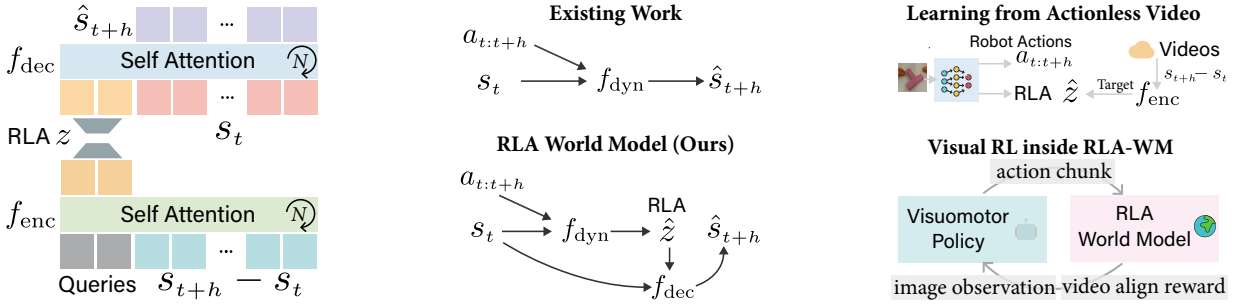
Learning Visual Feature-Based World Models via Residual Latent Action

Xinyu Zhang[1] Zhengtong Xu[2] Yutian Tao[3] Yeping Wang[3] Yu She[2] Abdeslam Boularias[1]
[1]Rutgers University [2]Purdue University [3]University of Wisconsin–Madison



(a) RLA Autoencoder Learning (b) Dynamics Learning (c) Applications

Figure 1. **Overview of our framework.** We introduce the Residual Latent Action (RLA), which compresses the DINO token residual $s_{t+h} - s_t$ into a compact latent z. We discover that RLA is predictive, generalizable, and encodes temporal progression. Next, we propose the RLA World Model (RLA-WM), which learns from offline videos and predicts RLA z instead of s_{t+h} directly. RLA-WM achieves accurate future prediction while being more efficient than state-of-the-art feature-based and video diffusion world models. Our approach enables two applications: learning policies from actionless videos and visual reinforcement learning via interaction entirely within RLA-WM. Extensive visualizations, source code, and a video overview are available on our project page: https://mlzxy.github.io/rla-wm.

Abstract

Visual feature-based world models predict future visual features rather than raw pixels, offering an efficient alternative to video diffusion. Yet existing feature-based approaches either rely on direct regression, which collapses on complex 3D interactions, while generative modeling in high-dimensional feature spaces still remains challenging. We introduce Residual Latent Action *(RLA), a simple representation distilled from DINO token residuals that is predictive, generalizable, and encodes temporal progression. RLA enables* RLA World Model *(RLA-WM), which predicts RLA values via flow matching. RLA-WM outperforms both state-of-the-art feature-based and video-diffusion world models on simulation and real-world datasets, while being orders of magnitude faster than video diffusion. We further demonstrate two applications: (1) a minimalist world action model with RLA that learns from actionless video; (2) the first visual RL framework trained entirely inside a world model learned from offline videos only, using a video-aligned reward and no online interactions or handcrafted rewards.*

1. Introduction

World models have recently received increasing research attention due to their great potential for policy learning and reasoning through future state prediction [8]. Currently, the predominant paradigm in world modeling relies on video generation, by predicting future trajectories in pixel-aligned VAE latent spaces [4, 12, 21, 23–25]. While visually compelling, this approach is prone to hallucination [11] and suffers from a heavy computational overhead [29]. As a result, downstream applications of world models remain largely constrained to open-loop robot data generation [20, 22], policy pretraining [14, 15], and planning for specific tasks [5, 9, 30].

Visual feature-based world models predict features of future frames, such as DINO tokens, rather than just videos [1]. DINO-WM [2, 6, 30] shows that direct regression of future DINO tokens leads to efficient and accurate world models for 2D manipulation tasks. However, despite these advantages, feature-based world models remain far less adopted, as predictions often become blurry or even collapse in complex 3D interactions [22]. A seemingly straightforward solution is to use generative models in feature space. However, feature-space generation is even more difficult than in pixel space due to the higher dimensionality [26, 28]. More importantly, heavy generative pipelines undermine the very advantages that feature-based models should provide.

Our work answers two key questions: (1) how to develop an efficient yet accurate world model in a visual feature space that scales to complex 3D manipulation? (2) how to leverage

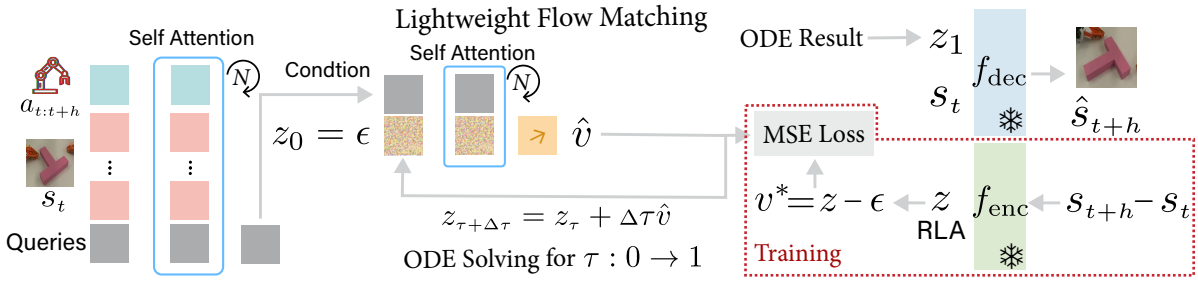


Figure 2. **RLA World Model.** RLA-WM predicts future states via residual latent action z. Robot actions $a_{t:t+h}$ are embedded (MLP), concatenated with DINO tokens s_t and learnable queries, then self-attention yields condition tokens. In flow matching, this condition is fixed and concatenated with noisy z_τ (starting from Gaussian noise $z_0 = \epsilon$). The flow network predicts velocity $\hat{v}$ to iteratively transform z_0 to final z_1. Decoder f_{dec} outputs $\hat{s}_{t+h}$ from z_1 and s_t. The model is trained with MSE loss against ground truth velocity $v^* = z - \epsilon$.

such world models to improve downstream policies?

2. Proposed Approach and Results

Problem Formulation. Let $x_t \in \mathbb{R}^{H\times W\times 3}$ denote the raw image observation at time t. We represent the DINO patch tokens as $s_t \in \mathbb{R}^{L\times C}$. C is the feature dimension and $L = \frac{H\times W}{P^2}$ is the sequence length for a given patch size P. We define an action chunk of horizon h as $a_{t:t+h}$. Our objective is to learn a dynamics function $f_{\text{dyn}} : (s_t, a_{t:t+h}) \mapsto s_{t+h}$ using only raw offline videos. This function acts as a direct, multi-step world model in the feature space.

Residual Latent Action. Instead of directly generating $\hat{s}_{t+h}$, we propose to learn a representation that captures the transition from s_t to s_{t+h} from DINO token residuals $s_{t+h} - s_t$, which also corresponds to the flow matching velocity of a Schrödinger bridge [7, 18] from s_t to s_{t+h}. Specifically, we feed these residuals along with learnable queries into an encoder f_{enc}, project the output queries to a low-dimensional space to obtain z, and pass z along with s_t into a decoder f_{dec} to reconstruct s_{t+h} (Fig. 1a). We refer to z as a *Residual Latent Action* (RLA). The RLA autoencoder uses almost only self-attention and a single regression loss on s_{t+h}. Three key properties of RLA set it apart from prior work, making it an ideal representation for dynamics modeling: (1) *Predictive Sufficiency*; Unlike in prior work, where latent actions serve only as weak conditioning for diffusion, RLA does not require iterative generation. We find that our RLA decoder f_{dec}, when conditioned on a compact RLA z, is able to reconstruct future DINO tokens with high fidelity in a single feedforward pass. (2) *Generalizability*; RLA autoencoder generalizes to novel scenes and interactions. (3) *Temporal Topology*; An emergent property of RLA is the topology of its learned latent space. Although the autoencoder is only trained on frame pairs (s_t, s_{t+h}), the RLA latent space naturally encodes temporal progression. Interpolating between a Gaussian noise and RLA produces frames that correspond to temporally intermediate states.

RLA World Model. Learning neural dynamics in RLA space encourages the model to capture state evolution rather than absolute states. Motivated by this, instead of generating high-dimensional s_{t+h}, we propose a world model to predict the compact RLA z, which is then decoded with current state s_t to reconstruct s_{t+h}. Specifically, learnable queries are concatenated with s_t and embedded actions $a_{t:t+h}$ and transformed via self-attention. These query tokens are then concatenated with a noisy RLA $z_\tau = \tau z + (1-\tau)\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$, through subsequent self-attention layers to predict the velocity $\hat{v}$. We supervise with ground truth velocity $v^* = z - \epsilon$, where $z = f_{\text{enc}}(s_{t+h} - s_t)$. At inference, we sample $z_0 = \epsilon$ and solve the ODE with $z_{\tau+\Delta\tau} = z_\tau + \Delta\tau\hat{v}$ from $\tau = 0$ to 1. The final z_1 is the predicted RLA, and is decoded via f_{dec} with s_t. Since condition network is executed once and iterative generation remains within the compact RLA space, the flow matching is lightweight, as shown by the floating point operations (FLOPs) reported in Tab. 1.

Table 1. **World Model Quality.** We measure the quality of generated future frames. Best **bold**, 2nd underlined.

	ManiSkill [16]		IWS [22]		FLOPs
Model	LPIPS↓	DINO-L1↓	LPIPS↓	DINO-L1↓	↓
DINO-WM [30]	0.156	0.078	0.223	0.058	**2.1T**
RAE [26]	0.324	0.143	0.550	0.159	14.3T
FM-WM [13]	0.127	0.063	0.360	0.119	14.3T
Vid2World [12]	0.199	0.084	0.388	0.139	1.1P
RLA-WM (Ours)	**0.071**	**0.030**	**0.196**	**0.053**	3.5T

Table 2. **(a) Learning from Actionless Videos**: average SR (%) over 5 ManiSkill tasks. **(b) WMRL**: 1500-seed evaluation (SR averaged for each robot).

(a) Learning from Actionless Videos

Method	Avg SR↑
BC-ResNet	27.2
DINO CLS [19]	27.4
UniVLA [3]	28.7
AdaWorld [10]	33.7
RLA (Ours)	**35.6**

(b) World Model-based RL (WMRL)

Method	XArm	UR10e	Panda	Avg
BC-ResNet	89.9	41.4	**62.8**	59.6
RLA-WMRL	**95.9**	**46.9**	57.0	**60.7**

Results (project page). On ManiSkill [16] and IWS [22], RLA-WM beats DINO-WM [30], RAE [26], FM-WM [13], and Vid2World [12] on LPIPS/SSIM/DINO-L1 at ∼300× fewer FLOPs (Tab. 1). A BC-ResNet with one RLA head trained on 5–15% action-labeled plus actionless videos beats DINO-CLS [27], UniVLA [3], and AdaWorld [10] (Tab. 2a). PPO [17] inside RLA-WM with a video-aligned reward yields +1.1% over BC under 1500 seeds (Tab. 2b).

References

[1] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

[2] Federico Baldassarre, Marc Szafraniec, Basile Terver, Vasil Khalidov, Francisco Massa, Yann LeCun, Patrick Labatut, Maximilian Seitzer, and Piotr Bojanowski. Back to the features: Dino as a foundation for video world models. *arXiv preprint arXiv:2507.19468*, 2025.

[3] Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025.

[4] Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, et al. Worldvla: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*, 2025.

[5] Delong Chen, Theo Moutakanni, Willy Chung, Yejin Bang, Ziwei Ji, Allen Bolourchi, and Pascale Fung. Planning with reasoning using vision language world model. *arXiv preprint arXiv:2509.02722*, 2025.

[6] Junha Chun, Youngjoon Jeong, and Taesup Kim. Sparse imagination for efficient visual world model planning. *arXiv preprint arXiv:2506.01392*, 2025.

[7] Valentin De Bortoli, Iryna Korshunova, Andriy Mnih, and Arnaud Doucet. Schrodinger bridge flow for unpaired data translation. *Advances in Neural Information Processing Systems*, 37:103384–103441, 2024.

[8] Jiahua Dong, Qi Lyu, Baichen Liu, Xudong Wang, Wenqi Liang, Duzhen Zhang, Jiahang Tu, Hongliu Li, Hanbin Zhao, Henghui Ding, et al. Learning to model the world: A survey of world models in artificial intelligence. 2026.

[9] Yilun Du, Mengjiao Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B Tenenbaum, et al. Video language planning. In *International Conference on Learning Representations (ICLR)*, 2024.

[10] Shenyuan Gao, Siyuan Zhou, Yilun Du, Jun Zhang, and Chuang Gan. Adaworld: Learning adaptable world models with latent actions. In *International Conference on Machine Learning (ICML)*, 2025.

[11] Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Yufan Deng, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, et al. Video-bench: Human-aligned video generation benchmark. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18858–18868, 2025.

[12] Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2world: Crafting video diffusion models to interactive world models. *arXiv preprint arXiv:2505.14357*, 2025.

[13] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*.

[14] Guanxing Lu, Baoxiong Jia, Puhao Li, Yixin Chen, Ziwei Wang, Yansong Tang, and Siyuan Huang. Gwm: Towards scalable gaussian world models for robotic manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9263–9274, 2025.

[15] Russell Mendonca, Shikhar Bahl, and Deepak Pathak. Structured world models from human videos. In *Robotics: Science and Systems (RSS)*, 2023.

[16] Tongzhou Mu, Zhan Ling, Fanbo Xiang, Derek Cathera Yang, Xuanlin Li, Stone Tao, Zhiao Huang, Zhiwei Jia, and Hao Su. Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[17] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[18] Yuyang Shi, Valentin De Bortoli, Andrew Campbell, and Arnaud Doucet. Diffusion schrödinger bridge matching. *Advances in neural information processing systems*, 36:62183–62223, 2023.

[19] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[20] GigaBrain Team, Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Haoyun Li, Jie Li, Jiagang Zhu, Lv Feng, et al. Gigabrain-0: A world model-powered vision-language-action model. *arXiv preprint arXiv:2510.19430*, 2025.

[21] Yuang Wang, Chao Wen, Haoyu Guo, Sida Peng, Minghan Qin, Hujun Bao, Xiaowei Zhou, and Ruizhen Hu. Precise action-to-video generation through visual action prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12713–12724, 2025.

[22] Yixuan Wang, Rhythm Syed, Fangyu Wu, Mengchao Zhang, Aykut Onol, Jose Barreiros, Hooshang Nayyeri, Tony Dear, Huan Zhang, and Yunzhu Li. Interactive world simulator for robot policy training and evaluation. *arXiv preprint arXiv:2603.08546*, 2026.

[23] Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. *arXiv preprint arXiv:2507.07982*, 2025.

[24] Jialong Wu, Shaofeng Yin, Ningya Feng, and Mingsheng Long. Rlvr-world: Training world models with reinforcement learning. *arXiv preprint arXiv:2505.13934*, 2025.

[25] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *International Conference on Learning Representations (ICLR)*, 2024.

[26] Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

[27] Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loic Magne, et al. Flare: Robot learning with implicit world modeling. *arXiv preprint arXiv:2505.15659*, 2025.

[28] Zhenxin Zheng and Zhenjie Zheng. Rethinking diffusion model in high dimension. *arXiv preprint arXiv:2503.08643*, 2025.

[29] Zangwei Zheng, Xiangyu Peng, Yuxuan Lou, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, et al. Open-sora 2.0: Training a commercial-level video generation model in $200 k. *arXiv preprint arXiv:2503.09642*, 2025.

[30] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. In *International Conference on Machine Learning (ICML)*, 2025.