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

Supplementary Material

Contents

A. 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 [5, 24, 66, 67, 88, 89, 105] are primarily trained using autoregressive modeling. Unlike autoregressive approaches, discrete diffusion models [32, 77] offer an alternative paradigm for language modeling. Masked diffusion models [2, 6], a specific variant of discrete diffusion, have shown impressive results across multiple domains [7, 33, 53, 63–65, 72, 76, 82, 110].

Among them, LLaDA [64] has demonstrated comparable performance with strong AR models like LLaMA3-8B-Instruct [24], 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 B.

Visual Instruction Tuning [41, 48, 49] 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 [68] or SigLIP [90, 114]) 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.

B. The Formulation of Masked Diffusion Models

In this section, we present the main formulation of masked diffusion models for completeness. Please refer to Ou et al. [65], Sahoo et al. [72], Shi et al. [76] 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.

$$q_{t|0}(\boldsymbol{x}_t|\boldsymbol{x}_0) = \prod_{i=0}^{N-1} q_{t|0}(\boldsymbol{x}_t^i|\boldsymbol{x}_0^i), \tag{2}$$

$$q_{t|0}(\boldsymbol{x}_t^i|\boldsymbol{x}_0^i) = \begin{cases} \alpha_t, & \boldsymbol{x}_t^i = \boldsymbol{x}_0^i, \\ 1-\alpha_t, & \boldsymbol{x}_t^i = \text{[M]}. \end{cases} \tag{3}$$

In LLaDA-V, we choose $\alpha_t = 1 - t$ following LLaDA [64] 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 $0 \le s < t \le 1$, each sampling step in the reverse process is characterized by

$$q_{s|t}(\boldsymbol{x}_s|\boldsymbol{x}_t) = \prod_{i=0}^{N-1} q_{s|t}(\boldsymbol{x}_s^i|\boldsymbol{x}_t), \tag{4}$$

$$q_{s|t}(\boldsymbol{x}_s^i|\boldsymbol{x}_t) = \begin{cases} 1, & \boldsymbol{x}_t^i \neq \text{[M]}, \boldsymbol{x}_s^i = \boldsymbol{x}_t^i, \\ \frac{1-\alpha_s}{1-\alpha_t}, & \boldsymbol{x}_t^i = \text{[M]}, \boldsymbol{x}_s^i = \text{[M]}, \\ \frac{\alpha_s-\alpha_t}{1-\alpha_t} p_{\boldsymbol{\theta}}(\boldsymbol{x}_0^i|\boldsymbol{x}_t), & \boldsymbol{x}_t^i = \text{[M]}, \boldsymbol{x}_s^i \neq \text{[M]}, \\ 0, & \text{otherwise}. \end{cases} \tag{5}$$

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:

$$\mathcal{L}_{\boldsymbol{\theta}} = \int_0^1 \frac{1}{t} \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} \left[\sum_{\{i|\boldsymbol{x}_t^i = m\}} -\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0^i|\boldsymbol{x}_t) \right] dt. \tag{6}$$

For each sampling step in the reverse process (Eq. (4)), given $\boldsymbol{x}_t$, we first identify masked positions i (where $\boldsymbol{x}_t^i = \text{[M]}$) 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. [8] 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 [64] 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.

C. Experiments

The implementation of LLaDA-V leverages official codebases and datasets from MAmmoTH [27], VisualWebInstruct [34], LLaVA-NeXT [50], and LMMS-EVAL [116], with details of the corresponding links provided in Tab. 5.

C.1. Model Architecture

The language tower of LLaDA-V strictly follows the architecture of LLaDA [64]. The architecture of LLaDA is largely based on LLaMA3 [88], 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 [1] in LLaMA3. Aside from these changes, both models employ widely used techniques in large language models, including RMSNorm [115], SwiGLU [75], and RoPE [80]. 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.

C.2. Detailed training settings

We summarize the detailed training configuration of LLaDA-V in Tab. 6, including the datasets, model backbones, and optimization hyperparameters used at each stage. LLaDA-V is trained sequentially on the first five datasets, while the last dataset is used only in the ablation study.

C.3. Attention Mask

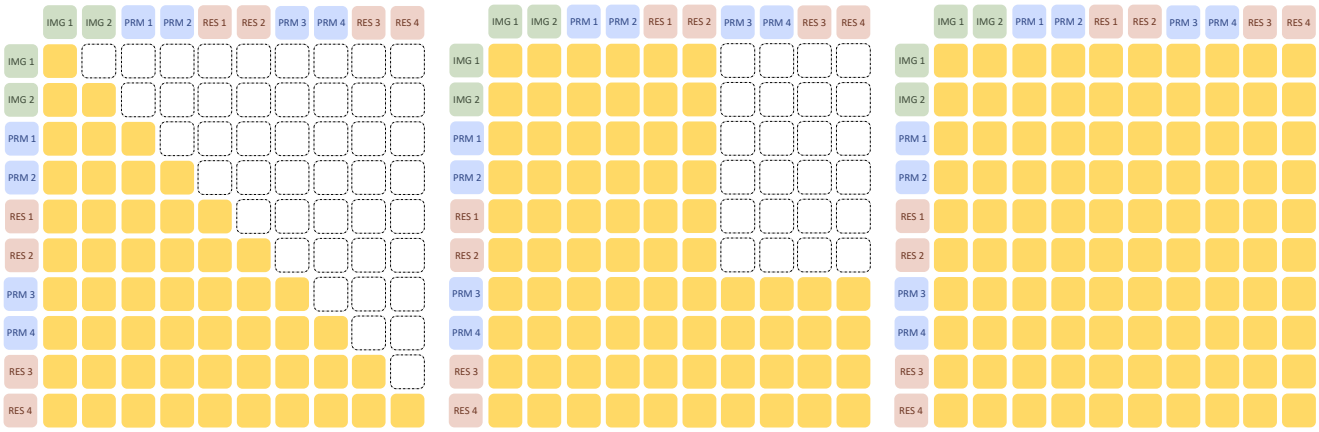
In Fig. 4, we summarize the attention masks discussed in this work. Conventional autoregressive MLLMs utilize a standard causal mask, as shown in Fig. 4a, which restricts each token's attention to itself and all previous tokens. LLaDA-V explores two additional alternatives: the *Dialogue Causal Mask*, 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. 3.4, both attention mask strategies demonstrate strong

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

Table 6. **Training Settings.** Here M-SI and M-OV represent the single image data and onevision data of MAmmoTH [27], while VW represents the data of VisualWebInstruct [34]. We train LLaDA-V sequentially through the first five datasets (LLaVA-Pretrain [48], M-SI, M-OV, VW, and M-OV+VW), while the last dataset (LLaVA-NeXT [50]) is used for ablation study in Sec. 3.4.

Training data	LLaVA-Pretrain	M-SI	M-OV	VW	M-OV+VW	LLaVA-NeXT
Vision tower	Siglip2-so400m-patch14-384 [90]					
Language tower	LLaDA-8B-Instruct [64]					
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



(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.

Table 7. Quantitative comparison of attention patterns between LLaMA3-V and LLaDA-V.

Model	Future-Looking Attention Ratio	Attention Entropy	Attention Sink (σ)
LLaMA3-V	$0.00\% \pm 0.00\%$	3.41 ± 0.17	100% (0.6)
LLaDA-V	$53.41\% \pm 0.61\%$	8.40 ± 0.09	0% (0.1)

Table 8. **Effect of Refresh Interval on MathVista.** We report accuracy (%) and throughput (tokens/s) under different cache refresh intervals r of Fast-dLLM [97].

Refresh Interval (r)	2	4	8	16	32	48
Accuracy (%)	59.1	59.1	58.2	57.0	56.6	56.8
Throughput (tokens/s)	10.6	16.9	24.0	29.8	34.2	35.5

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.

C.4. Extended Data Scaling Results

To complement the representative scaling results shown in Fig. 3, we further evaluate data scaling on additional benchmarks covering chart/document understanding (DocVQA, InfoVQA, ChartQA), mathematical reasoning (MathVista), and multi-image or vision-centric reasoning (MuirBench, MMMU-Pro-Vision) in Fig. 5. Overall, these additional results are broadly consistent with the trends observed in the main paper: LLaDA-V benefits from increased data and shows encouraging scaling behavior on several benchmarks. These results broaden the coverage of our scaling study and suggest that the observed scaling behavior is not limited to the representative tasks shown in the main paper.

C.5. Attention Pattern Analysis

To further explain the superior performance of LLaDA-V and gain deeper insights into its unique characteristics and underlying mechanisms, we conduct a comprehensive analysis of the attention patterns in both LLaDA-V and LLaMA3-V. Specifically, we analyze 100 multimodal sequences constructed by pairing randomly selected ImageNet images with the prompt "Please describe the image in detail." For each sequence, three metrics are computed from the attention matrices and averaged across all layers and heads, providing a quantitative comparison between diffusion-based and autoregressive-based models.

Future-Looking Attention Ratio. We evaluate this metric to measure whether earlier tokens exhibit attention to later ones. For a given token i in a sequence of length L, the future-looking attention ratio F_i is calculated as:

$$F_i = \sum_{j=i+1}^{L-1} A_{ij}, \tag{7}$$

where A_{ij} denotes the attention weight from token i to token j. We report the average value of F_i across all tokens.

Attention Entropy. This metric quantifies the distributional characteristics of attention weights, measuring the extent to which attention is uniformly allocated across the token sequence. For each token's attention distribution A_i, the entropy is defined as

$$H(A_i) = -\sum_{j=0}^{L-1} A_{ij} \log_2(A_{ij}), \tag{8}$$

where A_{ij} denotes the attention weight from token i to token j. The final entropy score is obtained by averaging $H(A_i)$ across all tokens.

Attention Sink. Inspired by observations in autoregressive language models, where the first token often receives a large amount of attention [26, 101], we quantify this phenomenon by computing the average attention received by each token j:

$$S_j = \frac{1}{L} \sum_{i=0}^{L-1} A_{ij}, \tag{9}$$

where A_{ij} denotes the attention weight from token i to token j. An *attention sink* is identified when the first token satisfies $S_0 = \max(S)$ and its score exceeds a predefined threshold σ.

Results and Discussion. As shown in Tab. 7, the Future-Looking Attention Ratio for LLaMA3-V is 0.00%, consistent with its strictly autoregressive architecture, where causal masking prevents tokens from attending to future positions. In contrast, LLaDA-V achieves a ratio of 53.41%, revealing a non-autoregressive and bidirectional attention

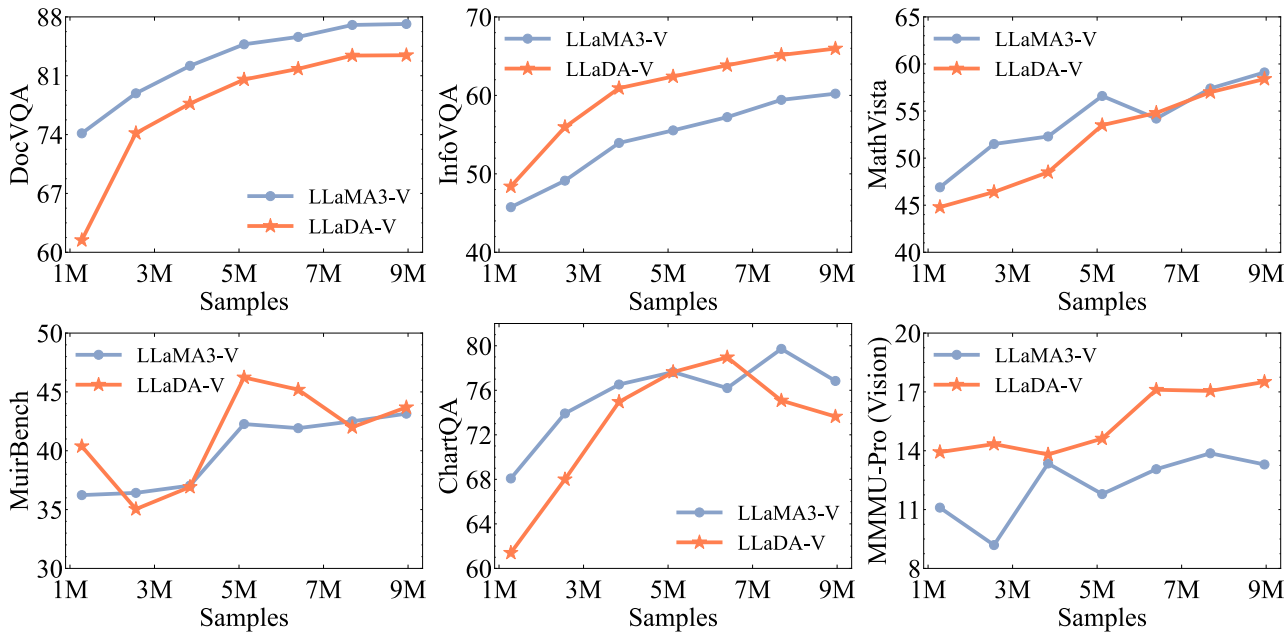


Figure 5. **Extended scaling curves on additional benchmarks.** We report results on chart/document understanding (DocVQA, InfoVQA, ChartQA), mathematical reasoning (MathVista), and multi-image or vision-centric reasoning (MuirBench, MMMU-Pro-Vision). These results broaden the coverage of the scaling study beyond the representative tasks shown in Fig. 3.

mechanism. This property is particularly beneficial for modeling visual data (e.g., images or videos), where spatial relationships are independent of sequence order.

Moreover, LLaDA-V exhibits substantially higher Attention Entropy (8.40 vs. 3.41) and entirely avoids the attention sink phenomenon (0% vs. 100% in LLaMA3-V), indicating a more global and flexible attention distribution. This enables LLaDA-V to attend broadly across image regions, while LLaMA3-V exhibits more concentrated attention that focuses on specific "sink" tokens. Such differences in attention behavior may explain LLaDA-V's superior performance on vision-language tasks, as it more effectively captures complex spatial dependencies in images.

C.6. Trade-off Between Accuracy and Latency

We adapt Fast-dLLM [97] in LLaDA-V, which enables caching of both generated text tokens and image tokens with a configurable cache refresh strategy. When the cache is refreshed too frequently, latency increases due to repeated updates, but accuracy tends to improve because the approximation error in the cached KV is reduced.

To investigate this trade-off between accuracy and inference efficiency, we conduct an ablation study on the cache refresh strategy within LLaDA-V. Instead of updating the cache at every decoding step, we adopt a refresh-based approach where the cache is updated only every r decoding steps (the *refresh interval*). We evaluate different interval values to analyze their effect on performance, as summarized in Tab. 8.

The results clearly illustrate a trade-off between accuracy and inference speed. Increasing the refresh interval (r) from 2 to 48 improves throughput by nearly 3.3× (from 10.6 to 35.5 tokens/s) while incurring only a minor accuracy drop (from 59.1% to 56.8%). This makes the refresh interval function as a practical *dial* that can be adjusted based on task requirements—set lower for accuracy-critical scenarios or higher for time-sensitive applications.

C.7. Limitations

Our work has two main limitations. First, our image processing strategy is still relatively simple. For high-resolution images, we split and resize image segments, process them through the SigLIP2 [90] vision tower, and then concatenate the resulting features. Unlike models with native dynamic-resolution support such as Qwen2-VL, this design may reduce both efficiency and visual fidelity. Second, the inference efficiency of LLaDA-V remains limited in practice. Although Fast-dLLM [97] provides an effective acceleration strategy and enables a practical accuracy–latency trade-off, our model still relies on such acceleration for efficient deployment; without it, the decoding throughput is relatively low due to the iterative denoising process and the lack of native autoregressive-style KV caching. We leave more efficient multimodal diffusion architectures and inference algorithms for future work.

C.8. Case Studies

We present several qualitative examples to illustrate LLaDA-V's multimodal reasoning capabilities across different types of visual understanding tasks (Tables 9–11).

Table 9. **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 10. **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 11. **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.