv2 ↑	68.3	68.4	68.5	67.8	67.5	67.6	**67.7**

Table 5: Performance across different diffusion timesteps.

clothing, expressions, background context, and object relationships. Fig. 4 illustrates our model's ability to accurately answer visual questions across various domains, including object counting, color recognition, material identification, and compositional reasoning.

4.4 Ablation Study and Analysis

Analysis of the inference timesteps. As shown in Tab. 5, increasing the number of diffusion steps generally improves performance, with most metrics plateauing around $T = 32$–50. In particular, GenEval and CIDEr scores improve substantially from $T = 8$ to $T = 32$, though the marginal gains diminish thereafter. VQAv2 remains largely stable across timesteps, suggesting that fewer steps suffice for discriminative tasks. Overall, a moderate number of steps offers a favorable trade-off between accuracy and efficiency.

Analysis of the text loss weight. As shown in Tab. 3, moderate text loss weights (approximately 0.6) yield the best overall performance. Both CIDEr and GenEval scores peak near this value, indicating that placing either too little or too much emphasis on text can impair generation quality. Notably, VQAv2 performance continues to improve with increased text supervision, but begins to plateau beyond 0.6. These observations suggest that while stronger textual guidance benefits discriminative tasks, generative tasks require a balanced integration of visual and textual signals—underscoring the notion that effective multimodal models must not only learn language, but also learn to ground it.

Analysis of joint training. "With joint training" denotes the use of cross-entropy loss on both image token prediction and text token prediction, whereas "without joint training" refers to applying the loss only on text token prediction. As shown in Tab. 4, removing joint training results in a dramatic drop in GenEval performance—from 61.6 to 28.3—highlighting a more than two-fold decrease that exceeds any other variation. Meanwhile, CIDEr remains nearly unchanged (59.4 → 58.38), suggesting that language quality is preserved, and VQAv2 declines only marginally (69.2 → 67.8), representing a minimal cost relative to the gains in cross-modal alignment. This ablation underscores that decoupling the training objectives significantly impairs the model's ability to integrate vision and language, reinforcing the necessity of unified optimization for multimodal coherence.

4.5 Inference Time Analysis

As shown in Fig. 6, autoregressive multimodal models are inherently limited by token-by-token decoding, which constrains their inference speed. Muddit overcomes this bottleneck with a parallel discrete diffusion decoder, reducing average latency to just 1.49 seconds, achieving a 4× to 11× speedup over competitive baselines (4.2× faster than Qwen-2.5-VL, 5.6× than Show-o, 8.1× than BLIP-2, and 10.9× than LLaVA-1.6).

Besides, we present detailed FLOPs comparison between Autoregressive and Discrete Diffusion.

Autoregressive (AR) without KV Cache:

- At step t, the model attends over t previous tokens.
- Per-step attention FLOPs: $O(t^2 D)$.

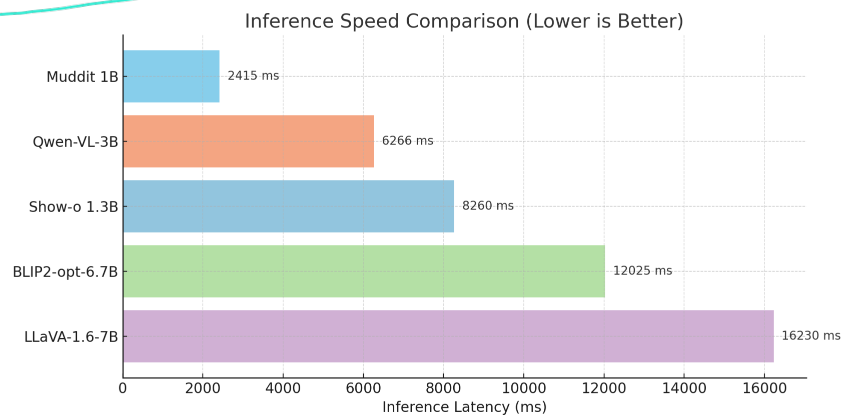


Figure 6: Inference speed comparison. We use 32 inference steps for Muddit and fix the sequence length to 77 across all models.

- Total FLOPs:

$$\sum_{t=1}^{L} O(t^2 D) = O\left(D \sum_{t=1}^{L} t^2\right) = O\left(D \cdot \frac{L(L+1)(2L+1)}{6}\right) = O(L^3 D)$$

Autoregressive (AR) with KV Cache:

- At step t, Q is computed for 1 token, and attends to t K/V keys.
- Per-step attention FLOPs: $O(tD)$.
- Total FLOPs:

$$\sum_{t=1}^{L} O(tD) = O\left(D \sum_{t=1}^{L} t\right) = O\left(D \cdot \frac{L(L+1)}{2}\right) = O(L^2 D)$$

Discrete Diffusion:

- Each step updates the full sequence (length L) in parallel.
- Per-step attention FLOPs: $O(L^2 D)$.
- Total FLOPs:

$$T \cdot O(L^2 D) = O(TL^2 D), \quad T \ll L$$

While discrete diffusion may appear less efficient than autoregressive (AR) models with KV caching in terms of theoretical FLOPs, it offers a significant advantage over AR without caching—achieving an L/T speedup by updating the full token sequence in parallel over T iterations. In practice, the higher degree of parallelism leads to competitive, and often faster, inference speed compared to AR models, especially when considering real-world GPU throughput. As KV cache techniques for discrete diffusion are rapidly evolving [37], we expect further acceleration in the near future, narrowing the theoretical speed gap even with KV-cache AR baselines.

4.6 Generated Results Step by Step

Muddit frames text generation as reverse discrete diffusion over a fixed-length sequence of 77 token indices. At inference time, the model performs $16 \leq T \leq 32$ denoising steps, starting from a maximally entropic prior where every token is masked. At each step t, a parameter-shared transformer $\mathtt{G}$ predicts a categorical distribution over all positions in parallel, and a sampler $\mathtt{S}$ selects the next sequence:

$$\mathbf{x}_{t-1} = \mathtt{S}(\mathtt{G}(\mathbf{x}_t, \mathtt{c}, t), \mathbf{x}_t, t), \qquad t = T, \ldots, 1, \tag{12}$$

where $\mathbf{x}_t \in \mathbb{V}^{77}$ is the token sequence at step t, and $\mathtt{c}$ denotes conditioning inputs. The logits can be tempered or top-k filtered before sampling each token independently. The resulting sequence $\mathbf{x}_{t-1}$ seeds the next step, enabling fast, parallel decoding without autoregressive constraints.

Because all positions are updated in parallel, Muddit preserves global syntactic and semantic structure throughout the reverse diffusion process—unlike left-to-right autoregressive models, which can only condition on past predictions. Empirically, as few as $16 \leq T \leq 32$ steps are sufficient to approximate the natural language distribution with high fidelity. Thus, Muddit unifies diffusion generation with parallel decoding, effectively overcoming the serial bottleneck that limits conventional autoregressive multimodal models.

We present 2 examples in Fig. 11 and Fig. 12.

5 Discussion

5.1 Limitations

While Muddit advances discrete diffusion for unified multimodal generation, it still presents several limitations. First, due to its token-level discrete representation, the model may underperform continuous diffusion models in generating photorealistic or high-resolution images. Second, Muddit is initialized from a pretrained text-to-image foundation model, which offers strong visual priors but limits its capacity for rich text understanding and generation compared to the latest large language models. This makes it less suitable for tasks that require long-form understanding and generation or deep linguistic reasoning.

5.2 Broader Impacts

Muddit explores a new paradigm in multimodal generation by leveraging a strong visual prior as the backbone, in contrast to the prevailing trend of scaling large language models. This offers a complementary path toward efficient, grounded multimodal generation, particularly in vision-centric applications. The model's ability to generate aligned visual and textual outputs in a fast, parallel manner could benefit downstream tasks, especially in completion-based scenarios such as masked captioning, image editing, and code implementation. However, as with all generative models, there remains a risk of misuse in synthetic content creation.

5.3 Additional Qualitative Results

Image-to-text Generation. We present more examples for image-to-text generation in Fig. 7.

Text-to-image Generation. We present more examples for text-to-image generation in Fig. 8.

Visual Question Answering. We present more examples for visual question answering in Fig. 9. Muddit reliably identifies fine-grained attributes (*e.g.*, "blonde" hair), object categories (*e.g.*, "beagle"), and physical affordances (*e.g.*, answering "No" to crossing at a red light). Notably, it also handles commonsense reasoning and spatial localization, such as inferring traffic legality or locating vehicles on the street.

Image-guided text editing. Zero-shot text-guided image editing performance is already verified and presented in Meissonic [5]. As the successor to Meissonic, we present Muddit's performance on the image-guided text editing task, where the model completes a masked sentence based on the input image. As shown in Fig. 10, given a partially masked caption and an image, Muddit fills in the blanks with semantically and visually grounded phrases.

6 Conclusion

In this work, we present Muddit, a unified generative framework that employs discrete diffusion to bridge text and image modalities. By unifying image and text generation within a single model, Muddit demonstrates strong performance across text-to-image, image-to-text, and VQA tasks. Notably, it outperforms or matches the capabilities of significantly larger autoregressive models, while enabling fast, parallel inference. Our results validate the effectiveness of discrete denoising as a general-purpose modeling strategy and highlight its potential to serve as a scalable backbone for future multimodal systems.

Assistant: The image depicts a charming street scene with a cobblestone pavement and tall buildings. A white dog is resting. The dog is lying down near the bicycle on the street. To the left of the street, there is an old bicycle parked to the side of the street.

Assistant: The image is a black - and - white photograph featuring three individuals sitting together on a bench. They appear to be in an urban setting, as indicated by the trees and buildings in the background. The individuals are dressed in police uniforms, with one person looking behind the first individual from the second individual.

Assistant: The image depicts an elderly man standing on a city street, wearing a red jacket and a black cap talking on cell phone to his ear. The background shows a crowd of people walking around, some of whom are handbags and others visible in the distance. The crowd appears to be the middle of a busy area.

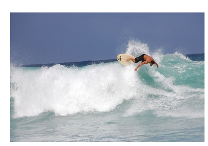
Assistant: The image depicts a close - up of a large, round pizza placed on a wooden table. The pizza is topped with cheese, and there are several pieces of basil leaves. In the background, there are other people sitting around.

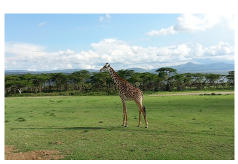
Assistant: The image depicts a tennis match on a blue court. the tennis match is set against a backdrop of a blue court. The player, wearing a white shirt and white shorts, is in mid action and holding a tennis racket, possibly preparing for a shot. the court has a row of seats.

Assistant: The image depicts a vibrant and colorful arrangement of umbrellas hanging from the branches of trees. The umbrellas come in color and shape, including pink, red, and yellow, blue, and purple. They are arranged in a circular pattern across the umbrellas, creating a visually appealing display. The background features a mix of green trees and foliage, suggesting an outdoor.

Assistant: The image depicts a bustling urban street scene with several notable elements. In the foreground, there is a tall green clock tower with a clock face and a pole, which stands prominently on the street. The clock has a clock face on it, indicating the time. In the background, there are several parked cars.

Assistant: The image depicts a person surfing on a surfboard on a large wave. The surfer is in mid-air, with his arms bent, navigating the balance of the wave. The background shows a clear blue sky, adding depth to the overall scene. The surfer seems to be in the middle of the the wave.





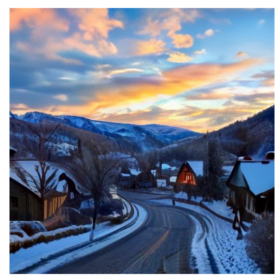
Assistant: The image depicts a giraffe standing on a grassy field in a rural outdoor setting. The background features a clear blue sky with some scattered clouds, and there are trees in the distant distance. The giraffe appears to be looking towards a distant horizon.

Assistant: The image depicts a stop sign in front of a grassy field with majestic mountains in the background. The sky is clear with a scattered clouds, and snow-capped mountains rise in the background. The scene appears to be in a rural setting.

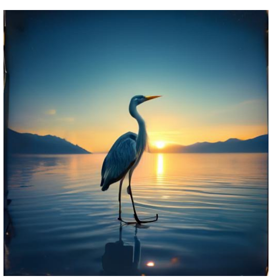
Figure 7: Image-to-text generated results.

Human: An editorial fashion photo portrait of a striking avant-garde model with bubble details. Iridescent art and pop surrealism influence the image's ethereal feel. Pastelaesthetic hues of soft pink, tangerine, and green create a dreamy backdrop. Seamless pink background, studio lighting emphasizes the model's silhouette against the soft pastel palette.

Human: Dark, heavy rainclouds gather over a rocky mountain range, obscuring the peaks. A weathered canvas tent flaps open in the wind, its silhouette stark against the gray backdrop. Water pours down in sheets, blurring the image and creating a misty effect.



Human: A quaint small town nestled amidst snow-capped hills, bathed in the soft, golden hues of dusk. The sky is a tapestry of twilight blue and orange, casting long shadows across cobblestone streets as frosted trees stand bare against the fading light. A sense of quiet peace emanates from the scene.



Human: An old cyanotype photograph capturing a serene lake at twilight. A majestic heron stands in the tranquil waters, its long legs poised, as the last golden rays of sunlight paint the sky with hues of blue and purple. The distant mountain range glows warmly with the soft, crepuscular light. Sharp focus on the heron and water reflection. Wide-angle lens captures a panoramic scene.

Human: A vibrant Japanese garden, inspired by Van Gogh's swirling brushstrokes. Vivid reds, yellows, blues, and greens dominate the scene. Cherry blossoms bloom in full splendor against a backdrop of ancient stone lanterns, rendered with lush foliage and dappled sunlight. The effect is heightened by 4K resolution and cinematic depth.

Figure 8: Text-to-image generation results.

Figure 9: Visual question answering results.

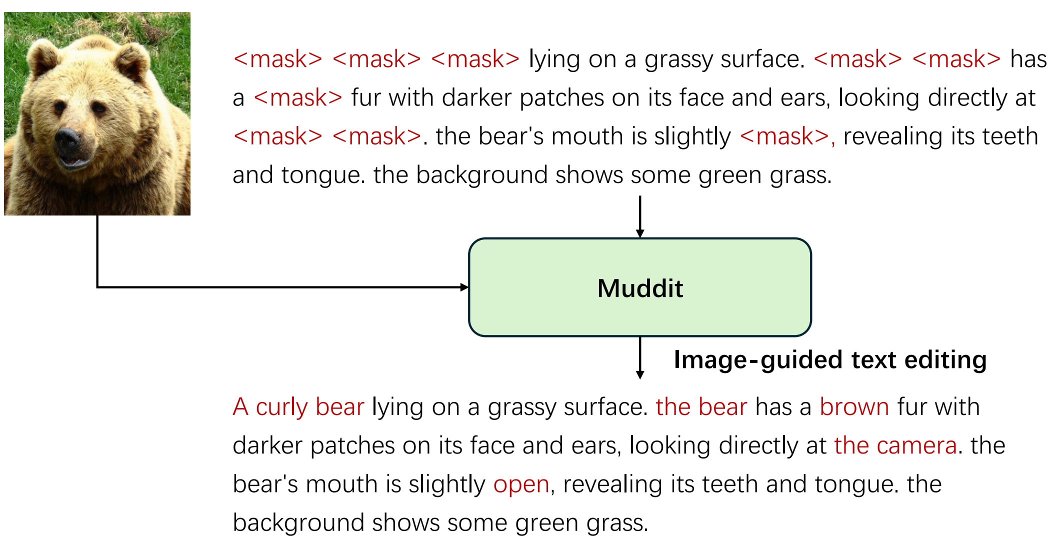


Figure 10: Image-guided text editing results.

User: Please describe this image.

T=1

<mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask>

<mask> <mask> depicts <mask> <mask> and whimsical <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask>, <mask> <mask> <mask> <mask> <mask> <mask>. <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask>

The <mask> depicts a <mask> and whimsical <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask>, <mask> <mask> <mask> <mask> <mask> <mask> <mask>. the <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask>, <mask> <mask> <mask> <mask> <mask> <mask>

The image depicts a surreal and whimsical scene in what appears to be a <mask> <mask>, possibly <mask> <mask> or a dining room. the floor is covered with <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask>, <mask> <mask> <mask> <mask> <mask>

The image depicts a surreal and whimsical scene in what appears to be a domestic setting, possibly a room or a dining room. the floor is covered with numerous pink flowers, adding a touch of <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask> <mask>, <mask> to the <mask> <mask> <mask> <mask> <mask> <mask> <mask>

The image depicts a surreal and whimsical scene in what appears to be a domestic setting, possibly a room or a dining room. the floor is covered with numerous pink flowers, adding a touch of <mask>. the petals are scattered throughout the room, adding to the dreamlike quality of the scene.

The image depicts a surreal and whimsical scene in what appears to be a domestic setting, possibly a room or a dining room. the floor is covered with numerous pink flowers, adding a touch of color. the petals are scattered throughout the room, adding to the dreamlike quality of the scene.

The image depicts a surreal and whimsical scene in what appears to be a domestic setting, possibly a room or a dining room. the floor is covered with numerous pink flowers, adding a touch of color. the petals are scattered throughout the room, adding to the dreamlike quality of the scene.

T=0

Figure 11: Image-to-text generated results in each step.

Figure 12: Image-to-text generated results in each step.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 9

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 8

[3] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models, 2023. 9

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. 9

[5] Jinbin Bai, Tian Ye, Wei Chow, Enxin Song, Qing-Guo Chen, Xiangtai Li, Zhen Dong, Lei Zhu, and Shuicheng Yan. Meissonic: Revitalizing masked generative transformers for efficient high-resolution text-to-image synthesis. *ICLR*, 2025. 2, 3, 4, 6, 7, 13

[6] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 3

[7] 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.*, 2:3, 2023. 7

[8] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11315–11325, 2022. 2, 3, 5

[9] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset, 2025. 1

[10] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*, 2024. 7

[11] Liang Chen, Shuai Bai, Wenhao Chai, Weichu Xie, Haozhe Zhao, Leon Vinci, Junyang Lin, and Baobao Chang. Multimodal representation alignment for image generation: Text-image interleaved control is easier than you think. *arXiv preprint arXiv:2502.20172*, 2025. 1, 2

[12] Sixiang Chen, Jinbin Bai, Zhuoran Zhao, Tian Ye, Qingyu Shi, Donghao Zhou, Wenhao Chai, Xin Lin, Jianzong Wu, Chao Tang, et al. An empirical study of gpt-4o image generation capabilities. *arXiv preprint arXiv:2504.05979*, 2025. 3

[13] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 3

[14] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025. 9

[15] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3, 4

[16] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023. 2

[17] 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. 7

[18] Lijie Fan, Tianhong Li, Siyang Qin, Yuanzhen Li, Chen Sun, Michael Rubinstein, Deqing Sun, Kaiming He, and Yonglong Tian. Fluid: Scaling autoregressive text-to-image generative models with continuous tokens. *arXiv preprint arXiv:2410.13863*, 2024. 3

[19] 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. 8

[20] 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. 2, 7

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

2024. 7
[22] Jianyuan Guo, Hanting Chen, Chengcheng Wang, Kai Han, Chang Xu, and Yunhe Wang. Vision superalignment: Weak-to-strong generalization for vision foundation models. *arXiv preprint arXiv:2402.03749*, 2024. 3
[23] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 16000–16009, 2022. 3
[24] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 8
[25] Inception. Inception labs. `https://www.inceptionlabs.ai/`, Feb 2025. Accessed: 2025-05-16. 2
[26] Black Forest Labs. Announcing black forest labs, 2024. `https://blackforestlabs.ai/announcing-black-forest-labs/`. 4
[27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 9
[28] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37:56424–56445, 2024. 3
[29] Zijie Li, Henry Li, Yichun Shi, Amir Barati Farimani, Yuval Kluger, Linjie Yang, and Peng Wang. Dual diffusion for unified image generation and understanding. *arXiv preprint arXiv:2501.00289*, 2024. 1, 3, 9
[30] Zijie Li, Henry Li, Yichun Shi, Amir Barati Farimani, Yuval Kluger, Linjie Yang, and Peng Wang. Dual diffusion for unified image generation and understanding, 2024. 7
[31] 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 *ECV*, 2014. 8
[32] Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining, 2024. 3
[33] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. 9
[34] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv preprint arXiv:2402.08268*, 2024. 3, 9
[35] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv preprint*, 2024. 7
[36] Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*, 2023. 3
[37] Xinyin Ma, Runpeng Yu, Gongfan Fang, and Xinchao Wang. dkv-cache: The cache for diffusion language models. *arXiv preprint arXiv:2505.15781*, 2025. 11
[38] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Liang Zhao, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. *arXiv preprint arXiv:2411.07975*, 2024. 3
[39] Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025. 2
[40] OpenAI. Addendum to gpt-4o system card: 4o image generation, 2025. Accessed: 2025-04-02. 3
[41] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025. 1
[42] 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. 7
[43] 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*, 2021. 6
[44] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 7
[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 7
[46] 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. 1
[47] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023. 2
[48] Alexander Swerdlow, Mihir Prabhudesai, Siddharth Gandhi, Deepak Pathak, and Katerina Fragkiadaki. Unified multimodal discrete diffusion. *arXiv preprint arXiv:2503.20853*, 2025. 2, 3
[49] Alexander Swerdlow, Mihir Prabhudesai, Siddharth Gandhi, Deepak Pathak, and Katerina Fragkiadaki. Unified multimodal discrete diffusion. *arXiv preprint arXiv:2503.20853*, 2025. 7, 9

[50] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 3, 7, 9

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

[52] 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. 1

[53] Chunwei Wang, Guansong Lu, Junwei Yang, Runhui Huang, Jianhua Han, Lu Hou, Wei Zhang, and Hang Xu. Illume: Illuminating your llms to see, draw, and self-enhance. *arXiv preprint arXiv:2412.06673*, 2024. 2

[54] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 3

[55] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024. 3

[56] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 3, 7, 9

[57] Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *arXiv preprint arXiv:2505.15809*, 2025. 2

[58] Chuyang Zhao, Yuxing Song, Wenhao Wang, Haocheng Feng, Errui Ding, Yifan Sun, Xinyan Xiao, and Jingdong Wang. Monoformer: One transformer for both diffusion and autoregression. *arXiv preprint arXiv:2409.16280*, 2024. 3

[59] Mengyu Zheng, Yehui Tang, Zhiwei Hao, Kai Han, Yunhe Wang, and Chang Xu. Adapt without forgetting: Distill proximity from dual teachers in vision-language models. In *European Conference on Computer Vision*, pages 109–125. Springer, 2024. 3

[60] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024. 3, 7, 9