IMAGINE A LITTLE, THEN COORDINATE AND RETHINK: MODALITY-ADAPTIVE ASYNCHRONOUS DENOISING FOR UNIFIED DIFFUSION VISION-LANGUAGE-ACTION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified discrete-diffusion vision-language-action (VLA) models bridge world modeling and robot control by jointly generating future visual observations and actions within a single multimodal Transformer, providing explicit visual foresight for robotic manipulation. However, these models treat multimodal outputs as a homogeneous generation process, denoising them synchronously under a single shared schedule and overlooking the inherent heterogeneity across modalities: future-visual predictions are not only generation targets but also evolving conditions for action generation. Our analysis reveals distinct recovery dynamics across modalities and sample-dependent action responses to visual recovery. Motivated by these findings, we introduce **MAAD-VLA**, a modality-adaptive asynchronous denoising framework that preserves modality-specific progression while adaptively coordinating vision and action. During training, MAAD-VLA independently samples visual and action noise levels, exposing the model to diverse combinations of cross-modal corruption so that it learns to denoise even when the two modalities are at different recovery stages. At inference, a response-guided two-stage denoising process first establishes partial visual foresight and then adaptively initiates action denoising alongside continued visual recovery under modality-specific schedules. We further introduce Action-centric Reconsideration, which enables self-correction by selectively remasking committed action tokens whose confidence drops sharply as the visual context evolves. Extensive experiments in both simulation and real-world environments demonstrate the effectiveness of MAAD-VLA, notably achieving an average task length of 4.82 on the CALVIN ABCD→D benchmark.

1 INTRODUCTION

Vision-language-action (VLA) models have made substantial progress in robotic manipulation tasks by transferring rich pretrained vision-language representations to robot control Kim et al. (2025a); Black et al. (2026). However, conventional action-only VLAs primarily learn mappings from observations and instructions to actions, leaving the dynamics of physical interaction implicit in the policy rather than explicitly modeling how their actions shape future states of the environment.

Recent efforts seek to bridge world modeling and action generation by incorporating future-state prediction into VLA learning CHEN et al. (2026); Yuan et al. (2026); Ma et al. (2026), allowing anticipated scene evolution to guide robot control. Among these approaches, unified discrete-diffusion VLAs, exemplified by UD-VLA CHEN et al. (2026), provide an end-to-end paradigm bridging perception, predictive foresight, and action generation within a unified generative process, enabling tightly coupled optimization and coherent cross-modal coordination. Despite these benefits, these models treat heterogeneous future-visual and action tokens as a single concatenated sequence and denoise them under a shared schedule, which may not adequately accommodate their distinct recovery dynamics and evolving conditioning requirements. In particular, the two modalities serve distinct functional roles: future-visual predictions are not only outputs to be recovered but also evolving conditions for action generation, reflecting an inherent heterogeneity between vision and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

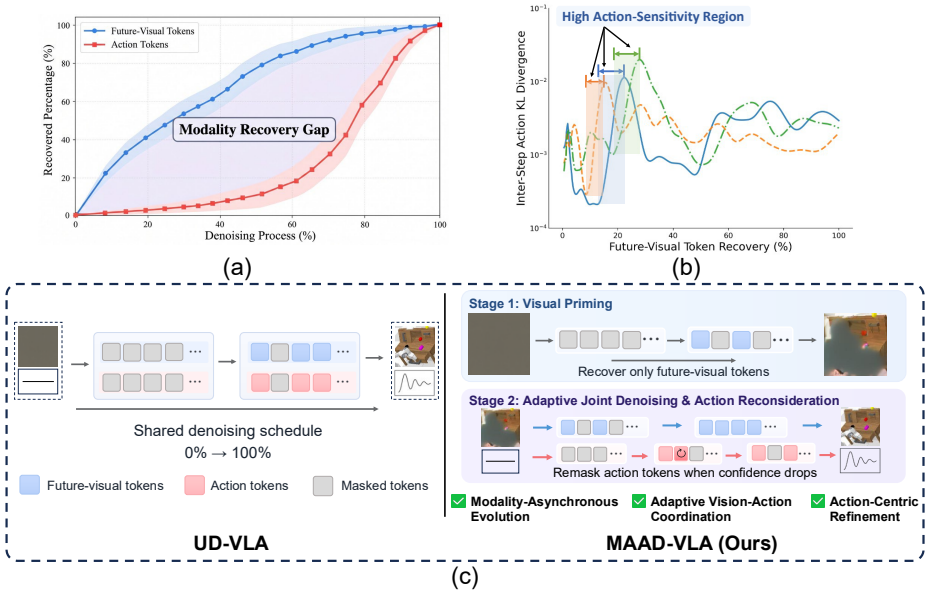


Figure 1: **(a) Modality-specific recovery dynamics.** Under joint confidence-guided decoding, future-visual and action tokens exhibit distinct denoising rates and recovery patterns. **(b) Action responses (KL) to visual recovery conditions.** We use KL divergence to quantify shifts in action predictive distributions as future-visual tokens are progressively recovered, with larger values indicating stronger action responses to newly recovered visual information. **(c) Denoising paradigm comparison.** UD-VLA follows a shared denoising schedule over future-visual and action tokens. In contrast, our MAAD-VLA first establishes partial future-visual context through *visual priming*, followed by *asynchronous joint denoising* that allows modality-specific progression while adaptively coordinating action generation with the evolving visual conditions.

action. This raises a central question: ***how should visual and action denoising be coordinated to exploit evolving visual foresight while accommodating modality-specific recovery dynamics?***

To address this question, we first examine how modality heterogeneity manifests in the joint vision–action denoising process of UD-VLA. As shown in Fig. 1(a), the two output modalities exhibit markedly different recovery dynamics under the same confidence-guided denoising process. Visual tokens gain confidence earlier and thus dominate recovery in the early iterations, while action tokens remain largely masked before being recovered rapidly at later stages. These distinct recovery patterns reveal heterogeneous denoising dynamics across modalities, suggesting that a uniform denoising schedule may not adequately accommodate their differences in recovery timing and pace. To probe this heterogeneity from an action-centric perspective, we examine both how action predictions respond to progressive future-visual recovery and how these responses vary across samples. Concretely, we keep action tokens fully masked and measure the KL divergence between consecutive action predictive distributions, capturing how action predictions evolve as additional visual context is progressively revealed Wu & Huang (2026); Kim et al. (2025b). As shown in Fig. 1(b), pronounced action-response regions emerge at different visual recovery stages across trajectories, revealing sample-dependent action sensitivity to evolving visual conditions and heterogeneous vision–action coordination dynamics. Taken together, treating heterogeneous outputs as a single denoising sequence can overlook the distinct evolution of each modality, particularly in robotic manipulation where action generation is the ultimate objective. This motivates a more adaptive, action-centric denoising paradigm that exploits evolving visual foresight to better guide action generation.

Therefore, we propose **MAAD-VLA**, a **M**odality-**A**daptive **A**synchronous **D**enoising framework for unified diffusion **VLA**s. To explicitly accommodate modality heterogeneity, MAAD-VLA combines decoupled noise sampling during training with adaptive asynchronous denoising at inference, coordinating action refinement with the progressive recovery of future visual information. During training, we independently sample noise levels for the two output modalities and apply *vision-first ordering* to provide more informative visual conditioning for action prediction. This design exposes the model to diverse corruption pairs and lays the foundation for asynchronous denoising at inference. At inference, as illustrated in Fig. 1(c), MAAD-VLA follows a two-stage *imagine-*

coordinate&rethink process, in contrast to UD-VLA's shared denoising schedule. It first performs *Visual Priming* to establish partial denoised future-visual foresight while keeping action tokens fully masked. When the action-response signal begins to rise sharply, it transitions to *Asynchronous Joint Denoising*, where vision and action are denoised concurrently at modality-specific rates. To further correct premature action commitments under continuously evolving visual conditions, we introduce *Action-centric Reconsideration*, a selective remasking mechanism that identifies unreliable commitments through cross-step confidence degradation and remasks them for subsequent refinement.

Overall, MAAD-VLA unifies modality and instance-wise adaptivity within a coherent vision–action denoising framework, enabling modality-specific progression while coordinating action generation with diverse and evolving visual conditions. Our contributions are fourfold:

- We identify heterogeneous recovery patterns across modalities and samples in unified diffusion VLAs, which motivates us to propose **MAAD-VLA**, a modality-adaptive asynchronous denoising framework.
- We introduce modality-decoupled learning that independently samples visual and action timesteps during training, covering diverse cross-modal noise-level combinations to support asynchronous inference.
- We develop response-guided two-stage inference that first builds partial visual foresight, then adaptively initiates action denoising. Vision and action then recover concurrently under modality-specific schedules, with Action-centric Reconsideration enabling self-correction by remasking committed action tokens whose confidence drops sharply.
- Extensive experiments on multiple simulation benchmarks and real-world robotic tasks demonstrate the effectiveness of MAAD-VLA, achieving an average successful sequence length of 4.82 on the long-horizon CALVIN ABCD→D.

2 Related Work

2.1 World-Model-Enhanced VLAs

Vision-language-action (VLA) models have achieved strong progress in robotic manipulation by transferring vision-language priors to embodied control Brohan et al. (2023); Kim et al. (2025a); Black et al. (2026). However, most action-centric VLAs directly map observations and instructions to actions without explicitly modeling future environmental dynamics, often leading to myopic decisions and limited temporal reasoning Zhao et al. (2025); Lv et al. (2025). Recent studies therefore incorporate world modeling into policy learning to provide foresight for action generation. One line of work couples video-generative world models with dedicated action experts or action-specific branches, enabling predicted future dynamics to facilitate action learning and generation Li et al. (2026); Ma et al. (2026); Bi et al. (2026); Yuan et al. (2026). While these methods provide high-fidelity visual predictions, many of them retain modality-specific pathways for world prediction and action generation. Consequently, their interaction relies on cross-branch information exchange, limiting fine-grained co-adaptation between future prediction and control. Another line pursues *native unified architectures*, discretizing multimodal inputs and outputs into a shared token space for joint vision-action modeling and optimization Cen et al. (2025); Wang et al. (2026b); CHEN et al. (2026); Yu et al. (2026). Within this token-unified paradigm, UniVLA Wang et al. (2026b) models vision, language, and action autoregressively, but its token-by-token decoding limits inference parallelism. UD-VLA CHEN et al. (2026) addresses this limitation through joint discrete denoising of future visual and action tokens, enabling parallel prediction and iterative visual conditioning for action generation; however, it refines both modalities under a shared synchronous schedule that may overlook their heterogeneous generation dynamics. Our MAAD-VLA elaborates the unified discrete paradigm by enabling vision and action to evolve independently while adaptively coordinating cross-modal interaction through response-guided asynchronous denoising.

2.2 Modality Heterogeneity in Multimodal Generation

Native unified multimodal models Team (2024); Xie et al. (2025); Yang et al. (2026) have shown strong potential for jointly modeling multimodal understanding and generation, yet sharing a common backbone does not imply homogeneous modeling or generation requirements across modalities.

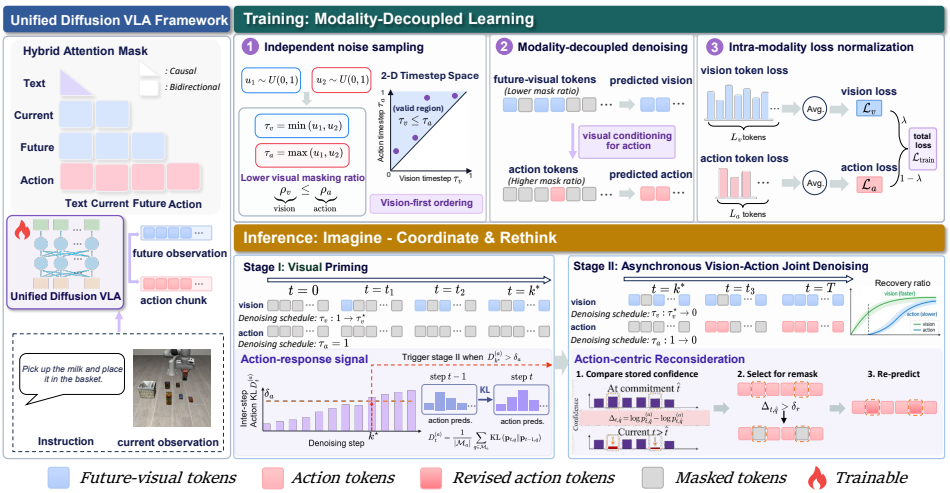


Figure 2: **Overview of MAAD-VLA.** Built upon a unified diffusion VLA, MAAD-VLA introduces modality-adaptive denoising at both training and inference. During training, MAAD-VLA combines modality-specific timestep sampling under a vision-first ordering with modality-wise loss normalization, exposing the model to diverse corruption distributions. At inference, it follows an *Imagine–Coordinate & Rethink* process: *Visual Priming* builds partial visual foresight while keeping action tokens fully masked. An inter-step action KL signal triggers *Asynchronous Vision–Action Joint Denoising* with modality-specific schedules, while *Action-centric Reconsideration* selectively remasks committed action tokens whose confidence drops as visual context evolves.

Transfusion Zhou et al. (2025) combines next-token prediction for text with diffusion for images, demonstrating that architectural unification can coexist with modality-specific objectives. Along this direction, UniDiffuser Bao et al. (2023) assigns independent diffusion timesteps to different modalities, while MMaDA-Parallel Tian et al. (2026) adopts modality-specific recovery schedules for parallel text-image generation, using linear recovery for text and cosine recovery for images while preserving cross-modal interaction throughout denoising. Modality heterogeneity has also received increasing attention in robotic manipulation. UWM Zhu et al. (2025) models video and action diffusion with independent modality timesteps, while DUST Won et al. (2025) introduces independent modality perturbations and asynchronous sampling with predefined modality-wise update rates; X-WAM Guo et al. (2026) similarly allocates different denoising budgets to action and visual prediction. However, these approaches primarily handle heterogeneity through modality-level designs or predefined schedules, leaving trajectory-specific coordination timing underexplored. Our analysis shows that action predictions become responsive to progressively recovered visual foresight at different stages across trajectories. Accordingly, MAAD-VLA preserves modality-specific denoising progress and adaptively determines when action generation should engage according to trajectory-dependent action responses to evolving visual foresight.

3 METHOD

We propose MAAD-VLA, a modality-adaptive asynchronous denoising framework that enables modality-specific evolution with adaptive vision–action coordination. Its overall pipeline is illustrated in Fig. 2. We first review the preliminaries of the unified discrete denoising paradigm for VLA modeling (Sec. 3.1), then present our training and inference paradigms in Sec. 3.2 and 3.3, respectively. Our modality-decoupled training is tailored to support asynchronous denoising at inference, aligning the two within a coherent framework for adaptive vision–action coordination.

3.1 PRELIMINARIES

Unified Token Representation. Building on recent advances in unified multimodal models, exemplified by Emu3 Wang et al. (2024), a growing line of VLA frameworks Wang et al. (2026b); CHEN et al. (2026); Yu et al. (2026) has sought to unify language, vision, and action within a shared discrete token space. Following this paradigm, we represent the three modalities in a shared discrete vocab-

ulary Ω using Emu3 language tokenizer Wang et al. (2024), a VQ tokenizer Zheng et al. (2022), and FAST Pertsch et al. (2025), respectively. The conditioning context is defined as $[\mathbf{x}^{\text{lang}}; \mathbf{V}]$, comprising the language instruction tokens $\mathbf{x}^{\text{lang}}$ and the current observation tokens $\mathbf{V}$. Following UD-VLA CHEN et al. (2026), we generate future-visual tokens $\widetilde{\mathbf{V}}$ for explicit visual foresight and an action chunk $\mathbf{A}$, with respective token lengths L_v and L_a. The full multimodal sequence is:

$$\underbrace{[\texttt{BOS}]\ \mathbf{x}^{\text{lang}}; [\texttt{BOI}]\ \mathbf{V}\ [\texttt{EOI}];}_{\text{condition}} \underbrace{[\texttt{BOI}]\ \widetilde{\mathbf{V}}\ [\texttt{EOI}]; [\texttt{BOA}]\ \mathbf{A}\ [\texttt{EOA}]}_{\text{output}} \tag{1}$$

where [BOX] and [EOX] are special tokens that mark the beginning and end of a modality sequence.

Masked Diffusion and Scheduling. Following absorbing-state discrete diffusion Austin et al. (2021); Sahoo et al. (2024), we sample a training timestep $\tau \in [0, 1]$ and mask output content tokens with $[\texttt{M}] \in \Omega$ at ratio $g(\tau)$, where g is the mask scheduling function. We adopt the cosine schedule $g(\tau) = \cos\left(\frac{\pi}{2}(1-\tau)\right)$. At inference, generation starts from fully masked content and proceeds through T iterative denoising steps. Specifically, inference step $t \in \{0, \ldots, T\}$ is mapped to the normalized diffusion time $\tau_t = 1 - t/T$, decreasing linearly from $\tau_0 = 1$ (fully masked) to $\tau_T = 0$ (fully denoised). The corresponding target mask ratio is $\rho_t = g(\tau_t)$.

Conditional Prediction and Confidence-based Decoding. Following CHEN et al. (2026), we adopt hybrid attention that is bidirectional within each output modality and causal across blocks. Both output modalities attend to the conditioning prefix $[\mathbf{x}^{\text{lang}}; \mathbf{V}]$, while action tokens additionally attend to future vision $\widetilde{\mathbf{V}}$. This allows action prediction to leverage intermediate visual features throughout denoising. With the environment time-step index omitted for brevity, the model predicts categorical distributions over masked tokens in each modality at denoising iteration t:

$$\begin{aligned} \mathbf{p}_t^{(v)} &= p_\theta^{(v)}\left(\cdot \mid \mathbf{x}^{\text{lang}}, \mathbf{V}, \widetilde{\mathbf{V}}_{t-1}\right), \\ \mathbf{p}_t^{(a)} &= p_\theta^{(a)}\left(\cdot \mid \mathbf{x}^{\text{lang}}, \mathbf{V}, \widetilde{\mathbf{V}}_{t-1}, \mathbf{A}_{t-1}\right), \end{aligned} \tag{2}$$

where $\mathbf{p}_t^{(m)}$ denotes the conditional token distribution for modality $m \in \{v, a\}$ produced by the unified model with parameters θ, and $\mathbf{p}_{t,q}^{(m)}$ denotes the probability assigned to the token at position q. Masked tokens are progressively recovered via confidence-guided greedy decoding. Specifically, we rank masked positions in the unified output sequence by their maximum predicted token probabilities and unmask the highest-confidence candidates. Each masked token is updated as follows:

$$x_{t,q}^{(m)} = \begin{cases} [\texttt{M}], & \text{otherwise}, \\ \arg\max_x \mathbf{p}_{t,q}^{(m)}(x), & \text{if } \operatorname{rank}_{t,q}^{(m)}(x)_\uparrow > \lceil (L_v + L_a)\, \rho_t \rceil, \end{cases} \tag{3}$$

where $\operatorname{rank}_{t,q}^{(m)}$ denotes the confidence rank of position q among all currently masked output positions, with larger ranks indicating higher confidence. After the final step, the fully unmasked token sequences are decoded into future observations and robot actions using their respective tokenizers.

3.2 Training: Modality-Decoupled Learning

Modality-Specific Noise Sampling with Vision-First Ordering. Despite enabling iterative cross-modal interaction, UD-VLA synchronizes future vision and action under a shared diffusion time τ_t, overlooking their heterogeneous denoising dynamics and modality-specific conditioning requirements. Therefore, we decouple the diffusion times of future vision and action during training, exposing the model to independently corrupted modality states and enabling it to accommodate asynchronous vision–action denoising at inference. This strategy has also been adopted by a series of multimodal diffusion models to support flexible generation under varying levels of conditioning Bao et al. (2023); Zhu et al. (2025). Specifically, for each training sample, we independently draw two base noise coordinates:

$$u_1, u_2 \overset{\text{i.i.d.}}{\sim} \mathcal{U}(0, 1). \tag{4}$$

To reflect the dynamics observed in Fig. 1(a), where informative visual context is recovered at a faster pace than action prediction, we impose a *vision-first* ordering on the two sampled coordinates:

$$\tau^{(v)} = \min(u_1, u_2), \qquad \tau^{(a)} = \max(u_1, u_2). \tag{5}$$

Since scheduling function g is strictly increasing over $[0, 1]$, this ordering is directly preserved in the corresponding mask ratios:

$$\tau^{(v)} \leq \tau^{(a)} \quad \Longrightarrow \quad \rho^{(v)} = g\left(\tau^{(v)}\right) \leq g\left(\tau^{(a)}\right) = \rho^{(a)}. \tag{6}$$

Training on such asynchronous vision–action states enables the model to denoise actions using partially recovered visual context, paving the way for visual priming at inference.

Objective. Denote $\mathcal{M}_v$ and $\mathcal{M}_a$ as the masked content positions in $\widetilde{\mathbf{V}}$ and $\mathbf{A}$, respectively. We then average the prediction loss over masked positions within each modality:

$$\mathcal{L}_v = -\frac{1}{|\mathcal{M}_v|} \sum_{q \in \mathcal{M}_v} \log \mathbf{p}_{t,q}^{(v)}, \qquad \mathcal{L}_a = -\frac{1}{|\mathcal{M}_a|} \sum_{q \in \mathcal{M}_a} \log \mathbf{p}_{t,q}^{(a)}. \tag{7}$$

Unlike UD-VLA, which normalizes the loss over the entire output sequence, we normalize it within each modality by the number of masked tokens, preventing the objective from being dominated by future-visual tokens, which contain over ten times as many tokens as action sequences. The final training objective combines the two modality-specific losses as

$$\mathcal{L}_{\text{train}} = \lambda \mathcal{L}_v + (1 - \lambda)\mathcal{L}_a, \tag{8}$$

where $\lambda \in [0, 1]$ balances contribution of two modalities.

3.3 Inference: Modality-Adaptive Asynchronous Denoising

Modality-decoupled training enables action prediction conditioned on future-visual states at varying noise levels. As visual tokens tend to recover earlier than action tokens (Fig. 1(a)), we adopt two-stage denoising: partial visual recovery to establish foresight, followed by asynchronous joint denoising with modality-specific schedules. Formally, the target mask ratios under schedule g evolve across the two stages as follows:

$$(\rho_v, \rho_a): \quad (1, 1) \xrightarrow{\text{Stage I}} \left(g\left(\tau_v^\star\right), 1\right) \xrightarrow{\text{Stage II}} (0, 0) \tag{9}$$

where $\tau_v^\star$ denotes the normalized visual denoising timestep at the stage transition, and $g(\tau_v^\star)$ is the corresponding remaining visual mask ratio. An action-response criterion guides this stage transition, while action-centric reconsideration remasks and re-predicts action tokens based on inter-step confidence drop, enabling late-stage correction under progressively refined visual context.

Stage I: Visual Priming. Inspired by the cognitive notion of *priming*, where prior stimuli can facilitate subsequent processing (Dehaene et al., 1998; Vorberg et al., 2003), we regard this partially recovered future-visual state as a *visual primer* for subsequent action generation. During this stage, τ_v decreases from 1 to $\tau_v^\star$, while action tokens remain fully masked. Using action-token distributions from the same forward passes, we compute the mean inter-iteration Kullback–Leibler (KL) divergence over action positions as an online action-response signal:

$$D_t^{(a)} = \frac{1}{|\mathcal{M}_a|} \sum_{q \in \mathcal{M}_a} \mathrm{KL}\left(\mathbf{p}_{t,q}^{(a)} \| \mathbf{p}_{t-1,q}^{(a)}\right), \quad t \geq 2. \tag{10}$$

A larger $D_t^{(a)}$ indicates a stronger action response to evolving visual context, providing a cue for when to initiate asynchronous joint denoising.

Stage II: Asynchronous Vision–Action Joint Denoising. Motivated by the observation in Fig. 1(b) that the action-response signal $D_t^{(a)}$ rises sharply within a window of future-visual recovery, we trigger asynchronous joint denoising at the first step where this signal exceeds a predefined threshold:

$$k^\star = \min\left\{k \in \{1, \ldots, T\} \mid D_k^{(a)} > \delta_a\right\}, \quad \tau_v^\star = 1 - \frac{k^\star}{T} \tag{11}$$

where $\delta_a > 0$ is the KL response threshold. Since action denoising begins after iteration $k^\star$, its available denoising budget is $T - k^\star$ steps.

Action-centric Reconsideration. During Stage II, earlier action commitments may lose support as the joint denoising context evolves. We revisit these commitments through confidence-drop remasking. For a previously revealed action token at position q in $\mathbf{A}_{t-1}$, let $\hat{t}_q < t$ denote its most recent mask-to-token commitment iteration. We measure its confidence degradation as

$$\Delta_{t,q} = \log \mathbf{p}_{\hat{t}_q,q}^{(a)} - \log \mathbf{p}_{t,q}^{(a)}, \tag{12}$$

Table 1: **Comparison on the CALVIN ABCD → D benchmark.** We report the success rates of completing 1–5 consecutive tasks and the average successful sequence length. RT-1 results are from the reimplementation reported by Li et al. (2024).

Method	Tasks Completed in a Row					Avg. Len. ↑
	1	2	3	4	5	
MCIL (Lynch & Sermanet, 2020)	0.373	0.027	0.002	0.000	0.000	0.40
RT-1 (Brohan et al., 2022)	0.844	0.617	0.438	0.323	0.227	2.45
Robo-Flamingo (Li et al., 2024)	0.964	0.896	0.824	0.740	0.660	4.09
DeeR (Yue et al., 2024)	0.982	0.902	0.821	0.759	0.670	4.13
GR-1 (Wu et al., 2024)	0.949	0.896	0.844	0.789	0.731	4.21
ReconVLA (Song et al., 2026)	0.980	0.900	0.845	0.785	0.705	4.23
MoDE (Reuss et al., 2025)	0.971	0.925	0.879	0.835	0.779	4.39
UP-VLA (Zhang et al., 2025)	0.962	0.921	0.879	0.842	0.812	4.42
MDT (Reuss et al., 2024)	0.986	0.958	0.916	0.862	0.801	4.52
UniVLA (Wang et al., 2026b)	0.985	0.961	0.931	0.899	0.851	4.63
UD-VLA (CHEN et al., 2026)	0.992	0.968	0.936	0.904	0.840	4.64
MAAD-VLA	**0.996**	**0.986**	**0.974**	**0.954**	**0.914**	**4.82**

where both probabilities correspond to the same committed token, evaluated at commitment and at the current iteration, respectively. At each eligible update, we select up to K_r previously committed action positions with the largest drops satisfying $\Delta_{t,q} > \delta_r$. To avoid excessive revisions caused by unstable early predictions, we enable remasking only from iteration (t_{min}) within Stage II. Selected positions are additionally masked in $\mathbf{A}_t$ and regenerated in subsequent denoising updates. Different from prior remasking criteria based on instantaneous token confidence Liang et al. (2025); Yao (2026), our mechanism tracks cross-step confidence degradation after commitment.

4 Experiment

4.1 Benchmarks

CALVIN ABCD→D. CALVIN Mees et al. (2022) evaluates long-horizon, language-conditioned robotic manipulation across four simulated environments, denoted A–D. We adopt the ABCD→D setting, training on demonstrations from all four environments and evaluating the policy in D. Each evaluation trial comprises a sequence of five subtasks drawn from a diverse set of language-specified goals, totaling up to 1,000 unique sequences. Following the evaluation configuration of UD-VLA CHEN et al. (2026), we report the success rates for completing 1–5 consecutive subtasks (ranging from 0 to 1) and the average successful sequence length (ranging from 0 to 5). Implementation details are provided in Appendix B.

LIBERO Liu et al. (2023). LIBERO is a simulated robot manipulation benchmark comprising four task suites: LIBERO-Spatial, LIBERO-Object, LIBERO-Goal, and LIBERO-Long. Each suite contains 10 diverse tasks with 50 human-teleoperated demonstrations per task, evaluating spatial reasoning, object manipulation, goal-conditioned control, and long-horizon task execution, respectively. We report the success rate on each suite, together with the overall average across all four suites.

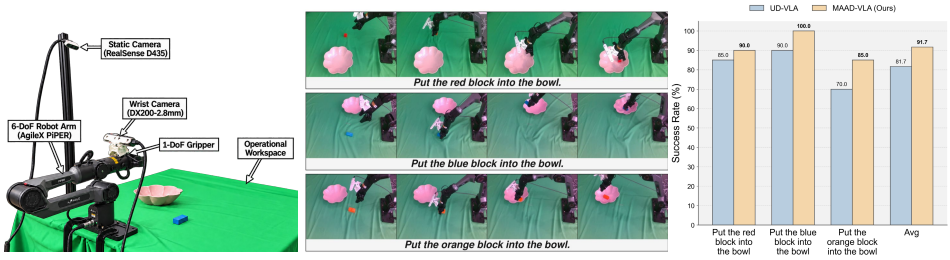
Real-World Evaluation. As shown in Fig. 3, we conduct real-world experiments using an AgileX PiPER 6-DoF robotic arm with a 1-DoF gripper. Visual observations are collected from two complementary viewpoints: a RealSense D435 captures the third-person view, while a gripper-mounted DX200-2.8mm camera provides the wrist view. The robot performs three language-conditioned tasks, each requiring it to place a block of the specified color into a bowl. For training, we collect 50 demonstration trajectories per task at 30 Hz. UD-VLA and MAAD-VLA are trained with the same optimization hyperparameters and each is evaluated over 20 rollouts per task with varied object placements, using an action chunk size of 20.

4.2 Main Results

CALVIN ABCD→D. As shown in Table 1, our MAAD-VLA outperforms all baselines in both consecutive-task completion and average successful sequence length, achieving an average length

Table 2: **Evaluation and comparison on the LIBERO benchmark.**

Method	Spatial	Object	Goal	Long	Average
Octo (Mees et al., 2024)	78.9%	85.7%	84.6%	51.1%	75.1%
SpatialVLA (Qu et al., 2025)	88.2%	89.9%	78.6%	55.5%	78.1%
CoT-VLA (Zhao et al., 2025)	87.5%	91.6%	87.6%	69.0%	83.9%
WorldVLA (Cen et al., 2025)	87.6%	96.2%	83.4%	60.0%	81.8%
ThinkAct (Huang et al., 2026)	88.3%	91.4%	87.1%	70.9%	84.4%
π_0-FAST (Pertsch et al., 2025)	96.4%	**96.8%**	88.6%	60.2%	85.5%
MolmoAct (Lee et al., 2025)	87.0%	95.4%	87.6%	77.2%	86.8%
FlowVLA (Zhong et al., 2026)	93.2%	95.0%	**91.6%**	72.6%	88.1%
DreamVLA (Zhang et al., 2026)	**97.5%**	94.0%	89.5%	89.5%	92.6%
UD-VLA (CHEN et al., 2026)	94.1%	95.7%	91.2%	89.6%	92.7%
MAAD-VLA	96.2%	94.8%	**91.6%**	**90.8%**	**93.4%**



(a) (b) (c)

Figure 3: **Real-world evaluation.** (a) PiPER robot setup. (b) Real-world manipulation tasks. (c) Success rates comparison.

of 4.82 and demonstrating stronger long-horizon execution capability. Notably, the gains over UD-VLA are particularly pronounced on longer task sequences, with improvements of 5.0 and 7.4 percentage points on 4/5 and 5/5 task completion, respectively. These increasingly larger gains suggest that MAAD-VLA effectively mitigates compounding errors across consecutive actions through adaptive cross-modal coordination, complemented by iterative action-token refinement that enables more reliable action generation during inference.

LIBERO. As presented in Table 2, MAAD-VLA achieves state-of-the-art performance on 2 out of the 4 task suites within the LIBERO benchmark, yielding the highest overall average success rate. In particular, on the more challenging LIBERO-Long suite, MAAD-VLA achieves the highest success rate of 90.8%, surpassing UD-VLA by 1.2 % and highlighting its effectiveness in complex manipulation tasks. Consistent with the larger gains observed on CALVIN, this result further supports the benefits of adaptive cross-modal coordination for reliable long-horizon manipulation.

Real-World Evaluation. As shown in Fig. 3(c), MAAD-VLA consistently outperforms UD-VLA across all three tasks, improving the overall success rate by 10% with varied object placements. We attribute these larger gains in the real world to the stronger need for adaptive vision–action coordination under varying object placements and execution uncertainty: MAAD-VLA allows action denoising to respond to the evolving visual context and selectively revises premature action commitments, leading to more reliable control than a fixed synchronous denoising process.

4.3 In-Depth Analysis

Component Ablations. Table 3 evaluates our training and inference components on CALVIN. Our training strategies, decoupled noise sampling and intra-modality loss normalization, accommodate heterogeneous corruption states and balance modality-wise supervision, jointly improving the average successful sequence length from 4.45 to 4.61 under a shared inference schedule. At inference, two-stage asynchronous denoising allows vision and action to evolve more flexibly at modality-specific paces while maintaining adaptive cross-modal coordination, further increasing the Avg. length to 4.73. Even without our training strategies, it improves the baseline from 4.45 to 4.60, demonstrating its standalone effectiveness. Moreover, Action-centric Reconsideration equips

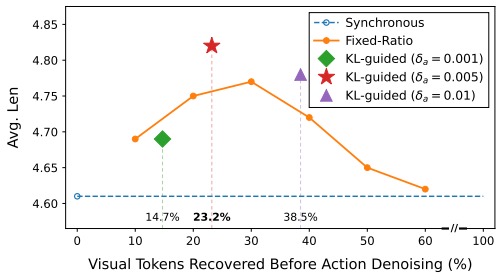


Figure 4: **Effect of Different Action-Denoising Start Strategies on CALVIN.**

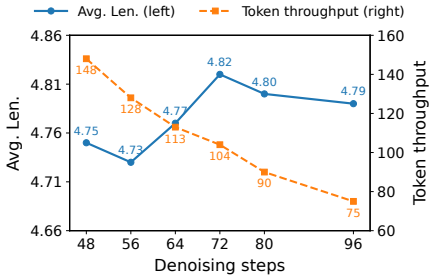


Figure 5: **Performance–efficiency trade-off.**

the model with self-correction capabilities as visual context evolves, yielding an additional improvement. These results highlight the complementary benefits of our training and inference strategies.

Table 3: **Component ablations on CALVIN.** Base/Full: baseline/full training.

Setting	Avg. Len. ↑	
Training (shared inference schedule)		
Baseline	4.45	
Independent noise only	4.52	
Intra-norm. only	4.56	
Both (ours)	4.61	
Inference	**Base**	**Full**
Two-stage Async.	4.60	4.73
+ Remask	4.66	**4.82**

Action-Denoising Transition Strategies. We compare synchronous, fixed-ratio, and KL-guided starts in Fig. 4. Synchronous denoises vision and action jointly from the beginning, while fixed-ratio delays action until a preset visual recovery ratio. *KL-guided* adaptively triggers the transition using threshold δ_a in Eq. 11, with dashed lines marking the mean visual recovery ratio at transition across evaluation samples. Fixed visual priming provides better initial grounding than synchronous decoding, but ignores sample-specific conditioning needs; under a fixed budget ($T = 72$), excessive visual priming also reduces the action-denoising budget and degrades performance. In contrast, our KL-guided strategy with $\delta_a = 0.005$ achieves the best performance of 4.82 by adapting the transition to sample-specific action responses, initiating action denoising at 23.2% visual recovery on average and balancing visual priming with action refinement.

Performance-efficiency trade-off. Fig. 5 shows that using 72 denoising steps strikes a favorable balance between performance and efficiency, while larger budgets reduce throughput with no further performance gain. Here, token throughput measures the number of output tokens per second. Unlike token-by-token autoregressive decoding, our diffusion-based formulation supports parallel token generation within each denoising step, offering the potential for high-throughput inference.

Table 4: **Remasking strategies.**

Remasking strategy	Avg. Len. ↑
No remasking	4.73
Random	4.54
Low conf.	4.76
Conf. drop (ours)	**4.82**

Remasking Strategy. Table 4 compares action-token remasking criteria. Random remasking selects tokens uniformly, while low-confidence remasking relies solely on current confidence. Our criterion instead prioritizes cross-step confidence drops relative to each token's most recent commitment. Its gains over the other strategies indicate that cross-step confidence degradation provides a more effective remasking signal than current confidence alone, helping identify earlier action commitments that lose support as visual context evolves. Further analysis of remasking is provided in Appendix E.

5 Conclusion

In this work, we propose MAAD-VLA, a modality-adaptive asynchronous denoising framework for unified diffusion VLAs. Motivated by distinct modality recovery dynamics and sample-dependent action responses to evolving visual foresight, MAAD-VLA combines modality-decoupled training with a two-stage asynchronous inference strategy: visual priming first establishes partial future-visual foresight, after which an online action-response signal triggers modality-specific joint denoising, complemented by action reconsideration. Extensive experiments demonstrate that MAAD-VLA improves manipulation performance through adaptive vision–action coordination, highlighting heterogeneity-aware modeling as a promising direction for world-model-based VLAs.

REFERENCES

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 17981–17993. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/958c530554f78bcd8e97125b70e6973d-Paper.pdf.

Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *International Conference on Machine Learning*, pp. 1692–1717. PMLR, 2023.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, Hongyan Zhao, Hanyu Liu, Zhizhong Su, Lei Ma, Hang Su, and Jun Zhu. Motus: A unified latent action world model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 35101–35113, June 2026.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control, 2026. URL https://arxiv.org/abs/2410.24164.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. 2022.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023. URL https://arxiv.org/abs/2307.15818.

Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, Deli Zhao, and Hao Chen. Worldvla: Towards autoregressive action world model, 2025. URL https://arxiv.org/abs/2506.21539.

JIAYI CHEN, Wenxuan Song, Pengxiang Ding, Ziyang Zhou, Han Zhao, Barrett Tang, Donglin Wang, and Haoang Li. Unified diffusion vla: Vision-language-action model via joint discrete denosing diffusion process. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 139291–139311, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/e187897ed7780a579a0d76fd4a35d107-Paper-Conference.pdf.

Sitan Chen, Kevin Cong, and Jerry Li. Optimal inference schedules for masked diffusion models. In Steve Hanneke and Tor Lattimore (eds.), *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 1279–1311. PMLR, 29 Jun–03 Jul 2026.

Thomas M Cover and Joy A Thomas. *Elements of information theory*, volume 2. wiley New York, 1991.

Stanislas Dehaene, Lionel Naccache, Gurvan Le Clec'H, Etienne Koechlin, Michael Mueller, Ghislaine Dehaene-Lambertz, Pierre-Francois van de Moortele, and Denis Le Bihan. Imaging unconscious semantic priming. *Nature*, 395(6702):597–600, 1998.

Jun Guo, Qiwei Li, Peiyan Li, Zilong Chen, Nan Sun, Yifei Su, Heyun Wang, Yuan Zhang, Xinghang Li, and Huaping Liu. Unified 4d world action modeling from video priors with asynchronous denoising. *arXiv preprint arXiv:2604.26694*, 2026.

Chi-Pin Huang, Yueh-Hua Wu, Min-Hung Chen, Frank Wang, and Fu-En Yang. Thinkact: Vision-language-action reasoning via reinforced visual latent planning. *Advances in Neural Information Processing Systems*, 38:82782–82802, 2026.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard (eds.), *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713. PMLR, 06–09 Nov 2025a.

Seo Hyun Kim, Sunwoo Hong, Hojung Jung, Youngrok Park, and Se-Young Yun. Klass: Kl-guided fast inference in masked diffusion models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 92267–92301. Curran Associates, Inc., 2025b. doi: 10.52202/085713-3087.

Jason Lee, Jiafei Duan, Haoquan Fang, Yuquan Deng, Shuo Liu, Boyang Li, Bohan Fang, Jieyu Zhang, Yi Ru Wang, Sangho Lee, et al. Molmoact: Action reasoning models that can reason in space. *arXiv preprint arXiv:2508.07917*, 2025.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, Yujun Shen, and Yinghao Xu. Causal world modeling for robot control, 2026. URL https://arxiv.org/abs/2601.21998.

Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, et al. Vision-language foundation models as effective robot imitators. In *International Conference on Learning Representations*, volume 2024, pp. 26703–26721, 2024.

Zhixuan Liang, Yizhuo Li, Tianshuo Yang, Chengyue Wu, Sitong Mao, Liuao Pei, Tian Nian, Shunbo Zhou, Xiaokang Yang, Jiangmiao Pang, et al. Discrete diffusion vla: Bringing discrete diffusion to action decoding in vision-language-action policies. *arXiv preprint arXiv:2508.20072*, 2025.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023.

Qi Lv, Weijie Kong, Hao Li, Jia Zeng, Zherui Qiu, Delin Qu, Haoming Song, Qizhi Chen, Xiang Deng, and Jiangmiao Pang. F1: A vision-language-action model bridging understanding and generation to actions, 2025. URL https://arxiv.org/abs/2509.06951.

Corey Lynch and Pierre Sermanet. Language conditioned imitation learning over unstructured data. 2020.

Teli Ma, Jia Zheng, Zifan Wang, Chunli Jiang, Andy Cui, Junwei Liang, and Shuo Yang. Dit4dit: Jointly modeling video dynamics and actions for generalizable robot control, 2026. URL https://arxiv.org/abs/2603.10448.

Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334, 2022. doi: 10.1109/LRA.2022.3180108.

Oier Mees, Dibya Ghosh, Karl Pertsch, Kevin Black, Homer Rich Walke, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, et al. Octo: An open-source generalist robot policy. In *First workshop on vision-language models for navigation and manipulation at ICRA 2024*, 2024.

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025.

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. 2025.

Moritz Reuss, Ömer Erdinç Yağmurlu, Fabian Wenzel, and Rudolf Lioutikov. Multimodal diffusion transformer: Learning versatile behavior from multimodal goals. 2024.

Moritz Reuss, Jyothish Pari, Pulkit Agrawal, and Rudolf Lioutikov. Efficient diffusion transformer policies with mixture of expert denoisers for multitask learning. In *International Conference on Learning Representations*, volume 2025, pp. 17247–17275, 2025.

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

Wenxuan Song, Ziyang Zhou, Han Zhao, Jiayi Chen, Pengxiang Ding, Haodong Yan, Yuxin Huang, Feilong Tang, Donglin Wang, and Haoang Li. Reconvla: Reconstructive vision-language-action model as effective robot perceiver. 40(22):18549–18557, 2026.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Ye Tian, Ling Yang, Jiongfan Yang, Anran Wang, Yu Tian, Haochen Wang, Zhiyang Teng, Zhuochen Wang, Yinjie Wang, Yunhai Tong, et al. Mmada-parallel: Multimodal large diffusion language models for thinking-aware editing and generation. In *International Conference on Learning Representations*, volume 2026, pp. 36089–36123, 2026.

Dirk Vorberg, Uwe Mattler, Armin Heinecke, Thomas Schmidt, and Jens Schwarzbach. Different time courses for visual perception and action priming. *Proceedings of the National Academy of Sciences*, 100(10):6275–6280, 2003.

Guanghan Wang, Yair Schiff, Subham Sahoo, and Volodymyr Kuleshov. Remasking discrete diffusion models with inference-time scaling. *Advances in Neural Information Processing Systems*, 38:147282–147339, 2026a.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

Yuqi Wang, Xinghang Li, Wenxuan Wang, Junbo Zhang, Yingyan Li, Yuntao Chen, Xinlong Wang, and Zhaoxiang Zhang. Unified vision-language-action model. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 80929–80944, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/82f625a28d822d2748b9f5c4f9a89bb9-Paper-Conference.pdf.

John Won, Kyungmin Lee, Huiwon Jang, Dongyoung Kim, and Jinwoo Shin. Dual-stream diffusion for world-model augmented vision-language-action model. 2025.

Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In *International Conference on Learning Representations*, volume 2024, pp. 10641–10662, 2024.

Yue Wu and Jian Huang. Stability-weighted decoding for diffusion language models, 2026. URL https://arxiv.org/abs/2604.17068.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2025, pp. 28240–28264, 2025.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. volume 38, pp. 138867–138907, 2026.

Lin Yao. Remask, don't replace: Token-to-mask refinement in diffusion large language models. *arXiv preprint arXiv:2604.18738*, 2026.

Quanfu Yu, Xian Wu, Hao Xu, and Liulong Ma. Hyworldvla: A vision-language-action model with hybrid world modeling for autonomous driving. *arXiv preprint arXiv:2607.20988*, 2026.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination?, 2026. URL https://arxiv.org/abs/2603.16666.

Yang Yue, Yulin Wang, Bingyi Kang, Yizeng Han, Shenzhi Wang, Shiji Song, Jiashi Feng, and Gao Huang. Deer-vla: Dynamic inference of multimodal large language models for efficient robot execution. volume 37, pp. 56619–56643, 2024.

Jianke Zhang, Yanjiang Guo, Yucheng Hu, Xiaoyu Chen, Xiang Zhu, and Jianyu Chen. Up-vla: A unified understanding and prediction model for embodied agent. 2025.

Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, Xinqiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, He Wang, Zhizheng Zhang, et al. Dreamvla: a vision-language-action model dreamed with comprehensive world knowledge. *Advances in Neural Information Processing Systems*, 38: 24195–24228, 2026.

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, Ankur Handa, Tsung-Yi Lin, Gordon Wetzstein, Ming-Yu Liu, and Donglai Xiang. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1702–1713, 2025. doi: 10.1109/CVPR52734.2025.00166.

Chuanxia Zheng, Tung-Long Vuong, Jianfei Cai, and Dinh Phung. Movq: Modulating quantized vectors for high-fidelity image generation. *Advances in Neural Information Processing Systems*, 35:23412–23425, 2022.

Zhide Zhong, Haodong Yan, Junfeng Li, Xiangchen Liu, Xin Gong, Tianran Zhang, Wenxuan Song, Jiayi Chen, Xinhu Zheng, Hesheng Wang, et al. Flowvla: Visual chain of thought-based motion reasoning for vision-language-action models. *IEEE Transactions on Robotics*, 2026.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *International Conference on Learning Representations*, volume 2025, pp. 6446–6469, 2025.

Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. 2025.

Appendix

Contents

A Pseudocode of MAAD-VLA 15

B Implementation Details 15

C Real-World Task Visualizations 16

D Theoretical Analysis of Asynchronous Denoising Mechanism 17

E More Analysis on Action-centric Reconsideration 19

F Visualization of Task-Dependent Denoising Dynamics 19

A PSEUDOCODE OF MAAD-VLA

Algorithm 1 summarizes the training and inference procedures of MAAD-VLA. Modality-decoupled training provides the foundation for asynchronous inference: ordered corruption exposes the model to more complete visual context alongside noisier actions, while modality-normalized losses balance their learning contributions. This prepares action prediction to exploit progressively recovered future observations. Building on this foundation, our inference framework makes the recovery process adaptive to each instance. Visual priming establishes predictive context, and the action response determines when joint denoising begins. Action-centric reconsideration then revisits earlier commitments as that context evolves. Together, training supports prediction under unequal modality corruption, while inference exploits this capability to coordinate heterogeneous vision and action modalities.

Algorithm 1 Training and Inference of MAAD-VLA

Input: model parameters θ, training dataset $\mathcal{D}$, learning rate η, batch size N_b, loss weight λ, conditioning tokens $(\mathbf{x}_{\text{lang}}, \mathbf{V})$, output lengths L_v, L_a, denoising steps T, cosine schedule g, thresholds δ_a, δ_r, remask budget K_r, start index $t_{\min}$.
// Training: modality-decoupled learning.
Sample a mini-batch of N_b examples from $\mathcal{D}$.
Independently sample $u_1, u_2 \sim \mathcal{U}(0,1)$ per example.
Set $\tau^{(v)} = \min(u_1, u_2)$ and $\tau^{(a)} = \max(u_1, u_2)$. ▷ Eq. (5)
Mask future-visual and action tokens using $g(\tau^{(v)})$ and $g(\tau^{(a)})$, respectively.
Predict both modalities from the clean prefix and masked outputs.
Compute modality-wise mean masked-token losses $\mathcal{L}_v, \mathcal{L}_a$. ▷ Eq. (7)
Compute $\mathcal{L}_{\text{train}} = \lambda\mathcal{L}_v + (1-\lambda)\mathcal{L}_a$. ▷ Eq. (8)
Update θ using η and the batch-averaged loss.

// Inference: Modality-Adaptive Asynchronous Denoising.
Initialize $\widetilde{\mathbf{V}}_0 = [\mathrm{M}]^{L_v}$, $\mathbf{A}_0 = [\mathrm{M}]^{L_a}$, and $k^\star = \perp$.
for $t = 1, \ldots, T$ **do**
  Compute $\mathbf{p}_t^{(v)}, \mathbf{p}_t^{(a)}$ in one forward pass. ▷ Eq. (2)
  Update $\widetilde{\mathbf{V}}_t$ by confidence, retaining $\lceil L_v g(1-t/T) \rceil$ masks.
  if $k^\star = \perp$ **then**
    // Stage I: Visual Priming.
    Keep $\mathbf{A}_t = \mathbf{A}_{t-1}$ fully masked.
    Compute $D_t^{(a)}$ for $t \geq 2$; set $D_1^{(a)} = 0$. ▷ Eq. (10)
    if $D_t^{(a)} > \delta_a$ **then**
      Set $k^\star = t$.
    end if
  else
    // Stage II: Asynchronous Vision–Action Joint Denoising.
    Set $\rho_t^{(a)} = g((T-t)/(T-k^\star))$.
    Update $\mathbf{A}_t$ by confidence, retaining at most $\lceil L_a \rho_t^{(a)} \rceil$ masks.
    For each newly revealed q, store $\hat{t}_q = t$ and $\mathbf{p}^{(a)}_{\hat{t}_q, q}$ until remasking.
    // Action-centric Reconsideration.
    if $t \geq k^\star + 1 + t_{\min} \ \wedge \ t < T$ **then**
      Compute $\Delta_{t,q}$ for each revealed position q in $\mathbf{A}_{t-1}$. ▷ Eq. (12)
      Select up to K_r such positions with the largest drops above δ_r.
      Remask the selected positions in $\mathbf{A}_t$.
    end if
  end if
end for
Decode $\widetilde{\mathbf{V}}_T, \mathbf{A}_T$ into the future observation and action chunk.

B IMPLEMENTATION DETAILS

We initialize MAAD-VLA from the pretrained world model of UniVLA Wang et al. (2026b) and perform post-training on downstream robotic tasks. For post-training, all experiments are conducted on 4 NVIDIA A800 GPUs with a global batch size of 64. The input observations consist of RGB and

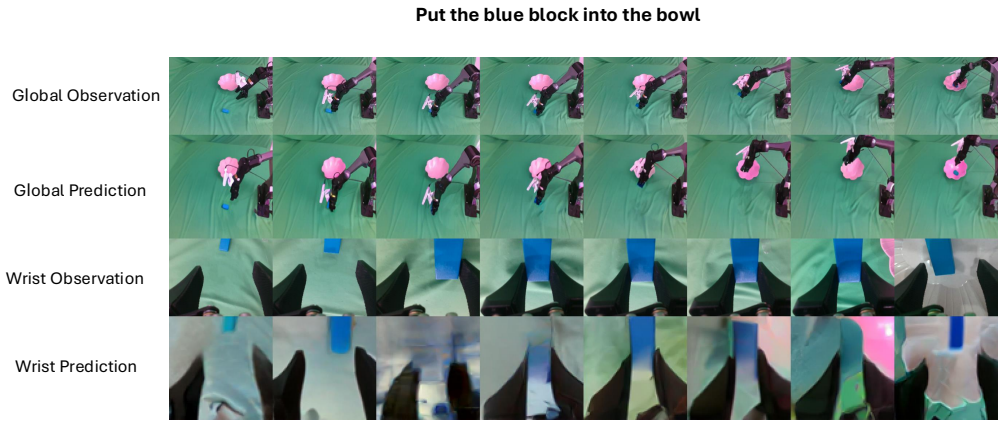


Figure 6: **Real-world task visualizations:** Put the blue block into the bowl

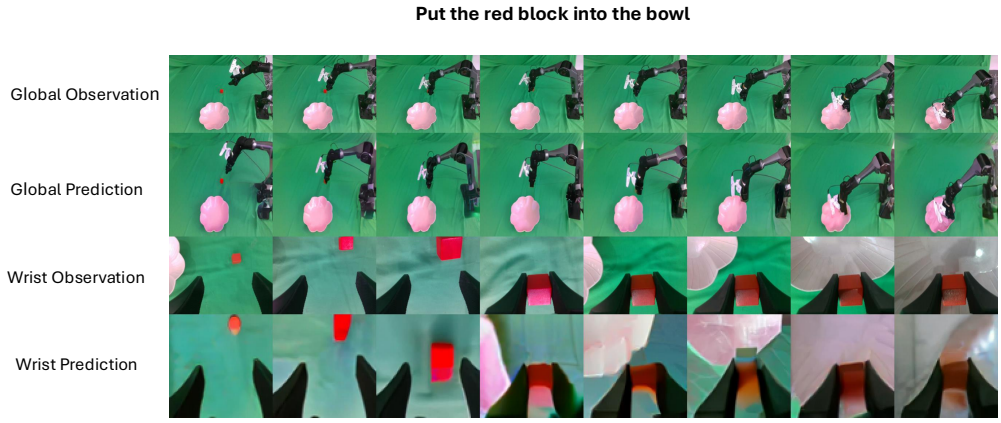


Figure 7: **Real-world task visualizations:** Put the red block into the bowl

wrist-view images with resolutions of 200×200 and 80×80, respectively. We train MAAD-VLA for 24k/30k/10k steps on CALVIN/LIBERO/real-world experiments. For LIBERO, we follow the joint training paradigm of UD-VLA. The action chunk sizes for CALVIN, LIBERO, and real-world experiments are set to 10, 10, and 20, respectively. The loss weighting coefficient λ is set to 0.5 to balance the contributions of visual and action objectives. We use the AdamW optimizer with an initial learning rate of 8×10^{-5} and cosine decay schedule. For inference, we deploy MAAD-VLA on a single NVIDIA RTX 4090 GPU with 24GB memory and use the default 72 denoising steps following UD-VLA CHEN et al. (2026) for fair comparison. The KL-response threshold δ_a is set to 0.005 based on statistics from a validation subset held out from the training data. During action-centric reconsideration, we remask up to $K_r = 2$ committed action tokens with the largest log-probability drops exceeding $\delta_r = 0.5$, starting after the first $t_{\min} = 16$ updates of Stage II. These inference hyperparameters are shared across datasets.

C Real-World Task Visualizations

Fig. 6 and Fig.7 visualize observations and predicted future images during blue- and red-block placement. The global-view predictions capture the overall scene layout and changes in robot arm pose,

while the wrist-view predictions retain the color of the target block and its approximate position relative to the gripper. Although local blur and geometric distortions remain, particularly in wrist views, the generated images preserve recognizable task-relevant content across the illustrated execution stages. These qualitative results suggest that MAAD-VLA generates visual foresight that captures both global scene structure and local manipulation context, providing sufficient task-relevant conditioning to support successful action execution in these tasks despite imperfect reconstruction. Moreover, this observation motivates future work on learning task-relevant predictive latents to condition action generation, rather than reconstructing future images in full visual detail.

D Theoretical Analysis of Asynchronous Denoising Mechanism

To gain deeper insights into the underlying mechanism of modality-adaptive asynchronous denoising, we investigate vision–action coordination from an information-theoretic perspective. During iterative denoising, progressively recovered future-visual tokens continually update the conditional distribution over actions. The key question is therefore not simply whether more visual context is available, but *when* this context becomes sufficiently action-relevant to justify committing action tokens. We formalize this through conditional mutual information (CMI) (Cover & Thomas, 1991), which connects the inter-step KL response used by our adaptive trigger to the arrival of action-relevant visual information.

Let C denote the language instruction and current visual observation, $V = (V_1, \dots, V_{L_v})$ the clean future-visual tokens, and $A = (A_1, \dots, A_{L_a})$ the clean action tokens. We analyze Stage I through an ***oracle reveal process***, in which the visual positions revealed at step k depend only on information already available:

$$U_k = f_k(C, H_{k-1}, \xi_k), \qquad \Delta V_k = (U_k, V_{U_k}), \qquad H_k = (H_{k-1}, \Delta V_k), \tag{13}$$

where $\xi = (\xi_1, \dots, \xi_K)$ is scheduler randomness with mutually independent components, jointly independent of (A, V) given C, and all visual tokens are revealed by step K. We define the filtration $\mathcal{F}_k = \sigma(C, H_k)$; during visual priming all action tokens remain masked. For an exact information-theoretic interpretation, we further assume *posterior consistency* of the action head:

$$p_{t,q}^{(a)} = P(A_q \in \cdot \mid \mathcal{F}_{t-1}), \tag{14}$$

which follows the indexing of Eq. 2: the action distribution predicted at iteration t is conditioned on the visual state obtained after iteration $t-1$. The exact identities below concern this oracle process; they serve as an idealized interpretation of the inference-time statistics, rather than a guarantee for the greedy procedure, which reveals model-predicted instead of ground-truth tokens and may reveal several tokens in parallel. For compactness, let $I_{\mathcal{F}}(X;Y)$ denote the mutual information between X and Y under the conditional law given the realized filtration $\mathcal{F}$.

Proposition 1. *Under posterior consistency, the conditional expected KL response equals the average action-relevant information contributed by the newly revealed visual evidence:*

$$\mathbb{E}\Big[D_t^{(a)} \mid \mathcal{F}_{t-2}\Big] = \frac{1}{L_a}\sum_{q=1}^{L_a} I_{\mathcal{F}_{t-2}}(A_q; \Delta V_{t-1}) \le I_{\mathcal{F}_{t-2}}(A; \Delta V_{t-1})\,. \tag{15}$$

Proof. The posterior-update identity expresses mutual information as the expected KL divergence between the updated and previous posteriors (Cover & Thomas, 1991). Applying this identity under the conditional law given the realized $\mathcal{F}_{t-2}$ to each A_q gives the equality in Eq. 15. □

Thus, $D_t^{(a)}$ is not merely a heuristic measure of prediction instability: under the oracle posterior, its conditional expectation is the per-sample rate at which newly revealed visual evidence contributes action-relevant information. This provides a principled interpretation of the transition criterion in Eq. 11.

We next quantify the residual action information that remains unavailable when an action is committed before subsequent visual evidence has arrived. Let $b_{k,q} = P(A_q \in \cdot \mid \mathcal{F}_k)$ and define the residual commitment cost after visual step s as

$$R_q(s) \triangleq \mathbb{E}\, D_{\mathrm{KL}}\left(b_{K,q} \| b_{s,q}\right) = \mathbb{E}\big[-\log b_{s,q}(A_q)\big] - \mathbb{E}\big[-\log b_{K,q}(A_q)\big], \tag{16}$$

i.e., the excess expected log-loss of predicting A_q from $b_{s,q}$ instead of the fully informed belief $b_{K,q}$.

Proposition 2. *The expected cost of committing A_q at step s equals the action information that remains to arrive through subsequent visual recovery:*

$$R_q(s) = H(A_q \mid \mathcal{F}_s) - H(A_q \mid \mathcal{F}_K) = \sum_{r=s+1}^{K} \mathbb{E}\, D_{\mathrm{KL}}\left(b_{r,q} \| b_{r-1,q}\right), \tag{17}$$

and consequently

$$\frac{1}{L_a} \sum_{q=1}^{L_a} R_q(s) = \sum_{t=s+2}^{K+1} \mathbb{E}[D_t^{(a)}]. \tag{18}$$

Proof. The posterior-update identity with $Y = \Delta V_{s+1:K}$ gives $R_q(s) = I(A_q; \Delta V_{s+1:K} \mid \mathcal{F}_s)$, which is the entropy difference; the chain rule over $\Delta V_{s+1}, \ldots, \Delta V_K$ gives the sum, and $b_{r,q} = p^{(a)}_{r+1,q}$ gives Eq. 18. □

Eq. 18 explains the role of Visual Priming: committing an action after step s incurs, in expected log-loss, exactly the action-relevant posterior update that remains to arrive; averaged over action tokens, this equals the future cumulative KL response. The KL trigger therefore withholds action commitment until the response becomes non-negligible, providing a sample-dependent onset, after which action tokens are committed progressively as the remaining visual evidence arrives.

Finite denoising budget. Proposition 2 alone would favor delaying action commitment indefinitely. Inference, however, operates under a fixed number of denoising steps. If action denoising starts later, the remaining action tokens must be recovered in fewer steps and hence in larger parallel blocks. Let B_r denote the positions committed in parallel at step r, given the previously revealed state G_{r-1} of both modalities. If positions in B_r are sampled independently from their exact conditional marginals, the expected forward-KL gap from their true joint conditional distribution to the product of these marginals equals the conditional total correlation

$$\mathrm{TC}(X_{B_r} \mid G_{r-1}) = \sum_{i \in B_r} H(X_i \mid G_{r-1}) - H(X_{B_r} \mid G_{r-1}), \tag{19}$$

which vanishes if and only if the simultaneously committed tokens are conditionally independent, consistent with recent analyses of how masked diffusion error depends on the unmasking schedule (Chen et al., 2026). For visual and action blocks B_r^v and B_r^a,

$$\mathrm{TC}(V_{B_r^v}, A_{B_r^a} \mid G_{r-1}) = \mathrm{TC}(V_{B_r^v} \mid G_{r-1}) + \mathrm{TC}(A_{B_r^a} \mid G_{r-1}) + I(V_{B_r^v}; A_{B_r^a} \mid G_{r-1}). \tag{20}$$

Hence the finite budget induces two opposing pressures. Under our hybrid attention mask, starting action denoising *too early* incurs the residual-information cost in Eq. 17. Because future-visual tokens do not attend to committed actions, later visual recovery cannot adapt to those commitments; hence the residual mismatch is not automatically repaired through the visual branch. Starting *too late* leaves fewer action updates and requires more tokens to be committed in parallel, which can increase the factorization cost in Eq. 19 when these tokens remain conditionally dependent. An intermediate priming point can therefore be preferable to either immediate joint denoising or fully recovering vision before actions.

Conclusion. Eq. 15–Eq. 18 show that, under the oracle posterior, the action KL response measures the arrival of action-relevant visual information, while the cumulative future KL exactly characterizes the excess expected log-loss of committing actions before this evidence is available. Meanwhile, under a finite denoising budget, Eq. 19 introduces the opposing cost of delaying action generation: fewer remaining steps require larger parallel commitments whose conditional dependencies can increase the factorization burden.

Consequently, MAAD-VLA's adaptive visual priming can be interpreted as balancing *premature-commitment cost* against *finite-budget parallelization cost*, rather than assuming that more visual priming is always beneficial. The KL trigger identifies sample-dependent changes in the action belief, while Action Reconsideration allows later visual evidence to revise commitments that lose support as the context evolves. Together, these mechanisms provide a unified information-theoretic rationale for coordinating when actions should begin to denoise and when earlier action decisions should be revisited, without requiring a guarantee that every posterior update necessarily translates into lower realized task error.

Table 5: **Ablation of the remasking budget.** K_r is the maximum number of action tokens remasked per eligible update. Best results are highlighted in bold.

Budget K_r	Avg. len. ↑	Gain ↑
0	4.73	–
1	4.78	+0.05
2	**4.82**	**+0.09**
3	4.79	+0.06

E More Analysis on Action-centric Reconsideration

We examine the effect of different remasking budget K_r by comparing no remasking with $K_r \in \{0, 1, 2, 3\}$. The budget limits the number of committed action tokens remasked per eligible update; fewer tokens may be selected when fewer positions satisfy the confidence-drop threshold δ_r. Table 5 reports the average successful sequence length on CALVIN.

The best results are obtained with $K_r = 2$, reaching an average length of 4.82. Increasing the budget to $K_r = 3$ reduces the average length to 4.79, indicating that a larger revision budget does not necessarily improve performance. These results support the default $K_r = 2$ reported in Appendix B among the evaluated settings.

E.1 Remasking Implementation Details

Confidence computation and selection. Action-centric reconsideration operates only on action tokens during Stage II. We compute the confidence drop in Eq. 12 using the probabilities of the same held token at its most recent commitment and at the current update. At each eligible update, we first identify committed action positions whose confidence drops strictly exceed $\delta_r = 0.5$. Among these candidates, we remask up to $K_r = 2$ positions with the largest drops. The budget is an upper bound, and no tokens are remasked if no position satisfies the threshold. Selected tokens are returned to the masked state for subsequent denoising under the updated visual context, allowing targeted revision of earlier action commitments.

EOA protection. The first committed [EOA] determines the end of the encoded action-token sequence. Once it is committed, all following positions are masked in the model input and excluded from action decoding. We therefore restrict remasking to committed action tokens preceding this delimiter, excluding [BOA], the first EOA, and its entire suffix. Protecting EOA prevents remasking from reopening the established boundary and treating the masked suffix as an unresolved action continuation. The boundary is determined from the state entering each update; if no EOA is present, no suffix restriction is applied. A subsequent reveal may introduce an earlier EOA, but an existing EOA cannot be remasked to extend the sequence.

Remasking window. Empirically, confidence estimates can fluctuate more during rapid late-stage action-token commitment. To limit late-stage perturbations, we enable remasking by default only within $k^\star + t_{\min} + 1 \le t \le T - 3$, where $k^\star$ is the final Stage-I update and T is the total number of denoising updates. The final three updates perform ordinary denoising without new remasks, consistent with the remasking-free completion phase of ReMDM-loop Wang et al. (2026a).

F Visualization of Task-Dependent Denoising Dynamics

Fig. 8 shows that both transition timing and action-response magnitude vary across task lengths on CALVIN. Specifically, longer-horizon settings (4/5 and 5/5) transition to asynchronous joint denoising later on average (27.79 and 27.54 steps) than the 1/5 setting (24.33 steps), suggesting that more challenging tasks require a stronger future-visual foundation before action denoising begins. Meanwhile, compared with simpler settings, the most challenging 5/5 setting exhibits the highest mean KL at transition (0.0156), suggesting that action predictions are less stable across consecutive denoising steps at this stage as the future-visual context evolves. These observations are consistent

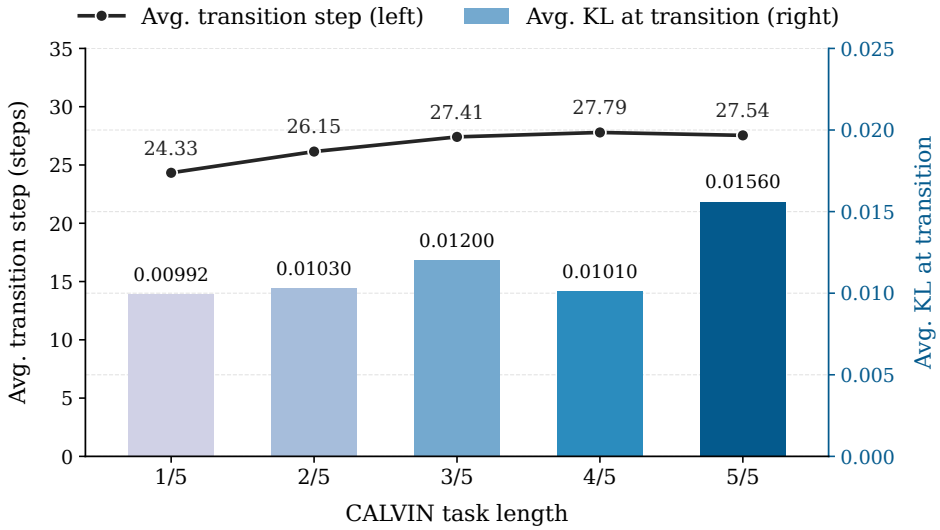


Figure 8: **Task-dependent denoising dynamics on CALVIN.** The line (left axis) shows the mean transition step from visual priming to asynchronous joint denoising, while the bars (right axis) report the mean action-response KL at transition.

with our motivation for adaptive vision–action coordination, with transition timing adapting to task demands and visual recovery dynamics rather than remaining fixed across samples.