SDAR-VL: Stable and Efficient Block-wise Diffusion for Vision-Language Understanding

Anonymous ACL submission

Abstract

Block-wise discrete diffusion offers an attractive balance between parallel generation and causal dependency modeling, making it a promising backbone for vision-language modeling. However, its practical adoption has been limited by high training cost, slow convergence, and instability, which have so far kept it behind strong autoregressive (AR) baselines. We present **SDAR-VL**, the first systematic application of block-wise discrete diffusion to large-scale vision-language understanding (VLU), together with an *integrated framework for efficient and stable training*. This framework unifies three components: 1) **Asynchronous Block-wise Noise Scheduling** to diversify supervision within each batch; 2) **Effective Mask Ratio Scaling** for unbiased loss normalization under stochastic masking; and 3) a **Progressive Beta Noise Curriculum** that increases effective mask coverage while preserving corruption diversity. Experiments on 21 single-image, multi-image, and video benchmarks show that SDAR-VL consistently improves *training efficiency*, *convergence stability*, and *task performance* over conventional block diffusion. On this evaluation suite, SDAR-VL sets a new state of the art among diffusion-based vision-language models and, under matched settings, matches or surpasses strong AR baselines such as LLaVA-OneVision as well as the global diffusion baseline LLaDA-V, establishing block-wise diffusion as a practical backbone for VLU.

1 Introduction

Vision-Language Models (VLMs) are becoming the backbone of multimodal AI, and robust multimodal *understanding* is a necessary foundation for future unified systems that can both understand and generate visual–textual content (Comanici et al., 2025; Hurst et al., 2024). Current state-of-the-art vision-language understanding models are dominated by autoregressive (AR) decoders such as Qwen-VL, InternVL, and the LLaVA family (Bai et al., 2025; Wang et al., 2025a; Li et al., 2024a). Their left-to-right factorization is well aligned with text generation, but also enforces a strict causal order that limits parallelism and makes it difficult to support the flexible decoding patterns desired in multimodal generation (Xin et al., 2025).

Discrete diffusion models (Austin et al., 2021; Gulrajani and Hashimoto, 2023) offer an alternative sequence modeling paradigm: sequences are generated by iterative denoising from corrupted inputs, enabling parallel decoding and bidirectional context. They have already shown strong performance in multimodal *generation* tasks such as image synthesis and infilling (Xin et al., 2025; Li et al., 2025a; Yang et al., 2025a), and are therefore promising as a unified backbone for both understanding and generation (Zhu et al., 2025a; Ye et al., 2025). However, existing *global* discrete diffusion approaches face two key obstacles when scaled to vision-language understanding: (1) full-sequence bidirectional denoising leads to per-step quadratic complexity and low throughput, motivating KV-cache–style accelerations (Peng et al., 2025; Wu et al., 2025b; Ma et al., 2025); and (2) the lack of causal inductive bias makes long-form text generation difficult, so practical systems often fall back to semi-autoregressive schemes (Zhu et al., 2025a; Yang et al., 2025c). These adaptations implicitly approximate block-wise causal processing, suggesting that a more principled block-wise diffusion formulation could be better suited to large-scale multimodal understanding.

Block-wise diffusion models (Arriola et al., 2025; Cheng et al., 2025; Wang et al., 2025b) precisely follow this idea by partitioning sequences into autoregressive blocks, while performing masked diffusion in parallel *within* each block. This preserves inter-block causal dependencies, reduces the computational cost of global diffusion, and still benefits from local bidirectional context.

Despite encouraging results in pure-language settings (Cheng et al., 2025), it remains unclear whether block-wise discrete diffusion can serve as a competitive and scalable backbone for vision-language understanding.

In this work, we introduce **SDAR-VL**, the first block-wise discrete diffusion framework for large-scale vision-language understanding. Built on the Block Discrete Denoising Diffusion (BD3) formulation (Arriola et al., 2025; Cheng et al., 2025), SDAR-VL targets the understanding side and addresses three BD3 training bottlenecks—high gradient variance, biased loss scaling under stochastic masking, and suboptimal noise scheduling—via an integrated framework (Sec. 3): (1) **Asynchronous Block-wise Noise Schedule** to smooth optimization by sampling noise per block; (2) **Effective Mask Ratio Scaling** to normalize loss using the realized mask ratio; and (3) **Progressive Beta Noise Curriculum** to shift toward higher mask ratios while preserving corruption diversity.

We instantiate SDAR-VL with 4B- and 8B-parameter models coupling a block-diffusion language backbone with a SigLIP-2 vision encoder and train them under a four-stage curriculum that progressively introduces alignment, capability expansion, reasoning, and long chain-of-thought (CoT) signals. Across 21 single-image, multi-image, and video benchmarks, SDAR-VL achieves state-of-the-art performance among diffusion-based multimodal models and, under matched vision towers, data, and training budgets, consistently outperforms the best global diffusion baseline LLaDA-V (You et al., 2025). Moreover, under comparable settings, SDAR-VL matches or surpasses the autoregressive LLaVA-OneVision (Li et al., 2024a), and long CoT distillation further strengthens performance on challenging multimodal reasoning benchmarks. In summary, our main contributions are:

- We provide the first systematic study of block-wise discrete diffusion for vision-language understanding, showing it matches or surpasses strong autoregressive VLMs while retaining diffusion's structural flexibility.
- We propose an integrated training framework—asynchronous block-wise noise scheduling, effective mask ratio scaling, and a progressive Beta noise curriculum—that jointly improves stability and efficiency for large-scale block-wise diffusion.
- Through extensive experiments on 21 benchmarks, we demonstrate that SDAR-VL sets a new state of the art among diffusion-based vision-language models and, under matched conditions, consistently outperforms both the global diffusion baseline LLaDA-V and the autoregressive LLaVA-OneVision. With long chain-of-thought distillation, the SDAR-VL-Think variants further reach parity with or surpass CoT-enhanced autoregressive baselines such as the Qwen2.5-VL–based R1-OneVision on math benchmarks.

2 Preliminary

We build our work on the Block Discrete Denoising Diffusion framework (BD3) (Arriola et al., 2025; Cheng et al., 2025), which interpolates between autoregressive and diffusion paradigms. A token sequence $\mathbf{x}$ of length L is partitioned into B non-overlapping blocks $\{x^1, \ldots, x^B\}$, each of length $L' = L/B$. The model factorizes the likelihood autoregressively over blocks:

$$\log p_\theta(\mathbf{x}) = \sum_{b=1}^{B} \log p_\theta(x^b \mid x^{<b}), \qquad (1)$$

where $x^{<b}$ denotes all preceding clean blocks. Each conditional distribution $p_\theta(x^b \mid x^{<b})$ is modeled via a discrete diffusion process: a forward process gradually masks tokens in x^b, and a reverse process, parameterized by Transformer with block-causal attention, reconstructs the original block.

The standard BD3 training objective applies the Negative Evidence Lower Bound (NELBO) to each block-conditional distribution, scaling the per-block loss by the inverse of the noise ratio t:

$$\mathcal{L}(\theta) = \mathbb{E}_{x,b,t}\left[\frac{-1}{t}\sum_{\ell \in \mathcal{M}_t^b} \log p_\theta\big(x_0^{b,\ell} \mid x_t^b, x^{<b}\big)\right] \qquad (2)$$

where $\mathcal{M}_t^b$ is the set of masked positions in block b at noise level t. Here, $1/t$ compensates for reduced contributing tokens, but—as we discuss in Sec. 3.2—this practice can introduce bias when t and the realized mask proportion differ.

3 Method: The SDAR-VL Framework

SDAR-VL augments the BD3 training paradigm to reduce gradient variance and improve convergence under fluctuating corruption difficulty. It consists of three components: (1) **Asynchronous**

Block-wise Noise Scheduling (ABNS) to diversify within-step training signals and stabilize optimization, (2) **Effective Mask Ratio Scaling** (EMRS) to provide unbiased loss normalization, and (3) **Progressive Beta Distribution Noise Curriculum** (PBNC) to balance noise coverage and diversity, yielding a more stable and efficient block-wise discrete diffusion training framework.

3.1 Asynchronous Block-wise Noise Schedule

The conventional BD3 framework uses a *synchronous* noise schedule: a single noise level t is sampled and applied to all blocks in a training sample, so every block experiences the same masking ratio. As shown in Sec. 4.4 (Fig. 2), the average loss increases monotonically with t, so sampling one t per step introduces large step-to-step difficulty fluctuations and high gradient variance.

We propose an *asynchronous* block-wise noise schedule, where each block b independently samples its corruption level:

$$t_b \sim \mathcal{P}(\cdot \mid \tau), \quad b = 1, \ldots, B. \tag{3}$$

Let

$$\ell_b = \sum_{\ell \in \mathcal{M}^b_{t_b}} \log p_\theta\big(x_0^{b,\ell} \mid x^b_{t_b}, x^{<b}\big) \tag{4}$$

denote the negative log-likelihood for block b. The modified asynchronous training objective is:

$$\mathcal{L}_{\text{async}}(\theta) = \mathbb{E}_{x,\, b,\, t_b}\left[-\frac{\ell_b}{t_b}\right]. \tag{5}$$

Block-specific corruption exposes the model to a richer distribution of reconstruction difficulties within each step, averaging losses over diverse t_b and smoothing fluctuations from extreme draws. Appendix B.2 shows that this asynchronous sampling leaves the expected objective unchanged while reducing the variance of the mini-batch estimator, explaining the improved training stability.

3.2 Effective Mask Ratio for Unbiased Loss Scaling

Under ABNS, the standard BD3 loss scaling multiplies each block's loss by $1/t_b$, implicitly assuming that the sampled ratio t_b matches the realized masked proportion in that block. However, discrete stochastic masking often causes deviations between t_b and the realized ratio, especially for small blocks or high-variance schedules. This mismatch yields biased NELBO estimates and noisier scaling; with independent t_b across blocks, it can further increase gradient variance and offset the stability gains of asynchronous scheduling. Appendix B.3 shows that the discrepancy grows with the variance of the discrete masking process.

To resolve this issue, we introduce the **effective mask ratio**:

$$t'_b = \frac{\|\mathbf{m}^b\|_1}{L'}, \tag{6}$$

where $\mathbf{m}^b \in \{0,1\}^{L'}$ is the binary mask vector for block b, and L' is the block length. By replacing the sampled t_b with the realized t'_b in the normalization term, we obtain an *unbiased* scaling:

$$\mathcal{L}_{\text{async+unbiased}}(\theta) = \mathbb{E}_{x,\, b,\, t_b}\left[-\frac{\ell_b}{t'_b}\right], \tag{7}$$

Intuitively, this simply weights each block by its *realized* corruption level rather than the nominal target. Appendix B.3 shows that the resulting $1/t'_b$ scaling is unbiased for the ideal objective (unlike the standard $1/t_b$ scaling), and empirically it removes systematic mis-scaling, lowers gradient variance, and stabilizes training.

3.3 Progressive Beta Distribution Noise Curriculum

A key lever to accelerate diffusion training is to *increase the effective mask ratio* so that more tokens contribute to the loss in each step, improving sample efficiency (Arriola et al., 2025). However, naively shifting all sampled noise ratios toward high values risks collapsing noise diversity, reducing the model's exposure to varied reconstruction difficulties and potentially harming generalization.

We address this trade-off with a progressive Beta scheduler for $\mathcal{P}(t_b \mid \tau)$, which preserves a wide support over t_b while smoothly increasing the mean μ_τ and concentration C_τ:

$$t_b \sim \text{Beta}(\alpha_\tau, \beta_\tau) \tag{8}$$

$$\mu_\tau = \frac{\alpha_\tau}{\alpha_\tau + \beta_\tau}, \qquad C_\tau = \alpha_\tau + \beta_\tau. \tag{9}$$

We start from a moderate mean μ_τ and low concentration C_τ, yielding a broad distribution over mid-range corruption levels for stable early training. Over a warmup horizon, both parameters increase: the mean raises the expected mask ratio (more supervised tokens per step), while the concentration tightens the distribution around this mean without

collapsing its support. PBNC thus maintains noise diversity while gradually shifting toward higher-coverage, harder reconstructions. Figure 1 visualizes this curriculum from easier, varied tasks to harder, high-coverage ones.

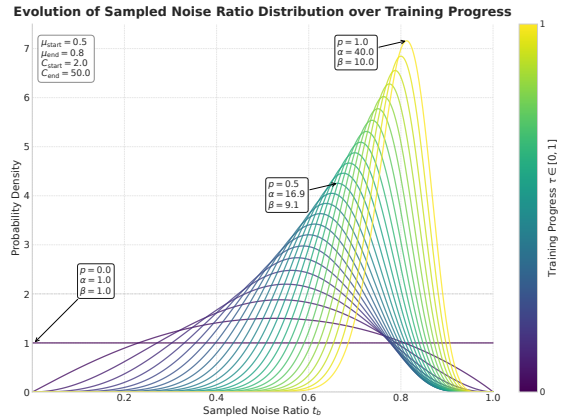


Figure 1: **Evolution of the progressive Beta noise scheduler.** Probability density functions of the block-level noise ratio t_b at different normalized training progress values u_τ (color-coded). As training progresses, the distribution shifts and sharpens toward higher t_b, increasing masked tokens per sample while preserving corruption diversity.

4 Experiment

4.1 Training Setting

Model Architecture Our experiments leverage the SDAR-Chat model (Cheng et al., 2025), a novel block diffusion-based large language model, in its 4B- and 8B-parameter variants. For the vision encoder, we utilize SigLIP-2-So400m-patch14-384 (Tschannen et al., 2025), renowned for its robust visual representation. The vision and language modalities are bridged by a projector, implemented as a randomly initialized two-layer MLP, to align the feature spaces.

Datasets and Training Stages We adopt a four-stage curriculum for vision–language alignment, multimodal capability, and long chain-of-thought (CoT) reasoning. Stage 1 trains only the projector on LLaVA-Pretrain (Liu et al., 2023) for vision–language alignment. Stage 2 improves general multimodal ability on the MAmmoTH-VL corpus (Guo et al., 2024) using M-SI followed by M-OV. Stage 3 strengthens reasoning on Visual-WebInstruct (VW) (Jia et al., 2025) and a mixture of M-OV and VW. Stage 4 distills long CoT from R1-OneVision (Yang et al., 2025b) to obtain SDAR-VL-Think. Full dataset statistics, stage-wise recipes, and hyperparameters are provided in Appendix C.1 (Table 5).

Chain-of-Thought Template Masking To encourage chain-of-thought-style responses, we introduce two special tokens, `<think>` and `</think>`, as delimiters of reasoning steps in the output. During training on *R1-OneVision*, these tokens are always masked under the masked reconstruction objective, providing direct supervision over reasoning-boundary placement and enabling reliable long-CoT template generation. This design improves template adherence, interpretability, and performance on complex multimodal tasks.

Implementation Details Our 4B and 8B models share the four-stage pipeline above. We apply sequence packing to reduce padding and increase GPU utilization, using boundary-aware attention and loss masks to ensure packed samples remain isolated. In the initial alignment stage (Stage 1), we exclusively train the projector while keeping both the vision and language towers frozen. In the subsequent fine-tuning stages, we unfreeze all components and perform full-parameter fine-tuning of the vision tower, language tower, and projector, using distinct learning rates to optimize overall performance, as specified in Table 5.

For the ablation studies presented in Section 4.4, we employ a two-stage training protocol. First, we perform vision-language alignment by training only the projector on the LLaVA-Pretrain dataset, identical to Stage 1 of our main pipeline. Second, we conduct full-parameter fine-tuning using the LLaVA-NeXT dataset (Liu et al., 2024a). The training hyperparameters are detailed in Table 5.

4.2 Evaluation Setting

Evaluation Benchmarks We evaluate SDAR-VL-4B/8B on 21 benchmarks spanning four multimodal ability categories. (1) *General visual understanding and reliability*: core perception, factuality, and robustness on MMBench (Liu et al., 2024b), SEEDBench (Li et al., 2023), MME (Fu et al., 2025a), MM-Vet (Yu et al., 2023), MM-Star (Chen et al., 2024a), MMMU (Yue et al., 2024), and HallusionBench (Guan et al., 2024). (2) *Mathematical and scientific reasoning*: visual math and science reasoning on MathVista (Lu et al., 2023), MathVerse (Zhang et al., 2024), MathVision (Wang et al., 2024b), and ScienceQA (Lu et al., 2022). (3) *Document-centric VQA*: text-rich

Table 1: **Performance on single-image benchmarks.** Results marked with * are taken from prior work (see Appendix C); all others are from our evaluations. Light-gray rows indicate the controlled group, which shares the same vision tower, image processor, and closely matched training data as our models (external baselines additionally marked with †). Within this group, the best score for each benchmark is in **bold**.

Model	MMBench	SEEDBench	MME	MMVet	MMStar	MMMU	HallBench	MathVista
	en-dev	image	test	test	test	val	avg	mini
Closed-source Models								
GPT-4V*	75.0	49.9	517/1409	49.9	57.1	56.8	-	49.9
GPT-4o*	-	76.2	-	76.2	-	69.1	-	63.8
Open-source Autoregressive Models								
InternVL2-4B	77.4	73.3	1535/531	55.7	54.2	46.1	41.2	58.6
InternVL2-8B	82.7	75.4	1642/577	60.8	61.2	47.9	45.3	58.2
InternVL2_5-4B	82.6	74.9	1678/633	63.3	58.5	50.2	46.8	61.6
InternVL2_5-8B	84.1	77.1	1688/643	68.1	62.6	53.4	49.5	63.2
Qwen2-VL-7B	81.7	76.4	1673/641	66.5	61.1	50.4	49.4	61.3
Qwen2.5-VL-3B	79.8	73.9	1564/589	66.9	56.1	49.6	44.6	61.1
Qwen2.5-VL-7B	84.2	77.1	1683/624	70.7	65.4	55.1	52.3	69.7
LLaVA-OV-7B†	82.1	**76.7**	1570/415	58.1	**62.0**	46.4	38.1	62.0
Discrete Diffusion Models								
Dimple-8B*	74.6	-	1514/432	-	-	45.2	-	42.3
LaViDA-L-8B*	70.5	-	1366/341	-	-	43.3	-	44.8
LaViDA-O-8B*	76.4	-	1431/488	-	-	45.1	-	56.9
MMaDA-8B*	68.5	-	1410/242	-	-	30.2	-	33.7
LLaDA-V-8B*†	**82.9**	74.8	1507/491	-	60.1	48.6	-	59.7
Block-wise Discrete Diffusion Models								
SDAR-VL-4B-Inst	78.7	75.4	1562/546	59.8	56.0	49.4	**46.8**	57.6
SDAR-VL-8B-Inst	82.2	75.5	**1632/569**	**62.6**	59.9	**53.0**	44.4	**62.5**

Model	MathVision	MathVerse	ScienceQA	RealworldQA	InfoVQA	DocVQA	ChartQA	AI2D
	testmini	mini-vo/vd	test	test	val/test	val/test	test	test
Closed-source Models								
GPT-4V*	-	32.8/-	75.7	61.4	-	-/88.4	78.5	78.2
GPT-4o*	-	50.2/-	-	58.6	-	-/92.8	85.7	94.2
Open-source Autoregressive Models								
InternVL2-4B	16.7	24.7/28.8	96.2	61.2	66.1/-	88.1/-	80.9	78.9
InternVL2-8B	19.2	24.2/32.2	97.2	64.7	73.0/-	91.0/-	82.2	83.7
InternVL2_5-4B	19.9	25.2/35.0	97.3	64.2	72.2/-	91.1/-	78.0	81.4
InternVL2_5-8B	18.8	27.2/39.9	98.1	69.4	75.5/-	92.0/-	82.8	84.5
Qwen2-VL-7B	18.7	25.7/32.9	86.1	70.4	76.3/-	93.9/-	83.1	83.1
Qwen2.5-VL-3B	21.0	35.6/38.6	81.7	65.2	76.1/-	92.9/-	84.1	81.2
Qwen2.5-VL-7B	25.2	42.7/44.6	89.3	68.5	82.2/-	94.9/-	86.2	84.8
LLaVA-OV-7B†	17.2	23.7/31.5	**95.3**	**69.7**	66.3/-	87.0/-	80.2	**82.6**
Discrete Diffusion Models								
Dimple-8B*	-	-	77.1	-	-	-	63.4	74.4
LaViDA-8B*	20.4	-/27.2	80.2	-	-/34.2	-/59.0	64.6	70.0
LaViDA-O-8B*	-	-/36.9	84.6	-	-/44.6	-/73.7	80.0	76.7
MMaDA-8B*	-	-/13.5	55.8	-	-/14.9	-/10.9	9.8	66.6
LLaDA-V-8B*†	-	-/28.5	-	63.2	66.3/-	83.9/-	78.3	77.8
Block-wise Discrete Diffusion Models								
SDAR-VL-4B-Inst	21.3	33.5/38.9	89.8	67.6	70.5/-	86.9/-	81.8	79.1
SDAR-VL-8B-Inst	**24.2**	**36.5/41.0**	92.2	66.5	**73.2/-**	**88.3/-**	**82.7**	79.9

understanding on DocVQA (Mathew et al., 2021), ChartQA (Masry et al., 2022), InfoVQA (Mathew et al., 2022), and AI2D (Kembhavi et al., 2016). (4) *Multi-image and video understanding*: cross-image and temporal reasoning on MuirBench (Wang et al., 2024a), BLINK (Fu et al., 2024), MVBench (Li et al., 2024b), VideoMME (Fu et al., 2025b), and MLVU (Zhou et al., 2024). Details of the evaluation setup, metrics, and prompts are provided in Appendix C.

Baselines We primarily compare to autoregressive vision–language models trained under settings closely aligned with ours. Our training uses the MAmmoTH (Guo et al., 2024) dataset, which also underlies LLaVA-OneVision (Li et al., 2024a), making model architecture, base language model, vision encoder, and training budget broadly comparable. Accordingly, our main baselines are models of similar scale and configuration, including InternVL-2 (OpenGVLab Team, 2024), Qwen2-VL (Wang et al., 2024c), and LLaVA-OneVision-OV (Li et al., 2024a), as well as newer iterations Qwen2.5-VL (Bai et al., 2025) and InternVL-2.5 (Chen et al., 2024b) to track progress across model generations.

We also compare against diffusion-based language models to contrast learning paradigms, including Dimple (Yu et al., 2025b), LLaDA-V (You et al., 2025), LaViDA (Li et al., 2025b), LaViDA-O (Li et al., 2025a), and MMaDA (Yang et al., 2025a). Among these, LLaDA-V uses the same vision tower and training data configuration as our models, enabling a controlled comparison that isolates model design and optimization strategies.

Table 2: **Performance on multi-image and video benchmarks.** Same notations as in Table 1. Rows shaded in light gray indicate the controlled comparison group, and the best results within this group are bold.

Model	MuirB	BLINK	VideoMME	LVBench	MLVU
	multi-img	multi-img	wo-subs	val	m-avg
GPT-4V*	62.3	51.1	59.9	61.3	49.2
GPT-4o*	-	-	71.9	66.7	64.6
InternVL-2-4B	40.3	47.4	53.6	51.0	58.0
InternVL-2-8B	48.5	50.3	56.1	53.6	61.5
InternVL-2.5-4B	45.1	50.7	61.6	54.2	66.7
InternVL-2.5-8B	51.2	54.8	63.8	59.5	67.2
Qwen2-VL-7B	39.9	52.9	59.7	55.1	64.3
Qwen2.5-VL-3B	46.5	48.8	58.4	52.7	63.7
Qwen2.5-VL-7B	58.2	55.6	61.9	59.7	64.3
LLaVA-OV-7B†	40.5	52.4	58.3	57.9	**66.8**
LLaDA-V-8B*†	48.3	-	56.1	-	59.5
SDAR-VL-4B-Inst	44.8	52.8	57.3	56.2	61.7
SDAR-VL-8B-Inst	**50.2**	**52.9**	**60.8**	**59.8**	65.0

4.3 Main Results

We present the main evaluation results in Table 1 and Table 2. For fair comparisons, we highlight a **controlled group (shaded in gray)** that shares the same vision tower, image processor, and largely matched training data configuration with our models; the best results within this group are boldfaced. Across single-image, multi-image, and video settings, SDAR-VL delivers strong and stable performance, demonstrating the effectiveness of block-wise discrete diffusion for multimodal learning.

Single-image benchmarks. As shown in Table 1, SDAR-VL achieves strong performance across general visual understanding, hallucination robustness, mathematical/scientific reasoning, and document-centric VQA. Under the controlled setting, our models consistently outperform the closely matched diffusion baseline LLaDA-V and lead the group on most benchmarks (14/21), indicating robust perception and reliable reasoning under matched architectures and data conditions. Notably, our advantages are more pronounced on reasoning- and text-rich tasks: the 8B model tops the controlled group on MathVista, MathVision, and MathVerse, while also achieving the best results on InfoVQA, DocVQA, and ChartQA. Beyond controlled comparisons, SDAR-VL surpasses all discrete diffusion multimodal models and remains competitive with strong autoregressive baselines. This competitiveness is notable given our smaller training scale ($\sim$57B tokens) versus reported scales of Qwen2-VL ($\sim$1.4T), Qwen2.5-VL ($\sim$4.1T), and InternVL2.5 (252B/142B for 4B/8B). We also observe consistent gains from 4B to 8B, highlighting the benefit of scaling under the block-wise discrete diffusion formulation.

Multi-image and video benchmarks. Table 2 further shows that SDAR-VL generalizes effectively to multi-image and video understanding. Within the controlled group, our 8B model achieves the best performance on BLINK, VideoMME, and LVBench, and remains competitive on MLVU, demonstrating strong cross-image and cross-frame aggregation ability. Compared with the matched diffusion baseline LLaDA-V, SDAR-VL shows consistent improvements across shared tasks, suggesting that block-wise diffusion provides a stronger foundation for multi-context visual reasoning. Against broader open-source baselines, our results are competitive overall, out-

Table 3: **Comparison of reasoning model performance before and after long Chain-of-Thought (CoT) distillation.** 'Inst' denotes models prior to long CoT distillation, while models with long CoT are indicated accordingly. R1-OneVision is a long-CoT-distilled variant built upon a Qwen2.5-VL-7B model.

Model	MathVista	MathVerse	MathVision
	mini	mini-vision	testmini
Qwen2.5-VL-7B*	63.7	38.2	25.4
R1-Onevision-7B*	64.1 (↑0.6%)	40.0 (↑4.7%)	**29.9 (↑17.7%)**
Lavida-8B*	44.8	27.2	20.4
Lavida-8B-Reason*	45.2 (↑0.9%)	29.3 (↑7.7%)	24.0 (↑17.6%)
SDAR-VL-4B-Inst	57.6	33.5	21.3
SDAR-VL-4B-Think	60.4 (↑4.9%)	38.6 (↑15.2%)	25.7 (↑20.7%)
SDAR-VL-8B-Inst	62.5	36.5	24.2
SDAR-VL-8B-Think	**64.2 (↑2.7%)**	**41.8 (↑14.5%)**	28.0 (↑15.7%)

performing Qwen2.5-VL on LVBench and MLVU while remaining close on VideoMME.

Effect of long CoT distillation. Table 3 shows that SDAR-VL consistently benefits from long CoT distillation, with clear gains on MathVista, MathVerse, and MathVision for both 4B and 8B variants. After distillation, SDAR-VL reaches strong competitiveness against CoT-enhanced autoregressive baselines, achieving near-parity on MathVision and surpassing R1-OneVision on MathVista/-MathVerse. In contrast, the full discrete diffusion baseline (LaViDA) exhibits smaller improvements, suggesting long CoT signals are easier to exploit under the block-wise formulation. Overall, these results indicate that SDAR-VL offers a more effective interface to distill long-form reasoning into diffusion-based multimodal models.

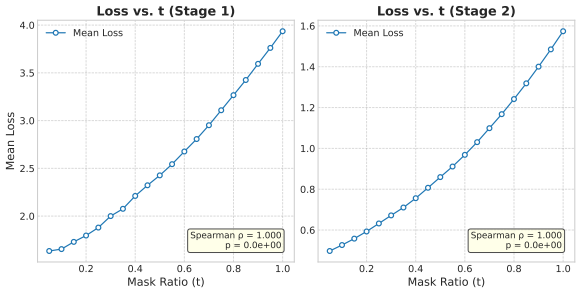


Figure 2: **Correlation between mask ratio t and average training loss at two stages.** Each point is the mean loss over samples with a fixed t; loss grows with t, indicating that higher noise increases reconstruction difficulty and motivating asynchronous noise scheduling for more stable training.

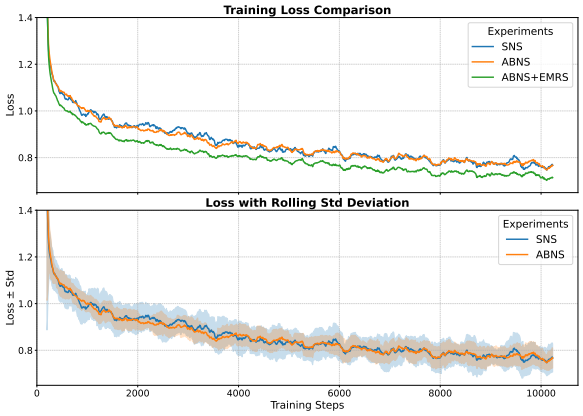


Figure 3: **Training loss of ablation variants on the 4B model during LLaVA-NeXT fine-tuning.** Top: training loss for synchronous noise scheduling (SNS), asynchronous scheduling (ABNS), and ABNS with effective mask ratio scaling (EMRS). Bottom: loss variance for SNS and ABNS, showing reduced fluctuation and improved stability with ABNS.

4.4 Ablation Studies

To isolate the individual and combined effects of the proposed SDAR-VL components (Sec. 3), we perform controlled ablations on the 4B variant. All models follow the same two-stage protocol (Sec. 4.1): projector-only alignment and subsequent full-parameter fine-tuning on LLaVA-NeXT.

We start from a BD3 (Arriola et al., 2025; Cheng et al., 2025) baseline with *synchronous noise scheduling* (SNS), where a single mask ratio t is sampled per step for all blocks with $1/t$ loss scaling. We then introduce a sequence of incremental variants. *ABNS* replaces SNS with Asynchronous Block-wise Noise Scheduling (Sec. 3.1), independently sampling a noise ratio t_b for each block to increase in-step supervision diversity. *ABNS+EMRS* further applies Effective Mask Ratio Scaling (Sec. 3.2), normalizing the loss by the realized mask ratio t'_b instead of the sampled t_b to reduce bias and gradient variance. *ABNS+EMRS+Clamp* serves as a strong static baseline that restricts the noise distribution to a high-noise uniform range $t_b \sim \mathcal{U}(0.45, 0.95)$, emphasizing challenging reconstructions but reducing exposure to low-noise cases. Finally, *ABNS+EMRS+PBNC* corresponds to the full SDAR-VL configuration and adopts the Progressive Beta Distribution Noise Curriculum (PBNC, Sec. 3.3), which gradually increases both the mean and concentration (μ_τ, C_τ) of the noise distribution to balance diversity and difficulty; we report results for two curriculum sharpness settings

Table 4: **Ablation study of SDAR-VL components on the 4B model.** We progressively add each component and evaluate performance on nine benchmarks after fine-tuning on LLaVA-NeXT. Scores are reported as percentages. The best result in each column is **bolded**. All models in this table use the same training budget.

Method	SEEDBench	MMStar	MME	DocVQA	ChartQA	AI2D	MathVista	ScienceQA	HallBench
	image	test	perp/cog	val	test	test	mini	test	avg
Baseline (SNS)	71.6	46.2	1480/326	68.2	65.4	71.9	41.6	75.4	36.8
ABNS	71.8	47.0	1482/327	68.1	67.4	72.0	40.2	74.9	37.5
ABNS + EMRS	72.1	47.5	1489/344	68.8	68.9	74.6	42.2	77.6	37.8
ABNS + EMRS + Clamp	72.6	48.2	1490/344	**70.2**	68.1	73.6	41.9	75.4	36.5
ABNS + EMRS + PBNC (c=25)	**73.5**	47.7	1501/331	69.4	**69.3**	74.5	**43.8**	**77.8**	38.4
ABNS + EMRS + PBNC (c=50)	**73.5**	**48.5**	**1515/336**	**70.2**	68.2	**74.4**	42.7	77.6	**40.1**

with C_{final}=25 and C_{final}=50.

4.4.1 Mask Ratio-Loss Analysis

To motivate asynchronous noise scheduling, we analyze how reconstruction difficulty varies with the mask ratio t. Using checkpoints from two stages (the initial stage and LLaVA-NeXT fine-tuning), we evaluate a fixed 10k subset of LLaVA-NeXT and compute the mean loss for each discrete t.

As shown in Figure 2, the loss increases monotonically with t, with strong positive Spearman correlations in both stages. This indicates that **higher mask ratios correspond to harder reconstructions and thus larger losses.** Consequently, under synchronous scheduling, sampling a single t per step can translate directly into substantial step-to-step loss variance, potentially destabilizing optimization. In contrast, asynchronous block-wise noise scheduling (ABNS) assigns diverse mask ratios t_b across blocks within a step, effectively averaging over a richer t distribution and smoothing loss fluctuations to improve training stability.

4.4.2 Training Loss Dynamics

We next analyze training loss curves of three configurations: synchronous scheduler (SNS), asynchronous block-wise scheduler (ABNS), and ABNS with mask ratio scaling (EMRS).

Figure 3 (top) shows the average training loss during LLaVA-NeXT fine-tuning. ABNS and SNS exhibit similar mean loss trends, whereas EMRS consistently lowers the loss and improves convergence stability, indicating more effective optimization via unbiased loss scaling. The bottom subplot shows that ABNS markedly reduces loss variance despite comparable mean loss, underscoring its key role in stabilizing training.

Due to the strong dependence of loss magnitude on t, plotting loss curves across different noise schedulers with varying mask ratio scalings (e.g., clamp or progressive beta) is not meaningful. Hence, those are excluded from loss curve plots but evaluated in downstream benchmarks.

4.4.3 Downstream Task Performance

Table 4 presents ablations on the 4B model with identical training budget across nine benchmarks. Relative to SNS, **ABNS** yields modest gains on ChartQA, MMStar, and HallBench, indicating limited but positive effects from increased in-step noise diversity. **EMRS** brings broader improvements across most tasks, notably on MME cognition and ScienceQA, suggesting more stable optimization via unbiased scaling. **Clamp** improves select scores (e.g., DocVQA) but hurts robustness, revealing a non-trivial coverage–diversity trade-off. **PBNC** best resolves this trade-off and delivers the strongest overall performance; C_{final}=25 tends to favor reasoning- and text-rich tasks, while C_{final}=50 strengthens general understanding and hallucination robustness. Overall, ABNS, EMRS, and PBNC provide complementary benefits, and the full SDAR-VL setup achieves the best results.

5 Conclusion

We introduced SDAR-VL, a block-wise discrete diffusion framework for vision-language understanding that stabilizes BD3 training via asynchronous block-wise noise scheduling, effective mask ratio scaling, and a progressive Beta noise curriculum. Across 21 benchmarks, SDAR-VL sets a new state of the art among diffusion-based vision-language models and, under matched settings, outperforms the diffusion baseline LLaDA-V as well as autoregressive LLaVA-OneVision. With long CoT distillation, SDAR-VL-Think further reaches parity with or surpasses CoT-enhanced autoregressive baselines such as the Qwen2.5-VL–based R1-OneVision on math-centric benchmarks, highlighting block-wise discrete diffusion as a practical backbone for multimodal reasoning.

6 Limitations

While SDAR-VL demonstrates that block-wise discrete diffusion is a competitive backbone for vision-language understanding, several limitations remain that merit future investigation:

Model Scaling and Compute Constraints. Due to academic computational resource constraints, our experiments are limited to the 4B and 8B parameter scales. While we observe positive scaling trends between these two sizes, it remains to be verified whether the stability and efficiency gains of the SDAR-VL framework hold at the scales of 30B or 70B+ parameters, where diffusion-based loss scaling dynamics can become more complex and unpredictable.

Scope of Generality and Unified Generation. The fundamental advantage of diffusion models lies in their modeling universality—the ability to handle continuous and discrete modalities within a single framework. However, due to scope and resource considerations, this work prioritizes solving the fundamental bottlenecks of training efficiency and stability for block-wise diffusion, restricting our evaluation to multimodal *understanding* (i.e., text generation). Consequently, we have not yet exploited the model's capacity for generating visual or auditory content. We believe the next frontier will likely be Multimodal Native Diffusion, where a single diffusion transformer generates text, image, and audio tokens in a unified denoising process, finally fulfilling the promise of true general-purpose generative models.

References

Marianne Arriola, Aaron Gokaslan, Justin T Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sekhar Sahoo, and Volodymyr Kuleshov. 2025. Block diffusion: Interpolating between autoregressive and diffusion language models. *arXiv preprint arXiv:2503.09573*.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024a. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Shuang Cheng, Yihan Bian, Dawei Liu, Yuhua Jiang, Yihao Liu, Linfeng Zhang, Wenhai Wang, Qipeng Guo, Kai Chen, Biqing Qi, and 1 others. 2025. Sdar: A synergistic diffusion-autoregression paradigm for scalable sequence generation. *arXiv preprint arXiv:2510.06303*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Haodong Duan, Xinyu Fang, Junming Yang, Xiangyu Zhao, Yuxuan Qiao, Mo Li, Amit Agarwal, Zhe Chen, Lin Chen, Yuan Liu, Yubo Ma, Hailong Sun, Yifan Zhang, Shiyin Lu, Tack Hwa Wong, Weiyun Wang, Peiheng Zhou, Xiaozhe Li, Chaoyou Fu, and 13 others. 2025. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. *Preprint*, arXiv:2407.11691.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. 2025a. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025b. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer.

Shansan Gong, Shivam Agarwal, Yizhe Zhang, Jiacheng Ye, Lin Zheng, Mukai Li, Chenxin An, Peilin Zhao, Wei Bi, Jiawei Han, and 1 others. 2024. Scaling diffusion language models via adaptation from autoregressive models. *arXiv preprint arXiv:2410.17891*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.

Ishaan Gulrajani and Tatsunori B Hashimoto. 2023. Likelihood-based diffusion language models. *Advances in Neural Information Processing Systems*, 36:16693–16715.

Jarvis Guo, Tuney Zheng, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Yizhi Li, Graham Neubig, Wenhu Chen, and Xiang Yue. 2024. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. *arXiv preprint arXiv:2412.05237*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Yatai Ji, Teng Wang, Yuying Ge, Zhiheng Liu, Sidi Yang, Ying Shan, and Ping Luo. 2025. From denoising to refining: A corrective framework for vision-language diffusion model. *arXiv preprint arXiv:2510.19871*.

Yiming Jia, Jiachen Li, Xiang Yue, Bo Li, Ping Nie, Kai Zou, and Wenhu Chen. 2025. Visualwebinstruct: Scaling up multimodal instruction data through web search. *arXiv preprint arXiv:2503.10582*.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Shufan Li, Jiuxiang Gu, Kangning Liu, Zhe Lin, Zijun Wei, Aditya Grover, and Jason Kuen. 2025a. Lavida-o: Elastic large masked diffusion models for unified multimodal understanding and generation. *arXiv e-prints*, pages arXiv–2509.

Shufan Li, Konstantinos Kallidromitis, Hritik Bansal, Akash Gokul, Yusuke Kato, Kazuki Kozuka, Jason Kuen, Zhe Lin, Kai-Wei Chang, and Aditya Grover. 2025b. Lavida: A large diffusion language model for multimodal understanding. *arXiv preprint arXiv:2505.16839*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llavanext: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024b. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Aaron Lou, Chenlin Meng, and Stefano Ermon. 2024. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Proceedings of the 41st International Conference on Machine Learning*, pages 32819–32848.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *CoRR*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Xinyin Ma, Runpeng Yu, Gongfan Fang, and Xinchao Wang. 2025. dkv-cache: The cache for diffusion language models. *arXiv preprint arXiv:2505.15781*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. Large language diffusion models. *arXiv preprint arXiv:2502.09992*.

OpenGVLab Team. 2024. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy. https://internvl.github.io/blog/2024-07-02-InternVL-2.0/. Accessed 2024-04-27.

Han Peng, Peiyu Liu, Zican Dong, Daixuan Cheng, Junyi Li, Yiru Tang, Shuo Wang, and Wayne Xin Zhao. 2025. How efficient are diffusion language models? a critical examination of efficiency evaluation practices. *arXiv preprint arXiv:2510.18480*.

Subham Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin Chiu, Alexander Rush, and Volodymyr Kuleshov. 2024. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184.

Qingyu Shi, Jinbin Bai, Zhuoran Zhao, Wenhao Chai, Kaidong Yu, Jianzong Wu, Shuangyong Song, Yunhai Tong, Xiangtai Li, Xuelong Li, and 1 others. 2025. Muddit: Liberating generation beyond text-to-image with a unified discrete diffusion model. *arXiv preprint arXiv:2505.23606*.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, and 1 others. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, and 1 others. 2024a. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024b. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024c. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025a. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Xu Wang, Chenkai Xu, Yijie Jin, Jiachun Jin, Hao Zhang, and Zhijie Deng. 2025b. Diffusion llms can do faster-than-ar inference via discrete diffusion forcing. *arXiv preprint arXiv:2508.09192*.

Chengyue Wu, Hao Zhang, Shuchen Xue, Shizhe Diao, Yonggan Fu, Zhijian Liu, Pavlo Molchanov, Ping Luo, Song Han, and Enze Xie. 2025a. Fast-dllm v2: Efficient block-diffusion llm. *arXiv preprint arXiv:2509.26328*.

Chengyue Wu, Hao Zhang, Shuchen Xue, Zhijian Liu, Shizhe Diao, Ligeng Zhu, Ping Luo, Song Han, and Enze Xie. 2025b. Fast-dllm: Training-free acceleration of diffusion llm by enabling kv cache and parallel decoding. *arXiv preprint arXiv:2505.22618*.

Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, and 1 others. 2025. Lumina-dimoo: An omni diffusion large language model for multimodal generation and understanding. *arXiv preprint arXiv:2510.06308*.

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

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, and 1 others. 2025b. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*.

Yicun Yang, Cong Wang, Shaobo Wang, Zichen Wen, Biqing Qi, Hanlin Xu, and Linfeng Zhang. 2025c. Diffusion llm with native variable generation lengths: Let [eos] lead the way. *arXiv preprint arXiv:2510.24605*.

Jiacheng Ye, Jiahui Gao, Shansan Gong, Lin Zheng, Xin Jiang, Zhenguo Li, and Lingpeng Kong. 2024a. Beyond autoregression: Discrete diffusion for complex reasoning and planning. *arXiv preprint arXiv:2410.14157*.

Jiacheng Ye, Shansan Gong, Liheng Chen, Lin Zheng, Jiahui Gao, Han Shi, Chuan Wu, Xin Jiang, Zhenguo Li, Wei Bi, and 1 others. 2024b. Diffusion of thought: Chain-of-thought reasoning in diffusion language models. *Advances in Neural Information Processing Systems*, 37:105345–105374.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. 2025. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*.

Zebin You, Shen Nie, Xiaolu Zhang, Jun Hu, Jun Zhou, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. 2025. Llada-v: Large language diffusion models with visual instruction tuning. *arXiv preprint arXiv:2505.16933*.

Runpeng Yu, Qi Li, and Xinchao Wang. 2025a. Discrete diffusion in large language and multimodal models: A survey. *arXiv preprint arXiv:2506.13759*.

Runpeng Yu, Xinyin Ma, and Xinchao Wang. 2025b. Dimple: Discrete diffusion multimodal large language model with parallel decoding. *arXiv preprint arXiv:2505.16990*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Lingzhe Zhang, Liancheng Fang, Chiming Duan, Minghua He, Leyi Pan, Pei Xiao, Shiyu Huang, Yunpeng Zhai, Xuming Hu, Philip S Yu, and 1 others. 2025. A survey on parallel text generation: From parallel decoding to diffusion language models. *arXiv preprint arXiv:2508.08712*.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, and 1 others. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2024. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv e-prints*, pages arXiv–2406.

Fengqi Zhu, Rongzhen Wang, Shen Nie, Xiaolu Zhang, Chunwei Wu, Jun Hu, Jun Zhou, Jianfei Chen, Yankai Lin, Ji-Rong Wen, and 1 others. 2025a. Llada 1.5: Variance-reduced preference optimization for large language diffusion models. *arXiv preprint arXiv:2505.19223*.

Fengqi Zhu, Zebin You, Yipeng Xing, Zenan Huang, Lin Liu, Yihong Zhuang, Guoshan Lu, Kangyu Wang, Xudong Wang, Lanning Wei, and 1 others. 2025b. Llada-moe: A sparse moe diffusion language model. *arXiv preprint arXiv:2509.24389*.

A Related Work

A.1 Large Language Diffusion Models

The paradigm of sequence generation has recently seen the emergence of diffusion models as a compelling alternative to traditional autoregressive approaches (Nie et al., 2025; Gong et al., 2024; Zhu et al., 2025b; Cheng et al., 2025). Unlike AR models that predict sequences token-by-token, discrete diffusion models treat generation as a denoising process, learning to reverse a gradual corruption of data (Austin et al., 2021; Gulrajani and Hashimoto, 2023; Lou et al., 2024; Sahoo et al., 2024; Arriola et al., 2025). Among these, masked diffusion has become the *de facto* standard, where the model learns to restore original tokens from a sequence corrupted with '[MASK]' tokens at various noise levels (Austin et al., 2021; Sahoo et al., 2024). This architectural shift offers several fundamental advantages. First, it enables parallel decoding, directly addressing the high latency inherent in the sequential nature of AR inference (Zhang et al., 2025). Second, by relaxing the strict left-to-right causal bias, diffusion models possess a bidirectional context, an inductive bias better suited for tasks requiring holistic reasoning and non-local dependency modeling (Ye et al., 2024a, 2025). This flexibility also hints at broader capabilities, such as iterative self-correction, flexible generation orders (e.g., infilling) and a more generalized modeling framework (Yu et al., 2025a; Ji et al., 2025; Ye et al., 2024b). This trend has recently culminated in hybrid architectures like block-wise diffusion, which fuse a global autoregressive structure with local parallel diffusion, preserving causal flow for variable-length generation while benefiting from intra-block parallelism (Arriola et al., 2025; Cheng et al., 2025; Wang et al., 2025b; Wu et al., 2025a).

A.2 Multimodal Large Language Diffusion Models

Inspired by successes in language, discrete diffusion has been extended to the multimodal domain, first targeting understanding tasks. Pioneering works like LLaDA (Zhu et al., 2025a), LaViDA (Li et al., 2025b), and Dimple (Yu et al., 2025b) established the paradigm: image and text features are concatenated and denoised as a single sequence via a global masked diffusion process. This foundation was quickly built upon by more ambitious efforts toward a unified architecture for both understanding and generation, such as Lumina-DiMOO (Xin et al., 2025), MMaDA (Yang et al., 2025a), Lavida-O (Li et al., 2025a) and Muddit (Shi et al., 2025), which frame all tasks within a single, elegant denoising framework.

However, this rapid pursuit of unification has exposed a foundational weakness. These models, whether for understanding or generation, predomi-

nantly rely on a global diffusion process. This approach suffers from crippling quadratic complexity and a notoriously difficult learning objective when applied to long, mixed-modality sequences, often resulting in "suboptimal performance" on complex reasoning tasks (You et al., 2025; Xin et al., 2025). We argue that **a robust and efficient understanding model is a prerequisite for any unified system**. Consequently, we are the first to systematically apply the block-wise diffusion paradigm (Arriola et al., 2025; Cheng et al., 2025) to this problem. Our approach circumvents the scalability issues of global diffusion by combining intra-block parallel denoising with an efficient inter-block autoregressive structure, creating a more powerful and practical foundation for multimodal understanding.

B Theoretical Analysis of SDAR-VL

In this section, we formalize the training dynamics of BD3 and SDAR-VL and show that: (i) asynchronous block-wise noise scheduling strictly reduces the variance of the stochastic gradient estimator (thus stabilizing training), and (ii) the standard BD3 loss scaling based on the sampled noise ratio t_b is a biased estimate of an ideal NELBO objective, while our effective mask ratio t'_b yields an unbiased estimator.

Throughout, we focus on a single block b within a sequence and treat the contribution of that block to the loss or to a particular gradient coordinate as a scalar random variable Z. All statements therefore apply coordinate-wise to the full gradient vector. For notational simplicity, we omit the explicit dependence on x and θ whenever it is clear from context.

B.1 Theoretical Setup and Notation

For a given training step, consider a sequence x decomposed into B blocks, and a noise level t sampled from a scheduling distribution $\mathcal{P}(\cdot \mid \tau)$. Given t, a binary mask $\boldsymbol{m} \in \{0,1\}^{L'}$ of block length L' is sampled (e.g., via independent Bernoulli or more structured masking), inducing the *realized* or *effective* mask ratio

$$t' = \frac{\|\boldsymbol{m}\|_1}{L'}. \tag{10}$$

For block b, we denote the corresponding block-level loss by

$$\ell_b = \sum_{\ell \in \mathcal{M}^b_{t_b}} -\log p_\theta\big(x_0^{b,\ell} \mid x^b_{t_b}, x^{<b}\big), \tag{11}$$

where $\mathcal{M}^b_{t_b}$ is the set of masked positions in block b under the mask $\boldsymbol{m}^b$.

To analyze variance and bias, we abstract the contribution of block b at a given step into a scalar random variable

$$Z = Z(t_b, \boldsymbol{m}^b), \tag{12}$$

which can be either the block loss itself $Z = \ell_b$ or any fixed linear functional of the gradient, e.g. $Z = u^\top \nabla_\theta \ell_b$ for some fixed vector u. For a given noise level t, define

$$\mu(t) := \mathbb{E}[Z \mid t], \qquad \sigma^2(t) := \mathrm{Var}(Z \mid t). \tag{13}$$

Expectations are taken over data, masks, and any other randomness conditioned on t.

B.2 Variance reduction via asynchronous block-wise noise scheduling

We compare two ways of sampling the noise levels for a batch of B blocks:

- **Synchronous schedule:** A single noise level t is sampled from $\mathcal{P}(\cdot \mid \tau)$ and shared by all B blocks in the sequence. Conditioned on t, we obtain i.i.d. block contributions $Z_1, \dots, Z_B \overset{\text{i.i.d.}}{\sim} Z \mid t$, and the batch-level estimator is

$$\bar{Z}_{\text{sync}} = \frac{1}{B}\sum_{b=1}^{B} Z_b. \tag{14}$$

- **Asynchronous block-wise schedule:** Each block b independently samples its own noise level $t_b \sim \mathcal{P}(\cdot \mid \tau)$ and mask $\boldsymbol{m}^b$. Thus, (t_b, Z_b) are i.i.d. across blocks, and the batch-level estimator is

$$\bar{Z}_{\text{async}} = \frac{1}{B}\sum_{b=1}^{B} Z_b. \tag{15}$$

Note that the same notation Z_b is used in both cases, but in the synchronous setting all Z_b share the same t, whereas in the asynchronous setting each Z_b is associated with an independently drawn t_b.

Lemma 1 (Unchanged expectation). Under the two sampling schemes above, the estimators $\bar{Z}_{\text{sync}}$ and $\bar{Z}_{\text{async}}$ have the same expectation:

$$\mathbb{E}[\bar{Z}_{\text{sync}}] = \mathbb{E}[\bar{Z}_{\text{async}}] = \mathbb{E}[Z]. \tag{16}$$

Proof. By the law of total expectation,

$$\mathbb{E}[\bar{Z}_{\text{sync}}] = \mathbb{E}_t\big[\mathbb{E}[\bar{Z}_{\text{sync}} \mid t]\big] = \mathbb{E}_t[\mu(t)] = \mathbb{E}[Z]. \tag{17}$$

In the asynchronous setting, $(Z_b)_{b=1}^B$ are i.i.d. copies of Z, so

$$\mathbb{E}[\bar{Z}_{\text{async}}] = \mathbb{E}\Big[\frac{1}{B}\sum_{b=1}^{B} Z_b\Big] = \mathbb{E}[Z]. \tag{18}$$

Lemma 2 (Variance reduction). Assume $B \geq 2$ and that the conditional mean $\mu(t)$ is not almost surely constant as a function of t, i.e. $\mathrm{Var}_t(\mu(t)) > 0$. Then

$$\mathrm{Var}(\bar{Z}_{\text{async}}) = \frac{1}{B}\Big(\mathbb{E}_t[\sigma^2(t)] + \mathrm{Var}_t(\mu(t))\Big), \tag{19}$$

$$\mathrm{Var}(\bar{Z}_{\text{sync}}) = \mathrm{Var}_t(\mu(t)) + \frac{1}{B}\mathbb{E}_t[\sigma^2(t)]. \tag{20}$$

Subtracting (19) from (20) gives

$$\mathrm{Var}(\bar{Z}_{\text{sync}}) - \mathrm{Var}(\bar{Z}_{\text{async}}) = \Big(1 - \tfrac{1}{B}\Big)\,\mathrm{Var}_t(\mu(t)). \tag{21}$$

This quantity is strictly positive whenever $B > 1$ and $\mathrm{Var}_t(\mu(t)) > 0$.

Proof. For the synchronous scheme, we apply the law of total variance:

$$\mathrm{Var}(\bar{Z}_{\text{sync}}) = \mathrm{Var}\big(\mathbb{E}[\bar{Z}_{\text{sync}} \mid t]\big) + \mathbb{E}\big[\mathrm{Var}(\bar{Z}_{\text{sync}} \mid t)\big]. \tag{22}$$

Conditioned on t, the Z_b are i.i.d. with conditional mean $\mu(t)$ and variance $\sigma^2(t)$, whence

$$\mathbb{E}[\bar{Z}_{\text{sync}} \mid t] = \mu(t), \qquad \mathrm{Var}(\bar{Z}_{\text{sync}} \mid t) = \frac{1}{B}\sigma^2(t). \tag{23}$$

Substituting gives (20). In the asynchronous scheme, $(Z_b)_{b=1}^B$ are i.i.d. draws from the marginal distribution of Z, so

$$\mathrm{Var}(\bar{Z}_{\text{async}}) = \frac{1}{B}\mathrm{Var}(Z). \tag{24}$$

Again by the law of total variance,

$$\mathrm{Var}(Z) = \mathbb{E}_t[\sigma^2(t)] + \mathrm{Var}_t(\mu(t)), \tag{25}$$

which yields (19). The difference (21) then follows directly. □

Remark 1 (Training stability). Lemma 2 shows that for any fixed noise distribution $\mathcal{P}(\cdot \mid \tau)$ and model parameters θ, asynchronous block-wise noise scheduling leaves the expectation of the batch estimator unchanged (Lemma 1) while reducing its variance by $(1 - 1/B)\mathrm{Var}_t(\mu(t))$. Empirically, we observe a strong dependence of the loss on the noise level t, which implies that $\mathrm{Var}_t(\mu(t))$ is non-negligible. In this regime, asynchronous scheduling stabilizes training by averaging over multiple noise levels *within* a single step, rather than concentrating the variability induced by different t across different steps.

B.3 Bias of standard loss scaling and unbiasedness of the effective mask ratio

We now formalize the bias introduced by the standard BD3 loss scaling using the sampled ratio t_b, and show that replacing t_b with the realized mask ratio t'_b restores an unbiased estimate of a natural NELBO objective.

Ideal objective based on realized corruption. Intuitively, the "strength" of supervision contributed by block b should be proportional to the *actual* fraction of tokens that are masked, rather than the nominal target t_b. This motivates the following ideal objective:

$$\mathcal{L}^\star(\theta) := \mathbb{E}\Big[-\frac{\ell_b}{t'_b}\Big], \qquad t'_b = \frac{\|\boldsymbol{m}^b\|_1}{L'}. \tag{26}$$

Here t_b is drawn from the scheduling distribution $\mathcal{P}(\cdot \mid \tau)$, and $\boldsymbol{m}^b$ is sampled conditioned on t_b.

Standard BD3 scaling as a biased estimator. In the standard BD3 framework, the block-wise contribution is scaled by the *sampled* noise ratio t_b:

$$V := -\frac{\ell_b}{t_b}, \qquad \mathcal{L}_0(\theta) := \mathbb{E}[V]. \tag{27}$$

Comparing $\mathcal{L}_0(\theta)$ with the ideal objective $\mathcal{L}^\star(\theta)$ in (26) gives a precise expression for the bias.

Lemma 3 (Bias of standard BD3 scaling). The standard BD3 objective $\mathcal{L}_0(\theta)$ is in general a biased approximation of $\mathcal{L}^\star(\theta)$:

$$\mathcal{L}_0(\theta) - \mathcal{L}^\star(\theta) = \mathbb{E}\Big[-\ell_b\Big(\frac{1}{t_b} - \frac{1}{t'_b}\Big)\Big]. \tag{28}$$

Unless $t_b = t'_b$ almost surely, the right-hand side is non-zero whenever ℓ_b correlates with the realized corruption level.

Proof. By definition,

$$\mathcal{L}_0(\theta) = \mathbb{E}\Big[-\frac{\ell_b}{t_b}\Big], \qquad \mathcal{L}^\star(\theta) = \mathbb{E}\Big[-\frac{\ell_b}{t'_b}\Big]. \tag{29}$$

Subtracting these gives (28). If $t_b \neq t'_b$ with positive probability and the loss ℓ_b depends monotonically on the realized corruption level t'_b, then ℓ_b is non-trivially correlated with $\frac{1}{t_b} - \frac{1}{t'_b}$, and the expectation in (28) does not vanish.

It is sometimes useful to decompose the bias in terms of mean and covariance:

$$\mathcal{L}_0(\theta) - \mathcal{L}^\star(\theta) = \mathbb{E}[\ell_b]\,\mathbb{E}\Big[\frac{1}{t'_b} - \frac{1}{t_b}\Big] + \mathrm{Cov}\Big(\ell_b, \frac{1}{t_b} - \frac{1}{t'_b}\Big), \tag{30}$$

highlighting that the mismatch between t_b and t'_b distorts the relative weighting of easier and harder blocks.

A simple approximation illustrating the dependence on block size. To obtain further intuition, consider the following stylized approximation. Suppose that, for a fixed block length L', the block loss can be expressed as

$$\ell_b \approx L'\, t'_b\, h(t'_b), \tag{31}$$

where $h(t')$ denotes the average negative log-likelihood per masked token at realized mask ratio

t'. Empirically, h is non-decreasing and often convex in t': higher corruption ratios lead to more difficult reconstructions. Assume that the mask is sampled via independent Bernoulli(t_b) trials, so that

$$t'_b = t_b + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid t_b] = 0, \qquad \mathrm{Var}(\varepsilon \mid t_b) = \frac{t_b(1-t_b)}{L'}. \tag{32}$$

In this approximation, the difference between the two block contributions becomes

$$V - U := -\frac{\ell_b}{t_b} + \frac{\ell_b}{t'_b} \approx -L' h(t'_b)\Big(\frac{t'_b}{t_b} - 1\Big), \quad U := -\frac{\ell_b}{t'_b}. \tag{33}$$

A second-order Taylor expansion of $h(t'_b)$ around t_b and taking conditional expectations over ε yields a bias term that scales proportionally to $\mathrm{Var}(\varepsilon \mid t_b) \propto 1/L'$. Thus, the discrepancy between t_b and t'_b—and hence the bias in (28)—is most pronounced for small blocks (small L') or high-variance masking schemes.

Effective mask ratio as an unbiased estimator. Our proposed effective mask ratio replaces t_b with the realized ratio t'_b in the scaling factor:

$$\mathcal{L}_{\mathrm{AU}}(\theta) := \mathbb{E}\Big[-\frac{\ell_b}{t'_b}\Big]. \tag{34}$$

Lemma 4 (Unbiasedness of effective mask ratio). For any parameter θ,

$$\mathcal{L}_{\mathrm{AU}}(\theta) = \mathcal{L}^\star(\theta). \tag{35}$$

Under standard regularity assumptions (e.g., dominated convergence),

$$\nabla_\theta \mathcal{L}_{\mathrm{AU}}(\theta) = \mathbb{E}\Big[-\frac{\nabla_\theta \ell_b}{t'_b}\Big], \qquad \nabla_\theta \mathcal{L}_{\mathrm{AU}}(\theta) = \nabla_\theta \mathcal{L}^\star(\theta). \tag{36}$$

In other words, using t'_b in the normalization yields an unbiased estimator of the ideal objective (26) and its gradient.

Proof. By definition, $\mathcal{L}_{\mathrm{AU}}(\theta) = \mathbb{E}[-\ell_b/t'_b]$, which coincides with $\mathcal{L}^\star(\theta)$ in (26). Under the stated regularity conditions, differentiation and expectation can be interchanged, giving

$$\nabla_\theta \mathcal{L}_{\mathrm{AU}}(\theta) = \mathbb{E}\Big[-\frac{\nabla_\theta \ell_b}{t'_b}\Big] = \nabla_\theta \mathcal{L}^\star(\theta). \tag{37}$$

Remark 2 (Interaction with asynchronous scheduling). Combining Lemma 2 and Lemma 4, SDAR-VL can be viewed as optimizing the ideal objective $\mathcal{L}^\star(\theta)$ using: (i) an unbiased, variance-reduced gradient estimator achieved by asynchronous block-wise noise sampling, and (ii) an unbiased scaling of each block's contribution based on its realized corruption level t'_b. Together, these modifications remove a systematic mis-scaling present in the standard BD3 objective and reduce gradient variance, leading to more stable training dynamics in practice.

C Experiments

C.1 Training Setting

Datasets and Training Stages Our training curriculum follows a sequential, multi-stage approach designed to progressively build the model's capabilities. The main training pipeline consists of six steps, which are organized into four stages, as detailed below. The specific hyperparameters for each step are provided in Appendix C.1, Table 5.

- **Stage 1: Vision-Language Alignment.** We begin by pre-training on the LLaVA-Pretrain dataset (Liu et al., 2023). This initial stage is dedicated to aligning the projector, connecting the visual features from the vision tower to the language model's embedding space.

- **Stage 2: Capability Expansion.** In the second stage, we fine-tune the model on the comprehensive MAmmoTH-VL dataset (Guo et al., 2024). This is done sequentially, first on M-SI (10M single-image samples) and then on M-OV (2M diverse-modality samples), to equip the model with a broad understanding of multimodal concepts.

- **Stage 3: Reasoning Enhancement.** This stage focuses on enhancing complex reasoning. We first fine-tune on the reasoning-centric VisualWebInstruct (VW) dataset (Jia et al., 2025). To further cultivate balanced reasoning skills, we conclude training with a mixture of M-OV and VW data.

- **Stage 4: Long CoT Distillation.** We further distill long-form chain-of-thought reasoning by fine-tuning *SDAR-VL-Instruct* on *R1-OneVision* (Yang et al., 2025b). This step injects multi-step reasoning patterns and yields *SDAR-VL-Think*, which improves performance on challenging math benchmarks.

C.2 Evaluation Setting

Evaluation Protocol For SDAR-VL, we adopt block-inference-based low-confidence greedy static decoding, with both the block length and the number of denoising steps fixed to 4. The decoding procedure follows SDAR (Cheng et al., 2025). Unless otherwise specified, all reported results for SDAR-VL use this configuration.

For the open-source baselines in Table 1, including the InternVL series, Qwen-VL series, and

Table 5: **Training Hyperparameters and Datasets.** We detail the configurations for each sequential training stage. M-SI and M-OV denote the single-image and one-vision subsets of MAmmoTH (Guo et al., 2024), while VW represents VisualWebInstruct (Jia et al., 2025). Our model is trained on the first five datasets sequentially. The final dataset, LLaVA-NeXT (Liu et al., 2024a), is used for full fine-tuning in our ablation studies (Sec. 4.4). **The presented settings are applied consistently for training both the 4B and 8B versions of our model.**

Training Data	**LLaVA-Pretrain**	**M-SI**	**M-OV**	**VW**	**M-OV+VW**	**R1-Onevision**	**LLaVA-NeXT**
Vision Tower	SigLIP-2-So400m-patch14-384 (Tschannen et al., 2025)						
Language Tower	SDAR-Chat (4B and 8B) (Cheng et al., 2025)						
Attention Mechanism	Blockwise Attention						
Batch Size	32	64	64	64	64	32	32
Packing Length	8192	16384	16384	16384	16384	16384	8192
# Samples	558K	10M	2M	900K	3M	155K	738K
Epochs	2	1	1	1	1	3	1
Training Tokens	1.1B	33.8B	9.3B	1.8B	11.1B	2.6B	2.7B
LR of Vision Tower	—	2×10^{-6}	2×10^{-6}	2×10^{-6}	2×10^{-6}	2×10^{-6}	2×10^{-6}
LR of Language Tower	—	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-5}
LR of Projector	1×10^{-3}	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-5}

LLaVA-OneVision-OV models, we run local evaluations using the VLMEval-Kit framework (Duan et al., 2025) with standard greedy decoding and the default evaluation configurations provided by the toolkit. The multi-image and video results in Table 2 follow the same evaluation setup and sources as Table 1. For closed-source models, the GPT-4V and GPT-4o results are directly taken from the LLaVA-OneVision paper (Li et al., 2024a). For discrete diffusion baselines in Table 1, the LLaDA-V scores are from its original paper (You et al., 2025), while the remaining diffusion-based results are taken from LaViDA-O (Li et al., 2025a). For the reasoning-focused comparison in Table 3, we use the Qwen2.5-VL and R1-OneVision-7B results reported in R1-OneVision (Yang et al., 2025b), and the LaViDA-8B results reported in LaViDA (Li et al., 2025b).

Evaluation Benchmarks To comprehensively assess the capabilities of our SDAR-VL models (4B and 8B), we conduct an extensive evaluation across a diverse suite of 21 established benchmarks. These benchmarks are carefully selected to probe a wide spectrum of multimodal abilities, from foundational visual perception to complex, high-level reasoning. The evaluation suite is organized into four key categories:

- **General Visual Understanding & Reliability:** We assess core vision-language capabilities and factual correctness using established benchmarks, including MMBench (Liu et al., 2024b), SeedBench (Li et al., 2023), MME (Fu et al., 2025a), MM-Vet (Yu et al., 2023), and MMStar (Chen et al., 2024a). To assess knowledge-intensive understanding, we evaluate on MMMU (Yue et al., 2024). In addition, we evaluate robustness against object hallucination using HallusionBench (Guan et al., 2024).

- **Mathematical and Scientific Reasoning:** We test advanced reasoning skills on benchmarks requiring complex problem-solving. This includes MathVista (Lu et al., 2023), MathVerse (Zhang et al., 2024) (for which we use chain-of-thought (CoT) decoding; in Table 1, "vo" and "vd" denote the vision-only and vision-dominant settings, respectively), and MathVision (Wang et al., 2024b) for mathematical reasoning in visual contexts, and ScienceQA (Lu et al., 2022) for multimodal scientific question answering.

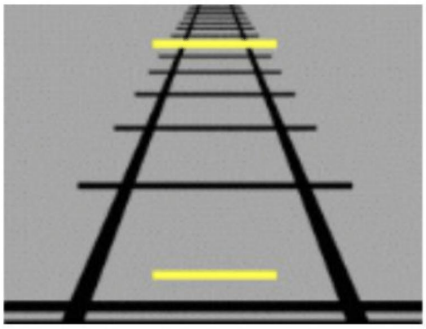
- **Document-centric VQA:** We evaluate the model's ability to understand and reason over text-rich images across diverse formats using DocVQA (Mathew et al., 2021), ChartQA (Masry et al., 2022), InfoVQA (Mathew et al., 2022), and AI2D (Kembhavi et al., 2016).

- **Multi-image and Video Understanding:** We evaluate multi-image reasoning with MuirBench (Wang et al., 2024a) and BLINK (Fu et al., 2024). For video understanding, we use MVBench (Li et al., 2024b), VideoMME (Fu et al., 2025b), and MLVU (Zhou et al., 2024) to assess temporal

comprehension and cross-frame aggregation.

D Example Appendix

Case D.1: HallusionBench



Question: Are the two yellow lines the same length?

Ground-truth answer: Yes

Model output (SDAR-VL-Instruct-8B):
To determine if the two yellow lines are the same length, we need to analyze the image and compare the lengths of the yellow lines.
1. **Identify the Yellow Lines:**
The first yellow line is located near the top of the image.
The second yellow line is located near the bottom of the image.
2. **Compare the Lengths:**
Visually inspect the lengths of both yellow lines.
Note that both yellow lines appear to span the same horizontal distance across the image.
3. **Conclusion:**
Based on the visual comparison, it can be concluded that the two yellow lines are indeed the same length.
Therefore, the answer is yes.

Case D.2: MME: artwork

Question: Is this artwork created by donatello? Please answer yes or no.

Ground-truth answer: Yes

Model output (SDAR-VL-Instruct-8B):
To determine if the artwork in the image was created by Donatello, we need to analyze the style and characteristics of the artwork and compare them to Donatello's known works.
1. **Identify the Artwork**: The image shows a detailed relief sculpture with figures in dynamic poses, which is characteristic of Renaissance art.
2. **Donatello's Style**: Donatello was an Italian sculptor of the early Renaissance period, known for his detailed and expressive sculptures. His works often feature human figures with a high degree of realism and emotional expression.
3. **Comparison to Known Works**: Donatello's famous works include "David," "St. George," and "The Feast of Herod." These sculptures are known for their intricate details, dynamic compositions, and the use of perspective.
4. **Analysis of the Image**: The relief sculpture in the image exhibits similar characteristics to Donatello's works: The figures are depicted with a high level of detail and realism. The composition is dynamic, with figures arranged in a way that suggests movement and interaction.
The use of perspective and the three-dimensional of the figures are indicative of Donatello's style.
5. **Conclusion**: Based on the stylistic similarities between the artwork in the image and Donatello's known works, it is reasonable to conclude that this artwork was indeed created by Donatello.
Therefore, the answer is: Yes.

Case D.3: MMMU: Financial Accounting

Hudson Alexander, Capital Dec. 1	$175,300
Dec. revenue	56,400
Dec. expenses	59,800

Question: Here are facts for the Hudson Roofing Company for December. <image 1> Assuming no investments or withdrawals, what is the ending balance in the owners' capital account?
A $171,900 B $170,000 C $172,500 D $181,900
Ground-truth answer: Yes

Model output (SDAR-VL-Instruct-8B):
To determine the ending balance in the owners' capital account, we need to consider the initial capital, the revenue, and the expenses for the month of December.
1. **Initial Capital**: The starting capital is $175,300.
2. **Revenue**: The revenue for December is $56,400.
3. **Expenses**: The expenses for December are $59,800.
To find the ending balance, we add the revenue to the initial capital and then subtract the expenses:
Ending Balance $=$ Initial Capital $+$ Revenue $-$ Expenses
Ending Balance $= 175,300 + 56,400 - 59,800$
Ending Balance $= 171,900$
Therefore, the ending balance in the owners' capital account is $171,900. The correct answer is A.

Case D.4: MMbench: finegrained_perception

Question: What direction is Czechia in the Mediterranean Sea?

A. east B. south C. west D. north

Ground-truth answer: D

Model output (SDAR-VL-Instruct-8B):
To determine the direction of Czechia relative to the Mediterranean Sea, we need to analyze the geographical positions of both locations on the map.
1. **Identify the Mediterranean Sea**: The Mediterranean Sea is located in the southern part of Europe, bordered by countries such as Spain, France, Italy, Greece, and others.
2. **Locate Czechia**: Czechia is situated in Central Europe, bordered by countries such as Poland, Slovakia, Austria, Germany, and the Czech Republic.
3. **Determine the Direction**:
The Mediterranean Sea is to the south of Czechia.
Therefore, Czechia is to the north of the Mediterranean Sea.
Given these observations, the correct answer is:
D. north

Case D.5: SeedBench: instance reasoning

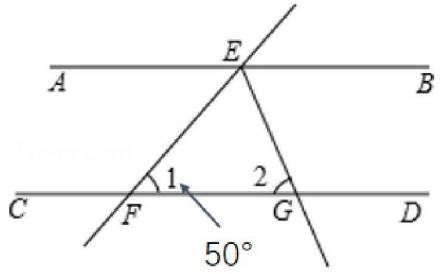
Question: Which object is closest to the crowd in the image?
A Airliner B TrainTrain C Jet D None of the above

Ground-truth answer: A

Model output (SDAR-VL-Instruct-8B):
To solve the problem, I will analyze the image to identify the object closest to the crowd. I will describe the image, focusing on the relevant details, and then use logical reasoning to determine the correct answer from the given options.
The image shows a large crowd of people gathered on a grassy area near a runway. A large airliner is visible on the runway, with its nose pointing towards the crowd. There is also a train track adjacent to the crowd, but a train is not visible in the image.
The question asks which object is closest to the crowd. The airliner is positioned on the runway, close to the crowd. The train track is next to the crowd, but no train is present. Therefore, the airliner is the object closest to the crowd.
A

Case D.6: Mathverse: geometry (angle)

As shown in the figure, AB parallel CD, EG bisects angle BEF, then angle 2 is equal to ()
Choices: A:50° B:60° C:65° D:90°



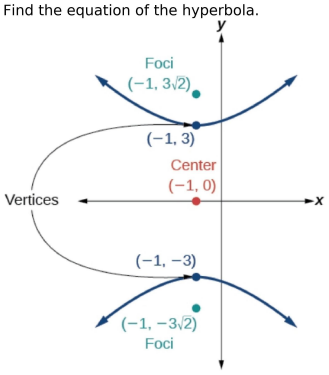
Question: According to the question shown in the image, first provide your reasoning, then give the measure of angle 2 and the correct option letter (A–D).

Ground-truth answer: 65° (option C).

Model output (SDAR-VL-Think-8B):
<think>
Okay, so I'm trying to solve this geometry problem. Let me analyze the image directly. The image shows two parallel lines, AB and CD, with a transversal EG intersecting them. Line EG bisects angle BEF, and there's an angle marked as 50 degrees at point F. The question asks for the measure of angle 2, which is equal to angle EGF.
First, let's focus on the parallel lines AB and CD. Since EG is a transversal, it creates corresponding angles and alternate interior angles. The image shows that EG bisects angle BEF, meaning it splits angle BEF into two equal parts.
Given that angle F is 50 degrees, and considering the properties of parallel lines, angle BEF must be supplementary to angle F. Therefore, angle BEF is 180 degrees minus 50 degrees, which equals 130 degrees. Since EG bisects angle BEF, each of the two angles created by the bisector will be half of 130 degrees, i.e., 65 degrees.
This means angles BEG and FEG are both 65 degrees. Now, consider angle EGF: since EG is a transversal and bisects angle BEF, angle EGF is one of the angles created by the bisector. Therefore, angle EGF is 65 degrees. Thus, angle 2, which is equal to angle EGF, is also 65 degrees.
</think>
The measure of angle 2 is 65 degrees, which corresponds to option C. Answer: C.

Case D.7: Mathverse: analytic geometry

Find the equation of the hyperbola.



Question: Find the equation of the hyperbola.

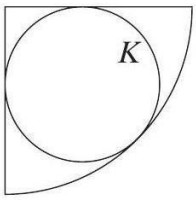
Ground-truth answer: $\frac{y^2}{9} - \frac{(x+1)^2}{9} = 1.$

Model output (SDAR-VL-Think-8B):
<think>
Okay, so I need to figure out the equation of the hyperbola based on the image provided. Let me start by recalling what I know about hyperbolas. A hyperbola is a type of conic section that has two separate branches. The standard form of its equation depends on whether it's oriented horizontally or vertically.
As seen in the image, the center is at $(-1, 0)$, which means the hyperbola is shifted to the left by 1 unit. The center is also the midpoint between the vertices, which are at $(-1, 3)$ and $(-1, -3)$. Since both vertices have the same x-coordinate, they lie vertically above and below the center. This indicates that the hyperbola is vertical, meaning its transverse axis is along the y-axis.
For a vertical hyperbola, the standard form of the equation is $\frac{(y-k)^2}{a^2} - \frac{(x-h)^2}{b^2} = 1$, where (h, k) is the center. In this case, $h = -1$ and $k = 0$, so the equation becomes $\frac{y^2}{a^2} - \frac{(x+1)^2}{b^2} = 1$.
Next, I need to determine the values of a and b. The distance from the center to each vertex is a. Since the vertices are at $(-1, 3)$ and $(-1, -3)$, the distance from the center $(-1, 0)$ to each vertex is 3 units. Therefore, $a = 3$, so $a^2 = 9$.
The foci of the hyperbola are at $(-1, 3\sqrt{2})$ and $(-1, -3\sqrt{2})$. The distance from the center to each focus is $c = 3\sqrt{2}$. For a hyperbola, the relationship between a, b, and c is $c^2 = a^2 + b^2$. Plugging in the known values gives $c^2 = 18$ and $a^2 = 9$, so $b^2 = 18 - 9 = 9$.
Putting everything together, the equation becomes $\frac{y^2}{9} - \frac{(x+1)^2}{9} = 1$.
</think>

Step-by-step summary. (1) Identify the center at $(-1, 0)$. (2) Use the vertical vertices $(-1, \pm 3)$ to get $a^2 = 9$. (3) Use the foci $(-1, \pm 3\sqrt{2})$ to get $c^2 = 18$ and hence $b^2 = 9$. (4) Substitute into the standard form of a vertical hyperbola.

Answer: $\frac{y^2}{9} - \frac{(x+1)^2}{9} = 1.$

Case D.8: Mathvision: metric geometry



Question: Question: A circle K is inscribed in a quarter circle with radius 6 as shown in the figure. What is the radius of circle K?
(A) $\frac{6-\sqrt{2}}{2}$ (B) $\frac{3\sqrt{2}}{2}$ (C) 2.5 (D) 3 (E) $6(\sqrt{2}-1)$

Ground-truth answer: E

Model output (SDAR-VL-Think-8B):
<think>
Okay, so I have this geometry problem involving a circle inscribed in a quarter circle. The quarter circle has a radius of 6, and I need to find the radius of the smaller circle K.
First, let's consider the image itself. The image shows a quarter circle with a radius of 6, and inside it, there's a smaller circle K that is tangent to both the arc and the two sides of the quarter circle. This means the smaller circle K is touching the curved part of the quarter circle and the two straight sides.
Next, let's focus on the positioning of the smaller circle K. As seen in the image, the center of the smaller circle K is at a distance equal to its radius from both sides of the quarter circle. If we place the quarter circle in a coordinate system with its center at the origin (0,0), the two sides of the quarter circle would be along the x-axis and y-axis. The center of the smaller circle K would then be at some point (r, r), where r is the radius of the smaller circle.
Now, let's consider the distance from the center of the smaller circle K to the arc of the quarter circle. This distance is the radius of the quarter circle minus the radius of the smaller circle, which is 6 - r. However, this distance is also the hypotenuse of a right triangle with legs of length r each. Therefore, we can use the Pythagorean theorem to express this distance as $\sqrt{r^2 + r^2} = \sqrt{2r^2} = r\sqrt{2}$.
Setting these two expressions for the distance equal to each other, we get:

$$6 - r = r\sqrt{2}$$

Solving for r, we rearrange the equation:

$$6 = r + r\sqrt{2}$$

$$6 = r(1 + \sqrt{2})$$

$$r = \frac{6}{1 + \sqrt{2}}$$

To rationalize the denominator, we multiply the numerator and the denominator by the conjugate of the denominator:

$$r = \frac{6(1 - \sqrt{2})}{(1 + \sqrt{2})(1 - \sqrt{2})}$$

$$r = \frac{6(1 - \sqrt{2})}{1 - 2}$$

$$r = \frac{6(1 - \sqrt{2})}{-1}$$

$$r = -6(1 - \sqrt{2})$$

$$r = 6(\sqrt{2} - 1)$$

Thus, the radius of the smaller circle K is $6(\sqrt{2} - 1)$.
Answer: E
</think>

The radius of the smaller circle inscribed in the quarter circle is calculated using geometric relationships and the Pythagorean theorem. By setting up an equation based on the distances from the center of the smaller circle to the sides and the arc, we determine that the radius is $6(\sqrt{2} - 1)$.

Answer: E