TouchWorld: A Predictive and Reactive Tactile Foundation Model for Dexterous Manipulation

Anonymous authors
Paper under double-blind review

Abstract

Dexterous manipulation relies on precise contact-rich interaction, where small errors in slip, alignment, or force can lead to task failure. However, these contact states are often difficult to infer reliably from vision alone. Tactile feedback provides direct information about physical interaction, yet effective tactile control involves both reasoning about desired contact outcomes and rapidly responding to unexpected deviations. Existing visuo-tactile policies typically handle these functions within a monolithic action model, coupling slow task-level reasoning with fast contact-level correction. We introduce TouchWorld, a predictive-and-reactive tactile foundation model that separates these functions across three control levels. A high-level planner selects executable subtasks and uses a tactile world model to predict visual-tactile subgoals. A goal-conditioned visuo-tactile policy generates nominal action chunks, while a tactile residual Transformer corrects execution using recent tactile and proprioceptive feedback. Predicted goals guide contact-aware action generation; residual feedback adapts those actions when actual contact deviates from expectation. The modules are trained in stages, including human tactile pretraining and robot adaptation of the world model. Across six real-robot manipulation tasks, TouchWorld achieves 65.0% success in clean conditions and 53.7% under human perturbations, exceeding the strongest baseline by 15.7 and 18.5 percentage points, respectively.

1 Introduction

Everyday manipulation, from inserting a power plug to pulling a tissue, requires anticipating contact and reacting when it changes. Vision and language provide semantic and geometric guidance but cannot reliably reveal slip, grasp stability, or contact force. Tactile sensing exposes these local physical states (Zhao et al., 2026; Huang et al., 2025b; Zhang et al., 2025; Yuan et al., 2026b; Huang et al., 2025a; Zheng et al., 2026b). Using this feedback effectively requires coordinating slow task reasoning, action-chunk generation, and rapid contact correction.

Recent vision-language-action (VLA) policies have advanced general manipulation (Black et al., 2024; Intelligence et al., 2025b; Bjorck et al., 2025; Zheng et al., 2026a; Liu et al., 2025; Wang et al., 2026; Intelligence et al., 2026; Kim et al., 2024; 2025; Dai et al., 2026; Intelligence et al., 2025a), while action chunking supports temporally coherent motions (Zhao et al., 2023; Chi et al., 2025). However, incorporating touch as additional input tokens in a monolithic policy (Yuan et al., 2026a) couples tactile feedback to nominal action generation. A plausible action chunk may therefore fail when contact changes before the next policy update, motivating faster tactile feedback (Xue et al., 2025; 2026). Contact-rich manipulation also benefits from anticipating the contact outcome toward which these local corrections should steer execution.

These requirements connect three lines of work. Hierarchical policies use intermediate plans to bridge task reasoning and control (Li et al., 2025), while predictive models supply future states for goal-conditioned manipulation (Zhou et al., 2025; Zang et al., 2026). Reactive tactile policies instead emphasize adaptation to immediate contact changes (Niu et al., 2026; Xue et al., 2025). Our focus is their integration: a predicted contact goal specifies the intended outcome of a subtask, and a separate feedback policy adjusts its execution as new tactile measurements arrive.

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.

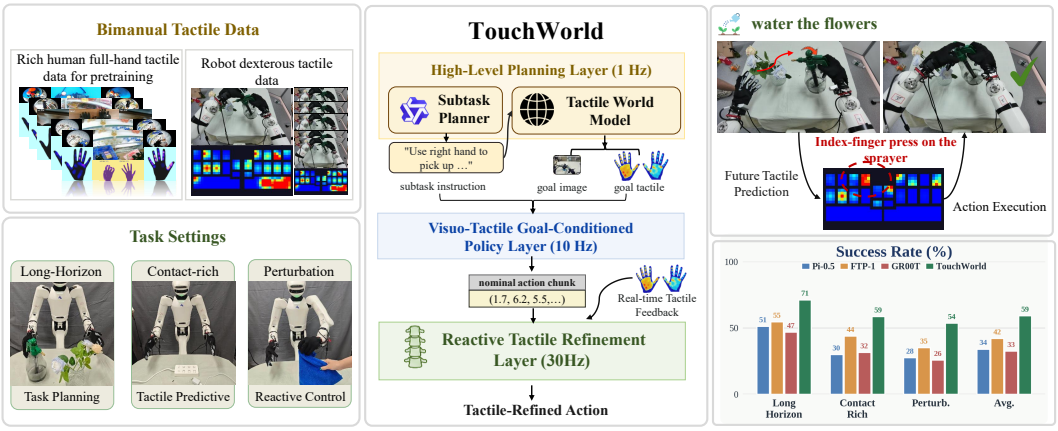

Figure 1: TouchWorld combines predictive tactile goals with reactive feedback. Long Horizon and Contact Rich report clean success over the two task groups in fig. 4; Perturb. averages all six tasks under perturbations, and Avg. averages all six tasks across both conditions. Values are rounded to integers.

We propose TouchWorld, a predictive-and-reactive tactile foundation model that uses touch as both a future contact reference and an online feedback signal (fig. 1). Its High-Level Planning Layer contains a Subtask Planner and a Tactile World Model, which select executable subtasks and predict their visual-tactile subgoals. A Visuo-Tactile Goal-Conditioned Policy generates nominal action chunks from these goals and current observations. A Tactile-Conditioned Refinement Policy then updates residual corrections within each chunk using recent tactile and proprioceptive histories. This hierarchy assigns task progress to the nominal policy and local contact adaptation to a faster feedback path.

On six real-robot tasks spanning long-horizon manipulation and precise contact control, TouchWorld achieves 65.0% average success in clean conditions and 53.7% under human perturbations, exceeding the strongest baseline by 15.7 and 18.5 percentage points.

This work makes three main contributions:

- We introduce a predictive tactile pathway that combines subtask planning with visual-tactile world modeling to guide goal-conditioned manipulation.
- We develop a tactile residual policy that corrects nominal action windows using fresh contact feedback without waiting for the next VLA update.
- We evaluate the complete hierarchy on six real-robot tasks in clean and perturbed settings, with component ablations and analyses of tactile prediction and subtask planning.

2 Method

Given an instruction ℓ, multi-view images $\mathcal{I}_t$, proprioception $\mathbf{s}_t$, tactile observations $\mathcal{X}_t$, and high-level memory m_t, TouchWorld selects a subtask, predicts its contact goal, and generates nominal actions (fig. 2):

$$\begin{aligned} \ell_t^{\text{sub}} &= \pi_{\text{subtask}}(\ell, \mathcal{I}_t, m_t), \\ g_t &= \pi_{\text{world}}(\ell, \ell_t^{\text{sub}}, \mathcal{I}_t, \mathcal{X}_t), \\ (\hat{\mathbf{A}}_{t:t+H-1}, \mathbf{c}_t) &= \pi_{\text{goal}}(\ell, \ell_t^{\text{sub}}, g_t, \mathcal{I}_t, \mathbf{s}_t, \mathcal{X}_t). \end{aligned} \tag{1}$$

Here ℓ_t^{sub} is the executable subtask, g_t its visual-tactile subgoal, and $\hat{\mathbf{A}}_{t:t+H-1}$ an H-step nominal action chunk. The action-generation context $\mathbf{c}_t$ also conditions the faster residual policy in section 2.3. These interfaces follow separate update schedules, with outputs cached between updates.

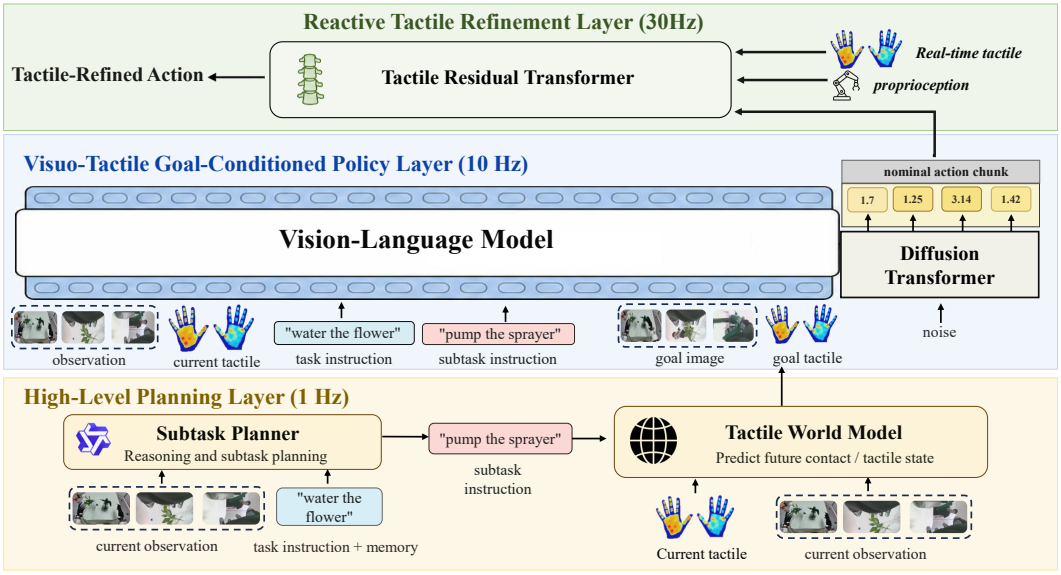

Figure 2: TouchWorld architecture: subtask planning and predicted goals guide the VLA, while TRT refines nominal actions using tactile and proprioceptive feedback. Frequencies denote model calls; 1 Hz applies to the planner, while world-model goals refresh on subtask or task-state changes.

2.1 High-Level Planning Layer

Subtask Planner. The planner selects the current executable subtask from the instruction, current images, and memory m_t, which stores previous subtasks, predicted goals, and execution status. This compact history supports phase tracking without reprocessing past images. The action policy's language input concatenates the original task instruction and current subtask, excluding the planner's reasoning text. This interface separates phase selection from motor execution: the planner identifies what the robot should accomplish next, while the action policy determines how to accomplish it from the current observations. Keeping the global instruction alongside the subtask retains the task context when several phases involve similar objects or motions.

Tactile World Model. The world model generates a visual-tactile sequence conditioned on the source observation; the final frame supplies the predicted subtask-terminal visual and tactile images forming the policy goal g_t. Its observation interface is a 2×2 grid of RGB views and tactile images, conditioned on task and subtask text. Predictions are refreshed when the planner detects a new subtask or meaningful task-state change; stable phases reuse the cached goal. The predicted goal thus describes a desired future observation rather than a low-level action command. The action policy must translate that reference into motions for the current robot state, while the feedback policy handles local deviations during execution. Language identifies the intended operation, whereas the visual-tactile goal specifies an anticipated physical outcome. The goal is a conditioning reference rather than a guarantee that future contact will match the prediction, motivating the separate feedback path.

2.2 Visuo-Tactile Goal-Conditioned Policy

The nominal policy uses a vision-language backbone and a diffusion Transformer action expert. Tactile readings are rendered or normalized into images and encoded alongside RGB observations, retaining compatibility with image-language pretraining without sensor-specific encoders in this branch. The language prompt concatenates `Task:` ℓ and `Current subtask:` ℓ_t^{sub}, preserving both overall intent and the immediate target. During both training and inference, the full policy receives predicted terminal visual and tactile images as goal context, alongside current observations and the task/subtask prompt.

Conditioned on this context and proprioception, the action expert generates nominal chunks through flow matching. Training interpolates demonstrated actions with Gaussian noise; inference integrates

the learned velocity field from noise to an action sequence. Visual, tactile-image, and language inputs form context tokens that the action expert attends to during generation. On the Wuji platform, each action has 120 dimensions: two 48-dimensional arm-hand representations, nine head-action dimensions, and 15 reserved dimensions.

2.3 Tactile-Conditioned Refinement Policy

We implement fast feedback with a Tactile Residual Transformer (TRT) f_ϕ. Let t be the current nominal chunk's starting action step and $\tau \geq t$ the current global action step; $j = \tau - t$ is the chunk-local offset. TRT conditions on a W-step nominal window, proprioceptive and tactile histories, and the context $\mathbf{c}_t$ cached for this chunk:

$$\begin{aligned} \Delta\mathbf{A}_{\tau:\tau+W-1} &= f_\phi(\hat{\mathbf{A}}_{\tau:\tau+W-1}, \mathbf{s}_{\tau-k:\tau}, \mathcal{X}_{\tau-k:\tau}, \mathbf{c}_t), \\ \tilde{\mathbf{A}}_{\tau:\tau+W-1} &= \hat{\mathbf{A}}_{\tau:\tau+W-1} + \Delta\mathbf{A}_{\tau:\tau+W-1}. \end{aligned} \tag{2}$$

Here k is the history lookback, so $[\tau - k, \tau]$ includes $k + 1$ time steps; $\tilde{\mathbf{A}}$ is the corrected action window. Unlike the nominal policy, TRT directly encodes high-frequency tactile histories using lightweight modality-specific encoders: image encoders for tactile maps, convolutions for matrix readings, and Fourier features with MLPs for low-dimensional signals. TRT uses a hidden width of 512, eight Transformer layers, eight attention heads, and a W-step residual query set. The histories capture contact changes, while nominal lookahead and VLA context identify the motion being refined.

The controller executes the first C corrected actions, then refreshes the residual with new feedback. This lets contact changes affect execution before the next nominal chunk arrives. Only the first C corrections are committed; the remainder are recomputed using new feedback. Let $\mathcal{R}$ index the 58 tactile-sensitive dimensions of the 120-dimensional action, covering wrist poses and selected hand joints. Residuals are zero outside $\mathcal{R}$ and are added in the nominal policy's normalized action space. The nominal policy therefore retains responsibility for task progress, while TRT adjusts local contact behavior. Model invocation rates and a reference commit schedule are given in section 3.1; executed feedback also depends on sensing and inference latency.

2.4 Training

We train the modules in four stages, without end-to-end back-propagation through the hierarchy. Additional hyperparameters are provided in Appendix B.

Subtask planning. We fine-tune Qwen3-VL-4B-Instruct (Bai et al., 2025) on 128,866 records from annotated teleoperation traces, predicting executable subtasks from instructions, current images, and recent planner memory. Samples match the deployment semantic update rate and include stale-history and phase-transition variants. We use supervised LoRA adaptation with rank 16, learning rate 10^{-4}, and 20 epochs; rare phases are oversampled to improve their representation in training.

Tactile world modeling. Starting from Wan2.2-TI2V-5B (Wan et al., 2025), we pretrain on EgoTouch human interaction data (Zhou et al., 2026), which pairs egocentric and wrist videos with bimanual hand poses and dense palm pressure maps. Human and robot tactile observations share an image-form goal-grid interface despite their different sensor layouts. We then adapt the model on 10 hours of robot demonstrations using 17-frame visual-tactile sequences conditioned on source observations and subtask labels. At 30 FPS, these demonstrations contain approximately 1.08 million robot frames. This transfers human contact priors to robot-specific subgoal prediction. Each sequence uniformly samples the interval from the source frame to the annotated target-subtask end, including both endpoints, at 384×224 resolution. We adapt the DiT attention and feed-forward projections with rank-64 LoRA, learning rate 10^{-4}, and 50 epochs.

Nominal action learning. We train the goal-conditioned action policy with flow-matching imitation learning on demonstrated action chunks, using the observation and task/subtask context in section 2.2 and predicted terminal visual and tactile goal images. Training uses 30,000 optimization steps, global batch size 32, and AdamW with a cosine learning-rate schedule from a peak of 2.5×10^{-5} to 2.5×10^{-6} after 1,000 warmup steps. We use bfloat16 precision, gradient clipping at 1.0, and action normalization statistics computed from the teleoperation training set.

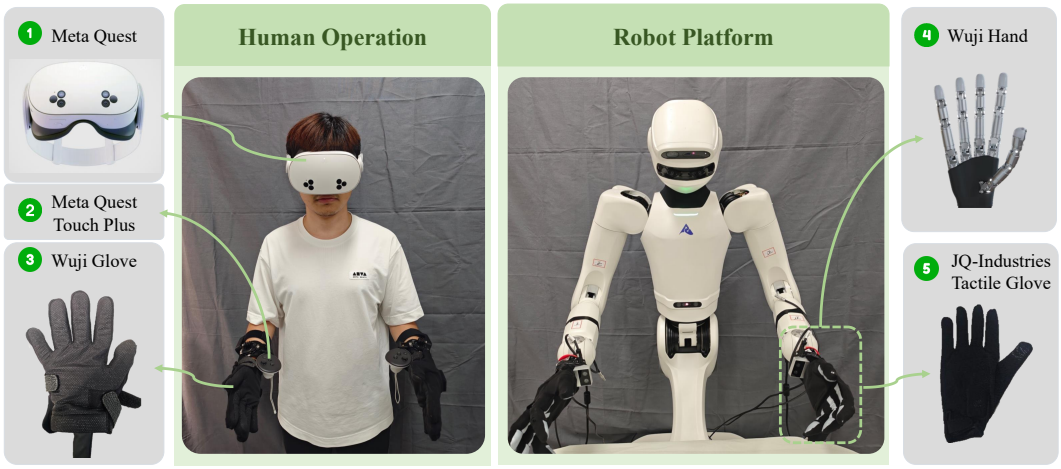

Figure 3: Teleoperation and robot hardware. A Meta Quest headset, Touch Plus controllers, and Wuji Glove support demonstration collection; Wuji hands with a JQ-Industries tactile glove execute the policy.

Residual refinement. We attach TRT to the trained nominal VLA and freeze the latter, optimizing the residual actor and tactile feedback encoder. The frozen policy supplies nominal action windows and context tokens; demonstrated high-frequency actions $\mathbf{A}^*$ define the target residual:

$$\mathcal{L}_{\text{fb}} = \left\| \Delta\mathbf{A}_{\tau:\tau+W-1} - (\mathbf{A}^*_{\tau:\tau+W-1} - \hat{\mathbf{A}}_{\tau:\tau+W-1}) \right\|^2_{\mathcal{R}}, \tag{3}$$

where $\|\cdot\|^2_{\mathcal{R}}$ sums squared errors over the W-step window and action dimensions in $\mathcal{R}$. Demonstrated and nominal actions use the same normalization; dimensions outside $\mathcal{R}$ retain the nominal output. TRT uses AdamW with learning rate and weight decay both set to 10^{-4}. Freezing the nominal policy keeps the residual targets tied to a fixed action generator throughout this stage.

3 Experiments

We evaluate task success and robustness on real robots, then examine component contributions, tactile prediction accuracy, and subtask executability.

3.1 Experimental Setup

Our humanoid platform uses Wuji dexterous hands and tactile gloves, with the same visual, proprioceptive, and tactile interfaces for teleoperation and policy evaluation (fig. 3).

Policy scheduling. The nominal policy predicts $H = 32$ actions; TRT refines sliding windows of $W = 16$ actions and executes $C = 4$ before refreshing. Planner, nominal VLA, and TRT are invoked at 1, 10, and 30 Hz, respectively; these are model invocation rates, not robot servo rates. Appendix B.5 gives a synchronous reference schedule for chunk replacement and its timing assumptions.

Tasks and data. The six tasks (fig. 4) cover long-horizon manipulation (Water Flower, Tabletop Clearing), precision contact (Cup Insertion, Power Plug Insertion), continuous contact regulation (Pot Wiping), and soft-object handling (Tissue Pulling). We evaluate clean execution and human perturbations, including target displacement, unstable contact, and grasp interference. Each task has 200 teleoperated training trajectories and 100 real-robot evaluation rollouts; we report success rates in both settings.

Training trajectories contain global instructions, manually annotated subtask segments, head/wrist images, proprioception, actions, and bilateral tactile pressure, supporting supervision of all four modules.

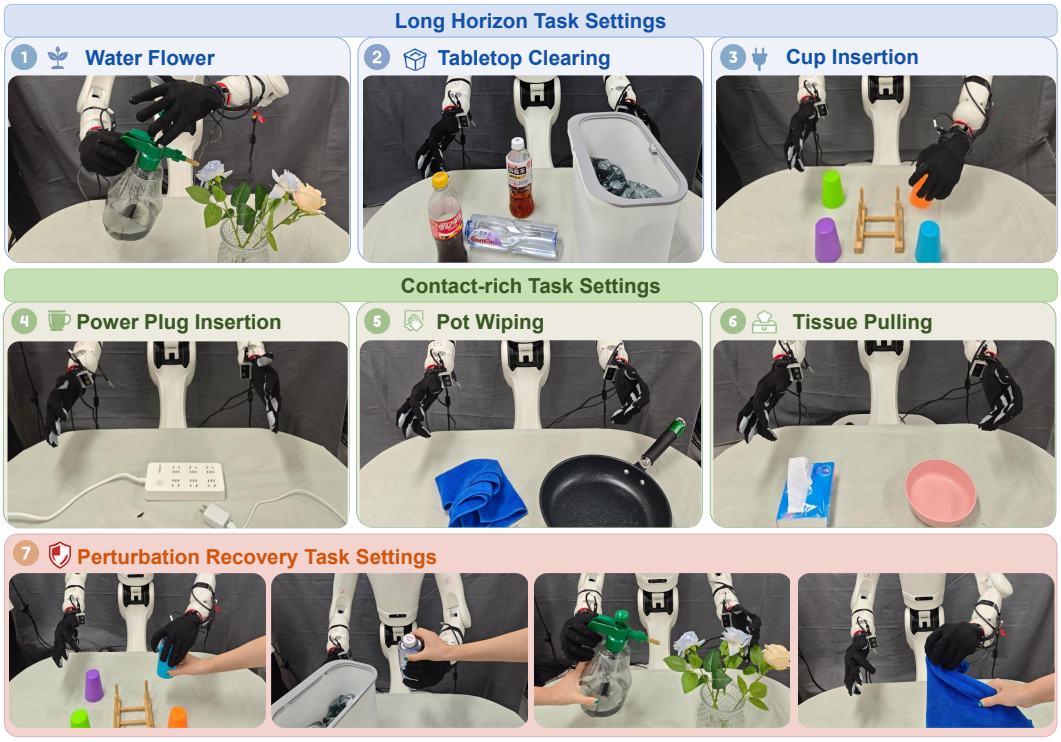

Figure 4: Six-task evaluation suite. The top two rows define the task groups used in fig. 1; task properties can overlap, as cup insertion also requires precise contact. The bottom strip (7) illustrates perturbations, not an additional task.

Table 1: Real-robot task success (%) in clean and human-perturbed settings. Best and second-best results are bold and underlined.

Method	Water Flower	Tabletop Clearing	Cup Insertion	Power Plug Insertion	Pot Wiping	Tissue Pulling	Avg.
Clean Setting							
Pi-0.5	52	66	36	12	39	39	40.7
FTP-1	56	60	48	32	57	43	49.3
GR00T N1.7	50	58	33	18	36	41	39.3
TouchWorld	72	76	66	45	70	61	65.0
Human Perturbation Setting							
Pi-0.5	34	44	24	6	28	30	27.7
FTP-1	39	42	34	20	42	34	35.2
GR00T N1.7	32	36	21	9	26	32	26.0
TouchWorld	60	62	52	35	57	56	53.7

Baselines. We compare with Pi-0.5, a VLA using the same instructions and visual observations; FTP-1, a monolithic tactile policy; and GR00T N1.7, a generalist robot policy evaluated on the same task suite.

3.2 Main Results

TouchWorld leads on all six tasks in both settings (table 1). Average success reaches 65.0% in clean conditions and 53.7% under perturbations, improving over FTP-1, the strongest baseline, by 15.7 and 18.5 percentage points. The clean-to-perturbed drop is 11.3 points for TouchWorld versus 14.2 for FTP-1, showing improved robustness alongside higher clean-task performance. FTP-1 is the strongest baseline on average, although Pi-0.5 performs better on tabletop clearing in both settings. These cross-model comparisons evaluate complete policies, rather than isolating the contribution of tactile inputs.

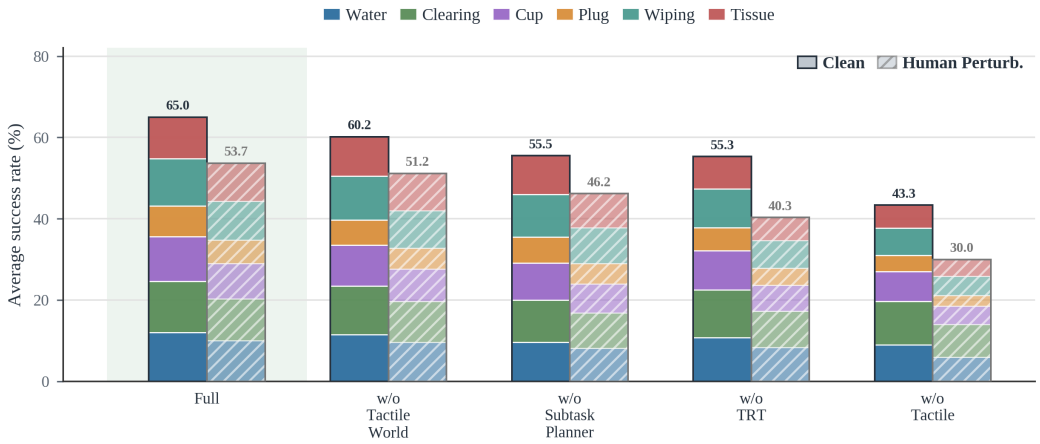

Figure 5: Component ablations. Bar heights show six-task average success, with task-wise contributions stacked. Solid/hatched bars denote clean/perturbed rollouts. TRT denotes the Tactile Residual Transformer; Tactile World denotes the World Model.

The improvements span different contact requirements. Under perturbations, power plug insertion rises from FTP-1's 20% to 35%, pot wiping from 42% to 57%, and tissue pulling from 34% to 56%. TouchWorld also reaches 72% and 76% on clean watering and tabletop clearing, respectively. Thus, the gains extend beyond precision insertion to sustained contact, soft-object handling, and tasks with multiple semantic phases.

3.3 Ablation Studies

We evaluate four ablation variants: without subtask planning, world-model goals, residual refinement, or tactile inputs (fig. 5). Without the planner, the Tactile World Model predicts a visual-tactile task-completion goal once from the original task instruction and initial observations; this goal remains fixed throughout the rollout. The action policy receives task-only text while retaining online visual, tactile, and proprioceptive inputs and TRT feedback. The world-model ablation removes predicted visual and tactile goal images while retaining current observations and the original-task/current-subtask prompt. The tactile ablation removes current tactile images from the VLA and world model, tactile history from TRT, and predicted tactile goals. The world model retains current visual observations and task/subtask text, providing only a predicted subtask-terminal visual goal to the VLA. This joint removal of tactile observations and goals causes the largest loss: success falls from 65.0% to 43.3% in clean trials and from 53.7% to 30.0% under perturbations. Without residual refinement, the corresponding reductions are 9.7 and 13.3 percentage points. The larger loss under disturbances supports the role of feedback in correcting execution when nominal actions no longer match contact conditions.

Removing subtask planning reduces clean and perturbed success by 9.5 and 7.5 points, respectively. Removing predicted visual-tactile goals produces smaller but consistent reductions of 4.8 and 2.5 points. Both predictive conditioning and reactive correction contribute to the complete system; residual feedback has the larger perturbation benefit in this evaluation. The planner ablation jointly changes subtask conditioning, goal granularity, and goal refresh, rather than isolating the effect of subtask text alone. These ablations do not isolate interaction effects or establish that individual gains would add linearly.

3.4 Tactile World Model Prediction Analysis

For held-out trajectories, we generate 17-frame visual-tactile clips from source observations and subtasks, aligning predictions with the recorded subtask segment. Ground-truth terminal goals are tactile pressure maps at the annotated subgoal state. We measure temporal contact accuracy over the clip and terminal contact IoU and volumetric IoU after pressure thresholding. Baselines copy

Table 2: Tactile prediction on held-out subgoal segments. All metrics are percentages; higher is better.

Method	Temporal Contact Acc.	Contact IoU	Volumetric IoU
Current tactile copy	70.4	31.8	24.6
Nearest-neighbor subgoal	77.5	39.2	31.0
Tactile World Model	86.3	52.7	43.8

Table 3: Subtask planner evaluation (%). Memory-augmented SFT improves subtask selection, execution, and phase transitions.

Planner	Subtask Acc.	Execution Success	Transition F1
Zero-shot Qwen3-VL-4B (Bai et al., 2025)	43	34	62
Zero-shot Qwen3-VL-32B (Bai et al., 2025)	69	54	71
SFT Qwen3-VL-4B w/o Memory	73	60	76
Memory-Augmented SFT Qwen3-VL-4B (Ours)	88	65	82

the current tactile observation or retrieve a nearest-neighbor subgoal from training data using task and subtask labels. Persistence tests whether the current contact state is already an adequate future reference. Retrieval provides a task-conditioned alternative without generating a new prediction from the source observation. These comparisons distinguish forecasting contact evolution from retaining the present state or selecting a previously observed goal.

The world model improves temporal accuracy from 77.5% to 86.3% and terminal contact IoU from 39.2% to 52.7% over retrieval (table 2), supporting its use as a predictive contact reference. Volumetric IoU also rises from 31.0% to 43.8%. These prediction metrics complement the policy ablation: they assess agreement with recorded contact outcomes, whereas removing goal conditioning measures its contribution to task execution. Together, the two evaluations examine both the quality and the utility of the predicted tactile goals.

3.5 Vision-Language Subtask Planner Analysis

A subtask can be incorrect for the scene or semantically correct but outside the action policy's learned behaviors. We therefore measure subtask accuracy, downstream execution success, and phase-transition F1, comparing zero-shot planning, supervised fine-tuning (SFT), and memory-augmented SFT.

Adding memory raises the fine-tuned 4B planner's subtask accuracy from 73% to 88% and transition F1 from 76% to 82% (table 3). It also exceeds the zero-shot 32B planner on all metrics, demonstrating the value of task-phase supervision and execution history in this setting. Downstream execution success increases from 34% for the zero-shot 4B model to 60% after SFT and 65% with memory. Reporting this metric alongside subtask accuracy distinguishes selecting the right semantic phase from successfully carrying it out with the learned action policy.

3.6 Qualitative Inference Analysis

Figure 6 illustrates subtask-conditioned prediction for watering and cup insertion: grasping, pressurizing, and spraying in the former; grasping, moving, and placing in the latter. Each phase pairs an executable subtask with predicted visual and tactile goals. The watering example separates acquiring the sprayer from operating its pump and trigger; the insertion example separates transporting the cup from placing it in the rack. These sequences illustrate how the same goal interface supports distinct phases within an instruction.

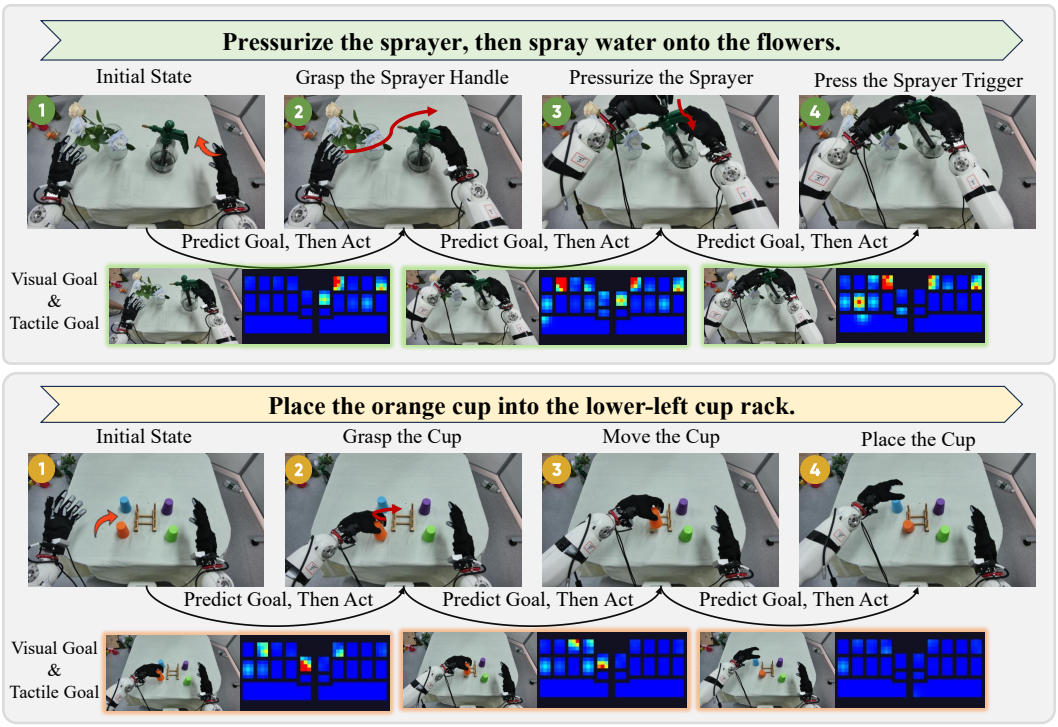

Figure 6: Subtask sequences and predicted visual-tactile goals for watering flowers (top) and cup insertion (bottom).

4 Discussion and Limitations

Visual-tactile goal conditioning and residual feedback both improve task success, with the larger perturbation benefit coming from feedback. Prediction thus complements online correction when contact deviates from expectation.

Our six-task evaluation leaves broader household and deformable-object generalization untested. Further tests should vary objects, workspaces, and task compositions. The full system predicts subtask-scale goals with variable temporal horizons; predictions remain uncertain under occlusion, object motion, and human disturbances. Uncertainty-aware prediction or multiple candidate goals may represent alternative futures, but require closed-loop validation.

Transfer to other tactile sensors or hand morphologies requires calibration, normalization, and likely adaptation data. The common image interface and separate feedback encoders provide modular interfaces, but do not by themselves establish cross-embodiment transfer. Finally, fixed planner and policy invocation rates and action windows coexist with event-triggered goal refreshes; their suitability may vary with contact dynamics. Adaptive scheduling could allocate more feedback updates near contact transitions and fewer during stable phases, trading computation against responsiveness.

5 Conclusion

TouchWorld combines predicted visual-tactile goals with tactile residual feedback for dexterous manipulation. Its planner and world model provide subtask-terminal contact references, the nominal policy converts them into action chunks, and TRT revises tactile-sensitive action dimensions using recent feedback. Across six real-robot tasks, this hierarchy outperforms the evaluated baselines in clean and human-perturbed settings. Component ablations show that predictive goal conditioning and residual refinement both contribute, with refinement providing larger gains under perturbations. Broader validation across tasks, objects, tactile sensors, and robot morphologies remains important.

AI use statement

We used generative AI tools to assist with manuscript preparation, primarily language polishing, including grammar correction, concise rephrasing, and improvements to clarity and organization. The tools also supported internal consistency checks, clarification of mathematical notation and control-scheduling examples, and inspection of relevant code to clarify implementation descriptions. The authors are responsible for verifying all AI-assisted content and take full responsibility for the final manuscript, including its methods, claims, results, and citations.

Ethics statement

This work involves human teleoperation data for collecting robot manipulation demonstrations and uses human interaction data for tactile representation pretraining. All data collection was conducted with informed consent from the participants, and no personally identifiable information is used as model input or released with the data.

All real-robot experiments were conducted in controlled laboratory environments with appropriate safety precautions and human supervision. The proposed system is intended for research on robotic manipulation and has not been evaluated for deployment in safety-critical or unsupervised real-world settings.

More broadly, advances in dexterous robotic manipulation may enable both beneficial and potentially harmful applications. We encourage responsible use of the proposed methods, particularly in settings involving close human–robot interaction. Any released data, models, or code will follow applicable privacy, licensing, and research-ethics requirements.

Reproducibility statement

The manuscript specifies the model interfaces, action and residual objectives, four-stage training procedure, and optimization settings in Sections 2 and 2.4, with additional implementation details in Appendix B. Section 3.1 documents the robot platform, task suite, training and evaluation data, deployment schedule, and baselines; Sections 3.3–3.5 describe the component ablations, tactile prediction metrics, and subtask-planner evaluation protocols. The tables report the task-level results and all evaluated variants, allowing the main comparisons to be reconstructed from the stated settings. We will provide any accompanying implementation materials and evaluation configurations through the submission's supplementary materials when available.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. arXiv preprint arXiv:2503.14734, 2025.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164, 2024.

Kevin Black, Manuel Galliker, and Sergey Levine. Real-time execution of action chunking flow policies. Advances in Neural Information Processing Systems, 38:33383–33407, 2026.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. The International Journal of Robotics Research, 44(10-11):1684–1704, 2025.

Weisheng Dai, Kai Lan, Jianyi Zhou, Bo Zhao, Xiu Su, Junwen Tong, Weili Guan, and Shuo Yang. Conla: Contrastive latent action learning from human videos for robotic manipulation. arXiv preprint arXiv:2602.00557, 2026.

Patrick Grady, Chengcheng Tang, Samarth Brahmbhatt, Christopher D Twigg, Chengde Wan, James Hays, and Charles C Kemp. Pressurevision: estimating hand pressure from a single rgb image. In European Conference on Computer Vision, pp. 328–345. Springer, 2022.

Patrick Grady, Jeremy A Collins, Chengcheng Tang, Christopher D Twigg, Kunal Aneja, James Hays, and Charles C Kemp. Pressurevision++: Estimating fingertip pressure from diverse rgb images. In 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 8683–8693. IEEE, 2024.

Jialei Huang, Shuo Wang, Fanqi Lin, Yihang Hu, Chuan Wen, and Yang Gao. Tactile-vla: unlocking vision-language-action model's physical knowledge for tactile generalization. arXiv preprint arXiv:2507.09160, 2025a.

Jialei Huang, Yang Ye, Yuanqing Gong, Xuezhou Zhu, Yang Gao, and Kaifeng Zhang. Spatially anchored tactile awareness for robust dexterous manipulation. arXiv preprint arXiv:2510.14647, 2025b.

Physical Intelligence, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Kevin Black, Ken Conley, Grace Connors, James Darpinian, Karan Dhabalia, Jared DiCarlo, et al. $\pi^*_{0.6}$: a vla that learns from experience. arXiv preprint arXiv:2511.14759, 2025a.

Physical Intelligence, 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, 2025b. URL https://arxiv.org/abs/2504.16054.

Physical Intelligence, Bo Ai, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Greg Balke, Kevin Black, George Bokinsky, Shihao Cao, Thomas Charbonnier, Vedant Choudhary, Foster Collins, Ken Conley, Grace Connors, James Darpinian, Karan Dhabalia, Maitrayee Dhaka, Jared DiCarlo, Danny Driess, Michael Equi, Adnan Esmail, Yunhao Fang, Chelsea Finn, Catherine Glossop, Thomas Godden, Ivan Goryachev, Lachlan Groom, Haroun Habeeb, Hunter Hancock, Karol Hausman, Gashon Hussein, Victor Hwang, Brian Ichter, Connor Jacobsen, Szymon Jakubczak, Rowan Jen, Tim Jones, Gregg Kammerer, Ben Katz, Liyiming Ke, Mairbek Khadikov, Chandra Kuchi, Marinda Lamb, Devin LeBlanc, Brendon LeCount, Sergey Levine, Xinyu Li, Adrian Li-Bell, Vladislav Lialin, Zhonglin Liang, Wallace Lim, Yao Lu, Enyu Luo, Vishnu Mano, Nandan Marwaha, Aikys Mongush, Liam Murphy, Suraj Nair, Tyler Patterson, Karl Pertsch, Allen Z. Ren, Gavin Schelske, Charvi Sharma, Baifeng Shi, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, Will Stoeckle, Jiaming Tang, Jimmy Tanner, Shalom Tekeste, Marcel Torne, Kyle Vedder, Quan Vuong, Anna Walling, Haohuan Wang, Jason Wang, XuDong Wang, Chris Whalen, Samuel Whitmore, Blake Williams, Charles Xu, Sukwon Yoo, Lili Yu, Wuming Zhang, Zhuoyang Zhang, and Ury Zhilinsky. $\pi_{0.7}$: a steerable generalist robotic foundation model with emergent capabilities, 2026. URL https://arxiv.org/abs/2604.15483.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246, 2024.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. arXiv preprint arXiv:2502.19645, 2025.

Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, et al. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. IEEE Robotics and Automation Letters, 5(3):3838–3845, 2020.

Yi Li, Yuquan Deng, Jesse Zhang, Joel Jang, Marius Memmel, Caelan Garrett, Fabio Ramos, Dieter Fox, Anqi Li, Abhishek Gupta, et al. Hamster: Hierarchical action models for open-world robot manipulation. In International Conference on Learning Representations, volume 2025, pp. 24040–24068, 2025.

Yunzhu Li, Jun-Yan Zhu, Russ Tedrake, and Antonio Torralba. Connecting touch and vision via cross-modal prediction. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10601–10610. IEEE, 2019.

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, volume 2025, pp. 29982–30009, 2025.

Dantong Niu, Zhuoyang Liu, Zekai Wang, Boning Shao, Zhao-Heng Yin, Anirudh Pai, Yuvan Sharma, Stefano Saravalle, Ruijie Zheng, Jing Wang, et al. T-rex: Tactile-reactive dexterous manipulation. arXiv preprint arXiv:2606.17055, 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.

Qiuyue Wang, Mingsheng Li, Jian Guan, Jinhui Ye, Sicheng Xie, Yitao Liu, Junhao Chen, Zhixuan Liang, Jie Zhang, Xintong Hu, et al. Qwen-vla: Unifying vision-language-action modeling across tasks, environments, and robot embodiments. arXiv preprint arXiv:2605.30280, 2026.

Han Xue, Jieji Ren, Wendi Chen, Gu Zhang, Yuan Fang, Guoying Gu, Huazhe Xu, and Cewu Lu. Reactive diffusion policy: Slow-fast visual-tactile policy learning for contact-rich manipulation. arXiv preprint arXiv:2503.02881, 2025.

Teng Xue, Alberto Rigo, Bingjian Huang, Jiayi Shen, Zhengtong Xu, Nick Colonnese, and Amirhossein H Memar. Tube diffusion policy: Reactive visual-tactile policy learning for contact-rich manipulation. arXiv preprint arXiv:2604.23609, 2026.

Chengbo Yuan, Zicheng Zhang, Mingjie Zhou, Wendi Chen, Yi Wang, Zhuoyang Liu, Dantong Niu, Shuo Wang, Hui Zhang, Wenkang Zhang, et al. Ftp-1: A generalist foundation tactile policy across tactile sensors for contact-rich manipulation. arXiv preprint arXiv:2606.13102, 2026a.

Haoran Yuan, Weigang Yi, Zhenyu Zhang, Wendi Chen, Yuchen Mo, Jiashi Yin, Xinzhuo Li, Xiangyu Zeng, Chuan Wen, Cewu Lu, et al. Vtam: Video-tactile-action models for complex physical interaction beyond vlas. arXiv preprint arXiv:2603.23481, 2026b.

Wenzhen Yuan, Siyuan Dong, and Edward H Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. Sensors, 17(12):2762, 2017.

Yujie Zang, Yuhang Zheng, Xian Nie, Yupeng Zheng, Shuai Tian, Songen Gu, Chen Gao, Zining Wang, Shuicheng Yan, and Wenchao Ding. Tacforesight: Force-guided tactile world model for contact-rich manipulation. arXiv preprint arXiv:2606.11184, 2026.

Zongzheng Zhang, Haobo Xu, Zhuo Yang, Chenghao Yue, Zehao Lin, Huan-ang Gao, Ziwei Wang, and Hao Zhao. Ta-vla: Elucidating the design space of torque-aware vision-language-action models. arXiv preprint arXiv:2509.07962, 2025.

Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. arXiv preprint arXiv:2304.13705, 2023.

Zifan Zhao, Siddhant Haldar, Jinda Cui, Lerrel Pinto, and Raunaq Bhirangi. Touch begins where vision ends: Generalizable policies for contact-rich manipulation. IEEE Robotics and Automation Letters, 2026.

Ruijie Zheng, Dantong Niu, Yuqi Xie, Jing Wang, Mengda Xu, Yunfan Jiang, Fernando Castañeda, Fengyuan Hu, You Liang Tan, Letian Fu, et al. Egoscale: Scaling dexterous manipulation with diverse egocentric human data. arXiv preprint arXiv:2602.16710, 2026a.

Yuhang Zheng, Songen Gu, Weize Li, Yupeng Zheng, Yujie Zang, Shuai Tian, Xiang Li, Ce Hao, Chen Gao, Si Liu, et al. Omnivta: Visuo-tactile world modeling for contact-rich robotic manipulation. arXiv preprint arXiv:2603.19201, 2026b.

Jianyi Zhou, Ziteng Gao, Feiyang Hong, Zirui Liu, Guannan Zhang, Weisheng Dai, Ruichen Zhen, Chuqiao Lyu, Haotian Wu, Yinian Mao, et al. Touchanything: A dataset and framework for bimanual tactile estimation from egocentric video. arXiv preprint arXiv:2605.13083, 2026.

Pengfei Zhou, Liliang Chen, Shengcong Chen, Di Chen, Wenzhi Zhao, Rongjun Jin, Guanghui Ren, and Jianlan Luo. Act2goal: From world model to general goal-conditioned policy. arXiv preprint arXiv:2512.23541, 2025.

A Related Work

Vision-language-action robot foundation policies. VLA policies combine large-scale robot data, vision-language pretraining, diffusion or flow-matching action generation, and cross-embodiment training (Intelligence et al., 2025b; Bjorck et al., 2025; Zheng et al., 2026a; Liu et al., 2025; Wang et al., 2026). These models improve semantic grounding and general manipulation, but vision alone does not reliably reveal slip, grasp stability, or contact force. Contact-rich dexterity therefore requires local physical feedback alongside semantic and geometric guidance.

Tactile representation learning and tactile manipulation policies. Sensors such as GelSight (Yuan et al., 2017) and DIGIT (Lambeta et al., 2020) provide dense evidence of local contact. Related work explores visuo-tactile prediction (Li et al., 2019), visual pressure estimation (Grady et al., 2022; 2024), and tactile estimation from egocentric human interaction (Zhou et al., 2026). At the policy level, FTP-1 scales tactile learning across heterogeneous sensors and contact-rich tasks (Yuan et al., 2026a), while T-Rex exploits high-frequency tactile signals for reactive dexterous control (Niu et al., 2026). These approaches motivate using touch both as action-generation context and as a fast feedback signal; TouchWorld additionally predicts visual-tactile goals to specify the contact outcomes toward which execution should progress.

Hierarchical, predictive, and reactive manipulation policies. Action chunking and diffusion-based visuomotor policies generate temporally coherent short-horizon sequences (Zhao et al., 2023; Chi et al., 2025; Black et al., 2026). Hierarchical policies use intermediate subtasks, paths, or visual goals to connect task reasoning with local execution (Li et al., 2025). Predictive policies and world models forecast future states or goals for action generation (Zhou et al., 2025; Zang et al., 2026), whereas reactive tactile policies correct local execution errors using recent contact feedback (Xue et al., 2025; Zhao et al., 2026). TouchWorld integrates these complementary roles: subtask planning and predicted visual-tactile goals guide nominal action chunks, while a separate, faster residual policy adapts them to changing contact.

B Implementation Details

The planner and world model form the high-level layer; the nominal action policy and tactile refinement policy form the other two control levels. We describe their implementation in the four-stage training order of section 2.4, without end-to-end back-propagation through the hierarchy.

B.1 Subtask Planner

We adapt Qwen3-VL-4B-Instruct with supervised LoRA fine-tuning. Its inputs are the original task instruction, current images, and memory of previous subtasks, predicted goals, and execution status, as in section 2.1. The action policy receives the original task instruction concatenated with the current subtask; planner reasoning is excluded and retained only for logging and analysis.

Samples follow the deployment semantic update cadence; the resulting SFT set contains 128,866 records with teacher labels, rare-phase oversampling, and stale-history, phase-transition, and noisy-memory variants. We use LoRA rank 16, alpha 32, dropout 0.05, learning rate 10^{-4} with cosine decay and a 0.1 warmup ratio, bfloat16 precision, and 20 epochs.

B.2 Tactile World Model

Starting from Wan2.2-TI2V-5B, we pretrain on EgoTouch human interaction data and then adapt on 10 hours of robot demonstrations. Human and robot observations use a common 2×2 grid interface containing RGB views and tactile images, conditioned on the original task and current subtask. The action policy receives the visual and tactile images from the final generated frame as its subtask-terminal goal, not the full sequence. Robot training sequences contain 17 frames sampled uniformly from the source observation to the annotated target-subtask end, with rounded indices and both endpoints included, at 384×224 resolution. The first frame provides image conditioning; the temporal span varies with the remaining interval to the subtask end. At 30 FPS, the robot demonstrations contain approximately 1.08 million frames. LoRA adaptation targets the DiT attention and feed-forward projections $\{q, k, v, o, \mathrm{ffn}.0, \mathrm{ffn}.2\}$, with rank 64, learning rate 10^{-4}, and 50 epochs.

B.3 Visuo-Tactile Goal-Conditioned Policy

The nominal policy combines a vision-language backbone with a diffusion Transformer action expert. It encodes multi-view RGB, image-form tactile observations, and task/subtask text as context tokens, without sensor-specific tactile encoders in this branch. Proprioception and predicted terminal visual and tactile goal images condition the full policy during both training and inference, as in section 2.2. On the Wuji platform, the policy predicts $H = 32$ actions of 120 dimensions: two 48-dimensional arm-hand representations, nine head-action dimensions, and 15 reserved dimensions.

Training uses flow-matching imitation learning: demonstrated actions are interpolated with Gaussian noise, and the action expert learns the velocity field integrated at inference. Action targets and normalization statistics come from the teleoperation zarr training set. We use 30,000 optimization steps, global batch size 32, bfloat16 precision, AdamW, gradient clipping at 1.0, and a cosine learning-rate schedule with 1,000 warmup steps, peak learning rate 2.5×10^{-5}, and final learning rate 2.5×10^{-6}.

B.4 Tactile-Conditioned Refinement Policy

TRT conditions on tactile and proprioceptive histories over global steps $[\tau - k, \tau]$, a sliding nominal-action window, and VLA context $\mathbf{c}_t$ cached at the current chunk's start t. Unlike the nominal VLA, it uses modality-specific tactile encoders: image encoders for tactile maps, convolutions for matrix readings, and Fourier features with MLPs for low-dimensional signals. The Transformer has hidden width 512, eight layers, eight attention heads, and $W = 16$ residual queries matching the lookahead window. Residuals affect only the 58 action dimensions indexed by $\mathcal{R}$, covering wrist poses and selected hand joints; they are scattered into the 120-dimensional action vector and are zero elsewhere.

We freeze the trained nominal VLA and optimize only the residual actor and tactile feedback encoder with the subspace squared-error objective in section 2.4. The target residual is the normalized demonstrated action window minus the normalized nominal prediction; squared errors are summed over the W steps and dimensions in $\mathcal{R}$. Residual addition uses the same normalized action space. TRT uses AdamW with learning rate and weight decay both 10^{-4}.

B.5 Deployment Schedule

The planner, nominal VLA, and TRT are invoked at 1, 10, and 30 Hz, respectively. These are model invocation rates, not robot servo rates. In the full system, the world model refreshes its prediction when the planner detects a new subtask or meaningful task-state change; stable phases reuse the cached goal. The nominal policy predicts $H = 32$ actions; TRT refines $W = 16$-step windows and commits $C = 4$ actions per update. The 10/30 Hz rates admit a synchronous reference schedule with three TRT updates per nominal update, executing 12 of the 32 actions. At chunk-local offsets $j = \tau - t \in \{0, 4, 8\}$, windows $[0, 15]$, $[4, 19]$, and $[8, 23]$ commit corrected actions $[0, 3]$, $[4, 7]$, and $[8, 11]$, respectively. Each nominal update replaces the unexecuted suffix and context, discards old residuals, and resets t to the new global start ($j = 0$). TRT recomputes corrections with fresh feedback; observation histories retain global indices. All windows stay within $H = 32$, requiring no cross-chunk stitching or padding. Uniform action steps and predictions ready at update boundaries imply 120 action steps/s, approximately 33.3 ms per commit, and 100 ms per nominal update. These are reference timings, not measured rollout or servo rates; asynchronous latency handling is separate.

In the planner ablation, the world model predicts a fixed visual-tactile task-completion goal once from the original task and initial observations. The VLA receives task-only text; online visual, tactile, and proprioceptive observations and TRT feedback are retained. World-model, VLA, and TRT checkpoints are reused without retraining or fine-tuning. The world-model ablation removes both predicted goal modalities, retaining current observations and task/subtask text. The tactile ablation removes the VLA's current tactile images, TRT's tactile history, and the world model's current tactile observations. The world model retains the original-task/current-subtask prompt and current visual observations to predict future visual frames; only the predicted subtask-terminal visual image conditions the VLA as a goal. No predicted tactile goal is supplied.