FUTUREDUET: DECOUPLING OBSERVATION ACCESS FROM FUTURE SUPERVISION IN WORLD ACTION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

World Action Models (WAMs) augment robot action generation with future visual supervision. Existing WAMs commonly fuse main and wrist observations into one visual stream and train both with the same future-video objective, despite their different visual dynamics. A stable main camera reveals scene-level task evolution, whereas wrist cameras move with the end effector, mixing local interaction changes with viewpoint shifts and self-occlusion. These contrasting predictive demands suggest that the two views may benefit from different future objectives. We introduce FutureDuet, which retains both views for control, while allowing each visual stream to receive a different future objective. For the main view, future RGB models task evolution, while interaction masks and robot skeletons focus supervision on task objects and robot motion. For the wrist stream, future latent prediction models short-horizon interaction changes without requiring pixel-level reconstruction. ActionDiT jointly reads the resulting Task State and Interaction State, combining scene-level progress with close-range interaction evidence. All auxiliary prediction modules are training-only, adding no inference overhead. FutureDuet achieves 94.2% clean and 94.1% randomized success on RoboTwin50 and 99.2% average success on LIBERO. The improvements are most pronounced on six RoboTwin50 tasks that require precise interaction, averaging gains of 9.2% and 12.8% over Fast-WAM in clean and randomized settings. Controlled studies further show complementary gains from separating the wrist pathway and designing future supervision separately for the two views.

1 INTRODUCTION

World Action Models (WAMs) augment robot action generation with future visual supervision, leveraging predicted scene evolution to improve policy learning (Bi et al., 2025; Yuan et al., 2026b; Cen et al., 2025b; Kim et al., 2026; Ma et al., 2026; Cheang et al., 2024; Cen et al., 2025a; Zhang et al., 2025; Pai et al., 2025). Fast-WAM further shows that this benefit remains largely when future observations are not generated at deployment (Yuan et al., 2026b). This finding shifts attention from the generation of futures at test time to how future supervision should be designed during training.

For WAMs that use both main and wrist cameras, a common design is to fuse their observations into a shared visual stream and train both views under a uniform future-video objective, as illustrated in Fig. 1(a) (Bi et al., 2025; Yuan et al., 2026b; Cen et al., 2025b). This makes incorporating both views straightforward, but entangles observation access with future supervision: every camera stream available to the policy is forced to predict the same form of future target.

However, the camera's frame of reference changes the prediction problem itself. A main camera provides a relatively stable scene frame in which object relations, robot motion, and task progress can be tracked over time. Wrist cameras instead move with the end effector, bringing relative pose, contact, and object response into close view while also introducing rapid viewpoint shifts and self-occlusion. From the main view, future changes reveal how the task unfolds across the scene; in wrist views, they reflect both the local interaction and the motion of the camera itself. The two views therefore support complementary state representations for control: one tracks scene-level task progress, while the other captures the ongoing close-range interaction. A shared future video

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.

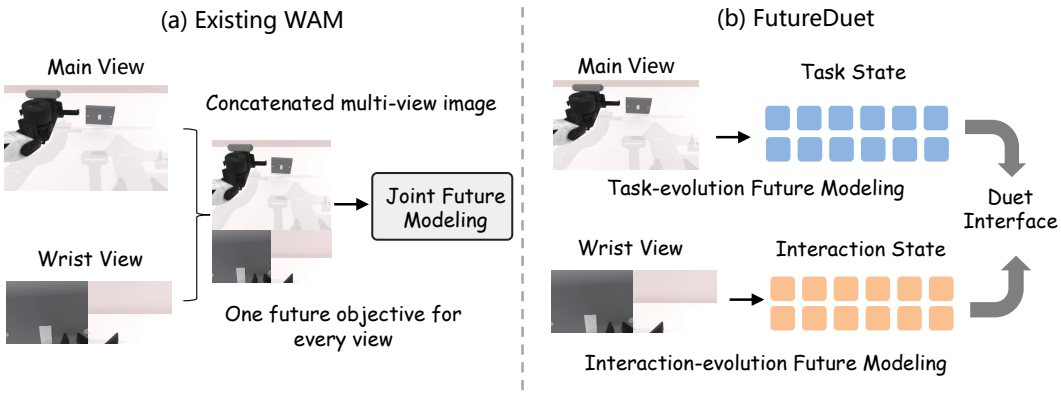

Figure 1: **Decoupling future supervision across views.** (a) A common multi-view WAM design concatenates observations and applies one future objective across views. (b) FutureDuet separately models task evolution from the main view and interaction evolution from wrist views; the resulting states jointly support action generation.

objective nevertheless treats these different visual changes as the same form of prediction target, raising a central question: should each view model the same kind of future?

We therefore introduce FutureDuet (Fig. 1(b)), a WAM that keeps both main and wrist observations available for action generation while allowing their future supervision to be designed separately. For the main view, future RGB prediction captures how the task evolves across the scene, while interaction masks and robot skeletons focus the representation on task objects and robot motion. For the wrist views, future latent prediction models short-horizon interaction changes without requiring pixel-level reconstruction of appearance changes caused by camera motion and self-occlusion. The two streams consequently provide complementary control information: the *Task State* tracks scene-level task progress, while the *Interaction State* captures close-range interaction changes. ActionDiT reads both states jointly to generate actions. All auxiliary prediction modules operate exclusively during training, adding no inference overhead.

FutureDuet achieves 94.2% clean and 94.1% randomized success on RoboTwin50 and 99.2% average success on LIBERO, with its largest RoboTwin50 gains on tasks requiring precise interaction completion. Controlled ablations show gains from wrist-pathway separation; with the deployed architecture fixed, future supervision provides further gains and latent targets perform best. Structured main-view supervision provides complementary gains, while readout diagnostics connect masks and skeletons to object- and robot-related evidence.

Our contributions are:

- We propose FutureDuet, a WAM that decouples observation access from future supervision. It retains main and wrist observations for action generation while allowing their future objectives to be designed separately. The auxiliary prediction modules operate only during training, adding no inference overhead.
- We design two complementary streams for joint control. The main stream learns a *Task State* through future RGB prediction, with masks and skeletons emphasizing task objects and robot motion. The wrist stream learns an *Interaction State* through future latent prediction to capture close-range interaction changes.
- We validate FutureDuet on RoboTwin50 and LIBERO, with the largest RoboTwin50 gains on tasks requiring precise interaction completion. Our controlled wrist-pathway study shows a 13.9-percentage-point improvement over unified RGB modeling, highlighting the benefits of separate wrist modeling and future supervision.

2 Related Work

Vision-Language-Action Policies and World Action Models. Vision-Language-Action policies map visual observations and language instructions to robot actions, drawing on pretrained vision-language representations and diverse robot demonstrations (Brohan et al., 2023; Octo Model Team et al., 2024; Kim et al., 2025b; Black et al., 2025; NVIDIA et al., 2025; Qu et al., 2025; Liu et al., 2025; Song et al., 2025; Wu et al., 2026). World Action Models augment this mapping with visual future modeling, allowing future observations or representations to shape action learning (Du et al., 2023; Zhou et al., 2024; Bharadhwaj et al., 2025; Feng et al., 2025; Jang et al., 2025). Existing methods use this predictive component in different ways. Some generate future observations as intermediate plans for action prediction (Du et al., 2023; Zhou et al., 2024; Bharadhwaj et al., 2025; Liang et al., 2025; Wu et al., 2023), while others jointly model video and action trajectories (Bi et al., 2025; Cen et al., 2025b; Zhu et al., 2025; Won et al., 2025). Fast-WAM further shows that future prediction can serve as an effective training signal without generating future observations during deployment (Yuan et al., 2026b). Future modeling in WAMs therefore serves not only as an inference mechanism, but also as a source of supervision for learning robot policies.

Predictive Targets for Robot Control. The choice of future target determines which aspects of change a policy is trained to model. RGB targets describe future scene appearance and motion (Hu et al., 2025; Yuan et al., 2026b), whereas latent and joint-embedding objectives model future visual representations without reconstructing every pixel (Sun et al., 2026; Zhang et al., 2026). Recent WAMs use compact latent futures or pretrained joint-embedding spaces to capture predictive dynamics (Chen et al., 2026a; Huang et al., 2026; Lin et al., 2026). Concurrent work W2-VLA predicts task-conditioned future wrist latents and exposes the resulting future-aware context to action generation (Pan et al., 2026). Other approaches emphasize action-relevant structure through object masks, point trajectories, or complementary semantic and geometric targets (Yu et al., 2026; Bharadhwaj et al., 2024; Haldar & Pinto, 2025; Yuan et al., 2026a). FutureDuet uses wrist-latent prediction only as training supervision and examines a broader design question: how different future objectives can be assigned across main and wrist streams that are jointly used for action generation.

3 Method

3.1 Overview

At control step t, FutureDuet accepts a language instruction c, proprioception p_t, main-view history $o^m_{\leq t}$, and wrist histories $o^w_{\leq t}$ to generate an action chunk $a = a_{t:t+H-1}$. Both views support control, but their future objectives are designed separately: the main stream models scene-level task evolution, while the wrist stream models short-horizon interaction changes.

As shown in Fig. 2, VideoDiT extracts main-view features $H^m_t = F_m(o^m_{\leq t}, c, p_t)$, and a lightweight online encoder extracts wrist features $H^w_t = E^w_\omega(o^w_{\leq t})$. Learned-query adapters summarize these features into a *Task State* and an *Interaction State*:

$$\begin{aligned} S^{\text{task}}_t &= A_{\text{task}}(Q^{\text{task}}, H^m_t) \in \mathbb{R}^{N_{\text{task}} \times d}, \\ S^{\text{int}}_t &= A_{\text{int}}(Q^{\text{int}}, H^w_t) \in \mathbb{R}^{N_{\text{int}} \times d}. \end{aligned} \tag{1}$$

Each adapter uses learned queries Q to attend to its stream and produce N_{task} or N_{int} tokens of width d. The Task State carries scene context from the instruction and proprioception, while the Interaction State summarizes close-range wrist observations.

The Duet Interface projects and fuses the two states for ActionDiT, which also reads current main-view key/value features K^m_t, V^m_t (Yuan et al., 2026b; Wan et al., 2025; Liang et al., 2024):

$$\begin{aligned} C^{\text{duet}}_t &= A_{\text{duet}}([S^{\text{task}}_t; S^{\text{int}}_t]), \\ \widehat{v}^a_\tau &= \Pi_\eta(a^\tau, \tau; C^{\text{duet}}_t, K^m_t, V^m_t, c, p_t). \end{aligned} \tag{2}$$

Rather than fusing main and wrist observations at early visual stages, FutureDuet defers their combination until each stream has been specialized and summarized into its dedicated state. The Duet Interface projects the Task State and Interaction State into a common conditioning space and supplies the fused tokens to ActionDiT. This interface brings the separately supervised representations

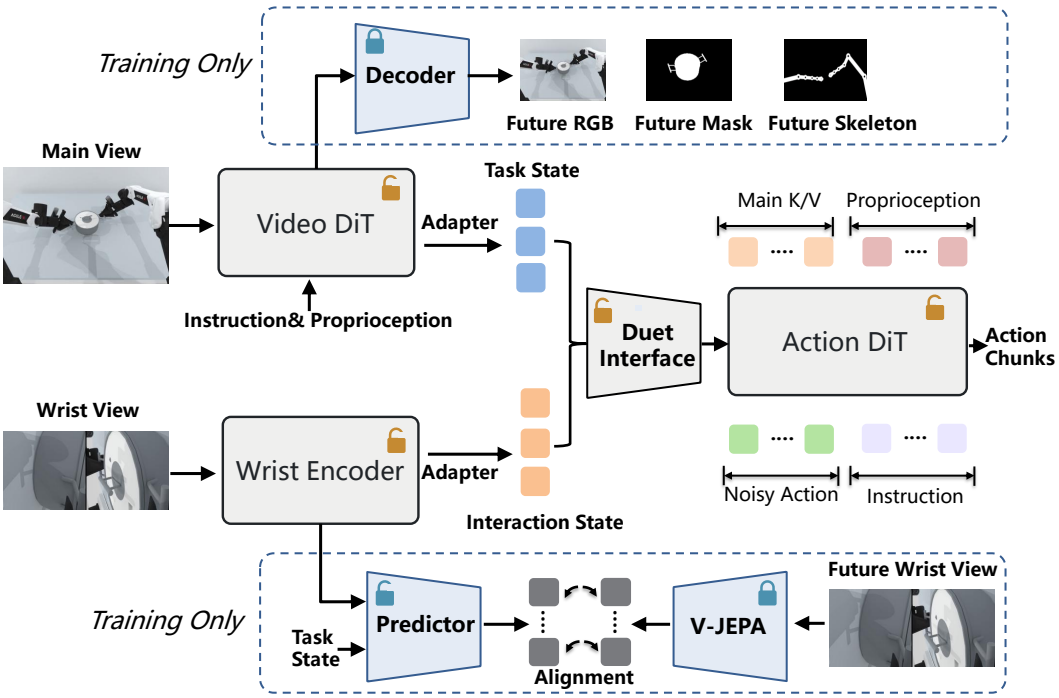

Figure 2: **Overview of FutureDuet.** Different training-only future objectives shape the Task State and Interaction State. ActionDiT reads both through the Duet Interface; auxiliary predictors are removed at deployment.

together for joint action generation. Here $[\,;\,]$ denotes token-wise concatenation, and a^τ is a noisy action chunk at flow time τ. ActionDiT generates actions by integrating its predicted velocity from noise. This joint readout combines scene-level task context with local interaction evidence.

3.2 Task State and Interaction State

Task State from the main view. Future RGB prediction models scene evolution, while interaction masks and robot skeletons focus supervision on task objects and articulated robot motion. The main stream retains the underlying WAM objective: future frames are encoded into VAE latents z^m, and VideoDiT learns to predict their flow velocity under Gaussian noise.

A shared DPT-style decoder reads noise-conditioned intermediate VideoDiT features $\widetilde{H}_t^{m,\tau}$. Separate output heads predict interaction masks and skeleton maps at sampled future offsets δ_k. Through the shared decoder, both auxiliary losses supervise these features, emphasizing object extent and robot configuration alongside the appearance information captured by future RGB prediction. Offline SAM 3 (Carion et al., 2025) masks mark the manipulated object and, when required, its target region; skeleton maps project recorded robot configurations into the calibrated main view. All targets are cached offline. The decoder and output heads are trainable and used only during training; App. A.1 provides details on interaction-mask generation and robot-skeleton projection.

Interaction State from the wrist views. Wrist motion mixes local interaction changes with camera-induced appearance variation. We use future latent prediction to supervise temporal structure without requiring pixel reconstruction. A frozen V-JEPA 2.1 (Mur-Labadia et al., 2026) encoder processes future wrist clips offline, producing cached targets $Z^w_{t,\mathrm{fut}}$.

Wrist histories provide local evidence but may not reveal the task context governing the next interaction. The Task State supplies this context to a shared training-only predictor:

$$\widehat{Z}^w_{t,\mathrm{fut}} = P_\psi(H^w_t, \mathrm{sg}[S^{\mathrm{task}}_t]). \tag{3}$$

The predictor jointly predicts left- and right-wrist latents, using view embeddings to distinguish the cameras. Future actions are outputs of the policy, not inputs to this predictor. The stop-gradient operator sg prevents the wrist objective from updating the main stream through the conditioning connection, while allowing it to train the online wrist encoder through P_ψ. We minimize the mean absolute error $\mathcal{L}_{\text{wrist}} = |\Omega|^{-1} \sum_{i \in \Omega} |\widehat{Z}^w_{t,\text{fut},i} - Z^w_{t,\text{fut},i}|$, where Ω indexes valid latent elements across the batch and both views, excluding padded future target positions.

3.3 Joint Action Readout and Training

Main view RGB prediction and action generation use flow matching. Their noisy inputs are $z^{m,\tau} = (1-\tau)z^m + \tau\epsilon^m$ and $a^\tau = (1-\tau)a + \tau\epsilon^a$, with independent standard Gaussian noise. The corresponding objectives are

$$\begin{aligned} \mathcal{L}^m_{\text{rgb}} &= \mathbb{E}\big[w(\tau)\|\widehat{v}^m_\tau - (\epsilon^m - z^m)\|_2^2\big], \\ \mathcal{L}_{\text{act}} &= \mathbb{E}\big[w(\tau)\|\widehat{v}^a_\tau - (\epsilon^a - a)\|_2^2\big]. \end{aligned} \tag{4}$$

Here $\widehat{v}^m_\tau$ and $\widehat{v}^a_\tau$ are the VideoDiT and ActionDiT velocity predictions. The two objectives use their respective noise-level samples and scheduler weights, denoted by $\tau \in [0,1]$ and $w(\tau)$ for brevity. Mask prediction uses focal–Tversky losses, and skeleton prediction uses focal–Dice losses. We jointly optimize

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\text{act}} + \lambda_{\text{rgb}}\mathcal{L}^m_{\text{rgb}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} \\ + \lambda_{\text{ske}}\mathcal{L}_{\text{ske}} + \lambda_{\text{wrist}}\mathcal{L}_{\text{wrist}}. \end{aligned} \tag{5}$$

Main-view future losses train VideoDiT, the wrist objective trains the online wrist encoder, and action loss updates both streams through the Duet Interface. At inference, VideoDiT and the wrist encoder form the two states from observed histories, and ActionDiT reads them jointly as in Eq. 2. The VAE decoder, DPT decoder with prediction heads, V-JEPA teacher, and wrist predictor are pruned during evaluation, ensuring that auxiliary future prediction adds no inference overhead. App. A.1 provides detailed training settings.

4 Experiments

Our evaluation first measures the complete model across broad bimanual and single-arm benchmarks, then examines three design questions: **Q1**: Why use a separate wrist pathway? **Q2**: Why design structured prediction targets for the main view? **Q3**: And how does this supervision shape the information read by ActionDiT? We answer these questions with benchmark results, controlled component studies, and frozen-checkpoint diagnostics of attention allocation and feature influence.

4.1 Experimental Setup

Benchmarks. RoboTwin 2.0 (Chen et al., 2026b) contains 50 bimanual tasks in clean and randomized settings and serves as our primary benchmark. We jointly train one model on 50 clean and 500 randomized demonstrations per task, totaling 27,500 demonstrations. This setup matches Fast-WAM in data, backbone, preprocessing, training budget, checkpoint selection, and evaluation. Published baseline values come from the cited benchmark reports; specifically, the UWM result follows the RoboTwin50 evaluation in X-WAM (Guo et al., 2026). LIBERO (Liu et al., 2023) comprises four ten-task, language-conditioned single-arm suites. We use 50 demonstrations per task to evaluate the same design beyond bimanual manipulation.

Controlled studies. The main-view factorial study uses six tasks with 500 randomized demonstrations per task. A separate three-task wrist-pathway study uses the same number of demonstrations and reports three-seed means over 100 randomized evaluation episodes per task and seed. The wrist study fixes main-view supervision to RGB to distinguish the effects of pathway separation and future prediction. The factorial study then tests whether structured main-view targets provide complementary gains under both unified and separate wrist pathways. The action-readout analysis uses another six-task set. App. A.1.2 provides the task lists, protocols, and training settings. Unless stated otherwise, entries are success rates (SR, %); ranked tables mark the best and second-best distinct values with green and yellow backgrounds.

Table 1: LIBERO benchmark comparison across four single-arm suites. Success rates (%); higher is better.

Method	Spatial ↑	Object ↑	Goal ↑	Long ↑	Avg. ↑
π_0 (Black et al., 2025)	96.8	98.8	95.8	85.2	94.1
GR00T-N1 (NVIDIA et al., 2025)	94.4	97.6	93.0	90.6	93.9
$\pi_{0.5}$ (Physical Intelligence et al., 2025)	98.6	98.2	98.0	92.4	96.8
OpenVLA-OFT (Kim et al., 2025a)	97.6	98.4	97.9	94.5	97.1
Motus (Bi et al., 2025)	96.8	99.8	96.6	97.6	97.7
Fast-WAM (Yuan et al., 2026b)	98.2	**100.0**	97.0	95.2	97.6
MaskWAM (Yu et al., 2026)	98.8	**100.0**	98.2	96.4	98.4
LingBot-VA (Li et al., 2026)	98.5	99.6	97.2	**98.5**	98.5
FutureDuet	**99.2**	**100.0**	**99.0**	98.4	**99.2**

Table 2: RoboTwin50 benchmark comparison in clean and randomized settings.

Method	Clean ↑	Rand. ↑
π_0 (Black et al., 2025)	65.9	58.4
X-VLA (Zheng et al., 2025)	72.9	72.8
$\pi_{0.5}$ (Physical Intelligence et al., 2025)	82.7	76.8
UWM (Zhu et al., 2025)	81.7	78.6
Motus (Bi et al., 2025)	88.7	87.0
X-WAM (Guo et al., 2026)	89.8	90.7
Fast-WAM (Yuan et al., 2026b)	91.9	91.8
LingBot-VA (Li et al., 2026)	92.9	91.5
FutureDuet	**94.2**	**94.1**

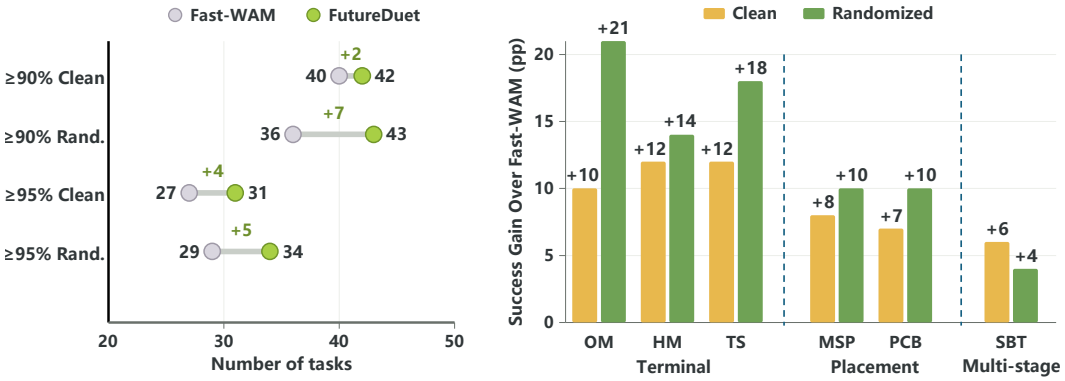

Figure 3: **RoboTwin50 performance varies substantially across tasks.** (a) Number of tasks reaching at least 90% and 95% success under clean and randomized evaluation. (b) Gains over Fast-WAM on the six tasks that remain below 90% for FutureDuet in both settings. Gains are measured in percentage points (pp). OM, HM, TS, MSP, PCB, and SBT denote Open Microwave, Hanging Mug, Turn Switch, Move Stapler Pad, Place Can Basket, and Stack Bowls Three, respectively.

4.2 Overall Performance

LIBERO. Tab. 1 compares FutureDuet with source-reported results across four single-arm suites. FutureDuet achieves 99.2% average success, with the best reported results on Spatial, Object, and Goal, demonstrating its effectiveness beyond the bimanual RoboTwin setting.

RoboTwin50. FutureDuet achieves 94.2% success in the clean setting and 94.1% under randomization, surpassing the reported Fast-WAM results by 2.3 percentage points in both settings (Tab. 2).

4.3 Task-Level Analysis on RoboTwin50

Performance across tasks. The high benchmark average coexists with a small set of challenging tasks (Fig. 3). FutureDuet reaches at least 90% success on 41 of 50 tasks in both settings, while six remain below this threshold in both. These tasks require precise terminal interaction, constrained placement, or precision across multiple stages. On this subset, FutureDuet improves over Fast-WAM by 9.2 percentage points in clean evaluation and 12.8 percentage points under randomization, compared with 2.3 points benchmark-wide. Complete results and task groupings appear in App. A.3.

Gains on challenging interactions. The six remaining challenging tasks account for 48% of the total net success gain over Fast-WAM in clean evaluation and 66% under randomization. Open Microwave, Hanging Mug, and Turn Switch show the largest gains within this group in both settings. Their illustrated failures share a common structure: the policy reaches the relevant object but does not complete the final local transition—the door must open far enough, the mug handle must engage the hook, and the end effector must align with and actuate the switch (Fig. 4). These examples suggest that success depends not only on reaching the interaction region, but also on resolving and executing the contact geometry required for completion. Because wrist views provide

close-range evidence of these terminal transitions, these cases motivate the controlled wrist-pathway study in Sec. 4.4, where we isolate the contributions of wrist access, pathway separation, and future supervision on the same three tasks.



Figure 4: **Failure cases on three challenging interaction tasks.** The policy reaches the relevant interaction region but fails to complete the required opening, engagement, or contact.

4.4 Why Use a Separate Wrist Pathway?

We isolate three factors on the wrist side: observation access, pathway separation, and future supervision (Fig. 5). Main-view supervision remains RGB-only throughout this study. We first compare unified and separate wrist pathways, then hold the lightweight pathway fixed to evaluate different prediction targets.

Wrist access and pathway separation. Incorporating wrist observations into the unified pathway boosts mean success from 34.3% to 55.8%, validating the value of close-range wrist cues. Remarkably, a lightweight separate pathway reaches 60.9% even without auxiliary wrist future prediction, demonstrating the intrinsic benefit of a decoupled stream.

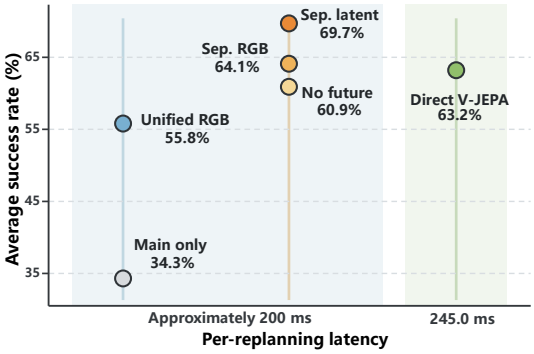

Figure 5: Three-seed wrist-pathway study with RGB-only main-view supervision.

Future supervision under a fixed pathway. Applying future RGB supervision boosts success from 60.9% to 64.1%, whereas future-latent supervision further elevates performance to 69.7%. The RGB result confirms that wrist future prediction is also beneficial, rather than being restricted to latent targets. Nevertheless, the additional 5.6-point gain from latent supervision demonstrates that the choice of target representation is critical within the same pathway. While both objectives provide auxiliary temporal signals beyond action loss, the latent target effectively leverages rich pretrained video features without demanding pixel-level reconstruction. Overall, these findings validate pairing a decoupled wrist pathway with future-latent prediction, achieving optimal performance while preserving deployment efficiency.

Deployment efficiency. Deploying a heavy encoder like V-JEPA highlights the trade-off between model capacity and efficiency, increasing latency from 196.86 ms to 244.99 ms for a 63.2% success rate. We bypass this limitation through training-only auxiliary supervision: confining future-latent prediction strictly to training enables our lightweight pathway to achieve 69.7% success at near-baseline latency ($\sim$200 ms). Since all auxiliary targets share the same deployed execution graph, upgrading the training target enriches feature representations with zero inference overhead.

4.5 Designing Structured Supervision for the Main View

Decoupling future supervision allows us to design the main and wrist pathways around the information they provide for control. With the wrist pathway established, we now investigate supervision strategies for the main view. Future RGB prediction captures scene evolution but does not explicitly prioritize the task objects and robot motion that guide manipulation. We therefore study whether masks and skeletons strengthen this supervision and complement the separate wrist pathway (Tab. 3).

Table 3: **Factorial study of main-view supervision and wrist pathways.** The study crosses four main-view target configurations with unified future-RGB and separate future-latent wrist pathways. Fine-grained execution comprises Open Microwave (OM), Hanging Mug (HM), and Turn Switch (TS); additional tasks comprise Adjust Bottle (AB), Blocks Ranking RGB (BR), and Grab Roller (GR). Green and yellow shading indicate the best and second-best result in each column.

Predictive configuration		Fine-grained execution			Additional tasks			Overall
Main-view targets	**Wrist pathway**	**OM**	**HM**	**TS**	**AB**	**BR**	**GR**	**Avg.**
RGB	Unified (future RGB)	43	60	57	88	90	88	71.0
RGB+Mask	Unified (future RGB)	43	60	58	**92**	93	89	72.5
RGB+Skel.	Unified (future RGB)	44	61	59	89	91	**92**	72.7
RGB+Mask+Skel.	Unified (future RGB)	45	62	59	91	**94**	90	73.5
RGB	Separate (future latent)	59	66	68	88	88	87	76.0
RGB+Mask	Separate (future latent)	60	66	69	**92**	93	87	77.8
RGB+Skel.	Separate (future latent)	60	68	**70**	88	89	90	77.5
RGB+Mask+Skel.	**Separate (future latent)**	**65**	**70**	**70**	91	**94**	**92**	**80.3**

Complementarity with wrist modeling. Adding masks and skeletons raises average success from 71.0% to 73.5% under the unified pathway and from 76.0% to 80.3% under the separate future-latent pathway, with gains spanning both fine-grained execution and the additional tasks. With RGB-only main-view supervision, the gains from the separate future-latent pathway are concentrated on OM, HM, and TS. Adding structured main-view targets also improves AB, BR, and GR under both pathways, broadening the benefits beyond the fine-grained tasks most responsive to wrist modeling. These improvements show that the value of separately designing future supervision extends beyond the wrist pathway. Alongside wrist-side prediction, explicitly emphasizing task objects and robot motion in the main-view objective provides additional gains, with the combined design achieving the highest average success under both wrist pathways.

Distinct contributions of masks and skeletons. Under both wrist pathways, masks provide larger gains than skeletons on Adjust Bottle and Blocks Ranking RGB, whereas skeletons provide larger gains on Grab Roller. Neither target is consistently better across these tasks, and combining them achieves the highest average success under either pathway. This pattern supports supervising both task objects and robot motion rather than relying on either target alone.

Together, these results show that structured main view supervision complements wrist modeling, with masks and skeletons contributing differently across tasks. We next examine whether this functional division is mirrored in how ActionDiT allocates attention to task objects and robot features.

4.6 Understanding the Effects of Structured Prediction Targets

The factorial study shows that masks and skeletons produce different task-level gains. We therefore examine whether they change the main-view evidence used by ActionDiT. Across 48 matched queries from separately trained six-task checkpoints, we first use action-to-visual attention to measure where action queries retrieve information, and then apply matched K/V interventions to test whether features from those regions affect the action output.

Where does ActionDiT read? We extract action-query-to-main-view attention from the final two joint VideoDiT–ActionDiT attention blocks. RR(O) and RR(G) quantify action attention directed to task-object and executing-robot regions relative to a size-matched background, whereas OCE measures object-center alignment, with lower values indicating better alignment.

The effects are target-specific. Under the separate wrist pathway, mask supervision raises object readout RR(O) from 1.10 to 1.46 and improves object-center alignment, whereas skeleton supervision produces the strongest robot-side change, raising RR(G) from 1.20 to 2.00. The same division appears under unified wrist modeling, indicating that it is induced by the main-view targets rather than the wrist pathway. Fig. 6 qualitatively illustrates this effect, showing how joint supervision concentrates action readout on the task-relevant interaction.

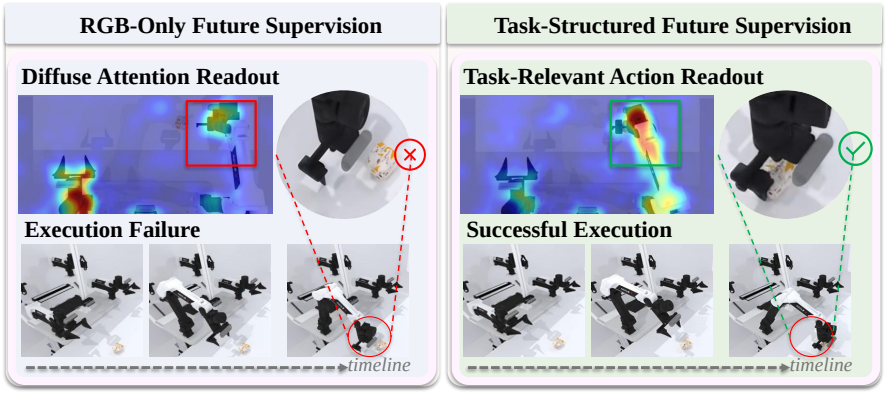

Figure 6: **Structured targets focus action readout.** Compared with RGB-only future supervision, joint mask-and-skeleton supervision focuses readout on the task-relevant interaction.

Which regions influence action generation? Attention indicates where action queries retrieve information, but not whether the retrieved features affect the output. We therefore replace object- or robot-region K/V features with the mean main-view K/V and measure the resulting action change using R(O) and R(G).
The intervention results mirror the attention patterns. Mask supervision primarily increases sensitivity to object features, whereas skeleton supervision primarily increases sensitivity to robot features; joint supervision preserves both effects under either wrist pathway. Thus, the target-specific readout is reflected not only in attention patterns but also in the action model's response to the corresponding visual features.

	Task-Object Readout			Robot-State Readout	
	RR(O)	R(O)	OCE↓	RR(G)	R(G)
Mask	+0.36	+0.63	+0.68	+0.04	+0.06
Skel.	+0.03	+0.10	+0.13	+0.80	+0.89
Both	+0.47	+1.00	+0.92	+0.96	+1.17

Figure 7: **Target-specific changes under separate wrist modeling.** Positive values indicate increases in attention readout and feature sensitivity, as well as reductions in OCE.

Together, the two diagnostics provide complementary evidence: attention identifies where Action-DiT reads, while K/V intervention shows that the corresponding features affect its output. Their agreement suggests that mask and skeleton supervision shape complementary sides of the Task State—the task object and the robot executing the motion—helping explain why their combination performs best in the factorial study. Full results appear in Appendix Tab. A10.

5 Conclusion

FutureDuet decouples two choices that existing WAMs often entangle together: which observations support control and what future each visual stream is trained to predict. By using main and wrist observations for action generation while predicting structured main-view targets and future wrist latents, our framework learns complementary representations of overall task progress and local interaction. This decoupled design improves performance across RoboTwin50 and LIBERO, with the largest RoboTwin50 gains appearing on tasks requiring precise interaction completion. Controlled studies isolate two key sources of wrist-side improvement: pathway separation provides a fundamental architectural benefit, while supervision target selection favors future-latent prediction over pixel-level RGB reconstruction under the same deployed architecture. Meanwhile, structured main-view supervision offers complementary gains. While our evaluation is currently limited to simulation, future work will test FutureDuet on real robots and explore broader view-target configurations. Overall, our results demonstrate that distinct camera streams can jointly guide action control without forcing every view to predict a uniform future, highlighting the value of decoupling observation access from training-time auxiliary future supervision in world action models.

AI USE STATEMENT

Generative AI tools were used to assist with code development for experiment execution and to refine the text for readability. The authors reviewed and verified all AI-assisted code and text, and take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Secs. 3–4 and the appendix document the architecture, objectives, training and evaluation protocols, target construction, and diagnostic definitions. The appendix also provides implementation details and complete task-level results.

REFERENCES

Homanga Bharadhwaj, Roozbeh Mottaghi, Abhinav Gupta, and Shubham Tulsiani. Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation. In *European Conference on Computer Vision*, pp. 306–324. Springer, 2024.

Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2Act: Human video generation in novel scenarios enables generalizable robot manipulation. In *Proceedings of the 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 3936–3951. PMLR, 2025. URL https://proceedings.mlr.press/v305/bharadhwaj25a.html.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, Hongyan Zhao, Hanyu Liu, Zhizhong Su, Lei Ma, Hang Su, and Jun Zhu. Motus: A unified latent action world model. *arXiv preprint arXiv:2512.13030*, 2025.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, Laura Smith, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control. In *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.010.

Anthony Brohan et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. URL https://arxiv.org/abs/2307.15818.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Jun Cen, Siteng Huang, Yuqian Yuan, Hangjie Yuan, Chaohui Yu, Yuming Jiang, Jiayan Guo, Kehan Li, Hao Luo, Fan Wang, Xin Li, Deli Zhao, and Hao Chen. Rynnvla-002: A unified vision-language-action and world model. *arXiv preprint arXiv:2511.17502*, 2025a.

Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, Deli Zhao, and Hao Chen. WorldVLA: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*, 2025b.

Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, Hanbo Zhang, and Minzhao Zhu. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*, 2024.

Jialei Chen, Kai Wang, Kang Chen, Shuaihang Chen, Feng Gao, Wenhao Tang, Zhiyuan Li, Weilin Liu, Zhuyu Yao, Boxun Li, Yuanbo Xu, and Chao Yu. LaWAM: Latent world action models for efficient dynamics-aware robot policies. *arXiv preprint arXiv:2606.15768*, 2026a. URL https://arxiv.org/abs/2606.15768.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, Weiliang Deng, Yubin Guo, Tian Nian, Xuanbing Xie, Qiangyu Chen, Kailun Su, Tianling Xu, Guodong Liu, Mengkang Hu, Huan-ang Gao, Kaixuan Wang, Zhixuan Liang, Yusen Qin, Xiaokang Yang, Ping Luo, and Yao Mu. RoboTwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026b. URL https://icml.cc/virtual/2026/poster/62192.

Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0403. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1d5b9233ad716a43be5c0d3023cb82d0-Abstract-Conference.html.

Yao Feng, Hengkai Tan, Xinyi Mao, Guodong Liu, Shuhe Huang, Chendong Xiang, Hang Su, and Jun Zhu. Vidar: Embodied video diffusion model for generalist bimanual manipulation, 2025. URL https://arxiv.org/abs/2507.12898.

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

Siddhant Haldar and Lerrel Pinto. Point policy: Unifying observations and actions with key points for robot manipulation. *arXiv preprint arXiv:2502.20391*, 2025.

Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 24328–24346. PMLR, 2025. URL https://proceedings.mlr.press/v267/hu25g.html.

Jiakai Huang, Zhongbo Wu, Zheng Zhang, Zihan Wang, Shan You, and Tao Huang. Foresight without seeing: Latent futures for world action models. *arXiv preprint arXiv:2608.11605*, 2026. URL https://arxiv.org/abs/2608.11605.

Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, Loic Magne, Ajay Mandlekar, Avnish Narayan, You Liang Tan, Guanzhi Wang, Jing Wang, Qi Wang, Yinzhen Xu, Xiaohui Zeng, Kaiyuan Zheng, Ruijie Zheng, Ming-Yu Liu, Luke Zettlemoyer, Dieter Fox, Jan Kautz, Scott Reed, Yuke Zhu, and Linxi Fan. Dreamgen: Unlocking generalization in robot learning through video world models, 2025. URL https://arxiv.org/abs/2505.12705.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. In *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2025a. doi: 10.15607/RSS.2025.XXI.017.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P. Foster, Pannag R. Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713. PMLR, 2025b. URL https://proceedings.mlr.press/v270/kim25c.html.

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. *arXiv preprint arXiv:2601.16163*, 2026.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026.

Junbang Liang, Pavel Tokmakov, Ruoshi Liu, Sruthi Sudhakar, Paarth Shah, Rares Ambrus, and Carl Vondrick. Video generators are robot policies. *arXiv preprint arXiv:2508.00795*, 2025.

Weixin Liang, Lili Yu, Liang Luo, Srinivasan Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen-tau Yih, Luke Zettlemoyer, et al. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *arXiv preprint arXiv:2411.04996*, 2024.

Yihan Lin, Jiawei He, Shifeng Bao, Chen Zhao, Yang Li, Xiaobo Wang, Yan Wang, Cheng Chi, and Jing Zhang. JEPA-WAM: Learning vision-language-action policies with joint-embedding world modeling. *arXiv preprint arXiv:2608.09381*, 2026. URL https://arxiv.org/abs/2608.09381.

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

Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. RDT-1B: A diffusion foundation model for bimanual manipulation. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=yAzN4tz7oI.

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.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-JEPA 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.

NVIDIA et al. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024. URL https://arxiv.org/abs/2405.12213.

Jonas Pai, Liam Achenbach, Victoriano Montesinos, Benedek Forrai, Oier Mees, and Elvis Nava. mimic-video: Video-action models for generalizable robot control beyond vlas. *arXiv preprint arXiv:2512.15692*, 2025.

Yuhao Pan, Haosong Peng, Zhengshen Zhang, Zhengyang Yan, Yalun Dai, Fushuo Huo, Chujie Wang, Tianyu Qi, Xiucheng Wang, Nan Cheng, and Wenchao Xu. World-to-Wrist: Task-conditioned future wrist modeling for fine-grained robot manipulation. *arXiv preprint arXiv:2608.05369*, 2026. URL https://arxiv.org/abs/2608.05369.

Physical Intelligence et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Jiayuan Gu, Zhigang Wang, Yan Ding, Bin Zhao, Dong Wang, and Xuelong Li. SpatialVLA: Exploring spatial representations for visual-language-action models. In *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.011.

Wenxuan Song, Ziyang Zhou, Han Zhao, Jiayi Chen, Pengxiang Ding, Haodong Yan, Yuxin Huang, Feilong Tang, Donglin Wang, and Haoang Li. Reconvla: Reconstructive vision-language-action model as effective robot perceiver, 2025. URL https://arxiv.org/abs/2508.10333.

Jingwen Sun, Wenyao Zhang, Zekun Qi, Shaojie Ren, Zezhi Liu, Hanxin Zhu, Guangzhong Sun, Xin Jin, and Zhibo Chen. VLA-JEPA: Enhancing vision-language-action model with latent world model. *arXiv preprint arXiv:2602.10098*, 2026.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

John Won, Kyungmin Lee, Huiwon Jang, Dongyoung Kim, and Jinwoo Shin. Dual-stream diffusion for world-model augmented vision-language-action model, 2025. URL https://arxiv.org/abs/2510.27607.

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.

Wei Wu, Fan Lu, Yunnan Wang, Shuai Yang, Shi Liu, Fangjing Wang, Qian Zhu, He Sun, Yong Wang, Shuailei Ma, et al. A pragmatic vla foundation model. *arXiv preprint arXiv:2601.18692*, 2026.

Hanyang Yu, Haitao Lin, Jingbo Zhang, Wenyao Zhang, Chenghao Gu, Heng Li, and Ping Tan. MaskWAM: Unifying mask prompting and prediction for world-action models. *arXiv preprint arXiv:2606.13515*, 2026.

Shanglin Yuan, Weiheng Zhao, Xin Shi, Haoyi Jiang, Xianda Guo, Liu Liu, Wenyu Liu, Wei Sui, and Xinggang Wang. DreamWAM: Beyond RGB future prediction for world action models. *arXiv preprint arXiv:2608.04996*, 2026a. URL https://arxiv.org/abs/2608.04996.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-WAM: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026b.

Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, Xinqiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, Fan Lu, He Wang, Zhizheng Zhang, Li Yi, Wenjun Zeng, and Xin Jin. Dreamvla: A vision-language-action model dreamed with comprehensive world knowledge. *CoRR*, abs/2507.04447, 2025. doi: 10.48550/ARXIV.2507.04447. URL https://doi.org/10.48550/arXiv.2507.04447.

Wenyao Zhang, Bozhou Zhang, Zekun Qi, Wenjun Zeng, Xin Jin, and Li Zhang. Disentangled robot learning via separate forward and inverse dynamics pretraining. *arXiv preprint arXiv:2604.16391*, 2026.

Jinliang Zheng, Jianxiong Li, Zhihao Wang, Dongxiu Liu, Xirui Kang, Yuchun Feng, Yinan Zheng, Jiayin Zou, Yilun Chen, Jia Zeng, Ya-Qin Zhang, Jiangmiao Pang, Jingjing Liu, Tai Wang, and Xianyuan Zhan. X-VLA: Soft-prompted transformer as scalable cross-embodiment vision-language-action model, 2025. URL https://arxiv.org/abs/2510.10274. arXiv preprint arXiv:2510.10274.

Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. RoboDreamer: Learning compositional world models for robot imagination. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61885–61896. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhou24f.html.

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. In *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.015.

A SUPPLEMENTARY MATERIAL

The supplementary material provides implementation and protocol details, complete task-level results, controlled wrist-pathway and main-view studies, and the full action-readout diagnostics.

A.1 IMPLEMENTATION AND TRAINING DETAILS

A.1.1 MODEL COMPONENTS AND TARGET CONSTRUCTION

Tab. A1 records the inputs, training targets, and deployed representations for the two visual streams.

Table A1: Inputs, future targets, and deployed representations in FutureDuet.

Stream	**Current policy input**	**Training-only targets**	**State exposed to ActionDiT**
Main view	Main-view history, language, and proprioception	Future video, interaction masks, and articulated skeleton maps	Task State and current main-view K/V features
Wrist views	Left/right wrist histories	Cached future wrist latents from a frozen V-JEPA 2.1 teacher	Interaction State

Target construction. Future main view frames are encoded into VAE latents for the future-video flow-matching objective. Interaction masks are generated offline from recorded RGB frames with SAM 3 (Carion et al., 2025); they retain the manipulated object and, when required by the task stage, its target container or region. Skeleton maps are generated by applying forward kinematics to recorded joint states, projecting robot links and joints with the calibrated main camera, and rasterizing the articulated structure:

$$q_{t+k} \xrightarrow{\text{FK}} X^{\text{3D}}_{t+k} \xrightarrow{K_{\text{cam}}[R|T]} X^{\text{2D}}_{t+k} \xrightarrow{\text{rasterize}} S_{t+k}.$$

The left and right arms use separate channels. Future wrist targets are constructed from the same recorded clips and cached offline. For future-latent supervision, a frozen V-JEPA 2.1 teacher encodes the clips into representation targets. For future-RGB supervision, the same clips are encoded by the video VAE and the predictor targets their VAE latents. The online wrist encoder, adapter, predictor backbone, detached Task State conditioning, and deployed ActionDiT interface are held fixed; only the target representation and its training-only output projection differ. Both wrist target types are precomputed, so neither target encoder is part of the deployed policy.

Interaction masks use an equally weighted combination of focal and Tversky losses; skeleton maps use an equally weighted combination of focal and Dice losses. Wrist regression averages absolute errors over valid future target elements, excluding padded target positions. In the flow objectives, expectations cover training samples, Gaussian noise, and scheduler timesteps; squared errors are averaged over valid elements before timestep weighting. The objectives are defined in Sec. 3.

Training and deployment. Action generation and the enabled future objectives are optimized jointly. Learned-query adapters produce the Task State and Interaction State from the main and wrist features, respectively; the Duet Interface combines these states for ActionDiT, which also reads current main-view K/V features. The adapters and Duet Interface remain active at deployment. The DPT decoder and mask/skeleton heads are trainable, whereas the V-JEPA teacher is frozen and used offline. The VAE decoder, DPT decoder, auxiliary output heads, and wrist predictor are omitted from the deployed policy.

A.1.2 OPTIMIZATION DETAILS

All matched variants within a study use the same demonstrations, preprocessing, optimization budget, and evaluation protocol. The study-specific sections below state which pathway or supervision factor changes in each comparison. Tab. A2 lists the shared settings.

Table A2: Training configuration used for the reported experiments.

Configuration	LIBERO	RoboTwin50	Six-task factorial	Three-task wrist-pathway
Training data	50 demos/task	50 clean + 500 rand./task	500 rand./task	500 rand./task
Evaluation	50 rollouts/task	100 rollouts/task/setting	100 rand./task	100 rand./task/seed
Training epochs	5	5	3	3
Precision	bf16	bf16	bf16	bf16
Optimizer	AdamW	AdamW	AdamW	AdamW
AdamW β / ϵ	$(0.9, 0.95)$ / 10^{-8}			
Base learning rate	2×10^{-5}	1×10^{-4}	1×10^{-4}	1×10^{-4}
Dense-head learning rate (when enabled)	1×10^{-4}	1×10^{-4}	1×10^{-4}	–
Wrist encoder/predictor learning rate (when enabled)	2×10^{-4}	2×10^{-4}	2×10^{-4}	2×10^{-4}
Weight decay	10^{-2}	10^{-2}	10^{-2}	10^{-2}
Gradient clipping	1.0	1.0	1.0	1.0
LR schedule	5% linear warm-up, then cosine decay to $0.01\times$ the base rate			

In configurations with structured main-view supervision, the dense heads predict eight sampled future mask and skeleton targets. The corresponding losses use weights of 0.1, while the wrist latent-regression loss uses weight 1.0. Auxiliary objectives use the same early-training ramp across matched variants.

Table A3: Training scale for the RoboTwin studies.

Study	Demonstrations	Optimizer steps	Global batch	Seeds
RoboTwin50	27,500	29,800	1,024	42
Six-task factorial ablation	3,000	33,005	128	42
Three-task wrist-pathway study	1,500	19,350	128	42, 43, 44

RoboTwin training uses eight NVIDIA H100 80GB GPUs with DeepSpeed ZeRO-1. A three-epoch six-task factorial run takes approximately 8–11 hours, while training the full RoboTwin50 model takes approximately 3–4 days.

Temporal and model configuration. For the RoboTwin controlled studies, the three input cameras are recorded at 480×640 and resized to 240×320; VideoDiT operates on a 384×320 canvas. Each sample contains a 32-step action chunk and nine video frames at action offsets $0, 4, \ldots, 32$. The wrist route uses eight observed frames at offsets $[-14, -12, \ldots, 0]$ and eight future target frames at offsets $[2, 4, \ldots, 16]$, each resized to 224×224. Robot state and action are both 14-dimensional and use precomputed z-score normalization statistics.

The shared backbone is Wan2.2-TI2V-5B with a 30-layer, 3072-dimensional VideoDiT and a 30-layer, 1024-dimensional ActionDiT. Both use 24 attention heads of dimension 128; ActionDiT projects Q/K/V to an internal width of 3072 despite its 1024-dimensional hidden states. ActionDiT is initialized from an interpolated Wan2.2 checkpoint, VideoDiT uses gradient checkpointing, and text is supplied as cached T5 context with a maximum length of 128. The wrist route uses a lightweight online encoder and a shared training-only predictor. Under future-RGB supervision, the predictor targets cached video-VAE latents; under future-latent supervision, it targets cached V-JEPA 2.1 representations through a target-specific output projection. In the RoboTwin controlled studies, the wrist future loss uses weight 1.0 with a 500-step ramp. All deployed modules are optimized jointly; the teacher and predictor are not used at inference.

A.2 Evaluation Protocols and Study Index

Tab. A4 separates the task scope, checkpoint family, and comparison reported by each study. "Queries" are fixed contact-adjacent observations used for representation diagnostics rather than policy-success evaluation. The three-task wrist-pathway study, six-task factorial ablation, and action-readout diagnostics use separate checkpoint sets.

Table A4: Scope, checkpoint family, and comparison reported by each study.

Study	Task scope	Models or checkpoints	Comparison
RoboTwin50 main result	50 tasks; 100 clean and 100 randomized rollouts per task	One jointly trained 50-task model	Benchmark performance and task-level distribution
LIBERO evaluation	Four single-arm suites; 50 rollouts per task	Models trained with the reported protocol	Performance on single-arm manipulation suites
Three-task wrist-pathway study	Open Microwave, Hanging Mug, and Turn Switch; 100 randomized rollouts per task and seed	Six variants, each trained with three seeds	Wrist observation, pathway configuration, and future target
Six-task factorial ablation	The three tasks from the wrist-pathway study plus Adjust Bottle, Blocks Ranking RGB, and Grab Roller; 100 randomized rollouts per task	Eight independently trained variants (seed 42)	Structured main-view targets under two predefined wrist configurations
Action-readout diagnostics	Adjust Bottle, Beat Block Hammer, Blocks Ranking RGB, Click Alarmclock, Click Bell, and Grab Roller; 48 matched queries	Eight frozen diagnostic checkpoints	Main-view regional readout and action-output sensitivity

Benchmark protocols. The RoboTwin50 model is jointly trained on 50 clean and 500 randomized demonstrations per task, totaling 27,500 demonstrations. Its dataset release, demonstration split, Wan2.2 and ActionDiT backbones, observation preprocessing, training budget, checkpoint-selection protocol, and evaluation follow Fast-WAM. Randomized evaluation varies scene clutter, appearance, illumination, tabletop conditions, object placement, and language instantiation. LIBERO uses 50 demonstrations and 50 evaluation episodes for each task in its four single-arm suites. Benchmark values for prior methods are taken from the evaluation sources identified in the main tables.

A.3 RoboTwin50 Task-Level Performance

A.3.1 Complete task-level results

Tab. A5 reports all 50 task-level results for FutureDuet alongside the reported RoboTwin50 evaluations of $\pi_{0.5}$, LingBot-VA, and Fast-WAM (Physical Intelligence et al., 2025; Li et al., 2026; Yuan et al., 2026b). Each task is evaluated over 100 episodes in each setting. For FutureDuet, the 50-task sums are 4708 clean successes and 4706 randomized successes over 5000 episodes, corresponding to 94.2% and 94.1%. Fig. 3 in the main paper summarizes the corresponding task-count distribution at the 90% and 95% thresholds.

Table A5: Per-task RoboTwin50 success rates (%). Bold light-green and underlined light-yellow cells mark the best and second-best distinct results; orange shading marks the six remaining challenging tasks.

	$\pi_{0.5}$		LingBot-VA		Fast-WAM		FutureDuet	
Task	**Clean**	**Rand.**	**Clean**	**Rand.**	**Clean**	**Rand.**	**Clean**	**Rand.**
Adjust Bottle	**100**	99	90	94	**100**	**100**	**100**	**100**
Beat Block Hammer	96	93	96	**98**	**99**	97	98	97
Blocks Ranking RGB	92	85	99	98	**100**	**100**	**100**	98
Blocks Ranking Size	49	26	**94**	96	**94**	**98**	**94**	**98**
Click Alarmclock	98	89	99	**100**	**100**	**100**	**100**	98
Click Bell	99	66	**100**	**100**	**100**	**100**	**100**	98
Dump Bin Bigbin	92	**97**	89	96	97	96	**98**	96
Grab Roller	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
Handover Block	66	57	**99**	78	95	81	96	**87**
Handover Mic	98	97	94	96	**99**	**100**	**99**	**100**
Hanging Mug	18	17	40	28	58	62	**70**	**76**
Lift Pot	96	85	**100**	99	**100**	**100**	**100**	98
Move Can Pot	51	55	**94**	**97**	90	88	**94**	94
Move Pillbottle Pad	84	61	99	**99**	**100**	**99**	99	**99**
Move Playingcard Away	96	84	**100**	99	**100**	**100**	**100**	**100**
Move Stapler Pad	56	42	**91**	**79**	77	64	85	74
Open Laptop	90	96	92	94	98	**100**	**99**	**100**
Open Microwave	34	77	**82**	**86**	62	45	72	66
Pick Diverse Bottles	81	71	**89**	82	80	85	84	**90**
Pick Dual Bottles	93	63	**100**	**99**	**100**	96	**100**	96
Place A2B Left	87	82	**97**	93	95	93	**97**	**97**
Place A2B Right	87	84	**97**	95	93	**99**	95	**99**
Place Bread Basket	77	64	**97**	95	91	93	94	**97**
Place Bread Skillet	85	66	**95**	90	90	93	93	**95**
Place Burger Fries	94	87	**97**	95	96	**99**	**97**	**99**
Place Can Basket	62	62	**81**	**84**	71	69	78	79
Place Cans Plasticbox	94	84	**100**	**99**	99	96	**100**	97
Place Container Plate	**99**	95	**99**	97	96	**100**	96	**100**
Place Dual Shoes	75	75	94	89	94	88	**96**	**90**
Place Empty Cup	**100**	99	**100**	**100**	**100**	**100**	**100**	**100**
Place Fan	87	85	**99**	93	96	96	98	**98**
Place Mouse Pad	60	39	**93**	**96**	83	89	88	95
Place Object Basket	80	76	91	88	89	88	**94**	**94**
Place Object Scale	86	80	**96**	95	90	**97**	92	**97**

Table A5: Per-task RoboTwin50 success rates (%), continued.

Task	$\pi_{0.5}$		LingBot-VA		Fast-WAM		FutureDuet	
	Clean	Rand.	Clean	Rand.	Clean	Rand.	Clean	Rand.
Place Object Stand	91	85	**99**	96	90	94	92	**99**
Place Phone Stand	81	81	**97**	97	**97**	**99**	**97**	**99**
Place Shoe	92	93	**98**	98	96	**99**	**98**	**99**
Press Stapler	87	83	85	82	90	97	**93**	**98**
Put Bottles Dustbin	84	79	87	91	95	90	**97**	**96**
Put Object Cabinet	80	79	85	87	94	89	**96**	**93**
Rotate QRcode	89	87	**96**	**91**	93	89	**96**	**91**
Scan Object	72	65	**96**	91	89	92	92	**93**
Shake Bottle	99	97	**100**	97	**100**	**100**	**100**	96
Shake Bottle Horizontally	99	99	**100**	99	**100**	**100**	**100**	97
Stack Blocks Three	91	76	**99**	**98**	95	97	96	94
Stack Blocks Two	97	**100**	**100**	98	**100**	**100**	**100**	96
Stack Bowls Three	77	71	**86**	83	80	81	**86**	**85**
Stack Bowls Two	**95**	96	94	**98**	92	**98**	94	**98**
Stamp Seal	79	55	**96**	**97**	90	94	92	93
Turn Switch	62	54	44	45	61	59	**73**	**77**
Average	82.7	76.8	92.9	91.5	91.9	91.8	**94.2**	**94.1**

A.3.2 Remaining challenging tasks

Six tasks remain below 90% for FutureDuet in both evaluation settings: Hanging Mug, Move Stapler Pad, Open Microwave, Place Can Basket, Stack Bowls Three, and Turn Switch. Across this set, Fast-WAM averages 68.2% clean and 63.3% randomized success, whereas FutureDuet reaches 77.3% and 76.2%. Tab. A6 groups these tasks by their primary execution demand.

Table A6: Execution demands of the six remaining challenging tasks in RoboTwin50.

Execution demand	Tasks	Why precision is decisive
Terminal interaction completion	Open Microwave; Hanging Mug; Turn Switch	Success depends on completing a small local transition after reaching the correct interaction region.
Constrained object-to-target placement	Move Stapler Pad; Place Can Basket	The manipulated object must satisfy a narrow target-region or support constraint at release.
Multi-stage precision accumulation	Stack Bowls Three	Alignment and stability must be preserved across successive placements, so local errors compound.

Fine-grained gains and remaining failures. The terminal-interaction rollouts show a common outcome: the microwave handle is reached but the door is not opened far enough, the mug reaches the rack but fails to engage the hook, or the correct switch is approached without completing the final actuation. These cases motivate the controlled three-task wrist-pathway study in Sec. A.4, which examines wrist observation, pathway configuration, and future supervision. Improvements on the constrained-placement and multi-stage tasks further show that the benefit extends beyond the three motivating examples. Future work can examine failure stage, contact quality, and recovery behavior to better understand the remaining errors.

A.4 Three-Task Wrist-Pathway Study

A.4.1 Controlled configuration and complete results

Controlled three-task configuration. We study Open Microwave, Hanging Mug, and Turn Switch using the same 500 randomized demonstrations per task for every variant. Each model is trained for three epochs (19,350 optimizer steps) with global batch size 128 and evaluated over 100 randomized episodes per task. Results are averaged over seeds 42, 43, and 44; no validation split or rollout-based model selection is used. Main-view supervision remains future RGB throughout, and mask and skeleton supervision are disabled.

Unified future RGB reproduces Fast-WAM within this pipeline. The main-view-only ablation does not change the image resolution or visual canvas: both wrist observations are replaced with constant white frames before the standard view concatenation and preprocessing, preserving the input shape, backbone, and training configuration. The direct V-JEPA control instead supplies frozen V-JEPA wrist features to the action model.

The three lightweight variants share the same online wrist encoder, adapter, Task State conditioning, and deployed ActionDiT interface. The no-future variant omits the auxiliary wrist objective. The other two use the same future wrist clips and predictor backbone, but target either cached video-VAE latents or cached V-JEPA latents through target-specific training-only output projections. Thus, their deployed wrist pathway is identical; only the training-time future supervision changes.

Tab. A7 reports the complete task-level results.

Table A7: Three-task wrist-pathway study. Values are three-seed mean success rates (%) over 100 randomized episodes per task and seed.

Wrist configuration	OM	HM	TS	Avg.
Main view only (white wrist frames)	25.0	38.3	39.7	34.3
Unified multi-view (Fast-WAM)	43.7	63.0	60.7	55.8
Direct V-JEPA wrist encoding	49.3	69.3	71.0	63.2
Lightweight wrist, no future loss	49.3	67.0	66.3	60.9
Separate future RGB	56.7	67.7	68.0	64.1
Separate future latent	**67.3**	**70.3**	**71.3**	**69.7**

OM: Open Microwave; HM: Hanging Mug; TS: Turn Switch.

The main-view-only and unified conditions isolate the contribution of wrist observations while preserving the unified input shape. The no-future lightweight variant shows that a dedicated wrist pathway can perform strongly even without a wrist future loss; relative to unified future RGB, however, this is a configuration-level comparison because both the pathway and wrist supervision change. Under the fixed lightweight pathway, both VAE-latent and V-JEPA-latent future targets improve over no future loss, with the V-JEPA target performing best. Direct V-JEPA encoding provides a higher-capacity reference but incurs substantially greater deployment latency.

A.4.2 Inference efficiency

We measure one policy call that generates a 32-step action chunk on an RTX 5090. Each deployed graph is evaluated in BF16 with 24 executed actions before replanning, two warm-up calls, and 100 timed calls. A fixed synthetic request is used across variants; wrist-enabled graphs additionally receive the raw camera tensors and run their standard resize, crop, normalization, and deployed wrist encoder.

Server-side timing covers model-side preprocessing, host-to-device transfer, wrist encoding when present, ActionDiT generation, action denormalization, and transfer back to the CPU. The final transfer synchronizes GPU computation. Serialization and transport are excluded. The three lightweight variants share the same deployed computation graph and are timed in independent runs; their small latency differences therefore reflect measurement variation rather than supervision-specific inference operations.

Table A8: Inference latency on an RTX 5090, averaged over 100 replanning calls. Amortized compute divides latency by 24 executed actions.

Variant	Server latency (ms)	Amortized compute (ms/step)
Main view only (white wrist frames)	196.86	8.20
Unified future RGB	196.86	8.20
Direct V-JEPA wrist encoding	244.99	10.21
Lightweight wrist, no future loss	199.72	8.32
Separate future RGB	199.77	8.32
Separate future latent	197.75	8.24

Direct V-JEPA wrist encoding increases per-replanning latency by 24.5% over unified future RGB. The lightweight wrist graph runs at approximately 200 ms, adding about 1.5% latency. Its no-future, future-RGB, and future-latent variants differ only in training-time supervision; the observed differences of at most 2.0 ms are within run-to-run timing variation. The V-JEPA teacher, future predictor, mask head, and skeleton head are disabled after training.

A.5 Factorial Study of Main-View Supervision and Wrist Pathways

Controlled configuration. The six-task study crosses four main-view target sets—RGB only, RGB+mask, RGB+skeleton, and RGB+mask+skeleton—with two predefined wrist configurations: unified future-RGB modeling and the separate future-latent pathway. All eight models use the same six tasks and 500 randomized demonstrations per task, the same observation and action preprocessing, the same pretrained initialization for shared modules, 33,005 optimizer steps, global batch size 128, seed 42, and the same evaluation protocol of 100 randomized rollouts per task.

Within each wrist configuration, the wrist pathway and wrist objective remain fixed; the four cells differ only in whether mask and skeleton supervision is enabled for the main view. The correspond-

ing dense heads are training-only and do not change the deployed architecture. Comparisons within each four-row block therefore isolate structured main-view supervision under that wrist configuration. Comparisons between the two blocks instead contrast the complete unified and separate wrist configurations, because both the wrist pathway and its future target change.

The main paper reports all eight task-level configurations. Tab. A9 aggregates them over the three fine-grained execution tasks (Open Microwave, Hanging Mug, and Turn Switch) and the three additional tasks (Adjust Bottle, Blocks Ranking RGB, and Grab Roller).

Table A9: Grouped view of the six-task factorial results from Tab. 3. Success rates (%).

Main-view targets	Wrist pathway	Fine-grained execution	Additional tasks	All six
RGB only	Unified future RGB	53.3	88.7	71.0
RGB+Mask	Unified future RGB	53.7	91.3	72.5
RGB+Skeleton	Unified future RGB	54.7	90.7	72.7
RGB+Mask+Skeleton	Unified future RGB	55.3	91.7	73.5
RGB only	Separate future latent	64.3	87.7	76.0
RGB+Mask	Separate future latent	65.0	90.7	77.8
RGB+Skeleton	Separate future latent	66.0	89.0	77.5
RGB+Mask+Skeleton	**Separate future latent**	**68.3**	**92.3**	**80.3**

The grouped results show that the separate wrist pathway primarily benefits the fine-grained tasks, while joint mask-and-skeleton supervision adds gains in both task groups. The full configuration achieves the highest average in each group.

A.5.1 Alternative View–Target Assignments

The chosen assignment follows what each stream shows consistently over the prediction horizon. The main camera usually retains the manipulated object, target region, and robot configuration in a common frame, supporting future masks and skeleton maps. Wrist views resolve local interaction details but may crop or occlude the broader task structure, so we use a latent target defined directly from each future wrist clip. The controlled wrist-pathway study compares RGB-reconstructive and latent future targets under the same lightweight pathway. This comparison tests target choice for the wrist stream but does not exhaust broader mappings, such as latent supervision for the main stream or structured pixel targets for wrist views. Such assignments also change target geometry, prediction heads, and loss formulation, and remain outside the present evaluation.

A.6 Action-Readout Diagnostics

The diagnostics use a separate six-task checkpoint grid comprising Adjust Bottle, Beat Block Hammer, Blocks Ranking RGB, Click Alarmclock, Click Bell, and Grab Roller. Eight disjoint evaluation episodes are sampled per task at fixed contact-adjacent phases, yielding 48 matched queries from the frozen checkpoints.

A.6.1 Attention readout

We extract action-query-to-main-view attention from the final two joint VideoDiT–ActionDiT attention blocks. Attention is averaged over heads, action tokens, denoising steps, and the two selected blocks, then renormalized over main-view video keys and mapped to the common spatial token grid. RR(O) is attention in the stage-relevant object region divided by attention in a size-matched background region. RR(G) is defined analogously for the executing-robot region. OCE is the distance between the attention centroid and object-region centroid in token-grid cells. Higher RR values and lower OCE indicate more localized main-view readout.

A.6.2 Matched K/V intervention

Object, robot, and background token sets are constructed on the same main-view grid. For each query, the selected K/V entries are replaced with the mean main-view K/V while the observation, language, proprioception, action noise, denoising settings, and wrist context remain fixed. R(O) is the action change caused by object-region replacement divided by the change from size-matched background replacement; R(G) is the corresponding robot-region ratio. Values above one indicate greater action-output sensitivity than matched background.

These interventions measure sensitivity of the generated action to the selected visual regions; they do not by themselves determine whether the changed action is more or less correct.

A.6.3 COMPLETE AGGREGATE DIAGNOSTICS

Table A10: Complete action-readout diagnostics grouped by wrist pathway. Higher RR and R values and lower OCE indicate stronger localized readout and action-output sensitivity.

Predictive design		Attention readout			K/V sensitivity	
Structured targets	**Wrist pathway**	**RR(O) ↑**	**RR(G) ↑**	**OCE ↓**	**R(O) ↑**	**R(G) ↑**
None	Unified future RGB	1.06	1.16	3.02	1.18	1.26
Mask	Unified future RGB	1.43	1.21	2.31	1.83	1.33
Skeleton	Unified future RGB	1.10	1.95	2.86	1.29	2.14
Mask+Skeleton	Unified future RGB	1.55	2.13	2.06	2.20	2.43
None	Separate future latent	1.10	1.20	2.95	1.24	1.30
Mask	Separate future latent	1.46	1.24	2.27	1.87	1.36
Skeleton	Separate future latent	1.13	2.00	2.82	1.34	2.19
Mask+Skeleton	**Separate future latent**	**1.57**	**2.16**	**2.03**	**2.24**	**2.47**

A.6.4 PER-TASK DIAGNOSTIC VALUES

Tab. A11 reports the task-wise values behind Tab. A10 for the RGB-only and RGB+mask+skeleton conditions under unified wrist modeling.

Table A11: Per-task diagnostics under unified future RGB; "M+S": mask and skeleton supervision.

Task	**RR(O)**		**RR(G)**		**R(O)**		**R(G)**	
	Base	M+S	Base	M+S	Base	M+S	Base	M+S
Adjust Bottle	1.02	1.49	1.12	2.00	1.12	2.14	1.18	2.34
Beat Block Hammer	1.05	1.54	1.18	2.17	1.17	2.21	1.29	2.47
Blocks Ranking RGB	1.12	1.62	1.19	2.25	1.27	2.35	1.32	2.58
Click Alarmclock	1.06	1.56	1.17	2.12	1.17	2.20	1.26	2.39
Click Bell	1.08	1.59	1.17	2.17	1.22	2.26	1.28	2.49
Grab Roller	1.03	1.50	1.13	2.07	1.13	2.04	1.23	2.31

Under unified future RGB, joint mask-and-skeleton supervision increases object- and robot-region readout and intervention sensitivity on all six diagnostic tasks. The aggregate results in Tab. A10 show the same component-specific pattern under both wrist pathways: mask prediction produces the larger object-side change, while skeleton prediction produces the larger robot-side change.