arXiv:2505.16933v2 [cs.LG] 4 Jun 2025

LLaDA-V: Large Language Diffusion Models with Visual Instruction Tuning

Zebin You^{1,2,3}*, **Shen Nie**^{1,2,3}, **Xiaolu Zhang**^{4}, **Jun Hu**^{4}, **Jun Zhou**^{4},
Zhiwu Lu^{1,2,3}, **Ji-Rong Wen**^{1,2,3}, **Chongxuan Li**^{1,2,3†}
^{1} Gaoling School of AI, Renmin University of China ^{2} Beijing Key Laboratory of Research on Large Models and Intelligent Governance ^{3} Engineering Research Center of Next-Generation Intelligent Search and Recommendation, MOE ^{4} Ant Group

Abstract

In this work, we introduce **LLaDA-V**, a purely diffusion-based Multimodal Large Language Model (MLLM) that integrates visual instruction tuning with masked diffusion models, representing a departure from the autoregressive paradigms dominant in current multimodal approaches. Built upon LLaDA, a representative large language diffusion model, LLaDA-V incorporates a vision encoder and MLP connector that projects visual features into the language embedding space, enabling effective multimodal alignment. Our empirical investigation reveals several intriguing results: First, LLaDA-V demonstrates promising multimodal performance despite its language model being weaker on purely textual tasks than counterparts like LLaMA3-8B and Qwen2-7B. When trained on the same instruction data, LLaDA-V is highly competitive to LLaMA3-V across multimodal tasks with better data scalability. It also narrows the performance gap to Qwen2-VL, suggesting the effectiveness of its architecture for multimodal tasks. Second, LLaDA-V achieves state-of-the-art performance in multimodal understanding compared to existing hybrid autoregressive-diffusion and purely diffusion-based MLLMs. Our findings suggest that large language diffusion models show promise in multimodal contexts and warrant further investigation in future research. Project page and codes: `https://ml-gsai.github.io/LLaDA-V-demo/`.

1 Introduction

Multimodal Large Language Models (MLLMs) are capable of processing multiple input modalities—including images [1–6], audio [7–9], and video [10–12]—alongside text, and can generate natural language responses that follow given diverse instructions. Despite significant advancements in MLLMs, most existing approaches predominantly rely on autoregressive models [13–21], leaving substantial room for exploring alternative probabilistic modeling approaches.

Recent attempts to incorporate diffusion models [22–26] into MLLMs have predominantly adopted one of two strategies: either leveraging autoregressive models to provide strong language modeling capabilities [27–32], or employing discrete diffusion-based approaches with limited language modeling capacity, which consequently leads to suboptimal performance [33, 34].

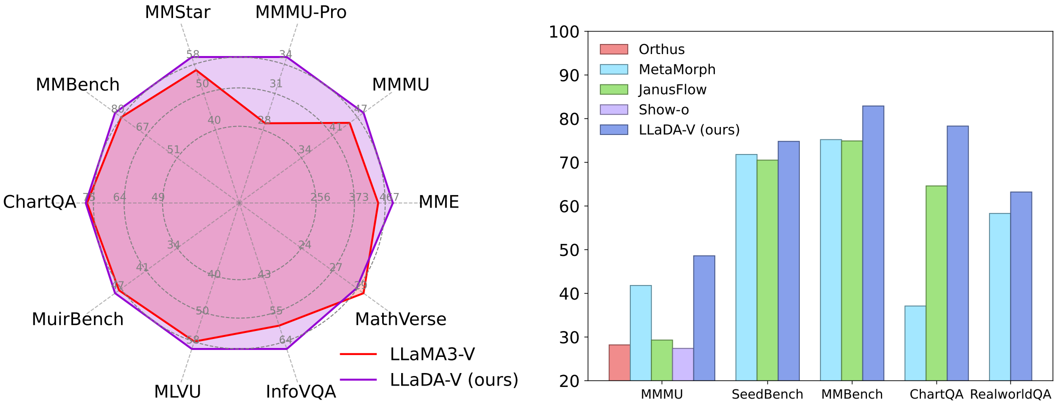
Encouragingly, recent advances in discrete diffusion models [25, 26, 35–43] have shown promising potential to overcome these limitations. In particular, LLaDA [42] has demonstrated performance competitive with LLaMA3-8B-Instruct [18] through large-scale pre-training and SFT, while retaining favorable scaling properties. Nevertheless, while LLaDA has shown remarkable progress in language

---

[*] Work done during an internship at Ant Group.

[†] Correspondence to Chongxuan Li.

Preprint. Under review.



(a) LLaDA-V vs. LLaMA3-V

(b) LLaDA-V vs. non-autoregressive MLLMs

Figure 1: **Benchmark Results.** (a) LLaDA-V demonstrates superior performance on more benchmarks compared to LLaMA3-V when trained on the same dataset, particularly excelling in multidisciplinary knowledge and mathematical reasoning tasks. (b) LLaDA-V achieves state-of-the-art performance in multimodal understanding among both hybrid autoregressive-diffusion (such as MetaMorph [31] and Show-o [28]) and purely diffusion-based models.

modeling, its capabilities and potential in multimodal understanding remain largely unexplored. Therefore, this naturally raises a key research question: Can an purely diffuision based MLLM (both training and sampling) achieve performance compared to autoregressive-based models?

In this paper, we explore how to effectively extend large language diffusion models to encompass strong multimodal understanding capabilities, focusing on the visual instruction tuning framework [1], which has demonstrated remarkable effectiveness across various autoregressive-based MLLMs. In particular, we introduce a vision encoder (e.g., SigLIP 2 [44]) and an MLP connector to map visual features into the LLaDA language embedding space, allowing joint processing of visual and textual inputs. Furthermore, we extend LLaDA's training objective to handle multi-turn multimodal dialogues, investigate various attention mechanism structures, adapt inference procedures for multimodal conversations, and develop a multi-stage training strategy. These comprehensive investigations result in **LLaDA-V**, a purely diffusion-based MLLM.

We first compare the data scalability of LLaDA-V to that of LLaMA3-V (our autoregressive baseline with LLaMA3-8B as the language tower) by varying the amount of instruction tuning data. LLaDA-V demonstrates stronger data scalability on several benchmarks, particularly excelling in tasks involving multidisciplinary knowledge and mathematical reasoning (see Fig. 3).

Furthermore, we benchmark LLaDA-V against autoregressive, hybrid autoregressive-diffusion, and pure diffusion models across 18 diverse multimodal tasks. Notably, when comparing with LLaMA3-V, we observe an interesting and promising phenomenon: despite with a slightly weaker language tower, our model achieves superior performance across 11 tasks (see partial results in Fig. 1 (a) and more details in Section 4). Similarly, when compared to the powerful autoregressive Qwen2-VL [5], despite LLaDA being considerably weaker than Qwen2-7B, LLaDA-V narrows the performance gap significantly, achieving comparable results on some benchmarks such as MMStar [45] (60.1 vs. 60.7). Furthermore, our model achieves state-of-the-art performance compared to existing hybrid autoregressive-diffusion models [28–32] and pure diffusion models [33, 34](see Fig. 1 (b)). Collectively, all these findings demonstrate not only the effectiveness of the LLaDA-V framework but also the promise of diffusion models on multimodal understanding.

In summary, our key contributions are as follows:

- We introduce LLaDA-V, a purely diffusion-based MLLM for multimodal understanding.
- We demonstrate that LLaDA-V benefits from data scaling and achieves superior scalability across multiple benchmarks when compared to our autoregressive baseline, LLaMA3-V.
- LLaDA-V achieves state-of-the-art results among both hybrid and purely diffusion-based MLLMs.

2 Preliminaries

In this section, we briefly introduce large language diffusion models, which serve as the language tower in our work, and visual instruction tuning, which forms the basis of our multimodal framework.

Large Language Diffusion Models. Large language models (LLMs) are currently experiencing rapid development. The predominant LLMs [13–19] are primarily trained using autoregressive modeling. Unlike autoregressive approaches, discrete diffusion models [22, 25] offer an alternative paradigm for language modeling. Masked diffusion models [26, 35], a specific variant of discrete diffusion, have shown impressive results across multiple domains [37–42, 46–48, 43].

Among them, LLaDA [42] has demonstrated comparable performance with strong AR models like LLaMA3-8B-Instruct [18], while maintaining the unique properties of masked diffusion models. Specifically, LLaDA employs a masked diffusion process that differs fundamentally from autoregressive approaches. Formally, let $\boldsymbol{x}_0 = [\boldsymbol{x}^i]_{i=1}^N$ represent a sentence comprising N tokens, and let [M] denote a special mask token. LLaDA defines a model distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x}_0)$ through a forward and a reverse process. In the forward process, LLaDA first samples a time step t uniformly from the interval $[0, 1]$. Subsequently, each token in $\boldsymbol{x}_0$ is replaced by [M] with probability t, yielding the corrupted sentence $\boldsymbol{x}_t$. In the reverse process, LLaDA commences with a sentence composed entirely of [M] tokens and iteratively predicts these masked tokens to reconstruct the original sentence. We provide detailed formulations and sampling processes of masked diffusion models in Appendix A.

Visual Instruction Tuning [1–3] is a mainstream Multimodal Large Language Model (MLLM) architecture, recognized for its powerful performance and data efficiency. Specifically, it comprises a vision tower (e.g., CLIP [49] or SigLIP [50, 44]) that converts images into visual representations, an MLP connector that projects these representations into an LLM's word embedding space, and the LLM itself. Through visual instruction tuning, this setup enables LLMs to achieve strong multimodal understanding capabilities with less than 1M image-text pairs.

3 Method

In contrast to predominant approaches that rely on autoregressive language models [1–6], our research explores how to perform visual instruction tuning [1] in language diffusion models [42] for multimodal understanding. To this end, we formulate a training objective for multi-turn multimodal dialogues and explore the attention mechanism architectures (Sec. 3.1), detail the inference process (Sec. 3.2), and design a multi-stage training strategy (Sec. 3.3). These components collectively enable diffusion language models to effectively process multimodal inputs.

3.1 Training Objective and Architecture

As with most MLLMs, the training of LLaDA-V utilizes multimodal understanding data involving multi-turn dialogues. For simplicity, we use a sample consisting of a single image and a two-turn dialogue as an example. As LLaDA-V represents an early exploration into applying large language diffusion models for multimodal understanding, its design prioritizes simplicity, effectiveness, and alignment with established training methodologies of autoregressive-based MLLMs. Consequently, we adopt the seminar visual instruction tuning framework [3], comprising a language tower, a vision tower, and an MLP projector. For the language tower, we selected LLaDA [42], a representative large language diffusion model with language performance comparable to LLaMA3-8B, enabling us to explore the capabilities of purely diffusion-based MLLMs. For the vision tower and MLP projector, we selected SigLIP 2 [44] and a two-layer MLP, respectively, due to their demonstrated effectiveness across various MLLMs.

For training the aforementioned models within LLaDA-V, we now present the necessary notations and training objective. Let $\boldsymbol{v}$ denote the image representation from the vision tower and MLP projector and [M] denote a special mask token. For a two-turn dialogue, we denote the data instance as $(\boldsymbol{v}, \boldsymbol{p}_0^1, \boldsymbol{r}_0^1, \boldsymbol{p}_0^2, \boldsymbol{r}_0^2)$, where $\boldsymbol{p}_0^1 = [\boldsymbol{p}_0^{1,i}]_{i=1}^{L_{p1}}$ and $\boldsymbol{p}_0^2 = [\boldsymbol{p}_0^{2,i}]_{i=1}^{L_{p2}}$ are the prompts for the first and second turns, while $\boldsymbol{r}_0^1 = [\boldsymbol{r}_0^{1,i}]_{i=1}^{L_{r1}}$ and $\boldsymbol{r}_0^2 = [\boldsymbol{r}_0^{2,i}]_{i=1}^{L_{r2}}$ are their corresponding ground-truth responses.

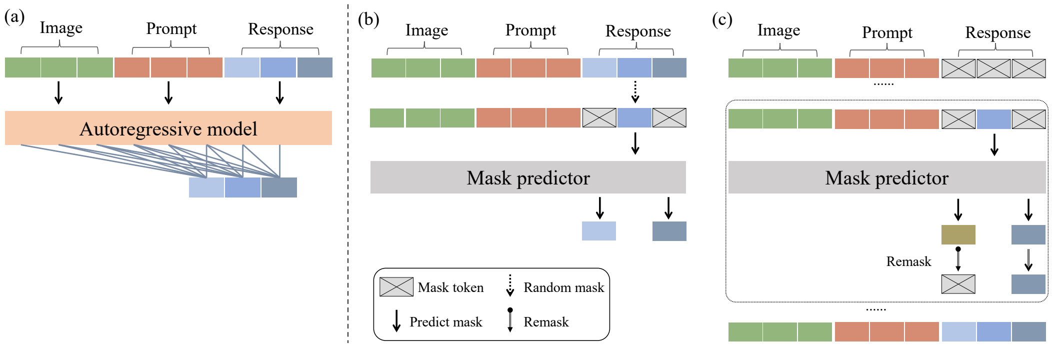


Figure 2: **Overview of Autoregressive Approaches and LLaDA-V.** Image representations are generated by an encoder and an MLP projector (not explicitly shown). (a) Autoregressive Training: Given image features and the input prompt, autoregressive models are trained to predict the response through next-token prediction. (b) LLaDA-V's Training: Image features and the input prompt remain unmasked, while only the response is randomly masked. (c) LLaDA-V's Inference: As time step t decreases from 1 to 0, generation begins with a fully masked response and iteratively predicts tokens.

Formally, the training objective for LLaDA-V, $\mathcal{L}(\boldsymbol{\theta})$, is defined as:

$$-\mathbb{E}_{\substack{\boldsymbol{v},t,\boldsymbol{p}_0^1,\boldsymbol{r}_0^1,\boldsymbol{r}_t^1,\\ \boldsymbol{p}_0^2,\boldsymbol{r}_0^2,\boldsymbol{r}_t^2}} \left[\frac{1}{t}\sum_{i=1}^{L_{r1}}\sum_{j=1}^{L_{r2}} \mathbf{1}[\boldsymbol{r}_t^{1,i}=[\mathrm{M}] \wedge \boldsymbol{r}_t^{2,j}=[\mathrm{M}]]\log p_{\boldsymbol{\theta}}(\boldsymbol{r}_0^{1,i},\boldsymbol{r}_0^{2,j}|\boldsymbol{v},\boldsymbol{p}_0^1,\boldsymbol{r}_t^1,\boldsymbol{p}_0^2,\boldsymbol{r}_t^2)\right], \tag{1}$$

where $\boldsymbol{r}_t^1$ and $\boldsymbol{r}_t^2$ denote the masked response.

Theoretically, the training objective in Eq. (1) has been proven to be an upper bound of the negative log-likelihood for masked tokens [41, 39]. Intuitively, as shown in Fig. 2 (b), the training objective aims to predict masked tokens within the response, given clean image features and prompts. Through Eq. (1) and visual instruction tuning framework [1], we effectively extend the large language diffusion model to encompass multimodal understanding capabilities.

Regarding the architecture of LLaDA-V, our primary focus is on exploring the attention mechanism design within the language tower. To mitigate the potential gap between training and inference (see Sec. 3.2), one might expect to use a causal attention structure during training for multi-turn dialogues (i.e., preventing an earlier turn like $\boldsymbol{p}_0^1, \boldsymbol{r}_0^1$ from accessing a later turn such as $\boldsymbol{p}_0^2, \boldsymbol{r}_0^2$). However, a bidirectional attention mechanism enables comprehensive understanding of the entire dialogue context during mask prediction, which has demonstrated its effectiveness in recent video diffusion models [51–53] where it enhances temporal consistency of generated video. Therefore, we conduct ablation studies on these two attention mechanism choices in Sec. 4.4, and observe that the bidirectional attention mechanism achieves superior results across more benchmarks. Based on these findings, we adopt the bidirectional attention mechanism in LLaDA-V.

3.2 Inference Process

Once the model is trained with the objective in Eq. (1), LLaDA-V can generate multi-turn dialogues through iterative response generation. When given a new prompt, the model leverages previous prompts and responses to generate appropriate subsequent responses. While dialogue generation proceeds turn by turn, LLaDA-V differs by generating each response via the reverse process of a masked diffusion model, rather than next-token prediction used in autoregressive models.

As shown in Fig. 2 (c), we illustrate the inference process using a one-turn dialogue example. Following this process, we generate samples from the distribution $p_{\boldsymbol{\theta}}(\boldsymbol{r}_0|\boldsymbol{v},\boldsymbol{p}_0)$ by initializing with a fully masked response $\boldsymbol{r}_1$ and applying the reverse process of the masked diffusion model, as detailed in Appendix A. Sampling starts by setting a target generation length and initializing the response $\boldsymbol{r}_1$ entirely with [M] tokens. The sequence is iteratively refined by transitioning from a state $\boldsymbol{r}_t$ to a state $\boldsymbol{r}_s$ (representing decreasing mask levels, with $s < t$). Each step involves two main phases: first, LLaDA-V, conditioned on $\boldsymbol{v}, \boldsymbol{p}_0$, and $\boldsymbol{r}_t$, predicts all [M] tokens in $\boldsymbol{r}_t$. Second, to form $\boldsymbol{r}_s$, a fraction s/t of these predictions are re-masked to [M], while the remainder $(1 - s/t)$ are kept as predicted, consistent with the reverse process of masked diffusion models. For the remasking strategy, rather than using standard random selection, we primarily adopt LLaDA's [42] *low-confidence* strategy, which preferentially re-masks low-confidence predictions while preserving high-confidence ones. We choose this approach based on its consistently demonstrated improvements across various tasks.

3.3 Training Strategies

We adopt a multi-stage training paradigm for LLaDA-V, with the first two stages following established practices in MLLMs like LLaVA-NeXT [54] to establish language-vision alignment and build visual instruction following abilities. We further enhance this paradigm with a third stage focused on multimodal reasoning, enabling comprehensive capabilities across diverse tasks.

Stage 1: Language-Image Alignment. In this stage, we train the MLP projector to align visual representations with LLaDA's word embeddings, following established MLLM practices [2, 3, 55]. The language and vision towers remain frozen throughout this process. We utilize the LLaVA-Pretrain dataset [1] for this alignment stage.

Stage 2: Visual Instruction Tuning. Following language-image alignment, Stage 2 focuses on developing LLaDA-V's comprehensive multimodal understanding capabilities by fine-tuning the entire model on large-scale instruction data. This fine-tuning, which utilizes high-quality, large-scale multimodal instruction data from MAmmoTH-VL [55], aims to establish strong visual instruction-following abilities and enable the model to handle diverse scenarios involving single images, multiple images, or video inputs. Stage 2 is conducted in two distinct phases as follows.

- *Single Image Training*: The model is trained on 10M single-image multimodal data to establish image understanding capabilities. In this phase, LLaDA-V develops strong performance in recognizing and interpreting single images to respond to diverse instructions.
- *OneVision Training*: Following single-image training, the model is further trained on approximately 2M diverse multimodal samples (single-image, multi-image, and video data). This phase expands LLaDA-V's capabilities to handle complex scenarios involving multiple images and temporal information beyond single-image contexts.

Stage 3: Multimodal Reasoning Enhancement. Following visual instruction-following, Stage 3 focuses on enhancing multimodal reasoning capabilities for complex tasks through two key steps:

- *Reasoning Training*: In this step, we trained LLaDA-V on reasoning-focused multimodal data from VisualWebInstruct [56], which contains 900K QA pairs featuring detailed reasoning chains and final answers. This training phase is designed to enhance the model's ability to perform complex multimodal reasoning.
- *Balanced Reasoning Training*: Following *reasoning training*, LLaDA-V consistently provided explicit reasoning before answers. To enhance response flexibility, a subsequent phase, inspired by Qwen 3's hybrid thinking mechanism [57], utilized a mixed dataset: reasoning-focused VisualWebInstruct combined with MAmmoTH-VL's OneVision data. In this mixed training, '/no_think' tags were appended to OneVision prompts to encourage direct answers, while '/think' tags were applied to 50% of reasoning-data prompts.

4 Experiment

This section presents our experimental setup and results, including: experimental settings (Sec. 4.1); data scaling experiments (Sec. 4.2); comprehensive benchmark evaluations (Sec. 4.3); and ablation studies on attention mask selection (Sec. 4.4).

4.1 Experimental Settings

Model. We use LLaDA-8B-Instruct [42] for the language tower of LLaDA-V, an open-source diffusion-based large language model with extensive pre-training and supervised fine-tuning (SFT). However, it lacks preference alignment techniques [58–61] that enhance conversational and reasoning capabilities in contemporary LLMs [17, 18]. Consequently, its performance falls behind Qwen2.5-7B-Instruct [19] and is marginally inferior to LLaMA3-8B-Instruct [18]. For a fair comparison between

Table 1: **Training Settings.** Here M-SI and M-OV represent the single image data and onevision data of MAmmoTH [55], while VW represents the data of VisualWebInstruct [56]. We train LLaDA-V sequentially through the first five datasets (LLaVA-Pretrain [1], M-SI, M-OV, VW, and M-OV+VW), while the last dataset (LLaVA-NeXT [54]) is used for ablation study in Sec. 4.4.

Training data	LLaVA-Pretrain	M-SI	M-OV	VW	M-OV+VW	LLaVA-NeXT
Vision tower	Siglip2-so400m-patch14-384 [44]					
Language tower	LLaDA-8B-Instruct [42]					
Attention	Bidirectional attention					
Batch size	64	256	256	256	256	64
Model max length	8192	8192	16384	8192	16384	8192
#Samples	558K	10M	2M	900K	3M	738K
LR of vision tower	-	2×10^{-6}		2×10^{-6}		2×10^{-6}
LR of language tower	-	1×10^{-5}		1×10^{-5}		1×10^{-5}
LR of projector	1×10^{-3}	1×10^{-5}		1×10^{-5}		1×10^{-5}
Epoch	1	1		1		1

LLaDA-V and autoregressive approaches, we use LLaMA3-8B-Instruct as the language tower in our primary baseline model, while maintaining all other components identical to LLaDA-V. For the vision tower, we utilize siglip2-so400m-patch14-384 [44], which offers robust visual representation capabilities. The projector is implemented as a randomly initialized two-layer MLP.

Data. For Stage 1, we employ the alignment dataset from LLaVA-Pretrain [1]. In Stage 2, we leverage the comprehensive MAmmoTH-VL [55] dataset, which consists of two primary components: SI-10M, comprising 10 million single-image multimodal samples, and OV-2M, containing 2 million diverse samples across single-image, multi-image, and video modalities. For Stage 3, we utilize the reasoning-focused multimodal dataset VisualWebInstruct [56]. To achieve balanced reasoning capabilities, we further incorporate OV-2M into this stage of training. A comprehensive description of these training strategies can be found in Section 3.3.

Training. As detailed in Sec. 3.3, the LLaDA-V training process consists of three stages. In the first stage, only the Projector is trained. Subsequently, the full model is trained during the second and third stages. Detailed training settings can be found in Tab. 1.

Evaluation. To comprehensively evaluate LLaDA-V's performance, we considered multiple vision-language benchmarks across several categories:

- *Multidisciplinary Knowledge & Mathematical Reasoning:* MMMU [62], MMMU-Pro [63], MMStar [45], MME [64], SeedBench [65], MMBench [66], MathVerse [67], and MathVista [68].
- *Chart & Doc Understanding:* AI2D [69], ChartQA [70], DocVQA [71], and InfoVQA [72].
- *Real-world Scene Understanding:* RealworldQA [73].
- *Multi-image & Video Understanding:* MuirBench [74], MLVU [75], and VideoMME [76].

4.2 Data Scalability of LLaDA-V

In order to demonstrate the effectiveness of LLaDA-V, we first evaluate the data scalability of LLaDA-V in comparison with the autoregressive baseline LLaMA3-V. To ensure a fair comparison between LLaDA-V and LLaMA3-V, we adopted an identical training pipeline for both models. The training process consisted of two main phases: first, we pretrained the projectors using LLaVA-Pretrain data [1]; then, we conducted full model training (including vision tower, language tower, and projector) on the single-image data of MAmmoTH-VL [55]. We evaluated the models' performance at various data scales using six carefully selected multimodal benchmarks.

As shown in Fig. 3, we observe two key findings: First, LLaDA-V's performance consistently improves with increasing training data, demonstrating that LLaDA-V benefits from data scalability. Second, despite using a slightly weaker language tower, LLaDA-V shows superior scalability compared to LLaMA3-V on multidisciplinary knowledge benchmarks such as MMMU [62] and

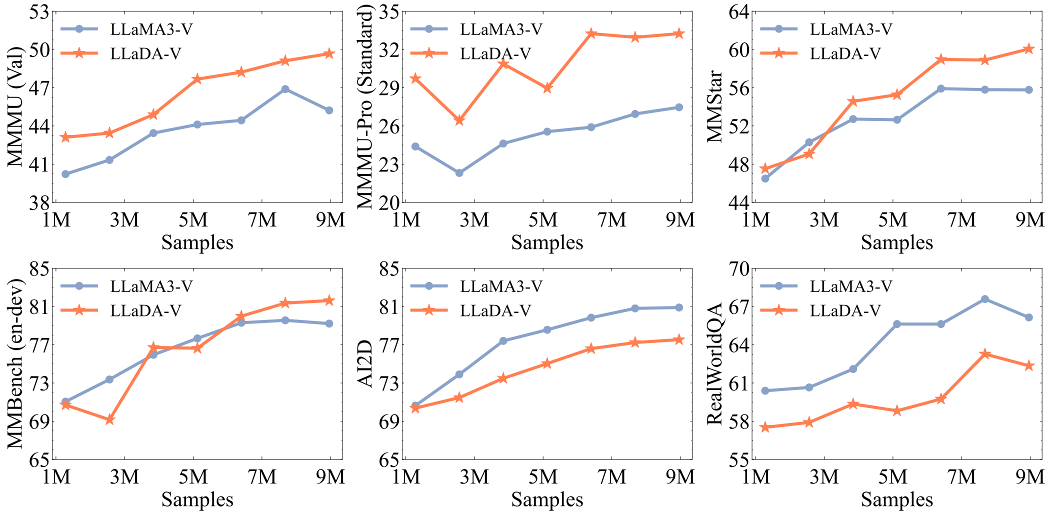


Figure 3: **Data Scalability of LLaDA-V.** Both LLaDA-V and LLaMA3-V were trained on MAmmoTH-VL-SI10M, with performance evaluated across six multimodal benchmarks. Despite having a weaker language tower, LLaDA-V shows superior data scalability across more tasks, especially excelling in multidisciplinary knowledge and mathematical reasoning.

Table 2: **Benchmark Results for Multidisciplinary Knowledge and Mathematical Reasoning Tasks.** "Diffusion" here encompasses both continuous and discrete diffusion models. *Notably, LLaDA-V outperforms all other hybrid and pure diffusion MLLMs, surpassing LLaMA3-V on 6 of 9 benchmarks despite having a relatively weaker language tower.* For comparison, we list each model's language tower, as this significantly impacts MLLM performance. "-" indicates unavailable data.

Model	Type	LLM Tower	MMMU val	MMMU-Pro standard	MMMU-Pro vision	MMStar test	MME cog./perp.	SeedB image	MMB en-dev	MathVerse mini-vision	MathVista testmini
ShareGPT4V[77]	AR	Vicuna-7B	-	-	-	-	376/1567	69.7	68.8	-	-
Cambrian-1[78]	AR	LLaMA3-8B	42.7	-	-	-	-/1547	74.7	75.9	-	49.0
LLaVA[1]	AR	Vicuna-7B	-	-	-	-	-/809	37.0	38.7	-	-
LLaVA-1.5[2]	AR	Vicuna-7B	-	-	-	-	-/1510	66.1	64.3	-	-
Qwen2-VL[5]	AR	Qwen2-7B	54.1	43.5	-	60.7	-	-	-	-	58.2
DeepSeek-VL[79]	AR	DeepSeek-7B	36.6	-	-	-	-	70.4	73.2	-	-
DeepSeek-VL2[80]	AR	-	51.1	-	-	61.3	-	-	-	-	62.8
Janus[81]	AR	DeepSeek-1.3B	30.5	-	-	-	-/1338	63.7	69.4	-	-
Janus-Pro[82]	AR	DeepSeek-7B	41.0	-	-	-	-/1567	72.1	79.2	-	-
Emu3[83]	AR	-	31.6	-	-	-	-	68.2	58.5	-	-
MAmmoTH[55]	AR	Qwen2.5-7B	50.8	-	25.3	63.0	-	76.0	-	34.2	67.6
LLaVA-OV[3]	AR	Qwen2-7B	48.8	-	-	61.7	418/1580	75.4	80.8	26.2	63.2
MetaMorph[31]	AR+Diff.	LLaMA3.1-8B	41.8	-	-	-	-	71.8	75.2	-	-
Show-o[28]	AR+Diff.	Phi1.5-1.3B	27.4	-	-	-	-/1232	-	-	-	-
JanusFlow[30]	AR+Diff.	DeepSeek-1.3B	29.3	-	-	-	-/1333	70.5	74.9	-	-
Orthus[32]	AR+Diff.	Chameleon-7B	28.2	-	-	-	-/1265	-	-	-	-
D-DiT[34]	Diff.	-	-	-	-	-	-/1124	-	-	-	-
LLaMA3-V	AR	LLaMA3-8B	45.4	28.3	14.5	56.5	446/1581	**76.6**	79.8	**29.0**	**62.1**
LLaDA-V	Diff.	LLaDA-8B	**48.6**	**35.2**	**18.6**	**60.1**	**491/1507**	74.8	**82.9**	28.5	59.7

MMMU-Pro [63]. Notably, for MMMU-Pro, LLaDA-V trained with merely 1M samples outperforms LLaMA3-V trained with 9M samples. However, on benchmarks assessing chart/document understanding (e.g., AI2D) and real-world scene understanding (e.g., RealworldQA), LLaDA-V lags behind LLaMA3-V.

4.3 Benchmark Results

To comprehensively assess LLaDA-V's multimodal understanding capabilities, we evaluated it against three different model architectures—autoregressive, hybrid autoregressive-diffusion, and pure diffusion models—across a diverse set of 18 benchmarks (detailed results in Tab. 2 and Tab. 3). These benchmarks encompass areas such as multidisciplinary knowledge, mathematical reasoning, chart/document understanding, real-world scene understanding, and multi-image/video tasks.

Table 3: **Benchmark Results for Chart, Document, Real-world Scene, Multi-image, and Video Tasks.** “Diffusion” here encompasses both continuous and discrete diffusion models. *Compared to LLaMA3-V, LLaDA-V shows comparable performance on chart/document tasks, performs less well on real-world scenes, but excels in multi-image and video tasks.* “-” indicates missing data.

Model	Type	LLM Tower	AI2D	ChartQA	DocVQA val	InfoVQA val	RealworldQA	SeedB video	MuirBench	MLVU dev	VideoMME
Cambrian-1[78]	AR	LLaMA3-8B	73.0	73.3	-	-	64.2	-	-	-	-
LLaVA[1]	AR	Vicuna-7B	-	-	-	-	-	23.8	-	-	-
LLaVA-1.5[2]	AR	Vicuna-7B	-	-	-	-	-	37.3	-	-	-
Qwen2-VL[5]	AR	Qwen2-7B	83.0	83.0	-	-	70.1	-	-	-	-
DeepSeek-VL2[80]	AR	-	81.4	86.0	-	-	68.4	-	-	-	-
Emu3[83]	AR	-	70.0	68.6	-	-	57.4	-	-	-	-
MAmmoTH[55]	AR	Qwen2.5-7B	84.0	86.2	-	-	69.9	57.1	55.1	64.7	58.8
LLaVA-OV[3]	AR	Qwen2-7B	81.4	80.0	-	-	66.3	56.9	41.8	64.7	58.2
MetaMorph[31]	AR+Diff.	LLaMA3.1-8B	-	37.1	-	-	58.3	-	-	-	-
JanusFlow[30]	AR+Diff.	DeepSeek-1.3B	-	64.6	-	-	-	-	-	-	-
LLaMA3-V	AR	LLaMA3-8B	**81.1**	77.8	**86.2**	58.9	**66.0**	**55.0**	47.4	57.5	55.8
LLaDA-V	Diff.	LLaDA-8B	77.8	**78.3**	83.9	**66.3**	63.2	53.7	**48.3**	**59.5**	**56.1**

Notably, in these comparative evaluations, LLaDA-V consistently demonstrates superior performance among hybrid autoregressive-diffusion and pure diffusion models, such as MetaMorph [31] and D-DiT [34]. Furthermore, when compared with our autoregressive baseline LLaMA3-V, LLaDA-V exhibited strengths in some tasks: it outperformed LLaMA3-V on most multidisciplinary knowledge and mathematical reasoning benchmarks (e.g. MMMU, MMMU-Pro), while also achieving superior performance in multi-image and video understanding tasks (e.g., MuirBench, MLVU). These results are impressive considering LLaDA-V uses a relatively weaker language tower (see results in Tab.2 of [42]). However, its performance remained less competitive on tasks focused on chart/document understanding (e.g., AI2D, DocVQA) and real-world scene comprehension (e.g., RealworldQA). For a fair comparison, our autoregressive baseline LLaMA3-V shares identical training protocols with LLaDA-V (see Sec. 3.3), with the only difference being the language tower.

When compared with the strong autoregressive-based MLLM Qwen2-VL [5], LLaDA-V generally underperforms across most benchmarks, only achieving comparable results on a limited number of specific tasks such as MMStar. The performance difference primarily stems from LLaDA-V’s weaker language backbone (LLaDA-8B) compared to Qwen2-VL’s Qwen2-7B (see results in Tab.2 of [42]), since the language model’s perfomance is crucial for MLLM’s performance [84]. However, as language diffusion models continue to improve, diffusion-based MLLMs are expected to achieve better performance, gradually narrowing the gap with leading models such as Qwen2-VL.

4.4 Ablation Study

We adopt the two-stage training paradigm of LLaVA-NeXT [54] for our ablation study. First, we train the MLP projector on the LLaVA-Pretrain dataset [1], then further fine-tune the entire model on the LLaVA-NeXT dataset [54]. Training hyperparameter details are provided in Tab. 1.

We consider two attention mask strategies: dialogue causal and no mask (i.e., bidirectional attention). In the dialogue causal approach, earlier dialogue turns cannot attend to later turns. Conversely, the no mask strategy employs bidirectional attention, allowing attention across all turns. Further details on these masking architectures are available in Appendix B. As shown in Tab. 4, the no mask strategy achieves superior performance, outperforming on 7 of the 12 benchmarks. We hypothesize that its underlying bidirectional attention mechanism provides a more comprehensive understanding of the entire dialogue context, thus improving model performance. This bidirectional attention mechanism is also widely adopted in recent video diffusion models [51–53] to improve temporal consistency. We thus adopt the no mask strategy in LLaDA-V.

5 Related Work

Diffusion Language Models. Recently, diffusion language models have attracted increasing attention, including both continuous [85–100] and discrete [25, 35, 36, 101–110] variants. Among them, the masked diffusion models, a subclass of discrete diffusion models, have achieved the best performance. Ou et al. [41], Shi et al. [39], Shao et al. [61] established the theoretical foundations

Table 4: **Ablation Studies on Attention Mask.** Comparison of LLaDA-8B using different attention masking strategies (dialogue causal vs. no mask) across 12 benchmarks. We adopt the no mask strategy in LLaDA-V as it shows slightly better performance on most benchmarks.

LLM Backbone Attention Mask	LLaDA [46] Dialogue Causal Mask	LLaDA [46] No Mask
MMMU [62](val)	42.89	**44.67**
MMMU-Pro [63](standard)	26.01	**26.59**
MMMU-Pro [63](vision)	11.56	**11.68**
MMStar [45]	49.60	**49.79**
MME [64](cog./perp.)	**365/1412**	352/1370
SeedBench [65](image)	**72.16**	71.59
SeedBench [65](video)	**45.75**	45.54
MMBench [66](en-dev)	75.42	**76.71**
AI2D [69]	70.89	**71.47**
ChartQA [70]	**55.20**	54.88
RealworldQA [73]	**61.18**	60.26
MuirBench [74]	28.69	**33.88**

of masked diffusion models and demonstrated their competitiveness with autoregressive models at the GPT-2 scale. LLaDA [42] scales masked diffusion models to 8B parameters, making it the first diffusion-based language model that can rival modern LLMs such as LLaMA3 across a wide range of downstream tasks. While LLaDA's language performance remains slightly inferior to LLaMA3-8B, LLaDA-V shows superior performance across more tasks compared to our LLaMA3 baseline. This suggests LLaDA-V's framework may offer inherent advantages for multimodal applications.

Multimodal Understanding. Multimodal Large Language Models (MLLMs) have made significant strides by integrating multiple input modalities with strong Large Language Models (LLMs) [13–21]. From the perspective of the probabilistic modeling methods, MLLMs are primarily classified into three categories: autoregressive models [1–12], autoregressive-diffusion hybrid models [27–30], and pure diffusion models [33, 34]. The most closely related work, D-DiT [34], combines continuous diffusion for visual content with discrete diffusion for text. However, its limited language modeling capacity results in performance that falls significantly behind autoregressive and hybrid approaches. In contrast, LLaDA-V leverages a powerful language diffusion model [42] with an effective training framework to achieve state-of-the-art results among both hybrid and purely diffusion-based MLLMs

6 Conclusion

We present **LLaDA-V**, a purely diffusion-based Multimodal Large Language Model (MLLM) for both training and sampling, which builds upon the visual instruction tuning framework [1] and the large language diffusion model [42]. LLaDA-V demonstrates superior performance among hybrid autoregressive-diffusion and purely diffsion-based model. Besides, LLaDA-V achieve better data scalability and performance across more benchmarks than LLaMA3-V, which employs a different language tower but shares the same training strategy. We effectively extend the large language diffusion model to encompass multimodal understanding capabilities.

Limitations. A limitation of our work is the image processing strategy. For high-resolution images, we split and resize image segments, process them through our SigLIP2 [44] vision tower, and concatenate the features. Unlike Qwen2-VL with native dynamic resolution support, this approach may reduce efficiency and accuracy in visual representation. We leave the development of more advanced image processing strategies for future work.

Broader Impacts. We believe that LLaDA-V can inspire further exploration of probabilistic modeling approaches for multimodal understanding. However, like many advanced Multimodal Large Language Models (MLLMs), LLaDA-V may generate hallucinations—factually incorrect content or information not present in the input. Nonetheless, approaches such as scaling up data and developing more advanced alignment techniques may help mitigate this problem.

References

[1] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.

[2] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 296–26 306.

[3] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu *et al.*, "Llava-onevision: Easy visual task transfer," *arXiv preprint arXiv:2408.03326*, 2024.

[4] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu *et al.*, "Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 24 185–24 198.

[5] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[6] C. Team, "Chameleon: Mixed-modal early-fusion foundation models," *arXiv preprint arXiv:2405.09818*, 2024.

[7] D. Ding, Z. Ju, Y. Leng, S. Liu, T. Liu, Z. Shang, K. Shen, W. Song, X. Tan, H. Tang *et al.*, "Kimi-audio technical report," *arXiv preprint arXiv:2504.18425*, 2025.

[8] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," *arXiv preprint arXiv:2311.07919*, 2023.

[9] S. Ghosh, S. Kumar, A. Seth, C. K. R. Evuru, U. Tyagi, S. Sakshi, O. Nieto, R. Duraiswami, and D. Manocha, "Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities," *arXiv preprint arXiv:2406.11768*, 2024.

[10] Y. Wang, X. Li, Z. Yan, Y. He, J. Yu, X. Zeng, C. Wang, C. Ma, H. Huang, J. Gao *et al.*, "Internvideo2. 5: Empowering video mllms with long and rich context modeling," *arXiv preprint arXiv:2501.12386*, 2025.

[11] L. Chen, X. Wei, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, Z. Tang, L. Yuan *et al.*, "Sharegpt4video: Improving video understanding and generation with better captions," *Advances in Neural Information Processing Systems*, vol. 37, pp. 19 472–19 495, 2024.

[12] Y. Zhang, J. Wu, W. Li, B. Li, Z. Ma, Z. Liu, and C. Li, "Video instruction tuning with synthetic data," *arXiv preprint arXiv:2410.02713*, 2024.

[13] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[14] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[15] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[16] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[17] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[18] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[19] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei *et al.*, "Qwen2. 5 technical report," *arXiv preprint arXiv:2412.15115*, 2024.

[20] Y. Li, S. Bubeck, R. Eldan, A. Del Giorno, S. Gunasekar, and Y. T. Lee, "Textbooks are all you need ii: phi-1.5 technical report," *arXiv preprint arXiv:2309.05463*, 2023.

[21] X. Bi, D. Chen, G. Chen, S. Chen, D. Dai, C. Deng, H. Ding, K. Dong, Q. Du, Z. Fu *et al.*, "Deepseek llm: Scaling open-source language models with longtermism," *arXiv preprint arXiv:2401.02954*, 2024.

[22] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*. pmlr, 2015, pp. 2256–2265.

[23] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[24] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[25] E. Hoogeboom, D. Nielsen, P. Jaini, P. Forré, and M. Welling, "Argmax flows and multinomial diffusion: Learning categorical distributions," *NeurIPS*, vol. 34, pp. 12 454–12 465, 2021.

[26] J. Austin, D. D. Johnson, J. Ho, D. Tarlow, and R. van den Berg, "Structured denoising diffusion models in discrete state-spaces," in *Advances in Neural Information Processing Systems*, 2021.

[27] F. Bao, S. Nie, K. Xue, C. Li, S. Pu, Y. Wang, G. Yue, Y. Cao, H. Su, and J. Zhu, "One transformer fits all distributions in multi-modal diffusion at scale," in *International Conference on Machine Learning*. PMLR, 2023, pp. 1692–1717.

[28] J. Xie, W. Mao, Z. Bai, D. J. Zhang, W. Wang, K. Q. Lin, Y. Gu, Z. Chen, Z. Yang, and M. Z. Shou, "Show-o: One single transformer to unify multimodal understanding and generation," *arXiv preprint arXiv:2408.12528*, 2024.

[29] C. Zhou, L. Yu, A. Babu, K. Tirumala, M. Yasunaga, L. Shamis, J. Kahn, X. Ma, L. Zettlemoyer, and O. Levy, "Transfusion: Predict the next token and diffuse images with one multi-modal model," *arXiv preprint arXiv:2408.11039*, 2024.

[30] Y. Ma, X. Liu, X. Chen, W. Liu, C. Wu, Z. Wu, Z. Pan, Z. Xie, H. Zhang, L. Zhao *et al.*, "Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation," *arXiv preprint arXiv:2411.07975*, 2024.

[31] S. Tong, D. Fan, J. Zhu, Y. Xiong, X. Chen, K. Sinha, M. Rabbat, Y. LeCun, S. Xie, and Z. Liu, "Metamorph: Multimodal understanding and generation via instruction tuning," *arXiv preprint arXiv:2412.14164*, 2024.

[32] S. Kou, J. Jin, Z. Liu, C. Liu, Y. Ma, J. Jia, Q. Chen, P. Jiang, and Z. Deng, "Orthus: Autoregressive interleaved image-text generation with modality-specific heads," *arXiv preprint arXiv:2412.00127*, 2024.

[33] A. Swerdlow, M. Prabhudesai, S. Gandhi, D. Pathak, and K. Fragkiadaki, "Unified multimodal discrete diffusion," *arXiv preprint arXiv:2503.20853*, 2025.

[34] Z. Li, H. Li, Y. Shi, A. B. Farimani, Y. Kluger, L. Yang, and P. Wang, "Dual diffusion for unified image generation and understanding," *arXiv preprint arXiv:2501.00289*, 2024.

[35] A. Campbell, J. Benton, V. D. Bortoli, T. Rainforth, G. Deligiannidis, and A. Doucet, "A continuous time framework for discrete denoising models," in *Advances in Neural Information Processing Systems*, 2022.

[36] Z. He, T. Sun, K. Wang, X. Huang, and X. Qiu, "Diffusionbert: Improving generative masked language models with diffusion models," *arXiv preprint arXiv:2211.15029*, 2022.

[37] H. Sun, L. Yu, B. Dai, D. Schuurmans, and H. Dai, "Score-based continuous-time discrete diffusion models," in *The Eleventh International Conference on Learning Representations*, 2023.

[38] A. Lou, C. Meng, and S. Ermon, "Discrete diffusion modeling by estimating the ratios of the data distribution," in *Forty-first International Conference on Machine Learning*, 2024.

[39] J. Shi, K. Han, Z. Wang, A. Doucet, and M. K. Titsias, "Simplified and generalized masked diffusion for discrete data," *arXiv preprint arXiv:2406.04329*, 2024.

[40] S. S. Sahoo, M. Arriola, Y. Schiff, A. Gokaslan, E. Marroquin, J. T. Chiu, A. Rush, and V. Kuleshov, "Simple and effective masked diffusion language models," *arXiv preprint arXiv:2406.07524*, 2024.

[41] J. Ou, S. Nie, K. Xue, F. Zhu, J. Sun, Z. Li, and C. Li, "Your absorbing discrete diffusion secretly models the conditional distributions of clean data," *arXiv preprint arXiv:2406.03736*, 2024.

[42] S. Nie, F. Zhu, Z. You, X. Zhang, J. Ou, J. Hu, J. Zhou, Y. Lin, J.-R. Wen, and C. Li, "Large language diffusion models," *arXiv preprint arXiv:2502.09992*, 2025.

[43] Z. You, J. Ou, X. Zhang, J. Hu, J. Zhou, and C. Li, "Effective and efficient masked image generation models," *arXiv preprint arXiv:2503.07197*, 2025.

[44] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa *et al.*, "Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features," *arXiv preprint arXiv:2502.14786*, 2025.

[45] L. Chen, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, J. Wang, Y. Qiao, D. Lin *et al.*, "Are we on the right way for evaluating large vision-language models?" *arXiv preprint arXiv:2403.20330*, 2024.

[46] S. Nie, F. Zhu, C. Du, T. Pang, Q. Liu, G. Zeng, M. Lin, and C. Li, "Scaling up masked diffusion models on text," *arXiv preprint arXiv:2410.18514*, 2024.

[47] A. Campbell, J. Yim, R. Barzilay, T. Rainforth, and T. Jaakkola, "Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design," 2024.

[48] V. T. Hu and B. Ommer, "[mask] is all you need," 2024. [Online]. Available: https://arxiv.org/abs/2412.06787

[49] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[50] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 11 975–11 986.

[51] A. Wang, B. Ai, B. Wen, C. Mao, C.-W. Xie, D. Chen, F. Yu, H. Zhao, J. Yang, J. Zeng *et al.*, "Wan: Open and advanced large-scale video generative models," *arXiv preprint arXiv:2503.20314*, 2025.

[52] Z. Yang, J. Teng, W. Zheng, M. Ding, S. Huang, J. Xu, Y. Yang, W. Hong, X. Zhang, G. Feng *et al.*, "Cogvideox: Text-to-video diffusion models with an expert transformer," *arXiv preprint arXiv:2408.06072*, 2024.

[53] W. Kong, Q. Tian, Z. Zhang, R. Min, Z. Dai, J. Zhou, J. Xiong, X. Li, B. Wu, J. Zhang *et al.*, "Hunyuanvideo: A systematic framework for large video generative models," *arXiv preprint arXiv:2412.03603*, 2024.

[54] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, "Llava-next: Improved reasoning, ocr, and world knowledge," January 2024. [Online]. Available: https://llava-vl.github.io/blog/2024-01-30-llava-next/

[55] J. Guo, T. Zheng, Y. Bai, B. Li, Y. Wang, K. Zhu, Y. Li, G. Neubig, W. Chen, and X. Yue, "Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale," *arXiv preprint arXiv:2412.05237*, 2024.

[56] Y. Jia, J. Li, X. Yue, B. Li, P. Nie, K. Zou, and W. Chen, "Visualwebinstruct: Scaling up multimodal instruction data through web search," *arXiv preprint arXiv:2503.10582*, 2025.

[57] Q. Team, "Qwen3: Think deeper, act faster," 2025, https://qwenlm.github.io/blog/qwen3/. [Online]. Available: https://qwenlm.github.io/blog/qwen3/

[58] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[59] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," *Advances in Neural Information Processing Systems*, vol. 36, pp. 53 728–53 741, 2023.

[60] Y. Meng, M. Xia, and D. Chen, "Simpo: Simple preference optimization with a reference-free reward," *Advances in Neural Information Processing Systems*, vol. 37, pp. 124 198–124 235, 2024.

[61] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu *et al.*, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," *arXiv preprint arXiv:2402.03300*, 2024.

[62] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun *et al.*, "Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9556–9567.

[63] X. Yue, T. Zheng, Y. Ni, Y. Wang, K. Zhang, S. Tong, Y. Sun, B. Yu, G. Zhang, H. Sun *et al.*, "Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark," *arXiv preprint arXiv:2409.02813*, 2024.

[64] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, J. Yang, X. Zheng, K. Li, X. Sun, Y. Wu, and R. Ji, "Mme: A comprehensive evaluation benchmark for multimodal large language models," *arXiv preprint arXiv:2306.13394*, 2023.

[65] B. Li, R. Wang, G. Wang, Y. Ge, Y. Ge, and Y. Shan, "Seed-bench: Benchmarking multimodal llms with generative comprehension," *arXiv preprint arXiv:2307.16125*, 2023.

[66] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu *et al.*, "Mmbench: Is your multi-modal model an all-around player?" in *European conference on computer vision*. Springer, 2024, pp. 216–233.

[67] R. Zhang, D. Jiang, Y. Zhang, H. Lin, Z. Guo, P. Qiu, A. Zhou, P. Lu, K.-W. Chang, Y. Qiao *et al.*, "Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?" in *European Conference on Computer Vision*. Springer, 2024, pp. 169–186.

[68] P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, and J. Gao, "Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models," *CoRR*, 2023.

[69] A. Kembhavi, M. Salvato, E. Kolve, M. Seo, H. Hajishirzi, and A. Farhadi, "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*. Springer, 2016, pp. 235–251.

[70] A. Masry, D. X. Long, J. Q. Tan, S. Joty, and E. Hoque, "Chartqa: A benchmark for question answering about charts with visual and logical reasoning," *arXiv preprint arXiv:2203.10244*, 2022.

[71] M. Mathew, D. Karatzas, and C. Jawahar, "Docvqa: A dataset for vqa on document images," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 2200–2209.

[72] M. Mathew, V. Bagal, R. Tito, D. Karatzas, E. Valveny, and C. Jawahar, "Infographicvqa," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1697–1706.

[73] x.ai, "Grok-1.5 vision preview," 2024, https://x.ai/news/grok-1.5v/. [Online]. Available: https://x.ai/news/grok-1.5v/

[74] F. Wang, X. Fu, J. Y. Huang, Z. Li, Q. Liu, X. Liu, M. D. Ma, N. Xu, W. Zhou, K. Zhang *et al.*, "Muirbench: A comprehensive benchmark for robust multi-image understanding," *arXiv preprint arXiv:2406.09411*, 2024.

[75] J. Zhou, Y. Shu, B. Zhao, B. Wu, S. Xiao, X. Yang, Y. Xiong, B. Zhang, T. Huang, and Z. Liu, "Mlvu: A comprehensive benchmark for multi-task long video understanding," *arXiv preprint arXiv:2406.04264*, 2024.

[76] C. Fu, Y. Dai, Y. Luo, L. Li, S. Ren, R. Zhang, Z. Wang, C. Zhou, Y. Shen, M. Zhang *et al.*, "Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis," *arXiv preprint arXiv:2405.21075*, 2024.

[77] L. Chen, J. Li, X. Dong, P. Zhang, C. He, J. Wang, F. Zhao, and D. Lin, "Sharegpt4v: Improving large multi-modal models with better captions," in *European Conference on Computer Vision*. Springer, 2024, pp. 370–387.

[78] P. Tong, E. Brown, P. Wu, S. Woo, A. J. V. IYER, S. C. Akula, S. Yang, J. Yang, M. Middepogu, Z. Wang *et al.*, "Cambrian-1: A fully open, vision-centric exploration of multimodal llms," *Advances in Neural Information Processing Systems*, vol. 37, pp. 87 310–87 356, 2024.

[79] H. Lu, W. Liu, B. Zhang, B. Wang, K. Dong, B. Liu, J. Sun, T. Ren, Z. Li, H. Yang *et al.*, "Deepseek-vl: towards real-world vision-language understanding," *arXiv preprint arXiv:2403.05525*, 2024.

[80] Z. Wu, X. Chen, Z. Pan, X. Liu, W. Liu, D. Dai, H. Gao, Y. Ma, C. Wu, B. Wang *et al.*, "Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding," *arXiv preprint arXiv:2412.10302*, 2024.

[81] C. Wu, X. Chen, Z. Wu, Y. Ma, X. Liu, Z. Pan, W. Liu, Z. Xie, X. Yu, C. Ruan *et al.*, "Janus: Decoupling visual encoding for unified multimodal understanding and generation," *arXiv preprint arXiv:2410.13848*, 2024.

[82] X. Chen, Z. Wu, X. Liu, Z. Pan, W. Liu, Z. Xie, X. Yu, and C. Ruan, "Janus-pro: Unified multimodal understanding and generation with data and model scaling," *arXiv preprint arXiv:2501.17811*, 2025.

[83] X. Wang, X. Zhang, Z. Luo, Q. Sun, Y. Cui, J. Wang, F. Zhang, Y. Wang, Z. Li, Q. Yu *et al.*, "Emu3: Next-token prediction is all you need," *arXiv preprint arXiv:2409.18869*, 2024.

[84] H. Laurençon, L. Tronchon, M. Cord, and V. Sanh, "What matters when building vision-language models?" *Advances in Neural Information Processing Systems*, vol. 37, pp. 87 874–87 907, 2024.

[85] X. Li, J. Thickstun, I. Gulrajani, P. S. Liang, and T. B. Hashimoto, "Diffusion-lm improves controllable text generation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 4328–4343, 2022.

[86] S. Gong, M. Li, J. Feng, Z. Wu, and L. Kong, "Diffuseq: Sequence to sequence text generation with diffusion models," *arXiv preprint arXiv:2210.08933*, 2022.

[87] X. Han, S. Kumar, and Y. Tsvetkov, "Ssd-lm: Semi-autoregressive simplex-based diffusion language model for text generation and modular control," *arXiv preprint arXiv:2210.17432*, 2022.

[88] R. Strudel, C. Tallec, F. Altché, Y. Du, Y. Ganin, A. Mensch, W. Grathwohl, N. Savinov, S. Dieleman, L. Sifre *et al.*, “Self-conditioned embedding diffusion for text generation,” *arXiv preprint arXiv:2211.04236*, 2022.

[89] T. Chen, R. Zhang, and G. Hinton, “Analog bits: Generating discrete data using diffusion models with self-conditioning,” *arXiv preprint arXiv:2208.04202*, 2022.

[90] S. Dieleman, L. Sartran, A. Roshannai, N. Savinov, Y. Ganin, P. H. Richemond, A. Doucet, R. Strudel, C. Dyer, C. Durkan *et al.*, “Continuous diffusion for categorical data,” *arXiv preprint arXiv:2211.15089*, 2022.

[91] P. H. Richemond, S. Dieleman, and A. Doucet, “Categorical sdes with simplex diffusion,” 2022.

[92] T. Wu, Z. Fan, X. Liu, Y. Gong, Y. Shen, J. Jiao, H.-T. Zheng, J. Li, Z. Wei, J. Guo, N. Duan, and W. Chen, “Ar-diffusion: Auto-regressive diffusion model for text generation,” 2023.

[93] R. K. Mahabadi, H. Ivison, J. Tae, J. Henderson, I. Beltagy, M. E. Peters, and A. Cohan, “Tess: Text-to-text self-conditioned simplex diffusion,” 2024.

[94] J. Ye, Z. Zheng, Y. Bao, L. Qian, and M. Wang, “Dinoiser: Diffused conditional sequence learning by manipulating noises,” *arXiv preprint arXiv:2302.10025*, 2023.

[95] Y. Zhang, J. Gu, Z. Wu, S. Zhai, J. Susskind, and N. Jaitly, “Planner: Generating diversified paragraph via latent language diffusion model,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 80 178–80 190, 2023.

[96] A. Lou and S. Ermon, “Reflected diffusion models,” 2023.

[97] A. Graves, R. K. Srivastava, T. Atkinson, and F. Gomez, “Bayesian flow networks,” *arXiv preprint arXiv:2308.07037*, 2023.

[98] Z. Lin, Y. Gong, Y. Shen, T. Wu, Z. Fan, C. Lin, N. Duan, and W. Chen, “Text generation with diffusion language models: A pre-training approach with continuous paragraph denoise,” in *International Conference on Machine Learning*. PMLR, 2023, pp. 21 051–21 064.

[99] K. Xue, Y. Zhou, S. Nie, X. Min, X. Zhang, J. Zhou, and C. Li, “Unifying bayesian flow networks and diffusion models through stochastic differential equations,” *arXiv preprint arXiv:2404.15766*, 2024.

[100] R. Zhang, S. Zhai, Y. Zhang, J. Thornton, Z. Ou, J. Susskind, and N. Jaitly, “Target concrete score matching: A holistic framework for discrete diffusion,” *arXiv preprint arXiv:2504.16431*, 2025.

[101] E. Hoogeboom, A. A. Gritsenko, J. Bastings, B. Poole, R. v. d. Berg, and T. Salimans, “Autoregressive diffusion models,” *arXiv preprint arXiv:2110.02037*, 2021.

[102] C. Meng, K. Choi, J. Song, and S. Ermon, “Concrete score matching: Generalized score matching for discrete data,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 34 532–34 545, 2022.

[103] M. Reid, V. J. Hellendoorn, and G. Neubig, “Diffuser: Discrete diffusion via edit-based reconstruction,” 2022.

[104] H. Sun, L. Yu, B. Dai, D. Schuurmans, and H. Dai, “Score-based continuous-time discrete diffusion models,” *arXiv preprint arXiv:2211.16750*, 2022.

[105] O. Kitouni, N. Nolte, J. Hensman, and B. Mitra, “Disk: A diffusion model for structured knowledge,” *arXiv preprint arXiv:2312.05253*, 2023.

[106] L. Zheng, J. Yuan, L. Yu, and L. Kong, “A reparameterized discrete diffusion model for text generation,” *ArXiv*, vol. abs/2302.05737, 2023.

[107] Z. Chen, H. Yuan, Y. Li, Y. Kou, J. Zhang, and Q. Gu, “Fast sampling via de-randomization for discrete diffusion models,” *arXiv preprint arXiv:2312.09193*, 2023.

[108] J. Ye, Z. Zheng, Y. Bao, L. Qian, and Q. Gu, "Diffusion language models can perform many tasks with scaling and instruction-finetuning," *arXiv preprint arXiv:2308.12219*, 2023.

[109] I. Gat, T. Remez, N. Shaul, F. Kreuk, R. T. Chen, G. Synnaeve, Y. Adi, and Y. Lipman, "Discrete flow matching," *arXiv preprint arXiv:2407.15595*, 2024.

[110] K. Zheng, Y. Chen, H. Mao, M.-Y. Liu, J. Zhu, and Q. Zhang, "Masked diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical sampling," 2024. [Online]. Available: https://arxiv.org/abs/2409.02908

[111] H. Chang, H. Zhang, L. Jiang, C. Liu, and W. T. Freeman, "Maskgit: Masked generative image transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 315–11 325.

[112] K. Zhang, B. Li, P. Zhang, F. Pu, J. A. Cahyono, K. Hu, S. Liu, Y. Zhang, J. Yang, C. Li, and Z. Liu, "Lmms-eval: Reality check on the evaluation of large multimodal models," 2024. [Online]. Available: https://arxiv.org/abs/2407.12772

[113] J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebron, and S. Sanghai, "Gqa: Training generalized multi-query transformer models from multi-head checkpoints," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 4895–4901.

[114] B. Zhang and R. Sennrich, "Root mean square layer normalization," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[115] N. Shazeer, "Glu variants improve transformer," *arXiv preprint arXiv:2002.05202*, 2020.

[116] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 568, p. 127063, 2024.

Contents

1 Introduction **1**

2 Preliminaries **3**

3 Method **3**
3.1 Training Objective and Architecture 3
3.2 Inference Process 4
3.3 Training Strategies 5

4 Experiment **5**
4.1 Experimental Settings 5
4.2 Data Scalability of LLaDA-V 6
4.3 Benchmark Results 7
4.4 Ablation Study 8

5 Related Work **8**

6 Conclusion **9**

A The Formulation of Masked Diffusion Models **18**

B Experiments **18**
B.1 Model Architecture 18
B.2 Attention Mask 19
B.3 Case Studies 19

A The Formulation of Masked Diffusion Models

In this section, we present the main formulation of masked diffusion models for completeness. Please refer to Shi et al. [39], Sahoo et al. [40], Ou et al. [41] for theoretical details.

In masked diffusion models, the forward process independently masks each token in a sentence $\boldsymbol{x}_0 \in \{0, 1, \ldots, K-1\}^N$, based on a given noise level $t \in [0, 1]$, where K and N denote the vocabulary size and sentence length, respectively.

$$q_{t|0}(\boldsymbol{x}_t|\boldsymbol{x}_0) = \prod_{i=0}^{N-1} q_{t|0}(\boldsymbol{x}_t^i|\boldsymbol{x}_0^i) \quad \text{and} \quad q_{t|0}(\boldsymbol{x}_t^i|\boldsymbol{x}_0^i) = \begin{cases} \alpha_t, & \boldsymbol{x}_t^i = \boldsymbol{x}_0^i, \\ 1 - \alpha_t, & \boldsymbol{x}_t^i = \text{[M]}. \end{cases} \tag{2}$$

In LLaDA-V, we choose $\alpha_t = 1 - t$ following LLaDA [42] due to its demonstrated superior empirical performance. Intuitively, during the forward process, each token independently has a probability t of being masked (replaced with [M]) and a probability $1 - t$ of remaining unchanged.

Masked diffusion models generate text by simulating a reverse process that gradually transforms masked tokens into meaningful content, starting from a fully masked sequence. Given $0 \leq s < t \leq 1$, each sampling step in the reverse process is characterized by

$$q_{s|t}(\boldsymbol{x}_s|\boldsymbol{x}_t) = \prod_{i=0}^{N-1} q_{s|t}(\boldsymbol{x}_s^i|\boldsymbol{x}_t) \text{ and } q_{s|t}(\boldsymbol{x}_s^i|\boldsymbol{x}_t) = \begin{cases} 1, & \boldsymbol{x}_t^i \neq \text{[M]}, \boldsymbol{x}_s^i = \boldsymbol{x}_t^i, \\ \frac{1-\alpha_s}{1-\alpha_t}, & \boldsymbol{x}_t^i = \text{[M]}, \boldsymbol{x}_s^i = \text{[M]}, \\ \frac{\alpha_s - \alpha_t}{1-\alpha_t} p_{\boldsymbol{\theta}}(\boldsymbol{x}_0^i|\boldsymbol{x}_t), & \boldsymbol{x}_t^i = \text{[M]}, \boldsymbol{x}_s^i \neq \text{[M]}, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where $p_{\boldsymbol{\theta}}$ is modeled by a Transformer. When using $\alpha_t = 1 - t$, the reverse process has an intuitive interpretation: at each generation step, tokens that are already meaningful content remain unchanged, while masked tokens [M] either stay masked with probability s/t or are replaced with meaningful content predicted by the model with probability $1 - s/t$.

The training objective of masked diffusion models is the following upper bound on negative log-likelihood:

$$\mathcal{L}_{\boldsymbol{\theta}} = \int_0^1 \frac{1}{t} \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} \left[\sum_{\{i|\boldsymbol{x}_t^i = m\}} - \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0^i|\boldsymbol{x}_t) \right] dt. \tag{4}$$

For each sampling step in the reverse process (Eq. (3)), given $\boldsymbol{x}_t$, we first identify masked positions i (where $\boldsymbol{x}_t^i$ = [M]) and then sample a token $\boldsymbol{x}_0^i$ for each such position from the distribution $p_\theta(\boldsymbol{x}_0^i \mid \boldsymbol{x}_t)$. Subsequently, a fraction s/t of these newly sampled tokens are typically selected randomly for re-masking. However, Chang et al. [111] introduced a deterministic re-masking strategy that selects tokens with the lowest confidence scores (i.e., the smallest $p_\theta(\boldsymbol{x}_0^i \mid \boldsymbol{x}_t)$ values) for re-masking, comprising the s/t proportion. LLaDA [42] adopts this low-confidence re-masking approach and demonstrates consistent improvements across various downstream tasks. In LLaDA-V, we also employ this low-confidence re-masking strategy following LLaDA.

B Experiments

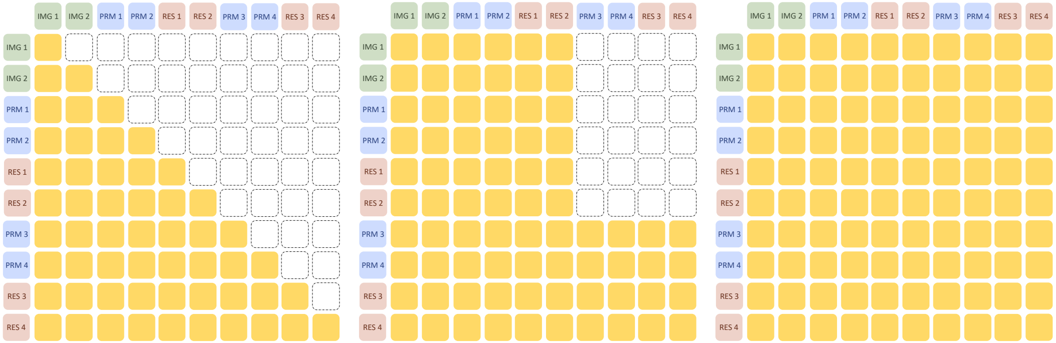
The implementation of LLaDA-V leverages official codebases and datasets from MAmmoTH [55], VisualWebInstruct [56], LLaVA-NeXT [54], and LMMS-EVAL [112], with details of the corresponding links provided in Tab. 5.

B.1 Model Architecture

The language tower of LLaDA-V strictly follows the architecture of LLaDA [42]. The architecture of LLaDA is largely based on LLaMA3 [16], with the main difference being the removal of the causal mask: LLaDA replaces the causal transformer in LLaMA3 with a bidirectional transformer. As a result, LLaDA does not support KV caching and uses standard multi-head attention, in contrast to the grouped query attention [113] in LLaMA3. Aside from these changes, both models employ

Table 5: **Code repositories and datasets leveraged in our implementation**

Code	**URL**
LMMs-Eval	`https://github.com/EvolvingLMMs-Lab/lmms-eval`
LLaVA-NeXT	`https://github.com/LLaVA-VL/LLaVA-NeXT`
MAmmoTH-VL	`https://github.com/MAmmoTH-VL/MAmmoTH-VL`
VisualWebInstruct	`https://github.com/TIGER-AI-Lab/VisualWebInstruct`
Data	**URL**
LLaVA-Pretrain	`https://huggingface.co/datasets/liuhaotian/LLaVA-Pretrain`
LLaVA-NeXT	`https://huggingface.co/datasets/lmms-lab/LLaVA-NeXT-Data`
MAmmoTH-VL	`https://huggingface.co/datasets/MAmmoTH-VL/MAmmoTH-VL-Instruct-12M`
VisualWebInstruct	`https://huggingface.co/datasets/TIGER-Lab/VisualWebInstruct`



(a) Causal Mask (b) Dialogue Causal Mask (c) No Mask

Figure 4: **Overview of Attention Masks.** (a) Standard causal mask used in autoregressive models like Qwen2-VL and LLaMA3-V, where tokens attend only to themselves and previous tokens. (b) Dialogue causal mask allowing full attention within turns while maintaining causality between turns. (c) Bidirectional attention in LLaDA-V, enabling tokens to attend to all tokens in the sequence. Note: In the figure, PRM represents prompt and RES represents response.

widely used techniques in large language models, including RMSNorm [114], SwiGLU [115], and RoPE [116]. For the vision tower in LLaDA-V, we employ the siglip2-so400m-patch14-384 model, which processes visual inputs with a resolution of 384×384 pixels and produces 729 visual tokens per image. For the projector in LLaDA-V, we employ a randomly initialized two-layer MLP.

B.2 Attention Mask

In Fig. 4, we summarize the attention masks discussed in this work. Conventional autoregressive MLLMs utilize a standard causal mask, as shown in Fig. 4a, which restricts each token's attention to itself and all previous tokens. LLaDA-V explores two additional alternatives: the *Dialogue Causal Mask*, which allows bidirectional attention within each dialogue turn while preserving causality across turns, which effectively aligns with the structure of multi-turn conversations, and the *No Mask* approach, which enables fully bidirectional attention, allowing all tokens to attend to every other token in the sequence. As discussed in Sec. 4.4, both attention mask strategies demonstrate strong performance. However, the no mask strategy achieves superior results, outperforming the alternative on 7 out of 12 benchmarks. Consequently, we adopt the no mask strategy as the default in LLaDA-V.

B.3 Case Studies

Table 6: **Case Study (1)**

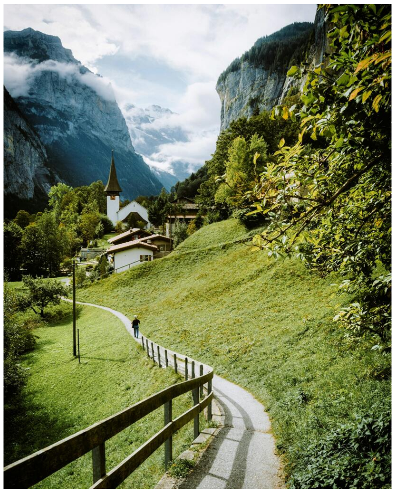
User	Please describe the image in detail 
LLaDA-V	The image captures a serene and picturesque landscape in the Swiss Alps. The the scene is is a of a lush, green path that meanders through a grassy area. The path is bordered by a wooden fence, and a person can be seen walking along it, adding a sense of scale to the scene. In the background, a white church with a pointed steeple stands out against the natural backdrop. The church is nestled among a valley, surrounded by green trees and trees. The valley is framed by towering mountains, their peaks shrouded in mist, adding a sense of depth and tranquility to the image. The sky above is a clear blue, with a few clouds lazily scattered across the sky, contributing to the overall tranquility of the scene. The image is a snapshot of a moment of peace and tranquility, in the natural beauty of the Swiss Alps. The objects in the image are arranged in layers, with the path in the foreground, the valley in the middle, and the mountains in the background. The path is in the foreground, and the church is in the distance, contributing to the overall tranquility of the scene. The image is a snapshot of a moment of peace and tranquility, in the natural beauty of the Swiss Alps.

Table 7: **Case Study (2)**

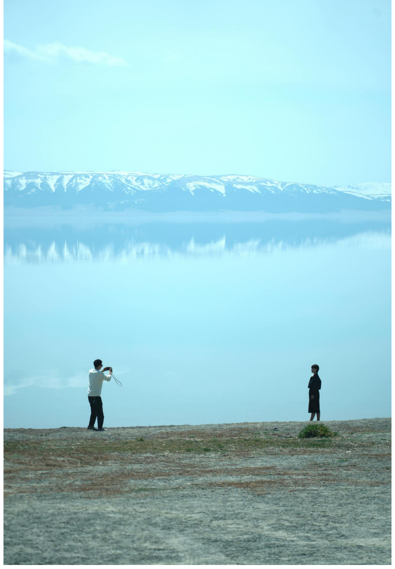
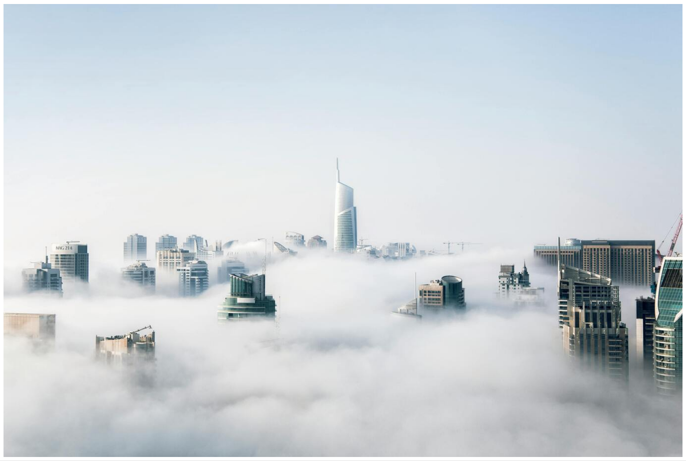
User	How many people are there in the picture? 
LLaDA-V	To determine the number of people in the picture, we need to carefully observe the image and count each individual present. 1. Observation of the image shows a scenic landscape with a lake of water surrounded by snow-capped mountains. The are two individuals visible in the image. One person is taking a photograph, and the other is standing still. 2. Counting the individuals: - The person person taking a photograph is located on the left side of the image. - The person standing still is located on the right side of the image. 3. Conclusion: - By counting these two individuals, we confirm that there are indeed two people present in the picture. Therefore, the detailed answer is: The image shows a scenic landscape with a lake of water surrounded by snow-capped mountains. There are two individuals visible in the image. - We see one person is taking a photograph on the left side of the image. - There is another person standing on the right side of the image. By counting these two individuals, we confirm that there are 2 people in the picture. Therefore, the number of people in the picture is 2.

Table 8: **Case Study (3)**

User	Is this picture real? 
LLaDA-V	To determine if the picture is real, we need to analyze several aspects of the image: 1. **Resolution and Detail**: The image shows a cityscape with high resolution and detail. The buildings have distinct outlines, and the fog is visible in detail, suggesting a high-quality photograph. 2. **Lighting and Shadows**: The lighting in the image appears natural, with shadows cast by the buildings indicating a strong light source, likely the sun. This is consistent with real-world lighting. 3. **Color and Contrast**: The colors in the image are vibrant and natural, with the contrast between the buildings and the fog being well-defined. This suggests that the image has been captured with a camera. 4. **Composition and Perspective**: The composition of the image is well-balanced, with the cityscape framed in a way that draws the viewer's attention. The perspective is also realistic, consistent with a real photograph. 5. **Artistic Elements**: There are no obvious signs of digital manipulation or artifacts, further supporting the conclusion that this is a real photograph. Given these observations, it is reasonable to conclude that the image is indeed a real photograph. Therefore, the answer is: Yes, the picture is real.