GigaWorld-Policy: An Efficient Action-Centered World–Action Model

Chaojun Ni[1,2*], Xinyu Zhou[2,3*], Yukun Zhou[2*], Jingyu Liu[2], Xiaofeng Wang[2], Zheng Zhu[2†], Yang Wang[2], Qiuping Deng[2], Yun Ye[2], Hao Li[2], Zhichao Liu[2], Jindi Lv[2], Boyuan Wang[2], Guosheng Zhao[2], Guan Huang[2], Min Cao[3], and Wenjun Mei[1†]

[1]Peking University, Beijing, China [2]GigaAI, Beijing, China
[3]School of Computer Science and Technology, Soochow University, Suzhou, China
[*]These authors contributed equally to this work.
[†] Corresponding authors: zhengzhu@ieee.org, mei@pku.edu.cn.
Project Page: https://gigaai-research.github.io/GigaWorld-Policy/

Abstract. World–Action Models (WAMs) initialized from pre-trained video generation backbones have demonstrated remarkable potential for robot policy learning. However, existing approaches face two critical bottlenecks: joint reasoning over future visual dynamics and actions incurs substantial inference overhead, and joint modeling entangles visual and motion representations, making motion prediction dependent on video forecasts. To address these issues, we introduce GigaWorld-Policy, an action-centered WAM that learns 2D pixel–action dynamics while enabling efficient action decoding with optional video generation. Specifically, the model predicts future action sequences conditioned on the current observation and, during training, generates future videos conditioned on the predicted actions and the same observation. Supervision from both action prediction and video generation provides richer learning signals and encourages physically plausible actions through visual-dynamics constraints. With a causal design that prevents future-video tokens from influencing action tokens, explicit future-video generation is optional at inference time, enabling faster deployment. To support this paradigm, we curate a diverse, large-scale robot dataset to pre-train the model. Experiments on real-world robotic platforms show that GigaWorld-Policy runs 9× faster than the WAM baseline Motus while improving task success rates by 7%; compared with $\pi_{0.5}$, it improves performance by 95% on RoboTwin 2.0.

Keywords: Video Generation· World–Action Models · Robot Policy Learning · Large-Scale Robot Dataset

1 INTRODUCTION

Vision–Language–Action (VLA) models [3, 7, 23, 24, 39, 51, 65–67, 78] based on Vision–Language Models (VLMs) [1, 34, 38, 48, 60, 66, 69, 70, 79, 80] have achieved

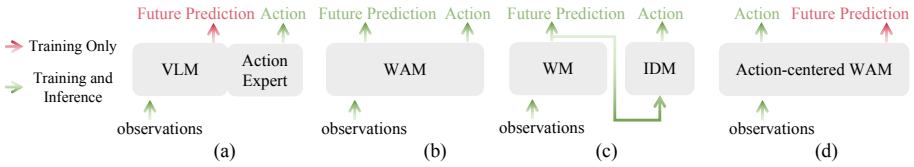

Fig. 1: (a) VLM-based VLA with auxiliary future supervision to densify training signals, but limited by the discriminative nature of VLMs [5,39,71]. (b) Joint action–video prediction with bidirectional attention [2, 26, 47], requiring future video generation at inference. (c) Two-stage design [30, 45] that first generates future video and then gets actions via an Inverse Dynamics Model (IDM), inheriting video prediction errors and incurring additional inference cost. (d) GigaWorld-Policy: an action-centered WAM that leverages future visual dynamics as supervision during training, while making future-video prediction optional at inference for low-latency action generation.

strong performance in robotic manipulation and control [9, 36, 63, 75, 76]. However, a major challenge remains: supervision sparsity. While observations and task conditioning are high-dimensional and semantically rich, action supervision is sparse and low-diversity. As a result, models may rely on contextual shortcuts, collapsing many situations into a small set of action prototypes instead of learning a physically consistent action distribution.

Therefore, some works [5, 6, 39, 71] attempt to inject future-state supervision into existing VLA frameworks by predicting future visual observations, as illustrated in Fig. 1 (a). However, VLM-based VLA [13–15, 29, 50, 53, 59] models are typically optimized for discriminative reasoning rather than high-fidelity generation, making it non-trivial for these additional losses to enforce continuity and physical consistency in the predicted actions.

bestIn contrast, recent efforts incorporate the World Model (WM) from video generation [35, 46, 57, 68] into robot policy learning [2, 26, 30, 47] to further increase supervision density and improve scalability. Leveraging video generation is appealing because it provides temporally dense supervision in the observation space beyond sparse action labels and injects strong spatiotemporal priors learned from large-scale video data. These methods commonly optimize joint objectives of future visual dynamics and action prediction, explicitly coupling future observation forecasting with action selection, thereby leveraging the representational and generative capacity of video models to guide action learning (Fig. 1 (b–c)). However, these approaches often require iterative sampling to roll out future videos at inference time, leading to high latency. Moreover, errors in video prediction can propagate to action decoding, causing mistakes and degraded long-horizon control, particularly when small early inaccuracies compound over time.

To address these limitations, we introduce GigaWorld-Policy, an action-centered and efficient World–Action model. Instead of making action prediction overly reliant on explicit video generation, GigaWorld-Policy leverages future visual dynamics as a reasoning signal and a source of dense supervision. Specifically,

GigaWorld-Policy is implemented as a causal sequence model that represents action tokens and future-visual tokens under a causal mask. During training, the model learns to predict future action sequences from the current observation context, and in parallel learns an action-conditioned visual dynamics model that forecasts future visual observations given the same current observation together with the predicted actions, thereby coupling action learning with explicit 2D pixel-level state evolution. These two learning signals are optimized within the same model, allowing future visual dynamics to regularize action plausibility and provide substantially denser supervision, which improves learning efficiency. Crucially, at inference time, explicit future-video prediction is optional: the model can be executed in an action-only mode that directly produces control commands without rolling out long sequences of video tokens. This design substantially reduces compute and memory overhead, avoids compounding errors from extended visual rollouts, and enables low-latency closed-loop control, as illustrated in Fig. 1 (d). To obtain stronger pre-trained weights, we use a curriculum training pipeline that injects physics priors from diverse video sources before any task-specific supervision. GigaWorld-Policy is initialized from a large-scale web-video foundation model [54], and then further pre-trained on embodied, robot-centric data that combines real robot recordings with large-scale egocentric human videos, improving robustness to embodiment-specific viewpoints and interaction dynamics. Finally, we post-train the model on target-robot trajectories that align images, language, and actions, specializing it for instruction-conditioned action prediction under the target robot's control interface and state distribution.

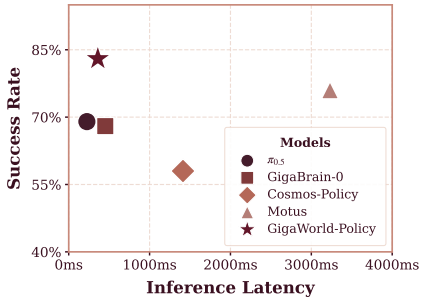

Fig. 2: Comparison of GigaWorld-Policy with baselines on inference frequency and task success rate in real-world settings and on an A100 GPU.

We validate GigaWorld-Policy through experiments in both simulation and real-world environments. GigaWorld-Policy outperforms strong baselines, delivering efficiency gains without sacrificing control performance. As shown in Fig. 2, it achieves a trade-off between success and efficiency, delivering 9× faster inference and 7% higher task success rates than the state-of-the-art WAM Motus [2]. Moreover, under comparable speed and VLA settings, GigaWorld-Policy improves performance by 20%.

The main contributions of this paper are summarized as follows:

– We propose GigaWorld-Policy, an action-centered and efficient World–Action model. During training, future visual dynamics provide dense supervision and a reasoning signal for action learning, without over-reliance on explicit video

synthesis. At inference time, future-visual prediction is optional and we can decode actions only, enabling low-latency control.
- We propose a pre-training paradigm that converts a generic video generation model into a strong initialization for robot policy learning, fully leveraging complementary data sources across stages.
- Experiments on real robotic platforms show that GigaWorld-Policy achieves a 9× inference speedup (down to 0.36 s per inference) while improving task success rates by 7% over baseline methods; compared with $\pi_{0.5}$, GigaWorld-Policy matches the inference speed and improves performance by 95% on RoboTwin 2.0.

2 Related Work

2.1 World Models for Robotic Video Generation

Recent advances in world models [28, 42, 43, 56, 57, 62, 72, 73] have improved advanced robotic video generation and prediction, driven by progress in video generation, representation learning, and temporal modeling [12,16,18,22,31–33, 36,40,55,62,74,81]. The central goal is to learn a generative model that captures the temporal evolution of the environment, enabling the prediction of future visual sequences.

Pandora [64] proposes a hybrid autoregressive–diffusion world model that generates videos while enabling real-time control via free-text actions. FreeAction [27] explicitly exploits continuous action parameters in diffusion-based robot video generation by using action-scaled classifier-free guidance to better control motion intensity. GigaWorld-0-video [52] is a high-fidelity world-model data engine that synthesizes temporally coherent, high-quality 2D video sequences with fine-grained control over appearance and scene layout. Some methods [10, 11] also explore explicit video world models, which aim to construct structured and manipulable 3D scene representations [41,61]. Aether [49] unifies geometry-aware world modeling by jointly optimizing 4D dynamic reconstruction, action-conditioned video prediction, and goal-conditioned visual planning.

However, most existing efforts improve the fidelity, consistency, and controllability of video world models, while largely overlooking how to adapt generic video generators into action-centered models that directly support policy learning under tight latency constraints. In contrast, we treat the video generator as policy initialization and propose an action-centered training recipe that aligns the backbone with robotic observations and action conditioned dynamics.

2.2 World–Action Models for Robotic Control

World–Action Models (WAM) [2, 26, 47, 58], grounded in the video generation paradigm, aim to predict robot actions and future visual dynamics within a unified framework. By modeling action-conditioned future observations, WAMs

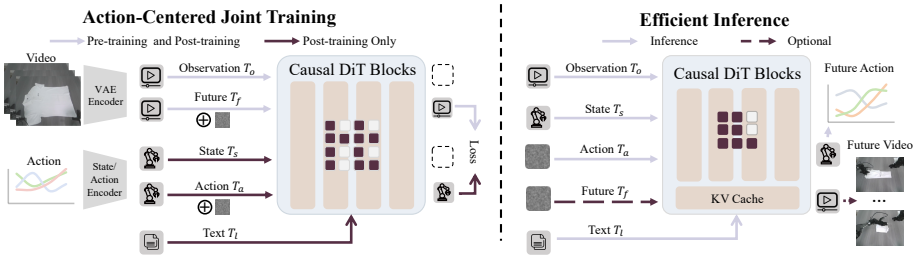

Fig. 3: Overview of GigaWorld-Policy, built on a pre-trained video generation backbone. During pre-training, the model learns action-relevant representations from large-scale videos. During post-training for policy learning, it jointly performs action-chunk prediction from the current observation and future video prediction as auxiliary supervision. At inference time, the future-video prediction branch is optional, enabling faster inference.

provide dense temporal supervision and a learned predictive prior that regularizes policy learning.

As shown in Fig. 1 (b), VideoVLA [47] directly utilizes video generation models as pre-trained weights to explore the transformation of large-scale video generation models into robotic manipulation models, employing multimodal diffusion Transformers to jointly model video, language, and action modalities, achieving dual prediction of actions and future visual outcomes. Motus [2] proposes a unified world model that leverages existing general pre-trained models and rich, shareable motion information, introducing a Mixture-of-Transformer (MoT) architecture to integrate three expert modules and adopting a UniDiffuser-style scheduler to enable flexible switching between different modeling modes. In contrast, as shown in Fig. 1 (c), Mimic-video [45] adopts a two-stage pipeline: it first leverages an Internet-scale pre-trained video backbone to predict future visual observations, and then uses a flow-matching inverse-dynamics action decoder to map the resulting video latents into low-level robot actions.

However, these methods all typically require iterative diffusion sampling to generate future videos during inference, which introduces significant latency and limits real-time deployment. Meanwhile, an over-reliance on explicit video prediction can be fragile: pixel-level forecasting is sensitive to stochasticity observability, and small visual prediction errors may compound over long horizons, weakening the usefulness of the learned dynamics for robust action generation.

3 Method

3.1 Problem Statement and Approach Overview

We formulate robotic manipulation as a sequential decision-making task. At each time step t, the robot receives multi-view RGB observations $o_t = \{o_t^v\}_{v \in S}$ from a fixed set of camera viewpoints $S = \{left, front, right\}$, a natural-language

instruction l, and the proprioceptive state s_t. Conditioned on these inputs, the policy predicts an action chunk of length p, $a_{t:t+p-1} = (a_t, a_{t+1}, \ldots, a_{t+p-1})$.

Vision-Language-Action Policies. Most existing VLA policies are trained via imitation learning to model and sample an action chunk conditioned on the observation, the robot state, and a language instruction:

$$a_{t:t+p-1} \sim q_\Theta(\cdot \mid o_t,\, s_t,\, l\,). \tag{1}$$

The distribution $q_\Theta(\cdot \mid o_t, s_t, l)$ parameterizes the policy over the next p actions. In this paradigm, learning is driven solely by action supervision from demonstrations, without any explicit supervision in the observation space.

Our Approach. Unlike approaches that only model an action distribution, we adopt a world-modeling perspective that learns how visual observations evolve under an executed action chunk. We implement our method as a single unified model g_Θ that parameterizes two complementary conditional distributions. For action modeling, anchored by demonstrations, the model learns to sample an action chunk conditioned on the observation, robot state, and language:

$$\big(a_{t:t+p-1},\, c_t\big) \sim g_\Theta(\cdot \mid o_t,\, s_t,\, l\,), \tag{2}$$

Here, c_t is an action latent conditioning signal that is used to guide visual forecasting. For visual feedforward dynamics modeling, given the same context and the predicted action conditioning signal, the model learns to sample a future observation sequence that captures the evolution of visual observations:

$$(o_{t+\Delta},\, o_{t+2\Delta},\, \ldots,\, o_{t+K\Delta}) \sim g_\Theta\big(\cdot \mid o_t,\, s_t,\, l,\, c_t\big), \tag{3}$$

where Δ denotes the temporal stride between predicted observations, and $K = \lfloor p/\Delta \rfloor$ so that the model predicts $\{o_{t+k\Delta}\}_{k=1}^{K}$ within the p-step horizon.

3.2 The Architecture of GigaWorld-Policy

As shown in Fig. 3, GigaWorld-Policy adapts a 5B-parameter diffusion Transformer [54], pre-trained via an action-centered objective to serve as a World–Action model for robotic manipulation. By concatenating multi-view inputs, the framework enables joint cross-view reasoning with consistency, and uses a causal masking scheme to unify action generation and visual dynamics.

Input Tokens. For visual token inputs, to enable multi-view generation without modifying the backbone while encouraging cross-view consistency, we merge the three camera views into a single composite image of the same resolution as a standard input:

$$o_t^{comp} = Compose\Big(o_t^{left},\, o_t^{front},\, o_t^{right}\Big)\,. \tag{4}$$

This composite representation preserves the spatial structure of each view in a shared coordinate frame, facilitating cross-view consistency. Meanwhile, since dense frame-by-frame prediction is frequently unnecessary due to strong temporal continuity and redundancy in adjacent observations, we only forecast a sparse

set of future frames $\{ o^{comp}_{t+k\Delta} \}_{k=1}^{K}$ using a fixed stride. Concretely, we predict one future observation every Δ steps along the action horizon, which preserves the key evolution of the scene while reducing supervision redundancy.

Shared Transformer Blocks. We then encode both the current observation o^{comp}_t and the predicted future observations $\{ o^{comp}_{t+k\Delta} \}_{k=1}^{K}$ using the same pre-trained Variational Autoencoder (VAE), and tokenize the resulting latents into spatiotemporal visual tokens T_o and T_f. In parallel, we embed proprioceptive states and actions into the pre-trained model's hidden dimension via linear projections, yielding state tokens T_s and action tokens T_a, respectively. The language instruction l is encoded by a pre-trained language encoder to obtain the instruction token sequence T_l.

Unlike MoE-based designs, we process all token types with a single shared stack of Transformer blocks. In particular, all tokens share the same query, key, and value projection matrices at every layer, which tightly couples action tokens with visual evidence while preserving the computational profile of the pre-trained backbone. Meanwhile, we use different positional encodings for different token types to respect their underlying structures: visual tokens adopt a 2D positional encoding over the image grid, whereas proprioceptive state and action tokens use a 1D temporal positional encoding.

Causal Self-Attention for Video and Action Modeling. To unify action generation and feedforward visual-dynamics modeling within a single diffusion Transformer, we pack all modalities into one token sequence and use a causal attention mask to control information flow. Concretely, at each diffusion step t, we concatenate modality-specific tokens into a unified sequence:

$$T_t = \left[\, T_o \, ; \, T_s \, ; \, T_a \, ; \, T_f \, \right]. \tag{5}$$

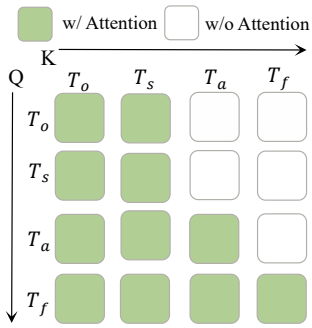

Fig. 4: Attention mask for GigaWorld-Policy: action tokens T_a attend to states T_s and current observations T_o only, while future video tokens T_f also attend to actions.

As shown in Fig. 4, we then impose a blockwise causal attention mask to enforce the following dependencies: (i) T_s and T_o may attend to each other, but cannot attend to T_a or T_f; (ii) action tokens in T_a may attend to the tokens $\{T_s, T_o\}$, but cannot attend to T_f; (iii) future-video tokens in T_f may attend to $\{T_s, T_o, T_a\}$, enabling feedforward dynamics prediction conditioned on the action chunk. This masking scheme prevents information leakage from predicted future frames into action generation, while allowing future-frame prediction to leverage both observations and actions, consistent with Eq. 2 and Eq. 3.

Notably, the language instruction is not included in the unified self-attention sequence; instead, l is provided as an external conditioning signal via cross-attention, and therefore does not participate in the causal ordering above.

Table 1: Datasets and their estimated collection hours used in embodied data pre-training, covering egocentric human demonstrations and real-world robot manipulation videos.

Dataset	Hours	Dataset	Hours	Dataset	Hours
EgoDex [21]	800	Agibot [4]	2,500	EGO4D [20]	3,500
RoboMind [63]	300	RDT [37]	25	Open X-Embodiment [44]	3,500
DROID [25]	350	ATARA [17]	10	Something-Something V2 [19]	200

3.3 GigaWorld-Policy: Training

Training Process and Data. We pre-train GigaWorld-Policy to progressively inject physics priors from diverse video sources, enabling the model to acquire generalizable visual-dynamics knowledge before any task-specific supervision.

We first initialize GigaWorld-Policy from a large-scale pretrained video model, trained on diverse web videos. Building on this foundation, we perform embodied data pre-training, where the model is further pre-trained on robot-centered video data spanning real-world robot videos and large-scale egocentric human videos. On the robot side, we aggregate real-world robot videos from multiple sources (e.g., Agibot [4], RDT [37], RoboMind [63], ATARA [17]), which capture robot-specific imaging characteristics, embodiment and workspace constraints, and the distinctive visual patterns induced by arms and end-effectors during interaction. In parallel, we include large-scale egocentric human demonstration videos (e.g., EgoDex [21], Ego4D [20]) to broaden coverage of everyday interaction primitives and long-horizon activity structure, improving robustness to diverse scenes, tools, and task contexts. Overall, as shown in Tab. 1, we collect approximately 10,000 hours of data, and apply unified cleaning, formatting, and sampling across sources to ensure quality and a controllable data distribution. This embodied stage adapts the representation to embodiment-specific viewpoints and manipulation-relevant interaction patterns, improving robustness to viewpoint-induced appearance variations.

After pre-training, the model is post-trained on target-robot task trajectory data that pairs images, language, and actions. This stage specializes the model to the target robot by learning instruction-conditioned action prediction under the robot's control interface and state distribution.

Training Objective. We use flow matching to optimize both action prediction and visual feedforward dynamics modeling. For either modality x (action tokens or future video tokens), we sample a flow time $s \sim \mathcal{U}(0,1)$ and noise $\epsilon \sim \mathcal{N}(0, I)$, and construct the interpolated noised variable

$$x^{(s)} = (1-s)\epsilon + sx, \tag{6}$$

with target velocity:

$$\dot{x}^{(s)} = x - \epsilon. \tag{7}$$

Let z_f denote the VAE latents corresponding to the future observation tokens T_f. We train the model to predict the velocity field of future latents conditioned

on history and the executed action chunk:

$$\mathcal{L}_{video} = \mathbb{E}_{s,\epsilon}\Big[\big\| g_\Theta\big(z_f^{(s)},\, s \mid T_s,\, T_o,\, T_a,\, T_l\big) - \dot{z}_f^{(s)} \big\|^2\Big]. \tag{8}$$

Similarly, letting a denote the action-token representation for the action chunk T_a, we optimize an action flow-matching objective conditioned on history:

$$\mathcal{L}_{action} = \mathbb{E}_{s,\epsilon}\Big[\big\| g_\Theta\big(a^{(s)},\, s \mid T_s,\, T_o,\, T_l\big) - \dot{a}^{(s)} \big\|^2\Big]. \tag{9}$$

For pre-training, we optimize only the video flow-matching objective. For post-training, we combine the video and action objectives using scalar weights λ_{video} and λ_{action} to balance their contributions:

$$\mathcal{L}_{all} = \lambda_{video}\,\mathcal{L}_{video} + \lambda_{action}\,\mathcal{L}_{action}. \tag{10}$$

3.4 GigaWorld-Policy: Inference

At inference time, our goal is to generate actions with low latency. Directly running the unified video–action diffusion Transformer would require sampling the future video-token stream at every control step, which is costly because video tokens are typically much longer than action tokens. Moreover, predicting future frames is not necessary for executing the policy. We therefore adopt action decoding and optionally decode video, preserving the same backbone and tokenization as in training while avoiding video-generation overhead.

We keep the same conditioning pipeline for language, observations, and proprioception. At each control step t, we first form the context from the instruction, the current proprioceptive state, and the multi-view observation (concatenated as in Eq. 4):

$$w_t = \big(T_l,\ T_s,\ T_o\big). \tag{11}$$

We then sample only the action chunk tokens T_a from the learned action flow model conditioned on w_t, without instantiating any future video tokens. Concretely, we initialize $a^{(0)} \sim \mathcal{N}(0, I)$ and integrate the learned velocity field from $s = 0$ to $s = 1$:

$$\frac{da^{(s)}}{ds} = g_\Theta\big(a^{(s)},\, s \mid w_t\big), \qquad s \in [0, 1], \tag{12}$$

obtaining $a^{(1)}$, which is then decoded to the continuous action chunk $\hat{a}_{t:t+p-1}$. Finally, we execute the action, observe the new state and images, and repeat the above procedure at the next control step. If future prediction is needed, we can enable the video branch at inference time. This can be done either by including the future video tokens in the input and denoising them jointly with the action tokens, or by reusing the KV cache saved during action denoising and then denoising the video tokens conditioned on that cached context.

4 Experiment

Evaluation Metrics. We primarily report the Success Rate (SR) to evaluate task completion performance. In simulation, SR is a binary indicator: SR = 1 if the task is completed and SR = 0 otherwise. For real-world experiments, we adopt a graded evaluation protocol. Specifically, for the pick-and-place task, we assign a score of 0.5 for a successful grasp and an additional 0.5 for successfully placing the object at the designated target location.

Table 2: Inference latency on an NVIDIA A100, and Success Rate in simulation and real world. The best results are marked in **bold**, and the second-best results are underlined.

Method	Time (ms)	SR in Simulation	SR in Real-World
Vision–Language–Action Models			
$\pi_{0.5}$ [23]	**225**	0.48	0.69
GigaBrain-0 [51]	452	–	0.68
World–Action Models			
Motus [2]	3231	**0.88**	<u>0.76</u>
Cosmos-Policy [26]	1413	–	0.58
Ours	<u>360</u>	<u>0.86</u>	**0.83**

Baselines. We compare GigaWorld-Policy against several state-of-the-art baselines spanning two dominant paradigms. For VLM-based VLA methods, we include $\pi_{0.5}$ [23], GigaBrain-0 [51], and X-VLA [77], which leverage large VLM backbones to align visual observations with language instructions and decode low-level control commands via an action head, serving as strong representatives of end-to-end perception-to-action pipelines. For WAM approaches, we evaluate Cosmos-Policy [26] and Motus [2], which explicitly learn environment dynamics in a latent space and utilize predictive rollouts to support decision making.

Implementation Details. Wan 2.2 5B [54] serves as the diffusion-Transformer backbone. We set the action chunk length to $p = 48$, and adopt a future-observation stride of $\Delta = 12$ to provide auxiliary visual-dynamics supervision by sparsely predicting future observations along the action horizon. During post-training, we balance objectives with loss weights $\lambda_{action} = 5$ and $\lambda_{video} = 1$, emphasizing action prediction while retaining the video-consistency regularizer. When reporting inference-time latency, we use the action-only decoding path, where future-video decoding is disabled and the model outputs actions directly. For real-world evaluations, each method is tested with 20 trials per task; in each trial, the robot is allowed up to five attempts to finish the task under the same instruction. In simulation, we evaluate each task over 100 test episodes.

4.1 Inference Speed Comparison

To assess deployment efficiency, we benchmark per-step inference latency on an NVIDIA A100 GPU and report it together with success rates in simulation and real-world experiments in Tab. 2. During inference, GigaWorld-Policy directly decodes actions without requiring explicit future-video generation, which substantially reduces computation compared with prior World–Action Models. As a result, it achieves an approximately 9× speedup over Motus [2] and a 3.9× speedup over Cosmos-Policy [26]. Despite the reduced latency, GigaWorld-Policy maintains competitive simulation performance, only slightly below Motus, and

Table 3: Evaluation on RoboTwin 2.0 Simulation. The best results are marked in **bold**, and the second-best results are underlined. GigaWorld-Policy achieves performance comparable to Motus while providing a 9× inference speedup.

Simulation Task	$\pi_{0.5}$		X-VLA		Motus		Our	
	Clean	Rand.	Clean	Rand.	Clean	Rand.	Clean	Rand.
Adjust Bottle	0.79	0.83	**1.00**	0.99	0.89	0.93	**1.00**	**1.00**
Place A2b Left	0.62	0.60	0.48	0.49	0.88	0.79	**0.94**	**0.88**
Place Bread Skillet	0.38	0.46	0.77	0.67	0.86	0.83	**0.94**	**0.90**
Place Cans Plasticbox	0.40	0.47	0.97	0.98	0.98	0.94	**1.00**	**1.00**
Place Fan	0.25	0.36	0.80	0.75	0.91	0.87	**0.92**	**0.94**
Place Mouse Pad	0.21	0.26	0.70	0.70	0.66	0.68	**0.88**	**0.90**
Place Object Basket	0.43	0.36	0.44	0.39	0.81	0.87	**0.90**	**0.92**
Place Object Stand	0.74	0.65	0.86	0.88	0.98	0.97	**1.00**	**0.98**
Rotate Qrcode	0.47	0.56	0.34	0.33	0.89	0.73	**0.90**	**0.84**
Shake Bottle	0.91	**1.00**	0.99	**1.00**	**1.00**	0.97	**1.00**	**1.00**
Stamp Seal	0.36	0.23	0.76	0.82	0.93	0.92	**0.96**	**0.98**
Dump Bin Bigbin	0.30	0.42	0.79	0.77	**0.95**	0.91	0.92	**1.00**
Handover Block	0.18	0.19	0.73	0.37	**0.86**	0.73	0.80	**0.80**
... (50+ tasks in total)								
Handover Mic	0.28	0.18	0.00	0.00	**0.78**	0.63	0.72	**0.72**
Move Stapler Pad	0.16	0.18	0.78	0.73	0.83	**0.85**	**0.92**	0.82
Open Laptop	0.19	0.35	0.93	**1.00**	0.95	0.91	**0.96**	0.98
Place A2b Right	0.62	0.57	0.36	0.36	**0.91**	0.87	0.90	**0.92**
Place Burger Fries	0.66	0.70	0.94	0.94	**0.98**	**0.98**	**0.98**	0.96
Place Container Plate	0.71	0.78	0.97	0.95	**0.98**	**0.99**	**0.98**	0.96
Press Stapler	0.80	0.70	0.92	**0.98**	0.93	**0.98**	**0.96**	0.96
Put Object Cabinet	0.24	0.15	0.46	0.48	**0.88**	0.71	0.74	**0.74**
Place Dual Shoes	0.12	0.07	0.79	**0.88**	0.93	0.87	**0.96**	0.84
Stack Blocks Two	0.48	0.56	0.92	0.87	**1.00**	**0.98**	**1.00**	0.94
Turn Switch	0.05	0.06	0.40	0.61	**0.84**	0.78	0.82	**0.84**
Average	0.43	0.44	0.73	0.73	**0.89**	**0.87**	0.87	0.85

obtains the best real-world success rate, improving over Motus by 7 percentage points. Compared with VLA baselines, GigaWorld-Policy is slower than $\pi_{0.5}$ but remains efficient, while yielding substantially higher success rates in both simulation and real-world settings. These results demonstrate that GigaWorld-Policy achieves a favorable trade-off between the capability of world-model-based policies and the inference efficiency required for practical robot deployment.

4.2 Simulation Benchmark Experiments

Simulation Setup. We evaluate single-task performance on RoboTwin 2.0 [8] over 50 representative manipulation tasks under domain randomization. To assess cross-task generalization and the effect of pre-training, we train all baselines and GigaWorld-Policy on a mixed dataset of 2,500 demonstrations from clean

Table 4: Comparison of task success rates in real-world experiments. The best results are marked in **bold**, and the second-best results are underlined.

Method	Clean the Desk	Scan a QR Code	Sweep up Trash	Stack Bowls	Average
Vision–Language–Action Models					
$\pi_{0.5}$ [23]	0.75	0.55	0.65	0.80	0.69
GigaBrain-0 [51]	0.70	0.65	0.60	0.75	0.68
World–Action Models					
Motus [2]	0.80	**0.75**	0.70	0.80	0.76
Cosmos-Policy [26]	0.65	0.50	0.45	0.70	0.58
Our	**0.90**	**0.75**	**0.75**	**0.90**	**0.83**

scenes (50 per task) and 25,000 demonstrations from heavily randomized scenes (500 per task). Randomization includes backgrounds, table clutter, table height perturbations, and lighting changes, testing robustness to distribution shifts.

Results. As shown in Tab. 3, GigaWorld-Policy consistently achieves strong performance across the 50 RoboTwin 2.0 [8] manipulation tasks. Compared with the $\pi_{0.5}$ [23] model, our method improves the average success rate by over 44 percentage points in the multi-task setting, demonstrating the benefit of large-scale action-centered pre-training. GigaWorld-Policy also performs competitively with Motus [2], reaching second-best average success rates under both evaluation settings, while achieving a 9× inference speedup. These results suggest that our action-centered World–Action modeling paradigm effectively strengthens policy learning via joint supervision, combining action prediction with a visual-dynamics consistency auxiliary constraint during training to better align learned representations with downstream control. At inference time, future-video prediction becomes optional, enabling low-latency action generation and substantially accelerating deployment without sacrificing overall task performance.

4.3 Real-World Experiment

To evaluate real-world effectiveness, we conducted gripper-based manipulation experiments on an AgileX PiPER 6-DoF robotic arm. We further designed four real-world tasks—Clean the Desk, Scan a QR Code, Stack Bowls, and Sweep up Trash—with detailed descriptions provided in the supplementary material. As summarized in Tab. 4, GigaWorld-Policy achieves stronger overall performance than representative world-model-based robot policies. Notably, although GigaWorld-Policy runs nine times faster than Motus [2], it still attains a 7% higher success rate. We attribute this advantage to the tight coupling between success and execution efficiency in real deployments: slower inference introduces control latency, reduces the effective action update rate, and weakens closed-loop error correction. GigaWorld-Policy also maintains a clear advantage over VLA baselines: compared with $\pi_{0.5}$ [23], it improves the average real-world success rate by about 14% and achieves consistently higher success across all four tasks.

Table 5: Effect of the future-frame sampling interval on real-world task success rate. For $\Delta > 0$, the model predicts $K = \lfloor 48/\Delta \rfloor$ future frames. "No pred." denotes the setting without future video prediction, i.e., $K = 0$. The best results are marked in **bold**, and the second-best results are underlined.

Setting	No pred.	$\Delta = 4$	$\Delta = 8$	$\Delta = 12$	$\Delta = 24$	$\Delta = 48$
K	0	12	6	4	2	1
Success Rate ↑	0.60	0.76	0.78	**0.83**	0.80	0.76

Table 6: Comparison of real-world OOD generalization. We evaluate models under unseen environments, novel objects, changed lighting, and modified tablecloths.

Method	Clean the Desk	Fold Clothes	Sweep up Trash	Avg.
$\pi_{0.5}$	0.55	0.45	0.50	0.50
Motus	0.65	0.55	0.60	0.60
Cosmos-Policy	0.45	0.35	0.40	0.40
Ours	**0.80**	**0.65**	**0.70**	**0.72**

Table 7: Success rate on challenging real-world tasks, including high-dexterity folding and long-horizon stacking.

Method	Stack Boxes	Fold Clothes	Avg.
$\pi_{0.5}$	0.55	0.55	0.55
Motus	0.60	0.55	0.58
Cosmos-Policy	0.45	0.50	0.48
Ours	**0.65**	**0.70**	**0.68**

We further evaluate GigaWorld-Policy in OOD real-world settings and harder manipulation tasks. OOD tests include variations in object color/size, clothing shape, unseen trash, and novel tablecloth/lighting conditions. Challenging tasks include high-dexterity Fold Clothes and long-horizon Stack Boxes. As shown in Tabs.6 and4, GigaWorld-Policy consistently outperforms all baselines, achieving average success rates of 0.72 on OOD tasks and 0.68 on challenging tasks.

4.4 Data Efficiency

Collecting real-world robot data is expensive, making sample-efficient policy learning critical. We study the data efficiency of GigaWorld-Policy and compare it against a VLA baseline. For each real-world task, we subsample the training demonstrations to multiple fractions, train all methods under the same setting, and evaluate success rates using the same protocol. As shown in Fig. 5, conditioning on the video-prior representations learned by our World–Action model yields an improvement in sample efficiency. Specifically, GigaWorld-Policy reaches the maximum success rate achieved by the VLA using only 10% of the training data.

4.5 Ablation Study

In this section, we conduct real-world ablation studies to isolate the contribution of each component to task success and robustness.

Importance of Pre-training. We analyze our pre-training pipeline by ablating its components. We compare four post-training settings: (i) training from

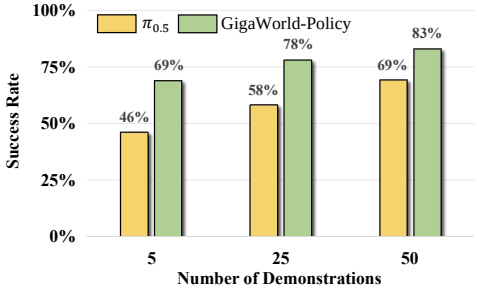

Fig. 5: Success rate as a function of training data fraction on real-world tasks. GigaWorld-Policy matches the $\pi_{0.5}$ with only 10% of the data.

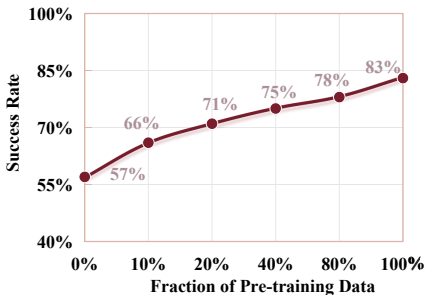

Fig. 6: Increasing the fraction of embodied data pre-training consistently improves real-world success.

Table 8: Ablation study on the causal self-attention mask. We report SR and generation quality on the real-world test set. The best results are **bold**.

Method	SR ↑	PSNR ↑	SSIM ↑
Self-Attn	0.81	27.87	0.892
Ours	**0.83**	**28.41**	**0.901**

Table 9: Effect of different pre-training configurations on SR.

Pretrained Init	Embodied Data Pre-training	SR ↑
		0.45
✓		0.57
	✓	0.73
✓	✓	**0.83**

scratch, (ii) initializing from the pretrained video model only, (iii) using embodied data pre-training only, and (iv) combining pretrained video initialization with embodied data pre-training. As shown in Tab. 9, training from scratch achieves an SR of 0.45. Video-model initialization improves SR to 0.57, while embodied data pre-training alone reaches 0.73. Combining both gives the best performance, showing their complementary benefits. We further study the effect of embodied pre-training data scale by fixing video initialization and post-training settings while varying the embodied data fraction. As shown in Fig. 6, real-world success steadily increases from 57% without embodied pre-training to 83% with the full dataset. This consistent trend suggests that larger embodied coverage strengthens action-relevant representations and improves real-world robustness.

Effect of Future-Frame Sampling Interval. We study whether feed-forward dynamics modeling improves action understanding and execution. As shown in Tab. 5, we fix the action chunk length to 48 and vary the number of predicted future frames by adjusting the sampling interval Δ. For $\Delta > 0$, the model predicts $K = \lfloor 48/\Delta \rfloor$ future frames. We additionally include a "No pred." setting with $K = 0$, where the model performs no future video prediction and degenerates to an action decoder conditioned only on the current observation. The results show that enabling feed-forward dynamics modeling consistently improves the success rate, increasing it from 0.60 in the no-prediction setting to 0.83 at $\Delta = 12$, yielding an absolute gain of 0.23. This suggests that anticipating future observations

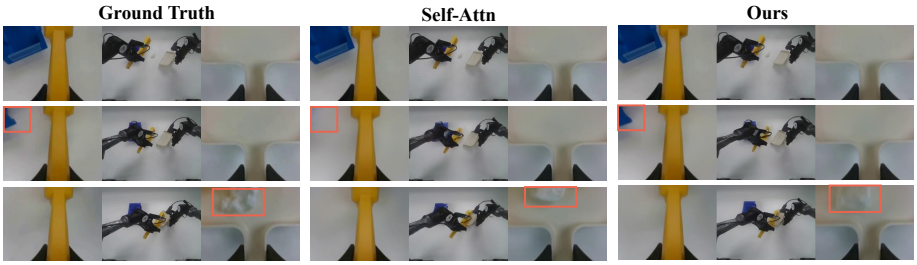

Fig. 7: Qualitative comparison between the self-attention baseline and our method. In the red boxes, our method more accurately predicts object state changes.

provides useful temporal priors for decision making, leading to more accurate and stable action execution. We observe that denser future-frame prediction is not always better: predicting too many frames, e.g., $\Delta = 4$, may introduce unnecessary modeling difficulty, while predicting too few frames, e.g., $\Delta = 48$, provides limited temporal supervision. A moderate sampling interval, $\Delta = 12$, achieves the best trade-off between dynamics modeling and action prediction.

Role of Causal Self-Attention for Video–Action Modeling. Our causal design prevents leakage by restricting action tokens to attend only to the state and current observation. We compare this causal mask with a self-attention variant. In the self-attention variant, all tokens can attend to each other without constraints. As shown in Tab. 8, both achieve similar SR under matched settings. However, causal self-attention is more deployable: it ensures the video prediction is optional at inference. We also compare future video generation quality on a held-out real-world test set from the target platform, using standard reconstruction metrics. As shown in Tab.8, our method achieves higher PSNR and SSIM; furthermore, in Fig.7, the red boxed regions indicate that our method more accurately predicts fine-grained object state and appearance changes. We attribute this improvement to the proposed causal masking, which enforces a physically meaningful factorization between action prediction and forward visual dynamics.

5 Conclusion

We introduced GigaWorld-Policy, an action-centered World–Action Model built on an action-conditioned video generation backbone pre-trained on a multi-level large-scale robot dataset. By decomposing policy learning into future action sequence prediction together with action-conditioned video generation during training, GigaWorld-Policy learns 2D pixel–action dynamics with supervision, producing more accurate and plausible action plans. Experiments on real-world robotic platforms show that GigaWorld-Policy reduces inference overhead while improving task performance, achieving a 9× inference speedup (down to 0.36 s per inference) and up to a 7% increase in task success rates over prior baselines.

References

1. Beyer, L., Steiner, A., Pinto, A.S., Kolesnikov, A., Wang, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschannen, M., Bugliarello, E., et al.: Paligemma: A versatile 3b vlm for transfer. arXiv preprint arXiv:2407.07726 (2024)
2. Bi, H., Tan, H., Xie, S., Wang, Z., Huang, S., Liu, H., Zhao, R., Feng, Y., Xiang, C., Rong, Y., et al.: Motus: A unified latent action world model. arXiv preprint arXiv:2512.13030 (2025)
3. Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al.: pi0: A vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164 (2024)
4. Bu, Q., Cai, J., Chen, L., Cui, X., Ding, Y., Feng, S., He, X., Huang, X., et al.: Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. In: IROS (2025)
5. Cen, J., Yu, C., Yuan, H., Jiang, Y., Huang, S., Guo, J., Li, X., Song, Y., Luo, H., Wang, F., et al.: Worldvla: Towards autoregressive action world model. arXiv preprint arXiv:2506.21539 (2025)
6. Chang, Y., Qin, J., Qiao, L., Wang, X., Zhu, Z., Ma, L., Wang, X.: Scalable training for vector-quantized networks with 100% codebook utilization. arXiv preprint arXiv:2509.10140 (2025)
7. Cheang, C., Chen, S., Cui, Z., Hu, Y., Huang, L., Kong, T., Li, H., Li, Y., Liu, Y., Ma, X., et al.: Gr-3 technical report. arXiv preprint arXiv:2507.15493 (2025)
8. Chen, T., Chen, Z., Chen, B., Cai, Z., Liu, Y., Liang, Q., Li, Z., Lin, X., Ge, Y., Gu, Z., et al.: Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. arXiv preprint arXiv:2506.18088 (2025)
9. Chen, Y., Cao, Z., Ren, H., Yang, C., Li, W., Wang, S., Wang, Y., Zhang, L., Shao, Y., Zhao, Z., et al.: Roborouter: Training-free policy routing for robotic manipulation. arXiv preprint arXiv:2603.07892 (2026)
10. Chen, Z., Luo, X., Li, D.: Visrl: Intention-driven visual perception via reinforced reasoning. arXiv preprint arXiv:2503.07523 (2025)
11. Chen, Z., Zhang, M., Yu, X., Luo, X., Sun, M., Pan, Z., Feng, Y., Pei, P., Cai, X., Huang, R.: Think with 3d: Geometric imagination grounded spatial reasoning from limited views. arXiv preprint arXiv:2510.18632 (2026)
12. Chen, Z., Zhao, R., Luo, C., Sun, M., Yu, X., Kang, Y., Huang, R.: Sifthinker: Spatially-aware image focus for visual reasoning. arXiv preprint arXiv:2508.06259 (2025)
13. Cui, C., Ding, P., Song, W., Bai, S., Tong, X., Ge, Z., Suo, R., Zhou, W., Liu, Y., Jia, B., et al.: Openhelix: A short survey, empirical analysis, and open-source dual-system vla model for robotic manipulation. arXiv preprint arXiv:2505.03912 (2025)
14. Ding, P., Ma, J., Tong, X., Zou, B., Luo, X., Fan, Y., Wang, T., Lu, H., Mo, P., Liu, J., et al.: Humanoid-vla: Towards universal humanoid control with visual integration. arXiv preprint arXiv:2502.14795 (2025)
15. Ding, P., Zhao, H., Zhang, W., Song, W., Zhang, M., Huang, S., Yang, N., Wang, D.: Quar-vla: Vision-language-action model for quadruped robots. In: ECCV (2024)
16. Dong, Z., Wang, X., Zhu, Z., Wang, Y., Wang, Y., Zhou, Y., Wang, B., Ni, C., Ouyang, R., Qin, W., et al.: Emma: Generalizing real-world robot manipulation via generative visual transfer. arXiv preprint arXiv:2509.22407 (2025)

17. Feng, Y., Tan, H., Mao, X., Xiang, C., Liu, G., Huang, S., Su, H., Zhu, J.: Vidar: Embodied video diffusion model for generalist manipulation. arXiv preprint arXiv:2507.12898 (2025)
18. Fu, Z., Li, Z., Chen, Z., Wang, C., Song, X., Hu, Y., Nie, L.: Pair: Complementarity-guided disentanglement for composed image retrieval. In: ICASSP (2025)
19. Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The" something something" video database for learning and evaluating visual common sense. In: ICCV (2017)
20. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: CVPR (2022)
21. Hoque, R., Huang, P., Yoon, D.J., Sivapurapu, M., Zhang, J.: Egodex: Learning dexterous manipulation from large-scale egocentric video. arXiv preprint arXiv:2505.11709 (2025)
22. Hu, Y., Li, Z., Chen, Z., Huang, Q., Fu, Z., Xu, M., Nie, L.: Refine: Composed video retrieval via shared and differential semantics enhancement. ACM Transactions on Multimedia Computing, Communications and Applications (2026)
23. Intelligence, P., Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., et al.: $\pi_{0.5}$: a vision-language-action model with open-world generalization. arXiv preprint arXiv:2504.16054 (2025)
24. Jiang, T., Yuan, T., Liu, Y., Lu, C., Cui, J., Liu, X., Cheng, S., Gao, J., Xu, H., Zhao, H.: Galaxea open-world dataset and g0 dual-system vla model. arXiv preprint arXiv:2509.00576 (2025)
25. Khazatsky, A., Pertsch, K., Nair, S., Balakrishna, A., Dasari, S., Karamcheti, S., Nasiriany, S., Srirama, M.K., Chen, L.Y., Ellis, K., et al.: Droid: A large-scale in-the-wild robot manipulation dataset. arXiv preprint arXiv:2403.12945 (2024)
26. Kim, M.J., Gao, Y., Lin, T.Y., Lin, Y.C., Ge, Y., Lam, G., Liang, P., Song, S., Liu, M.Y., Finn, C., et al.: Cosmos policy: Fine-tuning video models for visuomotor control and planning. arXiv preprint arXiv:2601.16163 (2026)
27. Kim, S., Lee, S., Cho, M.: Freeaction: Training-free techniques for enhanced fidelity of trajectory-to-video generation. arXiv preprint arXiv:2509.24241 (2025)
28. Li, H., Zhang, I., Ouyang, R., Wang, X., Zhu, Z., Yang, Z., Zhang, Z., Wang, B., Ni, C., Qin, W., et al.: Mimicdreamer: Aligning human and robot demonstrations for scalable vla training. arXiv preprint arXiv:2509.22199 (2025)
29. Li, H., Ding, P., Suo, R., Wang, Y., Ge, Z., Zang, D., Yu, K., Sun, M., Zhang, H., Wang, D., et al.: Vla-rft: Vision-language-action reinforcement fine-tuning with verified rewards in world simulators. arXiv preprint arXiv:2510.00406 (2025)
30. Li, L., Zhang, Q., Luo, Y., Yang, S., Wang, R., Han, F., Yu, M., Gao, Z., Xue, N., Zhu, X., et al.: Causal world modeling for robot control. arXiv preprint arXiv:2601.21998 (2026)
31. Li, Z., Chen, Z., Wen, H., Fu, Z., Hu, Y., Guan, W.: Encoder: Entity mining and modification relation binding for composed image retrieval. In: AAAI (2025)
32. Li, Z., Fu, Z., Hu, Y., Chen, Z., Wen, H., Nie, L.: Finecir: Explicit parsing of fine-grained modification semantics for composed image retrieval. https://arxiv.org/abs/2503.21309 (2025)
33. Li, Z., Hu, Y., Chen, Z., Huang, Q., Qiu, G., Fu, Z., Liu, M.: Retrack: Evidence-driven dual-stream directional anchor calibration network for composed video retrieval. In: AAAI (2026)

34. Liu, J., Han, J., Liu, L., Aviles-Rivero, A.I., Jiang, C., Liu, Z., Wang, H.: Mamba4d: Efficient 4d point cloud video understanding with disentangled spatial-temporal state space models. In: CVPR (2025)
35. Liu, J., Liu, M., Zhu, S., Zhang, Y., Li, J., Yang, M.Y., Nex, F., Cheng, H., Wang, H.: Arflow: Auto-regressive optical flow estimation for arbitrary-length videos via progressive next-frame forecasting. In: ICLR (2026)
36. Liu, L., Wang, X., Zhao, G., Li, K., Qin, W., Qiu, J., Zhu, Z., Huang, G., Su, Z.: Robotransfer: Geometry-consistent video diffusion for robotic visual policy transfer. arXiv preprint arXiv:2505.23171 (2025)
37. Liu, S., Wu, L., Li, B., Tan, H., Chen, H., Wang, Z., Xu, K., Su, H., Zhu, J.: Rdt-1b: a diffusion foundation model for bimanual manipulation. arXiv preprint arXiv:2410.07864 (2024)
38. Marafioti, A., Zohar, O., Farré, M., Noyan, M., Bakouch, E., Cuenca, P., Zakka, C., Allal, L.B., Lozhkov, A., Tazi, N., et al.: Smolvlm: Redefining small and efficient multimodal models. arXiv preprint arXiv:2504.05299 (2025)
39. Ni, C., Chen, C., Wang, X., Zhu, Z., Zheng, W., Wang, B., Chen, T., Zhao, G., Li, H., Dong, Z., et al.: Swiftvla: Unlocking spatiotemporal dynamics for lightweight vla models at minimal overhead. arXiv preprint arXiv:2512.00903 (2025)
40. Ni, C., Li, J., Li, H., Liu, H., Wang, X., Zhu, Z., Zhao, G., Wang, B., Li, C., Huang, G., et al.: Wonderfree: Enhancing novel view quality and cross-view consistency for 3d scene exploration. arXiv preprint arXiv:2506.20590 (2025)
41. Ni, C., Wang, X., Zhu, Z., Wang, W., Li, H., Zhao, G., Li, J., Qin, W., Huang, G., Mei, W.: Wonderturbo: Generating interactive 3d world in 0.72 seconds. arXiv preprint arXiv:2504.02261 (2025)
42. Ni, C., Zhao, G., Wang, X., Zhu, Z., Qin, W., Chen, X., Jia, G., Huang, G., Mei, W.: Recondreamer-rl: Enhancing reinforcement learning via diffusion-based scene reconstruction. arXiv preprint arXiv:2508.08170 (2025)
43. Ni, C., Zhao, G., Wang, X., Zhu, Z., Qin, W., Huang, G., Liu, C., Chen, Y., Wang, Y., Zhang, X., et al.: Recondreamer: Crafting world models for driving scene reconstruction via online restoration. In: CVPR (2025)
44. O'Neill, A., Rehman, A., Maddukuri, A., Gupta, A., Padalkar, A., Lee, A., Pooley, A., Gupta, A., Mandlekar, A., Jain, A., et al.: Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In: ICRA (2024)
45. Pai, J., Achenbach, L., Montesinos, V., Forrai, B., Mees, O., Nava, E.: mimic-video: Video-action models for generalizable robot control beyond vlas. arXiv preprint arXiv:2512.15692 (2025)
46. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
47. Shen, Y., Wei, F., Du, Z., Liang, Y., Lu, Y., Yang, J., Zheng, N., Guo, B.: Videovla: Video generators can be generalizable robot manipulators. arXiv preprint arXiv:2512.06963 (2025)
48. Steiner, A., Pinto, A.S., Tschannen, M., Keysers, D., Wang, X., Bitton, Y., Gritsenko, A., Minderer, M., Sherbondy, A., Long, S., et al.: Paligemma 2: A family of versatile vlms for transfer. arXiv preprint arXiv:2412.03555 (2024)
49. Team, A., Zhu, H., Wang, Y., Zhou, J., Chang, W., Zhou, Y., Li, Z., Chen, J., Shen, C., Pang, J., et al.: Aether: Geometric-aware unified world modeling. arXiv preprint arXiv:2503.18945 (2025)

50. Team, G., Wang, B., Ni, C., Huang, G., Zhao, G., Li, H., Li, J., Lv, J., Liu, J., Feng, L., et al.: Gigabrain-0.5 m*: a vla that learns from world model-based reinforcement learning. arXiv preprint arXiv:2602.12099 (2026)
51. Team, G., Ye, A., Wang, B., Ni, C., Huang, G., Zhao, G., Li, H., Li, J., Zhu, J., Feng, L., et al.: Gigabrain-0: A world model-powered vision-language-action model. arXiv preprint arXiv:2510.19430 (2025)
52. Team, G., Ye, A., Wang, B., Ni, C., Huang, G., Zhao, G., Li, H., Zhu, J., Li, K., Xu, M., et al.: Gigaworld-0: World models as data engine to empower embodied ai. arXiv preprint arXiv:2511.19861 (2025)
53. Tian, J., Wang, L., Zhou, S., Wang, S., Li, J., Sun, H., Tang, W.: Pdfactor: Learning tri-perspective view policy diffusion field for multi-task robotic manipulation. In: CVPR (2025)
54. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
55. Wang, B., Meng, X., Wang, X., Zhu, Z., Ye, A., Wang, Y., Yang, Z., Ni, C., Huang, G., Wang, X.: Embodiedreamer: Advancing real2sim2real transfer for policy training via embodied world modeling. arXiv preprint arXiv:2507.05198 (2025)
56. Wang, B., Ouyang, R., Wang, X., Zhu, Z., Zhao, G., Ni, C., Huang, G., Liu, L., Wang, X.: Humandreamer-x: Photorealistic single-image human avatars reconstruction via gaussian restoration. arXiv preprint arXiv:2504.03536 (2025)
57. Wang, B., Wang, X., Ni, C., Zhao, G., Yang, Z., Zhu, Z., Zhang, M., Zhou, Y., Chen, X., Huang, G., Liu, L., Wang, X.: Humandreamer: Generating controllable human-motion videos via decoupled generation. arXiv preprint arXiv:2503.24026 (2025)
58. Wang, S., Tian, J., Wang, L., Liao, Z., Li, J., Dong, H., Xia, K., Zhou, S., Tang, W., Gang, H.: Sampo: Scale-wise autoregression with motion prompt for generative world models. arXiv preprint arXiv:2509.15536 (2025)
59. Wang, S., Wang, L., Zhou, S., Tian, J., Li, J., Sun, H., Tang, W.: Flowram: Grounding flow matching policy with region-aware mamba framework for robotic manipulation. In: CVPR (2025)
60. Wang, S., Pei, M., Sun, L., Deng, C., Shao, K., Tian, Z., Zhang, H., Wang, J.: Spatialviz-bench: An mllm benchmark for spatial visualization. arXiv preprint arXiv:2507.07610 (2025)
61. Wang, W., Zhu, J., Zhang, Z., Wang, X., Zhu, Z., Zhao, G., Ni, C., Wang, H., Huang, G., Chen, X., et al.: Drivegen3d: Boosting feed-forward driving scene generation with efficient video diffusion. arXiv preprint arXiv:2510.15264 (2025)
62. Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., Lu, J.: Drivedreamer: Towards real-world-drive world models for autonomous driving. In: ECCV (2024)
63. Wu, K., Hou, C., Liu, J., Che, Z., Ju, X., Yang, Z., Li, M., Zhao, Y., Xu, Z., Yang, G., et al.: Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. arXiv preprint arXiv:2412.13877 (2024)
64. Xiang, J., Liu, G., Gu, Y., Gao, Q., Ning, Y., Zha, Y., Feng, Z., Tao, T., Hao, S., Shi, Y., et al.: Pandora: Towards general world model with natural language actions and video states. arXiv preprint arXiv:2406.09455 (2024)

65. Xiao, C., Dou, J., Lin, Z., Ke, Z., Hou, L.: From points to coalitions: Hierarchical contrastive shapley values for prioritizing data samples. arXiv preprint arXiv:2512.19363 (2025)
66. Xiao, C., Xu, T., Ma, S., Jiang, Y., Gao, H., Wu, Y.: Reversible primitive–composition alignment for continual vision–language learning. In: ICLR (2026)
67. Ye, A., Zhang, Z., Wang, B., Wang, X., Zhang, D., Zhu, Z.: Vla-r1: Enhancing reasoning in vision-language-action models. arXiv preprint arXiv:2510.01623 (2025)
68. Ye, H., Chen, S., Zhong, Z., Xiao, C., Zhang, H., Wu, Y., Shen, F.: Seeing through the conflict: Transparent knowledge conflict handling in retrieval-augmented generation. arXiv preprint arXiv:2601.06842 (2026)
69. Zeng, S., Chang, X., Xie, M., Liu, X., Bai, Y., Pan, Z., Xu, M., Wei, X.: Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving. arXiv preprint arXiv:2505.17685 (2025)
70. Zeng, S., Qi, D., Chang, X., Xiong, F., Xie, S., Wu, X., Liang, S., Xu, M., Wei, X.: Janusvln: Decoupling semantics and spatiality with dual implicit memory for vision-language navigation. arXiv preprint arXiv:2509.22548 (2025)
71. Zhang, W., Liu, H., Qi, Z., Wang, Y., Yu, X., Zhang, J., Dong, R., He, J., Wang, H., Zhang, Z., et al.: Dreamvla: a vision-language-action model dreamed with comprehensive world knowledge. arXiv preprint arXiv:2507.04447 (2025)
72. Zhao, G., Ni, C., Wang, X., Zhu, Z., Zhang, X., Wang, Y., Huang, G., Chen, X., Wang, B., Zhang, Y., et al.: Drivedreamer4d: World models are effective data machines for 4d driving scene representation. In: CVPR (2025)
73. Zhao, G., Wang, X., Ni, C., Zhu, Z., Qin, W., Huang, G., Wang, X.: Recondreamer++: Harmonizing generative and reconstructive models for driving scene representation. arXiv preprint arXiv:2503.18438 (2025)
74. Zhao, G., Wang, X., Zhu, Z., Chen, X., Huang, G., Bao, X., Wang, X.: Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. In: AAAI (2025)
75. Zhao, Z., Chen, B.M.: Benchmark for evaluating initialization of visual-inertial odometry. In: 2023 42nd Chinese Control Conference (CCC) (2023)
76. Zhao, Z., Yang, H., Liao, B., Zeng, Y., Yan, S., Gu, Y., Liu, P., Zhou, Y., Li, H., Civera, J.: Advances in global solvers for 3d vision. arXiv preprint arXiv:2602.14662 (2026)
77. Zheng, J., Li, J., Wang, Z., Liu, D., Kang, X., Feng, Y., Zheng, Y., Zou, J., Chen, Y., Zeng, J., et al.: X-vla: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. arXiv preprint arXiv:2510.10274 (2025)
78. Zhou, H., Tang, J., Zhang, J., Li, Y., Xiao, C., Hou, L., Ke, Z., Yao, J.: Comem: Compositional concept-graph memory for vision–language adaptation. In: ICLR (2026)
79. Zhou, S., Wang, S., Yuan, Z., Shi, M., Shang, Y., Yang, D.: GSQ-tuning: Group-shared exponents integer in fully quantized training for LLMs on-device fine-tuning. In: Findings of ACL (2025)
80. Zhou, S., Yuan, Z., Yang, D., Zhao, Z., Hu, X., Shi, Y., Lu, X., Wu, Q.: Information entropy guided height-aware histogram for quantization-friendly pillar feature encoder. arXiv preprint arXiv:2405.18734 (2024)
81. Zhou, S., Du, Y., Chen, J., Li, Y., Yeung, D.Y., Gan, C.: Robodreamer: Learning compositional world models for robot imagination. arXiv preprint arXiv:2404.12377 (2024)