

LLaDA-o: An Effective and Length-Adaptive Omni Diffusion Model

Zebin You^{1,2,3*} Xiaolu Zhang⁴ JUN ZHOU⁴ Chongxuan Li^{1,2,3†} Ji-Rong Wen^{1,2,3}

¹Gaoling School of AI, Renmin University of China ²Beijing Key Laboratory of Research on Large Models and Intelligent Governance ³Engineering Research Center of Next-Generation Intelligent Search and Recommendation, MOE ⁴Ant Group
{zebin, chongxuanli}@ruc.edu.cn

Abstract

We present **LLaDA-o**, an effective and length-adaptive omni diffusion model for multimodal understanding and generation. LLaDA-o is built on a Mixture of Diffusion (MoD) framework that decouples discrete masked diffusion for text understanding and continuous diffusion for visual generation, while coupling them through a shared, simple, and efficient attention backbone that reduces redundant computation for fixed conditions. Building on MoD, we further introduce a data-centric length adaptation strategy that enables flexible-length decoding in multimodal settings without architectural changes. Extensive experiments show that LLaDA-o achieves strong performance on multimodal understanding and generation benchmarks, reaching 87.04 on DPG-Bench and 0.86 on GenEval. To support future work on omni-diffusion models, we release the model and training/evaluation code at <https://github.com/ML-GSAI/LLaDA-o>.

1 Introduction

Masked diffusion models (MDMs) (Austin et al., 2021; Lou et al., 2024; Shi et al., 2024; Sahoo et al., 2024; Ou et al., 2025) have recently emerged as a promising alternative to autoregressive (AR) language models. By iteratively denoising masked tokens in parallel, MDMs have demonstrated strong language modeling performance (Nie et al., 2025; Ye et al., 2025; Cheng et al., 2025; Bie et al., 2025) and offer appealing properties such as bidirectional context modeling and improved inference parallelism (Song et al., 2025; Labs et al., 2025; Google, 2025), drawing increasing attention to diffusion-based language modeling.

Despite this progress, the potential of diffusion language models as *omni* models for multimodal

understanding and generation remains largely underexplored. A key challenge is that text and images favor different diffusion dynamics: masked diffusion naturally operates over discrete language tokens, whereas for images, continuous diffusion in latent space has become the de facto standard (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020b). Although attempts exist (Pynadath et al., 2025; Chang et al., 2022), these modality-specific preferences do not transfer trivially across domains, motivating a hybrid design that models text discretely and images continuously.

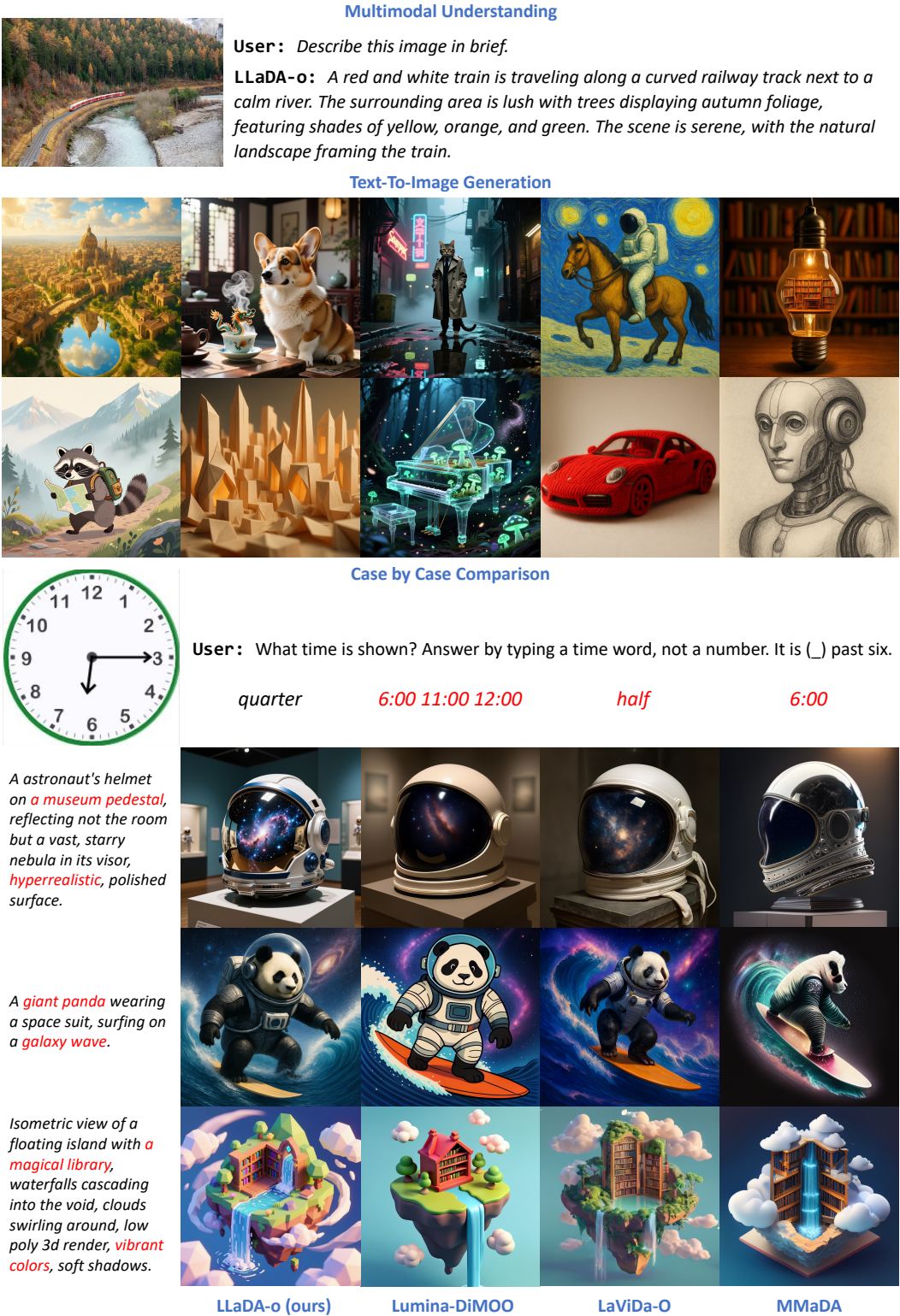
However, naively co-training both modalities within a single dense diffusion model is often ineffective. The heterogeneous state spaces and corruption processes can induce objective mismatch and gradient interference, leading to unstable optimization and suboptimal performance (Li et al., 2025c). In addition, existing omni-diffusion models (Xin et al., 2025; Li et al., 2025a; Yang et al., 2025a) often assume a fixed length for understanding, constraining their applicability open-ended settings.

To address these challenges, we propose **LLaDA-o**, an effective and length-adaptive omni diffusion model built upon the *Mixture of Diffusion* (MoD) framework. MoD decouples modality-appropriate diffusion processes into specialized experts while maintaining a shared attention backbone for cross-modality interaction. Specifically, it assigns masked diffusion to an *understanding expert* for text and visual encoder tokens, and continuous diffusion to a *generation expert* for visual latent tokens, mitigating optimization conflicts in dense co-training. Building on MoD, we further introduce an efficient attention mechanism and a data-centric length adaptation strategy, enabling scalable inference and flexible-length generation in multimodal settings.

We evaluate LLaDA-o from both qualitative and quantitative perspectives. Qualitatively, Fig. 1 shows that LLaDA-o achieves stronger multi-

*Work done during an internship at Ant Group.

†Correspondence to: Chongxuan Li.



among omni-diffusion models. We further evaluate text-to-image generation on the widely used GenEval (Ghosh et al., 2023) and DPG-Bench (Hu et al., 2024) benchmarks against strong generation-only and unified multimodal models. LLaDA-o obtains the highest DPG-Bench score (87.04), while its overall GenEval score is lower than Lumina-DiMOO (0.86 vs. 0.91). Besides, we present a comprehensive analysis for LLaDA-o. Overall, these results highlight the performance of LLaDA-o as a unified omni-diffusion model.

2 Preliminaries

We present preliminaries on diffusion models.

2.1 Continuous Diffusion Models

Continuous diffusion models (CDMs) (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020b) constitute a core paradigm in modern image generation, particularly when combined with diffusion Transformers (Peebles and Xie, 2023; Bao et al., 2023). Conceptually, a CDM specifies a forward stochastic process that corrupts data into noise, and learns to reverse this process to generate samples from the noise distribution. This mechanism can be characterized using stochastic differential equations (Song et al., 2020b). At inference, the same dynamics admit an equivalent ordinary differential equation (ODE) formulation, leading to deterministic samplers (Song et al., 2020a; Lu et al., 2022). This ODE perspective has inspired a series of variants, including rectified flow (RF) (Liu, 2022) and Flow Matching (Lipman et al., 2022).

In particular, RF (Liu, 2022) connects the noise distribution π_0 and data distribution π_1 via a deterministic linear path. Given a data-noise pair (x_0, x_1) , the intermediate state at time $t \in [0, 1]$ is defined as strict linear interpolation:

$$x_t = (1 - t)x_0 + tx_1. \quad (1)$$

RF learns a velocity field $p_\theta(x, t)$ to match the constant flow direction $x_1 - x_0$ by minimizing the following objective:

$$\mathcal{L}(\theta) = \mathbb{E}_{(x_0, x_1), t} \left[\left\| (x_1 - x_0) - p_\theta(x_t, t) \right\|_2^2 \right]. \quad (2)$$

Sampling corresponds to solving the ODE $dz_t/dt = p_\theta(z_t, t)$ initialized from $z_0 \sim \pi_0$. Formally, RF discretizes the following integral of the learned velocity field over the time interval (via the

Euler method, for instance):

$$z_1 = z_0 + \int_0^1 p_\theta(z_t, t) dt. \quad (3)$$

2.2 Diffusion Language Models

Discrete diffusion models (Sohl-Dickstein et al., 2015; Austin et al., 2021) are generative models tailored to discrete data, such as token sequences. Analogous to CDMs, they define a forward Markov process of discrete state that progressively corrupts the data toward a simple prior, typically a uniform distribution over tokens or an absorbing mask state. The model then learns the corresponding reverse dynamics to generate discrete data from this prior. A special case of discrete diffusion, masked diffusion models (MDMs) (Lou et al., 2024; Shi et al., 2024; Sahoo et al., 2024; Ou et al., 2025), have demonstrated strong potential and favorable scaling properties in language modeling.

Formally, let $x_0 = [x^i]_{i=1}^N$ denote a sequence of N tokens, and let $[M]$ represent a special mask token. In the forward process, tokens are independently corrupted based on a time step t uniformly sampled from $[0, 1]$. Let α_t be a continuous monotonically decreasing function with $\alpha_0 = 1$ and $\alpha_1 = 0$. Each token either remains unchanged with probability α_t or is replaced by $[M]$ with probability $1 - \alpha_t$ as follows:

$$q_{t|0}(x_t|x_0) = \prod_{i=0}^{N-1} q_{t|0}(x_t^i|x_0^i), \quad (4)$$

$$q_{t|0}(x_t^i|x_0^i) = \begin{cases} \alpha_t, & x_t^i = x_0^i, \\ 1 - \alpha_t, & x_t^i = [M]. \end{cases}$$

The reverse process reconstructs the sequence iteratively. In particular, consider the transformation from time t to s (where $0 \leq s < t \leq 1$). Tokens that are already predicted remain fixed. Conversely, masked tokens either stay masked with probability $\frac{1-\alpha_s}{1-\alpha_t}$ or are decoded with probability $\frac{\alpha_s-\alpha_t}{1-\alpha_t}$ based on the model prediction $p_\theta(x_0^i|x_t)$:

$$q_{s|t}(x_s|x_t) = \prod_{i=0}^{N-1} q_{s|t}(x_s^i|x_t^i),$$

$$q_{s|t}(x_s^i|x_t^i) = \begin{cases} 1, & x_t^i \neq [M], x_s^i = x_t^i, \\ \frac{1-\alpha_s}{1-\alpha_t}, & x_t^i = [M], x_s^i = [M], \\ \frac{\alpha_s-\alpha_t}{1-\alpha_t} p_\theta(x_0^i|x_t), & x_t^i = [M], x_s^i \neq [M], \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

During inference, MDMs generate text by iteratively simulating this reverse transition, converting a fully masked sequence into coherent text.

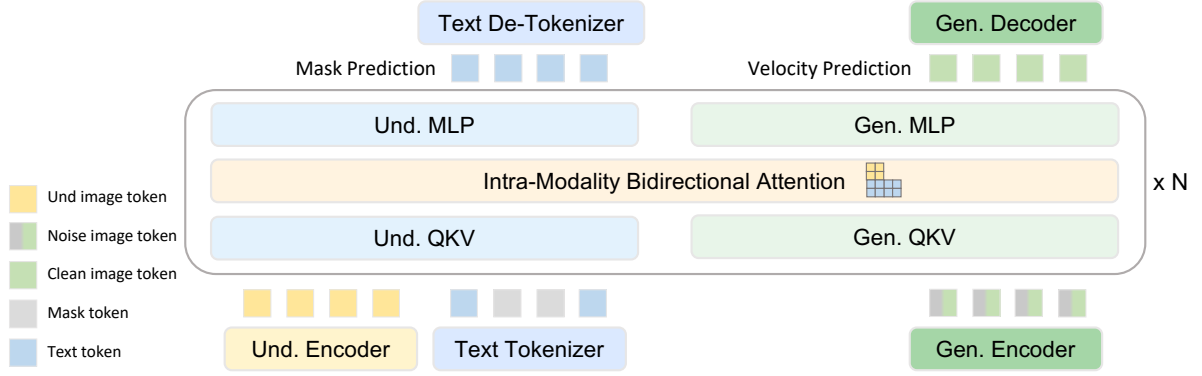


Figure 2: Overview of LLaDA-o: the mixture of diffusion framework.

Intuitively, MDMs function as mask predictors, aiming to recover masked tokens from the observed context. Formally, the training objective of MDMs is defined as:

$$\mathcal{L}(\theta) = \int_0^1 \frac{1}{t} \mathbb{E}_{\mathbf{x}_0, \mathbf{x}_t \sim q_t|0} \left[\sum_{i: \mathbf{x}_t^i = [\mathbf{M}]} -\log p_{\theta}(\mathbf{x}_0^i | \mathbf{x}_t) \right] dt. \quad (6)$$

Diffusion large language models (dLLMs) (Nie et al., 2025; Ye et al., 2025) demonstrate the viability of masked diffusion at scale, achieving competitive performance while enabling parallel decoding and flexible generation control.

3 Method

Motivated by the discussion in Sec. 1, LLaDA-o unifies multimodal understanding and generation within a single diffusion framework by separating discrete and continuous modalities, coupling them through a shared and efficient attention mechanism, and enabling adaptive-length training and inference, detailed as follows.

3.1 The Mixture of Diffusion Framework

At the core of LLaDA-o is a hybrid diffusion design that treats discrete tokens and continuous visual latents with their respective optimal parameterizations. However, naively co-training these modalities in a single dense model is often ineffective (Li et al., 2025c): the two branches operate on heterogeneous state spaces and corruption processes, which can induce objective mismatch and gradient interference, leading to training conflicts and suboptimal performance.

To address these challenges, we propose *Mixture of Diffusion* (MoD), a unified multi-modal diffusion framework illustrated in Fig. 2. Inspired by

modality-factorized designs (Liang et al., 2024; Deng et al., 2025), MoD employs two diffusion experts to decouple the processing of discrete and continuous modalities: an *understanding expert* that handles text and visual encoder tokens via masked diffusion, and a *generation expert* that handles visual latent tokens via continuous diffusion. While the experts allow for specialized processing, they share the self-attention backbone (see Sec. 3.2) to ensure cross-modality interaction.

Specifically, the understanding expert integrates a vision encoder (Zhai et al., 2023), a lightweight two-layer MLP, and a diffusion language model (Nie et al., 2025) sequentially. The image is encoded into semantic visual tokens by the encoder, projected into the language token space by the MLP, and jointly processed with prompt tokens by the language model. Given a training sample $(\mathbf{v}, \mathbf{p}, \mathbf{r}_0)$, where $\mathbf{v}$ denotes the projected image tokens, $\mathbf{p}$ denotes the prompt, and $\mathbf{r}_0$ denotes the ground-truth response, the entire expert is trained to optimize the variant of Eq. (6) as follows:

$$\mathcal{L}_{\text{und}} = \int_0^1 \frac{1}{t} \mathbb{E}_{\mathbf{v}, \mathbf{p}}_{\mathbf{r}_0, \mathbf{r}_t} \left[\sum_{i: \mathbf{r}_t^i = [\mathbf{M}]} -\log p_{\theta}(\mathbf{r}_0^i | \mathbf{v}, \mathbf{p}, \mathbf{r}_t) \right] dt. \quad (7)$$

In parallel, the generation expert comprises a variational autoencoder (VAE) (Kingma and Welling, 2013) and a diffusion Transformer (Peebles and Xie, 2023). The VAE maps between images and visual latent tokens and its parameters are kept frozen during training. Given the relevant variables $(\mathbf{p}, \mathbf{v}_0, \epsilon)$, where $\mathbf{p}$ denotes the prompt, $\mathbf{v}_0$ denotes the ground-truth image tokens from the VAE, and ϵ denotes a Gaussian noise, the training objective follows Eq. (2):

$$\mathcal{L}_{\text{gen}} = \mathbb{E}_{\mathbf{p}, \epsilon, \mathbf{v}_0, \mathbf{v}_t, t} \left[\left\| (\mathbf{v}_0 - \epsilon) - p_{\theta}(\mathbf{p}, \mathbf{v}_t, t) \right\|_2^2 \right]. \quad (8)$$

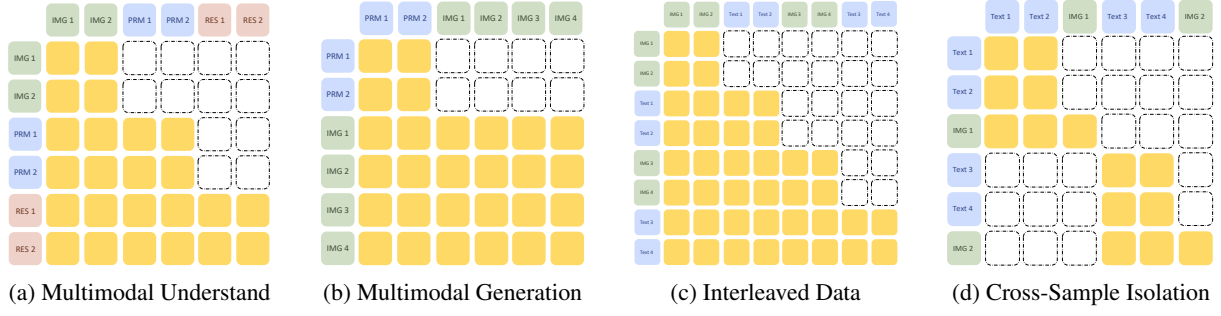


Figure 3: **Implementation of intra-modality bidirectional attention.** Yellow blocks indicate *unmasked* attention, while dashed white boxes denote *masked* attention. Text sequences are explicitly partitioned into Prompts (PRM) and Responses (RES) in cases (a–b).

It is worth noting that both the input image and text in the generation task are also processed by the understanding expert. As a result, the corresponding parameters are jointly trained. Detailed training protocols for multi-turn dialogue and interleaved data are provided in Appendix A.2.

3.2 Intra-Modality Bidirectional Attention

Although MoD decouples modality-specific processing via experts, they share a common attention backbone for cross-modality interaction. Using global attention in this setting is inefficient because it recomputes attention over the entire sequence at every denoising step, even when the condition (e.g., an input image or a text prompt) stays fixed.

To this end, we propose *intra-modality bidirectional attention*, a simple yet efficient attention scheme tailored for MoD. In particular, we partition an input sequence into modality blocks, apply full attention within each block, and enforce causal attention across blocks. This design preserves rich intra-modality context while enabling efficient inference: conditional blocks form a fixed prefix whose Key-Value (KV) cache can be reused across denoising steps, avoiding redundant computation on long sequences.

The proposed attention scheme can be implemented by introducing appropriate attention masks between modality blocks. As illustrated in Fig. 3(a–c), this principle naturally accommodates different data types, including multimodal understanding, generation, and interleaved data. We use sample packing with strict isolation between samples (see Fig. 3d) to prevent cross-sample interference.

In the evaluated MathVista/A100 setting, the full LLaDA-o inference pipeline achieves a 5.9 times throughput advantage over LLaDA-V with FastdLLM ($n = 48$) (You et al., 2025a) (see Fig. 4).

Algorithm 1 Training with Adaptive Length

- 1: **Input:** Model parameters θ ; training data (v, p, r_0) ; extension probability p_{ext} ; truncation probability p_{trunc} .
- 2: **repeat**
- 3: $\tilde{r}_0 \leftarrow r_0, u \sim \mathcal{U}(0, 1)$
- 4: **if** $u < p_{\text{ext}}$ **then**
- 5: Sample an integer k uniformly between $(1, |r_0|)$ and update $\tilde{r}_0$ by appending k EOS tokens.
- 6: **else if** $u < p_{\text{ext}} + p_{\text{trunc}}$ **and** $|r_0| > 16$ **then**
- 7: Sample an integer ℓ uniformly between $(1, |r_0| - 1)$ and update $\tilde{r}_0$ by truncating it to the first ℓ tokens.
- 8: **end if**
- 9: Train on $(v, p, \tilde{r}_0)$ with loss only on response tokens $\tilde{r}_0$; update θ
- 10: **until** convergence

3.3 Adaptive Length Augmentation

The understanding expert adopts masked diffusion for text modeling, typically assuming a fixed target length at inference time. To enable flexible generation without introducing architectural changes or violating sample isolation, we design a data-centric strategy called *adaptive length augmentation* that is fully compatible with our attention scheme.

As summarized in Alg. 1, during training, the target response in each individual sample is stochastically perturbed in two complementary ways. With probability p_{ext} , a bounded number of [EOS] tokens are appended to the original response, exposing the model to explicit termination at different positions. With probability p_{trunc} , the response is truncated to a random prefix, encouraging the model to learn proper continuation from partial targets. Importantly, both operations preserve strict sample iso-

Table 1: **Evaluation on multimodal understanding benchmarks.** The symbol $\dagger$ denotes results from LaViDa-O (Li et al., 2025a), while * indicates results we evaluated using official checkpoints and inference scripts. “-” represents missing data, and “Diff.” refers to diffusion language models. LLaDA-o achieves the highest score among diffusion-based unified models on 6 of the 11 reported metrics.

Model	MMMU val	MME cog.	SeedB perp.	MMB image	MathVerse en-dev	MathVista mini-vis.	MathVista testmini	AI2D -	ChartQA -	DocVQA val	InfoVQA val
AR Based											
Emu3 (Wang et al., 2024)	31.6	-	-	68.2	58.5	-	-	70.0	68.6	76.3	43.8
Janus-Pro (Chen et al., 2025c)	41.0	-	1567	72.1	79.2	-	-	-	-	-	-
MetaMorph (Tong et al., 2025)	41.8	-	-	71.8	75.2	-	-	-	37.1	-	-
Show-o (Xie et al., 2024)	27.4	-	1232	-	-	-	-	-	-	-	-
Show-o2 (Xie et al., 2025)	48.9	-	1620	-	79.3	-	-	78.6	-	-	-
BAGEL (Deng et al., 2025)	55.3	-	1687	-	85.0	-	73.1	-	-	-	-
Diff. Based											
LaViDa-L (Li et al., 2025b)	43.3	341	1365	-	70.5	27.2	44.8	70.0	64.6	59.0	34.2
Dimple (Yu et al., 2025)	45.2	432	1514	-	74.6	-	42.3	74.4	63.4	-	-
LLaDA-V (You et al., 2025a)	48.6	491	1507	74.8	82.9	28.5	59.7	77.8	78.3	83.9	66.3
MMaDA (Yang et al., 2025a)	30.2	242 $\dagger$	1410	64.2	68.5	13.5 $\dagger$	33.7 $\dagger$	66.6 $\dagger$	9.8 $\dagger$	10.9 $\dagger$	14.9 $\dagger$
Lumina-DiMOO (Xin et al., 2025)	58.6	-	1534	83.1	84.5	10.3*	30.3*	43.2*	8.3*	7.2*	6.2*
LaViDa-O (Li et al., 2025a)	45.1	488	1431	-	76.4	36.9	56.9	76.7	80.0	73.7	44.6
LLaDA-o	44.9	549	1412	75.3	71.1	37.1	66.1	79.3	87.9	91.5	54.7

lation and do not require modifying the attention structure or sequence packing strategy.

At inference (see Alg. 2 in Appendix. A.5), we perform block-wise generation (Arriola et al., 2025) under intra-modality bidirectional attention: the fixed conditioning blocks (images and prompt) are cached once, and response tokens are generated iteratively by appending a length- L masked block and denoising it. If [EOS] appears with high confidence, decoding terminates; otherwise, the completed block is cached and generation proceeds to the next block. This strategy enables efficient variable-length generation while fully reusing the KV cache of the fixed prefix.

Compared to approaches that rely on architectural modifications or multi-sample packing with globally bidirectional attention to handle variable-length outputs (Wu et al., 2026b; Kim et al., 2025; Yang et al., 2025b), our method remains lightweight, architecture-agnostic, and compatible with sample isolation and efficient inference.

4 Experiments

We present experimental settings and results.

4.1 Experimental Settings

Model. For the understanding expert, we use the representative LLaDA-8B-Instruct (Nie et al., 2025) to initialize the language model. As the vision encoder, we use SigLIP (Zhai et al., 2023), which has shown strong performance in many MLLMs, and adopt a randomly initialized two-

layer MLP as the projector. For the generation expert, we use the same diffusion transformer architecture as the masked predictor architecture in LLaDA (Nie et al., 2025) and is initialized from it, while additional conditional parameters for time embeddings are randomly initialized. We use VAE of FLUX (Labs, 2024) as the vision encoder for generation due to its strong reconstruction quality.

Training strategy. We train LLaDA-o in three stages to progressively scale both data difficulty and generation fidelity. In Stage 1, we use large-scale image understanding data together with image generation data, where generation is restricted to resolutions up to 512 to stabilize training. In Stage 2, we incorporate multimodal reasoning data and reuse a high-quality subset of the Stage 1 image generation data, while increasing the generation resolution to 1024 to improve high-resolution synthesis. Notably, we do not apply *adaptive length augmentation* for multimodal understanding in the first two stages. In Stage 3, we jointly apply *adaptive length augmentation* to activate variable-length generation for the understanding expert and add more high-quality image generation data, aligning the model with flexible text decoding and stronger visual generation. For more details on the training strategy and data, please refer to Appendix A.1.

Evaluation. We evaluate LLaDA-o on a broad set of benchmarks to reflect the main requirements of unified multimodal models: general knowledge understanding, reasoning, and fine-grained perception, as well as image generation. For multimodal

Table 2: **Evaluation of text-to-image generation on the GenEval benchmark (Ghosh et al., 2023).** “Gen.,” “Obj.,” and “Attr.” denote generation, object, and attribute, respectively, while “-” indicates missing data. LLaDA-o performs strongly on two-object and color-specific generation tasks, while its overall score is lower than Lumina-DiMOO. Following the protocols of BAGEL (Deng et al., 2025) and Show-o2 (Xie et al., 2025), we evaluate the results using rewritten prompts.

Model	# Params	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall↑
Gen. Only								
PixArt- α (Chen et al., 2023)	0.6B	0.98	0.50	0.44	0.80	0.08	0.07	0.48
DALL-E 3 (Betker et al., 2023)	-	0.96	0.87	0.47	0.83	0.43	0.45	0.67
SD3-Medium (Esser et al., 2024)	2B	0.99	0.94	0.72	0.89	0.33	0.60	0.74
FLUX.1-dev (Labs, 2024)	12B	0.98	0.81	0.74	0.79	0.22	0.45	0.66
Unified								
Emu3 (Wang et al., 2024)	8B	-	-	-	-	-	-	0.66
Janus (Wu et al., 2025b)	1.3B	0.97	0.68	0.30	0.84	0.46	0.42	0.61
Janus-Pro (Chen et al., 2025c)	7B	0.99	0.89	0.59	0.90	0.79	0.66	0.80
Mogao (Liao et al., 2025)	7B	1.00	0.97	0.83	0.93	0.84	0.80	0.89
MMaDA (Yang et al., 2025a)	8B	0.99	0.76	0.61	0.84	0.20	0.37	0.63
Show-o (Xie et al., 2024)	1.3B	0.98	0.80	0.66	0.84	0.31	0.50	0.68
BAGEL (Deng et al., 2025)	7B MoT	0.98	0.95	0.84	0.95	0.78	0.77	0.88
LaViDa-O (Li et al., 2025a)	-	-	-	-	-	-	-	0.89
Lumina-DiMOO (Xin et al., 2025)	8B	1.00	0.96	0.87	0.95	0.85	0.82	0.91
Show-o2 (Xie et al., 2025)	7B	1.00	0.87	0.58	0.92	0.52	0.62	0.76
LLaDA-o (ours)	8B MoT	0.99	0.98	0.73	0.96	0.69	0.83	0.86

Table 3: **Evaluation of text-to-image generation on DPG-Bench (Hu et al., 2024).** The symbol $\dagger$ denotes results from Lumina-DiMOO (Xin et al., 2025), while “Gen.” stands for generation and “-” indicates missing data. LLaDA-o achieves the highest overall score in this table (87.04).

Model	# Params	Global	Entity	Attribute	Relation	Other	Overall↑
Gen. Only							
PixArt- α (Chen et al., 2023)	0.6B	74.97	79.32	78.60	82.57	76.96	71.11
PixArt- Σ (Chen et al., 2024)	0.6B	86.89	82.89	88.94	86.59	87.68	80.54
DALL-E 3 (Betker et al., 2023)	-	90.97	89.61	88.39	90.58	89.83	83.50
SD3-Medium (Esser et al., 2024)	2B	87.90	91.01	88.83	80.70	88.68	84.08
FLUX.1-dev (Labs, 2024)	12B	74.35	90.00	88.96	90.87	88.33	83.84
Unified							
Emu3-DPO (Wang et al., 2024)	8B	-	-	-	-	-	81.60
Janus (Wu et al., 2025b)	1.3B	82.33	87.38	87.70	85.46	86.41	79.68
Janus-Pro (Chen et al., 2025c)	7B	86.90	88.90	89.40	89.32	89.48	84.19
Mogao (Liao et al., 2025)	7B	82.37	90.03	88.26	93.18	85.40	84.33
MMaDA † (Yang et al., 2025a)	8B	77.81	78.48	81.74	84.79	63.20	69.97
Show-o (Xie et al., 2024)	1.3B	-	-	-	-	-	67.48
BAGEL (Deng et al., 2025)	7B MoT	88.94	90.37	91.29	90.82	88.67	85.07
LaViDa-O (Li et al., 2025a)	-	-	-	-	-	-	83.20
Lumina-DiMOO (Xin et al., 2025)	8B	81.46	92.08	88.98	94.31	82.00	86.04
Show-o2 (Xie et al., 2025)	7B	89.00	91.78	89.96	91.81	91.64	86.14
LLaDA-o (ours)	8B MoT	92.91	93.30	90.40	91.75	92.79	87.04

understanding, we cover multidisciplinary knowledge (MMM (Yue et al., 2024), MME (Fu et al., 2023), SEED-Bench (Li et al., 2023), and MM-Bench (Liu et al., 2024)), mathematical reasoning (MathVerse (Zhang et al., 2024) and MathVista (Lu

et al., 2023)), and chart/document understanding (AI2D (Kembhavi et al., 2016), ChartQA (Masry et al., 2022), DocVQA (Mathew et al., 2021), and InfoVQA (Mathew et al., 2022)). For text-to-image generation, we use two widely used bench-

marks that test complementary aspects of generation: GenEval (Ghosh et al., 2023), which verifies fine-grained compositional attributes via an object-centric detection pipeline (e.g., object count, spatial relations, and color binding), and DPG-Bench (Hu et al., 2024), which evaluates faithful rendering of long, information-dense prompts with complex entity relationships and rich descriptions.

4.2 Benchmark Results

Multimodal understanding. We compare LLaDA-o with unified multimodal models and multimodal large language models in Tab. 1. LLaDA-o achieves the highest score among diffusion-based unified models on 6 of the 11 reported metrics, demonstrating strong performance of the integrated model. Compared with the autoregressive model BAGEL (Deng et al., 2025), LLaDA-o is generally weaker, which is expected given BAGEL’s stronger language backbone. BAGEL uses Qwen2.5-7B-Instruct (Team, 2024b) trained on 18T tokens, while the LLaDA-8B-Instruct (Nie et al., 2025) is trained on 2.3T tokens, and this gap is reflected in language capability (e.g., 84.8 vs. 49.4 on HumanEval). Despite this disadvantage, LLaDA-o narrows the gap; for example, on MathVista it achieves 66.1, improving over LLaDA-V (You et al., 2025a) (59.7) and approaching BAGEL (73.1). We believe masked diffusion backbones and the overall system can further improve with scale.

Text-to-image generation. We evaluate LLaDA-o against generation-only and unified multimodal models on GenEval (Ghosh et al., 2023) and DPG-Bench (Hu et al., 2024) (Tabs. 2 and 3). On GenEval, LLaDA-o performs strongly on two-object generation and color binding, although its overall score is lower than Lumina-DiMOO (0.86 vs. 0.91). On DPG-Bench, LLaDA-o obtains 87.04, surpassing Lumina-DiMOO (86.04), particularly on long, information-dense prompts. These results demonstrate strong generation quality of the integrated LLaDA-o model. Qualitatively, Fig. 1 shows that LLaDA-o produces more visually appealing images with richer fine-grained details than Lumina-DiMOO and LaViDa-O (Li et al., 2025a) while following the instructions.

We also provide additional qualitative text-to-image samples in Appendix. B.2.

Fig. 4 compares the throughput-accuracy trade-off on MathVista, showing a $5.9\times$ speedup of the full LLaDA-o pipeline over LLaDA-V with

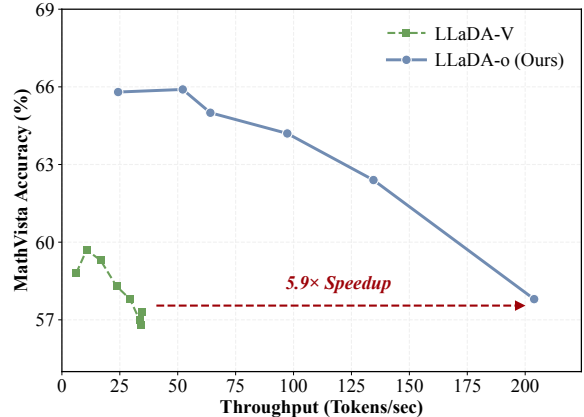


Figure 4: **Comparison of inference efficiency on MathVista.** We visualize the throughput-accuracy trade-off by varying the confidence threshold for LLaDA-o and the refresh interval (n) of Fast-dLLM applied to LLaDA-V. On a single NVIDIA A100, the full LLaDA-o pipeline achieves a $5.9\times$ throughput advantage over LLaDA-V with Fast-dLLM at comparable accuracy.

comparable performance. Detailed confidence-threshold ablations and analysis are provided in Appendix B.1.

5 Related Work

Diffusion large language models. Recently, diffusion large language models (dLLMs) (Nie et al., 2025; Ye et al., 2025; Zhu et al., 2025) have emerged based on masked diffusion models (MDMs) (Ou et al., 2025; Lou et al., 2024; Shi et al., 2024; Sahoo et al., 2024; You et al., 2025b), which are a special case of discrete diffusion models (Sohl-Dickstein et al., 2015; Hoogeboom et al., 2021). Through large-scale pretraining and supervised fine-tuning, these models have achieved performance comparable to strong autoregressive models such as LLaMA3. This demonstrates the practical applicability and strong potential of dLLMs as alternatives to ARMs. Most relevant to our work are unified multimodal models (Li et al., 2025a; Xin et al., 2025; Yang et al., 2025a). However, unlike these approaches that rely on masked diffusion models, we employ continuous diffusion models for image generation, thereby avoiding the information loss caused by image discretization.

Unified multimodal models are initially dominated by autoregressive architectures (Team, 2024a; Wang et al., 2024), which leverage structural consistency with LLMs via unified next-token prediction. Subsequently, hybrid approaches combine autoregressive text processing with diffusion-based image

generation (Zhou et al., 2024; Xie et al., 2024; Tong et al., 2025; Liao et al., 2025; Deng et al., 2025). Concurrently, Diffusion Large Language Models (dLLMs) (Nie et al., 2025; Ye et al., 2025; Zhu et al., 2025) have achieved performance comparable to strong autoregressive models via large-scale pretraining and supervised fine-tuning. Building on this success, recent dLLM-based unified multimodal models (Li et al., 2025a; Xin et al., 2025; Yang et al., 2025a; Swerdlow et al., 2025) employ masked diffusion mechanisms for both multimodal understanding and generation.

6 Conclusion

We presented LLaDA-o, a length-adaptive omni diffusion model for multimodal understanding and generation. Built on a Mixture of Diffusion framework with a shared efficient attention backbone and a data-centric adaptive length training strategy, LLaDA-o enables stable multimodal training and flexible-length generation. Experimental results demonstrate strong performance on multimodal understanding and text-to-image generation tasks. We believe that as masked diffusion models continue to advance in language modeling, LLaDA-o potentially provides a promising foundation for future omni diffusion approaches.

Limitations

A limitation of our work lies in the capability gap of the underlying language backbone. LLaDA-o is built on LLaDA-8B-Instruct, whereas strong autoregressive omni baselines often use more capable backbones and substantially larger pretraining corpora. Thus, part of the remaining performance gap may stem from differences in backbone scale and language capability, rather than the omni-diffusion formulation itself. We leave scaling masked diffusion backbones and more controlled comparisons to future work.

We do not include matched controls that separately remove MoD, replace the attention pattern with global attention, or disable adaptive-length augmentation. Therefore, the reported improvements should be interpreted as evidence for the complete LLaDA-o system rather than as isolated causal effects of individual components.

Acknowledgments

This work was supported by the Beijing Major Science and Technology Project under Contract

no. Z251100008425002; the National Natural Science Foundation of China (Nos. 62522609, 92470118); the Beijing Natural Science Foundation (No. L247030); the fund for building world-class universities (disciplines) of Renmin University of China and the research fund of Ant Group.

References

- Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, and 1 others. 2025. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*.
- Marianne Arriola, Aaron Gokaslan, Justin T Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sekhar Sahoo, and Volodymyr Kuleshov. 2025. Block diffusion: Interpolating between autoregressive and diffusion language models. *arXiv preprint arXiv:2503.09573*.
- Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993.
- Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. 2023. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22669–22679.
- James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, and 1 others. 2023. [Improving image generation with better captions](#). *Computer Science*.
- Tiwei Bie, Maosong Cao, Kun Chen, Lun Du, Mingliang Gong, Zhuochen Gong, Yanmei Gu, Jiaqi Hu, Zenan Huang, Zhenzhong Lan, and 1 others. 2025. Llada2. 0: Scaling up diffusion language models to 100b. *arXiv preprint arXiv:2512.15745*.
- Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. 2022. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11315–11325.
- Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, and 1 others. 2025a. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*.
- Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. 2024. Pixart- σ : Weak-to-strong training of diffusion transformer

- for 4k text-to-image generation. In *European Conference on Computer Vision*, pages 74–91. Springer.
- Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and 1 others. 2023. Pixart- α : Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*.
- Junying Chen, Zhenyang Cai, Pengcheng Chen, Shunian Chen, Ke Ji, Xidong Wang, Yunjin Yang, and Benyou Wang. 2025b. Sharegpt-4o-image: Aligning multimodal models with gpt-4o-level image generation. *arXiv preprint arXiv:2506.18095*.
- Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025c. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.
- Zhihong Chen, Xuehai Bai, Yang Shi, Chaoyou Fu, Huanyu Zhang, Haotian Wang, Xiaoyan Sun, Zhang Zhang, Liang Wang, Yuanxing Zhang, and 1 others. 2025d. Opengpt-4o-image: A comprehensive dataset for advanced image generation and editing. *arXiv preprint arXiv:2509.24900*.
- Shuang Cheng, Yihan Bian, Dawei Liu, Linfeng Zhang, Qian Yao, Zhongbo Tian, Wenhai Wang, Qipeng Guo, Kai Chen, Biqing Qi, and 1 others. 2025. Sdar: A synergistic diffusion-autoregression paradigm for scalable sequence generation. *arXiv preprint arXiv:2510.06303*.
- Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, and 1 others. 2025. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*.
- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, and 1 others. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.
- Rongyao Fang, Aldrich Yu, Chengqi Duan, Linjiang Huang, Shuai Bai, Yuxuan Cai, Kun Wang, Si Liu, Xihui Liu, and Hongsheng Li. 2025. Flux-reason-6m & prism-bench: A million-scale text-to-image reasoning dataset and comprehensive benchmark. *arXiv preprint arXiv:2509.09680*.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.
- Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. 2023. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152.
- Google. 2025. [Gemini diffusion](#).
- Shuhao Gu, Jialing Zhang, Siyuan Zhou, Kevin Yu, Zhaohu Xing, Liangdong Wang, Zhou Cao, Jintao Jia, Zhuoyi Zhang, Yixuan Wang, and 1 others. 2024. Infinity-mm: Scaling multimodal performance with large-scale and high-quality instruction data. *arXiv preprint arXiv:2410.18558*.
- Jiawei Guo, Tianyu Zheng, Yizhi Li, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Graham Neubig, Wenhui Chen, and Xiang Yue. 2025. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13869–13920.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.
- Emiel Hoogetboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. 2021. Argmax flows and multinomial diffusion: Learning categorical distributions. *NeurIPS*, 34:12454–12465.
- Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. 2024. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*.
- jackyhate. 2024. [jackyhate/text-to-image-2m](https://huggingface.co/datasets/jackyhate/text-to-image-2m). <https://huggingface.co/datasets/jackyhate/text-to-image-2m>.
- Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer.
- Jaeyeon Kim, Lee Cheuk-Kit, Carles Domingo-Enrich, Yilun Du, Sham Kakade, Timothy Ngotiaoco, Sitan Chen, and Michael Albergo. 2025. Any-order flexible length masked diffusion. *arXiv preprint arXiv:2509.01025*.
- Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Black Forest Labs. 2024. Flux. <https://github.com/black-forest-labs/flux>.
- Inception Labs, Samar Khanna, Siddhant Kharbanda, Shufan Li, Harshit Varma, Eric Wang, Sawyer

- Birnbaum, Ziyang Luo, Yanis Miraoui, Akash Palrecha, and 1 others. 2025. Mercury: Ultra-fast language models based on diffusion. *arXiv preprint arXiv:2506.17298*.
- Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.
- Shufan Li, Jiuxiang Gu, Kangning Liu, Zhe Lin, Zijun Wei, Aditya Grover, and Jason Kuen. 2025a. Lavidao: Elastic large masked diffusion models for unified multimodal understanding and generation. *arXiv preprint arXiv:2509.19244*.
- Shufan Li, Konstantinos Kallidromitis, Hritik Bansal, Akash Gokul, Yusuke Kato, Kazuki Kozuka, Jason Kuen, Zhe Lin, Kai-Wei Chang, and Aditya Grover. 2025b. Lavidao: A large diffusion language model for multimodal understanding. *arXiv preprint arXiv:2505.16839*.
- Zijie Li, Henry Li, Yichun Shi, Amir Barati Farimani, Yuval Kluger, Linjie Yang, and Peng Wang. 2025c. Dual diffusion for unified image generation and understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2779–2790.
- Weixin Liang, Lili Yu, Liang Luo, Srinivasan Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen-tau Yih, Luke Zettlemoyer, and 1 others. 2024. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *arXiv preprint arXiv:2411.04996*.
- Chao Liao, Liyang Liu, Xun Wang, Zhengxiong Luo, Xinyu Zhang, Wenliang Zhao, Jie Wu, Liang Li, Zhi Tian, and Weilin Huang. 2025. Mogao: An omni foundation model for interleaved multi-modal generation. *arXiv preprint arXiv:2505.05472*.
- Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.
- Qiang Liu. 2022. Rectified flow: A marginal preserving approach to optimal transport. *arXiv preprint arXiv:2209.14577*.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.
- Aaron Lou, Chenlin Meng, and Stefano Ermon. 2024. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Forty-first International Conference on Machine Learning*.
- Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems*, 35:5775–5787.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *arXiv preprint arXiv:2310.02255*.
- Jian Ma, Xujie Zhu, Zihao Pan, Qirong Peng, Xu Guo, Chen Chen, and Haonan Lu. 2025. X2edit: Revisiting arbitrary-instruction image editing through self-constructed data and task-aware representation learning. *arXiv preprint arXiv:2508.07607*.
- Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.
- Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.
- Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.
- Jordan Meyer, Nick Padgett, Cullen Miller, and Laura Exline. 2024. Public domain 12m: A highly aesthetic image-text dataset with novel governance mechanisms. *arXiv preprint arXiv:2410.23144*.
- Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. Large language diffusion models. *arXiv preprint arXiv:2502.09992*.
- Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. 2025. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. In *International Conference on Learning Representations*, volume 2025, pages 64972–65009.
- William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205.
- Patrick Pynadath, Jiabin Shi, and Ruqi Zhang. 2025. Candi: Hybrid discrete-continuous diffusion models. *arXiv preprint arXiv:2510.22510*.
- Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. 2024. Simple and effective masked diffusion language models. *arXiv preprint arXiv:2406.07524*.

- Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K Titsias. 2024. Simplified and generalized masked diffusion for discrete data. *arXiv preprint arXiv:2406.04329*.
- Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR.
- Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020a. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.
- Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020b. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.
- Yuxuan Song, Zheng Zhang, Cheng Luo, Pengyang Gao, Fan Xia, Hao Luo, Zheng Li, Yuehang Yang, Hongli Yu, Xingwei Qu, and 1 others. 2025. Seed diffusion: A large-scale diffusion language model with high-speed inference. *arXiv preprint arXiv:2508.02193*.
- Alexander Szwedlow, Mihir Prabhudesai, Siddharth Gandhi, Deepak Pathak, and Katerina Fragkiadaki. 2025. Unified multimodal discrete diffusion. *arXiv preprint arXiv:2503.20853*.
- Chameleon Team. 2024a. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*.
- Qwen Team. 2024b. [Qwen2.5: A party of foundation models](#).
- Shengbang Tong, David Fan, Jiachen Li, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. 2025. Metamorph: Multimodal understanding and generation via instruction tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17001–17012.
- Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiyang Yu, and 1 others. 2024. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*.
- Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, and 1 others. 2025a. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*.
- Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and 1 others. 2025b. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12966–12977.
- Chenyuan Wu, Jiahao Wang, Pengfei Zheng, Ruirui Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, and 1 others. 2026a. Omni-gen2: Towards instruction-aligned multimodal generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21964–21975.
- Zirui Wu, Lin Zheng, Zhihui Xie, Jiacheng Ye, Jiahui Gao, Shansan Gong, Yansong Feng, Zhenguo Li, Wei Bi, Guorui Zhou, and 1 others. 2026b. Dreamon: Diffusion language models for code infilling beyond fixed-size canvas. In *International Conference on Learning Representations*, volume 2026, pages 76626–76643.
- Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. 2024. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*.
- Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. 2025. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*.
- Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, and 1 others. 2025. Lumina-dimoo: An omni diffusion large language model for multimodal generation and understanding. *arXiv preprint arXiv:2510.06308*.
- Ling Yang, Ye Tian, Bowen Li, Xincheng Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. 2025a. Mmada: Multimodal large diffusion language models. *arXiv preprint arXiv:2505.15809*.
- Yicun Yang, Cong Wang, Shaobo Wang, Zichen Wen, Biqing Qi, Hanlin Xu, and Linfeng Zhang. 2025b. Diffusion llm with native variable generation lengths: Let [eos] lead the way. *arXiv preprint arXiv:2510.24605*.
- Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. 2025. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*.
- Zebin You, Shen Nie, Xiaolu Zhang, Jun Hu, Jun Zhou, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. 2025a. Llada-v: Large language diffusion models with visual instruction tuning. *arXiv preprint arXiv:2505.16933*.
- Zebin You, Jingyang Ou, Xiaolu Zhang, Jun Hu, Jun Zhou, and Chongxuan Li. 2025b. Effective and efficient masked image generation models. *arXiv preprint arXiv:2503.07197*.
- Runpeng Yu, Xinyin Ma, and Xinchao Wang. 2025. Dimple: Discrete diffusion multimodal large language model with parallel decoding. *arXiv preprint arXiv:2505.16990*.

- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.
- Hong Zhang, Zhongjie Duan, Xingjun Wang, Yuze Zhao, Weiyi Lu, Zhipeng Di, Yixuan Xu, Yingda Chen, and Yu Zhang. 2025a. Nexus-gen: Unified image understanding, generation, and editing via prefilled autoregression in shared embedding space. *arXiv preprint arXiv:2504.21356*.
- Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, and 1 others. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer.
- Yi Zhang, Bolin Ni, Xin-Sheng Chen, Heng-Rui Zhang, Yongming Rao, Houwen Peng, Qinglin Lu, Han Hu, Meng-Hao Guo, and Shi-Min Hu. 2025b. Bee: A high-quality corpus and full-stack suite to unlock advanced fully open mllms. *arXiv preprint arXiv:2510.13795*.
- Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. 2024. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*.
- Fengqi Zhu, Zebin You, Yipeng Xing, Zenan Huang, Lin Liu, Yihong Zhuang, Guoshan Lu, Kangyu Wang, Xudong Wang, Lanning Wei, and 1 others. 2025. Llada-moe: A sparse moe diffusion language model. *arXiv preprint arXiv:2509.24389*.
- Le Zhuo, Liangbing Zhao, Sayak Paul, Yue Liao, Renrui Zhang, Yi Xin, Peng Gao, Mohamed Elhoseiny, and Hongsheng Li. 2025. From reflection to perfection: Scaling inference-time optimization for text-to-image diffusion models via reflection tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15329–15339.

A Experimental Details

A.1 Training Stages and Data Setup

The training of LLaDA-o proceeds in a preliminary alignment phase followed by three stages.

Projector Alignment. Prior to the main training, we focus on aligning visual representations with the understanding expert. We train the MLP projector using the Stage 1 data from Infinity-MM (Gu et al., 2024), while keeping all other components frozen.

Stage 1: Foundation Setup. In this stage, we establish the baseline capabilities using large-scale image understanding and generation data. We restrict image generation to a resolution of 512 and do not apply *adaptive length augmentation* for multimodal understanding. For *text*, we use an in-house 10M SFT dataset. For *multimodal understanding*, we utilize Stage 2-4 data from Infinity-MM (Gu et al., 2024), MAmoTH-VL-Instruct-12M (Guo et al., 2025), LLaVA-OneVision-1.5 (An et al., 2025), and FineVision (An et al., 2025). For *text-to-image generation*, we combine image captioning data (from Infinity-MM (Gu et al., 2024) and LLaVA-OneVision-1.5 (An et al., 2025)) with generation datasets including PD12M (Meyer et al., 2024), Text-to-Image-2M (jackyhate, 2024), BLIP3o-Pretrain (Chen et al., 2025a), Nexus-Gen (Zhang et al., 2025a), FLUX-Reason-6M (Fang et al., 2025), and synthetic data from Qwen-Image (Wu et al., 2025a).

Stage 2: High-Resolution and Reasoning. We incorporate multimodal reasoning data and increase the generation resolution to 1024. *Adaptive length augmentation* remains disabled for understanding tasks. In terms of data, we switch the *understanding* source to Honey-Data-15M (Zhang et al., 2025b). For *generation*, we refine the dataset by removing image captions and PD12M (Meyer et al., 2024), while increasing the sampling ratio of the remaining high-quality data. We introduce *interleaved multimodal data* from X2Edit (Ma et al., 2025) and OmniGen2 (Wu et al., 2026a), along with editing data from ShareGPT-4o-Image (Chen et al., 2025b) and OpenGPT-4o-Image (Chen et al., 2025d). The text data remains unchanged.

Stage 3: Variable-Length Refinement. In the final stage, we jointly apply *adaptive length augmentation* to activate variable-length generation for the understanding expert and fine-tune the model with high-quality data. For *understanding*, we incorporate Honey-Data-1M and retain Honey-

Data-15M (Zhang et al., 2025b) with a reduced ratio. For *generation*, we add premium datasets including ShareGPT-4o-Image (Chen et al., 2025b), OpenGPT-4o-Image (Chen et al., 2025d), BLIP3o-60k (Chen et al., 2025a), and GenRef (Zhuo et al., 2025), while reducing the ratio of the Stage 2 generation data. For *interleaved data*, we remove X2Edit (Ma et al., 2025) and keep the rest consistent with Stage 2.

A.2 Multi-turn Dialogue Data and Interleaved Multimodal Data Setup

For multi-turn dialogues, LLaDA (Nie et al., 2025) randomly selects one turn for training. It concatenates the preceding dialogue history, including both prompts and responses, as the input context for that turn, and computes the loss only on the selected turn’s response. This setup is also used in LaViDa (Li et al., 2025b) and Dimple (Yu et al., 2025). In contrast, LLaDA-V (You et al., 2025a) computes the loss on the response of every turn in the dialogue. In LLaDA-o, we follow the LLaDA-V strategy, since it achieves strong performance and shows data scalability in LLaDA-V.

For interleaved multimodal data, we treat the text only as a condition (i.e., prompt) for image generation. Thus, we compute the continuous diffusion loss only on each turn of image tokens and mask out the loss on text tokens, preventing the image objective from interfering with text representations.

If no masked position reaches the threshold in an iteration, we unmask the remaining position with the highest prediction confidence. This fallback guarantees progress and ensures termination, including when $\tau = 1.0$.

A.3 Prompts of Selected Generated Images

Tab. 4 lists the text prompts used to generate the samples shown in Fig. 1. The Image IDs in the table correspond to the spatial arrangement of the images: IDs 1–5 represent the top row (from left to right), and IDs 6–10 represent the bottom row.

A.4 Computational Cost

Tab. 5 details the computational resources and time required for each training stage of LLaDA-o. The primary training phases (Stage 1 and Stage 2) are conducted on 256 NVIDIA H800 GPUs, accounting for the majority of the computational cost. The final refinement stage (Stage 3) is performed using 64 NVIDIA A100 GPUs.

Table 4: **Prompts for the generated samples.** The Image IDs correspond to the order of images in Fig. 1, arranged from left to right and top to bottom.

Image ID	Prompt
1	Bright scene, aerial view, ancient city, fantasy, gorgeous light, mirror reflection, high detail, wide angle lens.
2	A photorealistic corgi sitting calmly in a traditional Chinese tea house, steam from the teacup forming a tiny dragon, morning light, detailed fur.
3	A detective cat in a trench coat standing in a rainy alleyway, cyberpunk neon signs reflecting in puddles, film noir style, volumetric fog, dramatic shadows, cinematic shot.
4	An astronaut riding a horse on the moon, oil painting by Van Gogh.
5	A miniature library inside a vintage lightbulb, warm cozy light.
6	A raccoon wearing a tiny backpack and using a map to navigate a misty mountain trail, storybook illustration, soft colors.
7	A complex landscape of a futuristic city made entirely of folded origami paper, soft warm paper texture, subsurface scattering, dramatic lighting from inside the paper buildings, tilt-shift photography.
8	A transparent glass piano in a dark forest, glowing bioluminescent mushrooms growing inside the piano, magical dust floating in the air, fantasy illustration, intricate details.
9	A red sports car made entirely of knitted wool, soft studio lighting.
10	Portrait of a futuristic robot, pencil sketch by Leonardo da Vinci.

Table 5: **Computational resources and training cost.** The table details the GPU hardware and total GPU hours consumed in each training stage.

Training Stage	GPUs	GPU Hours
Stage 1	256 H800	55,296
Stage 2	256 H800	30,720
Stage 3	64 A100	1,536

Table 6: **Effect of confidence threshold on MathVista.** We report the accuracy (%) and throughput (tokens/s) across varying confidence thresholds, with the block length fixed at 64.

Confidence Threshold (τ)	0.2	0.4	0.6	0.8	0.9	1.0
Accuracy (%)	57.8	62.4	64.2	65.0	65.9	65.8
Throughput (tokens/s)	203.9	134.5	97.3	64.1	52.2	24.3

A.5 Text Generation Process of LLaDA-o

During inference, we adopt a blockwise sampling procedure (Arriola et al., 2025). As outlined in Algorithm 2, we first cache the fixed prefix (images and prompt). Subsequently, we extend the sequence by appending a block of length L , initialized with mask tokens. Within each block, we perform iterative denoising: tokens with prediction confidence exceeding a threshold τ are accepted, while others remain masked for the next step. If an [EOS] token is detected, we truncate the sequence and terminate; otherwise, we update the cache with the completed block and proceed to the next masked block.

B Additional Results

B.1 Further Analysis

Tab. 6 and Fig. 4 analyze the inference efficiency and generation quality. Unlike autoregressive models, LLaDA-o offers the flexibility to regulate this trade-off via the confidence threshold: raising the threshold prioritizes accuracy by selecting only high-confidence tokens, while lowering it accelerates generation. Empirically, a threshold of 0.9 achieves the optimal balance. When comparing with LLaDA-V on MathVista/A100 (Fig. 4), LLaDA-o demonstrates significant efficiency gains. Most notably, LLaDA-o delivers a $5.9\times$ speedup compared to LLaDA-V. This improvement reflects the complete LLaDA-o pipeline, whose intra-modality bidirectional attention reduces computational redundancy; its isolated contribution is not measured here.

Algorithm 2 Text Generation Process of LLaDA-o.

```
1: Input: Parameters  $\theta$ ; Images  $\mathcal{V}$ ; Prompt  $\mathcal{P}$ ; Block size  $L$ ; Steps per block  $S$ ; Threshold  $\tau$ .
2: Output: Generated sequence  $\mathbf{y}$ .
3:  $\mathcal{C} \leftarrow \text{ENCODE}(\mathcal{V}, \mathcal{P}; \theta)$ ;  $\mathbf{y} \leftarrow \emptyset$  {Initial Context Encoding}
4: loop
5:    $\mathbf{b} \leftarrow [\text{MASK}]^L$ ;  $\mathcal{M} \leftarrow \{1, \dots, L\}$  {Init Block & Mask indices}
6:    $S_{\max} \leftarrow 2S$  {Maximum iterations for this block}
7:   for  $j = 1, \dots, S_{\max}$  do
8:     if  $\mathcal{M} = \emptyset$  then
9:       break
10:    end if
11:     $\mathbf{p} \leftarrow f_{\theta}(\mathbf{b}, \mathcal{C})$  {Forward pass using parameters  $\theta$ }
12:     $\mathcal{A} \leftarrow \emptyset$  {Accepted positions in this step}
13:    for  $i \in \mathcal{M}$  do
14:      if  $\max(\mathbf{p}_i) \geq \tau$  then
15:         $\mathbf{b}_i \leftarrow \arg \max(\mathbf{p}_i)$  {Keep high-confidence tokens}
16:         $\mathcal{M} \leftarrow \mathcal{M} \setminus \{i\}$ 
17:         $\mathcal{A} \leftarrow \mathcal{A} \cup \{i\}$ 
18:      end if
19:    end for
20:    if  $\mathcal{A} = \emptyset$  then
21:       $i^* \leftarrow \arg \max_{i \in \mathcal{M}} \max(\mathbf{p}_i)$  {Fallback to ensure progress}
22:       $\mathbf{b}_{i^*} \leftarrow \arg \max(\mathbf{p}_{i^*})$ 
23:       $\mathcal{M} \leftarrow \mathcal{M} \setminus \{i^*\}$ 
24:    end if
25:  end for
26:  if EOS  $\in \mathbf{b}$  then
27:    return  $\mathbf{y} \parallel \text{TRUNC}(\mathbf{b}, \text{EOS})$ 
28:  end if
29:   $\mathbf{y} \leftarrow \mathbf{y} \parallel \mathbf{b}$ ;  $\mathcal{C} \leftarrow \text{UPDATECACHE}(\mathcal{C}, \mathbf{b})$ 
30: end loop
```

Table 7: **Effect of block length on MathVista.** We report the average number of generated tokens under different block lengths, with the confidence threshold fixed at 0.95.

Block length	32	64	96	128
Average tokens	165	148	145	154
Accuracy (%)	63.6	66.1	66.2	65.3

We provide comprehensive analysis of the variable-length behavior in Appendix B.3, B.4 and Tab. 7. Qualitatively, unlike LLaDA-V, which produces redundant or incomplete text depending on mismatched block settings, LLaDA-o generates content of appropriate length consistent across varying block sizes ($L \in \{16, \dots, 128\}$). Quantitatively, Tab. 7 reveals that the generated length remains relatively stable: increasing the block length

Table 8: **Text-to-image generation performance across training stages.** We report the evaluation results on GenEval and DPG-Bench to demonstrate the performance progression. In this table, we use the original prompts of GenEval.

Training Stage	GenEval	DPG-Bench
Stage 1	0.74	86.1
Stage 2	0.78	86.2
Stage 3	0.82	87.0

from 32 to 96 results in only a minor decrease in average token count (165 to 145) while improving accuracy (63.6% to 66.2%). These results demonstrate that the output length is mainly driven by the input content rather than the preset block size, confirming the effectiveness of our *adaptive length augmentation*.

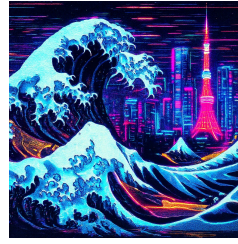
Finally, Tab. 8 studies the effect of each train-



A giant astronaut sitting on top of the Golden Gate Bridge fishing in the clouds, dreamlike atmosphere, vintage travel poster style, pastel colors, matte painting.



A cute tiny robot holding a red balloon, stop-motion claymation style, fingerprints visible on the clay, studio lighting, depth of field, plasticine texture.



The Great Wave off Kanagawa in a cyberpunk style, neon lights glowing on the waves, futuristic Tokyo skyline in the background, traditional woodblock print texture mixed with digital glitch art.



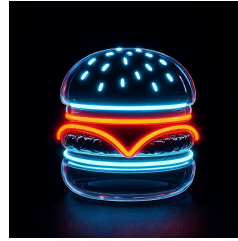
Color photo of a corgi made of transparent glass, standing on the riverside in Yosemite National Park.



A snail with a disco ball as its shell, macro photography, realistic.



A running horse formed by splashing blue water, frozen motion photography.



A transparent glass hamburger with neon lights glowing inside, dark background.



A blooming rose made of colorful smoke, black background.



A typewriter made of clear glass, with tiny, living fireflies trapped inside as the "letters", on a writer's wooden desk, soft glow.



An octopus playing a drum set in a jazz club, cinematic lighting.



Macro photography of a dewdrop on a spider web, reflecting a whole galaxy inside the droplet, extreme close-up, high detail, bokeh background, magical atmosphere.



A majestic horse made of splashing liquid water, galloping on a sea of fire, high speed photography, frozen motion, crystal clear droplets, cinematic lighting, strong contrast between blue and orange.

Figure 5: **Additional generated samples.** We present 12 randomly selected images generated by LLaDA-o. For each sample, the prompt used for generation is shown below the corresponding image. All results are produced under the same setting as in the main paper.

ing stage on text-to-image generation. Results on GenEval and DPG-Bench improve from Stage 1 to Stage 3, with Stage 3 performing best (0.82 and 87.0), supporting the effectiveness of our multi-stage training pipeline. For completeness, we report the computational cost of our three-stage training pipeline in Appendix A.4.

B.2 Additional Generated Images

We provide additional examples generated by LLaDA-o in Fig. 5. These samples demonstrate the

model’s capability to produce high-quality images that are semantically aligned with user prompts, exhibiting both high fidelity and diversity.

B.3 Qualitative Comparison with LLaDA-V under Mismatched Block Lengths

In this section, we present qualitative samples demonstrating the variable-length text generation capability of LLaDA-o. As shown in Tab. 9, we compare LLaDA-o with the strong LLaDA-V baseline in scenarios where the semantic requirement

of the prompt conflicts with the pre-defined generation block length.

First, as shown in the top row of Tab. 9, when the user requests simple text extraction but provides a long block length (i.e., $L = 64$), LLaDA-V tends to fill the entire window with redundant content. In contrast, LLaDA-o accurately extracts the text and correctly terminates, adhering to the user’s intent. Second, as shown in the bottom row, when the user requests a detailed image description but assigns a short block length (i.e., $L = 16$), LLaDA-V is constrained by the fixed window, resulting in an overly brief response. Conversely, LLaDA-o automatically extends the generation by appending additional blocks if no End-of-Sequence (EOS) token is detected within the current block, continuing until the generation is complete.

B.4 Effect of Block Length on Variable-Length Generation

In this section, we present qualitative samples to illustrate the variable-length generation behavior of LLaDA-o. As shown in Tab. 10, we maintain the same image and prompt while varying the pre-defined generation block length $L \in \{16, 32, 64, 128\}$. Across these different settings, the generated outputs remain largely consistent, with only minor fluctuations in response length. This suggests that LLaDA-o is not rigidly constrained by the specific block length and can dynamically adapt the effective output length based on the user prompt and image content.

Table 9: **Qualitative samples demonstrating variable-length generation.** **Top:** Under a long block length setting ($L = 64$), LLaDA-V generates redundant content for a simple prompt, while LLaDA-o correctly terminates. **Bottom:** Under a short block length setting ($L = 16$), LLaDA-V produces an overly brief response for a detailed prompt, whereas LLaDA-o dynamically adapts the output length to the user prompt.

User	What’s written on this image, just give me the answer.
	
LLaDA-V	The image displays a lightboard sign with the phrase “TIME FOR CHANGE” written in bold, colorful letters. The sign is set against a backdrop of blurred lights, suggesting a festive or celebratory atmosphere. The text on the sign reads “TIME FOR CHANGE,” indicating a call to action or a message for transformation.
LLaDA-o	TIME FOR CHANGE
User	Please describe this image in detail.
	
LLaDA-V	The image captures a serene landscape featuring a tranquil lake surrounded by towering mountains.
LLaDA-o	The image captures a serene mountain landscape, likely a popular tourist destination. Dominating the scene is a large lake, its surface a vibrant shade of green, reflecting the surrounding mountains. The lake’s edge is lined with a wooden fence, adding a rustic charm to the scene. Beyond the fence, a grassy area stretches out, dotted with trees and a small building, possibly a shelter or a visitor center. The building is nestled among the trees, blending seamlessly with the natural surroundings. The mountains rise majestically in the background, their peaks dusted with snow, suggesting a high-altitude location. The sky above is a clear blue, adding to the overall tranquility of the scene. The image does not contain any visible text.

Table 10: **Qualitative samples demonstrating variable-length generation.** We show generations from LLaDA-o for the same image and the same prompt under different block length settings ($L \in \{16, 32, 64, 128\}$). Despite changing L , the output length remains largely stable, indicating that LLaDA-o can dynamically adapt the output length to the user prompt and image.

User	Please describe this image in detail.
	
LLaDA-o (block:16)	The image captures a serene mountain landscape, likely a popular tourist destination. Dominating the scene is a large lake, its surface a vibrant shade of green, reflecting the surrounding mountains. The lake's edge is lined with a wooden fence, adding a rustic charm to the scene. Beyond the fence, a grassy area stretches out, dotted with trees and a small building, possibly a shelter or a visitor center. The building is nestled among the trees, blending seamlessly with the natural surroundings. The mountains rise majestically in the background, their peaks dusted with snow, suggesting a high-altitude location. The sky above is a clear blue, adding to the overall tranquility of the scene. The image does not contain any visible text.
LLaDA-o (block:32)	The image depicts a serene, picturesque natural landscape featuring a tranquil lake surrounded by rugged mountains and lush greenery. The lake is a vibrant shade of turquoise, reflecting the clear blue sky above. A wooden fence runs along the edge of the lake, separating the grassy area from the water. The fence is simple and rustic, with vertical wooden posts and horizontal rails. In the foreground, there is a well-maintained grassy area with a wooden fence. The grass is lush and green, indicating a healthy, natural environment. A few yellow flowers are scattered throughout the grass, adding a touch of color to the scene. The overall atmosphere of the image is peaceful and inviting, suggesting a perfect spot for relaxation and enjoyment of the beauty of nature.
LLaDA-o (block:64)	The image depicts a serene natural landscape featuring a small, clear lake surrounded by lush green grass and tall evergreen trees. The lake is nestled in a valley, with rugged, rocky mountains rising in the background. The sky is clear and blue, indicating a sunny day. In the foreground, there is a wooden fence that runs along the edge of the grassy area, separating it from the lake. A small wooden structure, possibly a cabin or a small house, is visible on the right side of the image, partially obscured by the tall trees. The overall scene is peaceful and idyllic.
LLaDA-o (block:128)	The image captures a serene mountain landscape, featuring a small, clear lake nestled in a valley. The lake is surrounded by lush green grass and trees, creating a stark contrast with the rocky mountains in the background. The mountains, with their rugged peaks, rise majestically against the clear blue sky. The perspective of the image is from a low angle, looking up at the mountains, which adds a sense of grandeur and majesty to the scene. The overall mood of the image is peaceful and tranquil, inviting the viewer to immerse themselves in the natural beauty of this mountainous paradise.