SEEING FARTHER TO ACT NOW: LONG-HORIZON FUTURE MODELING FOR ROBOT MANIPULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

World Action Models (WAMs) provide predictive context for robot manipulation by conditioning action generation on imagined future visual states. However, existing WAMs commonly couple the visual prediction horizon to the action horizon, restricting future information to the immediate interaction and potentially obscuring longer-term task consequences. We show that prediction horizon introduces a fundamental trade-off: short-horizon futures preserve fine-grained motion and interaction details, whereas longer-horizon futures capture broader task progression but with greater visual uncertainty, making the most useful temporal scale task dependent. Building on this insight, we propose **S**eeing **F**arther to **A**ct **N**ow (**SFAN**), a multi-timescale world-action model that decouples visual prediction from the executable action horizon and assigns complementary roles to different temporal scales. SFAN uses sparsely sampled long-horizon futures to provide broader task context and generate the primary action prediction, while a complementary short-horizon pathway densely samples nearby futures to predict a residual correction conditioned on the primary action. This asymmetric design allows longer-range futures to guide interaction selection while preserving the fine-grained local information needed to refine how the selected interaction is executed. Extensive experiments on simulated and real-world robot manipulation demonstrate that SFAN consistently outperforms existing baselines, with particularly strong gains on tasks that benefit from longer-horizon future context.

1 INTRODUCTION

World Action Models (WAMs) (Wu et al., 2023; Tian et al., 2024; Li et al., 2025; Zhu et al., 2025; Kim et al., 2026; Yuan et al., 2026) have emerged as an effective approach to robot manipulation by modeling how the environment evolves under interaction. By predicting future visual states, WAMs provide anticipatory context that helps policies reason about the consequences of ongoing interactions and generate appropriate actions. Beyond this direct predictive role, recent studies (Yuan et al., 2026; Zhao et al., 2026a) show that future prediction can also improve learned representations during training and enhance policy robustness under distribution shifts at inference, suggesting that future information provides benefits beyond explicit visual forecasting.

Existing WAMs commonly set the visual prediction horizon to match the temporal span of the action chunk, a design choice we refer to as *horizon coupling*. While this captures the immediate motion and interaction induced by the generated action chunk, the predicted future ends at the same temporal boundary and may therefore fail to reveal whether the action advances the task beyond its immediate effect. This limitation becomes particularly important when locally plausible actions can lead to different longer-term task outcomes. Figure 1(a) illustrates such a case, where the robot can perform a plausible local interaction with an incorrect target, while extending the prediction horizon reveals the downstream consequence and provides information that favors the correct target.

More broadly, the choice of prediction horizon determines what information is captured in imagined futures. As shown in Figure 1(b), short-horizon predictions remain close to the current state and preserve fine-grained interaction details, whereas longer-horizon predictions capture broader state transitions at the cost of greater visual uncertainty. Consistent with this trade-off, Figure 1(c) shows that task performance responds differently to increasing prediction horizons across tasks, indicating

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

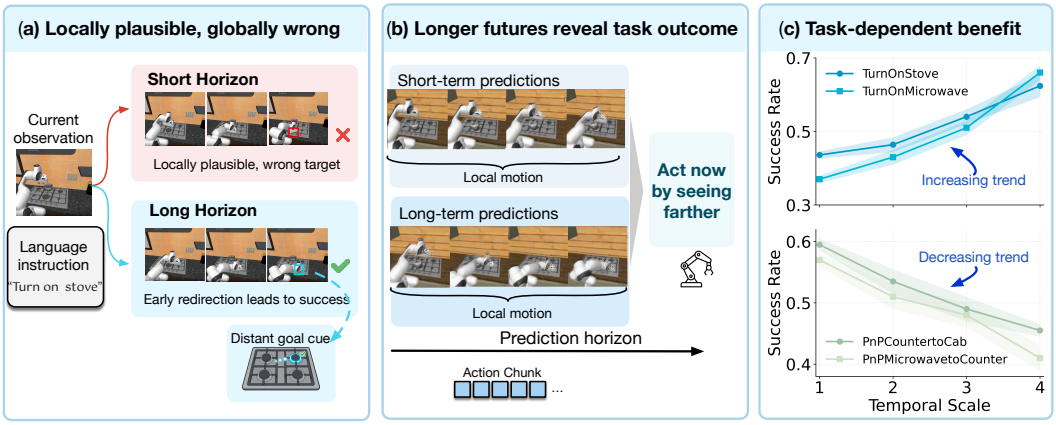

Figure 1: **Motivation for SFAN. (a)** Failure case demonstration under horizon coupling. Short-range prediction favors a locally plausible but incorrect target, whereas a longer horizon reveals downstream consequences and favors the correct object. **(b)** Predicted futures across temporal horizons. Short horizons preserve local interaction details, while longer horizons capture broader state transitions with greater visual uncertainty. **(c)** Success rates across four visual prediction horizons, with temporal scales 1–4 corresponding to $1\times$, $2\times$, $3\times$, and $4\times$ the action horizon, respectively. Performance varies differently with prediction horizon across tasks.

that the most useful temporal scale of future information is task dependent. Together, these results establish prediction horizon as an important yet underexplored design dimension in WAMs.

This horizon-dependent trade-off suggests that no single temporal scale is uniformly optimal for action generation. More importantly, the visual prediction horizon need not remain coupled to the executable action horizon, and future information can instead be organized across temporal scales according to its role in control. Short-term futures can preserve the fine-grained interaction cues needed for precise execution, while longer-term futures expose broader task evolution and provide guidance about which interaction to pursue. When short-term predictions remain compatible with multiple plausible interactions, longer-term predictions can reveal their downstream consequences and help identify the appropriate target. An effective WAM should therefore exploit these complementary roles across temporal scales rather than committing to a single fixed horizon.

Building on this insight, we propose **S**eeing **F**arther to **A**ct **N**ow (**SFAN**), which organizes future information across temporal scales to guide and refine current action generation. SFAN uses a sparsely sampled long-horizon future to provide broader task context and generate the primary action prediction. A complementary short-horizon pathway densely samples nearby futures and predicts a residual correction conditioned on the primary action, preserving fine-grained interaction cues without independently modeling the full action distribution. This division of labor allows long-horizon predictions to guide interaction selection while short-horizon predictions refine its execution. By assigning these complementary functions to different temporal scales, SFAN combines task-level foresight with local control precision for reliable manipulation across diverse tasks.

Our contributions are threefold. (1) We identify prediction horizon as an important yet underexplored design dimension in WAMs and characterize its trade-off between capturing longer-term task evolution and preserving the local detail needed for precise interaction. (2) Building on this insight, we propose SFAN, a multi-timescale future modeling framework that separates long-horizon action guidance from short-horizon residual refinement. (3) Extensive experiments in simulation and real-world settings demonstrate that SFAN consistently outperforms existing WAMs, with particularly strong gains on tasks that benefit from longer-horizon future context.

2 Related Work

2.1 World-Action Models

World Action Models (WAMs) have recently emerged as a promising paradigm for incorporating predictive information into robot manipulation. Existing approaches use predicted future visual

states to provide additional context for action generation (Zhen et al., 2024; Zhu et al., 2025; Liang et al., 2025; Ma et al., 2026; Kim et al., 2026; Zhou et al., 2026). Recent work has begun to reconsider how temporal structure is organized in WAMs. WALL-WM (Li et al., 2026) replaces fixed action chunks with event-grounded temporal segments, while AHA-WAM (Cai et al., 2026) decouples world prediction from action execution through asynchronous temporal modeling. Ctrl-World (Guo et al., 2026) studies controllable long-horizon imagination for policy evaluation and improvement, and ABot-M0.5 (Chen et al., 2026) introduces intermediate latent actions to connect coarse video prediction with fine-grained control. These approaches modify the temporal organization of world prediction or its interaction with action generation, but do not explicitly examine prediction horizon as a design dimension for current action generation. Our work focuses on this dimension and studies how future information at different temporal scales affects manipulation behavior.

2.2 Temporal Horizons in Action Generation

Action-chunking policies predict temporally extended action sequences to improve temporal consistency, but the executed portion of each chunk determines how frequently the policy receives closed-loop feedback (Chi et al., 2024; Liu et al., 2025). This creates a trade-off between longer execution horizons, which favor smoother open-loop behavior, and shorter horizons, which allow more frequent replanning. Recent work has therefore explored adaptive action horizons. AutoHorizon (Wang et al., 2026) dynamically determines the execution horizon for each action chunk, while Mixture of Horizons (Jing et al., 2026) combines predictions across multiple action horizons. Related methods further adjust replanning or execution according to uncertainty, interaction phase, or task progress (Zhao et al., 2026b; Nie et al., 2026). These works demonstrate that the appropriate temporal horizon for action execution can vary throughout manipulation. However, they primarily adapt temporal structure on the action side, while the visual prediction horizon in WAMs remains comparatively underexplored. Our work focuses on this dimension and studies how prediction horizon shapes the future information available for action generation.

3 Methodology

Overview. We propose **S**eeing **F**arther to **A**ct **N**ow (SFAN), a world-action modeling framework that integrates long-horizon future guidance with short-horizon residual refinement for robot manipulation, as illustrated in Figure 2. Within a shared video generation model, separate temporal modules produce predictive representations at different horizons. The long-horizon pathway uses temporally extended future information to generate the primary action prediction with broader task context, while the short-horizon pathway uses nearby future information to predict a residual correction conditioned on the primary action. This asymmetric design assigns long-range prediction to task-level guidance and short-range prediction to fine-grained local refinement.

3.1 Preliminaries

World Action Models combine a video generation model with an action model that uses predictive visual information for robot control (Li et al., 2025; Kim et al., 2026; Ma et al., 2026). Given the current observation $\mathbf{o}_s$ and language instruction c, the video model Φ_{vid} predicts future visual evolution. An intermediate predictive representation $\mathbf{F}$ is extracted from the video generation process and used as visual context for action generation. In our formulation, the video model is trained with a denoising objective $\mathcal{L}_{\text{video}}$. Conditioned on $\mathbf{F}$ and c, the action model predicts an action chunk $\mathbf{A} = [\mathbf{a}_s, \dots, \mathbf{a}_{s+L-1}] \in \mathbb{R}^{L \times d_a}$, where L is the action horizon and d_a is the action dimension.

Action generation is formulated with conditional flow matching (Lipman et al., 2023) between the expert action distribution and an isotropic Gaussian prior. During training, we sample a ground-truth action chunk $\mathbf{A}$, Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and a flow timestep $\tau \sim \mathcal{U}(0, 1)$. The intermediate action state follows the linear probability path $\mathbf{X}_\tau = (1-\tau)\mathbf{A} + \tau\boldsymbol{\epsilon}$, with target velocity $\mathbf{v}^* = \boldsymbol{\epsilon} - \mathbf{A}$. Conditioned on the predictive representation $\mathbf{F}$ and instruction c, the velocity field v_θ is trained to match the target velocity using

$$\mathcal{L}_{\text{act}} = \mathbb{E}_{\mathbf{A},\boldsymbol{\epsilon},\tau}\left[\left\| v_\theta(\mathbf{X}_\tau, \tau; \mathbf{F}, c) - (\boldsymbol{\epsilon} - \mathbf{A}) \right\|_2^2\right]. \tag{1}$$

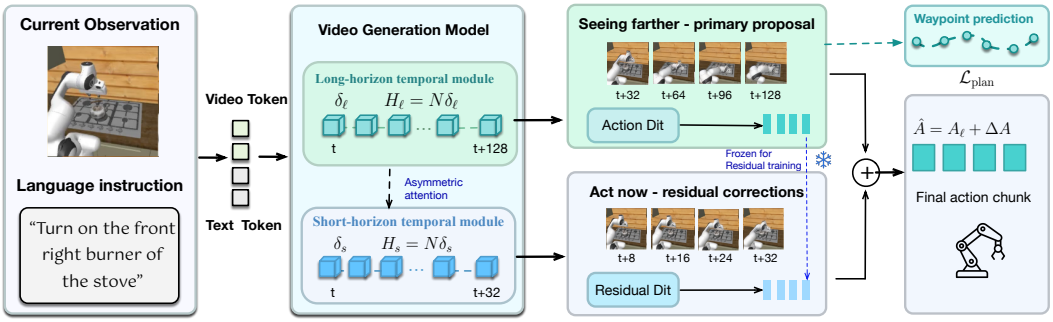

Figure 2: **Overview of SFAN.** Given the current visual observation and language instruction, the video generation model produces long- and short-horizon predictive representations using the same number of future tokens but different temporal strides. The sparsely sampled long-horizon representation captures broader task progression and conditions the action model to generate a primary action chunk. The densely sampled short-horizon representation preserves near-term interaction details and subsequently predicts a residual correction conditioned on the primary action. The final action chunk combines the primary action prediction with the short-horizon residual, providing long-horizon guidance with local refinement. Decoded future frames are shown for visualization.

The video and action models are jointly optimized with

$$\mathcal{L} = \mathcal{L}_{\text{video}} + \lambda_{\text{act}} \mathcal{L}_{\text{act}}, \tag{2}$$

where λ_{act} balances the two objectives. Gradients from $\mathcal{L}_{\text{act}}$ are propagated through $\mathbf{F}$ into the video model, allowing the predictive representation to be jointly shaped by future modeling and action generation.

3.2 Long-Horizon Predictive Guidance

The long-horizon pathway generates the primary action prediction using temporally extended future information. Given the current observation and language instruction, the long-horizon temporal module predicts N future visual tokens with a temporal stride δ_ℓ, yielding a prediction horizon $H_\ell = N\delta_\ell$. Keeping N fixed while increasing δ_ℓ extends temporal coverage independently of the action horizon, allowing farther-future information to guide the same action chunk. The resulting predictive representation $\mathbf{F}_\ell$ provides the broader future context for primary action generation.

To explicitly supervise long-range task progression, the action model predicts a set of future waypoints at selected temporal offsets. Let $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_K]$ denote the waypoint targets extracted from the expert trajectory, where $\mathbf{p}_k$ corresponds to the target at the k-th predefined future offset, and let $\hat{\mathbf{P}}$ denote the corresponding prediction. These waypoint targets provide sparse supervision of task progression across the extended prediction horizon, including stages beyond those covered by the current action chunk. We supervise waypoint prediction with

$$\mathcal{L}_{\text{plan}} = \mathbb{E}\left[\left\|\hat{\mathbf{P}} - \mathbf{P}\right\|_2^2\right]. \tag{3}$$

Waypoint prediction is used only as an auxiliary training signal and shares the representation used for primary action generation. This shared representation allows long-range task-progress supervision to directly shape the features used for action generation throughout policy learning. The action model is therefore encouraged to retain future information that supports both long-range task progression and current action generation.

Conditioned on $\mathbf{F}_\ell$ and c, the action model generates the primary action chunk using the flow-matching formulation in Eq. 1, with objective $\mathcal{L}_{\text{act}}^\ell$. The long-horizon temporal module and primary action model are jointly optimized with

$$\mathcal{L}_\ell = \mathcal{L}_{\text{video}}^\ell + \lambda_{\text{act}} \mathcal{L}_{\text{act}}^\ell + \lambda_{\text{plan}} \mathcal{L}_{\text{plan}}. \tag{4}$$

The resulting action chunk forms the primary action prediction of SFAN and is subsequently refined using short-horizon future information.

3.3 SHORT-HORIZON RESIDUAL REFINEMENT

The short-horizon pathway refines the primary action prediction using densely sampled near-term future information. Because the long-horizon prediction can lose fine-grained interaction changes that are important for precise local control, the short-horizon temporal module predicts the same number N of future visual tokens using a smaller temporal stride $\delta_s < \delta_\ell$, resulting in a prediction horizon $H_s = N\delta_s < H_\ell$. Given the current observation $\mathbf{o}_s$ and language instruction c, this module produces the predictive representation $\mathbf{F}_s$ that preserves finer-grained motion and interaction details than $\mathbf{F}_\ell$ over the near-term horizon.

Let $\hat{\mathbf{A}}_\ell \in \mathbb{R}^{L\times d_a}$ denote the primary action chunk sampled by the long-horizon pathway. Conditioned on $\mathbf{F}_s$, c, and $\hat{\mathbf{A}}_\ell$, the refinement module models the residual action distribution using conditional flow matching and generates a residual action chunk $\Delta\hat{\mathbf{A}}_s$. The final action prediction is obtained as

$$\hat{\mathbf{A}} = \hat{\mathbf{A}}_\ell + \Delta\hat{\mathbf{A}}_s. \tag{5}$$

Rather than independently generating another full action chunk, the short-horizon pathway models the correction to the primary action. Conditioning on $\hat{\mathbf{A}}_\ell$ allows the refinement module to account for the action already proposed by the long-horizon pathway, while $\mathbf{F}_s$ provides the local interaction information needed for local adjustment. This residual formulation preserves the long-horizon pathway as the primary action generator and specializes the short-horizon pathway for local refinement.

The short-horizon temporal module is trained with a video denoising objective $\mathcal{L}^s_{\text{video}}$, which supervises densely sampled near-term future dynamics. During residual training, the long-horizon pathway is frozen, while the short-horizon temporal module and refinement module are optimized. Freezing the long-horizon pathway preserves the guidance learned from extended future information and prevents residual training from altering the primary action generator. The detached primary action is denoted by $\bar{\mathbf{A}}_\ell = \text{sg}[\hat{\mathbf{A}}_\ell]$, where $\text{sg}[\cdot]$ is the stop-gradient operator that helps to prevent residual training from modifying the long-horizon pathway. Given the ground-truth action chunk $\mathbf{A}$, we define the target residual as $\mathbf{R} = \mathbf{A} - \bar{\mathbf{A}}_\ell$.

We formulate residual generation using conditional flow matching. Given Gaussian noise $\boldsymbol{\epsilon}_s \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and a flow timestep $\tau \sim \mathcal{U}(0, 1)$, the intermediate residual state follows $\mathbf{Y}_\tau = (1-\tau)\mathbf{R} + \tau\boldsymbol{\epsilon}_s$, with target velocity $\mathbf{u}^*_s = \boldsymbol{\epsilon}_s - \mathbf{R}$. Conditioned on $\mathbf{F}_s$, c, and $\bar{\mathbf{A}}_\ell$, the residual velocity field r_ϕ is trained with

$$\mathcal{L}_{\text{res}} = \mathbb{E}\left[\left\| r_\phi(\mathbf{Y}_\tau, \tau; \mathbf{F}_s, c, \bar{\mathbf{A}}_\ell) - \mathbf{u}^*_s \right\|_2^2\right]. \tag{6}$$

The short-horizon pathway is optimized with

$$\mathcal{L}_s = \mathcal{L}^s_{\text{video}} + \lambda_{\text{res}}\mathcal{L}_{\text{res}}, \tag{7}$$

where λ_{res} balances near-term future modeling and residual action learning. The short-horizon temporal module and refinement module are jointly optimized by $\mathcal{L}_s$. By modeling the residual action distribution, the refinement pathway focuses on corrections to the sampled primary action.

3.4 IMPLEMENTATION DETAILS

Model Architecture. Our video generation model is adapted from Cosmos-Predict2-2B (NVIDIA, 2025) and jointly models long- and short-horizon futures within a shared backbone. Separate temporal modules are used for the two horizons, allowing each to learn horizon-specific future information while remaining part of the same video generation model. We employ asymmetric attention between the two temporal streams: the short-horizon stream can access long-horizon context, while the long-horizon stream remains independent of short-horizon information. The primary action generator and residual refinement module adopt the same action DiT architecture adapted from GR00T-N1.5 (NVIDIA et al., 2025), with separate parameters. The primary action generator predicts the long-horizon action chunk $\hat{\mathbf{A}}_\ell$, while the refinement module predicts the residual correction $\Delta\hat{\mathbf{A}}_s$ conditioned on $\mathbf{F}_s$, c, and $\hat{\mathbf{A}}_\ell$. The final action is given by $\hat{\mathbf{A}} = \hat{\mathbf{A}}_\ell + \Delta\hat{\mathbf{A}}_s$.

Training Strategy. The long- and short-horizon pathways are optimized in two stages. We first train the long-horizon temporal module together with the primary action generator using the long-horizon video, action, and planning objectives. The trained long-horizon pathway is then frozen, and the short-horizon temporal module and refinement module are optimized using the original training

Table 1: **Results on RoboCasa.** We compare against state-of-the-art methods using success rates computed over 50 rollouts per task. The best, second-best, and third-best results are highlighted.

Category	Task	DP-ResNet	DP-CLIP	GR00T	FPV	DP-VLA	UVA	Video-Policy	Robo-Tales	Cosmos-Policy	Ours
Pick and Place	PnPCabToCounter	0.06	0.00	0.20	0.10	0.10	0.26	0.16	0.44	0.28	0.46
	PnPMicrowaveToCounter	0.06	0.06	0.16	0.12	0.18	0.18	0.16	0.26	0.36	0.46
	PnPCounterToStove	0.00	0.02	0.24	0.04	0.22	0.16	0.52	0.54	0.56	0.64
	PnPCounterToSink	0.14	0.08	0.10	0.08	0.30	0.16	0.32	0.50	0.62	0.58
	PnPCounterToCab	0.06	0.02	0.36	0.14	0.32	0.18	0.38	0.44	0.64	0.50
	PnPSinkToCounter	0.10	0.22	0.33	0.30	0.56	0.38	0.52	0.56	0.76	0.68
	PnPStoveToCounter	0.02	0.06	0.29	0.26	0.62	0.24	0.64	0.66	0.62	0.64
	PnPCounterToMicrowave	0.06	0.02	0.13	0.10	0.56	0.10	0.40	0.44	0.36	0.40
Doors	OpenSingleDoor	0.42	0.32	0.59	0.74	0.42	0.54	0.78	0.80	0.74	0.88
	OpenDoubleDoor	0.70	0.82	0.15	0.92	0.80	0.90	0.92	0.94	0.88	1.00
	CloseDoubleDoor	0.78	0.84	0.75	0.78	0.84	0.76	0.90	0.94	0.90	0.92
	CloseSingleDoor	0.78	0.48	0.83	0.84	1.00	0.88	0.98	0.98	0.98	0.94
Drawers	OpenDrawer	0.64	0.60	0.79	0.72	0.66	0.28	0.40	0.80	0.94	0.86
	CloseDrawer	0.82	0.96	0.99	0.94	1.00	0.72	0.88	1.00	0.98	1.00
Twisting Knobs	TurnOffStove	0.16	0.08	0.27	0.20	0.16	0.14	0.06	0.10	0.16	0.28
	TurnOnStove	0.38	0.28	0.56	0.66	0.64	0.50	0.34	0.40	0.54	0.66
Turning Levers	TurnOffSinkFaucet	0.68	0.70	0.73	0.78	0.72	0.64	0.68	0.78	0.72	0.88
	TurnOnSinkFaucet	0.66	0.66	0.63	0.70	0.56	0.62	0.80	0.84	0.64	0.78
	TurnSinkSpout	0.62	0.26	0.53	0.52	0.90	0.64	0.26	0.32	0.70	0.70
Pressing Buttons	CoffeePressButton	0.76	0.68	0.85	0.90	0.86	0.84	0.92	0.96	0.92	0.98
	TurnOnMicrowave	0.68	0.88	0.78	0.68	0.84	0.94	0.86	0.96	0.76	0.82
	TurnOffMicrowave	0.62	1.00	0.71	0.96	0.86	0.96	0.94	0.96	1.00	0.84
Insertion	CoffeeSetupMug	0.10	0.12	0.23	0.16	0.30	0.20	0.18	0.20	0.36	0.32
	CoffeeServeMug	0.44	0.60	0.73	0.48	0.64	0.78	0.74	0.60	0.54	0.44
Avg. Success		0.41	0.41	0.50	0.51	0.59	0.50	0.57	0.64	0.66	0.69

set together with additional trajectories, providing broader state coverage for residual learning. For each training sample, the frozen primary policy performs complete flow-based sampling to obtain $\hat{\mathbf{A}}_\ell$, from which the residual target $\mathbf{A} - \hat{\mathbf{A}}_\ell$ is constructed. This staged training preserves the long-horizon action behavior while allowing the short-horizon pathway to specialize in local refinement.

Inference. At inference, the shared video generation model jointly produces the long- and short-horizon predictive representations $\mathbf{F}_\ell$ and $\mathbf{F}_s$. The primary action generator first predicts the long-horizon action chunk $\hat{\mathbf{A}}_\ell$ from $\mathbf{F}_\ell$. The refinement module then uses $\mathbf{F}_s$, c, and $\hat{\mathbf{A}}_\ell$ to predict the residual correction $\Delta\hat{\mathbf{A}}_s$, without rerunning the video generation model. The refined action chunk $\hat{\mathbf{A}} = \hat{\mathbf{A}}_\ell + \Delta\hat{\mathbf{A}}_s$ is executed in a receding-horizon manner.

4 Experiments

In this section, we evaluate SFAN across standard and composite simulated manipulation benchmarks as well as real-world robot tasks. We compare against task-specific visuomotor policies, generalist vision-language-action models, and world-model-based methods. We further conduct ablation studies to examine the complementary roles of long- and short-horizon future information and different temporal fusion strategies.

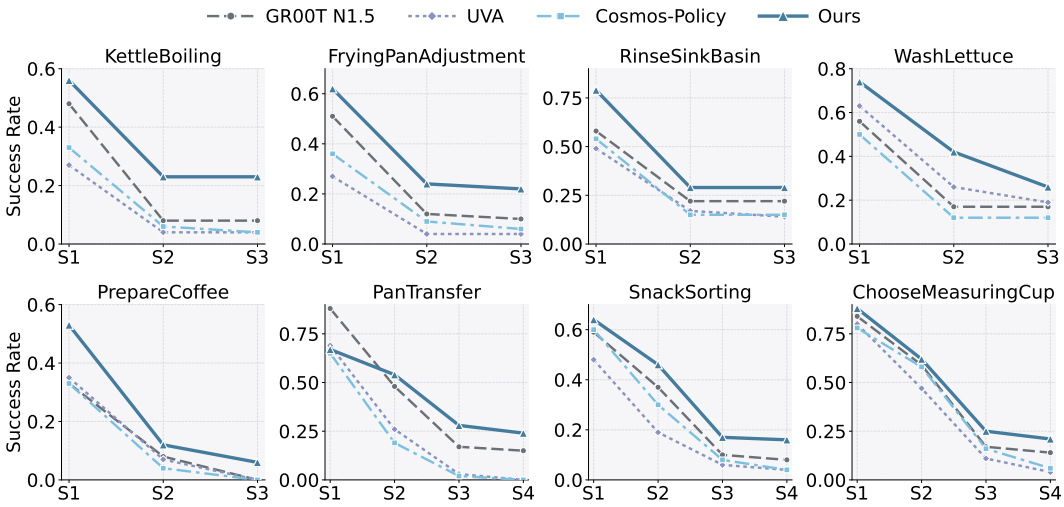

Figure 3: **Stage-wise performance on composite RoboCasa tasks.** We report success rates across the constituent stages of each task. SFAN maintains stronger performance as task execution progresses and achieves higher success rates at later stages across most tasks.

4.1 Simulation Experiments

We evaluate SFAN on two RoboCasa-based simulation benchmarks: the standard RoboCasa benchmark (Nasiriany et al., 2024) and composite tasks from RoboCasa365 (Nasiriany et al., 2026). For RoboCasa, we evaluate on 24 household manipulation tasks using the 7-DoF Franka Panda manipulator. The tasks cover diverse object manipulation and articulated-object interactions in visually rich household environments. For each task, we train on 50 demonstration trajectories. For RoboCasa365, we evaluate on eight composite tasks: *KettleBoiling*, *FryingPanAdjustment*, *RinseSinkBasin*, *WashLettuce*, *PrepareCoffee*, *PanTransfer*, *SnackSorting*, and *ChooseMeasuringCup*. These tasks consist of multi-stage household activities that require coordinated execution of manipulation skills over extended task sequences. For each task, we train on 100 demonstration trajectories. For both benchmarks, we evaluate over 50 trials under the same random seeds for fair comparison.

Baselines. We compare SFAN against three categories of methods that cover representative policy paradigms for robot manipulation. The first includes task-specific visuomotor policies, including DP-ResNet, DP-CLIP (Chi et al., 2024), and FPV (Donat et al., 2025), which directly map visual observations to actions without explicitly modeling future scene evolution. The second consists of generalist vision-language-action policies, including GR00T and DP-VLA (Han et al., 2024). We use GR00T-N1 (NVIDIA et al., 2025) for the standard RoboCasa benchmark and GR00T-N1.5 (Bjorck et al., 2025) for the composite and real-world evaluations. The third includes video-conditioned and world-model-based policies, including UVA (Li et al., 2025), Video Policy (Liang et al., 2025), RoboTales (Gani et al., 2026), and Cosmos Policy (Kim et al., 2026), which incorporate predicted future visual information or learned environment dynamics into action generation. Together, these baselines allow us to compare SFAN against both direct visuomotor policies and approaches that explicitly exploit future predictive information.

Results. Table 1 reports the results on the standard RoboCasa benchmark. SFAN achieves the highest average success rate of 69%, outperforming Cosmos Policy at 66% and RoboTales at 64%. Beyond the overall average, SFAN consistently ranks among the top-performing methods across most tasks, indicating strong performance across diverse manipulation settings. In particular, SFAN achieves 100% on *OpenDoubleDoor*, 66% on *TurnOnStove*, and 88% on *TurnOffSinkFaucet*, while also performing strongly on several pick-and-place tasks, including *PnPCabToCounter*, *PnPMicrowaveToCounter*, and *PnPCounterToStove*. These results indicate that combining long-horizon predictive guidance with short-horizon residual refinement is effective across tasks with different temporal and interaction requirements.

Figure 3 reports stage-wise performance on the eight RoboCasa365 composite tasks. These tasks require multiple sequential interactions, so errors at an early stage can alter the scene state and

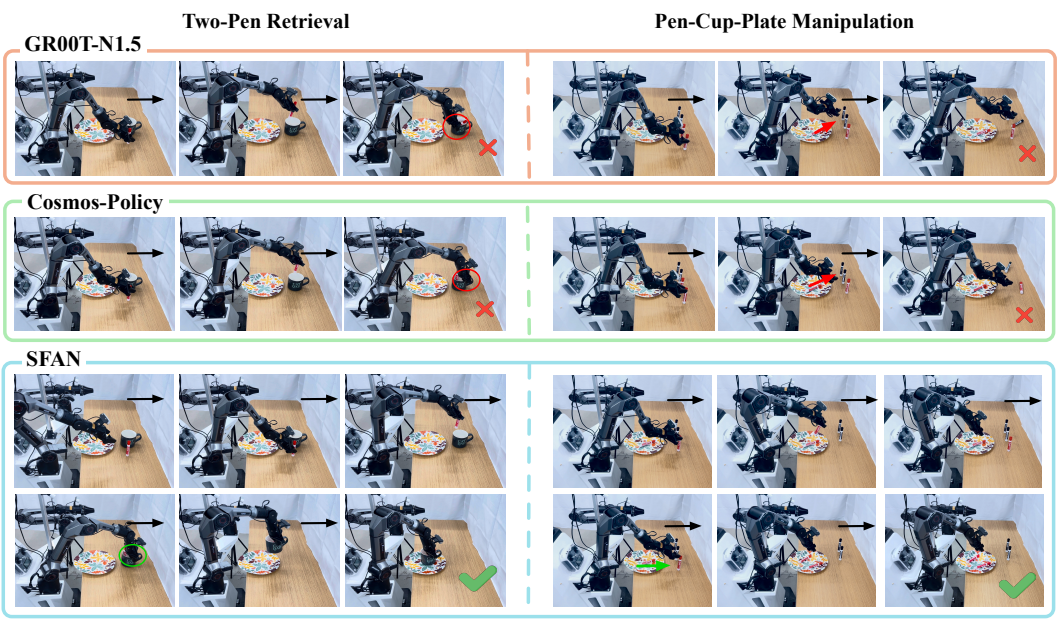

Figure 4: **Qualitative comparison on real-world manipulation tasks.** Baseline policies either fail at intermediate interactions or select incorrect targets, while SFAN completes the full task sequence.

prevent the remaining task from being completed. SFAN achieves higher stage-wise success rates across the majority of tasks, with improvements already evident at the first interaction stage. This suggests that long-horizon predictive guidance helps the policy identify task-relevant interaction targets early in the rollout. As execution progresses, SFAN maintains higher completion rates at later stages, while the baseline methods exhibit larger performance drops across successive stages. This indicates that SFAN more reliably preserves task progression over extended interaction sequences.

The stage-wise advantage also translates into stronger full-task performance. Averaged over all eight tasks, SFAN achieves a full-task success rate of 20.75%, compared with 11.75% for GR00T-N1.5, 9.25% for Cosmos-Policy, and 6.50% for UVA. Together with the standard RoboCasa results, these findings support the complementary roles of the two temporal pathways: long-horizon future information provides guidance for early target selection and subsequent task progression, while short-horizon refinement preserves the local interaction accuracy required to complete each stage.

4.2 Real-World Experiments

Setup. We further evaluate SFAN on an AgileX robot platform in a laboratory environment. We consider two multi-stage manipulation tasks, *Two-Pen Retrieval* and *Pen–Cup–Plate Manipulation*, and compare against GR00T-N1.5 (NVIDIA et al., 2025) and Cosmos-Policy (Kim et al., 2026). In *Two-Pen Retrieval*, the robot sequentially retrieves two red pens from two red and two black pens, with full success requiring both targets to be retrieved without grasping a distractor. In *Pen–Cup–Plate Manipulation*, the robot places a red pen into a cup and then places the cup onto a plate, with full success requiring the pen to remain inside the cup and the cup to be stably placed on the plate. For each task, we collect 100 demonstration trajectories for training and evaluate each method over 15 real-world rollouts. All methods are evaluated under the same initial configurations, and we report both stage-wise and full-task success rates.

Results. Table 2 shows that SFAN achieves the highest success rate at every evaluated stage on both tasks. On *Two-Pen Retrieval*, SFAN achieves 73% and 47% success on the first and second retrieval stages, respectively, resulting in a 40% full-task success rate, compared with 13% for GR00T-N1.5 and 33% for Cosmos-Policy. On *Pen–Cup–Plate Manipulation*, SFAN achieves 87% success for pen grasping and 47% for cup grasping, resulting in a 33% full-task success rate, compared with 7% for GR00T-N1.5 and 13% for Cosmos-Policy.

Figure 4 further illustrates the differences in execution behavior. On *Two-Pen Retrieval*, the baselines can deviate toward an incorrect target after the first retrieval, whereas SFAN more consistently

Table 2: **Real-world results.** We report both stage-wise and full-task success rates.

Two-Pen Retrieval			
Method	First Red	Second Red	Full
GR00T-N1.5	0.40	0.20	0.13
Cosmos-Policy	0.60	0.33	0.33
SFAN	**0.73**	**0.47**	**0.40**

Pen–Cup–Plate			
Method	Pen Grasp	Cup Grasp	Full
GR00T-N1.5	0.67	0.20	0.07
Cosmos-Policy	0.53	0.13	0.13
SFAN	**0.87**	**0.47**	**0.33**

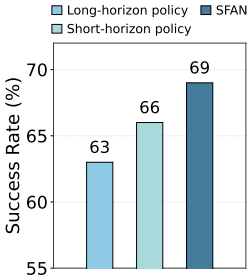

Figure 5: **Temporal horizon ablation.**

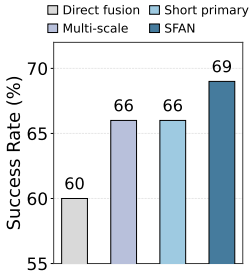

Figure 6: **Fusion strategy ablation.**

approaches the remaining red pen. On *Pen–Cup–Plate Manipulation*, the baselines can exhibit less reliable cup grasping or fail during intermediate interactions, while SFAN more consistently maintains the task sequence through grasping and final placement. Together, the quantitative results show that long-horizon guidance supports task progression across stages, while short-horizon refinement preserves the local precision required for grasping and placement.

4.3 Ablations and Analysis

Complementary Roles of Long- and Short-Horizon Prediction. We isolate the contribution of each temporal pathway by comparing the short-horizon policy, the long-horizon policy, and the full SFAN in Figure 5. The short- and long-horizon policies achieve average success rates of 66% and 63%, respectively. Short-horizon prediction preserves fine-grained interaction details but provides limited guidance for broader task progression, whereas long-horizon prediction captures extended task context at the cost of local motion and contact precision. This gap indicates that neither temporal scale alone provides all the information required for effective action generation. Combining the two pathways improves the average success rate to 69%, outperforming the short- and long-horizon variants by 3 and 6 percentage points. This improvement indicates that long-range guidance and local refinement provide complementary information for action generation.

Fusion Strategy Ablation. We further compare different strategies for combining long- and short-horizon predictive information, as shown in Figure 6. Directly conditioning the action model on both representations reduces the average success rate to 60%, suggesting interference between the two temporal scales. A multi-scale temporal representation reaches approximately 66%, remaining comparable to the short-horizon baseline. We also evaluate a short-primary design, where short-horizon prediction drives the base action and long-horizon information serves as auxiliary context. This variant provides little additional gain, suggesting that long-range guidance is less effective when it does not directly shape the primary action. These comparisons indicate that performance depends not only on which temporal scales are available, but also on how they contribute to action generation. In contrast, SFAN assigns distinct roles to the two pathways: long-horizon information determines the primary action prediction, while short-horizon information provides residual refinement. This asymmetric formulation achieves the highest average success rate of 69%, supporting the use of long-range context for action guidance and short-range information for local correction.

5 Conclusion

We study the visual prediction horizon as an important design dimension in world-action models for robot manipulation. Our analysis reveals a horizon-dependent trade-off between capturing longer-term task evolution and preserving local interaction detail. We further show that coupling the visual prediction horizon to the action chunk can restrict the future information available for action generation. Building on this insight, we propose **S**eeing **F**arther to **A**ct **N**ow (SFAN), which organizes future information across temporal scales using long-horizon predictions for primary action generation and short-horizon predictions for residual refinement. Experiments in simulation and real-world settings show that SFAN consistently improves over existing WAMs, highlighting the importance of structuring future information according to its role in action generation.

Ethics Statement

This work studies robot manipulation in simulated environments and controlled laboratory settings. The experiments do not involve human subjects, personal data, or deployment in safety-critical environments. All real-world experiments are conducted using robotic platforms under controlled conditions. We do not identify additional ethical concerns beyond the standard safety considerations associated with robotic manipulation research.

Reproducibility Statement

We provide the model architecture, training procedure, temporal configurations, and objective functions in Section 3 and Appendix A. Experimental settings, benchmark tasks, training data sizes, and evaluation protocols are described in Section 4. Additional task descriptions, ablations, and analyses are provided in Appendices B and C. These details are intended to support reproduction of the proposed method and the reported experiments.

AI Use Statement

In this work, we used generative AI tools for providing feedback on research methodology and for assisting in the implementation of baselines with unreleased code. We did not use generative AI tools for generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing research methodology or experiments, assisting with translation, cleaning or reformatting datasets, supporting qualitative or thematic data analysis, or interpreting results. Additionally, we used generative AI tools for sourcing and searching for information and for providing feedback to improve the readability of the paper. We have reviewed all AI-assisted work. We thoroughly checked the AI-generated feedback and carefully verified the corresponding code. We take responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

References

Johan Bjorck, Valts Blukis, Fernando Castañeda, et al. GR00T N1.5: An improved open foundation model for generalist humanoid robots. NVIDIA Research, June 2025.

Jisong Cai, Long Ling, Shiwei Chu, Zhongshan Liu, Jiayue Kang, Zhixuan Liang, Wenjie Xu, Yinan Mao, Weinan Zhang, Xiaokang Yang, Ru Ying, Ran Zheng, and Yao Mu. Aha-wam:asynchronous horizon-adaptive world-action modeling with observation-guided context routing, 2026. URL https://arxiv.org/abs/2606.09811.

Ronghan Chen, Yandan Yang, Zuojin Tang, Dongjie Huo, Tong Lin, Haoning Wu, Haoyun Liu, Yuzhi Chen, Lulu Zheng, Botai Yuan, Tianlun Li, Mingxin Wang, Dekang Qi, Bin Hu, Wei Mei, Yuze Xuan, Haolong Yang, Yanqing Zhu, Mu Xu, Zhiheng Ma, and Xinyuan Chang. Abot-m0.5: Unified mobility-and-manipulation world action model, 2026. URL https://arxiv.org/abs/2607.00678.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion, 2024. URL https://arxiv.org/abs/2303.04137.

A. Donat, X. Jia, X. Huang, A. Taranovic, D. Blessing, G. Li, H. Zhou, H. Zhang, R. Lioutikov, and G. Neumann. Towards fusing point cloud and visual representations for imitation learning. *arXiv preprint arXiv:2502.12320*, 2025.

Hanan Gani, Tejal Kulkarni, Madhoolika Chodavarapu, Nicklas Hansen, and Manmohan Chandraker. Robotales: Learning reasoning-guided robot policies via task-aligned simulated futures, 2026. URL https://arxiv.org/abs/2607.06018.

Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-world: A controllable generative world model for robot manipulation, 2026. URL https://arxiv.org/abs/2510.10125.

ByungOk Han, Jaehong Kim, and Jinhyeok Jang. A dual process vla: Efficient robotic manipulation leveraging vlm. In *Conference on Robot Learning*, 2024.

Dong Jing, Gang Wang, Jiaqi Liu, Weiliang Tang, Zelong Sun, Yunchao Yao, Zhenyu Wei, Yunhui Liu, Zhiwu Lu, and Mingyu Ding. Mixture of horizons in action chunking, 2026. URL https://arxiv.org/abs/2511.19433.

Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, and Jinwei Gu. Cosmos policy: Fine-tuning video models for visuomotor control and planning, 2026. URL https://arxiv.org/abs/2601.16163.

Shalfun Li, Victor Yao, Charles Yang, Truth Qu, Regis Cheng, Ryan Yu, Howard Lu, Newton Von, Vincent Chen, Yohann Tang, Maeve Zhang, Ellie Ma, Gody Li, Sage Yang, Lorien Shu, J. W. Gao, Ethan Chen, Colin Ye, Yu Sun, Elise Mon, PS Zhang, Neo Li, Lily Li, James Wang, Ping Yang, Chris Pan, Lucy Liang, Hang Su, Roy Gan, Hao Wang, and Qian Wang. Wall-wm: Carving world action modeling at the event joints, 2026. URL https://arxiv.org/abs/2606.01955.

Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model, 2025. URL https://arxiv.org/abs/2503.00200.

Junbang Liang, Pavel Tokmakov, Ruoshi Liu, Sruthi Sudhakar, Paarth Shah, Rares Ambrus, and Carl Vondrick. Video generators are robot policies, 2025. URL https://arxiv.org/abs/2508.00795.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

Yuejiang Liu, Jubayer Ibn Hamid, Annie Xie, Yoonho Lee, Maximilian Du, and Chelsea Finn. Bidirectional decoding: Improving action chunking via guided test-time sampling, 2025. URL https://arxiv.org/abs/2408.17355.

Teli Ma, Jia Zheng, Zifan Wang, Chunli Jiang, Andy Cui, Junwei Liang, and Shuo Yang. Dit4dit: Jointly modeling video dynamics and actions for generalizable robot control. *arXiv preprint arXiv:2603.10448*, 2026.

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots, 2024. URL https://arxiv.org/abs/2406.02523.

Soroush Nasiriany, Sepehr Nasiriany, Abhiram Maddukuri, and Yuke Zhu. Robocasa365: A large-scale simulation framework for training and benchmarking generalist robots, 2026. URL https://arxiv.org/abs/2603.04356.

Junnan Nie, Jiayi Li, Jiachen Zhang, Junyi Lao, Chenghao Liu, Tianle Zhang, Liang Lin, and Songfang Huang. Pace: Phase-aware chunk execution for robot policies with action chunking, 2026. URL https://arxiv.org/abs/2606.00537.

NVIDIA. Cosmos-predict2: World simulation model for physical ai, 2025. URL https://github.com/nvidia-cosmos/cosmos-predict2.

NVIDIA, :, Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. Gr00t n1: An open foundation model for generalist humanoid robots, 2025. URL https://arxiv.org/abs/2503.14734.

Yang Tian, Sizhe Yang, Jia Zeng, Ping Wang, Dahua Lin, Hao Dong, and Jiangmiao Pang. Predictive inverse dynamics models are scalable learners for robotic manipulation, 2024. URL https://arxiv.org/abs/2412.15109.

Haoxuan Wang, Gengyu Zhang, Yan Yan, Ramana Rao Kompella, and Gaowen Liu. Vla knows its limits: Adaptive execution horizons for robot policies, 2026. URL https://arxiv.org/abs/2602.21445.

Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation, 2023. URL https://arxiv.org/abs/2312.13139.

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

Weiheng Zhao, Haoyi Jiang, Xin Shi, Liu Liu, Fan Huang, Zhizhong Su, Wei Sui, and Xinggang Wang. Faster-wam: Efficient inference-time future conditioning for robust world action models, 2026a. URL https://arxiv.org/abs/2608.04404.

Yuchi Zhao, Miroslav Bogdanovic, Arjun Sohal, Liyu Tao, Kourosh Darvish, Alán Aspuru-Guzik, Florian Shkurti, and Animesh Garg. Dynamic execution horizon prediction for chunk-based robot policies, 2026b. URL https://arxiv.org/abs/2606.11408.

Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model, 2024. URL https://arxiv.org/abs/2403.09631.

Pengfei Zhou, Shengcong Chen, Di Chen, Jiaxu Wang, Rongjun Jin, Bingwen Zhu, Yike Pan, Songen Gu, Kuanning Wang, Shufeng Nan, Xingyu Qiu, Chenhao Qiu, Pu Yang, Yunuo Cai, Jianxiong Gao, Yifan Li, Yanwei Fu, Xiangyu Yue, Zhi Chen, and Jianlan Luo. τ_0-wm: A unified video-action world model for robotic manipulation, 2026. URL https://arxiv.org/abs/2606.01027.

Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets, 2025. URL https://arxiv.org/abs/2504.02792.

APPENDIX

A IMPLEMENTATION DETAILS

Video Generation Model. We build the predictive video model on Cosmos-Predict2-2B (NVIDIA, 2025) and retain the pretrained T5 encoder for language conditioning. Multi-view observations are spatially tiled into a single composite frame before being processed by the video model. Within the video generation model, the long- and short-horizon streams use separate horizon-specific temporal parameters while remaining part of a shared architecture. Both streams share the same observation anchor and produce predictive representations F_ℓ and F_s at different temporal scales. Both pathways predict $N = 4$ future visual tokens. The long-horizon pathway uses a temporal stride of $\delta_\ell = 32$, yielding a prediction horizon of $H_\ell = 128$ control steps, while the short-horizon pathway uses $\delta_s = 8$, yielding $H_s = 32$ control steps. We use asymmetric attention between the two streams: the long-horizon representation is formed independently of short-horizon information, whereas the short-horizon stream can additionally attend to long-horizon context. The video model is optimized using the standard latent-space EDM denoising objective.

Action Generation. The primary action generator and residual refinement module use the same action-DiT architecture adapted from GR00T-N1.5 (Bjorck et al., 2025), with separate parameters. Each module contains 16 Transformer layers with interleaved self-attention and cross-attention, and directly consumes intermediate predictive video features together with language embeddings without an additional visual encoder. Intermediate features sampled across the video model are provided to the corresponding cross-attention layers, while learnable bridge tokens aggregate predictive visual context within the action stream. The long-horizon representation F_ℓ conditions the primary action generator to predict an action chunk $\hat{\mathbf{A}}_\ell$ with horizon $L = 32$. The short-horizon representation F_s is then used by the refinement module, together with the language instruction and $\hat{\mathbf{A}}_\ell$, to predict a residual correction $\Delta\hat{\mathbf{A}}_s$. The final action is obtained as $\hat{\mathbf{A}} = \hat{\mathbf{A}}_\ell + \Delta\hat{\mathbf{A}}_s$. For the auxiliary planning objective, we use $K = 8$ waypoint targets defined as future end-effector poses uniformly sampled over the long-horizon prediction range. Training proceeds in two stages: the long-horizon pathway and primary action generator are optimized first, after which they are frozen while the short-horizon pathway and residual refinement module are trained.

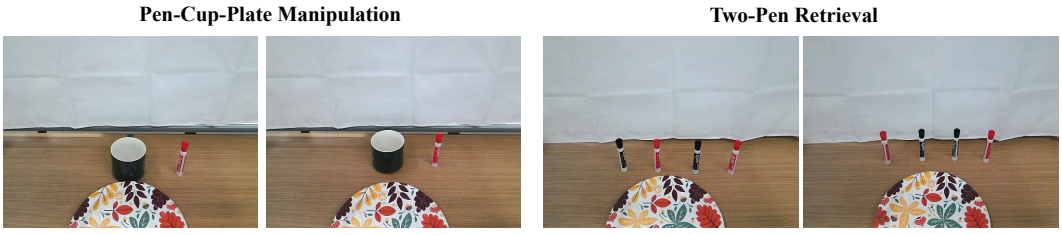

Figure 7: **Examples of the real-world manipulation tasks.** *Two-Pen Retrieval* varies the locations of the red target pens among distractors, requiring target identification and sequential retrieval. *Pen–Cup–Plate Manipulation* varies the initial pen and cup positions and requires precise grasping, insertion, and placement across multiple stages.

B TASK EXAMPLES

Figure 7 illustrates the two real-world manipulation tasks used in our evaluation. The tasks are designed to introduce complementary challenges in target selection, precise interaction, and multi-stage execution.

In *Two-Pen Retrieval*, two red pens and two black distractor pens are placed at varying positions on the table. The robot must sequentially identify and retrieve the two red targets while avoiding the distractors. Changing the locations of the red pens across trials requires the policy to rely on the current observation rather than memorizing a fixed interaction sequence.

In *Pen–Cup–Plate Manipulation*, the initial positions of both the pen and cup are varied across trials. The robot must first grasp the pen and place it into the cup, and then grasp the cup and place it onto

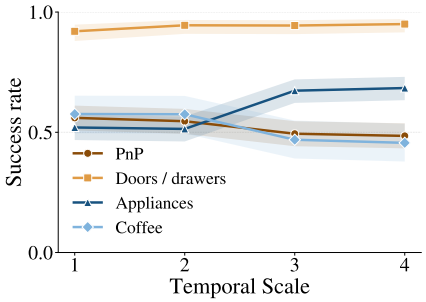

Figure 8: **Category-wise effect of prediction horizon on RoboCasa.** Success rates are averaged within each task category over temporal scales 1–4, corresponding to visual prediction horizons of $1\times$–$4\times$ the action horizon. Different task categories exhibit distinct trends as the prediction horizon increases.

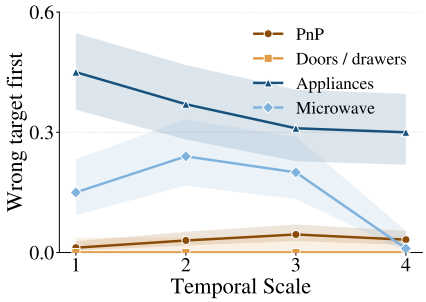

Figure 9: **Effect of prediction horizon on wrong-target failures.** We report the fraction of failed rollouts that first move toward an incorrect interaction target. Longer horizons reduce wrong-target failures for appliance-related tasks, while microwave tasks improve only at sufficiently long horizons.

the plate. This task combines precise grasping and placement with multi-stage task progression, since successful completion of later interactions depends on the outcome of the preceding stages.

C Additional Ablations and Analysis

Category-wise Effect of Prediction Horizon. We further analyze the effect of prediction horizon by grouping RoboCasa tasks according to interaction type and averaging performance within each category. As shown in Figure 8, temporal scales 1–4 correspond to visual prediction horizons of $1\times$–$4\times$ the action horizon, respectively. We restrict the analysis to temporal scales 1–4, since larger prediction horizons would exceed the valid trajectory range for a subset of demonstrations. Appliance-related tasks improve as the prediction horizon increases, consistent with their stronger dependence on downstream task states, such as whether an appliance has been correctly activated or turned off, which become more explicit at longer horizons. In contrast, pick-and-place and coffee-related tasks rely more heavily on precise geometric alignment, grasping, and contact dynamics. As the prediction horizon expands under a fixed number of future tokens, temporal sampling becomes sparser, which can weaken the fine-grained local information needed for these interactions and reduce success rates. Door and drawer manipulation remains comparatively stable across temporal scales, suggesting that these tasks are less sensitive to the trade-off between long-range task context and fine-grained local interaction detail. These results further support the complementary use of long- and short-horizon predictions in SFAN.

Failure Analysis of Target Selection. We further analyze whether prediction horizon affects target selection by measuring, among failed rollouts, the fraction in which the robot first moves toward an incorrect interaction target. As shown in Figure 9, appliance-related tasks exhibit a clear decrease in wrong-target failures as the prediction horizon increases. This suggests that longer-horizon future information better exposes the downstream consequences of the current interaction and helps determine which target should be approached. Microwave-related tasks show a non-monotonic trend: the wrong-target rate initially increases at intermediate horizons and then drops substantially at the longest horizon. One possible explanation is that moderately extending the horizon is still insufficient to disambiguate the correct interaction target, while the sparser temporal sampling can already weaken local action quality. Once the horizon becomes sufficiently long, the predicted future provides clearer information about later task progression, allowing the policy to better determine the appropriate direction for the current action. In contrast, pick-and-place and door/drawer tasks show relatively small changes, suggesting that target selection in these categories is less sensitive to long-range future context.

Effect of prediction density under a fixed horizon. To examine whether the benefit of long-horizon prediction simply results from its temporal sampling pattern, we fix the prediction horizon and vary the number of predicted future tokens. As shown in Table 3, using 4 or 8 future to-

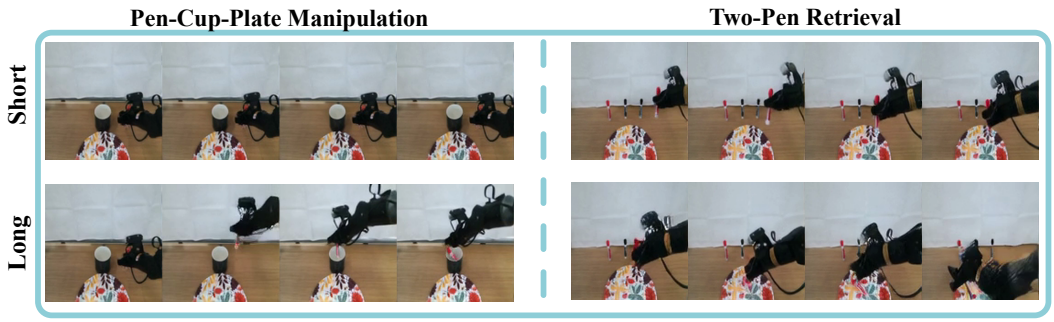

Figure 10: **Qualitative visualization of long- and short-horizon predictions on real-world tasks.** Short-horizon predictions remain closer to the current state and capture local interaction changes, while long-horizon predictions exhibit larger state transitions and advance farther along the task sequence.

Table 3: **Effect of prediction density.** We vary the number of future tokens under a fixed prediction horizon.

Future Tokens	Success Rate (%)
4	69
8	69
12	64
16	58

Table 4: **Inference cost comparison.** We report inference latency per action prediction, peak GPU memory usage, and model parameter count.

Method	Latency (s)	Memory (GB)	Params (B)
SFAN	1.77	7.7	2.52
Cosmos-Policy	0.70	8.9	1.96
UVA	0.28	2.1	0.33

kens achieves the same average success rate of 69%, while increasing the token count to 12 and 16 reduces performance to 64% and 58%, respectively. This result suggests that denser temporal sampling within the same long-range horizon does not necessarily provide more useful information for action generation, as predicting additional intermediate future states may increase modeling difficulty without providing commensurate control-relevant information. Together with our horizon analysis, these results indicate that temporal coverage and prediction density should be considered separately when structuring future information.

Qualitative Visualization of Predicted Futures. We visualize the predicted future frames produced by the long- and short-horizon pathways on the real-world tasks. As shown in Figure 10, the short-horizon predictions remain close to the current observation and mainly capture near-term motion and local interaction changes. In contrast, the long-horizon predictions exhibit substantially larger state transitions, with the robot and manipulated objects progressing farther along the task sequence. This qualitative difference is consistent across both *Pen–Cup–Plate Manipulation* and *Two-Pen Retrieval*, illustrating that the two pathways capture future information at distinct temporal scales. The long-horizon pathway provides a broader view of task progression, while the short-horizon pathway preserves finer local dynamics for action refinement.

Inference Overhead. We further compare the inference cost of SFAN with Cosmos-Policy and UVA in terms of latency, peak GPU memory, and model size. As shown in Table 4, SFAN requires 1.77,s per action prediction, uses 7.7,GB of peak GPU memory, and contains 2.52,B parameters. This improved temporal modeling comes with additional inference latency due to the two temporal pathways and residual refinement. Despite its larger model size, SFAN uses less peak GPU memory than Cosmos-Policy (7.7,GB vs. 8.9,GB).