TACWM: UNIFIED WORLD MODELING OF VISION, TOUCH, ACTION, AND OBJECT POSE FOR DEXTEROUS MANIPULATION

Anonymous authors
Paper under double-blind review

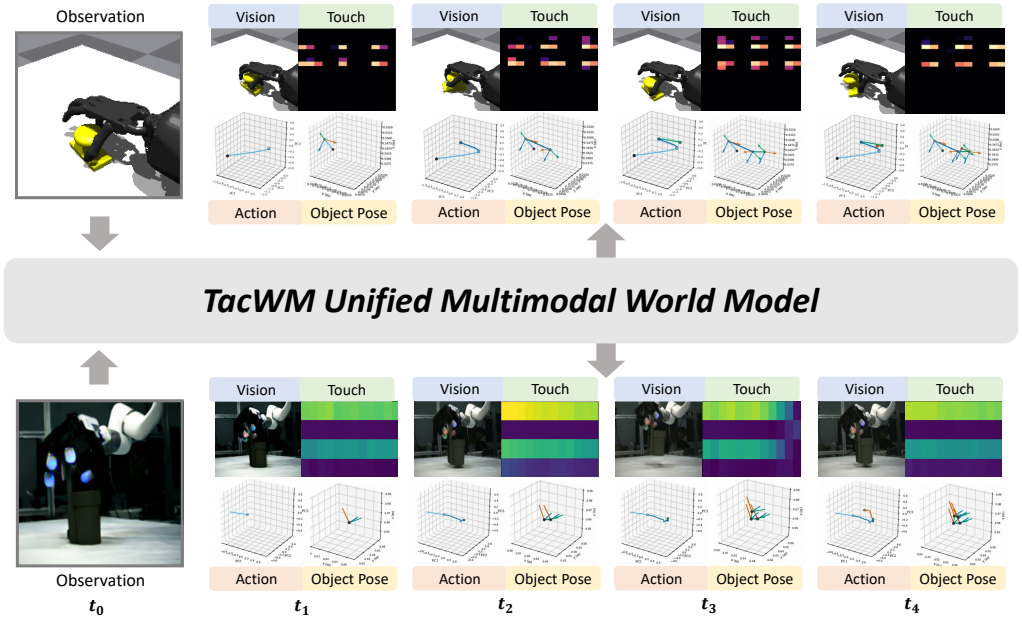

Figure 1: **Overview of TacWM.** TacWM is a multimodal world model that unifies vision, touch, action, and object pose modeling. Conditioned on historical context and a language instruction, it jointly predicts all four modalities over four future transitions, denoted by t_1–t_4.

ABSTRACT

High-performance dexterous manipulation requires accurate modeling of the contact dynamics between the hand and manipulated objects. However, existing methods either overlook the importance of tactile feedback in contact-rich tasks or focus solely on fitting action trajectories without modeling object status, thereby limiting manipulation performance. We introduce TacWM, a unified multimodal world model that jointly models visual observations, tactile feedback, actions, and object poses. Rather than modeling each modality independently, TacWM utilizes a unified diffusion transformer to model their joint distribution under causal temporal conditioning. Specifically, each modality is first encoded by a dedicated encoder, and TacWM is trained to predict the next-step representations of all modalities from multimodal history. This joint prediction allows different modalities to complement and reinforce one another. In particular, predicting future tactile feedback and object poses helps the model anticipate contact changes and object motion, thereby improving action generation for dexterous manipulation. To support large-scale training on heterogeneous datasets, we also propose a cross-embodiment mechanism that progressively aligns action and tactile signals from new embodiments in shared latent spaces. Extensive experiments demonstrate that TacWM achieves state-of-the-art performance in dexterous manipulation, marking a step toward general-purpose multimodal dexterous manipulation. Beyond enabling effective dexterous manipulation, TacWM can also effectively predict future videos, tactile feedback, and object poses.

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.

1 INTRODUCTION

Embodied intelligence has witnessed rapid progress recently thanks to the advances in policy learning and data availability. As one important part, dexterous manipulation serves as the interface between embodied systems and the physical world (Kannan et al., 2023; Xu et al., 2023). Vision-language-action (VLA) models form one major paradigm for general robot control, mapping visual observations and language instructions directly to actions with pretrained vision-language models (Brohan et al., 2023; Zitkovich et al., 2023; Driess et al., 2023; Kim et al., 2025; Black et al., 2025b). World-action models (WAMs) provide another by jointly predicting future visual states and actions, extending the long-standing idea of learning dynamics for control through imagined futures (Ha & Schmidhuber, 2018; Hafner et al., 2020; Ye et al., 2026a; Li et al., 2026a). Existing dexterous manipulation systems support in-hand reorientation, grasping and relocation, and learning from human demonstrations (OpenAI et al., 2020; Kannan et al., 2023; Xu et al., 2023; Qin et al., 2022).

Many learned manipulation policies rely mainly on visual observations, including action-chunking, diffusion, and VLA policies (Zhao et al., 2023; Chi et al., 2023; Kim et al., 2025). However, tactile feedback is often absent even though it provides direct evidence of contact, force, and grasp stability (Kappassov et al., 2015; Calandra et al., 2017; Lee et al., 2020; Guzey et al., 2023). Vision-only policies can therefore miss contact changes hidden by hand–object occlusion, while large visual backbones also create pressure on closed-loop inference latency (Guzey et al., 2023; Cai et al., 2026). Moreover, direct policy learning usually fits action trajectories without explicitly predicting how the manipulated object evolves (Zhao et al., 2023; Chi et al., 2023; Black et al., 2025b). Nonetheless, training a tactile-aware model is difficult because tactile data are limited and sensor-dependent (Higuera et al., 2024; Niu et al., 2026). The heterogeneity of robot hands further compounds the problem: platforms differ substantially in their action spaces, kinematics, and tactile sensor configurations (O'Neill et al., 2024; He et al., 2025; Kappassov et al., 2015).

To address these challenges, we introduce TacWM, a multimodal world model based on a unified diffusion transformer that jointly models vision, action, tactile feedback, and object poses. Tactile signals provide information about contact that may be unavailable from vision (Guzey et al., 2023; Niu et al., 2026), while explicit object pose modeling connects hand motion to manipulation outcomes. This supports a broader view of embodiment: an object coupled to the hand through contact becomes part of the dynamic system being modeled. By jointly predicting visual observations, tactile feedback, hand action, and object poses, TacWM captures a more complete account of dexterous manipulation dynamics. To tackle the rarity of tactile data, we unify data sources collected with heterogeneous dexterous hands and tactile sensors, following the broader direction of learning from multi-embodiment robot data (O'Neill et al., 2024; Ghosh et al., 2024; Niu et al., 2026). Specifically, we propose a cross-embodiment alignment mechanism that maps heterogeneous actions and tactile signals into shared latent spaces, allowing new robot embodiments and sensor configurations to be incorporated progressively. Extensive experiments on VTDexManip (Liu et al., 2025) demonstrate state-of-the-art performance on dexterous manipulation tasks. TacWM also supports multimodal generation through its joint modeling of vision, touch, action, and object pose. Our contributions are summarized as follows:

- We introduce TacWM, a unified world model of vision, touch, action, and object pose for dexterous manipulation. Tactile feedback helps capture contact-rich interactions, while object pose modeling incorporates the manipulated object into the hand–object dynamics.
- We propose a cross-embodiment alignment mechanism that maps heterogeneous action and tactile signals into shared latent spaces and supports the incremental integration of new robot embodiments and sensor configurations.
- We demonstrate state-of-the-art closed-loop dexterous manipulation performance, advancing toward more general-purpose dexterous manipulation.

2 RELATED WORK

Dexterous Manipulation. Dexterous manipulation spans learned in-hand control, grasp synthesis, and fine-tuning for precise hand–object interaction (OpenAI et al., 2020; Xu et al., 2023; Kannan et al., 2023). Direct policies include action chunking and diffusion (Zhao et al., 2023; Chi et al.,

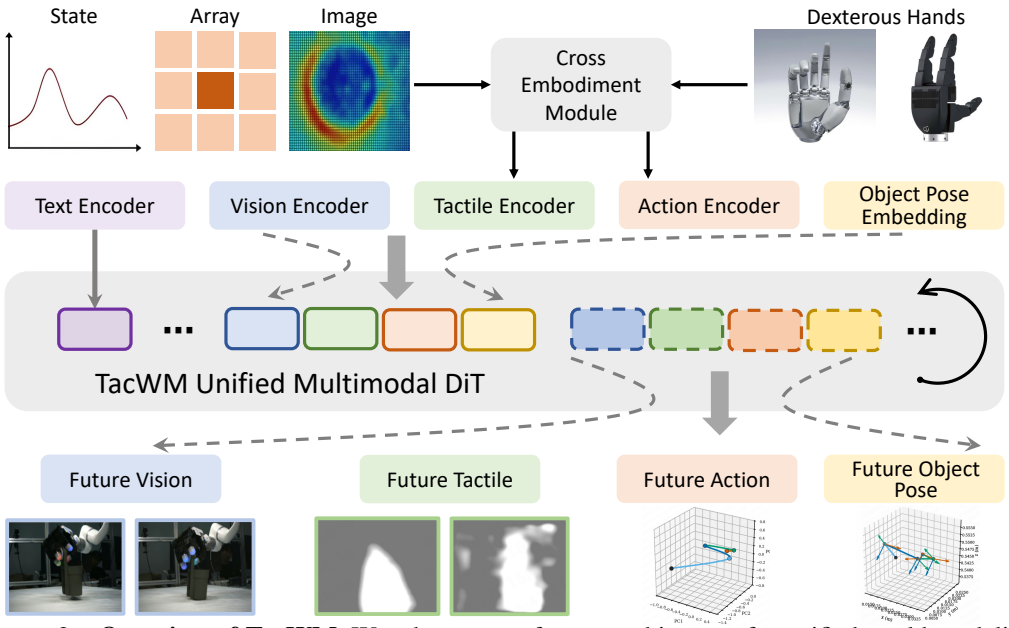

Figure 2: **Overview of TacWM.** We adopt a transformer architecture for unified world modeling of vision, touch, action and object pose. TacWM predicts the future observations of these modalities simultaneously from history context and language instruction. We also propose a cross-embodiment module to better utilize heterogeneous data from different platforms.

2023), while VLA models scale vision–language-conditioned control across data and model sizes (Brohan et al., 2023; Zitkovich et al., 2023; Kim et al., 2025; Black et al., 2025b); DexVLA adds a diffusion action expert (Wen et al., 2025). Human-video methods instead transfer motion priors through pose estimation, retargeting, or video pretraining (Qin et al., 2022; Shaw et al., 2023; Luo et al., 2025). Unlike direct action prediction, world models learn predictive dynamics in latent or visual spaces (Ha & Schmidhuber, 2018; Hafner et al., 2020; Wu et al., 2023; Wang et al., 2026). Video-based policies and world-action models generate future visual evolution before or jointly with actions (Du et al., 2023; Ye et al., 2026a; Li et al., 2026a). By predicting how observations change after an action, they expose physical constraints that imitation otherwise leaves implicit. For dexterous control, DexWM predicts future latent states, and VERA converts imagined videos into closed-loop actions (Goswami et al., 2025; Li et al., 2026b). However, vision provides only indirect evidence of force, slip, and grasp stability (Kappassov et al., 2015; Calandra et al., 2017).

Visual–tactile World Models. Touch exposes local pressure, force, and grasp stability when contact is occluded (Kappassov et al., 2015; Lambeta et al., 2020; Calandra et al., 2017). Visual–tactile research has progressed from joint representations and cross-modal prediction (Li et al., 2019; Lee et al., 2020; Guzey et al., 2023) to reusable pretraining across sensors and tasks (Higuera et al., 2024; Liu et al., 2024). VTDexManip provides a benchmark, while ViTacFormer predicts future touch for action generation (Liu et al., 2025; Heng et al., 2025). Predictive models jointly forecast vision and touch (Higuera et al., 2026), capture short-horizon contact or structured spatiotemporal dynamics (Zheng et al., 2026; Zhang et al., 2026b), and generate actions with tactile-aware attention (Wu et al., 2026). These methods improve physical consistency beyond static fusion, but generally leave object dynamics and pose implicit.

Cross Embodiment Alignment. Pooling data across embodiments increases interaction diversity, but incompatible action spaces and hardware dynamics hinder transfer (O'Neill et al., 2024; He et al., 2025; Ye et al., 2026b). Open X-Embodiment standardizes heterogeneous datasets, while Octo and π_0 support multiple observation and action spaces with shared backbones (O'Neill et al., 2024; Ghosh et al., 2024; Black et al., 2025b). Data collection faces this gap: DexPilot retargets human hand motion to one dexterous system, whereas AnyTeleop supports multiple arms, hands, simulators, and real platforms (Handa et al., 2020; Qin et al., 2023). Policy models use shared action experts, embodiment curricula, or latent action spaces (Wen et al., 2025; Jiang et al., 2026), while world models align dynamics through geometric hand and action representations (He et al., 2025). These representations must retain embodiment-specific information for executable commands, and heterogeneous tactile layouts remain a barrier (Kappassov et al., 2015; Higuera et al., 2024).

3 METHOD

3.1 TASK FORMULATION

Let O_t denote the world observation at time t and A_t the robot action. A manipulation world model learns the action-conditioned evolution

$$p_\theta(O_{t+1} \mid O_{\le t}, A_{\le t}), \tag{1}$$

to ground action generation in predicted consequences. Recent world-action models commonly set $O_t = V_t$ and learn dynamics through future video prediction (Ye et al., 2026a; Li et al., 2026a). Video captures scene geometry, appearance, and visible motion, but incompletely represents dexterous interaction. Contact and slip can remain hidden by hand–object occlusion, while the object's task-relevant state must be inferred indirectly from pixels.

TacWM instead represents scene evolution, hand–object contact, and object motion explicitly with $O_t = (V_t, T_t, P_t)$. Here, V_t provides global context and visible change, T_t captures local contact responses unavailable from vision, and P_t directly tracks the manipulation target. Their joint evolution links each action to its visual consequence, contact response, and task progress. Given history $X_{t-H:t-1} = (O_{t-H:t-1}, A_{t-H:t-1})$, current observation O_t, and instruction embedding E_{text}, TacWM models the next F steps jointly

$$p_\theta(A_{t:t+F-1},\, O_{t+1:t+F} \mid X_{t-H:t-1},\, O_t,\, E_{\text{text}}). \tag{2}$$

3.2 UNIFIED MULTIMODAL WORLD MODELING

Multimodal Tokenization and Flow Generation. TacWM realizes Eq. 2 with a shared generative transformer. A text encoder converts the instruction into language tokens. A frozen video VAE compresses RGB frames into visual latents, which are patchified into visual tokens. Embodiment-specific encoders map native actions and tactile streams into shared latent spaces. Their cross-embodiment alignment is detailed in Sec. 3.3. Object pose is represented by normalized 3D translation and 6D rotation and projected into a pose token. Modality and temporal embeddings identify the source and position of non-language tokens, while language tokens condition the model through cross-attention. TacWM uses flow matching to generate future vision, action, and pose. For each modality $m \in \{V, A, P\}$ and future transition k, training interpolates between a clean target $Z_0^{m,k}$ and Gaussian noise $\epsilon^{m,k}$,

$$Z^{m,k}_{\sigma_{m,k}} = (1-\sigma_{m,k})Z_0^{m,k} + \sigma_{m,k}\epsilon^{m,k}, \qquad U^{m,k} = \epsilon^{m,k} - Z_0^{m,k}. \tag{3}$$

The noise levels $\sigma_{m,k}$ are sampled independently across modalities and future transitions. Different parts of the predicted future can therefore appear at different corruption levels, requiring the Transformer to use cross-modal and temporal context instead of relying on synchronized denoising. It predicts $U^{m,k}$ from noisy future tokens, multimodal history, language, and the corresponding noise level. Future tactile prediction uses learned queries and directly regresses clean tactile latents. Because all future modalities interact inside the same transformer, each velocity or tactile prediction is conditioned on the evolving joint future. Training applies validity-masked squared error to the flow velocities and tactile latents, normalizes each modality by its valid targets, and averages the available modality losses. At inference, vision, action, and pose start from Gaussian noise and follow the learned velocity field toward clean samples over decreasing noise levels. Tactile latents are read from the final joint state, after which modality-specific decoders recover RGB frames, native actions, tactile signals, and camera-frame object poses.

Historical and Future Token Interaction. TacWM separates the observed history H from the generated future F with a structured attention mask. With rows denoting query tokens and columns denoting source tokens in the order $[H, F]$, the permitted interactions are

$$\mathcal{A}_{[H,F]\leftarrow[H,F]} = \begin{bmatrix} \mathcal{C}_{H\times H} & 0_{H\times F} \\ 1_{F\times H} & 1_{F\times F} \end{bmatrix}, \qquad (\mathcal{C}_{H\times H})_{ij} = \begin{cases} 1, & \tau(j) \le \tau(i), \\ 0, & \tau(j) > \tau(i), \end{cases} \tag{4}$$

where one denotes an allowed interaction, zero denotes a masked interaction, and $\tau(\cdot)$ is the transition index. Historical tokens can read all modalities in their current or earlier transitions, but

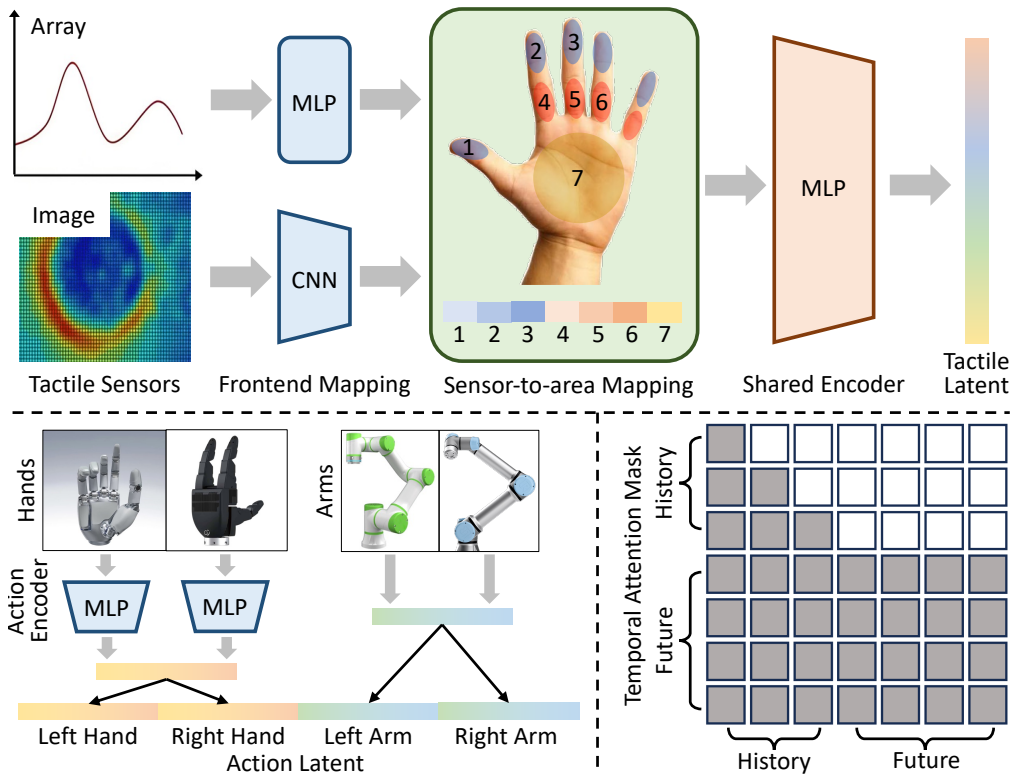

Figure 3: **Details for the cross-embodiment module and temporal attention mask.** Upper, The tactile alignment module uses separate frontend mapping layers for different tactile signals, maps different sensor installation to the canonical setup and derives the tactile latent with a shared encoder. Bottom Left, The action alignment module first maps different hand actions with seperate encoders and assemble the hand and arm actions to form the action latent. Bottom right, we adopt causal attention for history tokens and bidirectional attention among future tokens. Gray regions indicate allowed attention, while blank regions indicate masked attention.

never later transitions or future targets. This causal prefix is compatible with KV caching because its cached keys and values remain unchanged when observations are appended. Future tokens read the full history and interact bidirectionally across all future modalities and timesteps, promoting continuous video, consistent contact evolution, and mutually compatible action–object trajectories. Validity masks further remove placeholders introduced by heterogeneous data. Inactive action slots and unavailable tactile region–time positions are zeroed and excluded from their losses. Missing pose and padding tokens are removed from attention, and missing pose targets are excluded from training. Samples with different action dimensions, sensor counts, or available modalities can therefore share the same token layout without treating absent signals as observations.

3.3 Cross-Embodiment Alignment

Robot hands differ in joint structure, control commands, tactile hardware, and sensor placement. Directly mixing their raw signals would force the world model to learn both platform conversion and interaction dynamics. TacWM separates these roles: frozen cross-embodiment codecs translate native actions and tactile observations into common representations, while the shared Transformer focuses on how manipulation evolves. A new embodiment can therefore be added through its codecs, masks, and identity embedding without changing the world-model architecture.

Cross-Embodiment Action Alignment. Native actions first undergo embodiment-specific normalization so that different joint ranges have comparable scales. We then separate arm and hand controls. Arm commands retain their physical control representation, whereas the variable-dimensional joints of each hand are compressed by an autoencoder into a fixed 32-dimensional latent, following XL-VLA (Jiang et al., 2026). These components are placed into a common 78-dimensional layout consisting of left-arm, left-hand, right-arm, and right-hand slots. The complete mapping is

Table 1: **Overall closed-loop success rate (%) on VTDexManip.** Avg. is the macro average across the five tasks. **Bold** and underline indicate the best and second-best results, respectively.

Method	ReorientDown	ReorientUp	BottleCap	Sliding	Handover	Avg.
$\pi_{0.5}$ Black et al. (2025a)	20.7	23.7	11.0	**97.0**	16.0	33.7
ViTacFormer (Heng et al., 2025)	26.7	27.7	8.3	83.0	15.0	32.1
T-Rex (Niu et al., 2026)	24.3	25.3	18.0	79.0	15.0	32.3
TacWM (Ours)	**29.0**	**31.0**	**19.3**	86.0	**17.0**	**36.5**

summarized as

$$z^A = M^A \odot \Pi^A\big(a_{\text{arm}}, \mathcal{E}^A(a_{\text{hand}})\big) \in \mathbb{R}^{78}, \quad \hat{a} = \mathcal{D}^A(z^A), \tag{5}$$

where $\mathcal{E}^A$ compresses the hand controls, Π^A places arm and hand components into their shared slots, and M^A keeps only the slots used by the specific embodiment. An embodiment embedding identifies the source platform. The decoder $\mathcal{D}^A$ reads the generated active slots, reconstructs native hand commands, combines them with arm controls, and restores their physical scales. This design shares action structure while preserving control interface of each robot.

Cross-Embodiment Tactile Alignment. Tactile alignment must handle different signal types and sensor layouts. TacWM uses source-specific frontends for image-based tactile observations, scalar contact forces, and six-axis force–torque measurements. Each frontend maps its native stream to the same latent width, while a shared functional map preserves sensor meaning. Every physical sensor is assigned to one of 48 regions distinguishing the two hands and locations such as fingertips, finger segments, and palms. We retain 12 temporal positions within each transition instead of pooling them, preserving brief events such as contact onset and release. The tactile mapping is

$$Z^T = M^T \odot \mathcal{E}^T_{\text{can}}\big(\mathcal{F}^T(T) + R_\rho + Q_{\text{time}}\big), \quad \hat{T} = \mathcal{D}^T(Z^T), \tag{6}$$

where $\mathcal{F}^T$ is the source-specific frontend, ρ maps physical sensors to functional regions, and R and Q encode region and time. The canonical encoder $\mathcal{E}^T_{\text{can}}$ produces valid region–time latents selected by M^T, and $\mathcal{D}^T$ reconstructs native signals. TacWM predicts these canonical latents rather than hardware-specific readings, therefore a new embodiment can be integrated with its frontend, region mapping, and decoder.

3.4 Closed-Loop Inference

At each replanning step, TacWM conditions on the recent observation–action history, the current multimodal observation, and the language instruction. It initializes future visual, action, and object-pose representations from noise, introduces tactile query tokens, and jointly samples their future trajectories. The action decoder then converts the generated canonical action representation into commands for the current robot, while the remaining decoders recover the corresponding visual, tactile, and object-state predictions.

TacWM uses these predictions in a receding-horizon control loop. Rather than executing the entire imagined trajectory, the controller applies only the immediate action segment and then observes the resulting world state. The executed action and newly measured observations are appended to the history before generating another future. Predicted vision, touch, and object motion therefore guide action generation within each planning step, while real feedback anchors the next step. This repeated observe–predict–act cycle limits the accumulation of open-loop prediction errors and allows the policy to adjust when the physical interaction differs from its imagined evolution.

4 Experiments

We evaluate TacWM on closed-loop dexterous manipulation and future tactile prediction, covering both control performance and world-modeling accuracy. We also conduct ablation study to verify the effectiveness of the cross-embodiment module and the contribution of each modality.

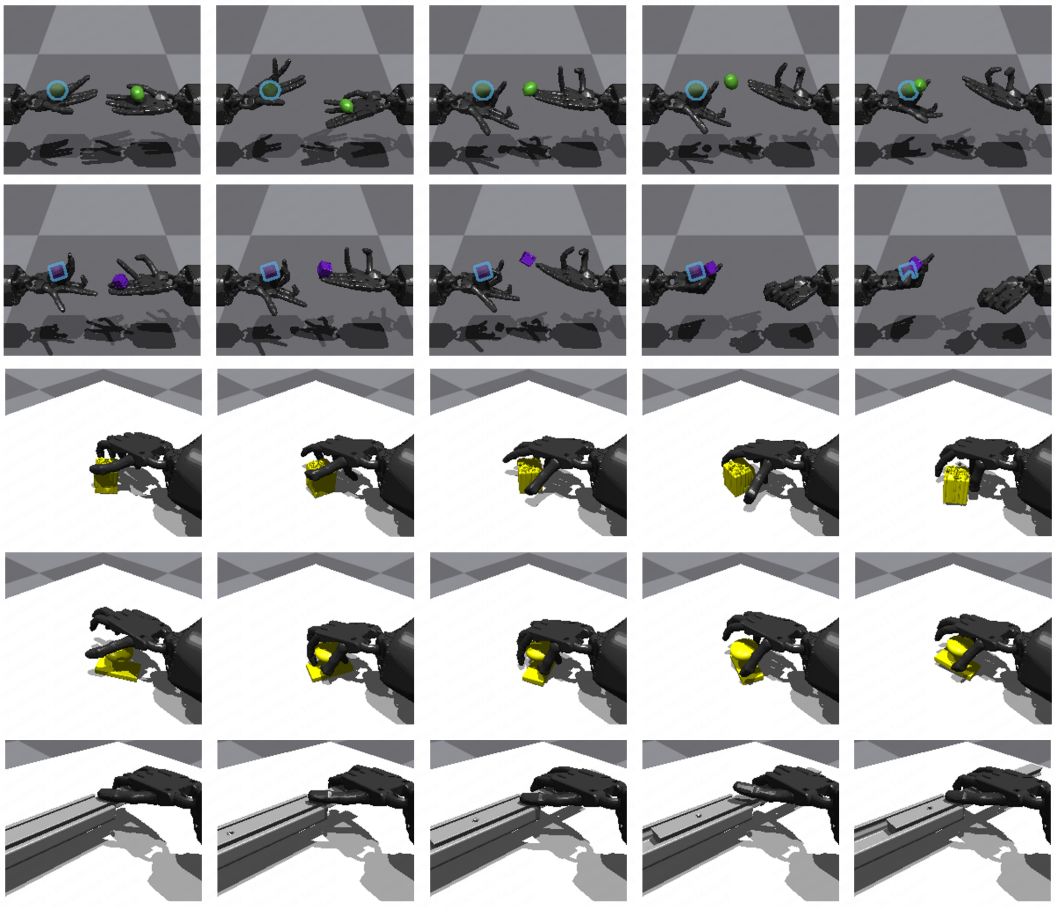

Figure 4: **Qualitative examples of closed-loop manipulation with TacWM**. The first two rows show HandOver: tossing an object from the right hand to the left hand to reach the target position outlined in light blue. The third and fourth rows show ReorientDown: rotating an object in hand by a target angle. The last row shows Sliding: sliding the upper component of an articulated object outward along its rail relative to the base. Time progresses from left to right.

4.1 Datasets and Implementation Details

We initialize the shared transformer backbone from LingBot-VA (Li et al., 2026a) and pretrain TacWM for 100k steps on DataDex (..., 2026), HRDexDB (Lim et al., 2026), and T-Rex (Niu et al., 2026), sampled at a ratio of 3:1:6. Pretraining takes 40 hours on eight NVIDIA A800 GPUs. Their visual observations, actions, tactile signals, and object poses are organized into a common six-transition history and four-transition future, with 12 control frames per transition. DataDex contributes image-based touch, HRDexDB provides scalar tactile signals, and T-Rex provides six-axis force–torque signals. We use ProxyPose(Zhang et al., 2026a) to annotate camera-frame object poses for DataDex and T-Rex. Their action and tactile autoencoders are kept frozen after reconstruction training, while only the multimodal transformer is trainable. From this pretrained checkpoint, we fine-tune task-specific models on successful VTDexManip (Liu et al., 2025) expert trajectories.

4.2 Closed-Loop Dexterous Manipulation

Benchmark and Evaluation. We evaluate closed-loop manipulation on five VTDexManip(Liu et al., 2025) tasks: ReorientDown, ReorientUp, BottleCap, Sliding, and HandOver. ReorientDown succeeds when the object orientation is within 0.1 rad of the target orientation, while ReorientUp requires a relative rotation of more than 3 rad about the object's initial z axis. BottleCap requires the cap joint to rotate by more than 6 rad, Sliding requires the prismatic joint displacement to exceed 0.15 m, and HandOver succeeds when the object is within 0.03 m of the transfer target. We use each environment's native success flag and report the fraction of successful closed-loop episodes.

Table 2: **Future tactile prediction on open-loop pretraining datasets. Bold** indicates the better result. TacWM consistently outperforms the persistent baseline on all three datasets.

Dataset	Method	Latent MSE ↓	Decoded MAE ↓
DataDex (..., 2026)	Persistence	0.042	0.179
	TacWM (Ours)	**0.019**	**0.112**
HRDexDB (Lim et al., 2026)	Persistence	0.448	0.472
	TacWM (Ours)	**0.301**	**0.378**
T-Rex (Niu et al., 2026)	Persistence	0.064	0.212
	TacWM (Ours)	**0.044**	**0.195**

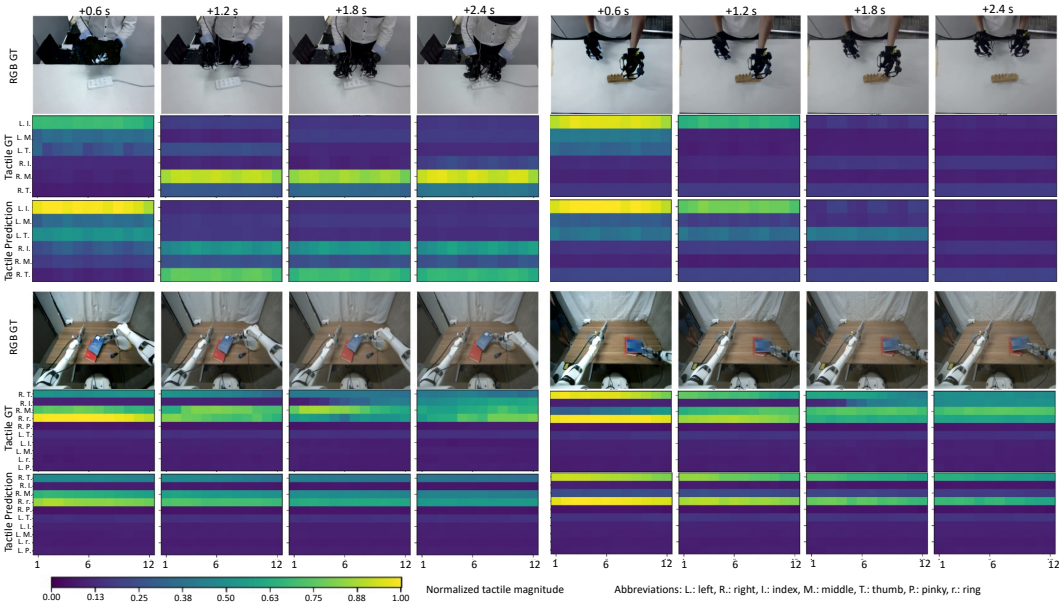

Figure 5: **Qualitative results of future tactile prediction.** For each case, the three rows show ground-truth vision, ground-truth tactile signals, and predicted tactile signals, respectively. Future visual observations are shown only for reference and are not used to condition tactile prediction. Tactile intensities are normalized to $[0, 1]$ within each case.

ReorientDown, ReorientUp, and BottleCap each contain 10 seen and 5 unseen evaluation objects. Sliding contains 3 objects with successful task-specific expert trajectories and 2 held-out objects, which we refer to as seen and unseen, respectively. HandOver contains 5 seen objects with successful expert trajectories and has no valid unseen split. One BottleCap object in the seen split of the benchmark has no successful demonstration and is excluded from task-specific fine-tuning but retained in evaluation. We evaluate 20 episodes per object, yielding 300 episodes for each of ReorientDown, ReorientUp, and BottleCap, and 100 episodes for Sliding and HandOver. The simulator runs at 60 Hz, while observations and actions are organized at 20 Hz. We use our standard predict-four-execute-one closed-loop inference protocol. We report the overall success rate for each task, and Avg. denotes the macro average across the evaluated tasks.

Baselines. We compare with three task-specific policies: $\pi_{0.5}$(Black et al., 2025a), a vision–language–action model; ViTacFormer(Heng et al., 2025), a visuo-tactile policy with a tactile prediction objective; and T-Rex(Niu et al., 2026), which uses tactile feedback to refine action predictions. Each baseline is evaluated in the same closed-loop environment using its supported modalities.

Results. Table 1 reports the overall closed-loop success rates. TacWM achieves the best performance on ReorientDown, ReorientUp, and BottleCap, and obtains the highest five-task macro average. These results show that jointly modeling vision, tactile, actions, and object poses leads to stronger closed-loop manipulation performance than vision-only and visuo-tactile baselines.

Table 3: **Overall closed-loop success rate (%) for ablation studies.** Modality ablations remove the specified modality during pretraining and task-specific fine-tuning. Shared Encoder shares each modality's parameters across embodiments. Avg. is the macro average across five tasks; "—" denotes unreported results. **Bold** and underline indicate the best and second-best results, respectively.

Variant	ReorientDown	ReorientUp	BottleCap	Sliding	HandOver	Avg.
w/o Vision	27.3	23.7	10.7	80.0	**18.0**	31.9
w/o Tactile	<u>29.0</u>	27.0	16.7	85.0	12.0	33.9
w/o Object Pose	**30.3**	<u>27.7</u>	<u>17.3</u>	**91.0**	9.0	<u>35.1</u>
Shared Encoder	—	26.3	16.3	—	—	—
TacWM (Ours)	<u>29.0</u>	**31.0**	**19.3**	<u>86.0</u>	<u>17.0</u>	**36.5**

4.3 Tactile Prediction

We assess the ability of TacWM to predict future tactile signals. Given six observed transitions and a language instruction, the model jointly generates four future transitions without access to ground-truth future modalities. We evaluate tactile predictions against held-out trajectories and compare them with a persistence baseline that repeats the final observed tactile state.

We measure tactile latent MSE in the standardized canonical latent space and decoded tactile MAE in the tactile signal space of the platform. Metrics are computed at valid positions over the four prediction horizons. As shown in Table 2, TacWM consistently outperforms the persistent baseline on both metrics across all three datasets.

4.4 Ablation Studies

Modality Ablations. We remove one modality at a time from the full TacWM model, including both historical inputs and future prediction targets, while keeping all other architecture and training settings unchanged throughout pretraining and task-specific fine-tuning. Table 3 shows that the full TacWM obtains the highest macro average across the five tasks. Removing vision causes the largest overall drop, especially on ReorientUp and BottleCap. Removing tactile feedback also lowers the macro average, while removing object pose has a smaller aggregate effect but substantially hurts Handover. Overall, the four-modality design achieves the strongest average performance.

Cross-Embodiment Alignment. We compare our alignment design with a shared-encoder variant on ReorientUp and BottleCap. This variant shares action and tactile encoders across embodiments without embodiment-specific parameters. TacWM improves success rates by 4.7 and 3.0 percentage points, respectively, supporting our alignment design.

5 Conclusion

We introduced TacWM, a multimodal world model for dexterous manipulation that jointly represents vision, tactile feedback, robot actions, and object poses. Unlike action-only policies or visually grounded world models, TacWM learns the coupled evolution of contact, object motion, scene appearance, and control. A unified Transformer trained with flow matching generates future modalities from multimodal history and language instructions, while structured attention preserves causal historical processing and allows future tokens to interact for coherent prediction. To exploit heterogeneous robot data, TacWM further aligns embodiment-specific action spaces and tactile sensor layouts into shared latent representations without changing the world-model backbone. Experiments show that this joint formulation benefits both control and prediction. On five VTDexManip tasks, TacWM achieves the highest average closed-loop success rate and leads on three tasks, establishing state-of-the-art overall performance among the evaluated methods. It also predicts future tactile signals more accurately than a persistence baseline across DataDex, HRDexDB, and T-Rex, despite their distinct tactile sensing formats. Modality ablations further show that the complete model provides the strongest average manipulation performance. Together, these results demonstrate that modeling what the robot will do alongside how contact, objects, and visual scenes will evolve provides a practical foundation for contact-rich dexterous manipulation across embodiments.

AI USE STATEMENT

In this work, we have not used generative AI tools for any tasks with required disclosure, and all such tasks are not applicable to this work. We used generative AI tools only to polish the writing of parts of the manuscript. The authors reviewed all AI-assisted edits to ensure that they preserve the intended technical meaning and contain no unsupported claims. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, Sections 3, 4.1 and 4.2 of the main paper provide detailed descriptions of our methodology and implementation. We will release the complete source code and model weights.

REFERENCES

... Tacexogaze: A human-centric glove interface for multi-view vision-tactile data collection in dexterous manipulation, 2026. Preprint forthcoming.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025a. URL https://arxiv.org/abs/2504.16054.

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 *RSS*, 2025b. doi: 10.15607/RSS.2025.XXI.010. URL https://roboticsproceedings.org/rss21/p010.html.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael S. Ryoo, Grecia Salazar, Pannag R. Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan H. Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-1: Robotics transformer for real-world control at scale. In *RSS*, 2023. doi: 10.15607/RSS.2023.XIX.025. URL https://roboticsproceedings.org/rss19/p025.html.

Rui Cai, Jun Guo, Xinze He, Piaopiao Jin, Jie Li, Bingxuan Lin, Futeng Liu, Wei Liu, Fei Ma, Kun Ma, Feng Qiu, Heng Qu, Yifei Su, Qiao Sun, Dong Wang, Donghao Wang, Yunhong Wang, Rujie Wu, Diyun Xiang, Yu Yang, Hangjun Ye, Yuan Zhang, and Quanyun Zhou. Xiaomi-Robotics-0: An open-sourced vision-language-action model with real-time execution. *arXiv preprint arXiv:2602.12684*, 2026. URL https://arxiv.org/abs/2602.12684.

Roberto Calandra, Andrew Owens, Manu Upadhyaya, Wenzhen Yuan, Justin Lin, Edward H. Adelson, and Sergey Levine. The feeling of success: Does touch sensing help predict grasp outcomes? In *CoRL*, volume 78 of *Proceedings of Machine Learning Research*, pp. 314–323, 2017. URL https://proceedings.mlr.press/v78/calandra17a.html.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin C. M. Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *RSS*, 2023. doi: 10.15607/RSS.2023.XIX.026. URL https://roboticsproceedings.org/rss19/p026.html.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An embodied multimodal language model. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8469–8488, 2023. URL https://proceedings.mlr.press/v202/driess23a.html.

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 *NeurIPS*, volume 36, 2023. doi: 10.52202/075280-0403. URL https://papers.nips.cc/paper_files/paper/2023/hash/1d5b9233ad716a43be5c0d3023cb82d0-Abstract-Conference.html.

Dibya Ghosh, Homer Rich Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Quan Vuong, Ted Xiao, Pannag R. Sanketi, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *RSS*, 2024. doi: 10.15607/RSS.2024.XX.090. URL https://roboticsproceedings.org/rss20/p090.html.

Raktim Gautam Goswami, Amir Bar, David Fan, Tsung-Yen Yang, Gaoyue Zhou, Prashanth Krishnamurthy, Michael Rabbat, Farshad Khorrami, and Yann LeCun. World models for learning dexterous hand-object interactions from human videos. *arXiv preprint arXiv:2512.13644*, 2025. URL https://arxiv.org/abs/2512.13644.

Irmak Guzey, Ben Evans, Soumith Chintala, and Lerrel Pinto. Dexterity from touch: Self-supervised pre-training of tactile representations with robotic play. In *CoRL*, volume 229 of *Proceedings of Machine Learning Research*, pp. 3142–3166, 2023. URL https://proceedings.mlr.press/v229/guzey23a.html.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *NeurIPS*, volume 31, 2018. URL https://papers.nips.cc/paper/7512-recurrent-world-models-facilitate-policy-evolution.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *ICLR*, 2020. URL https://openreview.net/forum?id=S1lOTC4tDS.

Ankur Handa, Karl Van Wyk, Wei Yang, Jacky Liang, Yu-Wei Chao, Qian Wan, Stan Birchfield, Nathan Ratliff, and Dieter Fox. DexPilot: Vision-based teleoperation of dexterous robotic hand-arm system. In *ICRA*, pp. 9164–9170, 2020. doi: 10.1109/ICRA40945.2020.9197124.

Zihao He, Bo Ai, Tongzhou Mu, Yulin Liu, Weikang Wan, Jiawei Fu, Yilun Du, Henrik I. Christensen, and Hao Su. Scaling cross-embodiment world models for dexterous manipulation. *arXiv preprint arXiv:2511.01177*, 2025. URL https://arxiv.org/abs/2511.01177.

Liang Heng, Haoran Geng, Kaifeng Zhang, Pieter Abbeel, and Jitendra Malik. ViTacFormer: Learning cross-modal representation for visuo-tactile dexterous manipulation. *arXiv preprint arXiv:2506.15953*, 2025. URL https://arxiv.org/abs/2506.15953.

Carolina Higuera, Akash Sharma, Chaithanya Krishna Bodduluri, Taosha Fan, Patrick Lancaster, Mrinal Kalakrishnan, Michael Kaess, Byron Boots, Mike Lambeta, Tingfan Wu, and Mustafa Mukadam. Sparsh: Self-supervised touch representations for vision-based tactile sensing. *arXiv preprint arXiv:2410.24090*, 2024. URL https://arxiv.org/abs/2410.24090.

Carolina Higuera, Sergio Arnaud, Byron Boots, Mustafa Mukadam, Francois Robert Hogan, and Franziska Meier. Visuo-tactile world models. *arXiv preprint arXiv:2602.06001*, 2026. URL https://arxiv.org/abs/2602.06001.

Guangqi Jiang, Yutong Liang, Jianglong Ye, Jia-Yang Huang, Changwei Jing, Rocky Duan, Pieter Abbeel, Xiaolong Wang, and Xueyan Zou. Cross-hand latent representation for vision-language-action models. *arXiv preprint arXiv:2603.10158*, 2026. URL https://arxiv.org/abs/2603.10158.

Aditya Kannan, Kenneth Shaw, Shikhar Bahl, Pragna Mannam, and Deepak Pathak. DEFT: Dexterous fine-tuning for hand policies. In *CoRL*, volume 229 of *Proceedings of Machine Learning Research*, pp. 928–942, 2023. URL https://proceedings.mlr.press/v229/kannan23a.html.

Zhanat Kappassov, Juan-Antonio Corrales, and Véronique Perdereau. Tactile sensing in dexterous robot hands—review. *Robot. Auton. Syst.*, 74:195–220, 2015. doi: 10.1016/j.robot.2015.07.015.

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 *CoRL*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713, 2025. URL https://proceedings.mlr.press/v270/kim25c.html.

Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian H. Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, Dinesh Jayaraman, and Roberto Calandra. DIGIT: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robot. Autom. Lett.*, 5(3):3838–3845, 2020. doi: 10.1109/LRA.2020.2977257.

Michelle A. Lee, Yuke Zhu, Peter Zachares, Matthew Tan, Krishnan Srinivasan, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Learning multimodal representations for contact-rich tasks. *IEEE Trans. Robot.*, 36(3):582–596, 2020. doi: 10.1109/TRO.2019.2959445.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, Yujun Shen, and Yinghao Xu. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026a. URL https://arxiv.org/abs/2601.21998.

Sizhe Lester Li, Evan Kim, Xingjian Bai, Tong Zhao, Tao Pang, Max Simchowitz, and Vincent Sitzmann. Turning video models into generalist robot policies. *arXiv preprint arXiv:2605.27817*, 2026b. URL https://arxiv.org/abs/2605.27817.

Yunzhu Li, Jun-Yan Zhu, Russ Tedrake, and Antonio Torralba. Connecting touch and vision via cross-modal prediction. In *CVPR*, pp. 10609–10618, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Li_Connecting_Touch_and_Vision_via_Cross-Modal_Prediction_CVPR_2019_paper.html.

Jongbin Lim, Taeyun Ha, Mingi Choi, Jisoo Kim, Byungjun Kim, Subin Jeon, and Hanbyul Joo. HRDexDB: A large-scale dataset of dexterous human and robotic hand grasps. *arXiv preprint arXiv:2604.14944*, 2026. URL https://arxiv.org/abs/2604.14944.

Qingtao Liu, Qi Ye, Zhengnan Sun, Yu Cui, Gaofeng Li, and Jiming Chen. Masked visual-tactile pre-training for robot manipulation. In *ICRA*, pp. 13859–13875, 2024. doi: 10.1109/ICRA57147.2024.10610933.

Qingtao Liu, Yu Cui, Zhengnan Sun, Gaofeng Li, Jiming Chen, and Qi Ye. VTDexManip: A dataset and benchmark for visual-tactile pretraining and dexterous manipulation with reinforcement learning. In *ICLR*, 2025. URL https://openreview.net/forum?id=jf7C7EGw21.

Hao Luo, Yicheng Feng, Wanpeng Zhang, Sipeng Zheng, Ye Wang, Haoqi Yuan, Jiazheng Liu, Chaoyi Xu, Qin Jin, and Zongqing Lu. Being-H0: Vision-language-action pretraining from large-scale human videos. *arXiv preprint arXiv:2507.15597*, 2025. URL https://arxiv.org/abs/2507.15597.

Dantong Niu, Zhuoyang Liu, Zekai Wang, Boning Shao, Zhao-Heng Yin, Anirudh Pai, Yuvan Sharma, Stefano Saravalle, Ruijie Zheng, Jing Wang, Ryan Punamiya, Mengda Xu, Yuqi Xie, Yunfan Jiang, Letian Fu, Konstantinos Kallidromitis, Matteo Gioia, Junyi Zhang, Jiaxin Ge, Haiwen Feng, Fabio Galasso, Wei Zhan, David M. Chan, Yutong Bai, Roei Herzig, Jiahui Lei, Fei-Fei Li, Ken Goldberg, Jitendra Malik, Pieter Abbeel, Yuke Zhu, Danfei Xu, Linxi Jim Fan, and Trevor Darrell. T-Rex: Tactile-reactive dexterous manipulation. *arXiv preprint arXiv:2606.17055*, 2026. URL https://arxiv.org/abs/2606.17055.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open X-embodiment: Robotic learning datasets and RT-X models. In *ICRA*, pp. 6892–6903, 2024. doi: 10.1109/ICRA57147.2024.10611477. URL https://arxiv.org/abs/2310.08864.

OpenAI, Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, Jonas Schneider, Szymon Sidor, Josh Tobin, Peter Welinder, Lilian Weng, and Wojciech Zaremba. Learning dexterous in-hand manipulation. *Int. J. Robot. Res.*, 39(1):3–20, 2020. doi: 10.1177/0278364919887447. URL https://doi.org/10.1177/0278364919887447.

Yuzhe Qin, Yueh-Hua Wu, Shaowei Liu, Hanwen Jiang, Ruihan Yang, Yang Fu, and Xiaolong Wang. DexMV: Imitation learning for dexterous manipulation from human videos. In *ECCV*, volume 13699 of *Lecture Notes in Computer Science*, pp. 570–587, 2022. doi: 10.1007/978-3-031-19842-7_33.

Yuzhe Qin, Wei Yang, Binghao Huang, Karl Van Wyk, Hao Su, Xiaolong Wang, Yu-Wei Chao, and Dieter Fox. AnyTeleop: A general vision-based dexterous robot arm-hand teleoperation system. In *RSS*, 2023. doi: 10.15607/RSS.2023.XIX.015. URL https://roboticsproceedings.org/rss19/p015.html.

Kenneth Shaw, Shikhar Bahl, and Deepak Pathak. VideoDex: Learning dexterity from internet videos. In *CoRL*, volume 205 of *Proceedings of Machine Learning Research*, pp. 654–665, 2023. URL https://proceedings.mlr.press/v205/shaw23a.html.

Fangyuan Wang, Ziyuan Wang, Guorui Pei, Mengshi Zhang, Canxi Liang, Jun Hu, Zhongxuan Li, Jinsong Wu, Ning Han, Zeqing Zhang, Jiaming Qi, Hongmin Wu, Shiyao Zhang, Pai Zheng, Jia Pan, David Navarro-Alarcon, Sichao Liu, and Peng Zhou. World models for robotic manipulation: A survey. *SmartBot*, pp. e70053, 2026. doi: 10.1002/smb2.70053.

Junjie Wen, Yichen Zhu, Jinming Li, Zhibin Tang, Chaomin Shen, and Feifei Feng. DexVLA: Vision-language model with plug-in diffusion expert for general robot control. In *CoRL*, volume 305 of *Proceedings of Machine Learning Research*, pp. 3094–3114, 2025. URL https://proceedings.mlr.press/v305/wen25b.html.

Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. DayDreamer: World models for physical robot learning. In *CoRL*, volume 205 of *Proceedings of Machine Learning Research*, pp. 2226–2240, 2023. URL https://proceedings.mlr.press/v205/wu23c.html.

Siyu Wu, Linjing You, Junjie Zhu, Yaozu Liu, Kaixiang Huang, Yonghang Chen, Jituo Li, Changhao Zhang, Jian Liu, Hengshuo Chu, Qi Li, and Hengshuang Zhao. Tactile-WAM: Touch-aware world action model with tactile asymmetric attention. *arXiv preprint arXiv:2606.26663*, 2026. URL https://arxiv.org/abs/2606.26663.

Yinzhen Xu, Weikang Wan, Jialiang Zhang, Haoran Liu, Zikang Shan, Hao Shen, Ruicheng Wang, Haoran Geng, Yijia Weng, Jiayi Chen, Tengyu Liu, Li Yi, and He Wang. UniDexGrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In *CVPR*, pp. 4737–4746, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Xu_UniDexGrasp_Universal_Robotic_Dexterous_Grasping_via_Learning_Diverse_Proposal_Generation_CVPR_2023_paper.html.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi Jim Fan, and Joel Jang. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026a. URL https://arxiv.org/abs/2602.15922.

Yifan Ye, Yankai Fu, Yaoxu Lv, Bohan Hou, Jun Cen, Lingdong Kong, Duo Zheng, Tianxing Chen, Jiaming Liu, Ziang Cao, Yunfan Lou, Wei Chow, Xian Sun, Yingshuo Wang, Kuangzhi Ge, Xiaowei Chi, Xidong Zhang, Zhibo Pang, Yiwu Zhong, Sirui Han, Zhihe Lu, Weihao Yuan, Qifeng Chen, Michael Yu Wang, Yao Mu, Ziwei Liu, Jianfei Yang, Ping Luo, and Shanghang Zhang. Data pyramid for embodied manipulation: A survey. *arXiv preprint arXiv:2607.24744*, 2026b. URL https://arxiv.org/abs/2607.24744.

Ruihang Zhang, Felix Taubner, Pooja Ravi, Kiriakos N Kutulakos, and David B Lindell. Proxypose: 6-dof pose tracking via video-to-video translation. *arXiv preprint arXiv:2607.06555*, 2026a.

Zhiyuan Zhang, Pokuang Zhou, Kaidi Zhang, Adeesh Desai, Temitope Amosa, Davood Soleymanzadeh, Jiuzhou Lei, Minghui Zheng, and Yu She. ContactWorld: What matters in vision-tactile world models for contact-rich manipulation. *arXiv preprint arXiv:2606.13877*, 2026b. URL https://arxiv.org/abs/2606.13877.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *RSS*, 2023. doi: 10.15607/RSS.2023.XIX.016. URL https://roboticsproceedings.org/rss19/p016.html.

Yuhang Zheng, Songen Gu, Yupeng Zheng, Weize Li, Yujie Zang, Shuai Tian, Xiang Li, Ce Hao, Chen Gao, Si Liu, Haoran Li, Yilun Chen, Shuicheng Yan, and Wenchao Ding. OmniVTA: Visuotactile world modeling for contact-rich robotic manipulation. *arXiv preprint arXiv:2603.19201*, 2026. URL https://arxiv.org/abs/2603.19201.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *CoRL*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183, 2023. URL https://proceedings.mlr.press/v229/zitkovich23a.html.