Bidimensional Reinforcement Learning for Multimodal Diffusion Language Models

Bowen Li[1⋆], Yinjie Wang[1⋆], Yunzhi Zhang[2], Junhong Liu[3], Yingqing Guo[1], Jiajun Wu[2], Mengdi Wang[1†], and Ling Yang[1†]

[1] Princeton University, Princeton, NJ, USA
[2] Stanford University, Stanford, CA, USA
[3] Independent Researcher, China

Abstract. Diffusion Large Language Models (dLLMs) have achieved remarkable success in pure language tasks and recently shown promising results in multimodal understanding. However, existing post-training methods for multimodal dLLMs have not yet fully exploited the unique structure and inference paradigm of diffusion models, leaving room for more tailored optimization strategies. We present **Bi-VRL**, a reinforcement learning framework specifically designed for multimodal dLLMs, featuring a **bidimensional optimization strategy** that enables comprehensive optimization across both reasoning and diffusion timestep dimensions. For the diffusion timestep dimension, our trajectory-level optimization learns a value function over the entire diffusion path, enabling holistic credit assignment across denoising timesteps. For the reasoning dimension, our thought-level optimization decomposes outputs into reasoning steps and delivers targeted rewards to enhance logical soundness. This core strategy is augmented by a fine-grained multimodal alignment mechanism that systematically evaluates both the logical correctness of reasoning steps and their grounding in visual evidence. Our framework is architecture-agnostic, successfully applied to models with both discrete vision tokens (MMaDA) and continuous vision embeddings (LLaDA-V). Extensive experiments across challenging multimodal reasoning benchmarks demonstrate that Bi-VRL significantly outperforms strong supervised fine-tuning baselines, establishing effective post-training methodology for multimodal dLLMs. We will release the code at `https://github.com/Gen-Verse/dLLM-RL`.

Keywords: Diffusion Language Models · Multimodal Reasoning · Reinforcement Learning · Bidimensional Optimization

1 Introduction

Vision-language models have shown remarkable ability to understand and reason about the world [19,30,46]. State-of-the-art systems use reinforcement learning to

⋆ Equal contribution. The work was completed when Bowen and Yinjie were visiting students at Princeton University.
† Corresponding authors.

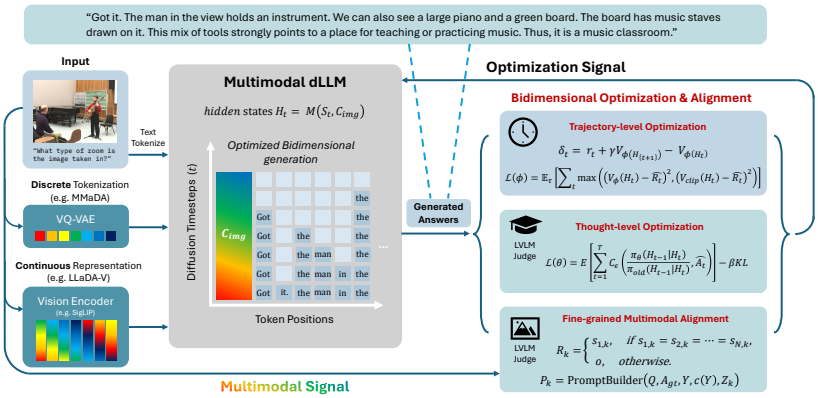


Fig. 1: An overview of Bi-VRL framework. The framework processes multimodal inputs (image and text) by supporting both discrete tokenization (via VQ-VAE) and continuous representation (via vision encoder). The multimodal dLLM performs optimized bidimensional generation across diffusion timesteps and token positions to produce the response (top: example output). The generated response is then evaluated to formulate a comprehensive **Bidimensional Optimization & Alignment** signal composed of three key components: (*i*) **Trajectory-level Optimization** that assigns credit across the diffusion trajectory, (*ii*) **Thought-level Optimization** that delivers rewards to reasoning steps via LLM judges, and (*iii*) **Fine-grained Multimodal Alignment** that verifies both logical soundness and visual grounding. These signals jointly optimize the policy through reinforcement learning.

align their outputs with complex instructions and human preferences [42]. Concurrently, diffusion models have become a strong alternative for generative tasks and have recently shown great promise in text generation [1,24,34,35,37,41,55]. These diffusion large language models (dLLMs) offer unique advantages, including iterative refinement and substantial parallel acceleration potential. However, reinforcement learning, key to activating the reasoning ability of language models, remains largely underexplored for multimodal dLLMs.

Applying reinforcement learning to multimodal dLLMs presents specific challenges. Unlike the linear generation in AR models, the iterative, non-autoregressive process of dLLMs creates a two-dimensional optimization landscape over token positions and denoising timesteps. Recent work on dLLM-RL/TraceRL [48] has addressed the temporal dimension by introducing trajectory-aware optimization that aligns training with inference traces across denoising timesteps in pure text settings. However, existing approaches primarily focus on unimodal scenarios and trajectory-level credit assignment, while providing limited supervision on the logical structure of the generated content, particularly in terms of grounding reasoning steps in visual evidence. This raises a key question for multimodal reasoning: How can we extend RL optimization beyond the temporal dimension to also address the spatial dimension of reasoning structure, while ensuring tight coupling between textual reasoning and visual grounding?

To address this question, we propose **Bi-VRL**, a **Bi**dimensional **V**isual **R**einforcement **L**earning framework for multimodal dLLMs. Building upon dLLM-RL/TraceRL's trajectory-level optimization infrastructure for the temporal dimension, our key contribution is a **Bidimensional Optimization Strategy** that unifies temporal and spatial optimization for multimodal reasoning. Specifically, we extend the existing trajectory-level optimization along the temporal dimension with a **Thought-level Optimization** along the spatial dimension, which decomposes the output into reasoning steps and leverages the diffusion generation schedule to identify causally significant tokens that mark the culmination of each reasoning step. To enable comprehensive evaluation at this granular level, we introduce a **Fine-grained Multimodal Alignment** mechanism that acts as a verifier, systematically assessing both the logical soundness of textual reasoning and its grounding in visual evidence. By concentrating rewards on these pivotal tokens based on their logical correctness and visual alignment, our framework creates a sparse yet precise optimization signal that couples the iterative refinement of diffusion with visually-grounded logical reasoning. We summarize our contributions as follows:

- We propose Bi-VRL, a reinforcement learning framework for multimodal diffusion language models that extends trajectory-level optimization to incorporate spatial reasoning structure.
- Our core contribution is a **bidimensional optimization strategy** that unifies temporal trajectory optimization with spatial thought-level optimization, augmented by fine-grained multimodal alignment, enabling comprehensive credit assignment across the diffusion process, logical reasoning steps, and visual grounding.
- We demonstrate the architectural generality of our framework by applying it to two distinct models: MMaDA (with discrete vision tokens) and LLaDA-V (with continuous vision embeddings).
- Through extensive experiments on challenging multimodal reasoning benchmarks, we show that Bi-VRL substantially outperforms both strong supervised fine-tuning baselines and alternative RL methods.

2 Related Work

Diffusion Language Model Early work on diffusion language models (DLMs) has split into two main lines: continuous methods that embed tokens in a continuous space and discrete methods that operate directly on the vocabulary through learned categorical transitions or masking schemes [1, 10, 13, 14, 34, 35, 37, 41, 55]. Masked diffusion language models have proved particularly effective and have now been scaled from billion-parameter to hundred-billion-parameter regimes, with open-source examples such as LLaDA [34], Dream [55], SDAR [8], and LLaDA2.X [2,3]. Building on this, MMaDA [54], LLaDA-V [56], Lumina-DiMOO [52, 53] and LaViDa-O [27] extend the paradigm to the multimodal setting, demonstrating its potential beyond unimodal language processing. Recent work

has also begun exploring reinforcement learning to further improve diffusion language models in both unimodal and multimodal settings.

Reinforcement Learning for Multimodal Reasoning Reinforcement learning (RL) has become a key technique for unlocking the reasoning capabilities of large language models [15, 17, 20, 21, 49, 58]. Recent work has started to extend this RL-based paradigm to the multimodal setting [36, 40], where models must not only follow complex instructions but also ground their reasoning in visual evidence, localize relevant regions, and maintain consistency between intermediate steps and final answers. Initial attempts have also explored applying RL to multimodal diffusion language models, with approaches employing random masking strategies [54] or coupled masking strategies [28]. However, these methods primarily rely on coarse outcome-level rewards and do not exploit the unique bidimensional structure of diffusion generation—specifically, the interplay between the temporal dimension of iterative refinement across denoising timesteps and the spatial dimension of reasoning structure across token positions. A central question remains: how to design a comprehensive optimization strategy that addresses both dimensions while ensuring fine-grained visual grounding and logical soundness?

3 Bi-VRL Framework

In this section, we present **Bi-VRL**, a **Bi**dimensional **V**isual **R**einforcement **L**earning framework designed to unlock the multimodal reasoning capabilities of large diffusion language models. An overview of the framework is shown in Figure 1. Bi-VRL is built to handle the challenges of applying reinforcement learning to the iterative, non-autoregressive generation process of diffusion models. We first describe a hybrid tokenization scheme that bridges the discrete nature of text with the continuous representation of visual data. We then introduce a bidimensional reward strategy that provides fine-grained optimization signals at both the full-generation trajectory level and the intermediate reasoning thought level. Finally, we detail a fine-grained multimodal alignment mechanism that tightly couples visual perception and textual generation throughout the reasoning process.

3.1 Hybrid Image Tokenization Scheme

To ground the model's reasoning in visual information, Bi-VRL supports two image tokenization strategies, discrete and continuous, allowing it to flexibly exploit the strengths of both representations.

Discrete Image Tokenization Creating unified architectures for both understanding and generation is common for multimodal dLLMs, given diffusion model's strength in image generation, as in MMaDA [54] and Lumina-DiMOO [53]. These models use a Vector-Quantized Variational Autoencoder (VQ-VAE) [45] to encode images into sequences of discrete integer indices from a learned codebook.

The resulting visual tokens are prepended to the text tokens, forming a unified input stream, so a single model processes visual and textual information in a shared format. To ensure the broad applicability of Bi-VRL, we explicitly support this discrete tokenization pathway.

Continuous Image Representation We also use the pre-trained vision encoder SigLIP [61] to map the entire image into dense, high-dimensional embeddings. These continuous embeddings are not part of the input token sequence; instead, they are injected into the diffusion model as conditioning via cross-attention. This provides a holistic semantic image representation that guides the iterative generation process. By supporting both discrete and continuous pathways, Bi-VRL can exploit fine-grained, localized information from discrete visual tokens and rich, global semantic context from continuous embeddings. This versatility is crucial for applying our reinforcement learning strategy across diverse multimodal reasoning tasks.

3.2 Bidimensional Optimization Strategy

The iterative refinement process of diffusion language models is not constrained by a fixed generation order, turning generation from a one-dimensional path into a two-dimensional space over token positions and diffusion timesteps. We argue that an effective RL framework must exploit this bidimensional structure, which existing methods largely overlook. To bridge this gap, we introduce a **Bidimensional Optimization Strategy** that delivers targeted optimization signals along both axes.

This strategy comprises three synergistic components. Along the temporal dimension, **trajectory-level optimization** evaluates the entire generation path, assigning credit and propagating value signals (returns and advantages) across diffusion timesteps. Along the spatial dimension, **thought-level optimization** provides fine-grained credit assignment by decomposing the output into reasoning steps and concentrating rewards on causally significant tokens that mark the culmination of each step. To enable this spatial optimization, **fine-grained multimodal alignment** acts as a verifier that systematically evaluates both the logical soundness of each reasoning step and its grounding in visual evidence, providing the reward signals for thought-level optimization. Further implementation details, including the pseudocode and evaluation prompts, are provided in the Appendix.

Trajectory-level Optimization For the temporal dimension, we build upon the TraceRL framework [48], which provides an efficient infrastructure for trajectory-aware optimization in diffusion language models. We extend TraceRL's trajectory representation and value learning mechanism to the multimodal setting, enabling our model to learn a temporally consistent value landscape that is tightly coupled to both visual and textual contexts. At its core is a value function V_ϕ, trained to estimate the expected future outcome from any given intermediate multimodal state.

We represent the visual input as a general visual context C_{img}, consistent with our hybrid tokenization scheme. It is either a sequence of discrete visual tokens, $C_{img} = \{v_1, v_2, \ldots, v_k\}$, or a set of continuous embedding vectors, $C_{img} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m\}$ extracted by a vision encoder. A state at diffusion timestep t is characterized by the textual component S_t, where tokens scheduled for generation at timesteps $t' \geq t$ are masked.

The value function does not operate on S_t in isolation, but instead on a fused multimodal representation. Let M denote the backbone of our diffusion model (e.g., a Transformer). Unlike autoregressive models, which rely on causal attention, M uses full (bidirectional) attention. This allows every element in the input sequence, including text tokens S_t and visual features C_{img}, to freely attend to one another, producing a tightly integrated hidden representation H_t:

$$H_t = M(S_t, C_{img}).$$

Our value function V_ϕ, parameterized by ϕ, is trained to estimate the expected future outcome from this rich multimodal hidden state, i.e., $V_\phi(H_t)$. A full generation trajectory is thus represented as a sequence of hidden states, $\tau = (H_0, H_1, \ldots, H_T)$. After the trajectory is completed, we evaluate the quality of the final reasoning output to obtain a reward signal.

To propagate this signal and obtain a dense, per-step optimization target, we adopt Generalized Advantage Estimation (GAE) [38]. For each multimodal state, the temporal-difference (TD) residual δ_t is defined as

$$\delta_t = r_t + \gamma V_\phi(H_{t+1}) - V_\phi(H_t),$$

where γ is the discount factor and r_t is the reward at step t. The GAE advantage $\hat{A}_t$ is then computed recursively as $\hat{A}_t = \delta_t + (\gamma\lambda)\hat{A}_{t+1}$. The learning target for the value function, i.e., the return $\hat{R}_t$, is given by $\hat{R}_t = \hat{A}_t + V_\phi(H_t)$. We will introduce how r_t is obtained in Sec. 3.2, as this plays a central role in aligning trajectory-level optimization with thought-level optimization.

The value function V_ϕ is updated by minimizing a clipped mean-squared error loss over the sequence of multimodal hidden states, through the following training objective [39]:

$$\mathcal{L}(\phi) = \mathbb{E}_\tau \left[\sum_{t=0}^{T} \max \left((V_\phi(H_t) - \hat{R}_t)^2, (V_{\text{clip}}(H_t) - \hat{R}_t)^2 \right) \right],$$

where $V_{\text{clip}}(H_t)$ denotes the clipped value prediction used to stabilize training. Through this process, V_ϕ learns to produce an accurate, step-by-step evaluation of the multimodal generation trajectory, yielding a dense optimization signal that guides the model toward high-quality reasoning grounded in the visual context.

Thought-level Optimization For the spatial dimension, we introduce the Thought-level Optimization strategy that provides fine-grained credit assignment across token positions. The key challenge is to deliver targeted supervision

to intermediate reasoning steps rather than relying solely on a coarse outcome-level reward. To address this, we leverage chunk-level reward signals R_k provided by our fine-grained multimodal alignment mechanism, which evaluates both the logical soundness and visual grounding of each reasoning step.

A key innovation of our approach lies in the design of the credit assignment mechanism. Let $T_{\text{map}} = (\tau_1, \ldots, \tau_L)$ denote the generation schedule of the diffusion process, where τ_j is the timestep at which token y_j is generated. Rather than uniformly distributing R_k across all tokens in chunk Z_k, we identify a small subset of causally significant tokens. To this end, we define a mapping function Φ that selects the indices of tokens generated at the latest timesteps within the concluding segment of each chunk:

$$\mathcal{K}_k = \Phi(Z_k, T_{map}).$$

These tokens, indexed by $\mathcal{K}_k$, mark the culmination of a reasoning step. By concentrating the reward signal on these pivotal points, we construct a sparse, token-level reward vector $\mathbf{r} = (r_1, \ldots, r_L)$, where each token reward r_j is defined as:

$$r_j = \begin{cases} R_k & \text{if } j \in \mathcal{K}_k \text{ for some } k \in \{1, \ldots, K\} \\ 0 & \text{otherwise} \end{cases}$$

This formulation yields a precise and effective gradient signal, encouraging the model to correctly complete each thought before moving on to the next. The resulting sparse reward vector $\mathbf{r}$ is then used to compute $\hat{A}_t$ (see Sec. 3.2). Finally, we perform thought-level optimization of the policy π_θ with the following objective:

$$\mathcal{L}(\theta) = \mathbb{E}_{\substack{Q \sim D_{\text{task}} \\ \tau \sim \pi_{\text{old}}(\cdot|Q)}} \Big[\sum_{t=1}^{T} C_\epsilon\Big(\frac{\pi_\theta(H_{t-1} \mid H_t)}{\pi_{\text{old}}(H_{t-1} \mid H_t)}, \hat{A}_t\Big) \Big] - \beta \text{KL},$$

where $C_\epsilon(r, A) \triangleq \min(rA, \text{clip}(r, 1-\epsilon, 1+\epsilon)A)$, $\text{KL} = \text{KL}(\pi_\theta || \pi_{\text{old}})$, π_{old} is the old policy, β and ϵ are hyperparameters.

By jointly optimizing $\mathcal{L}(\theta)$ and $\mathcal{L}(\phi)$, we align trajectory-level and thought-level optimization, with fine-grained multimodal alignment rewards, thereby achieving bidimensional optimization across both temporal and spatial dimensions.

Fine-grained Multimodal Alignment While the trajectory-level optimization provides a holistic value signal for the entire generation path, it does not explicitly penalize subtle logical errors or misalignments between textual assertions and visual evidence. To bridge this gap and provide the reward signals for thought-level optimization, our framework incorporates a dedicated Fine-grained Multimodal Alignment mechanism. This component acts as a verifier, systematically evaluating both the logical soundness of each reasoning step and its grounding in visual evidence.

Instead of relying on a single outcome reward o, this component delivers targeted rewards to intermediate reasoning steps. Given a complete generated response (at $t = 0$) represented as a token sequence $Y = (y_1, y_2, \ldots, y_L)$, we partition it into K contiguous, non-overlapping text chunks $Z_1, Z_2, \ldots, Z_K$, where each chunk Z_k serves as a proxy for one step in the model's chain of thought.

To assess both the logical correctness and visual grounding of each chunk Z_k, we employ a panel of powerful large vision-language models (LVLMs), $\mathcal{J}_1, \ldots, \mathcal{J}_N$, as automated judges. These judges receive both the input image and text, and are instructed to evaluate each chunk along several key dimensions:

- **Visual Alignment:** Does the chunk contradict or hallucinate information not present in the image?
- **Logical Soundness:** Does the chunk contain mathematical errors, flawed reasoning, or fallacious arguments?
- **Absence of Justification:** Does the chunk present an answer or conclusion without the necessary preceding steps or justification (i.e., "answer-only")?
- **Internal Consistency:** Does the chunk contradict information presented in previous chunks of the same response or the initial problem statement?

To facilitate this nuanced evaluation, for each chunk we construct a comprehensive prompt, P_k:

$$P_k = \text{PromptBuilder}(Q, A_{gt}, Y, c(Y), Z_k),$$

where Q is the textual problem statement, A_{gt} is the ground-truth answer, Y is the full generated response with its final correctness label $c(Y)$, and Z_k is the specific chunk to be scored.

Each judge $\mathcal{J}_n$ processes the prompt P_k along with the input image and outputs a score $s_{n,k} \in \{-1, 1\}$. To further improve robustness and reduce the bias of any single judge, we combine the thought-level rewards $s_{n,k}$ with the verifiable outcome reward $o \in \{-1, 1\}$ obtained in the RL training data. The final integrated reward for chunk k is then given by:

$$R_k = \begin{cases} s_{1,k}, & \text{if } s_{1,k} = s_{2,k} = \cdots = s_{N,k}, \\ o, & \text{otherwise.} \end{cases}$$

This evaluation is fully parallelized, with dedicated inference engines processing batches of prompts to maximize throughput. This mechanism creates a dense reward landscape that directly penalizes logical errors and deviations from visual evidence, compelling the model to generate text that is both logically coherent and rigorously grounded in the provided image. These chunk-level rewards R_k serve as the foundation for the thought-level optimization described in Sec. 3.2.

4 Experiments

4.1 Experimental Setup

Models To rigorously validate the architectural agnosticism of our Bi-VRL framework, we instantiate it on two distinct multimodal diffusion models, each representing one of the pathways in our proposed *Hybrid Image Tokenization Scheme*.

For the discrete tokenization pathway, we employ MMaDA [54], a model with a unified understanding-generation architecture. MMaDA leverages a pretrained MAGVIT-v2 [57] quantizer to transform a 512×512 pixel image into a sequence of 1024 discrete visual tokens, which are then prepended to the text sequence to form a homogeneous input stream. In contrast, for the continuous representation pathway, we utilize LLaDA-V [56]. This model adopts a powerful siglip2-so400m-patch14-384 vision tower [44] to encode the image into high-dimensional continuous embeddings, which are injected into the diffusion model as a conditioning signal via a two-layer MLP projector. By demonstrating the efficacy of Bi-VRL on these two architecturally disparate models, we substantiate its flexibility and broad applicability across different paradigms of visual data integration.

Datasets and Benchmarks Our training dataset is curated by sourcing visual reasoning problems from established collections, including GeoQA [4], CLEVR [22], MMK12 [33], and GQA [18]. We employ a rigorous filtering process, leveraging state-of-the-art models as judges to select questions of moderate difficulty. To ensure a rich diversity of problem types for robust learning, this filtered set is then programmatically processed to encompass three distinct formats: numeric answer, multiple-choice, and binary (true/false) questions. To monitor performance curves during training, we use the MathVerse-testmini. For the final, comprehensive performance evaluation, we assess our models on a wide array of challenging benchmarks: MMMU [59], MMMU-Pro [60], MMStar [6], MME [11], SEEDBench [26], MMBench [31], MathVerse [62].

Implementation Details During the RL fine-tuning process, we sample a mini-batch of 32 prompts at each optimization step. For each prompt, we generate 8 candidate responses and filter out instances where all responses are either correct or incorrect to ensure a meaningful training signal. Following LLaDA [34], we employ a semi-autoregressive generation scheme with a maximum length of 256 tokens and a block size of 32. To align with this paradigm, the chunk length for our thought-level optimization is set to 128. A sampling temperature of 0.8 is used during training. The policy model is updated with a learning rate of 1×10^{-6}, while the value model, when applicable, uses a learning rate of 5×10^{-6}. For evaluation, we maintain the same settings but reduce the temperature to 0.1 to favor more deterministic outputs. And we sample each question 3 times and take the average as the final score.

4.2 Main Results

We evaluate the effectiveness of Bi-VRL by applying it to two architecturally distinct multimodal diffusion models: LLaDA-V (continuous vision embeddings) and MMaDA (discrete vision tokens). Table 1 presents a comprehensive comparison across seven challenging multimodal reasoning benchmarks. Our framework delivers substantial and consistent improvements across both model architectures and nearly all evaluation benchmarks, validating that Bi-VRL is an effective post-training method for enhancing multimodal reasoning capabilities.

Table 1: Performance gains from the Bi-VRL framework on major multimodal reasoning benchmarks. Our enhanced models demonstrate significant improvements over their respective baselines. Other contemporary models are listed for contextual reference.

Model	Arch.	MMMU	MMMU-Pro	MMStar	MME	SeedB	MMB	MathVerse
LLaVA-Phi [64]	AR	-	-	-	-/1335	-	59.8	-
LLaVA-1.5 [29]	AR	-	-	-	-/1510	66.1	64.3	-
Qwen2-VL [46]	AR	54.1	43.5	60.7	-	-	-	-
DeepSeek-VL2 [50]	AR	51.1	-	61.3	-	-	-	-
Janus-Pro [7]	AR	41.0	-	-	-/1567	72.1	79.2	-
Emu3 [47]	AR	31.6	-	-	-	68.2	58.5	-
MAmmoTH [16]	AR	50.8	-	63.0	-	76.0	-	34.2
LLaVA-OV [25]	AR	48.8	-	61.7	418/1580	75.4	80.8	26.2
BAGEL [9]	AR+Diff.	55.3	-	-	-/1610	-	79.2	-
BLIP3-o [5]	AR+Diff.	50.6	-	-	647/1682	77.5	83.5	-
MetaMorph [43]	AR+Diff.	41.8	-	-	-	71.8	75.2	-
Show-o [51]	AR+Diff.	27.4	-	-	-/1232	-	-	-
JanusFlow [32]	AR+Diff.	29.3	-	-	-/1333	70.5	74.9	-
Orthus [23]	AR+Diff.	28.2	-	-	-/1265	-	-	-
LaViDa-R1 [28]	Diff.	-	32.8	-	-	-	-	38.7
LLaDA-V	Diff.	48.6	35.2	60.1	491/1507	74.8	82.9	28.5
LLaDA-V + Bi-VRL (ours)	Diff.	**53.4**	**38.3**	**63.2**	**511/1554**	**81.1**	**85.4**	**31.2**
MMaDA	Diff.	30.2	26.4	37.6	401/1410	64.2	**68.5**	23.5
MMaDA + Bi-VRL (ours)	Diff.	**33.8**	**30.2**	**40.0**	**420/1487**	**64.6**	68.1	**28.3**

Notably, the effectiveness of Bi-VRL across two fundamentally different architectures demonstrates the broad applicability of our bidimensional optimization strategy, showing that our framework can be readily adapted to various multimodal diffusion model designs without requiring architecture-specific modifications.

4.3 Ablation Study

To understand the contribution of each component in our Bi-VRL framework and validate the effectiveness of our bidimensional optimization strategy, we conduct comprehensive ablation studies. We analyze both the individual contributions of our three core components and compare our approach against alternative RL methods. The SFT baseline is obtained by briefly fine-tuning the original models to align their output format with our automated answer verification, which may result in slightly lower scores than originally reported.

Component Ablation We systematically ablate each of the three core components of our framework—trajectory-level optimization, thought-level optimization, and fine-grained multimodal alignment—to assess their individual contributions. Table 2 presents the results on both LLaDA-V and MMaDA across seven benchmarks.

The results reveal several key insights. First, removing trajectory-level optimization causes the most significant performance drop across most benchmarks,

Table 2: Component ablation study. We remove each component individually to assess its contribution. All three components provide complementary benefits, and their synergistic combination yields the best overall performance.

LLaDA-V	**MMMU**	**MMMU-P**	**MMStar**	**MME**	**SeedB**	**MMB**	**MathV**
SFT Baseline	48.6	35.2	60.1	491/1507	74.8	82.9	28.5
Ours (Full)	**53.4**	**38.3**	**63.2**	**511/1554**	81.1	**85.4**	**31.2**
- Trajectory-level Opt.	51.6	37.5	61.3	504/1540	78.8	84.5	29.7
- Fine-grained Align.	52.7	37.9	61.6	504/1550	78.6	84.1	30.5
- Thought-level Opt.	53.0	38.2	62.8	508/1534	**81.4**	**85.4**	30.9
MMaDA	**MMMU**	**MMMU-P**	**MMStar**	**MME**	**SeedB**	**MMB**	**MathV**
SFT Baseline	30.2	26.4	37.6	401/1410	64.2	68.5	23.5
Ours (Full)	**33.8**	**30.2**	**40.0**	**420/1487**	64.6	68.1	**28.3**
- Trajectory-level Opt.	31.6	28.6	39.1	409/1425	64.3	68.4	26.6
- Fine-grained Align.	32.6	28.5	38.9	408/1470	**64.9**	68.0	27.0
- Thought-level Opt.	33.5	29.7	39.8	416/1473	64.7	**68.8**	27.8

highlighting its foundational role in providing holistic credit assignment across the diffusion trajectory. Second, removing fine-grained multimodal alignment leads to notable degradation, particularly on benchmarks requiring strong visual grounding (e.g., MMMU, MMStar), confirming its effectiveness in penalizing visual hallucinations and enforcing faithful perception. Third, removing thought-level optimization results in smaller but consistent drops, especially on reasoning-intensive benchmarks like MathVerse, demonstrating its value in providing fine-grained supervision at the reasoning step level.

Importantly, the full model with all three components consistently achieves the best or near-best performance across all benchmarks, validating the synergistic effect of our bidimensional optimization strategy. Each component addresses a complementary aspect of the optimization challenge: trajectory-level optimization handles temporal credit assignment, thought-level optimization provides spatial granularity, and fine-grained multimodal alignment ensures visual grounding. Their combination yields a comprehensive and robust optimization signal.

Comparison with Alternative RL Methods To demonstrate the superiority of our bidimensional optimization strategy, we compare Bi-VRL against two simpler RL baselines representing existing approaches in the literature: (1) **Random Masking**, which randomly masks a subset of tokens and applies PPO with outcome-level rewards (analogous to methods used in d1 [63] and MMaDA [54]), and (2) **Coupled-GRPO**, which uses paired complementary masks to ensure coverage (similar to approaches in DiffuCoder [12] and LaViDa-R1 [28]). Table 3 presents the comprehensive comparison.

The results reveal a clear performance hierarchy. Random Masking, which applies RL without sophisticated credit assignment, provides only marginal improvements over the SFT baseline. Coupled-GRPO achieves better results through improved coverage but still falls significantly short of our method. In contrast,

Table 3: Comparison with alternative RL methods. Our Bi-VRL framework substantially outperforms simpler RL baselines that lack bidimensional optimization, demonstrating the importance of our trajectory-level and thought-level credit assignment strategy.

LLaDA-V	**MMMU**	**MMMU-P**	**MMStar**	**MME**	**SeedB**	**MMB**	**MathV**
SFT Baseline	48.6	35.2	60.1	491/1507	74.8	82.9	28.5
Random Masking	49.9	35.2	60.0	497/1534	75.2	84.0	28.8
Coupled-GRPO	51.8	37.0	60.6	499/1550	77.9	83.5	29.2
Bi-VRL (Ours)	**53.4**	**38.3**	**63.2**	**511/1554**	**81.1**	**85.4**	**31.2**
MMaDA	**MMMU**	**MMMU-P**	**MMStar**	**MME**	**SeedB**	**MMB**	**MathV**
SFT Baseline	30.2	26.4	37.6	401/1410	64.2	68.5	23.5
Random Masking	30.7	27.6	38.1	410/1443	63.9	68.0	25.5
Coupled-GRPO	30.9	28.1	38.6	417/1462	64.4	**68.8**	26.3
Bi-VRL (Ours)	**33.8**	**30.2**	**40.0**	**420/1487**	**64.6**	68.1	**28.3**

Bi-VRL substantially outperforms both baselines across nearly all benchmarks, validating our core hypothesis: effective RL for diffusion language models requires bidimensional optimization that addresses both the temporal dimension (via trajectory-level value learning) and the spatial dimension (via thought-level credit assignment with fine-grained multimodal alignment). Existing approaches that rely solely on outcome-level rewards with simple masking strategies fail to exploit the unique bidimensional structure of diffusion generation and consequently achieve limited improvements.

4.4 Qualitative Analysis

To provide deeper insight into how our fine-grained multimodal alignment mechanism improves reasoning quality, we present qualitative case studies in Figure 2. These examples illustrate how our framework corrects different types of reasoning failures by enforcing both logical soundness and visual grounding through the alignment scores produced by our LVLM judges.

Visual Hallucination Correction. The first example (top row) demonstrates how our framework addresses visual hallucination—a critical failure mode where the model generates content unsupported by the image. The base MMaDA model incorrectly claims the tree suffers from a "nematode attack," a concept not present in the visual evidence, resulting in a low alignment score of 0.694. This hallucination occurs because the model lacks fine-grained supervision on the visual grounding of its reasoning chunks.

In contrast, our Bi-VRL-enhanced model (MMaDA†) begins with a visually grounded observation: "The tree is placed on a piece of paper, which is not typical for a pot." This accurate perception is rewarded by our fine-grained multimodal alignment mechanism, which employs LVLM judges to evaluate each reasoning

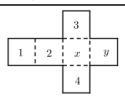
	Base MMaDA	**MMaDA† (with dLLM-VRL)**
 **Question:** What seems to be the issue with this young citrus tree? (Answer: E. Pot bound)	The image shows a young citrus tree with a visible root system. The roots appear to be thin and brittle, which is characteristic of a **nematode attack**. Nematodes are known to cause root issues in citrus trees. This suggests that there is a problem with the tree. Final Answer: **B**	To determine the issue with the young citrus tree, I need to analyze the image. The image shows a young citrus tree with a lot of thin, brown roots. These roots are visible and appear to be weak ... **The tree is placed on a piece of paper, which is not typical for a pot**. The presence of the roots and the paper suggests that the tree might not be in a pot or a container... Final Answer: **E**
	Alignment Score: 0.694	**Alignment Score:** 0.795
 **Question:** As shown in the figure, △ABC is inscribed in ⊙O, ∠BAC = 120°, AB = AC = 4, and BD is the diameter of ⊙O. Then BD = _____. (Answer: 8)	In triangle BAD, since BD is the diameter, ∠BAD = 90°. And we can see **AD = AB**. Given AB = 4, we assume AD = 4. By the Pythagorean theorem, BD = $\sqrt{(4^2 + 4^2)} = 4\sqrt{2}$. Final Answer: **$4\sqrt{2}$**	In △ABC, since AB = AC = 4 and ∠BAC = 120°, we find ∠BCA = (180° - 120°) / 2 = 30°. **Angles subtended by the same arc are equal, so ∠BDA = ∠BCA = 30°**. In △BAD, since BD is the diameter, ∠BAD = 90°. Using trigonometry, BD = AB / sin(∠BDA) = 4 / sin(30°) = 8. So BD = 8. Final Answer: **8**
	Alignment Score: 0.571	**Alignment Score:** 0.857
 **Question:** The figure shows the unfolded net of a cubic box, with different numbers and letters on its surfaces. When folded along the dotted lines to form a cube, the sum of the numbers on opposite faces is equal. Then, x+y= _____ . (Answer: 11)	When folded, the sum of opposite faces is constant. In this net, the top face '3' and the bottom face '4' are opposite. Therefore, the constant sum for any pair of opposite faces is 3 + 4 = 7. Face '1' is opposite to face 'x', so $1 + x = 7$, which means $x = 6$. **The face with the number '3' is opposite to face 'y'**, so $3 + y = 7$, which means $y = 4$. Therefore, $x + y = 6 + 4 = 10$. Final Answer: **10**	When the net is folded, **the pairs of opposite faces are (1, x), (2, y), and (3, 4)**. We can find the constant sum using the pair of faces with known numbers, which are '3' and '4'. The sum is 3 + 4 = 7. This means that the sum of numbers on every pair of opposite faces must be 7. For the pair (1, x), their sum must be 7. This gives us the equation 1 + x = 7, which means x = 6.For the pair (2, y), their sum must also be 7. This gives us the equation 2 + y = 7, which means y = 5. Final Answer: **11**
	Alignment Score: 0.615	**Alignment Score:** 0.846

Fig. 2: Qualitative comparison showing how Bi-VRL corrects reasoning failures through fine-grained multimodal alignment. Each row shows: (left) the question with image, (middle) base model output with alignment score, (right) our enhanced model output with improved alignment score. The alignment scores are computed by our LVLM judges evaluating both logical correctness and visual grounding of reasoning chunks. Green highlights indicate correctly grounded observations; red highlights indicate errors.

chunk against the image. Specifically, the judges assess whether the chunk contradicts or hallucinates information not present in the image (the *Visual Alignment* criterion). By assigning positive rewards R_k to chunks with faithful visual grounding and negative rewards to hallucinated content, our thought-level optimization concentrates these rewards on causally significant tokens, guiding the model toward correct reasoning. The improved alignment score of 0.795 quantitatively reflects this enhanced visual fidelity.

Logical Grounding Enforcement. The second example (middle row) highlights how our framework enforces logical consistency with visual evidence. The base model makes an unsubstantiated assumption ("AD = AB") that is not supported by the geometric properties shown in the image, leading to an alignment score of 0.571. This error stems from insufficient supervision on the logical soundness of intermediate reasoning steps.

Our enhanced model correctly applies a valid geometric theorem: "Angles subtended by the same arc are equal, so $\angle$BDA = $\angle$BCA = 30°." This reasoning is both logically sound and derivable from the visual configuration. Our fine-grained alignment mechanism rewards this correct reasoning through the *Logical*

Soundness criterion, which checks whether chunks contain mathematical errors or flawed reasoning. The trajectory-level optimization then propagates these rewards across the diffusion trajectory, enabling the model to learn that logical consistency with visual evidence leads to better outcomes. The alignment score improves to 0.857, confirming the enhanced logical grounding.

Spatial Understanding Improvement. The third example (bottom row) showcases improved spatial reasoning. The base model incorrectly identifies opposite faces of the unfolded cube ("The face with the number '3' is opposite to face 'y' "), achieving an alignment score of 0.615. This spatial misinterpretation indicates insufficient visual-spatial reasoning capability.

Our enhanced model correctly identifies all pairs of opposite faces: "the pairs of opposite faces are (1, x), (2, y), and (3, 4)." This accurate spatial interpretation is enabled by our bidimensional optimization strategy. The fine-grained alignment mechanism evaluates the spatial correctness of each reasoning chunk against the image, while the thought-level optimization delivers targeted rewards to the tokens marking the culmination of each reasoning step (as defined by the mapping function Φ in Sec. 3.2). The improved alignment score of 0.846 demonstrates the enhanced spatial understanding.

Summary. These examples collectively demonstrate that our fine-grained multimodal alignment mechanism acts as an effective verifier, systematically penalizing deviations from visual evidence across multiple error types. The alignment scores, computed by LVLM judges evaluating both textual reasoning and visual grounding, provide interpretable measures of improvement. Crucially, these qualitative improvements reflect the synergistic interplay of our three core components. Beyond these specific cases, we observe similar patterns of improvement across our training data, where the model progressively learns to anchor its reasoning in observable visual features rather than making unwarranted assumptions. The interpretability afforded by our chunk-level alignment scores also facilitates error diagnosis and iterative refinement, making our framework particularly valuable for applications requiring high reliability and transparency in multimodal reasoning.

5 Conclusion

We presented Bi-VRL, a reinforcement learning framework for multimodal diffusion language models. Our bidimensional optimization strategy addresses the iterative, non-autoregressive generation process by combining three synergistic components: trajectory-level optimization for holistic credit assignment across diffusion timesteps, thought-level optimization for fine-grained supervision on reasoning steps, and fine-grained multimodal alignment for visual grounding and logical soundness.

Experiments on two architecturally distinct models—LLaDA-V with continuous vision embeddings and MMaDA with discrete vision tokens—demonstrated

the effectiveness and broad applicability of our framework. Ablation studies confirmed that all three components contribute complementary benefits, yielding substantial performance improvements across seven challenging multimodal reasoning benchmarks. Qualitative analysis revealed that our framework effectively corrects visual hallucinations, logical inconsistencies, and spatial misinterpretations.

This work establishes bidimensional optimization as an effective paradigm for post-training multimodal diffusion language models, providing a foundation for future research in reinforcement learning for non-autoregressive generative models.

Acknowledgments

This work is in part supported by the Stanford Institute for Human-Centered Artificial Intelligence (HAI), Walter and Sandra Tseng, and the Clement and Xinxin Foundation.

References

1. Austin, J., Johnson, D.D., Ho, J., Tarlow, D., Van Den Berg, R.: Structured denoising diffusion models in discrete state-spaces. Advances in neural information processing systems **34**, 17981–17993 (2021)
2. Bie, T., Cao, M., Cao, X., Chen, B., Chen, F., Chen, K., Du, L., Feng, D., Feng, H., Gong, M., Gong, Z., Gu, Y., Guan, J., Guan, K., He, H., Huang, Z., Jiang, J., Jiang, Z., Lan, Z., Li, C., Li, J., Li, Z., Liu, H., Liu, L., Lu, G., Lu, Y., Ma, Y., Mou, X., Pan, Z., Qiu, K., Ren, Y., Tan, J., Tian, Y., Wang, Z., Wei, L., Wu, T., Xing, Y., Ye, W., Zha, L., Zhang, T., Zhang, X., Zhao, J., Zheng, D., Zhong, H., Zhong, W., Zhou, J., Zhou, J., Zhu, L., Zhu, M., Zhuang, Y.: Llada2.1: Speeding up text diffusion via token editing (2026), https://arxiv.org/abs/2602.08676
3. Bie, T., Cao, M., Chen, K., Du, L., Gong, M., Gong, Z., Gu, Y., Hu, J., Huang, Z., Lan, Z., Li, C., Li, C., Li, J., Li, Z., Liu, H., Liu, L., Lu, G., Lu, X., Ma, Y., Tan, J., Wei, L., Wen, J.R., Xing, Y., Zhang, X., Zhao, J., Zheng, D., Zhou, J., Zhou, J., Zhou, Z., Zhu, L., Zhuang, Y.: Llada2.0: Scaling up diffusion language models to 100b (2025), https://arxiv.org/abs/2512.15745
4. Chen, J., Tang, J., Qin, J., Liang, X., Liu, L., Xing, E., Lin, L.: Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 513–523 (2021)
5. Chen, J., Xu, Z., Pan, X., Hu, Y., Qin, C., Goldstein, T., Huang, L., Zhou, T., Xie, S., Savarese, S., et al.: Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. arXiv preprint arXiv:2505.09568 (2025)
6. Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., et al.: Are we on the right way for evaluating large vision-language models? Advances in Neural Information Processing Systems **37**, 27056–27087 (2024)
7. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)

8. Cheng, S., Bian, Y., Liu, D., Jiang, Y., Liu, Y., Zhang, L., Wang, W., Guo, Q., Chen, K., Qi, B., et al.: Sdar: A synergistic diffusion-autoregression paradigm for scalable sequence generation. arXiv preprint arXiv:2510.06303 (2025)
9. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., Shi, G., Fan, H.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025). https://doi.org/10.48550/arXiv.2505.14683
10. Dieleman, S., Sartran, L., Roshannai, A., Savinov, N., Ganin, Y., Richemond, P.H., Doucet, A., Strudel, R., Dyer, C., Durkan, C., et al.: Continuous diffusion for categorical data. arXiv preprint arXiv:2211.15089 (2022)
11. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2025)
12. Gong, S., Zhang, R., Zheng, H., Gu, J., Jaitly, N., Kong, L., Zhang, Y.: Diffucoder: Understanding and improving masked diffusion models for code generation. arXiv preprint arXiv:2506.20639 (2025)
13. Graves, A., Srivastava, R.K., Atkinson, T., Gomez, F.: Bayesian flow networks. arXiv preprint arXiv:2308.07037 (2023)
14. Gulrajani, I., Hashimoto, T.B.: Likelihood-based diffusion language models. Advances in Neural Information Processing Systems **36**, 16693–16715 (2023)
15. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
16. Guo, J., Zheng, T., Li, Y., Bai, Y., Li, B., Wang, Y., Zhu, K., Neubig, G., Chen, W., Yue, X.: 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). pp. 13869–13920 (2025)
17. Hu, J., Zhang, Y., Han, Q., Jiang, D., Zhang, X., Shum, H.Y.: Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. arXiv preprint arXiv:2503.24290 (2025)
18. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6700–6709 (2019)
19. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
20. Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., et al.: Openai o1 system card. arXiv preprint arXiv:2412.16720 (2024)
21. Jiang, P., Lin, J., Cao, L., Tian, R., Kang, S., Wang, Z., Sun, J., Han, J.: Deepretrieval: Hacking real search engines and retrievers with large language models via reinforcement learning. arXiv preprint arXiv:2503.00223 (2025)
22. Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2901–2910 (2017)
23. Kou, S., Jin, J., Liu, Z., Liu, C., Ma, Y., Jia, J., Chen, Q., Jiang, P., Deng, Z.: Orthus: Autoregressive interleaved image-text generation with modality-specific heads. arXiv preprint arXiv:2412.00127 (2024)

24. Labs, I., Khanna, S., Kharbanda, S., Li, S., Varma, H., Wang, E., Birnbaum, S., Luo, Z., Miraoui, Y., Palrecha, A., Ermon, S., Grover, A., Kuleshov, V.: Mercury: Ultra-fast language models based on diffusion (2025), https://arxiv.org/abs/2506.17298
25. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
26. Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., Shan, Y.: Seed-bench: Benchmarking multimodal llms with generative comprehension. arXiv preprint arXiv:2307.16125 (2023)
27. Li, S., Gu, J., Liu, K., Lin, Z., Wei, Z., Grover, A., Kuen, J.: Lavida-o: Elastic large masked diffusion models for unified multimodal understanding and generation (2025), https://arxiv.org/abs/2509.19244
28. Li, S., Zhu, Y., Gu, J., Liu, K., Lin, Z., Chen, Y., Tao, M., Grover, A., Kuen, J.: Lavida-r1: Advancing reasoning for unified multimodal diffusion language models (2026), https://arxiv.org/abs/2602.14147
29. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
30. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
31. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European conference on computer vision. pp. 216–233. Springer (2024)
32. Ma, Y., Liu, X., Chen, X., Liu, W., Wu, C., Wu, Z., Pan, Z., Xie, Z., Zhang, H., Yu, X., et al.: Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7739–7751 (2025)
33. Meng, F., Du, L., Liu, Z., Zhou, Z., Lu, Q., Fu, D., Han, T., Shi, B., Wang, W., He, J., et al.: Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. arXiv preprint arXiv:2503.07365 (2025)
34. Nie, S., Zhu, F., You, Z., Zhang, X., Ou, J., Hu, J., Zhou, J., Lin, Y., Wen, J.R., Li, C.: Large language diffusion models. arXiv preprint arXiv:2502.09992 (2025)
35. Ou, J., Nie, S., Xue, K., Zhu, F., Sun, J., Li, Z., Li, C.: Your absorbing discrete diffusion secretly models the conditional distributions of clean data. arXiv preprint arXiv:2406.03736 (2024)
36. Peng, Y., Zhang, G., Zhang, M., You, Z., Liu, J., Zhu, Q., Yang, K., Xu, X., Geng, X., Yang, X.: Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. arXiv preprint arXiv:2503.07536 (2025)
37. Sahoo, S., Arriola, M., Schiff, Y., Gokaslan, A., Marroquin, E., Chiu, J., Rush, A., Kuleshov, V.: Simple and effective masked diffusion language models. Advances in Neural Information Processing Systems **37**, 130136–130184 (2024)
38. Schulman, J., Moritz, P., Levine, S., Jordan, M., Abbeel, P.: High-dimensional continuous control using generalized advantage estimation. arXiv preprint arXiv:1506.02438 (2015)
39. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
40. Shen, H., Liu, P., Li, J., Fang, C., Ma, Y., Liao, J., Shen, Q., Zhang, Z., Zhao, K., Zhang, Q., et al.: Vlm-r1: A stable and generalizable r1-style large vision-language model. arXiv preprint arXiv:2504.07615 (2025)

41. Shi, J., Han, K., Wang, Z., Doucet, A., Titsias, M.: Simplified and generalized masked diffusion for discrete data. Advances in neural information processing systems **37**, 103131–103167 (2024)
42. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
43. Tong, S., Fan, D., Li, J., Xiong, Y., Chen, X., Sinha, K., Rabbat, M., LeCun, Y., Xie, S., Liu, Z.: Metamorph: Multimodal understanding and generation via instruction tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17001–17012 (2025)
44. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
45. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
46. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
47. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024)
48. Wang, Y., Yang, L., Li, B., Tian, Y., Shen, K., Wang, M.: Revolutionizing reinforcement learning framework for diffusion large language models. arXiv preprint arXiv:2509.06949 (2025)
49. Wang, Y., Yang, L., Tian, Y., Shen, K., Wang, M.: Cure: Co-evolving coders and unit testers via reinforcement learning. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
50. Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al.: Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. arXiv preprint arXiv:2412.10302 (2024)
51. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. arXiv preprint arXiv:2408.12528 (2024)
52. Xin, Y., Luo, S., Qin, Q., Chen, H., Zhu, K., Zhang, Z., He, Y., Zhang, R., Bai, J., Cao, S., et al.: dmllm-tts: Self-verified and efficient test-time scaling for diffusion multi-modal large language models. arXiv preprint arXiv:2512.19433 (2025)
53. Xin, Y., Qin, Q., Luo, S., Zhu, K., Yan, J., Tai, Y., Lei, J., Cao, Y., Wang, K., Wang, Y., et al.: Lumina-dimoo: An omni diffusion large language model for multimodal generation and understanding. arXiv preprint arXiv:2510.06308 (2025)
54. Yang, L., Tian, Y., Li, B., Zhang, X., Shen, K., Tong, Y., Wang, M.: Mmada: Multimodal large diffusion language models. arXiv preprint arXiv:2505.15809 (2025)
55. Ye, J., Xie, Z., Zheng, L., Gao, J., Wu, Z., Jiang, X., Li, Z., Kong, L.: Dream 7b: Diffusion large language models. arXiv preprint arXiv:2508.15487 (2025)
56. You, Z., Nie, S., Zhang, X., Hu, J., Zhou, J., Lu, Z., Wen, J.R., Li, C.: Llada-v: Large language diffusion models with visual instruction tuning. arXiv preprint arXiv:2505.16933 (2025)
57. Yu, L., Lezama, J., Gundavarapu, N.B., Versari, L., Sohn, K., Minnen, D., Cheng, Y., Birodkar, V., Gupta, A., Gu, X., et al.: Language model beats diffusion–tokenizer is key to visual generation. arXiv preprint arXiv:2310.05737 (2023)

58. Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, L., et al.: Dapo: An open-source llm reinforcement learning system at scale. arXiv preprint arXiv:2503.14476 (2025)
59. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., 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. pp. 9556–9567 (2024)
60. Yue, X., Zheng, T., Ni, Y., Wang, Y., Zhang, K., Tong, S., Sun, Y., Yu, B., Zhang, G., Sun, H., et al.: Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 15134–15186 (2025)
61. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
62. Zhang, R., Jiang, D., Zhang, Y., Lin, H., Guo, Z., Qiu, P., Zhou, A., Lu, P., Chang, K.W., Qiao, Y., et al.: Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In: European Conference on Computer Vision. pp. 169–186. Springer (2024)
63. Zhao, S., Gupta, D., Zheng, Q., Grover, A.: d1: Scaling reasoning in diffusion large language models via reinforcement learning. arXiv preprint arXiv:2504.12216 (2025)
64. Zhu, Y., Zhu, M., Liu, N., Xu, Z., Peng, Y.: Llava-phi: Efficient multi-modal assistant with small language model. In: Proceedings of the 1st International Workshop on Efficient Multimedia Computing under Limited. pp. 18–22 (2024)