TiRO: Fast Discrete-Diffusion VLA Inference via Teacher-Initialized Rollout

Yuze Lv[1,2] **Yao Mu**[2,†]

[1]Nanjing University [2]ScaleLab, Shanghai Jiao Tong University [†]Corresponding author

Abstract

Discrete-diffusion vision-language-action models generate action chunks through iterative token denoising, but multi-step decoding is too slow for closed-loop robot control. Speculative decoding seems attractive, yet we show that batched verification becomes expensive for long-context multimodal VLA models because each candidate must duplicate hundreds of visual, language, and state tokens. We propose **TiRO**, a **T**eacher-**i**nitialized **R**oll**o**ut distillation scheme that invokes the large teacher for only the first denoising step and then trusts a lightweight drafter to complete the remaining five steps without verification. Across four manipulation tasks, TiRO achieves a $3.39\times$ speedup over six-step MM-ACT while preserving 96.0% of its average success rate.

1 Introduction

Vision-language-action (VLA) models transfer vision-language reasoning to robot control (Zitkovich et al., 2023; Kim et al., 2024). Recent diffusion and flow-based policies generate action chunks through iterative refinement (Chi et al., 2025; Liu et al., 2025; Black et al., 2024), and discrete-diffusion VLA models further denoise action tokens within a multimodal Transformer (Liang et al., 2025b,a). While multi-step denoising improves action quality, six-step MM-ACT requires about one second to generate a single action chunk in our deployment setting.

A natural acceleration strategy is speculative decoding (Leviathan et al., 2023). A lightweight drafter proposes future states, and the large model verifies multiple candidates in one batched forward pass. This idea has also been adapted to VLA or diffusion-style inference (Wang et al., 2025; Li et al., 2025). However, the assumption that verification can be cheaply batched does not hold for long-context discrete-diffusion VLA models. Unlike decoupled continuous-action experts in diffusion or flow-based VLA models, MM-ACT denoises actions together with hundreds of visual, language, state, and action tokens. Thus, verifying multiple candidate states duplicates the full multimodal context for every candidate. The profiling inset in Figure 1 shows that short sequences partially amortize batching overhead, whereas MM-ACT-like long contexts approach linear latency growth.

We also examined two verification variants before choosing trust rollout. *Full-trajectory verification* lets the drafter predict the remaining five denoising states and verifies these candidates with the teacher in a batch. This closely follows speculative decoding, but each candidate still carries the full multimodal context, so its latency remains close to the six-step teacher. *Last-only verification* lets the drafter roll out to the final state and invokes the teacher only once more to verify or correct that state. Although this reduces the verification workload, it still requires a second teacher forward after the initial teacher step and can introduce unstable corrections on harder tasks. These observations suggest that, for discrete-diffusion VLA inference, the key question is not how to verify more candidates, but how to avoid verification after the first teacher step.

We propose **TiRO**, a teacher-conditioned distillation framework for the six-step MM-ACT action decoder. TiRO stands for **T**eacher-**i**nitialized **R**oll**o**ut: a teacher step initializes the denoising trajectory and provides action-position hidden states, after which a lightweight drafter completes the remaining steps without teacher validation. Rather than relying on speculative verification, TiRO targets the long-context discrete-diffusion VLA setting, where verification duplicates multimodal context and becomes the bottleneck. Experiments on four manipulation tasks show TiRO reduces average action-generation latency from 965 ms to 285 ms, yielding a $\mathbf{3.39}\times$ speedup, while average success drops by only 3.25 points.

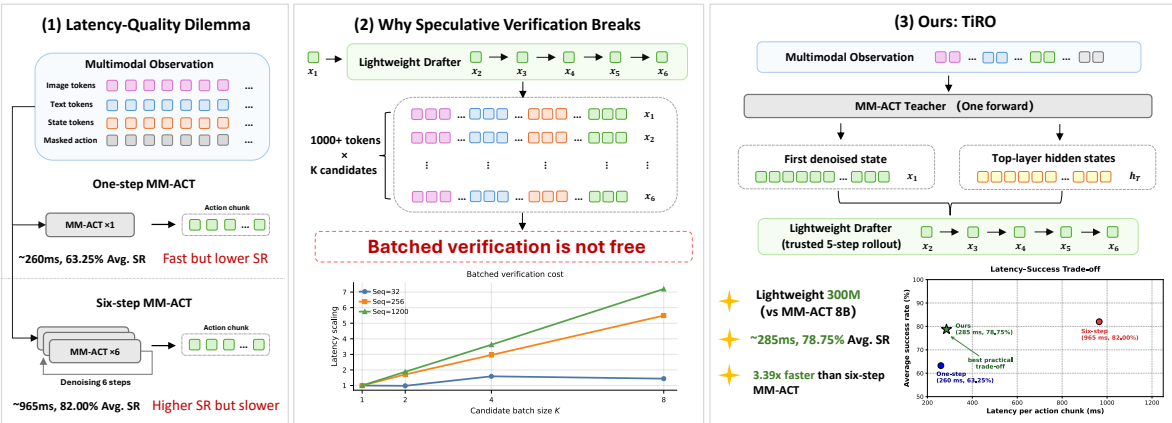

Figure 1: Latency–quality trade-off in discrete-diffusion VLA inference. One-step MM-ACT is fast but less reliable; six-step denoising improves success rate but takes nearly one second. Speculative verification duplicates the long multimodal context for every candidate, while TiRO invokes the teacher once and trusts a lightweight drafter for the remaining steps.

2 Method

As shown in Figure 2, TiRO keeps the first denoising step of the 8B MM-ACT teacher and replaces the remaining five with a lightweight $\sim$300M drafter. The drafter is not trained as a standalone policy; it learns to continue the denoising trajectory from the first state. We condition it only on action-position states, which aggregate multimodal context through self-attention while avoiding the cost and noise of feeding the full sequence. This design builds on target-model initialization for denoising-step drafting (Li et al., 2025) and hidden-state auxiliary prediction (Cai et al., 2024), but replaces speculative verification with trusted rollout to avoid duplicating long multimodal contexts.

Teacher-conditioned drafter. Let $\mathbf{c}$ denote the multimodal context and let $\mathbf{x}_0$ denote the initial all-mask action block, i.e., pure noise in the discrete action-token space. The six-step MM-ACT decoder produces a denoising trajectory $\mathbf{x}_0 \to \mathbf{x}_1 \to \cdots \to \mathbf{x}_6$. In TiRO, the 8B teacher T_θ performs only the first transition:

$$(\mathbf{x}_1, \mathbf{h}_T) = T_\theta(\mathbf{c}, \mathbf{x}_0), \tag{1}$$

where $\mathbf{x}_1$ is the remasked state after the first teacher denoising step and $\mathbf{h}_T$ is the slice of the teacher's top-layer hidden states at action-token positions. The drafter models the remaining five transitions: at step k, it takes $\mathbf{x}_k$ and predicts logits used to produce $\mathbf{x}_{k+1}$, for $k = 1, \ldots, 5$. Three token-aligned signals are fused by element-wise addition:

$$\mathbf{z}_k = E_x(\mathbf{x}_k) + W_h \mathbf{h}_T + E_t(k), \tag{2}$$

where E_x is the action-token embedding, W_h projects teacher hidden states to the drafter dimension, and $E_t(k)$ is a timestep embedding broadcast to all action-token positions. The lightweight drafter D_ϕ maps $\mathbf{z}_k$ to action-token logits ℓ_ϕ^k, from which sampled tokens and unknown-only remasking produce the next state $\mathbf{x}_{k+1}$.

Two-stage distillation. The drafter imitates the teacher's remaining five denoising transitions, with teacher targets $\mathbf{x}_2, \ldots, \mathbf{x}_6$. The two training stages use the same supervision but different input trajectories. During teacher-forcing warmup, the drafter receives teacher-generated intermediate states as inputs. During self-rollout distillation, it receives its own previous states as inputs, matching the inference path without teacher correction after $\mathbf{x}_1$. In both stages, cross-entropy (CE) and KL-based distillation are applied at every drafter step and averaged over $K = 5$. The step-wise objective is

$$\mathcal{L} = \frac{1}{K} \sum_{k=1}^{K} \Big[\alpha \, \mathrm{CE}(q_\phi^k, y_T^k) + \beta \, \tau^2 \mathrm{KL}(q_{T,\tau}^k \,\|\, q_{\phi,\tau}^k) \Big], \tag{3}$$

where $K = 5$, $\alpha = 1.0$, $\beta = 1.0$, and $\tau = 1.0$ in experiments. y_T^k denotes the teacher hard target corresponding to $\mathbf{x}_{k+1}$. It is known-preserving: unknown positions take the teacher-predicted tokens, while resolved positions keep their current tokens. We use $q_\phi^k = \mathrm{softmax}(\ell_\phi^k)$ for CE, and $q_{T,\tau}^k = \mathrm{softmax}(\ell_T^k/\tau)$, $q_{\phi,\tau}^k = \mathrm{softmax}(\ell_\phi^k/\tau)$ for KL distillation.

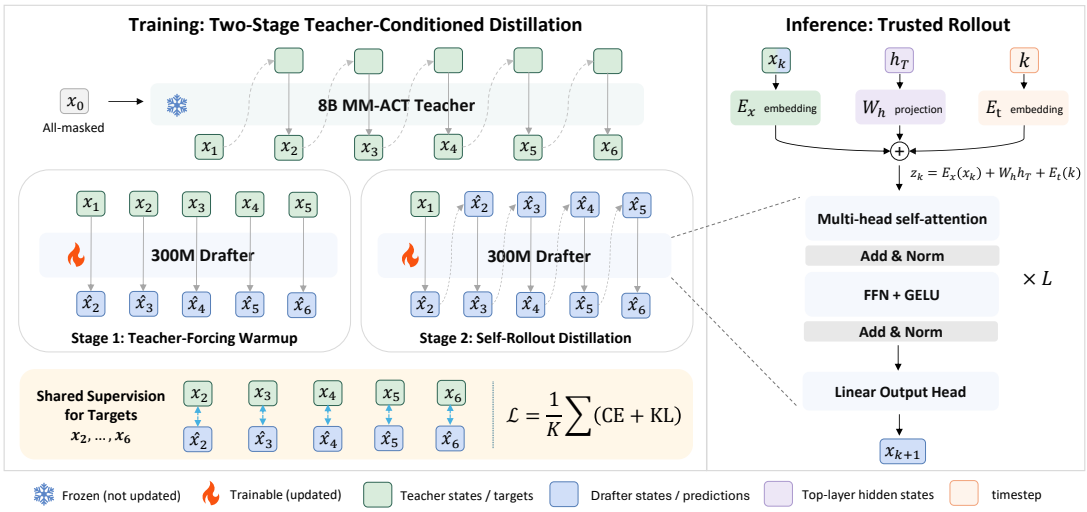

Figure 2: Training and inference pipeline of TiRO. The frozen teacher provides reference denoising states, while the trainable drafter learns them through teacher-forcing warmup and self-rollout distillation with shared CE and KL supervision. At inference, hidden-state conditioned input fusion drives the trusted drafter rollout without teacher.

Trust rollout and remasking. At inference, the teacher is invoked once for $\mathbf{x}_0 \to \mathbf{x}_1$ and to provide $\mathbf{h}_T$. The drafter performs the remaining five transitions $\mathbf{x}_1 \to \cdots \to \mathbf{x}_6$ and commits $\mathbf{x}_6$ as the final action chunk without teacher validation. Both teacher and drafter follow an unknown-only remasking constraint: resolved tokens are never remasked. They differ only in how unknown positions are selected. The teacher uses confidence-based remasking, while the drafter samples positions uniformly within the unknown set under the same cosine mask ratio. We use this uniform unknown-only selector as a stability-oriented engineering choice, since early drafter confidence can be poorly calibrated during self-rollout.

3 Experiments

We evaluate on four manipulation tasks from the RoboTwin2.0 benchmark (Chen et al., 2025): Adjust Bottle, Beat Hammer Block, Click Bell, and Place Burger Fries. We report task success rate and average wall-clock latency for generating one action chunk. We compare against one-step MM-ACT with chunk sizes 8 and 16, and the original six-step MM-ACT with chunk size 16 (Liang et al., 2025a). Our method uses chunk size 16 and the same evaluation protocol.

Table 1 shows that TiRO achieves the best practical latency–success trade-off. Compared with six-step MM-ACT, it reduces average latency from 965 ms to 285 ms, a 3.39× speedup. The average success rate decreases from 82.00% to 78.75%, preserving 96.0% of the six-step baseline. The one-step baselines reach similar latency but much lower success, confirming that simply reducing denoising steps is insufficient. Beat Hammer Block shows the largest drop, suggesting that some tasks may still benefit from selective teacher fallback.

The two one-step baselines represent the low-latency regime, while six-step MM-ACT represents the high-success but high-latency regime. Ours occupies the middle regime: it retains the first teacher step for global multimodal grounding and uses the drafter only for local denoising continuation.

4 Related Work

VLA and action diffusion. VLA models connect vision-language reasoning with robot control (Zitkovich et al., 2023; Kim et al., 2024). Continuous-action policies such as Diffusion Policy, RDT, and π_0 generate action chunks through diffusion or flow-matching experts (Chi et al., 2025; Liu et al., 2025; Black et al., 2024). Discrete-diffusion VLA models instead denoise discrete action tokens through a shared multimodal backbone (Liang et al., 2025b,a). Our work studies how to reduce the inference latency of this discrete-diffusion VLA regime.

Speculative acceleration. Speculative decoding accelerates language models by drafting future tokens and verifying them with the target (Leviathan et al., 2023). Medusa shows target-model hidden states can support auxiliary prediction (Cai et al., 2024). Spec-VLA adapts speculative decoding to

Method	Adjust	Hammer	Bell	Burger	Avg. SR ↑	ΔSR ↑	Time ↓	Speedup ↑
One-step MM-ACT, chunk size 8	61%	66%	91%	53%	67.75%	−14.25	∼258 ms	3.74×
One-step MM-ACT, chunk size 16	62%	47%	90%	54%	63.25%	−18.75	∼260 ms	3.71×
Six-step MM-ACT, chunk size 16	88%	**67%**	**93%**	**80%**	**82.00%**	0.00	∼965 ms	1.00×
Ours: TiRO, chunk size 16	**90%**	60%	90%	77%	78.75%	**−3.25**	∼285 ms	**3.39×**

Table 1: Main results over 100 evaluation episodes per task. Latency is measured as action-chunk generation time on an A800 80GB GPU. **Bold** values highlight the best task scores and the most favorable latency–success trade-off.

VLA with relaxed physical acceptance (Wang et al., 2025), and TS-DP applies speculative verification to diffusion policy acceleration (Li et al., 2025). In contrast, we target long-context discrete-diffusion VLA models and replace teacher validation with trust distillation.

5 Limitations and Future Work

As summarized in Table 1, TiRO trades a 3.25-point success-rate drop for a 3.39× speedup, but does not preserve the teacher's exact output distribution. After the first step, the drafter controls the rollout, which can fail if small token errors accumulate into large physical deviations. Future work could combine trust rollout with adaptive fallback based on drafter entropy or safety constraints. Broader manipulation tasks, fine-grained hyperparameter ablations, and validation on other discrete-diffusion VLA architectures are important next steps. Finally, reducing repeated multimodal computation via approximate KV caching or context compression is left for future work.

References

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, and 1 others. 2024. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*.

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D Lee, Deming Chen, and Tri Dao. 2024. Medusa: Simple llm inference acceleration framework with multiple decoding heads. *arXiv preprint arXiv:2401.10774*.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, and 1 others. 2025. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. 2025. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, and 1 others. 2024. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*, pages 19274–19286. PMLR.

Ye Li, Jiahe Feng, Yuan Meng, Kangye Ji, Chen Tang, Xinwan Wen, Shutao Xia, Zhi Wang, and Wenwu Zhu. 2025. Ts-dp: Reinforcement speculative decoding for temporal adaptive diffusion policy acceleration. *arXiv preprint arXiv:2512.15773*.

Haotian Liang, Xinyi Chen, Bin Wang, Mingkang Chen, Yitian Liu, Yuhao Zhang, Zanxin Chen, Tianshuo Yang, Yilun Chen, Jiangmiao Pang, and 1 others. 2025a. Mm-act: Learn from multimodal parallel generation to act. *arXiv preprint arXiv:2512.00975*.

Zhixuan Liang, Yizhuo Li, Tianshuo Yang, Chengyue Wu, Sitong Mao, Tian Nian, Liuao Pei, Shunbo Zhou, Xiaokang Yang, Jiangmiao Pang, and 1 others. 2025b. Discrete diffusion vla: Bringing discrete diffusion to action decoding in vision-language-action policies. *arXiv preprint arXiv:2508.20072*.

Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. 2025. Rdt-1b: a diffusion foundation model for bimanual manipulation. In *International Conference on Learning Representations*, volume 2025, pages 29982–30009.

Songsheng Wang, Rucheng Yu, Zhihang Yuan, Chao Yu, Feng Gao, Yu Wang, and Derek F Wong. 2025. Spec-vla: speculative decoding for vision-language-action models with relaxed acceptance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26916–26928.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, and 1 others. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR.