TAD: TEACHER-ASSISTANT DISTILLATION OF TACTILE PREDICTIONS FOR ROBOT CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Tactile sensing provides complementary information for contact-rich manipulation, and predictions of tactile change offer an additional source of supervision for robot policies. However, forecasting and control impose different requirements on a representation, making it challenging to preserve predictive knowledge while adapting it to action learning. We propose Teacher-Assistant Distillation (TAD), which transfers tactile predictions into a vision–language–action policy through a frozen teacher and an adaptive assistant. The teacher predicts features of future tactile changes from current observations and initially supervises intermediate predictive tokens in the policy. During training, a teacher-initialized assistant adapts through action supervision while remaining anchored to the teacher. An exponential moving average of the assistant's prediction parameters provides distillation targets, while the online assistant's action readout supplies residual corrections to the denoising field. Asymmetric attention isolates predictive tokens from noisy-action inputs while allowing action tokens to attend to them. Instantiated on $\pi_{0.5}$, TAD achieves mean success rates of 90.0% across eight UniVTAC tasks and 87.5% across six real-world tasks, exceeding N_0-VTLA by 6.9 and 10.0 percentage points, respectively. Ablations show lower success rates without the assistant, teacher-derived supervision, or future tactile targets. These findings support teacher-anchored adaptation as a mechanism for transferring tactile predictive knowledge into robot control.

1 INTRODUCTION

Vision–language–action (VLA) policies map visual observations and language instructions to robot actions using pretrained representations (Kim et al., 2024; Black et al., 2025). In contact-rich manipulation, however, visible geometry alone may not resolve the interaction state: an object can slip within a grasp, or a connector can encounter resistance despite appearing aligned. Tactile observations provide complementary information about these interactions (Higuera et al., 2024; George et al., 2024). Predicting their temporal evolution may further distinguish contacts that appear similar but lead to different outcomes. The challenge is to translate this predictive information into features useful for action generation.

Recent tactile policies incorporate future sensory prediction into learned representations or action conditioning (Liu et al., 2025; NeoteAI Team & Fudan TEAI Team, 2026b;a). Although tactile features can also adapt through action learning, accurate sensory prediction does not necessarily identify the features most useful for control. Distillation can transfer predictive knowledge, as illustrated by privileged foresight in robot learning (Fang et al., 2026), while teacher-assistant distillation introduces an intermediate model to bridge capacity differences (Mirzadeh et al., 2020). Here, we address the relationship between predictive supervision and an evolving control objective. Frozen teacher targets remain independent of action learning; unconstrained predictor adaptation may weaken the learned predictive structure. We therefore ask: *how can tactile predictive knowledge be preserved while its transfer to a robot policy adapts to the action objective?*

We propose Teacher-Assistant Distillation (TAD), which combines a fixed predictive reference with action-driven adaptation (Figure 1). A frozen pretrained teacher initially supervises predictive policy features. A teacher-initialized assistant then adapts through action supervision while remaining anchored to the teacher. Its exponential moving average supplies evolving distillation targets, and

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.

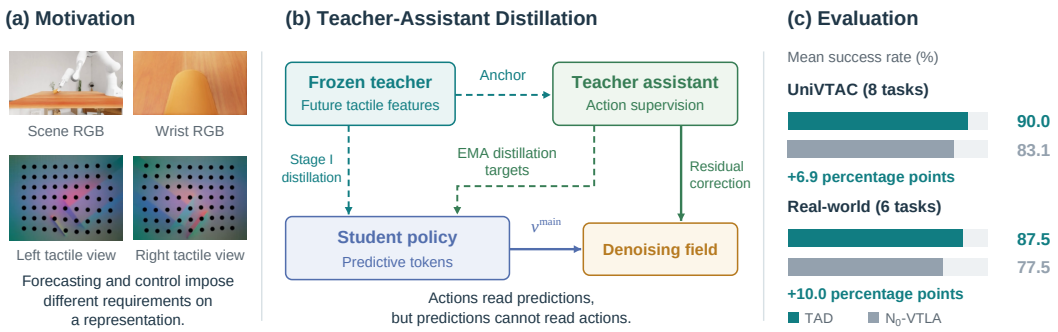

Figure 1: **Teacher-Assistant Distillation (TAD).** (a) Current visual and tactile observations. (b) A teacher-anchored assistant adapts through action supervision; its EMA supplies distillation targets, and its online features correct the denoising field. Asymmetric attention isolates predictive tokens from noisy action inputs. (c) Mean success rates on eight UniVTAC and six real-world tasks.

its online features provide residual corrections to action generation. Asymmetric attention permits actions to access predictive features while excluding noisy action inputs from those features. At deployment, the policy and assistant use only current observations and fixed parameters.

We instantiate TAD on a tactile-conditioned $\pi_{0.5}$ policy. Trained with 800 demonstrations across eight UniVTAC tasks (Chen et al., 2026), TAD achieves 90.0% mean success. Across six real-world tasks, it achieves 87.5% mean success. These results exceed N_0-VTLA by 6.9 and 10.0 percentage points, respectively. Physical ablations assess the contributions of the assistant, teacher-derived supervision, tactile observations, and future tactile targets. Environment variations and offline diagnostics further characterize robustness and the learned mechanism.

Our contributions are threefold: **(1)** a framework for transferring observation-conditioned tactile predictions into intermediate robot-policy features; **(2)** a teacher-anchored assistant that links action supervision to adaptive distillation targets and residual action corrections; and **(3)** an asymmetric attention constraint that isolates predictive features from noisy action inputs while preserving their availability to action generation.

2 Related Work

Tactile representations and policy learning. Tactile representation learning provides a foundation for contact-rich control through self-supervised and visuotactile pretraining (Higuera et al., 2024; George et al., 2024), multimodal alignment (Yang et al., 2024), and cross-sensor transfer (Zhao et al., 2024; Feng et al., 2025). At the policy level, 3D-ViTac fuses vision and touch in a shared 3D space for diffusion-based imitation learning (Huang et al., 2024), while Reactive Diffusion Policy combines slow action-chunk generation with fast tactile feedback (Xue et al., 2025). Within language-conditioned control, VTLA integrates visuotactile inputs with preference learning (Zhang et al., 2025); VLA-Touch uses tactile feedback to refine actions from a frozen VLA (Bi et al., 2025); and TaF-VLA incorporates force-aligned tactile representations into a VLA backbone (Huang et al., 2026). TAD complements current tactile conditioning with predictive supervision that transfers anticipated tactile changes into policy features.

Predictive representations for robot control. Future visual prediction has supported robot control through video-based planning, generative pretraining, and predictive feature conditioning (Finn & Levine, 2016; Du et al., 2023; Wu et al., 2023; Cheang et al., 2024; Hu et al., 2024). FLARE aligns intermediate policy features with embeddings of actual future observations, using an action-trained encoder whose EMA supplies the targets (Zheng et al., 2025). In multisensory manipulation, MLA combines sensory alignment with future prediction (Liu et al., 2025), while DreamTacVLA integrates spatial alignment with a tactile world model (Ye et al., 2025). N_0-VTLA conditions a flow-matching action expert on predicted tactile changes and jointly trains the policy (NeoteAI Team & Fudan TEAI Team, 2026b); N_0-TWAM combines predicted tactile trajectories with observed-touch features (NeoteAI Team & Fudan TEAI Team, 2026a). TAD supervises policy features with

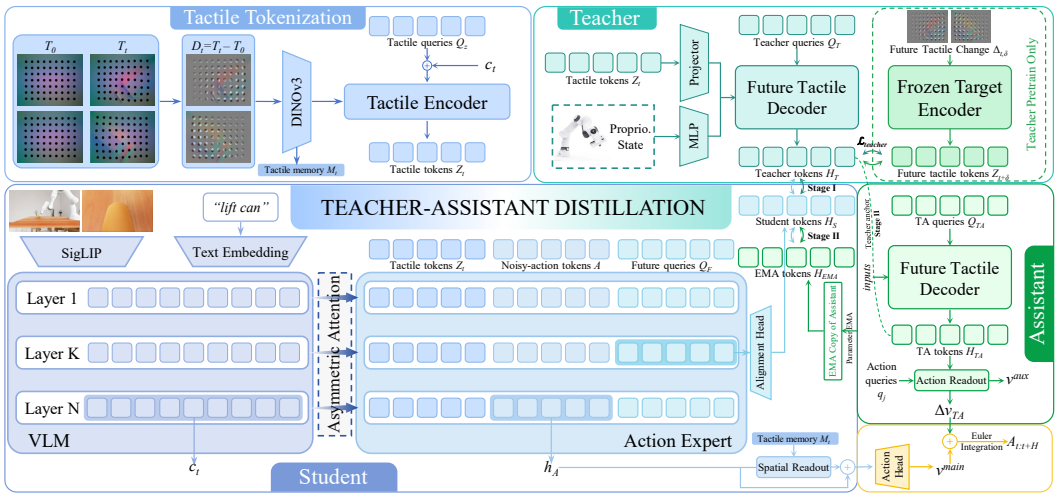

Figure 2: **Overview of Teacher-Assistant Distillation.** Current tactile differences are encoded into tokens that condition the action expert using visual–language context. A teacher predicts features of future tactile changes to supervise the student's predictive tokens through an Alignment Head. The teacher assistant adapts predictive supervision and contributes action corrections, while a separate Spatial Readout provides access to current tactile memory. The two-stage objectives and assistant EMA are specified in the text.

tactile forecasts generated from current observations by a teacher-anchored assistant. Its EMA tracks the assistant's prediction parameters, allowing action supervision to adapt the distillation targets while retaining a fixed predictive reference. An asymmetric attention constraint additionally isolates predictive features from noisy action inputs.

Knowledge distillation and teacher assistants. Knowledge distillation transfers teacher behavior through soft targets (Hinton et al., 2015), while generalized distillation incorporates privileged training information (Lopez-Paz et al., 2016). Extensions transfer intermediate features, attention maps, and representational structure through feature matching, relational objectives, or contrastive learning (Romero et al., 2014; Zagoruyko & Komodakis, 2016; Park et al., 2019; Tian et al., 2019). Collaborative and online distillation further allow supervision to evolve during training (Zhang et al., 2017; Lan et al., 2018). More specifically, teacher-assistant distillation bridges teacher–student capacity gaps through an intermediate model (Mirzadeh et al., 2020); REPA aligns denoising features with pretrained representations (Yu et al., 2025); and Mean Teacher derives consistency targets from averaged student weights (Tarvainen & Valpola, 2017). In robotics, privileged foresight distillation transfers action-denoising corrections from a teacher with access to future observations (Fang et al., 2026). In TAD, the assistant serves as a control-adaptive intermediary anchored to a frozen predictive teacher. It learns from action supervision, supplies distillation targets through averaged prediction parameters, and remains active at deployment to provide residual action corrections.

3 Method

Teacher-Assistant Distillation (TAD) transfers predictions of future tactile change into a robot policy. A teacher first learns to predict tactile representations from current observations and supervises predictive tokens in the student's action expert. A teacher assistant then adapts these representations through action supervision while remaining anchored to the frozen teacher. An exponential moving average (EMA) of the assistant's prediction parameters supplies the student's distillation targets, while the assistant's action readout contributes to the policy's denoising field (Figure 2).

3.1 CURRENT TACTILE CONDITIONING

At time t, the observation $o_t = (I_t, \ell, s_t, T_t, T_0)$ contains scene and wrist images, an instruction, proprioceptive state, two tactile views, and their episode-initial references. The policy generates an action chunk $A_{t:t+H}$, where H denotes the action horizon. We augment a visual–language backbone and flow-matching action expert with a tactile tokenization module. Visual features, language, and discretized state form a prefix P that the action expert attends to.

Current tactile inputs and future targets use distinct signed differences:

$$D_t^v = \mathcal{R}(T_t^v) - \mathcal{R}(T_0^v), \qquad \Delta_{t,\delta}^v = \mathcal{R}(T_{t+\delta}^v) - \mathcal{R}(T_t^v), \tag{1}$$

where v indexes sensors, $\mathcal{R}$ denotes image preprocessing, and δ is the tactile prediction offset, distinct from H. A frozen, tactile-adapted DINOv3 backbone E (Siméoni et al., 2025) extracts global and spatial features from each view. A trainable Tactile Encoder $\mathcal{P}$ uses five learned tactile queries Q_z to compress the projected memory into tactile latents:

$$M_t = \mathrm{concat}_v\, E(D_t^v), \qquad z_t = \mathcal{P}(Q_z, W_g M_t; c_t), \qquad Z_t = W_z z_t. \tag{2}$$

Here W_g projects the tactile memory into the Tactile Encoder's feature space, and W_z maps its output to the action expert's width. The context c_t is the masked mean of valid visual–language prefix states and conditions only the Tactile Encoder's attention queries; keys and values come from tactile memory. On this conditioning path, gradients through c_t are stopped. The resulting tokens Z_t condition the action expert, whose attention to the prefix provides a separate gradient path for visual–language policy adaptation. Appendix A specifies the conditioning operation and dimensions.

3.2 LEARNING THE TACTILE PREDICTION TEACHER

The teacher predicts features of future tactile changes from current observations. Its own Tactile Encoder and associated projections summarize current touch using the original frozen visual–language context. A Projector maps the resulting tactile latents into the decoder space, and an MLP embeds the continuous proprioceptive state as an additional memory token. Learnable teacher queries Q_T attend to this memory through the Future Tactile Decoder to produce teacher tokens:

$$H_T^\psi = \mathcal{T}_\psi(o_t; \kappa), \qquad Z_{t+\delta}^{\mathrm{future}} = \mathrm{concat}_v\, E_{\mathrm{target}}(\Delta_{t,\delta}^v). \tag{3}$$

Here κ denotes the frozen visual–language and DINOv3 backbones, while ψ comprises the trainable Tactile Encoder, Future Tactile Decoder, and their associated projections, state embedding, and learned queries. The target encoder E_{target} is a separately stored frozen copy of the tactile-adapted DINOv3 backbone, including the same feature pooling. It encodes observed future tactile changes without passing through the five-query Tactile Encoder. We use $\delta = 10$ frames, approximately 0.33 seconds. The target contains twenty features: one global feature and nine pooled spatial features per sensor, in a fixed sensor and spatial order. The twenty teacher outputs are aligned with these slots at the same future offset.

Let $\mathcal{C}(U, V; m)$ be cosine distance averaged over valid tokens and samples. Teacher pretraining minimizes

$$\mathcal{L}_{\mathrm{teacher}} = \mathcal{C}(H_T^\psi, \mathrm{sg}(Z_{t+\delta}^{\mathrm{future}}); m^{\mathrm{future}}), \tag{4}$$

where sg denotes stop-gradient and m^{future} marks valid sensor features with available future targets from the same episode. Future observations are used only to construct these pretraining targets. The teacher predicts tactile changes under the behavior represented in the demonstrations. After pretraining, we freeze the teacher as $\mathcal{T}_0$, with output $H_T = \mathcal{T}_0(o_t; \kappa)$. Student distillation uses this observation-conditioned forecast and does not require the future-target encoder.

3.3 DISTILLATION WITHIN THE ACTION EXPERT

The student introduces twenty learnable future queries Q_F, which initialize the predictive-token group F. These queries are distinct from the tactile queries Q_z and teacher queries Q_T. The action expert processes tactile, noisy-action, and predictive tokens $[Z_t; A; F]$, conditioned on the visual–language prefix. At an intermediate layer K, a linear Alignment Head projects the predictive hidden states into the teacher feature space:

$$H_S = W_F h_F^{(K)}, \qquad K = 10. \tag{5}$$

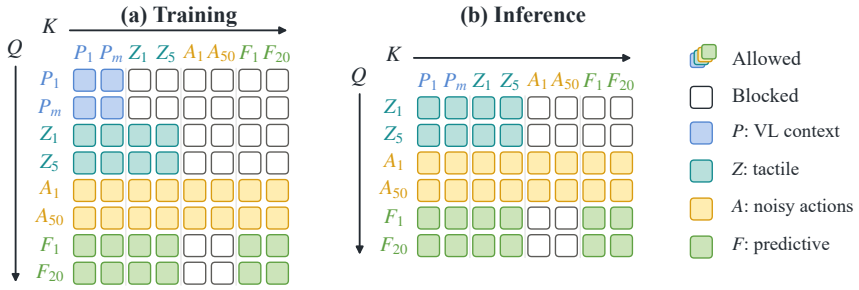

Figure 3: **Asymmetric attention in TAD.** Filled cells permit attention; empty cells block it. Representative valid tokens are shown. Action tokens can attend to predictive tokens, while predictive tokens cannot access noisy-action inputs directly or through other token groups. Inference caches prefix keys and values without changing visibility; tactile, action, and predictive states are recomputed at each denoising step.

Stage I aligns the student tokens H_S with the frozen teacher tokens H_T. The predictive hidden states continue through the remaining expert layers and remain available to action tokens. The twenty aligned slots correspond to the teacher's ordered global and spatial features at a single future offset. The teacher target is shared across flow times, while student predictive states may vary with flow-time conditioning.

Flow matching uses $x_\tau = (1-\tau)a + \tau\varepsilon$, where a is the normalized demonstration action chunk, $\varepsilon \sim \mathcal{N}(0, I)$, and the target velocity is $u = \varepsilon - a$ (Lipman et al., 2023). The noisy-action tokens A encode x_τ. Because x_τ contains demonstration-action information for $\tau < 1$, we exclude all direct and indirect paths from A to F at every layer. Figure 3 shows the attention pattern, with queries as rows and keys as columns. For fixed observations, model parameters, flow time, and any stochastic choices, predictive features are invariant to x_τ under this mask (Appendix D). This forward-input constraint still allows action supervision to update predictive-token parameters and shared transformations. Predictive-token visibility depends on current sensor availability rather than future-target availability, preserving the same information structure during training and inference.

Stage II also adds a Spatial Readout of current tactile memory:

$$\Delta h_{\text{loc}} = \text{SpatialReadout}(h_A, M_t), \qquad v^{\text{main}} = W_a(h_A + \Delta h_{\text{loc}}). \tag{6}$$

Here h_A denotes the final normalized action states. These states query tactile memory augmented with sensor and position embeddings, providing direct access to spatial tactile features beyond the compressed tokens. The readout's output projection is initialized to zero, so introducing this pathway initially preserves the student's action field. The Action Head W_a projects the updated action states to the main denoising field v^{main}.

3.4 Teacher-assistant adaptation

The assistant independently copies the teacher's prediction parameters and produces $H_{\text{TA}} = \mathcal{T}_\psi(o_t; \kappa)$ using the same frozen original context. It is trained jointly with the student on demonstrations. The term *online assistant* distinguishes this trainable branch from its EMA copy; it does not imply adaptation during deployment. The shared original context preserves a common input representation for the teacher, assistant, and EMA target while the student's visual–language parameters adapt. The Action Readout uses action queries to read the assistant's predictive memory:

$$\begin{aligned} q_j &= W_x x_{\tau,j} + W_s s_t + e(\tau) + p_j, \\ r &= \phi\big(q + \text{CrossAttn}(q, H_{\text{TA}})\big), \\ v^{\text{aux}} &= W_u r + b_u, \quad \Delta v_{\text{TA}} = W_r r, \quad v = v^{\text{main}} + \Delta v_{\text{TA}}. \end{aligned} \tag{7}$$

Here p_j identifies the action slot and ϕ is a feed-forward readout. Noisy actions and flow time enter the action queries, not the predictive memory. The residual projection W_r starts at zero, preserving the initial policy. A separate auxiliary head W_u provides action supervision from the outset; only the residual contributes to the deployed field.

An anchor to H_T constrains the assistant as it adapts to action supervision. Its EMA provides a gradually changing student target, $H_{\mathrm{EMA}} = \mathcal{T}_{\bar{\psi}}(o_t; \kappa)$:

$$\bar{\psi}_{k+1} = \beta\bar{\psi}_k + (1-\beta)\psi_{k+1}, \qquad \beta = 0.999, \quad \bar{\psi}_0 = \psi_0. \tag{8}$$

The average follows each optimizer update and includes only prediction parameters, excluding action readouts, the student, and frozen context. Student distillation does not update the assistant directly. The anchor constrains feature deviation from the teacher, whereas EMA averages assistant parameters. The real-world ablation study evaluates the assistant as a complete component.

3.5 Training and deployment

Writing $\mathcal{L}_{\mathrm{FM}}$ for squared flow-prediction error, Stage I minimizes

$$\mathcal{L}_{\mathrm{I}} = \mathcal{L}_{\mathrm{FM}}(v_S, u) + \lambda_{\mathrm{KD}}\mathcal{C}(H_S, \mathrm{sg}(H_T); m), \tag{9}$$

where v_S excludes the spatial readout and assistant. Stage II initializes from this student and minimizes

$$\begin{aligned}\mathcal{L}_{\mathrm{II}} = \mathcal{L}_{\mathrm{FM}}(v, u) + \lambda_{\mathrm{aux}}\mathcal{L}_{\mathrm{aux}}(v^{\mathrm{aux}}, u) \\ + \lambda_{\mathrm{KD}}\mathcal{C}(H_S, \mathrm{sg}(H_{\mathrm{EMA}}); m) + \lambda_{\mathrm{anchor}}\mathcal{C}(H_{\mathrm{TA}}, \mathrm{sg}(H_T); m).\end{aligned} \tag{10}$$

The auxiliary objective is also flow matching. Examples without either tactile view contribute zero auxiliary loss, and both additional action readouts vanish for these examples. We use $\lambda_{\mathrm{aux}} = 0.1$ and $\lambda_{\mathrm{KD}} = \lambda_{\mathrm{anchor}} = 0.01$. Current-view availability determines m; neither student stage requires future images. EMA supervision replaces direct teacher-to-student alignment, while the original teacher anchors the assistant. Appendix C gives the loss reductions and optimization settings.

At deployment, ten Euler steps integrate the combined field from Gaussian noise to an action chunk. Observation-dependent memories are shared across these steps, while student predictive states and action readouts are recomputed at each flow time. The trained assistant and its original frozen visual–language and DINOv3 context remain active; EMA targets, anchoring predictions, and the future-target encoder are unnecessary. All parameters remain fixed during deployment. The Spatial Readout modifies hidden action states before projection; the assistant's Action Readout corrects the velocity field at each denoising step. Both residual output projections are initialized to zero when introduced and are subsequently learned through the joint objective.

4 Experiments

We evaluate TAD on eight UniVTAC tasks and six physical manipulation tasks, followed by component ablations, environmental generalization tests, and analyses of teacher prediction, assistant correction, and tactile attention.

4.1 Experimental setup

Tasks and observations. UniVTAC (Chen et al., 2026) comprises eight tasks spanning lifting, extraction, placement, insertion, and grasp classification. We use 100 demonstrations per task: 800 episodes and 152,080 frames at 30 Hz. Inputs include scene and wrist RGB, bilateral tactile images, robot state, and language. Current tactile images are referenced to episode-initial observations; teacher targets encode change over the next ten frames.

Baselines. Simulation baselines comprise $\pi_{0.5}$ (Black et al., 2025), InternVLA-A1 (Cai et al., 2026), ACT + UniVTAC (Zhao et al., 2023; Chen et al., 2026), VITaL (George et al., 2024), and N_0-VTLA (NeoteAI Team & Fudan TEAI Team, 2026b), covering visual–language control, visual prediction, and direct or predictive tactile conditioning. Simulation baseline scores are taken from NeoteAI Team & Fudan TEAI Team (2026b) (Appendix E.3). Physical evaluations compare $\pi_{0.5}$, N_0-VTLA, and TAD on six tasks.

Training and metrics. Representation adaptation, policy initialization, teacher learning, and Stage I each use 20,000 updates (Appendix C). Success uses task-specific completion criteria and equal task weighting. Prediction error is cosine distance, averaged over valid tokens, then frames within tasks, and finally tasks. The appendix also specifies the earlier HDMI adaptation schedule used in the diagnostic analyses.

Table 1: **Success rates (%) on eight UniVTAC tasks for TAD and five baseline methods.** Avg denotes the mean across tasks. Best results are bold, including ties.

Method	Lift Bottle	Pull-out Key	Lift Can	Put Bottle	Insert Hole	Insert HDMI	Insert Tube	Grasp Classify	Avg
ACT + UniVTAC	71	46	29	31	24	28	56	99	48.0
VITaL	72	47	8	32	25	6	34	**100**	40.5
$\pi_{0.5}$	**100**	35	6	34	25	8	74	49	41.4
InternVLA-A1	58	66	90	37	93	12	95	86	67.1
N_0-VTLA	99	**99**	88	60	95	25	**99**	**100**	83.1
TAD (Ours)	**100**	**99**	**95**	**90**	**98**	**44**	95	99	**90.0**

Table 2: **Success rates on six real-world manipulation tasks (%).** Avg denotes the unweighted mean across the six tasks. Best results are bold, including ties.

Method	Insert Plug	Unscrew Cap	Lift Tube	Upright Bottle	Wipe Whiteboard	Pour Water	Avg
$\pi_{0.5}$	30	40	65	45	60	45	47.5
N_0-VTLA	55	75	90	75	90	80	77.5
TAD (Ours)	**70**	**85**	**95**	**90**	**95**	**90**	**87.5**

4.2 Simulation benchmark results

TAD achieves 90.0% task-averaged success, exceeding N_0-VTLA and $\pi_{0.5}$ by 6.9 and 48.6 percentage points, respectively (Table 1). It ranks first on six of eight tasks, including ties. The largest gains over N_0-VTLA occur on Put Bottle (+30 points) and Insert HDMI (+19 points); Insert Tube and Grasp Classify remain below this baseline.

4.3 Real-World Experiments

Across six physical tasks, TAD attains the highest success on every task and averages 87.5%, versus 77.5% for N_0-VTLA and 47.5% for $\pi_{0.5}$ (Table 2). The largest gains over N_0-VTLA are 15 percentage points on plug insertion and bottle uprighting. Plug insertion remains the lowest-success task for all three methods. Appendix F.2 provides autonomous executions with synchronized scene, wrist, and bilateral tactile observations (Figure 10).

4.4 Real-world ablations

We ablate four components under matched demonstrations, policy-update budgets, and twenty predefined configurations per task, retraining all affected stages (600 rollouts across five variants). Without the assistant, the policy retains direct frozen-teacher distillation; without the teacher, the assistant learns from actions without teacher initialization or anchoring. The tactile-free variant excludes touch during training and deployment. Current targets replaces future supervision with $E(D_t)$ while retaining the remaining mechanism (Appendix E.5). Full TAD performs best on every task (Table 3). Removing the assistant or teacher lowers mean success by 10.0 or 19.2 percentage points, respectively. Removing touch produces the largest drop, from 87.5% to 47.5%; using current targets lowers success to 63.3%. These comparisons support the assistant and teacher mechanisms as a whole and the use of future tactile targets; they do not isolate individual losses or readouts.

4.5 Generalization under environment variations

Following Dream-Tac (Lou et al., 2026), we compare fixed checkpoints of $\pi_{0.5}$, N_0-VTLA, and TAD without test-time adaptation, using unchanged success criteria and twenty trials per condition and method (Figure 4). **(a) Height.** Raising or lowering the bottle support by 5 cm in *Unscrew Cap* reduces TAD success from 85% to 80%, with the bottle and cap unchanged. **(b) Position.** Changing the *Wipe Whiteboard* region from *Standard* to *Random* reduces success from 95% to 90%. **(c) Object.** Success remains 90% for both *Seen* and *Unseen Object* geometries in *Adjust Bottle*. **(d) Background.** Success remains 85% under a *Changed* background in *Unscrew Cap*, holding the bottle, cap tightness,

Table 3: **Real-world component ablations.** Success rates (%) over twenty trials per task and variant. Current targets replaces future tactile targets with current-feature supervision. Avg is the unweighted mean across the six tasks. The highest success rate in each column is shown in bold.

Variant	Insert Plug	Unscrew Cap	Lift Tube	Adjust Bottle	Wipe Whiteboard	Pour Water	Avg
w/o assistant	60	75	85	80	85	80	77.5
w/o teacher	50	65	80	70	75	70	68.3
w/o tactile information	30	40	65	45	60	45	47.5
Current targets	45	60	75	65	70	65	63.3
Full TAD	**70**	**85**	**95**	**90**	**95**	**90**	**87.5**

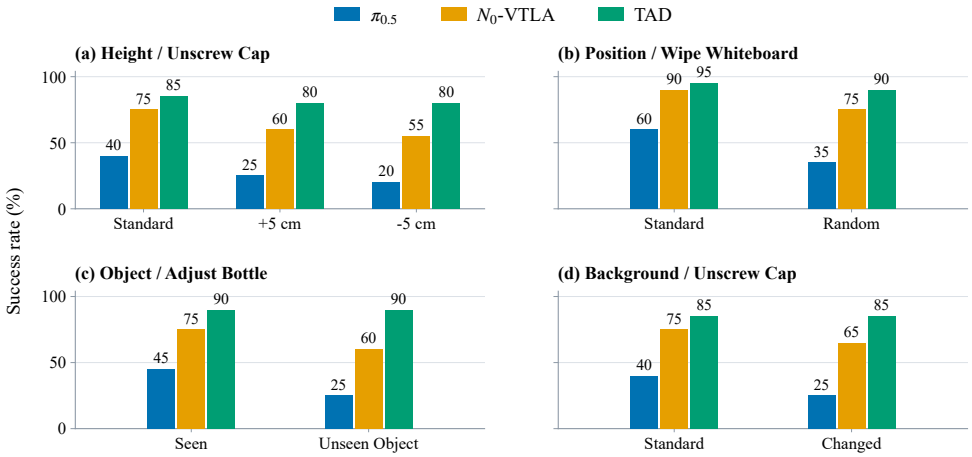

Figure 4: **Real-world generalization across four variations.** (a) Standard and shifted support heights in Unscrew Cap. (b) Standard and Random wiping-region positions in Wipe Whiteboard. (c) Seen and Unseen Object conditions in Adjust Bottle. (d) Standard and Changed backgrounds in Unscrew Cap. Numbers above bars indicate success rates (%).

illumination, and contact surfaces fixed. TAD exceeds both baselines in every tested condition, supporting robustness within these evaluated variations.

4.6 Analysis of Predictive Tactile Distillation

Predictive teacher. Across 14,187 teacher-stage holdout frames, the teacher achieves cosine distance 0.0357 versus 0.1325 for the training-target mean and ranks best on every task (Figure 5; Appendix E.2). **Assistant correction.** On training demonstrations, retaining the residual reduces paired flow-matching error by 67.9% (95% interval: 62.4–73.0%; Figure 6; Appendix E.8). **Tactile attention.** Assistant attention is more spatially concentrated (normalized-entropy difference: -0.139), and highest-attention key removal produces larger command perturbations than matched random removal (Figure 7; Appendix E.9). These offline diagnostics characterize fitted action correction and local sensitivity, without establishing closed-loop gains.

5 Conclusion

TAD transfers tactile predictions to robot control through a frozen teacher and an anchored assistant that adapts supervision and corrects actions. It achieves mean success rates of 90.0% across eight UniVTAC tasks and 87.5% across six real-world tasks, exceeding N_0-VTLA by 6.9 and 10.0 percentage points, respectively. Real-world ablations support the contributions of the assistant, teacher-derived supervision, tactile information, and future-oriented prediction, while evaluations under environment variations assess robustness within the tested conditions. Discussion and limitations are provided in Appendix G.

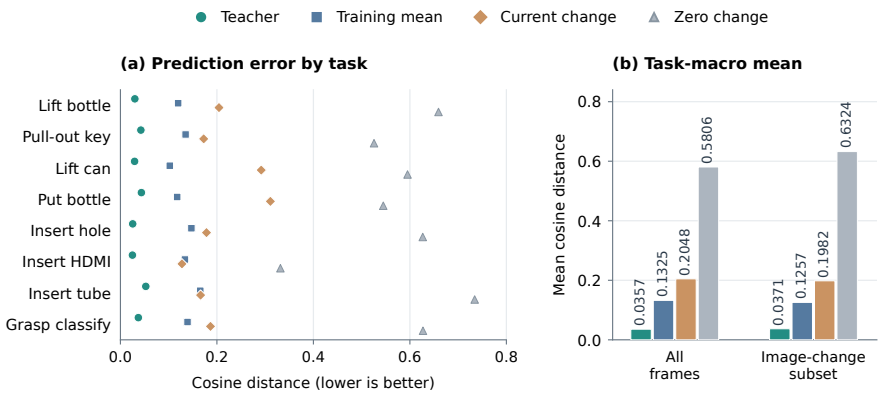

Figure 5: **Tactile prediction quality.** (a) Per-task cosine distance. (b) Task-macro error on all evaluation frames and the image-change subset. Lower is better; Appendix E.2 specifies the evaluation.

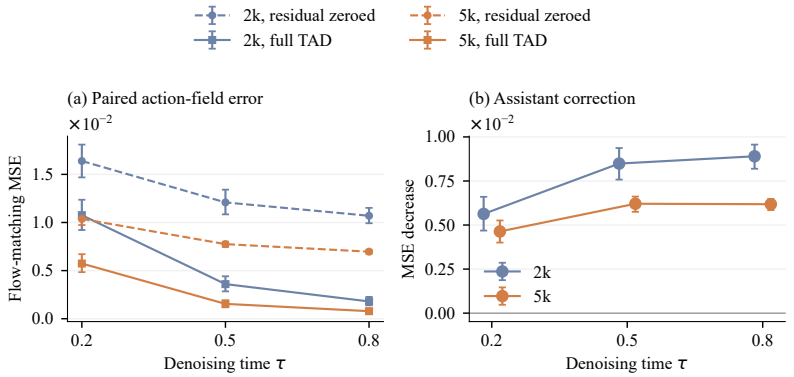

Figure 6: **Assistant action correction.** (a) Error with the residual retained (solid) or zeroed (dashed). (b) Paired error reduction. Colors identify checkpoints; bars show 95% episode-bootstrap intervals. Protocol: Appendix E.8.

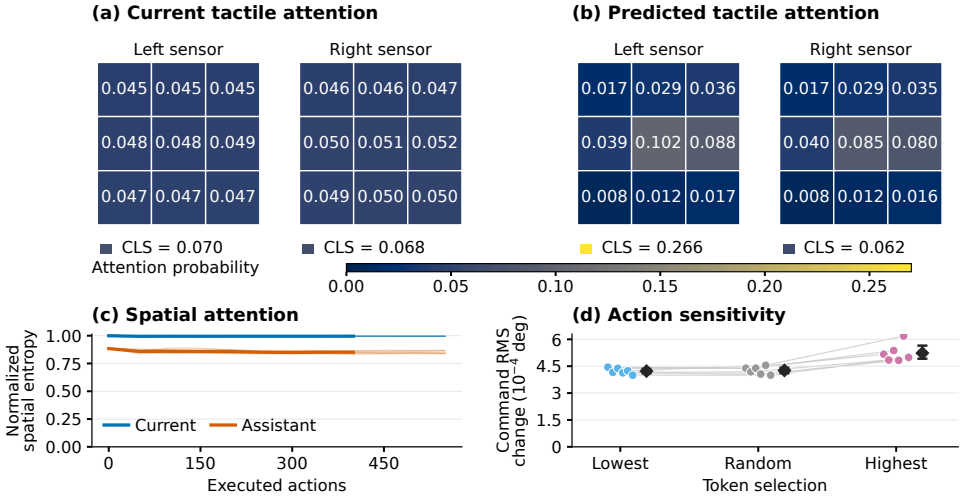

Figure 7: **Tactile attention and paired action sensitivity in six exploratory HDMI episodes. (a,b)** Current and predicted-tactile attention, with CLS weights reported separately. **(c)** Spatial-only normalized entropy; thin/thick lines denote episodes/equal-episode means. **(d)** Command perturbations under matched key removal; circles denote episode means, and diamonds/bars denote cohort means/95% episode-bootstrap intervals. Counterfactual actions were not executed. Appendix E.9 gives the protocol.

REPRODUCIBILITY STATEMENT

Section 3 defines the prediction and control objectives and the allowed information paths. Appendices A–C specify observation transformations, architecture, initialization, loss reduction, optimization, and deployment dependencies; Appendix D states the assumptions and proof of action-input isolation. Appendix E gives data partitions, prediction metrics, rollout settings, and comparison scope, including the inherited exposure of the teacher-stage holdout data. Appendix F documents observation selection and outcome visualization, including real-robot recording identifiers and frame indices. Published simulation baseline scores, eight-task TAD results, real-world comparisons, component ablations, generalization evaluations, and earlier checkpoint diagnostics are reported with their respective evaluation settings. The earlier diagnostic identifies its training checkpoint and episode count; it complements the retrained component ablations and does not establish independent training replication.

ETHICS STATEMENT

The reported experiments use existing demonstration data, simulated robot rollouts, and autonomous physical-robot evaluations; no human-subject study is reported. Tactile image differences and their learned features are not calibrated measurements of force, pressure, or safety. Performance in the evaluated settings does not establish reliable behavior under arbitrary physical sensor errors, unmodeled contacts, or human interaction. The manuscript distinguishes observed images, predicted features, published scores, and measured policy outcomes, and limits empirical claims to the evaluated tasks and conditions.

AI USE STATEMENT

Generative AI tools assisted the conceptual framing and exposition of the method, formulation and drafting of mathematical arguments, feedback on experimental design, interpretation of recorded results, and Chinese–English discussion and drafting. They also assisted literature identification, reference formatting, manuscript editing, figure construction, and visualization and consistency-checking scripts. Reported empirical results are drawn from author-conducted evaluations, recorded training outputs, or explicitly cited publications; language models did not synthesize the reported rollout outcomes or observation images. The real-world comparison and ablation entries in Tables 2 and 3, and the generalization results in Figure 4, report measured outcomes from physical-robot evaluations. This disclosure covers assistance used in preparing the manuscript and associated analysis artifacts. The authors remain responsible for the accuracy, attribution, and final verification of all claims and AI-assisted material.

REFERENCES

Jianxin Bi, Kevin Yuchen Ma, Ce Hao, Mike Zheng Shou, and Harold Soh. VLA-Touch: Enhancing vision-language-action models with dual-level tactile feedback. *arXiv preprint arXiv:2507.17294*, 2025. URL https://arxiv.org/abs/2507.17294.

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*, 2025. URL https://arxiv.org/abs/2504.16054.

Junhao Cai, Zetao Cai, Jiafei Cao, Yilun Chen, Zeyu He, Lei Jiang, Hang Li, Hengjie Li, Yang Li, Yufei Liu, et al. InternVLA-A1: Unifying understanding, generation and action for robotic manipulation. *arXiv preprint arXiv:2601.02456*, 2026. URL https://arxiv.org/abs/2601.02456.

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

Baijun Chen, Weijie Wan, Tianxing Chen, Xianda Guo, Congsheng Xu, Yuanyang Qi, Haojie Zhang, Longyan Wu, Tianling Xu, Zixuan Li, Yizhe Wu, Rui Li, Xiaokang Yang, Ping Luo, Wei Sui, and Yao Mu. UniVTAC: A unified simulation platform for visuo-tactile manipulation data generation, learning, and benchmarking. *arXiv preprint arXiv:2602.10093*, 2026. URL https://arxiv.org/abs/2602.10093.

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, 2023. URL https://arxiv.org/abs/2302.00111.

Pengcheng Fang, Hongli Chen, and Xiaohao Cai. Privileged foresight distillation: Zero-cost future correction for world action models. *arXiv preprint arXiv:2604.25859*, 2026. URL https://arxiv.org/abs/2604.25859.

Ruoxuan Feng, Jiangyu Hu, Wenke Xia, Tianci Gao, Ao Shen, Yuhao Sun, Bin Fang, and Di Hu. AnyTouch: Learning unified static-dynamic representation across multiple visuo-tactile sensors, 2025. URL https://arxiv.org/abs/2502.12191.

Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion, 2016. URL https://arxiv.org/abs/1610.00696.

Abraham George, Selam Gano, Pranav Katragadda, and Amir Barati Farimani. VITaL pretraining: Visuo-tactile pretraining for tactile and non-tactile manipulation policies. *arXiv preprint arXiv:2403.11898*, 2024. URL https://arxiv.org/abs/2403.11898.

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.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations, 2024. URL https://arxiv.org/abs/2412.14803.

Binghao Huang, Yixuan Wang, Xinyi Yang, Yiyue Luo, and Yunzhu Li. 3D-ViTac: Learning fine-grained manipulation with visuo-tactile sensing, 2024. URL https://arxiv.org/abs/2410.24091.

Yuzhe Huang, Pei Lin, Wanlin Li, Daohan Li, Jiajun Li, Jiaming Jiang, Chenxi Xiao, and Ziyuan Jiao. TaF-VLA: Tactile-force alignment in vision-language-action models for force-aware manipulation, 2026. URL https://arxiv.org/abs/2601.20321.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag 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. *arXiv preprint arXiv:2406.09246*, 2024. URL https://arxiv.org/abs/2406.09246.

Xu Lan, Xiatian Zhu, and Shaogang Gong. Knowledge distillation by on-the-fly native ensemble, 2018. URL https://arxiv.org/abs/1806.04606.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Zhuoyang Liu, Jiaming Liu, Jiadong Xu, Nuowei Han, Chenyang Gu, Hao Chen, Kaichen Zhou, Renrui Zhang, Kai Chin Hsieh, Kun Wu, Zhengping Che, Jian Tang, and Shanghang Zhang. MLA: A multisensory language-action model for multimodal understanding and forecasting in robotic manipulation. *arXiv preprint arXiv:2509.26642*, 2025. URL https://arxiv.org/abs/2509.26642.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.03643.

Yunfan Lou, Yifan Ye, Yankai Fu, Jun Cen, Xiaowei Chi, Yaoxu Lyu, Peidong Jia, Sirui Han, Zhihe Lu, and Shanghang Zhang. Dream-Tac: A unified tactile world action model for contact-rich robot manipulation. *arXiv preprint arXiv:2606.08737*, 2026. URL https://arxiv.org/abs/2606.08737.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5191–5198, 2020. doi: 10.1609/aaai.v34i04.5963. URL https://ojs.aaai.org/index.php/AAAI/article/view/5963.

NeoteAI Team and Fudan TEAI Team. N_0-TWAM: Scaling tactile-native world-action model for contact-rich manipulation. *arXiv preprint arXiv:2607.23783*, 2026a. URL https://arxiv.org/abs/2607.23783.

NeoteAI Team and Fudan TEAI Team. N_0-VTLA: Scaling vision-tactile-language-action model with latent tactile tokens. *arXiv preprint arXiv:2607.23782*, 2026b. URL https://arxiv.org/abs/2607.23782.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation, 2019. URL https://arxiv.org/abs/1904.05068.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets, 2014. URL https://arxiv.org/abs/1412.6550.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. URL https://arxiv.org/abs/2508.10104.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*, 2017. URL https://arxiv.org/abs/1703.01780.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation, 2019. URL https://arxiv.org/abs/1910.10699.

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

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, 2025. URL https://arxiv.org/abs/2503.02881.

Fengyu Yang, Chao Feng, Ziyang Chen, Hyoungseob Park, Daniel Wang, Yiming Dou, Ziyao Zeng, Xien Chen, Rit Gangopadhyay, Andrew Owens, and Alex Wong. Binding touch to everything: Learning unified multimodal tactile representations, 2024. URL https://arxiv.org/abs/2401.18084.

Guo Ye, Zexi Zhang, Xu Zhao, Shang Wu, Haoran Lu, Shihan Lu, and Han Liu. Learning to feel the future: DreamTacVLA for contact-rich manipulation. *arXiv preprint arXiv:2512.23864*, 2025. URL https://arxiv.org/abs/2512.23864.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=DJSZGGZYVi.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer, 2016. URL https://arxiv.org/abs/1612.03928.

Chaofan Zhang, Peng Hao, Xiaoge Cao, Xiaoshuai Hao, Shaowei Cui, and Shuo Wang. VTLA: Vision-tactile-language-action model with preference learning for insertion manipulation. *arXiv preprint arXiv:2505.09577*, 2025. URL https://arxiv.org/abs/2505.09577.

Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. Deep mutual learning, 2017. URL https://arxiv.org/abs/1706.00384.

Jialiang Zhao, Yuxiang Ma, Lirui Wang, and Edward H. Adelson. Transferable tactile transformers for representation learning across diverse sensors and tasks, 2024. URL https://arxiv.org/abs/2406.13640.

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. URL https://arxiv.org/abs/2304.13705.

Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loïc Magne, Avnish Narayan, You Liang Tan, Guanzhi Wang, Qi Wang, Jiannan Xiang, Yinzhen Xu, Seonghyeon Ye, Jan Kautz, Furong Huang, Yuke Zhu, and Linxi Fan. FLARE: Robot learning with implicit world modeling. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 3952–3971. PMLR, 2025. URL https://proceedings.mlr.press/v305/zheng25a.html.

A Additional methodological details

A.1 Observations and temporal correspondence

Each observation contains scene and wrist RGB images, a language instruction, robot state, and left and right tactile images. Each sensor uses the first frame of the same demonstration or evaluation episode as its reference. The reference is fixed within an episode and reset at the next episode. Sensor availability is represented explicitly and masks the corresponding tactile features.

Tactile images are decoded in RGB order, resized to 168×224, padded with 28 black rows above and below, and normalized to $[-1, 1]$. This preprocessing $\mathcal{R}$ precedes subtraction. Current conditioning uses the signed difference $\mathcal{R}(T_t^v) - \mathcal{R}(T_0^v)$, whereas teacher targets use $\mathcal{R}(T_{t+\delta}^v) - \mathcal{R}(T_t^v)$ for the same sensor v. Neither difference is converted to an absolute value or clipped after subtraction; matching padding regions cancel. These image differences describe changes in sensor appearance rather than calibrated force.

The future offset is $\delta = 10$ observation frames at 30 Hz. Teacher-training pairs require valid current and future views from the same episode and sensor; missing future frames are not replaced by the final episode frame. Current-view availability and future-target validity are separate masks. Thus observations near an episode boundary can condition the student even without a valid future target. The teacher, assistant, and assistant EMA all predict from current observations, so student distillation does not require future frames.

A.2 Tactile representation adaptation

The tactile backbone is DINOv3 ViT-L/16. A 224×224 difference image produces a 14×14 patch grid. We discard register tokens and retain the global class feature together with nine features obtained by adaptive average pooling to 3×3. Each sensor therefore contributes ten width-1024 features. Features are concatenated in a fixed sensor order, with the global feature first and spatial features in row-major order within each sensor.

Before policy and teacher learning, the backbone is adapted by reconstructing complete tactile difference images. A learned projection feeds the ten-token representation to a two-layer decoder of width 256 with eight attention heads. Its 196 learned patch queries reconstruct the image without a dense-patch bypass. For reconstructed differences $\widehat{D}$ and a valid, unpadded-pixel indicator r, the objective is

$$\mathcal{L}_{\text{rec}} = \frac{\sum_{b,v,p,c} r_{bvpc} (\widehat{D}_{bvpc} - D_{bvpc})^2}{\sum_{b,v,p,c} r_{bvpc}}. \tag{11}$$

The backbone, projection, and decoder are optimized jointly, without random masking, cropping, flipping, color perturbation, or noise augmentation. After adaptation, the reconstruction decoder is discarded and the backbone is frozen. Future targets are extracted before the trainable projection using an independently stored frozen copy of the adapted backbone, keeping the target representation fixed during teacher learning.

A.3 Query conditioning and architecture

The visual–language prefix combines image features, instruction embeddings, and discretized robot state. Its final normalized states are pooled over valid positions. Let h_i^P denote these states, m_i^P their validity mask, and $g_t = W_g M_t$ the projected tactile memory. Each of the two Tactile Encoder blocks computes

$$\begin{aligned} c_t &= \frac{\sum_i m_i^P h_i^P}{\max(1, \sum_i m_i^P)}, \qquad b_t = W_c \,\mathrm{sg}(c_t), \\ q^+ &= q + \mathrm{MHA}\big(\mathrm{LN}(q + b_t), \mathrm{LN}(g_t), \mathrm{LN}(g_t)\big), \\ q' &= q^+ + \mathrm{FFN}(\mathrm{LN}(q^+)). \end{aligned} \tag{12}$$

The conditioning vector is broadcast across five queries and affects only attention queries; keys and values remain tactile features. Adding the attention output to q, rather than $q + b_t$, avoids directly adding visual–language context to the latent representation. The student stops gradients through c_t

Table 4: Architectural dimensions. Prediction slots represent sensor and spatial features at one future offset; action slots represent positions in an action chunk.

Component	Configuration
Visual–language / action streams	Width 2048 / 1024; 18 layers each
Current tactile memory M_t	Two sensors; 10 features each; width 1024
Compression queries Q_z	5×2048; 2 blocks; 8 heads
Future-decoder queries Q_T	20×512; 2 blocks; 8 heads
Future-decoder memory / output	Five latents and one state token / 20×1024
Student predictive queries Q_F	20×1024; readout after layer 10
Spatial / assistant action attention	Width 256; 4 heads each
Action horizon / representation width	50 / 32; 8 physical coordinates

on this path, while its projections and Tactile Encoder remain trainable. Its prefix can still receive gradients through action-expert attention. Teacher and assistant prediction branches use the original frozen prefix.

Table 4 distinguishes the three query banks. The future decoder projects five compression latents to width 512 and appends a continuous-state token from a two-layer SiLU network. Twenty learned queries read these six memory tokens through two pre-normalized transformer-decoder blocks, followed by normalization and projection to width 1024. The decoder uses feed-forward width 2048, GELU activation, and zero dropout. Its outputs follow the target encoder's sensor and spatial ordering at a single future offset.

The student's predictive queries enter the action expert after the tactile and action tokens. Their layer-ten readout supplies the distillation features, while the states continue through the remaining layers. The action expert reads prefix keys and values despite the streams' different hidden widths. Compression, future-decoder, and student predictive queries are distinct parameters; the teacher, online assistant, and EMA retain separate copies of prediction parameters.

B Spatial and assistant action readouts

B.1 Reading current tactile detail

Let h_A denote final normalized action states and $M_{v,j}$ the frozen feature at position j of sensor v. Learned sensor embeddings e_v and within-sensor position embeddings e_j augment the spatial memory:

$$\begin{aligned} \widetilde{M}_{v,j} &= W_m \,\mathrm{LN}(M_{v,j}) + e_v + e_j, \\ \Delta h_{\mathrm{loc}} &= W_{\mathrm{loc}} \,\mathrm{Attn}_4\big(W_q \,\mathrm{LN}(h_A), W_k \widetilde{M}, W_v \widetilde{M}\big). \end{aligned} \tag{13}$$

There are two sensor embeddings and ten position embeddings, including the global-feature position. Attention has width 256 and four heads of width 64, with scores scaled by $1/\sqrt{64}$ and normalized over available keys. Concatenated head outputs are projected to width 1024 and added to the action states immediately before the original action-output projection. This readout introduces no extra expert tokens and does not modify tactile memory or predictive states.

The bias-free output map W_{loc} is initialized to zero, preserving the initial action field. The readout is trained through the main action objective without an additional spatial reconstruction or alignment loss. With one available sensor, attention reads its ten features; with neither sensor available, the residual is zero throughout training and deployment.

B.2 Reading assistant predictions

The assistant produces H_{TA} from current observations using its own projection, Tactile Encoder, and future decoder together with the original frozen visual–language and tactile context. Noisy actions, flow time, and demonstration-action labels do not enter this prediction branch. Action dependence

enters only through the readout:

$$
\begin{aligned}
q_j &= W_x x_{\tau,j} + W_s s_t + e(\tau) + p_j, \\
\widetilde{H} &= W_M \,\mathrm{LN}(H_{\mathrm{TA}}), \\
C &= \mathrm{Attn}_4\big(W_Q \,\mathrm{LN}(q), W_K \widetilde{H}, W_V \widetilde{H}\big), \\
r &= \mathrm{SiLU}\big(W_h \,\mathrm{LN}(q + C)\big).
\end{aligned} \tag{14}
$$

The time embedding is a two-layer SiLU network applied to $[\tau, \sin(\pi\tau), \cos(\pi\tau)]$, and p_j is learned for each of the fifty action positions. Attention has width 256 and four heads. Affine maps include biases unless stated otherwise. Separate output maps produce the auxiliary field $v^{\mathrm{aux}} = W_u r + b_u$ and the bias-free residual $\Delta v_{\mathrm{TA}} = W_r r$; only the residual enters the combined denoising field.

The residual map is initialized to zero, while the auxiliary map has a nonzero initialization. Auxiliary flow matching therefore provides a gradient path to the assistant prediction branch from the start of training. Both outputs are zero when all tactile views are absent. These action-conditioned queries read predictive memory without writing back to it, preserving the information paths established in Appendix D.

C Optimization, normalization, and deployment

C.1 Initialization and trainable components

Representation adaptation, eight-task policy initialization, teacher learning, and Stage I each use 20,000 updates, as specified in Section 4.1. Teacher learning initializes a separate current-observation Tactile Encoder from the policy and trains its tactile projection, five-query Tactile Encoder, and future decoder. Its visual–language and tactile backbones and target encoder remain frozen; the action expert does not participate in prediction. The resulting teacher is frozen before distillation.

Stage I combines action flow matching with teacher-feature distillation. The student's visual–language and action pathways can adapt, while the tactile backbone remains frozen. Stage II starts from this eight-task predictive student, adds the zero-output spatial and assistant residuals, and initializes the online assistant and EMA prediction parameters from the frozen teacher. The original teacher remains fixed as the assistant's anchor. The component comparisons follow the shared demonstration data and policy-update budget specified in Section 4.4.

The assistant-correction diagnostic in Section 4.6 uses earlier HDMI task-adaptation checkpoints at 2,000 and 5,000 updates; these checkpoints are distinct from the final benchmark evaluations.

The combined action objective trains the student and both action readouts, with gradients reaching the assistant Tactile Encoder through its residual. Auxiliary flow matching also trains the assistant Tactile Encoder and auxiliary readout. Student alignment uses detached EMA predictions; assistant anchoring uses detached original-teacher predictions. Thus student distillation does not directly update the online assistant, EMA, or original teacher. Shared student expert parameters can receive action gradients despite the forward restriction on predictive tokens.

C.2 Masks and objective reductions

For sample b, let $\mathcal{V}_b$ be its valid alignment-token set and $\mathcal{B}_+ = \{b : |\mathcal{V}_b| > 0\}$. Cosine alignment is

$$
\mathcal{C}(U, V; m) = \frac{1}{|\mathcal{B}_+|} \sum_{b \in \mathcal{B}_+} \frac{1}{|\mathcal{V}_b|} \sum_{j \in \mathcal{V}_b} \left[1 - \frac{U_{bj}^\top V_{bj}}{\max(\|U_{bj}\|_2, \epsilon) \max(\|V_{bj}\|_2, \epsilon)} \right], \tag{15}
$$

with $\epsilon = 10^{-8}$ and $\mathcal{C} = 0$ for an empty $\mathcal{B}_+$. Averaging within each sample before averaging contributing samples gives equal sample weight regardless of whether one or two sensors are valid. Teacher pretraining requires valid current and future views; student distillation and assistant anchoring require only current-view validity. Predictive-token visibility never depends on future-label availability.

Table 5: Optimization settings. The main schedule applies to teacher learning and both distillation stages; representation adaptation has the separate rates and schedule listed below.

Setting	Value
AdamW betas / epsilon	$(0.9, 0.95)$ / 10^{-8}
Main rate / warmup	5×10^{-5} / 1,000 updates
Main weight decay / gradient clipping	10^{-10} / 1.0
Reduced-rate multiplier	0.1
Global batch / precision	64 / BF16
Distillation / auxiliary / anchor weights	0.01 / 0.1 / 0.01
EMA decay	0.999
Representation backbone / decoder rate	10^{-5} / 10^{-4}
Representation projection rate / decay	10^{-4} / 0.05
Representation warmup / final rate	1,000 updates / 1% of peak
Training seed / hardware	42 / eight H100 GPUs

For batch size B, action length $L = 50$, padded action width $d_a = 32$, and current-touch indicator p_b, the action losses are

$$\mathcal{L}_{\mathrm{FM}} = \frac{1}{BLd_a} \sum_b \|v_b - u_b\|_F^2, \qquad \mathcal{L}_{\mathrm{aux}} = \frac{1}{BLd_a} \sum_b p_b \|v_b^{\mathrm{aux}} - u_b\|_F^2. \tag{16}$$

In Stage II, v is the combined field, including both residual pathways, and receives a single flow-matching loss. The auxiliary loss retains the full-batch denominator. All fifty action slots remain structurally valid for attention; the physical action consists of seven joint coordinates and one gripper coordinate, padded to width 32. Flow times are sampled as $0.001 + 0.999\,\mathrm{Beta}(1.5, 1)$.

The assistant-correction diagnostic in Figure 6 instead averages evaluation error over the eight physical action dimensions, excluding padding.

C.3 Optimizer and EMA updates

Table 5 summarizes optimization. Teacher learning and both distillation stages use warmup followed by a constant main learning rate. Tactile projections and Tactile Encoders use a rate multiplier of 0.1; the student also applies this multiplier to its latent projection. The future decoder, assistant action head, and spatial readout use the full rate. Representation adaptation uses separate backbone and reconstruction-module rates, the same AdamW betas and gradient clipping, and a cosine decay after warmup to one percent of the peak rate.

At step k, the original teacher and current EMA first produce detached anchor and distillation targets. The weighted Stage II objective then updates the student and online assistant once, after gradient clipping. Only after this optimizer step are the updated assistant prediction parameters incorporated into the EMA:

$$\bar{\psi}_{k+1} = 0.999\bar{\psi}_k + 0.001\psi_{k+1}. \tag{17}$$

The EMA includes the tactile projection, Tactile Encoder, and future decoder. It excludes action readouts, the student, and all frozen encoders. Each student update therefore uses targets from the preceding EMA state.

C.4 Normalization and inference

Task adaptation retains two transformations of the same raw observation. The student and assistant action queries use task-specific state and action normalization. The assistant prediction branch retains the original eight-task state normalization and tokenization used by its teacher. Images and sensor availability are shared. This preserves the teacher's input convention for predictive supervision while using task-specific statistics for action generation.

For each action chunk, observation-dependent memories and visual–language context are computed once. Ten Euler steps integrate the combined field from Gaussian noise:

$$x_{\tau - 0.1} = x_\tau - 0.1 v(o_t, x_\tau, \tau). \tag{18}$$

Expert suffix states and action readouts are recomputed at each flow time. Deployment uses the trained online assistant, not its EMA copy, together with the original frozen visual–language and tactile context. Original-teacher anchoring, future-target encoding, and auxiliary supervision are unnecessary at inference. All parameters remain fixed, and no future observations enter the deployed policy.

D INFORMATION FLOW AND ADAPTATION MECHANISMS

This section establishes the separation between predictive features and noisy action inputs, and explains how those features can nevertheless learn from action supervision.

D.1 ACTION-INPUT ISOLATION

For token groups $[P; Z; A; F]$, the attention mask in Figure 3 is

$$\mathcal{M} = \begin{array}{c|cccc} & P & Z & A & F \\ \hline P & 1 & 0 & 0 & 0 \\ Z & 1 & 1 & 0 & 0 \\ A & 1 & 1 & 1 & 1 \\ F & 1 & 1 & 0 & 1 \end{array}. \tag{19}$$

Rows denote queries and columns denote keys. A one permits the corresponding block, subject to current-observation validity masks; within-group attention is bidirectional. The same rule holds at every layer. Inference caches prefix keys and values once per observation, while suffix queries $[Z; A; F]$ retain the corresponding rows of this mask.

Let $\mathcal{S} = P \cup Z \cup F$ collect the prefix, tactile-latent, and predictive-token states. These groups need not have equal hidden dimensions.

Proposition 1 (Action-input isolation)**.** *Fix observations, model parameters, flow time, validity masks, and any auxiliary random choices. Suppose initial states in $\mathcal{S}$ are independent of x_τ, all attention operations obey Equation 19, and other transformations are tokenwise with conditioning independent of x_τ. Assume attention is well defined, including any empty valid-key set. Then any two noisy-action inputs x, x' give, at every layer,*

$$h_r^{(l)}(x) = h_r^{(l)}(x') \quad (r \in \mathcal{S}), \qquad \frac{\partial h_F^{(l)}}{\partial x_\tau} = 0 \tag{20}$$

where the derivative exists. The intermediate predictive readout H_S has the same invariance.

Proof. Initial prefix and tactile states depend on observations, and predictive queries are learned parameters, so the claim holds before the first layer. Suppose it holds for every state in $\mathcal{S}$ before an attention block. For a query $r \in \mathcal{S}$, the mask restricts its valid-key set to $\mathcal{K}_r \subseteq \mathcal{S}$. A head computes

$$a_r = \sum_{j \in \mathcal{K}_r} \frac{\exp(q_r^\top k_j / \sqrt{d})}{\sum_{i \in \mathcal{K}_r} \exp(q_r^\top k_i / \sqrt{d})} v_j. \tag{21}$$

By the induction hypothesis, queries, keys, and values are identical for x and x'. Fixed projections, positional information, and masks then give identical attention outputs. An action-independent convention for empty key sets preserves the same conclusion. Residual addition, tokenwise normalization, feed-forward transformations, and conditioning on fixed observations and flow time preserve equality. Induction through all layers proves the claim; the linear readout of predictive states inherits it. □

Why masking is required at every layer. Blocking only the final direct edge from A to F would still allow an earlier path such as $A \to Z \to F$. Equation 19 excludes such indirect routes by keeping $\mathcal{S}$ closed under incoming attention dependencies. Spatial readout modifies only action states, and assistant action queries cannot write back to predictive memory, so neither introduces a reverse path.

Scope of the guarantee. The proposition concerns forward dependence at fixed parameters and flow time. Shared expert transformations and predictive-token parameters can still learn from action loss because action states read predictive features. Student predictive states may depend on flow time, whereas assistant memory depends only on current observations. The guarantee does not establish counterfactual tactile prediction: the teacher learns tactile evolution in demonstrations without candidate-action conditioning.

D.2 Residual initialization and action supervision

The spatial and assistant residual projections are bias-free and initialized to zero, preserving the student's action field when introduced. For one action slot with $v = v^{\text{main}} + W_r r$ and $\ell = \frac{1}{2}\|v-u\|_2^2$, the residual gradients are

$$\nabla_{W_r}\ell = (v-u)r^\top, \qquad \nabla_r \ell = W_r^\top (v-u). \tag{22}$$

At initialization, this path can update W_r but supplies no gradient to r. The separately initialized auxiliary head supplies feature gradients proportional to $W_u^\top(v^{\text{aux}} - u)$, enabling action supervision of the assistant before its residual projection changes. The spatial residual similarly begins learning through its output projection, but has no auxiliary action objective. These properties explain initialization and gradient routing; subsequent improvement is an empirical question assessed by the experiments.

D.3 Teacher anchoring and EMA targets

The anchor penalizes cosine disagreement between the assistant and the frozen teacher for the same observation. It constrains feature direction rather than parameter distance or feature magnitude. EMA has a different role: it averages the assistant prediction parameters used to generate student targets. With $\bar{\psi}_0 = \psi_0$, Equation 8 expands to

$$\bar{\psi}_k = \beta^k \psi_0 + (1-\beta)\sum_{i=1}^{k} \beta^{k-i}\psi_i. \tag{23}$$

The coefficients sum to one. For a nonlinear predictor, this parameter average is not an average of past feature outputs. The teacher remains fixed, the online assistant learns from action and anchoring losses, and student alignment treats the EMA prediction as a detached target. Appendix C specifies the update order and parameter scope. The reported component ablation evaluates the assistant as a whole rather than isolating anchoring and EMA individually.

D.4 Missing touch and deployment inputs

When both tactile views are absent, the spatial and assistant residuals, auxiliary loss, and predictive alignment are zero; the main action objective remains active. This does not force equality with the original visual policy, since tactile supervision may have changed shared parameters. Missing touch is treated as unavailable information, not a zero-contact measurement.

Future-view validity affects teacher-pretraining targets only. It neither changes student predictive-token visibility nor supplies a future-dependent input at deployment. The deployed policy uses current observations, episode-initial tactile references, and internal denoising variables with fixed parameters. Assistant memory is reusable across denoising steps, while student predictive states are recomputed with the expert.

E Experimental details and evaluation

This appendix complements the results in Section 4 with the teacher-evaluation protocol, physical-task definitions, and component interventions. The simulation benchmark, real-world comparison, ablations, and generalization results use the values reported in the main text. Teacher prediction error and closed-loop task success measure different aspects of the method and are evaluated separately.

Table 6: Demonstration and teacher-evaluation coverage. Every task contains 100 demonstrations. "Valid holdout" counts frames with valid future targets; "Image change" denotes the subset in Equation 26. These counts are frames, not independent control trials.

Task	Demonstration frames	Valid holdout	Image change
Grasp and classify	5,265	422	404
Insert HDMI	11,700	1,070	132
Insert hole	16,713	1,546	1,193
Insert tube	18,679	1,678	1,612
Lift bottle	30,432	2,902	2,884
Lift can	19,145	1,825	1,673
Pull out key	20,126	1,863	1,417
Put bottle in shelf	30,020	2,881	2,319
Total	152,080	14,187	11,634

E.1 Demonstrations and temporal sampling

The UniVTAC dataset (Chen et al., 2026) contains eight tasks with 100 demonstration episodes per task: 800 episodes and 152,080 frames at 30 Hz. Each frame provides scene and wrist RGB observations, bilateral tactile images, robot state, and language conditioning. Table 6 summarizes coverage. Episode durations differ across tasks, so equal episode counts do not imply equal frame counts.

Current tactile conditioning compares each sensor image with its own episode-initial reference. Future supervision instead uses the difference between frames $t + 10$ and t, after the same preprocessing. The prediction horizon is approximately 0.33 s. A target is valid only when both observations belong to the same episode and sensor; episode boundaries are never crossed or padded with a terminal image. Missing future targets exclude a frame from teacher evaluation, but do not prevent action learning from observation-conditioned teacher or assistant predictions.

E.2 Teacher evaluation and comparison scope

The original teacher is evaluated on demonstrations held out from teacher fitting. This split applies to teacher fitting; the inherited tactile representation and base action policy had access to these demonstrations during their earlier training. The evaluation therefore measures teacher-stage prediction under an inherited representation, rather than end-to-end generalization to demonstrations unseen throughout training. All predictors are evaluated on the same valid frames, frozen target features, and sensor–spatial slots.

Let $\mathcal{V}_{qt}$ denote the valid feature slots for evaluation frame t of task q. Each available sensor contributes one global token and nine spatial tokens. For predictor b, the per-frame error is

$$e_{qt}^{(b)} = \frac{1}{|\mathcal{V}_{qt}|} \sum_{j \in \mathcal{V}_{qt}} \left[1 - \frac{\langle \widehat{Y}_{qtj}^{(b)}, Y_{qtj} \rangle}{\max(\|\widehat{Y}_{qtj}^{(b)}\|_2, \epsilon) \max(\|Y_{qtj}\|_2, \epsilon)} \right], \qquad \epsilon = 10^{-8}. \tag{24}$$

Frames without valid slots are excluded. For the valid-frame set $\mathcal{F}_q$, the task error and task-macro error are

$$E_q^{(b)} = \frac{1}{|\mathcal{F}_q|} \sum_{t \in \mathcal{F}_q} e_{qt}^{(b)}, \qquad E_{\text{macro}}^{(b)} = \frac{1}{8} \sum_{q=1}^{8} E_q^{(b)}. \tag{25}$$

Thus, available tokens are averaged within a frame, frames within a task, and then tasks with equal weight. Longer tasks do not dominate the final average. Within a task, longer demonstrations still contribute more frames; the metric is not an episode-balanced average.

We compare the teacher with three simple predictors. The *training-target mean* uses the mean frozen future-target feature for each of the twenty slots, accumulated from teacher-training samples rather than evaluation targets; repeated training samples may contribute to this mean. It makes a constant prediction across observations. The *current-change* predictor encodes the current-minus-initial tactile difference with the frozen target encoder. The *zero-change* predictor encodes an all-zero difference

Table 7: Per-task cosine distance on teacher-stage holdout frames. Lower is better. N counts frames; the final row sums counts and averages task errors. The inherited encoder and policy previously saw these demonstrations.

Task	N	Teacher	Training mean	Current change	Zero change
Grasp and classify	422	**0.0372**	0.1391	0.1867	0.6271
Insert HDMI	1,070	**0.0249**	0.1337	0.1275	0.3316
Insert hole	1,546	**0.0254**	0.1471	0.1784	0.6267
Insert tube	1,678	**0.0526**	0.1655	0.1662	0.7343
Lift bottle	2,902	**0.0299**	0.1197	0.2043	0.6593
Lift can	1,825	**0.0294**	0.1024	0.2918	0.5951
Pull out key	1,863	**0.0424**	0.1351	0.1726	0.5256
Put bottle in shelf	2,881	**0.0435**	0.1176	0.3108	0.5448
Total / macro mean	14,187	**0.0357**	0.1325	0.2048	0.5806

image; its output is an encoded feature, not an all-zero feature vector. All three preserve sensor and spatial correspondence.

The teacher outperforms all three predictors on every task in Table 7. Across 14,187 valid frames, its task-macro cosine distance is 0.0357, compared with 0.1325 for the training-target mean, 0.2048 for current-change features, and 0.5806 for zero-change-image features (Figure 5). These results support prediction beyond constant outputs and current-feature reuse. This metric measures agreement in the frozen tactile feature space. It does not measure image reconstruction or calibrated contact force, and the teacher does not condition its prediction on a proposed action.

Frames with tactile image change. We also evaluate a subset with measurable future image change. Let $\mathcal{U}_{qt}$ contain the available current/future sensor pairs and let $R_t^v = \mathcal{R}(T_t^v)$ be a preprocessed RGB image with values in $[-1, 1]$. Define

$$c_{qt} = \frac{1}{|\mathcal{U}_{qt}|} \sum_{v \in \mathcal{U}_{qt}} \text{mean}_{p,c} \left| R_{t+10}^v(p,c) - R_t^v(p,c) \right|, \qquad \mathcal{F}_q^{\text{chg}} = \{t \in \mathcal{F}_q : c_{qt} \geq 0.01\}. \quad (26)$$

The mean includes the full preprocessed image grid and weights available sensors equally. Subset membership depends on observed image change, not prediction error. Applying the same task-macro aggregation to the 11,634 selected frames gives errors of 0.0371, 0.1257, 0.1982, and 0.6324 for the teacher, training-target mean, current-change, and zero-change predictors, respectively. Coverage varies by task (Table 6). This appearance-based statistic is not a contact label or a measurement of force or pressure.

E.3 Control metrics and benchmark comparisons

For task q, success is the fraction of evaluated executions satisfying its task-completion criterion. The reported average is the unweighted mean of task success rates. Simulation uses the native task success criteria. Table 1 reports TAD's final eight-task results, with a mean of 90.0%. The five baseline rows are the published results from Table 1 of NeoteAI Team & Fudan TEAI Team (2026b); their task scores and reported averages are retained. This benchmark compares reported task performance and does not establish identical inherited pretraining data or total training compute across methods.

Physical evaluation uses fixed policy checkpoints. The online assistant denotes the assistant trained jointly with the student; it does not perform test-time adaptation. The trained assistant and its frozen context remain active during deployment, whereas the EMA supplies training targets only. Selected successful and unsuccessful recordings in Appendix F.2 illustrate execution behavior; their balanced selection is separate from the quantitative trial outcomes.

E.4 Real-world platform and task outcomes

The physical platform consists of a Franka Research 3 manipulator, a Robotiq 2F-85 gripper, and Xense Photon tactile sensors at the fingertips (Figure 8). Two RealSense D435i cameras provide wrist

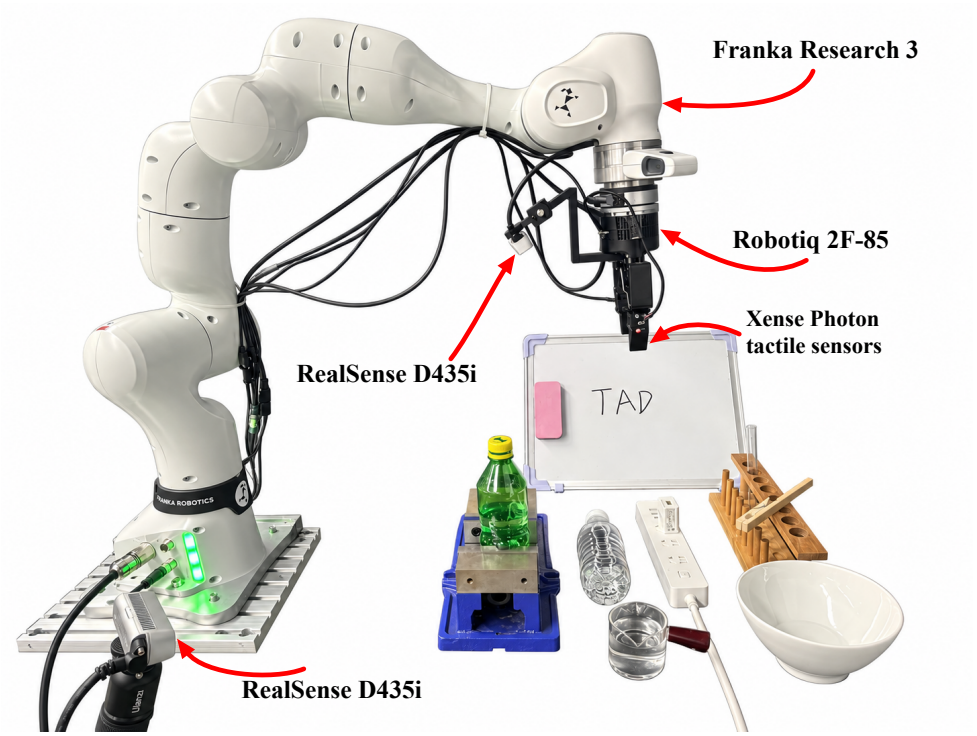

Figure 8: **Real-world experimental platform.** A Franka Research 3 manipulator with a Robotiq 2F-85 gripper, Xense Photon tactile sensors, and two RealSense D435i cameras providing wrist and external scene views. The workspace contains objects used in the six physical tasks.

Table 8: **Physical-task completion outcomes.** Task-level descriptions used to interpret the reported success rates.

Task	Completion outcome
Insert Plug	Seat the plug in the target socket.
Unscrew Cap	Fully unscrew the cap from the bottle.
Lift Tube	Lift the test tube using the gripper-held clamp without dropping it.
Upright Bottle	Bring the bottle into an upright, stable configuration.
Wipe Whiteboard	Remove the markings from the target board region.
Pour Water	Transfer water from the source container into the target vessel.

and external scene views. Observations combine these RGB images, bilateral touch, robot state, and a language instruction.

The six tasks are *Insert Plug*, *Unscrew Cap*, *Lift Tube*, *Upright Bottle*, *Wipe Whiteboard*, and *Pour Water*. *Adjust Bottle* in the ablation and generalization results denotes the same bottle-uprighting task. Table 8 summarizes their completion outcomes. Table 2 reports the completed comparison of $\pi_{0.5}$, N_0-VTLA, and TAD, with mean success rates of 47.5%, 77.5%, and 87.5%, respectively. The qualitative observations use the same task definitions.

E.5 Real-world ablation definitions

Table 3 compares Full TAD with four interventions on the same six tasks. All conditions use the same demonstrations and policy-update budget, with affected upstream stages retrained. Each condition is evaluated on the same twenty predefined test configurations per task, giving 600 rollouts across five conditions. Consequently, one successful trial changes a task score by five percentage points. The average weights the six tasks equally.

Without the assistant. This condition retains tactile input, the future teacher, student distillation, and the spatial readout. Direct supervision from the frozen teacher replaces EMA assistant targets. The assistant, its auxiliary objective and teacher anchor, and its residual action correction are removed. The comparison therefore tests the assistant mechanism as a whole rather than isolating its target adaptation from its action correction.

Without the teacher. This condition retains tactile input and an action-trained assistant, but removes teacher-derived initialization and supervision, including the teacher anchor. Affected training stages are restarted so that future-teacher information is not inherited through pretrained student or assistant weights. This intervention tests teacher-derived supervision as a whole, rather than the anchor alone.

Without tactile information. This condition excludes tactile observations, tactile targets, and tactile-derived initialization throughout training and deployment. It retains RGB, language, robot state, and the matched action-training data and policy-update budget. Touch-dependent conditioning and teacher–assistant pathways are removed. This is a modality-level intervention; it does not imply identical parameter counts or total training compute.

Current targets instead of future prediction. This condition retains the complete mechanism but replaces the teacher's future-change target with current tactile features $E(D_t)$. The affected teacher, student distillation, and assistant initialization are retrained using this target. It tests temporal content while retaining predictive supervision in feature space. Setting the future offset to zero would instead yield a zero difference and is not the current-target intervention.

The respective six-task averages are 87.5% for Full TAD, 77.5% without the assistant, 68.3% without the teacher, 47.5% without tactile information, and 63.3% with current targets. These contrasts support the component-level conclusions in Section 4.4; they do not independently identify the effect of every loss term or readout. Matching policy-update budgets also does not equate total training cost when an upstream component is removed.

E.6 Generalization under environment variations

Following the categories in Dream-Tac (Lou et al., 2026), Section 4.5 evaluates height, position, object geometry, and background variations. The compared methods are $\pi_{0.5}$, N_0-VTLA, and TAD. Checkpoints and success criteria remain fixed, with twenty trials per condition and method and no test-time adaptation. The conditions below correspond to Figure 4.

Height. For *Unscrew Cap*, the support is raised or lowered by 5 cm while retaining the same bottle and cap. TAD achieves 80% success for each shift, compared with 85% in the standard setting.

Position. For *Wipe Whiteboard*, the wiping-region location varies between the *Standard* and *Random* settings. TAD achieves 90% success under position variation, compared with 95% in the standard setting.

Object geometry. For bottle uprighting, the *Seen* and *Unseen Object* settings compare the training bottle with an unseen bottle geometry. TAD achieves 90% success in both settings. This result concerns the evaluated replacement object rather than all unseen geometries.

Background. For *Unscrew Cap*, the *Standard* and *Changed* settings vary the visual background while retaining the bottle, cap tightness, illumination, and contact surfaces. TAD achieves 85% success in both settings. Both comparisons use the standard Unscrew Cap setting as a reference.

E.7 Interpreting the complementary evidence

Teacher prediction, component ablations, and environment variations address complementary questions. The teacher comparison establishes that observation-conditioned predictions outperform simple alternatives in the frozen tactile feature space, within the teacher-stage holdout scope specified above. The current-target ablation connects the temporal content of supervision to downstream task success.

The no-assistant and no-teacher controls support the contribution of these complete components without attributing their gains to individual internal paths.

The generalization experiments characterize performance under the reported physical variations. They do not measure sensor dropout, tactile delay, or inference overhead. Neither equality of two observed success rates nor a selected successful recording establishes invariance to a perturbation. Trial success rates describe the evaluated policies and conditions; they do not estimate variation across independently trained models. Future-feature error, paired action-field error, physical task success, and qualitative sensor observations should therefore be read as distinct, complementary evidence.

E.8 Assistant correction: protocol and results

Paired intervention. We retain or zero the assistant velocity residual in fixed HDMI policies at 2,000 and 5,000 updates, holding the remaining policy, observations, noisy actions, and target velocities fixed (Figure 6). At each checkpoint, twelve frames from each of ten training episodes, three denoising times ($\tau \in \{0.2, 0.5, 0.8\}$), and two fixed noise draws yield 720 paired conditions. Normalized flow-matching MSE averages 50 action positions and eight physical dimensions, excluding padding. The horizontal axis in Figure 6 is denoising time, not execution time.

Results and uncertainty. At 5,000 updates, retaining the residual lowers mean error from 0.00836 to 0.00269, a 67.9% relative reduction. The 95% paired episode-bootstrap interval is 62.4–73.0%, using 5,000 resamples of the ten complete episodes. The mean decrease is positive at every evaluated denoising time and in all ten episode averages. At 2,000 updates, the relative reduction is 58.8%. This diagnostic supports a useful correction to the fitted action field. It uses training demonstrations and retains jointly learned parameters, so it neither estimates closed-loop performance gains nor replaces a retrained no-assistant ablation.

E.9 Tactile attention and action sensitivity

Exploratory cohort. Figure 7 uses a fixed 2,000-update checkpoint on six predeclared HDMI-insertion seeds. This separate exploratory cohort yields 1/6 baseline successes, 69 observations, 23 probes, and 115 offline counterfactual predictions.

Attention extraction and observation selection. Panels (a,b) use seed 2000102, zero-based request 7, with 50 executed action slots. Among noninitial observations, this case is nearest the median reference-to-current tactile RGB change averaged across sensors; ties are resolved by seed, then request. Selection is independent of attention and outcome. Each sensor has one CLS and nine spatial tokens. Unsmoothed maps average final-denoising attention over heads and executed slots, use a common absolute scale, and report CLS weights separately. Independent softmaxes describe within-readout allocation, not relative branch contributions.

Spatial concentration. For panel (c), entropy is conditioned on each sensor's nine spatial tokens, divided by $\log 9$, and averaged over sensors, heads, and executed slots. Thin lines represent episodes; thick lines represent equal-episode means over common action-count support. The episode-averaged normalized-entropy difference is -0.139 for assistant predicted-tactile attention relative to the current-tactile readout.

Paired counterfactual intervention. For panel (d), at reached requests 1, 4, 7, and 10, one spatial key per sensor is removed from assistant direct action-residual cross-attention. CLS is retained and weights are renormalized; predictive F-tokens and current-tactile pathways remain intact. Key rankings average baseline attention over heads and all 50 planned slots at each denoising step. Three random masks match the removal count. Paired predictions share observations and initial noise. Command RMS changes cover seven revolute joints and executed slots. Random repeats are averaged within probes, then probes within episodes.

Results and interpretation. Connected circles show paired episode means; diamonds and bars show cohort means and descriptive 95% episode-bootstrap intervals (20,000 resamples; $n = 6$). Highest-attention and random key removal yield mean command changes of 0.00052319° and 0.00042621°,

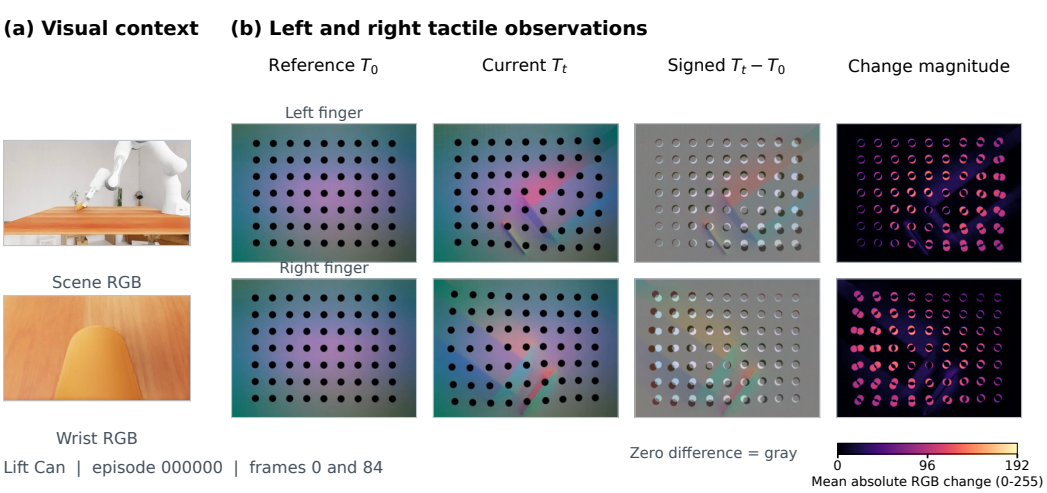

Figure 9: **Observed tactile change in a Lift Can demonstration.** Scene and wrist views accompany left (top) and right (bottom) tactile images at reference frame 0 and current frame 84. Signed RGB differences use gray for zero and a shared $[-192, 192]$ intensity range; magnitude is the channel mean of $|T_t - T_0|$. The native-resolution images show recorded observations, not future predictions.

respectively. The paired difference is 0.00009698° (95% interval [0.00007546, 0.00012571]°). The results associate structured spatial allocation with local command sensitivity. The absolute perturbations are small, and counterfactual actions were not executed; these measurements do not establish closed-loop performance gains.

F Qualitative observations and execution examples

This section complements the quantitative results with recorded visual and tactile observations and chronological success/failure examples. Selected recordings illustrate task execution; their number and outcome balance are not used to estimate the success rates in the main text.

F.1 Observed tactile change

Figure 9 uses a Lift Can demonstration, with episode-initial reference frame 0 and current frame 84. Scene and wrist RGB accompany left and right tactile observations. Signed RGB differences map zero to gray using a shared display range; the magnitude panel shows the channel-mean absolute difference. These panels visualize sensor appearance, not calibrated force or pressure.

The current input compares frame 84 with frame 0, whereas future supervision compares frame 94 with frame 84 when valid. Persistent deformation can therefore produce a large current-minus-reference difference with little subsequent change. Model inputs retain signed differences after the preprocessing in Appendix A; the magnitude panel is for visualization only. The teacher and assistant predict latent features rather than RGB tactile images.

F.2 Real-world successes and failures

The real-robot collection contains twelve autonomous executions of complete TAD: one successful and one unsuccessful recording for each of the six tasks in Table 2. Task identities, outcomes, and complete-model attribution were checked for the supplied recordings. The paired selection illustrates both completed and unsuccessful interactions and is separate from the quantitative evaluation denominator. *Upright Bottle* and *Adjust Bottle* refer to the same task; *Lift Tube* uses a clamp held by the robot gripper.

Views and temporal ordering. Each observation group in Figure 10 contains scene and wrist RGB and left/right tactile images at the same stored frame index. RGB crops remain fixed within a task

Table 9: **Frame selection for the real-robot examples.** Zero-based recording and frame indices in chronological display order. S and F denote the annotated outcome of the complete recording.

Task	Recording	Outcome	Selected frames
Insert Plug	0	S	350, 640, 800
	1	F	600, 800, 1020
Unscrew Cap	2	S	300, 555, 750
	3	F	640, 870, 885
Lift Tube	4	S	465, 635, 1060
	5	F	340, 555, 595
Upright Bottle	6	S	385, 405, 465
	7	F	610, 940, 1040
Wipe Whiteboard	8	S	310, 545, 935
	9	F	370, 495, 570
Pour Water	10	S	425, 970, 1350
	11	F	380, 565, 790

and camera. Figure 11 uses the wrist view for plug insertion and whiteboard wiping and the scene view for the other tasks. Each row presents three observations in chronological order. Frames were selected to show visibly changing interaction stages; they need not be equally spaced or coincide with termination. Table 9 records the selected indices.

Interpretation. Successful recordings show the required task outcome; unsuccessful recordings show attempts that do not complete it. The displayed tactile channels retain their recorded difference-image appearance without per-panel contrast enhancement. Their recording metadata do not establish the acquisition-time reference and scale, so they are not equated numerically with the simulation input differences. These observations can show contact-related visual changes but do not by themselves identify a failure's cause or measure prediction accuracy. In particular, a visually changing tactile image is neither a ground-truth contact annotation nor a force measurement.

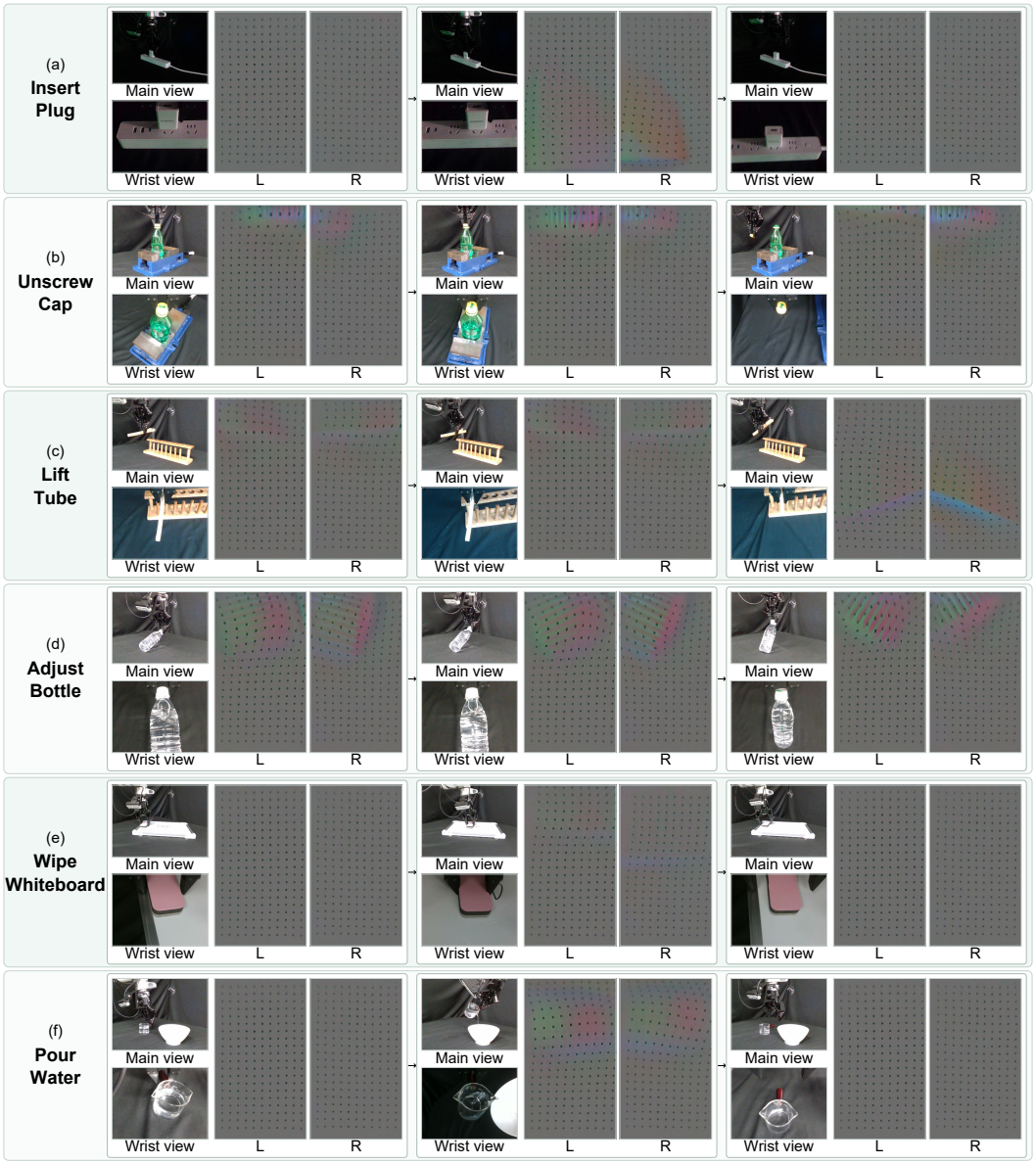

Figure 10: **Successful autonomous executions of complete TAD.** Each row shows one of the six physical tasks and three chronological observation groups. Each group contains scene and wrist RGB on the left and synchronized left (L) and right (R) tactile observations on the right. The tactile panels display recorded sensor observations, not predicted images.

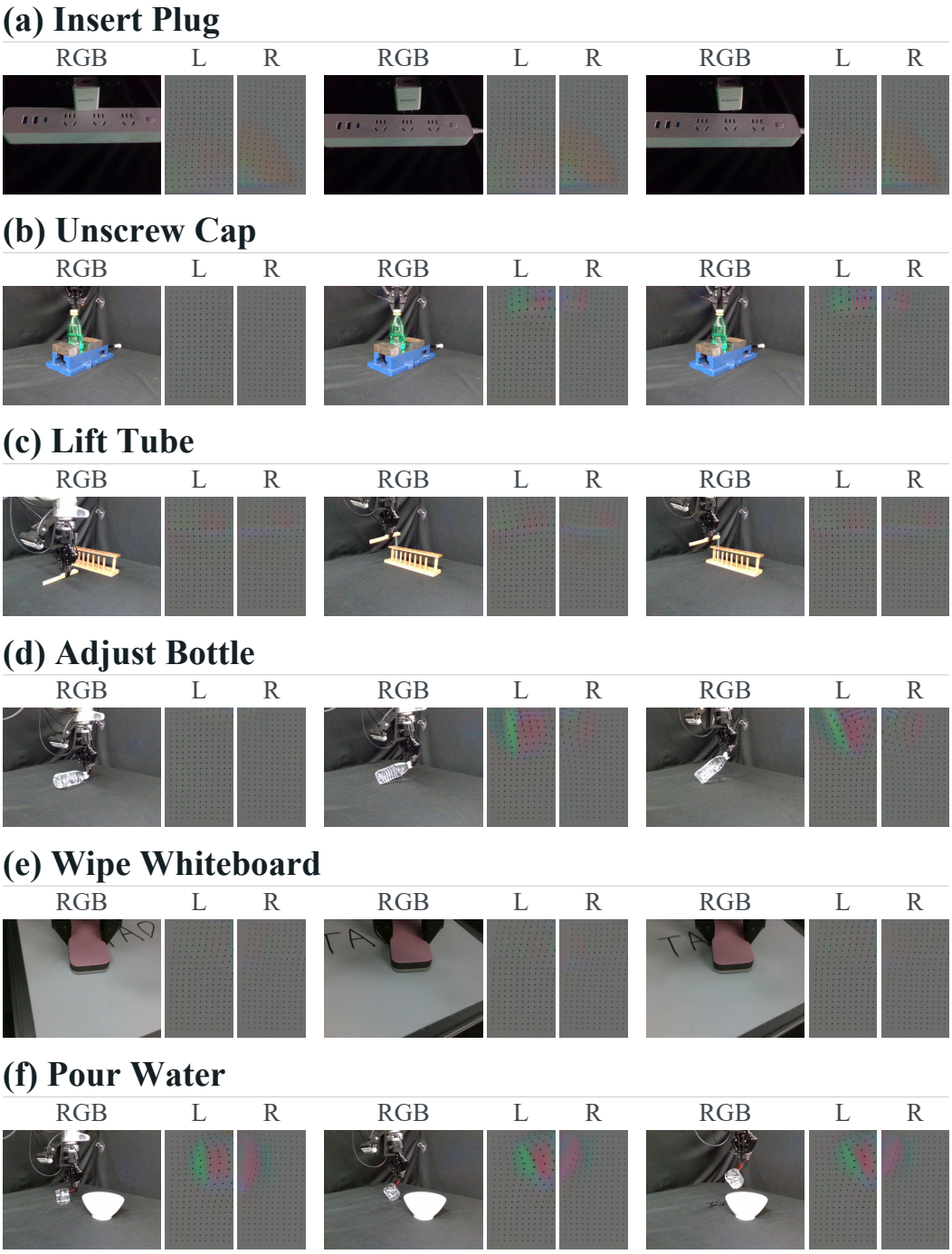

Figure 11: **Unsuccessful autonomous executions of complete TAD.** One unsuccessful recording for each task, in the same order as Figure 10. Each snapshot contains a selected RGB view and left/right tactile observations. The three snapshots show interaction stages, not necessarily the terminal state, and do not attribute failure to an unmeasured internal mechanism.

F.3 SIMULATION SUCCESSES AND FAILURES

Figures 12 and 13 show preflight recordings with model and checkpoint labels retained. The successful HDMI clip uses predictive-only R2; the unsuccessful HDMI clip uses complete TAD. Other clips use the full-reconstruction reference. They are not paired final-benchmark trials. Each row contains four chronological overview-camera frames; timestamps denote encoded video time. Outcomes follow the recorded native success flag. The failure examples show a can tipping over, a bottle dropping, and an incomplete insertion, without identifying an internal cause.

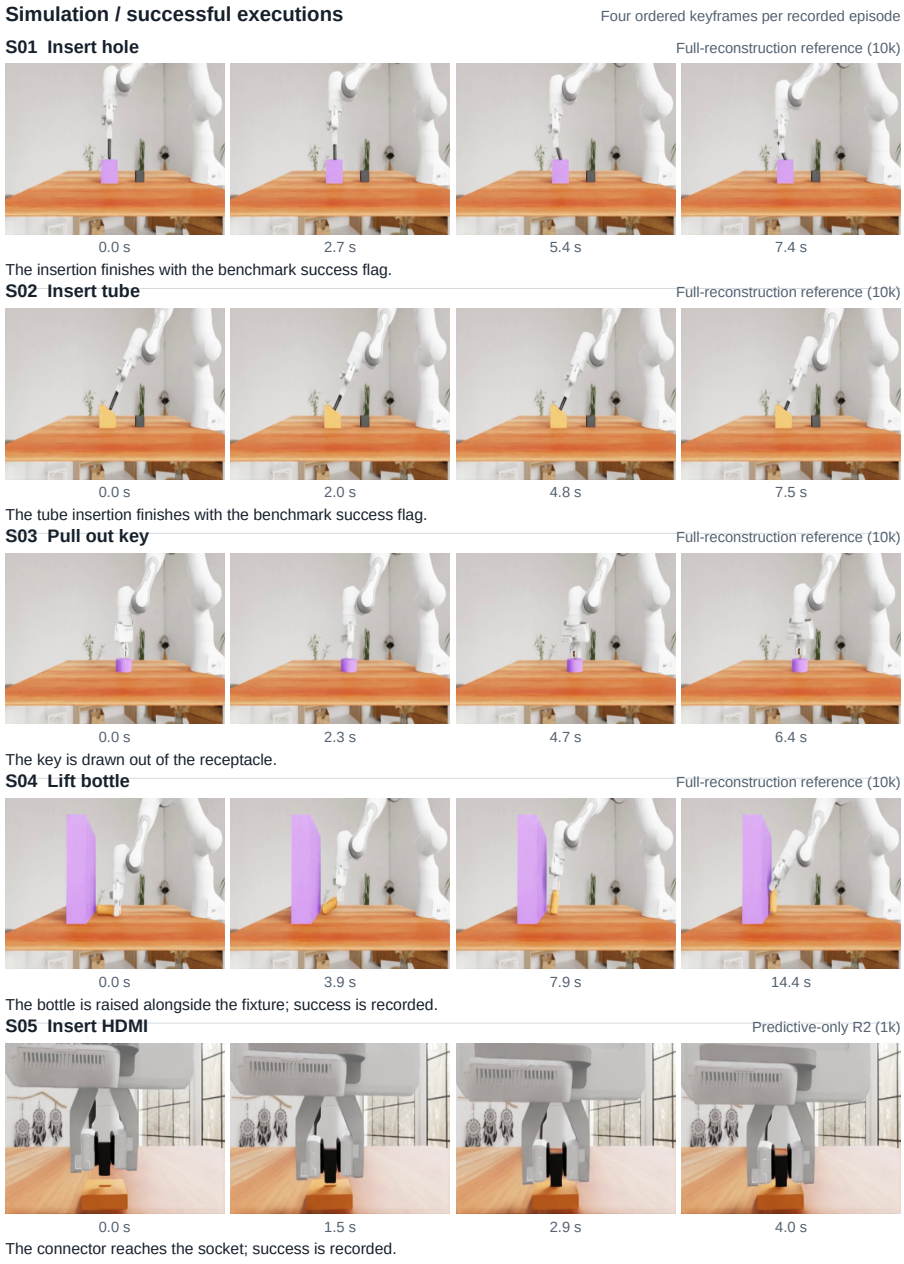

Figure 12: **Successful simulation examples.** Four ordered overview-camera frames per recording. The model and checkpoint are identified in each row: the full-reconstruction reference for four tasks and predictive-only R2 for HDMI insertion. These preflight examples are excluded from the final benchmark scores. Scene content is taken directly from the recordings.

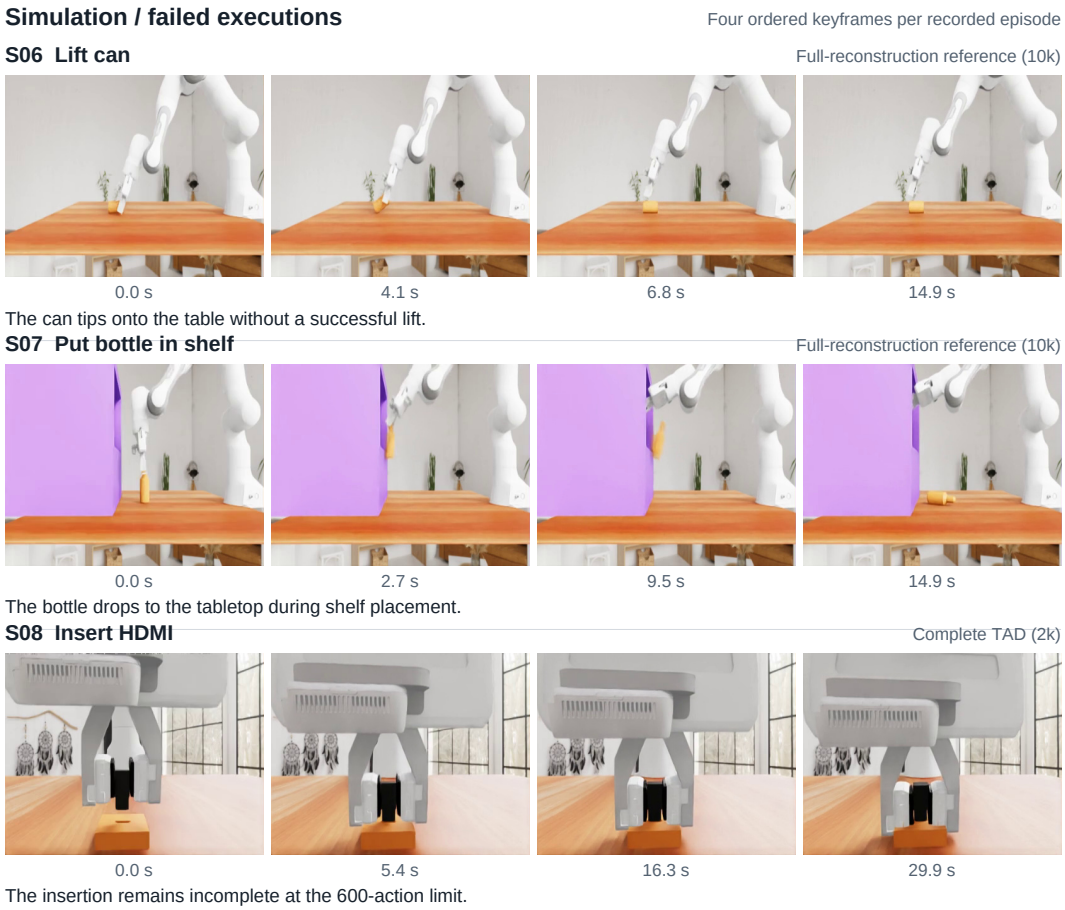

Figure 13: **Unsuccessful simulation examples.** The full-reconstruction reference fails to lift the can and place the bottle, while the displayed complete-TAD HDMI attempt does not finish insertion. Original model and checkpoint labels are retained. These selected recordings illustrate task-level failure behavior and do not add to or replace the quantitative results in the main text.

G DISCUSSION AND LIMITATIONS

The simulation and real-world results assess the complete TAD framework, while matched real-world ablations evaluate the contributions of the assistant, teacher-derived supervision, tactile information, and future-oriented prediction. The ablations do not separately isolate the effects of EMA targets, teacher anchoring, and residual action correction. Simulation baseline scores are taken from the cited publications. The real-world comparisons and environment-variation tests cover six tasks and the specified test conditions; they do not establish robustness beyond these settings. The teacher predicts tactile change under demonstration behavior without counterfactual action conditioning; its holdout is specific to teacher learning because inherited representations saw these demonstrations. Deployment retains the assistant and frozen context, requiring latency and memory measurements.