UTILITY-GUIDED WORLD FEEDBACK FOR VISUOTACTILE MANIPULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

World–action models jointly predict future observations and actions, but predictive accuracy alone does not determine which visual and tactile forecasts are useful for control. This gap is particularly important in contact-rich manipulation, where the utility of predicted evidence varies across sensors and time. We propose Utility-Guided World Feedback (UWF), a framework that learns to assess and selectively incorporate predicted evidence into action generation. UWF estimates the marginal utility of sensor–time evidence groups against a shared supervised action objective, converts utility-weighted evidence into action-chunk corrections, and regulates their adoption through sequence-level risk estimates. A shared action reference supports consistent evidence assessment, while joint world–action denoising updates feedback as predictions evolve. The framework learns from demonstrations without additional reward annotations or simulator rollouts for utility supervision. Across eight UniVTAC simulation tasks, UWF achieves a mean success rate of 90.1%. On eight real-robot tasks evaluated against four baselines with 20 trials per task and method, UWF achieves 91.3% mean success, exceeding the strongest baseline by 11.9 percentage points and obtaining six task wins and two ties. These results support utility-guided selection of predicted visuotactile evidence as an effective approach to contact-rich manipulation.

1 INTRODUCTION

A plausible forecast is not necessarily a useful guide to action. Action-chunk policies and generative controllers learn coherent behavior from demonstrations (Zhao et al., 2023; Chi et al., 2025; Intelligence et al., 2025); world–action models additionally couple behavior to predicted scene dynamics (Li et al., 2026; Ye et al., 2026a). For contact-rich manipulation, tactile representations and tactile predictions expose information that vision alone may miss (George et al., 2025; Chen et al., 2026; NeoteAI Team & Fudan TEAI Team, 2026b). Dream-Tac and N_0-TWAM bring tactile prediction into world–action modeling (Lou et al., 2026; NeoteAI Team & Fudan TEAI Team, 2026a), but joint prediction alone does not establish which forecast features improve an action or when their influence should be withheld.

Using world predictions effectively requires determining which predicted visual and tactile changes can improve actions, and how strongly they should influence corrections to an action sequence. Fast-WAM separates training-time video modeling from test-time imagination (Yuan et al., 2026b), while reactive tactile policies and TacForcing refresh feedback during execution (Xue et al., 2025; Zhou et al., 2026). We further examine how these predictions should influence action generation. Prediction error alone cannot capture the distinct requirements of predictability, utility and reliability.

Universal Planning Networks learn through imitation-supervised differentiable planning (Srinivas et al., 2018); TaskSense focuses world modeling through spatial attention and inverse-dynamics supervision (Islam & Huber, 2026); MOPO penalizes uncertain model rollouts in offline reinforcement learning (Yu et al., 2020). We assess predicted information by asking whether it reduces action-prediction error relative to demonstration targets. Building on marginal-contribution attribution (Shapley et al., 1953; Lundberg & Lee, 2017), we measure how predictions from the overhead and wrist cameras and two tactile sensors at different future time steps contribute to reducing action-prediction error. Quantile regression (Koenker & Bassett Jr, 1978) complements this assessment by estimating how much improvement a correction offers toward the lower end of its predicted ben-

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.

efit distribution. These estimates guide both which information to emphasize and how strongly to modify the action. Appendix G provides a fuller comparison.

Utility-Guided World Feedback (UWF) represents predicted visual and tactile changes relative to current observations as *world evidence*. A shared action reference enables consistent assessment of evidence groups, whose estimated utility guides corrections to the complete action sequence. A learned gate scales corrections by their predicted benefit across the sequence, limiting uncertain feedback. Rejected corrections remain supervised, allowing the feedback pathway to keep learning. Joint denoising updates feedback as world and action predictions evolve. All utility and risk targets are derived from existing demonstrations, without additional reward annotations or simulator rollouts for supervision.

Our contributions are threefold. First, we propose a selective world-feedback architecture that converts predicted visuotactile changes into action-sequence corrections under a shared action reference. Second, we develop demonstration-based marginal-utility supervision to learn the relative contribution of each sensor–time evidence group to action improvement. Third, sequence-level feedback adoption scales the correction added to the base action prediction according to estimated improvement. During training, demonstration-based supervision also optimizes the fully corrected prediction, allowing corrections to learn even when excluded from the final output. Experiments support this design: UWF achieves 90.1% mean simulation success and 91.3% real-robot success, exceeding the strongest real-robot baseline by 11.9 percentage points. Across three fixed training windows, visual/tactile latent MSE decreases by 42.1%/49.7% relative to N_0-TWAM, with higher mean spatial change-rank correlation and gradient alignment in both modalities.

2 Related Work

World–action models and tactile feedback. Action-chunk and generative policies learn behavior from demonstrations (Zhao et al., 2023; Chi et al., 2025; Intelligence et al., 2025); world–action models additionally predict visual dynamics (Li et al., 2026; Ye et al., 2026a; Yuan et al., 2026b). Xiaomi-Robotics-1 scales real-world VLA pretraining (Xiaomi Robotics Team et al., 2026); G0.5 unifies reasoning and action tokens (Liu et al., 2026). OpenWAM studies how architecture, information flow and pretraining shape world–action synergy (Wang et al., 2026). Dream-Tac and N_0-TWAM extend this prediction to tactile sensing (Lou et al., 2026; NeoteAI Team & Fudan TEAI Team, 2026a), while reactive tactile policies update actions from execution-time observations (Xue et al., 2025; Zhou et al., 2026). We focus on the complementary decision of which forecast evidence should modify an action chunk, and by how much.

Evidence utility and acceptance. TaskSense emphasizes task-relevant world prediction (Islam & Huber, 2026); MOPO penalizes uncertain model rollouts (Yu et al., 2020). We instead measure evidence utility against a shared supervised action target. Sampled marginal contributions and quantile regression (Shapley et al., 1953; Koenker & Bassett Jr, 1978) connect this measurement to an evidence-dependent residual and its adoption scale. Detailed comparisons with representation learning, tactile policies and attribution methods appear in Appendix G.

3 Method

3.1 Forecasting under a shared action reference

Following Figure 1(b), we adapt a pretrained visuotactile world–action model using task-specific demonstrations (NeoteAI Team & Fudan TEAI Team, 2026a). The backbone predicts future sensor observations and a base action flow velocity; a feedback head then corrects that velocity. Let h contain observed images, tactile inputs and available robot state, and D the predicted visual/tactile changes. We explicitly correct the base velocity v_0:

$$v_\lambda = v_0 + \lambda r, \qquad r = F_\theta(D; h, A_\sigma, \sigma), \qquad 0 \le \lambda \le 1. \tag{1}$$

Here A is a normalized demonstration sequence, ϵ Gaussian noise, and $\sigma \in [0, 1]$ the action noise level. The noisy input is $A_\sigma = (1-\sigma)A + \sigma\epsilon$, with flow target $u = \epsilon - A$ (Lipman et al., 2022). The trainable F_θ outputs a same-shaped correction r; λ is chunk-wide. Integrating v_λ produces actions;

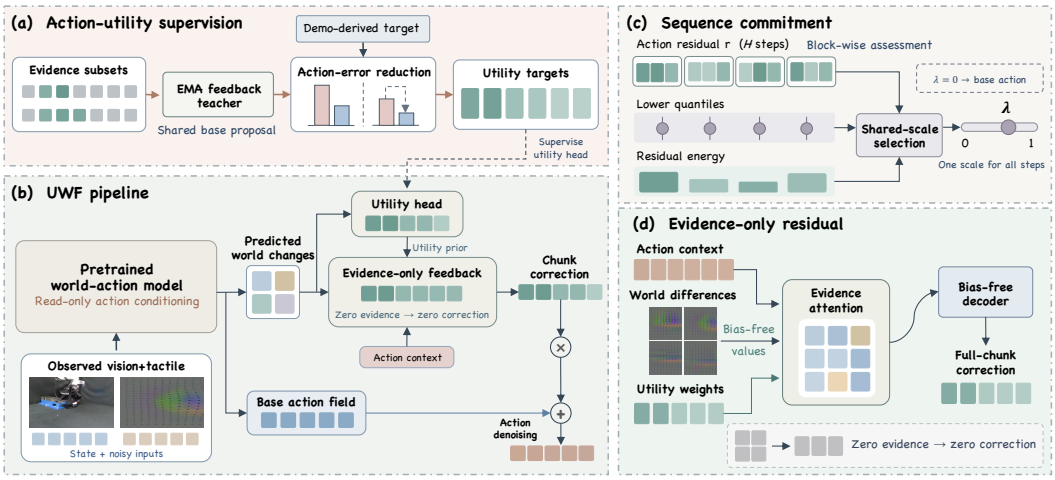

Figure 1: **UWF overview.** (a) EMA-teacher subset evaluations supervise evidence utility. (b) The jointly fine-tuned backbone uses read-only action conditioning; utility-weighted feedback modifies the base denoising velocity as $v_0 + \lambda r$. (c) Blockwise lower-quantile benefits and residual energies determine one chunk-wide scale. (d) Bias-free evidence values and decoding ensure zero correction for zero evidence. Task images illustrate real-robot visual inputs and tactile differences.

r is not a pose offset. We jointly train v_0. We use $\mathsf{D}_a(v, u)$ for the scheduler-weighted, frame-balanced action MSE and $\ell(v, u)$ for unweighted valid-coordinate MSE; their distinct reductions are specified in Appendix A.1.

Read-only world conditioning. At each layer, we reuse a copy of the input noisy-action embeddings rather than updated policy features, keeping the forecast's action reference fixed. This "read-only" rule keeps the forward-pass reference fixed while allowing world-prediction losses to train the conditioning embeddings and projections. Clean world context is restricted to observed history, and action queries cannot directly read predicted-world tokens in the shared backbone. Predicted-world feedback therefore reaches actions through F_θ; observed tactile input remains history-conditioned.

For each sensor and future frame, predicted velocity v^m converts noisy latent z_σ to $\widehat{z} = z_\sigma - \sigma v^m$, using that stream's noise level. We subtract the same sensor's preceding-context latent and pool the difference to a 2×2 grid: four tokens per group. Two future frames from the overhead/wrist cameras and two tactile sensors give eight groups (32 tokens). The world objective $\mathcal{L}_W$ regresses visual and tactile flow velocities against their demonstration-derived targets, weighting each valid group by the action-utility weights introduced next. Visual errors additionally retain the fixed scheduler weighting; tactile errors do not.

3.2 From evidence utility to action correction

An exponential-moving-average copy of the feedback head evaluates subsets of the *same forecast* (Figure 1(a)). A coalition S is simply a subset of these groups. Let $\bar{r}(S)$ be the teacher residual using only S, and define the supervised proposal error $R(S) = \ell(v_0 + \bar{r}(S), u)$. For a random permutation π, let S_k contain its first k groups, with $S_0 = \emptyset$. The marginal contribution of the newly admitted group is

$$c_{\pi_k} = R(S_{k-1}) - R(S_k). \tag{2}$$

A positive c_{π_k} means that adding group π_k to S_{k-1} reduces action-target error. For eight groups, the empty subset plus eight additions require nine teacher-head evaluations, all sharing v_0, u and the backbone forecast. This is sampled coalitional attribution (Shapley et al., 1953), not physical causal identification or exact permutation averaging.

The utility head maps detached context/evidence features to s_g by Smooth-L1 regression to normalized labels $\widetilde{c}_g = c_g / \max\{\ell(v_0, u), \epsilon_c\}$. We denote its valid-group mean loss by $\mathcal{L}_C$. For G available groups, $w_g = \frac{1}{2} + \frac{G}{2}\,\mathrm{softmax}(s)_g$: weights average to one and stay at least $1/2$. Each

Table 1: Action-velocity proposals and their supervision roles. All four proposals share the same sampled target $u = \epsilon - A$; sg denotes stop-gradient.

p	Action velocity v_p	Supervision role
base	v_0	Train the fallback prediction
commit	$v_0 + \mathrm{sg}[\lambda]r$	Train the output used by the sampler
full	$\mathrm{sg}[v_0] + r$	Learn feedback even when $\lambda = 0$
subset	$\mathrm{sg}[v_0] + F_\theta(D; S)$	Learn from a nonempty proper subset

weight scales its group's world loss, and $\log w_g$ is added to its feedback-attention logits. Teacher probes omit the attention prior. Detached labels and weights ensure that the utility head learns from measured contributions, not from reducing weights on hard predictions.

Evidence-only correction. The feedback head (Figure 1(d)) queries evidence using noisy actions, the base clean-action estimate and context. The base clean-action estimate is $A_\sigma - \sigma v_0$. Attention values are bias-free projections of world differences; the bounded residual decoder is also bias-free. This makes the evidence indispensable to the correction:

$$F_\theta(0; h, A_\sigma, \sigma) = 0. \tag{3}$$

An empty coalition has the same null output by explicit masking. Context controls attention, but only evidence supplies values. The zero-initialized decoder corrects every action position.

Besides supervising v_0 and the committed output v_λ, we also train the fully corrected proposal $v_{\text{full}} = \mathrm{sg}[v_0] + r$ and a partial-evidence proposal $v_S = \mathrm{sg}[v_0] + F_\theta(D; S)$ against the same target u. Here sg keeps the forward value but blocks its gradient. S contains some but not all available groups; $F_\theta(D; S)$ abbreviates the same head with only S visible and no utility prior. These two proposal losses are never multiplied by λ, so a closed gate does not eliminate their learning signal. In these auxiliary predictions, $\mathrm{sg}[v_0]$ blocks direct updates to the additive base; gradients still pass through evidence and queries.

3.3 Sequence-level commitment

Figure 1(c) selects one scale for the combined correction. Divide the action sequence into four contiguous blocks. For each block, define energy $E_b = \langle r_b, r_b \rangle$ and alignment $B_b = -2\langle v_{0,b} - u_b, r_b \rangle$, with averages over valid coordinates. The change in action error is exactly

$$\ell_b(v_\lambda, u) - \ell_b(v_0, u) = \lambda^2 E_b - \lambda B_b. \tag{4}$$

Here E_b measures correction magnitude; positive B_b favors error reduction. Only E_b is available without the demonstration target. A risk head predicts the lower tenth percentile of normalized B_b, providing a conservative improvement estimate; labels use the current feedback head's residual. Its pinball-regression loss (Koenker & Bassett Jr, 1978) is $\mathcal{L}_Q$, with detached features and targets and valid-coordinate weighting. Let q_b denote its prediction restored to the scale of B_b, and set $\widehat{\Delta}_b(\lambda) = \lambda^2 E_b - \lambda q_b$. We minimize aggregate predicted error with one chunk-wide coefficient, subject to nonincreasing predicted error in every included block:

$$\lambda^\star = \operatorname*{arg\,min}_{0 \le \lambda \le 1} \sum_{b \in \mathcal{B}} n_b \widehat{\Delta}_b(\lambda) \quad \text{s.t.} \quad \widehat{\Delta}_b(\lambda) \le 0, \quad \forall b \in \mathcal{B}. \tag{5}$$

Here $\mathcal{B}$ contains blocks with valid coordinates and sufficient residual energy; n_b counts valid coordinates. One λ scales all blocks: zero selects v_0, and one uses the full correction. The solution caps the aggregate optimum by the scale allowed in every block; a block with $q_b \le 0$ forces $\lambda = 0$. If $\mathcal{B}$ is empty, we also use zero (Appendix A.5). The constraint is on *predicted surrogate error*, not physical safety: learned quantiles are not certified joint lower bounds on the true alignments.

3.4 Joint optimization and inference

For each training sample, all four proposals use the same sampled noise and target $u = \epsilon - A$. Table 1 lists their roles; sg denotes stop-gradient.

Table 2: Eight-task UniVTAC success rates (%). † marks non-language action policies, included alongside VLAs but not claimed to be VLAs. Cross-source protocols are not verified as identical. Bold marks column maxima, including ties.

Method	Insert HDMI	Insert Hole	Insert Tube	Grasp Classify	Lift Can	Lift Bottle	Pull-out Key	Put Bottle in Shelf	Avg.-8
Vision-Language-Action Models									
ACT (vision only)† (Zhao et al., 2023)	15	19	45	50	20	42	28	28	30.9
ACT + UniVTAC† (Chen et al., 2026)	28	24	56	99	29	71	46	31	48.0
VITaL† (George et al., 2025)	6	25	34	**100**	8	72	47	32	40.5
$\pi_{0.5}$ (Intelligence et al., 2025)	8	25	74	49	6	**100**	35	34	41.4
StarVLA-α (Ye et al., 2026b)	24	52	69	68	65	32	51	**88**	56.1
InternVLA-A1 (Cai et al., 2026a)	12	93	95	86	90	58	66	37	67.1
Xiaomi-Robotics-0 (Cai et al., 2026b)	69	96	98	45	13	21	80	12	54.3
Xiaomi-Robotics-1 (Xiaomi Robotics Team et al., 2026)	62	90	95	73	40	64	86	38	68.5
GalaxeaVLA (G0.5) (Liu et al., 2026)	59	92	94	77	51	72	92	46	72.9
N_0-VTLA (NeoteAI Team & Fudan TEAI Team, 2026b)	25	95	**99**	**100**	88	99	**99**	60	83.1
World-Action Models									
GigaWorld-Policy (Ye et al., 2026a)	0	12	9	20	0	38	32	21	16.5
LingBot-VA (Li et al., 2026)	38	42	96	17	0	0	58	0	31.4
FastWAM (Yuan et al., 2026b)	19	66	98	72	0	21	73	35	48.0
OpenWAM-α (Wang et al., 2026)	55	89	93	82	69	37	83	76	73.0
N_0-TWAM (NeoteAI Team & Fudan TEAI Team, 2026a)	68	**99**	98	94	**93**	58	79	87	84.5
UWF (Ours)	**90**	96	95	99	88	67	98	**88**	**90.1**

Here S is sampled once per chunk and its residual uses no contribution prior. For each proposal, we first average squared velocity errors over valid action coordinates within a frame, multiply by the fixed action-scheduler weight, and then average valid frames:

$$\mathsf{D}_a(v_p, u) = \frac{1}{|\mathcal{F}_a|} \sum_{f \in \mathcal{F}_a} \omega_a(\sigma_f) \frac{\sum_{i,d} M_{fid}(v_{p,fid} - u_{fid})^2}{\sum_{i,d} M_{fid}}. \tag{6}$$

$\mathcal{F}_a$ contains valid sample–frame pairs, i, d index action slots and coordinates, and M is one on valid coordinates and zero on padding/inactive coordinates. The context action frame is excluded. Each branch therefore contributes one scalar loss for the same demonstration target.

The world loss similarly averages velocity errors, not decoded-image errors. Let $e_{mf} = \mathrm{mean}_{c,x,y}(v_f^m - u_f^m)^2$, where z^m is the encoded demonstration observation, ϵ^m is Gaussian noise and $u^m = \epsilon^m - z^m$. Here w_{mf} is the detached w_g for sensor m and frame f. For the two cameras $\mathcal{M}_v$ and two tactile streams $\mathcal{M}_t$,

$$\mathcal{L}_W = \frac{1}{2|\mathcal{F}_w|} \sum_{f \in \mathcal{F}_w} \left[\omega_v(\sigma_f) \sum_{m \in \mathcal{M}_v} \mathrm{sg}[w_{mf}] e_{mf} + \sum_{m \in \mathcal{M}_t} \mathrm{sg}[w_{mf}] e_{mf} \right]. \tag{7}$$

$\mathcal{F}_w$ includes all valid world frames, with unit contribution weight on context; absent tactile predictions contribute zero. Only visual error receives the fixed scheduler weight. The factor two averages the two sensors within each modality; the visual and tactile losses are then added.

The remaining terms supervise the assessment heads. For chunk j and available evidence group g, the detached EMA-chain label is $\widetilde{c}_{jg} = c_{jg} / \max\{\ell(v_0, u), \epsilon_c\}$. For temporal block b, the detached current-student label is $\widetilde{B}_{jb} = B_{jb} / (\sqrt{E_{jb}} + \epsilon_r)$. Their losses are

$$\mathcal{L}_C = \frac{\sum_{j,g} \chi_{jg} H_\kappa(s_{jg} - \mathrm{sg}[\widetilde{c}_{jg}])}{\sum_{j,g} \chi_{jg}}, \qquad \mathcal{L}_Q = \frac{\sum_{j,b} n_{jb} \nu_{jb} \rho_\alpha(\mathrm{sg}[\widetilde{B}_{jb}] - \widehat{b}_{jb})}{\sum_{j,b} n_{jb} \nu_{jb}}. \tag{8}$$

Binary masks χ, ν select valid groups/blocks; ν also excludes low residual energy, and n counts valid coordinates. $H_\kappa(x)$ is $x^2/(2\kappa)$ for $|x| < \kappa$ and $|x| - \kappa/2$ otherwise; $\rho_\alpha(x) = \max\{\alpha x, (\alpha - 1)x\}$. The risk output $\widehat{b}$ gives $q = (\sqrt{E} + \epsilon_r)\widehat{b}$ for computing λ. Empty reductions give zero.

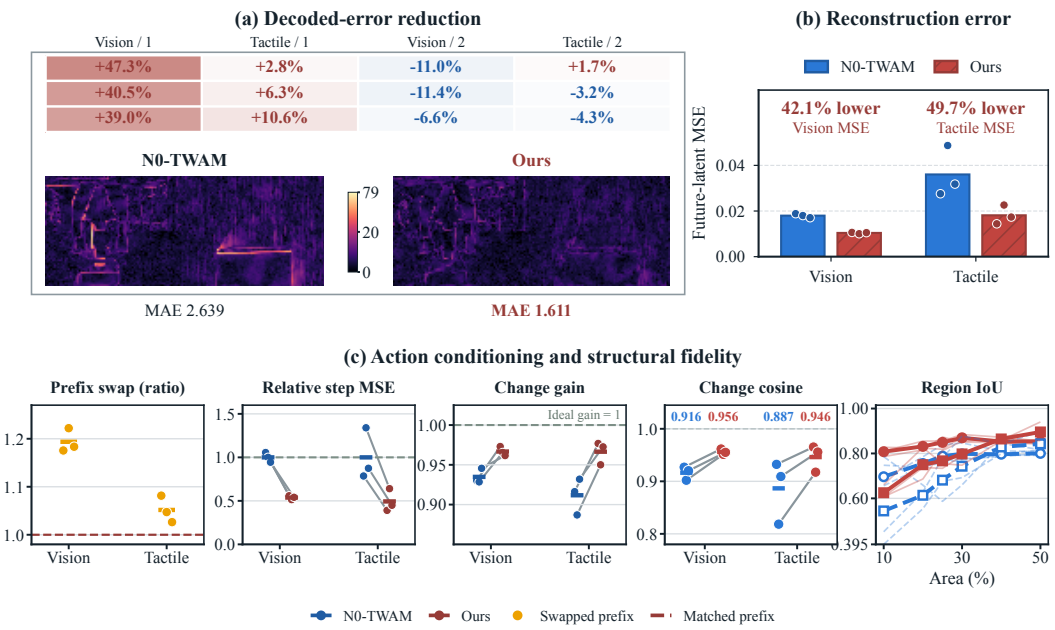

Figure 2: **Reconstruction gains, conditioning specificity and change fidelity.** (a) Top: decoded-crop MAE reductions relative to N_0-TWAM for fixed training windows 25/50/75 (rows), by modality and future slot (decoded indices 4/8); positive values indicate improvement. Bottom: window 75, first-slot channel-mean absolute-error maps, sharing a square-root color scale (0–79); labels give crop MAE (0–255). (b) Two-future-slot latent MSE: means and all three windows, with hatched bars for ours. (c) Left to right: prefix-swap error, future-slot difference error, change-projection gain, future-step change cosine similarity, and high-change region IoU across area fractions. All three windows are retained; IoU/cosine axes start at 0.395/0.78. Dashed and solid curves in the IoU panel denote baseline and Ours; circles and squares denote vision and tactile sensing. For this example, UWF reduces visual/tactile MAE by 39.0%/10.6% relative to N_0-TWAM.

With $\mathcal{P} = \{\text{base}, \text{commit}, \text{full}, \text{subset}\}$, the final scalar objective is

$$\mathcal{L} = \mathcal{L}_W + \sum_{p \in \mathcal{P}} \omega_p \mathsf{D}_a(v_p, u) + \eta_C \mathcal{L}_C + \eta_Q \mathcal{L}_Q. \tag{9}$$

The fixed η_C, η_Q scale utility/risk losses; ω_p scale proposal losses, unlike scheduler weights, evidence weights or λ (Appendix B). All proposals share targets and masks. With λ detached, a closed gate blocks the committed branch's direct gradient to r, but full/subset losses still train corrections through evidence and query paths. Detached assessment features and targets restrict $\mathcal{L}_C, \mathcal{L}_Q$ to their own heads. After each optimizer step, $\bar{\theta} \leftarrow 0.99\bar{\theta} + 0.01\theta$ updates the teacher feedback head from the student head; labels use its pre-update parameters.

At inference, actions start from noise. Three world steps and four action steps form six joint events. At each scheduled action step, $A_{\sigma'} = A_\sigma + (\sigma' - \sigma)v_\lambda$, with $\sigma' < \sigma$. Events recompute evidence, corrections and gates; only scheduled streams advance, with observed context fixed. Subsequent chunks use newly observed history, not imagined outcomes. No demonstration targets or coalition probes are needed; Appendix A.7 details both procedures.

4 Experiments

4.1 UniVTAC simulation

Table 2 summarizes published reference results on the eight UniVTAC tasks. The table groups VLA and WAM baselines, with 12 published score rows and three additionally evaluated methods. Published scores are not our matched reruns. Means weight all eight tasks equally. UWF achieves

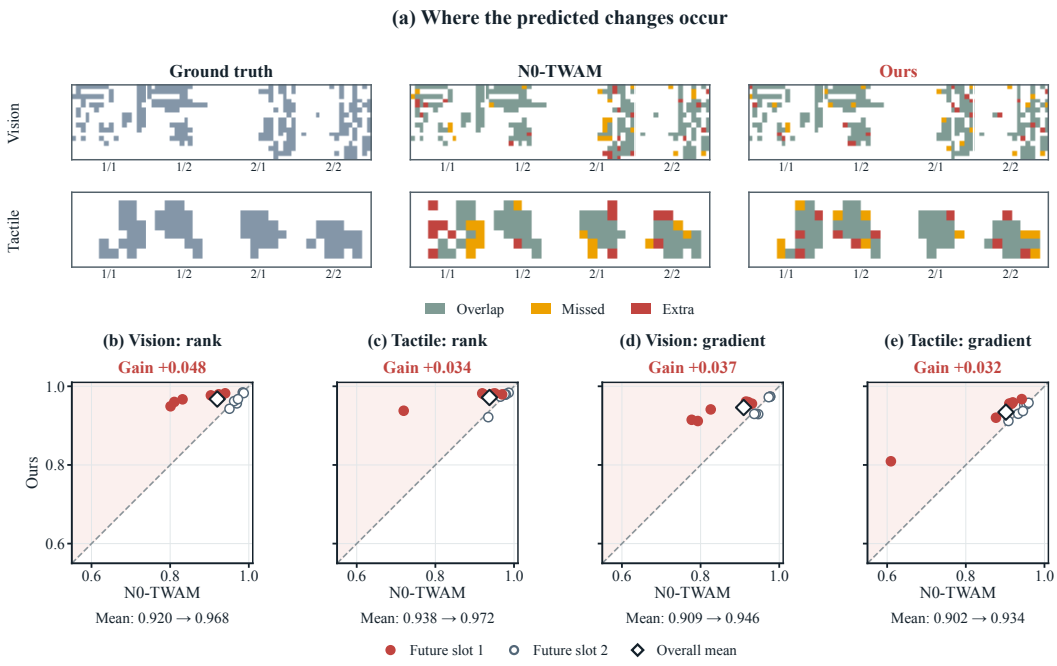

Figure 3: **Spatial localization and paired structural quality.** (a) Window 50: all sensor/future-slot high-change regions at 25% area; gold/red indicate missed/extra locations. (b,c) Visual/tactile change-rank correlations; (d,e) visual/tactile gradient cosines. Scatter axes compare baseline (horizontal) with Ours (vertical), with identical limits 0.55–1.01. Filled/open circles denote future slots 1/2; diamonds are overall means. Each scatter panel contains all 12 sensor/slot fields across three windows; points above the diagonal favor Ours. Labels report absolute mean gains and baseline-to-Ours means. These are training-window diagnostics.

an eight-task mean success rate of **90.1%**, with task-level rates ranging from 67% to 99%. These are simulation task-success results, distinct from the reconstruction diagnostics below.

Relative to N_0-TWAM, the strongest baseline by eight-task mean, UWF improves five tasks and raises the mean by 5.6 percentage points. Gains concentrate on Insert HDMI (68% to 90%) and Pull-out Key (79% to 98%), while Insert Hole, Insert Tube and Lift Can decrease by 3, 3 and 5 points. The aggregate improvement therefore reflects substantial gains on particular tasks rather than uniform gains across all tasks. Lift Bottle improves from 58% to 67% but remains UWF's lowest-scoring task. Appendix C additionally reports FTP-1, FTP-$\pi_{0.5}$, Tactile-VLA, RDP and TacForcing on their six-task protocols, with source-specific comparator rows. Six-task means are not inserted into this eight-task table. Illustrative task configurations are provided in Appendix D.

4.2 Diagnosing the reconstruction gains

On three fixed Insert HDMI training windows (indices 25, 50 and 75), our frozen 1000-step checkpoint reduces mean future-latent reconstruction MSE by **42.1% in vision** and **49.7% in tactile sensing** relative to post-trained N_0-TWAM. Figure 2 asks what changes accompany this improvement. Inputs, noise and targets match by hash; all three windows are included. The diagnostic uses teacher forcing on these three training windows. Shared latent-feature projections and full-dimensional error maps appear in Figure 4(a,b), with definitions in Appendix F.5.

Which structure is better preserved? In the quarter of latent locations with the largest GT change, error falls by 51.1% (vision) and 54.9% (tactile), compared with 38.4% and 47.8% elsewhere. Error in the difference between the two future slots decreases by 46.2% and 50.7%. Mean change-projection gains move from 0.935 to 0.967 in vision and from 0.912 to 0.966 in tactile sensing, toward the ideal value of one. Larger gains in high-change regions, together with projection

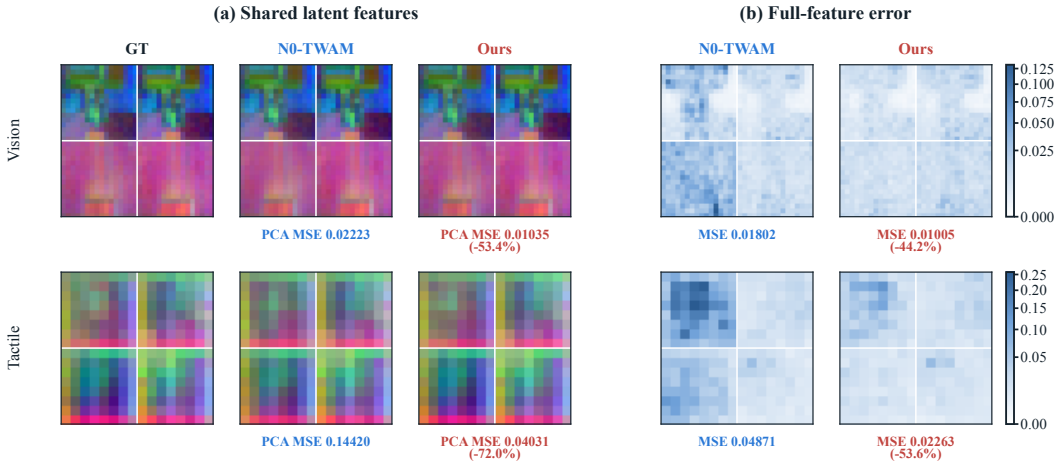

Figure 4: **Latent-feature fidelity.** Window 50, both sensors (tile rows) and future slots (tile columns). (a) Three-component PCA fitted to GT latents across three fixed windows, explaining 53.0%/49.4% of visual/tactile variance; projected MSE is measured before color rescaling. (b) Full 48-channel squared-error maps, sharing one square-root scale per modality. Labels show mean errors and baseline-relative reductions. Projected and full-dimensional errors measure different feature spaces. UWF reduces full-feature visual/tactile MSE by 44.2%/53.6%.

Table 3: Eight-task real-robot success rates. Task cells report successes out of 20 trials per method; Avg. is the mean task success rate (%). Tasks follow Figure 5(a–h). Bold marks the highest value in each column, including ties.

Method	Pick & place	Remove cap	Turn vise	Transfer pouch	Close box	Fill basket	Reinsert dropper	Erase board	Avg. (%)
OpenWAM-α (Wang et al., 2026)	17/20	10/20	6/20	15/20	18/20	16/20	13/20	9/20	65.0
Xiaomi-Robotics-1 (Xiaomi Robotics Team et al., 2026)	16/20	14/20	10/20	16/20	18/20	18/20	10/20	15/20	73.1
GalaxeaVLA (G0.5) (Liu et al., 2026)	19/20	12/20	14/20	19/20	19/20	18/20	12/20	11/20	77.5
N_0-TWAM (NeoteAI Team & Fudan TEAI Team, 2026a)	18/20	16/20	**17/20**	18/20	**20/20**	17/20	8/20	13/20	79.4
UWF (Ours)	**20/20**	**19/20**	**17/20**	**20/20**	**20/20**	**20/20**	**14/20**	**16/20**	**91.3**

gains closer to one, show that UWF preserves both changing latent structure and its amplitude more accurately.

Does the action condition actually matter? We freeze all weights and cyclically replace only the copied action-prefix tokens with those from another fixed window, leaving policy tokens, world inputs, masks, modulation and positions unchanged. All three windows worsen: the mean per-window relative error increase is 19.4% in vision and 5.2% in tactile sensing. Re-inserting each window's own prefix produces exactly the original output. The reconstruction therefore responds to the supplied action condition. Together with the 42.1%/49.7% latent-MSE reductions, this shows lower reconstruction error while retaining action-conditioned prediction. Task evaluations complement these diagnostics: UWF improves mean success over N_0-TWAM by 5.6 points in simulation and 11.9 points on real robots (Tables 2 and 3). Appendix F reports the complete intervention results.

Are the locations and directions of change preserved? Figure 3 compares spatial localization and structural agreement. At 25% selected area, mean region IoU improves from 0.788 to 0.849 in vision and 0.681 to 0.767 in tactile sensing; both modalities improve at all six evaluated area fractions. Future-step change cosine increases from 0.916 to 0.956 and 0.887 to 0.946, respectively (Figure 2(c)). The gains span spatial localization and the direction of predicted temporal changes. Appendix F.2 defines the metrics; Appendix F.4 further analyzes gains across change-intensity regions and decomposes reconstruction error.

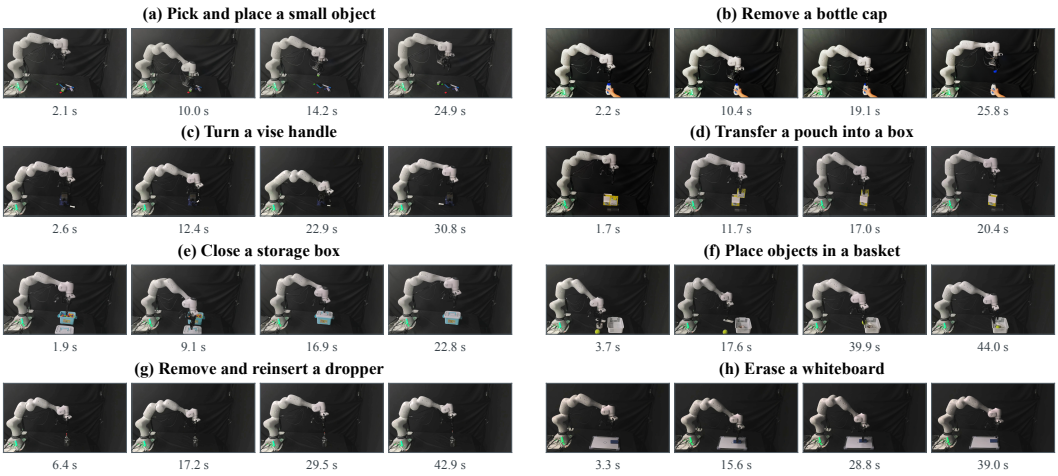

Figure 5: **Eight real-robot task sequences.** Each task has four timestamped, uncropped frames in their original aspect ratio; global tonal adjustments improve visibility. Two tasks share each row. Tasks span transfer, cap removal, rotation, lid closure, dropper reinsertion and wiping.

How much does spatial structure improve? Figure 4 shows window 50: projected visual/tactile MSE falls by 53.4%/72.0%, and full-feature MSE by 44.2%/53.6% (Appendix F.5). Across the three windows, mean spatial rank correlation improves from 0.920 to 0.968 in vision and 0.938 to 0.972 in tactile sensing; gradient cosine improves from 0.909 to 0.946 and 0.902 to 0.934, respectively (Figure 3(b–e)). All first-slot pairs improve; second-slot differences are small and mixed. All six mismatched prefixes increase error (17.9%/5.7% on average), showing that matching action prefixes consistently improves prediction within this checkpoint; full results appear in Appendix F.3.

4.3 Eight-task real-robot comparison

Table 3 reports 20 trials per task and method (800 total). UWF succeeds in 146/160 trials, averaging **91.3%** versus 79.4% for N_0-TWAM, the highest-scoring baseline: an 11.9-percentage-point gap. It leads on six tasks and ties for best on two. Figure 5 illustrates execution progress.

Task-level gains and interpretation. Against N_0-TWAM, dropper manipulation improves by 30 points and three tasks by 15 points; failures fall from 33 to 14 (57.6% fewer). Dropper manipulation and wiping account for 10 of the 14 remaining failures, identifying the main tasks for further improvement. With 20 trials, rates change in five-point increments, and perfect observed scores do not establish reliability. Appendix E (Figure 7, Table 6) reports mean visual/tactile SSIM of 0.855/0.857 for GT–UWF pairs, measuring selected-image fidelity rather than between-model prediction superiority. The worst-task success rate also rises from 40% to 70%, showing gains beyond near-saturated tasks; this finite-task result is not a reliability guarantee.

5 Conclusion

UWF connects world prediction to action generation through read-only conditioning, supervised evidence utility, evidence-only corrections and sequence-level commitment. The base action field remains available, and demonstration supervision trains corrections even when rejected. Mean success rates reach 90.1% in simulation and 91.3% on real robots, with six task wins and two ties against the strongest real-robot baseline. Reconstruction diagnostics further show improved spatial change localization and gradient alignment in both modalities. The gains span insertion, extraction and real-robot object handling. This separates evidence selection from correction strength. On real robots, failures fall from 33 to 14 and the lowest task success rate increases from 40% to 70%. Future work should address matched ablations, calibration under distribution shift and inference latency, including when selective feedback remains effective as prediction horizons grow.

AI USE STATEMENT

Generative AI tools assisted algorithm prototyping, code development and revision, experimental scripting and execution, analysis and visualization, literature organization, and manuscript drafting and editing. The reported diagnostic values are computed from recorded model outputs; AI-generated prose or illustrations are not used as experimental measurements. Verification performed during preparation includes paired-input and checkpoint checks, numerical consistency checks, and rendering checks, as documented in Appendix F. The authors have completed a thorough review of all AI-assisted technical content, citations, code, and empirical claims. The authors are responsible for the final submitted content, including AI-assisted text, code, figures, and claims.

ETHICS STATEMENT

This work focuses on robot manipulation learning, using demonstrations to train and evaluate manipulation policies. No experiments were conducted with humans as research subjects. Predicted visual and tactile observations can be inaccurate, and the learned acceptance rule constrains a surrogate prediction error rather than physical risk. Accordingly, neither improved reconstruction nor the proposed gate should be interpreted as a safety guarantee; deployment around people or fragile objects requires independent safeguards and task-specific validation. Existing demonstrations and pretrained models should be used and redistributed only under their applicable permissions, with attention to identifiable information in recorded observations. The current manuscript reports simulation task-success results, published reference scores, real-robot success counts and sequences, and limited training-window diagnostics. We distinguish published reference scores from matched experiments and retain the limits of the diagnostic findings to avoid overstating generalization or control performance.

REPRODUCIBILITY STATEMENT

Section 3 defines the method, and Appendix A specifies the losses, masking rules, gradient paths, and sampling procedure; Appendix B records objective coefficients and fixed scheduler weights. Section 4 and Appendix C identify evaluation scope and the sources of reference results. Appendix F documents the paired diagnostic setup and its limitations; Appendices F.2, F.3, and F.5 define the spatial metrics, complete prefix intervention, and shared feature projection. The project archive retains plotting scripts, paired arrays, checkpoint and source hashes, and the intervention log. No publicly accessible anonymous code or data package is linked in this draft; access to the underlying model weights and demonstration data remains necessary to reproduce the model forwards.

REFERENCES

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*, 2026a.

Rui Cai, Jun Guo, Xinze He, Piaopiao Jin, Jie Li, Bingxuan Lin, Futeng Liu, Wei Liu, Fei Ma, Kun Ma, et al. Xiaomi-robotics-0: An open-sourced vision-language-action model with real-time execution. *arXiv preprint arXiv:2602.12684*, 2026b.

Baijun Chen, Weijie Wan, Tianxing Chen, Xianda Guo, Congsheng Xu, Yuanyang Qi, Haojie Zhang, Longyan Wu, Tianling Xu, Zixuan Li, et al. Univtac: A unified simulation platform for visuo-tactile manipulation data generation, learning, and benchmarking. *arXiv preprint arXiv:2602.10093*, 2026.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.

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

Abraham George, Selam Gano, Pranav Katragadda, and Amir Barati Farimani. Vital pretraining: Visuo-tactile pretraining for tactile and non-tactile manipulation policies. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 258–264. IEEE, 2025.

Carolina Higuera, Sergio Arnaud, Byron Boots, Mustafa Mukadam, Francois Robert Hogan, and Franziska Meier. Visuo-tactile world models. 2026.

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

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

SM Islam and Manfred Huber. Tasksense: Focusing on what matters in world models. *arXiv preprint arXiv:2608.06544*, 2026.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pp. 33–50, 1978.

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

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Yicheng Liu, Zibin Dong, Baijun Ye, Tianyuan Yuan, Tao Jiang, Anqi Yang, Shicheng Cao, Haonan Liu, Yue Sun, Zihan Guo, et al. G0. 5: One autoregressive stream for robot reasoning and action. *arXiv preprint arXiv:2608.11739*, 2026.

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.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Lloyd S Shapley et al. A value for n-person games. 1953.

Aravind Srinivas, Allan Jabri, Pieter Abbeel, Sergey Levine, and Chelsea Finn. Universal planning networks: Learning generalizable representations for visuomotor control. In *International conference on machine learning*, pp. 4732–4741. PMLR, 2018.

NeoteAI Team and Fudan TEAI Team. n_0-twam: Scaling tactile-native world-action model for contact-rich manipulation, 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, 2026b. URL https://arxiv.org/abs/2607.23782.

Xiaomi Robotics Team, Jun Guo, Piaopiao Jin, Jason Li, Peiyan Li, Yingyan Li, Futeng Liu, Wanli Peng, Optimus Qin, Yifei Su, Nan Sun, Qiao Sun, Runze Suo, Heyun Wang, Yunhong Wang, Rujie Wu, Caoyu Xia, Lina Zhang, Jack Zhao, Guoliang Chen, Wenlong Chen, Xinze He, Bin Li, Qing Li, Zhuorong Li, Heng Qu, Wenxuan Song, Diyun Xiang, Yifan Xie, Peiran Xu, Hangjun Ye, Wen Ye, Han Zhao, and Quanyun Zhou. Xiaomi-robotics-1: Scaling vision-language-action models with over 100k hours of real-world trajectories, 2026. URL https://arxiv.org/abs/2607.15330.

Yuran Wang, Siqiao Huang, Mingleyang Li, Chenhao Zhang, Jiaqi Liang, Weiyang Jin, Yue Chen, Xuemin Chi, Donghao Zhou, Qize Yu, et al. Openwam: An open, modular exploration towards systematic world-action model pretraining. *arXiv preprint arXiv:2609.07398*, 2026.

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

Zihang Yao, Chaoyue Ding, and Yingying Yu. Disentangling visuo-tactile foresight: Oracle-guided interface discovery for world action models. *arXiv preprint arXiv:2608.00547*, 2026.

Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Hao Li, Hengtao Li, Jie Li, Jindi Lv, Jingyu Liu, et al. Gigaworld-policy: An efficient action-centered world–action model. *arXiv preprint arXiv:2603.17240*, 2026a.

Jinhui Ye, Ning Gao, Senqiao Yang, Jinliang Zheng, Zixuan Wang, Yuxin Chen, Pengguang Chen, Yilun Chen, Shu Liu, and Jiaya Jia. Starvla-α: Reducing complexity in vision-language-action systems. *arXiv preprint arXiv:2604.11757*, 2026b.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *Advances in neural information processing systems*, 33:14129–14142, 2020.

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

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

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.

Jianbo Zhou, Boyuan Zhao, Yuzheng Zhang, Yiyang Chen, Wenxin Chen, Qiuyue Li, Xiangyang Gu, Yuhan Cao, Xiao Xia, Yanzhe Hu, et al. Tacforcing: Streaming action generation with execution-time tactile feedback. *arXiv preprint arXiv:2608.25798*, 2026.

SUPPLEMENTARY MATERIAL (APPENDIX)

A EXPANDED FORMULATION AND IMPLEMENTATION DETAILS

A.1 PROBLEM SETTING AND DESIGN PRINCIPLE

We consider task-specific adaptation from demonstrations containing language, RGB observations, tactile observations, and robot actions. At deployment, the policy receives a history h_t of actual observations and predicts a complete future action chunk A_t. Our implementation builds on the pretrained N_0-TWAM mixture-of-transformers backbone (NeoteAI Team & Fudan TEAI Team, 2026a); we do not introduce a new foundation-model pretraining procedure. The original action interface is retained, including masks for inactive action coordinates. The world streams predict future visual and tactile latents, and the action stream predicts a conditional flow velocity, following the flow-matching formulation (Lipman et al., 2022). For a normalized clean action chunk A and Gaussian noise ϵ, the implemented interpolation is $A_\sigma = (1-\sigma)A + \sigma\epsilon$, with target velocity $u = \epsilon - A$. Consequently, the contribution and risk targets below evaluate a denoising surrogate at a specified noise level, not a physical reward.

Let v_0 denote the base action velocity and let D denote evidence extracted from the predicted world. UWF introduces a residual r and a scalar commitment coefficient λ for each chunk:

$$v_{\text{final}} = v_0 + \lambda r, \qquad r = F_\theta(D; h_t, A_\sigma, \sigma), \qquad 0 \le \lambda \le 1. \tag{10}$$

Here A_σ is the current noisy action iterate; σ abbreviates the modality-specific noise levels. The correction is applied to the *velocity field*, not added directly to a physical end-effector pose. The sampling procedure integrates this field to obtain the action chunk. The base branch is itself adapted during joint training: fallback to v_0 does not mean reverting to an untouched pretrained checkpoint.

Targets and the common action loss. All action proposals for a training example are compared with the *same* sampled flow target u, not with a newly sampled target for each branch. Let $\mathcal{F}_a$ index valid sample–frame pairs in a minibatch, i index action slots, d index coordinates, and M_{fid} be the action-validity mask. We use the frame-balanced, noise-weighted operator

$$\mathsf{D}_a(v,u) = \frac{1}{|\mathcal{F}_a|} \sum_{f \in \mathcal{F}_a} \omega_a(\sigma_f) \frac{\sum_{i,d} M_{fid}(v_{fid} - u_{fid})^2}{\sum_{i,d} M_{fid}}, \tag{11}$$

where ω_a is the fixed flow-scheduler training weight, not a learned gate. It is precomputed on the modality's noise grid and normalized to have grid mean one; its construction is given in Appendix B. Frames with no valid actions, including the initial context slot, are excluded. This defines the two action-supervision terms used throughout the method:

$$\mathcal{L}_{\text{act}}^{\text{base}} = \mathsf{D}_a(v_0, u), \qquad \mathcal{L}_{\text{act}}^{\text{gated}} = \mathsf{D}_a(v_0 + \text{sg}[\lambda] r, u)\,. \tag{12}$$

Here sg denotes stop-gradient and λ is computed by Section A.5. For contribution labels and blockwise risk targets, we instead use the unweighted masked chunk error $\ell_M(v,u) = \sum M(v-u)^2 / \sum M$; it is distinct from D_a, which includes scheduler weights and frame balancing. Zero-valid-element reductions contribute zero.

Equation 10 separates three roles. World prediction supplies candidate evidence, contribution learning determines its relative usefulness, and commitment determines whether to use the resulting proposal. The following components implement these roles without requiring unconditional trust in forecasts.

A.2 A READ-ONLY PREFIX FOR WORLD CONDITIONING

A direct feedback connection can obscure the distinction between an action-conditioned forecast and a forecast conditioned on intermediate policy features. We make that distinction explicit inside the transformer. At the start of each forward pass, the model copies the embedded *current noisy action inputs*, together with their time and position information, into a separate conditioning stream. This stream is restored from the same input representation at every transformer layer; it is not replaced by the layer's updated action features. It remains differentiable, allowing world-prediction losses

to train the shared embeddings and projections. "Read-only" therefore refers to its forward-pass information content, not to frozen parameters or detached gradients.

The attention mask separates observed context, predicted world tokens, policy tokens, and the copied action-conditioning stream. Observed world context is restricted to the corresponding history prefix. Predicted world queries may read the permitted noisy-action conditioning prefix, while policy queries do not directly attend to predicted world tokens in the shared backbone. The latter restriction routes predicted-world influence through the explicit feedback head instead. The observed local tactile pathway is likewise restricted to the preceding context of the chunk. This organization does not assert that a forecast is physically correct; it specifies where forecast information can enter action generation and keeps evolving policy features out of the read-only stream.

The prefix is reconstructed for each denoising forward pass. It is not a persistent key–value cache, nor a claim that the training and deployment control clocks are automatically aligned. Those are separate implementation and evaluation questions.

How the constrained world streams are trained. The prefix is an architectural restriction, not an additional regularization loss. Each visual or tactile stream m receives its own noised latent z_σ^m and predicts velocity v^m against $u^m = \epsilon^m - z^m$. Let e_{mf} be the mean squared velocity error over all spatial positions and latent channels of sensor m at valid sample–frame pair f. Writing $\mathcal{M}_v$ for the two camera streams and $\mathcal{M}_t$ for the two tactile streams, their objective is

$$\mathcal{L}^w_{\text{world}} = \frac{1}{2|\mathcal{F}_w|} \sum_{f \in \mathcal{F}_w} \left[\omega_v(\sigma_f) \sum_{m \in \mathcal{M}_v} \text{sg}[w_{mf}] e_{mf} + \sum_{m \in \mathcal{M}_t} \text{sg}[w_{mf}] e_{mf} \right]. \tag{13}$$

Here $\mathcal{F}_w$ contains valid world frames, and w_{mf} is the sensor–time contribution weight constructed in Section A.3; the initial context frame has weight one. The implementation retains the pretrained recipe's distinction between scheduler-weighted visual error and global tactile error without an additional scheduler weight. If tactile prediction is unavailable, its term is zero. These losses train the world streams under the prefix constraint, while action-feedback gradients provide the additional route described in Section A.4.

A.3 World evidence and action-conditioned contribution

We describe world evidence as a change relative to an observation, rather than as an unconstrained hidden representation. For each two-frame forecast chunk, the model converts its predicted visual and tactile velocities into clean-latent estimates and subtracts the preceding context latent from the corresponding sensor stream. Each difference map is pooled into a 2×2 spatial grid. With two camera streams and two tactile streams at two forecast positions, this yields eight sensor–time groups, each containing four evidence tokens. These are fixed, structured groups; the method does not infer an object graph or discover symbolic contact states. Unavailable modalities and padded frames are masked out. The two future positions preserve a minimal within-chunk temporal distinction instead of collapsing an entire forecast into one vector. Their preceding-context reference is taken from the permitted conditioning stream, including its training-time augmentation when enabled; at deployment it comes from actual observations. The feedback interface also receives the six modality–time noise levels, distinguishing an uncertain early denoising estimate from a later one without treating the noise level itself as a calibrated measure of model error.

Which of these groups should matter to the action proposal? We construct a training target using an exponential-moving-average copy of the feedback head. The lagged teacher reduces the target's immediate dependence on the same student update that it supervises, although the target still evolves during training. For a random permutation π of the evidence groups, let S_k contain the first k groups and let $\bar{r}(S_k)$ be the teacher's residual using that subset. Writing u for the demonstration-derived action flow target and ℓ_M for masked mean-squared error, the sampled marginal contribution is

$$c_{\pi_k} = \ell_M(v_0 + \bar{r}(S_{k-1}), u) - \ell_M(v_0 + \bar{r}(S_k), u)\,. \tag{14}$$

A positive contribution means that adding a group reduces the teacher proposal's action-target error in that sampled context. The construction follows the marginal-contribution perspective underlying Shapley values (Shapley et al., 1953), but uses a sampled permutation chain rather than an exact enumeration of all coalitions. It intervenes on which evidence the feedback head can access, not on

the physical environment. The backbone forecast is shared across the coalition probes; a separate transformer rollout is not required for every subset. For eight groups, one permutation requires nine nested coalitions, including the empty set. The teacher probes are evaluated without the learned contribution prior, so that the current prior is not directly used to manufacture its own target. Because the empty-coalition residual is zero, the sampled contributions telescope:

$$\sum_{g=1}^{8} c_g = \ell_M(v_0, u) - \ell_M(v_0 + \bar{r}(\{1, \dots, 8\}), u)\,. \tag{15}$$

This identity accounts for the teacher proposal's total surrogate improvement along a sampled chain; it does not establish a unique causal attribution or downstream policy improvement.

The utility head predicts normalized versions of these targets from detached context and evidence features. Its scores define positive, mean-one weights over available groups, with a nonzero floor so that low current utility does not eliminate supervision for a world stream. Specifically, the regression target is $\widetilde{c}_g = c_g / \max\{\ell_M(v_0, u), \epsilon_c\}$, where ϵ_c is a fixed normalization floor. Normalizing by the base error expresses contribution relative to the current prediction difficulty, while the denominator floor prevents amplification when that error is small. The marginal contribution c_g is a *label*, not itself an extra term added to the training objective. The score-regression loss is

$$\mathcal{L}_{\text{utility}} = \frac{\sum_{j,g} \chi_{jg} H_\kappa(s_{jg} - \text{sg}[\widetilde{c}_{jg}])}{\sum_{j,g} \chi_{jg}}, \tag{16}$$

where j indexes chunks, χ_{jg} marks an available group in a valid chunk, and $H_\kappa(x) = x^2/(2\kappa)$ for $|x| < \kappa$, and $|x| - \kappa/2$ otherwise. Thus $\mathcal{L}_{\text{utility}}$ is Smooth-L1 regression of the EMA-teacher labels; its inputs and labels are detached from the shared backbone. For scores s_g, available groups $\mathcal{A}$, and $G = |\mathcal{A}| > 0$, the weights are

$$w_g = \frac{1}{2} + \frac{G}{2} \frac{\exp(s_g/T)}{\sum_{j \in \mathcal{A}} \exp(s_j/T)}, \qquad g \in \mathcal{A}, \tag{17}$$

with $T = 1$ in the current implementation and zero weight outside $\mathcal{A}$. Available weights average to one and are at least $1/2$; negative contribution predictions therefore reduce relative emphasis rather than delete a modality. The weights influence both world-prediction losses and the feedback head's attention prior. In the latter, $\log w_g$ is added to the attention logits of all four patches in group g. They are detached when used in these two places: the contribution estimator is trained against explicit marginal-improvement labels, rather than simply learning to suppress whichever prediction loss is largest. This allocates supervision by action relevance, without guaranteeing task-optimality.

A.4 Evidence-only correction of the full action chunk

The feedback head uses attention to map sensor–time evidence into an action-sequence residual. Queries are formed from the noisy action iterate, the base clean-action estimate, noise levels, action positions, and any available masked state context. Keys carry the evidence representation and sensor–time/spatial position information. Crucially, values are obtained through a *bias-free* linear projection of the world-difference tokens. The residual decoder is also bias-free and uses a bounded output nonlinearity. Thus, conditioning information can determine *where to read*, but it cannot independently create a nonzero correction through the value pathway.

This construction gives a simple structural property: if all available evidence values are zero, the feedback residual is exactly zero. For an empty coalition, attention is evaluated safely and its output is explicitly suppressed, recovering the same null behavior. The residual decoder is zero-initialized so that the new pathway initially leaves the base proposal unchanged. These properties identify a controlled fallback; they do not imply that every nonzero residual is useful.

The feedback head predicts a correction at every position in the action sequence, with action-position embeddings interpolated to the sequence length. The real-robot deployment configuration used in this paper contains 24 action slots; the standalone joint sampler separately provides a default 16-step configuration (Appendix A.7). During training, predictions corrected using full evidence or sampled evidence subsets are directly supervised against the demonstration-derived flow target. Neither auxiliary loss is multiplied by the gate coefficient. Thus, even when a correction is excluded

from the final output, its prediction error still updates the feedback head. The full-evidence proposal uses the contribution prior; the auxiliary subset proposal uses a randomly chosen nonempty proper prefix of the sampled permutation without that prior. Both objectives use a detached base velocity in their additive proposal, removing the direct shortcut of satisfying the auxiliary target by updating the base output alone. This does not detach the shared backbone everywhere: gradients may still reach it through evidence and query conditioning.

Proposal supervision. Let $r^S = F_\theta(D; S)$ denote the student feedback residual when attention is restricted to the sampled subset S. Using the action-loss operator of Equation 11, we obtain

$$\begin{aligned} \mathcal{L}_{\text{act}}^{\text{raw}} &= \mathsf{D}_a(\text{sg}[v_0] + r, u)\,, \\ \mathcal{L}_{\text{act}}^{\text{subset}} &= \mathbb{E}_S\, \mathsf{D}_a\big(\text{sg}[v_0] + r^S, u\big)\,. \end{aligned} \tag{18}$$

The expectation is estimated with one sampled nonempty proper coalition per chunk. Neither term is multiplied by λ. Consequently, a closed commitment gate suppresses the residual in the executed proposal but does not suppress these learning signals. The base and gated outputs are separately supervised by Equation 12; no extra "zero-evidence loss" is needed because the null-residual property follows from the architecture.

A.5 Sequence-level commitment from predicted improvement

Contribution and commitment answer different questions. Contribution compares evidence subsets during training; commitment asks whether the *current combined proposal* should alter the base output. We divide a chunk into four contiguous temporal blocks and assess the residual relative to the base velocity in each block. For block b, let $e_b = v_{0,b} - u_b$, and use averages over its valid coordinates to define residual energy $E_b = \langle r_b, r_b \rangle$ and the linear improvement term $B_b = -2\langle e_b, r_b \rangle$. The exact change in this local squared-error surrogate is

$$\ell_b(v_0 + \lambda r, u) - \ell_b(v_0, u) = \lambda^2 E_b - \lambda B_b. \tag{19}$$

Residual energy is available without a target, but B_b is known only during training. A risk head therefore predicts a low conditional quantile, at level α, of the energy-normalized improvement term. Define $\widetilde{B}_b = B_b/(\sqrt{E_b} + \epsilon_r)$ and let $\widehat{b}_b$ be the risk head's prediction in this normalized scale. Unlike the contribution labels, which use the EMA teacher, this target is computed from the *current student's* full-evidence residual and the same demonstration flow target u. The head is trained by

$$\begin{aligned} \mathcal{L}_{\text{risk}} &= \frac{\sum_{j,b} n_{jb} \nu_{jb} \rho_\alpha\Big(\text{sg}[\widetilde{B}_{jb}] - \widehat{b}_{jb}\Big)}{\sum_{j,b} n_{jb} \nu_{jb}}, \\ \rho_\alpha(x) &= \max\{\alpha x, (\alpha - 1)x\}. \end{aligned} \tag{20}$$

Here ν_{jb} selects blocks with valid coordinates and $E_{jb} \geq \epsilon_E$, and n_{jb} counts their valid coordinates. Features, residual energy, and target construction are detached for this regression, so this term trains the risk head rather than allowing the proposal to change its own target through the risk loss. At both training and deployment, the scale is restored as $q_b = (\sqrt{E_b} + \epsilon_r)\widehat{b}_b$. Let q_b denote the prediction after restoring its energy scale, and let n_b be the number of valid coordinates in the block. Let $\mathcal{B}$ contain blocks with valid coordinates and $E_b \geq \epsilon_E$. For nonempty $\mathcal{B}$, ignoring numerical stabilizers, the implemented rule is

$$\lambda = \text{clip}_{[0,1]}\left[\min\left\{\frac{\sum_{b\in\mathcal{B}} n_b q_b}{2\sum_{b\in\mathcal{B}} n_b E_b}, \min_{b\in\mathcal{B}} \frac{q_b}{E_b}\right\}\right]. \tag{21}$$

The first term chooses a scale from the aggregate quadratic surrogate; the second limits that scale using the least favorable block. To obtain Equation 5, write the predicted change as $\lambda(\lambda E_b - q_b)$. For $E_b > 0$, any $q_b < 0$ leaves only zero feasible. Otherwise the feasible interval ends at the smaller of one and the minimum ratio q_b/E_b. Projecting the unconstrained aggregate minimizer onto this interval gives the rule above. This derivation ignores numerical stabilizers and uses only selected positive-energy blocks. A single coefficient is applied to the entire residual, so an apparently beneficial segment cannot independently authorize a correction elsewhere in the sequence. If no block has meaningful residual energy, the gate is zero. The coefficient is shared across positions

within a denoising evaluation and recomputed at the next solver event; it is not fixed for an entire episode. There is no separately labeled loss on λ: the commitment coefficient is computed from q_b and E_b and detached in the gated action loss. This separates a known geometric cost of the proposal from its unknown alignment with the target, rather than interpreting a raw prediction-error score as action risk.

Equation 19 explains the conservative form of the rule. If every q_b were a valid lower bound on B_b, the blockwise constraint would make the surrogate change nonpositive in every included block. Our learned quantile predictions do *not* establish that premise for each test example or jointly across blocks. We therefore interpret the gate as a learned conservative decision rule, not a calibrated safety certificate, a monotonic policy-improvement theorem, or a bound on physical task failure. Both calibration and realized action utility must be measured.

A.6 Joint learning and gradient routing

Equation 9 applies the same action-error operator to four supervised proposals. Its world term $\mathcal{L}_W$ is $\mathcal{L}^w_{\text{world}}$ in Section A.2; $\mathcal{L}_C$ and $\mathcal{L}_Q$ are the utility and risk losses in Equations 16 and 20. The base, committed, full-evidence and subset proposals correspond respectively to the base, gated, raw and subset losses defined above. One subset sample per chunk estimates the coalition expectation. Appendix B maps the coefficients to the original configuration. Action losses respect coordinate and horizon masks; world losses retain all valid frames with the declared contribution weighting. The feedback teacher is updated by an exponential moving average with decay 0.99.

Detachment is used selectively. Contribution labels, teacher proposals, and the commitment coefficient are not differentiated through as optimization decisions. The feedback values remain connected to predicted world latents, so the proposal losses can train world predictions through their effect on action-target error. The method consequently offers two distinct forms of supervision: an explicit contribution target for allocating world evidence, and an action gradient through the evidence-to-residual pathway. Neither substitutes for closed-loop evaluation.

For each minibatch, we first sample noise and construct the shared flow targets, then perform one backbone forward pass under the prefix mask. The student heads produce contribution scores, full/subset residuals, and the risk prediction; the EMA head supplies the detached coalition labels. Grouping visual and tactile supervision into one world term, we compute the seven objective terms above, combine them using Equation 9, and backpropagate the resulting scalar. The feedback teacher is updated only after the optimizer step. This order separates constructing a supervision target from optimizing against it.

A.7 Mixed-clock joint sampling and observed-history renewal

At deployment, coalition labels and demonstration targets are absent. The model predicts world and action velocities, forms the residual, applies the gate, and integrates the resulting field to produce an action chunk. World and action streams use different noise schedules: the world streams use the shift $s(\tau) = 5\tau/(1 + 4\tau)$, whereas the action schedule is unshifted. The current sampler combines the three-step world clock and the four-step action clock into one descending union of update events,

$$\mathcal{T} = \{0, \tfrac{1}{3}, \tfrac{2}{3}\} \cup \{0, \tfrac{1}{4}, \tfrac{1}{2}, \tfrac{3}{4}\}. \tag{22}$$

The elements of $\mathcal{T}$ are integration endpoints, not the noise levels at which model evaluations are performed. Starting from noise level one, each of the six distinct events evaluates the joint model at the streams' current noise levels, but advances only the stream or streams scheduled at that event. Only the two future world slots and their action slots are updated; observed context remains fixed throughout the solve. The default geometry uses eight action commands per future slot (16 total); an explicit training-geometry diagnostic mode uses 12 per slot (24 total). Their world-anchor offsets are 0/8/16 and 0/12/24, respectively, so decoded forecasts from these modes must not be evaluated against temporally mismatched targets. This 16-action default is a standalone sampler setting, distinct from the shared real-robot deployment configuration described below. Thus action corrections can be recomputed as the world estimate changes, without an inner trajectory optimizer or a separate backbone rollout for each coalition. This is the implemented joint mixed-clock sampler, not the released sequential world-then-action inference procedure; its effect must be isolated in matched inference comparisons. For the real-robot comparison in Table 3, UWF and N_0-TWAM share a

streaming deployment configuration with two blocks of 12 action commands (24 slots total), and three world and four action denoising steps. The cold-start chunk skips its first 12-slot seed block; subsequent complete chunks execute all 24 slots. Real visual and tactile keyframes are collected every three action slots and re-encoded to update the history cache after each completed chunk. An interrupted chunk is not committed to the cache. Thus subsequent predictions use observed history rather than imagined outcomes. This streaming execution protocol is distinct from the standalone joint sampler's history recomputation; equal denoising-step counts do not imply identical update schedules. This separation between a learned prediction-to-action interface and its execution protocol is important: comparisons of the method require the same controller, action budget, observation preprocessing, and trial-validity rules across policies.

B HYPERPARAMETERS AND FIXED SCHEDULER WEIGHTS

Equation 9 uses the following coefficients. This reorganizes the notation without changing the training objective.

Main-text coefficient	Equivalent parametrization	Value
ω_{base}	β_{base}	0.3
ω_{commit}	1	1
ω_{full}	η_{fb}	0.3
ω_{subset}	$\eta_{\text{fb}}\beta_{\text{sub}}$	0.1
η_C	η_{dec}	0.1
η_Q	$\eta_{\text{dec}}\beta_{\text{risk}}$	0.05

The auxiliary coefficients and stabilizers are summarized below. Table 4 records their implementation settings; no coefficient is learned by minimizing the objective that it weights.

Table 4: Hyperparameters for the objective and auxiliary targets.

Symbol	Role	Setting
η_{fb}	Feedback-proposal supervision	0.3
η_{dec}	Contribution and commitment supervision	0.1
β_{base}	Base-branch supervision	0.3
β_{sub}	Relative subset-proposal weight	$1/3$
β_{risk}	Relative quantile-risk weight	$1/2$
ϵ_c	Contribution normalization floor	0.05
κ	Smooth-L1 transition scale	0.1
α	Lower-quantile level	0.1
ϵ_r	Risk-target normalization stabilizer	10^{-4}
ϵ_E	Minimum residual energy	10^{-8}

The world and gated-action objectives have unit coefficients. The effective subset and risk weights are $\eta_{\text{fb}}\beta_{\text{sub}}$ and $\eta_{\text{dec}}\beta_{\text{risk}}$, respectively. Each constituent retains its own mask and stop-gradient conventions; grouping the objectives does not make their gradient paths interchangeable.

For completeness, the fixed visual/action flow-scheduler weights are constructed separately on each modality's $N = 1000$-point noise grid. The grid starts from $\tau_k = 1 - k/N$ for $k = 0, \ldots, N-1$, followed by the modality-specific noise shift. At grid value $\sigma_{m,k}$, the implementation uses

$$g_{m,k} = \exp\left[-2(\sigma_{m,k} - \tfrac{1}{2})^2\right], \qquad \omega_m(\sigma_{m,k}) = \frac{N\,(g_{m,k} - \min_l g_{m,l})}{\sum_l (g_{m,l} - \min_r g_{m,r})}.$$

The action grid is unshifted and the visual grid uses the shift specified in Section A.7; the noise lookup returns the nearest registered grid point. These weights are fixed preprocessing for flow supervision, not contribution scores, commitment coefficients, or additional trainable objectives.

C ADDITIONAL UNIVTAC REFERENCE RESULTS

Table 5 extends the reference coverage to tactile foundation and reactive policies. These publications evaluate six tasks, excluding Insert HDMI and Grasp Classify; their means must not be ranked against the eight-task means in Table 2. Published reference numbers and rounding are retained as reported. Repeated methods denote different source records, not pooled trials or estimates.

Table 5: Six-task UniVTAC success rates (%). Source D: FTP-1 (Yuan et al., 2026a), Table 1. Source E: TacForcing (Zhou et al., 2026), Table 1. Avg.-6 averages only the six displayed tasks. No cross-source best-value highlighting is applied.

Method	Source	Lift Bottle	Pull-out Key	Lift Can	Put Bottle in Shelf	Insert Hole	Insert Tube	Avg.-6
VLA / action-policy baselines								
VITaL† (George et al., 2025)	D	72	47	8	32	25	34	36.33
UniVTAC-ACT† (Chen et al., 2026)	D	71	46	29	31	25	56	43.00
$\pi_{0.5}$ (Intelligence et al., 2025)	D	97	38	72	16	31	41	49.16
Tactile-VLA (Huang et al., 2025)	D	97	32	15	10	41	56	41.83
FTP-$\pi_{0.5}$ (Yuan et al., 2026a)	D	77	30	26	19	47	72	45.16
FTP-1 (Yuan et al., 2026a)	D	97	48	65	47	64	79	66.66
$\pi_{0.5}$ (Intelligence et al., 2025)	E	88	43	46	43	39	48	51
UniVTAC-ACT (Chen et al., 2026)	E	59	41	24	6	36	58	37
RDP (Xue et al., 2025)	E	84	18	12	41	23	75	42
FTP-1 (Yuan et al., 2026a)	E	89	35	66	23	62	76	59
TacForcing (Zhou et al., 2026)	E	90	48	63	43	69	79	65
WAM								
UWF (Ours)	Ours	67	98	88	88	96	95	88.67

Author–year citations identify methods; source letters identify score reports.

The dagger marks entries that the FTP-1 paper attributes to the original benchmark. Its UniVTAC-ACT Insert Hole entry is 25, whereas the original UniVTAC Table I reports 24; we preserve this discrepancy rather than silently reconcile the records. Likewise, FTP-1 has a reported mean of 66.66 in its own paper and 59 in the TacForcing evaluation. TacForcing uses 50-action simulation horizons partitioned into blocks of five, unlike the 24-action training and 16-action execution horizons of the N_0-TWAM reference. These differences motivate matched reruns, not a pooled leaderboard.

Oracle-future interface studies, including OVTF/AFM (Yao et al., 2026), are relevant to the scientific question but are excluded from these deployment-policy tables because their oracle conditions supply future information unavailable to an ordinary deployed policy.

D QUALITATIVE TASK ILLUSTRATIONS

Figure 6 illustrates the insertion, extraction and object-handling configurations of the eight UniVTAC tasks.

E REAL-ROBOT VISUAL AND TACTILE PREDICTION EXAMPLES

Table 6 quantifies all 32 GT–prediction pairs: two visual and two tactile images per task. Visual images are 256×256 pixels; tactile difference visualizations are 128×128. Metrics are computed from the provided RGB images in the 0–255 range without rescaling, image registration or brightness adjustment. The tactile metrics assess agreement of the supplied difference visualizations, not physical force or displacement accuracy.

For RGB arrays $I, \widehat{I}$, MAE and MSE average $|I - \widehat{I}|$ and $(I - \widehat{I})^2$ over all pixels and channels. PSNR is $10 \log_{10}(255^2/\mathrm{MSE})$, computed per image before averaging. SSIM uses an 11×11 Gaussian window with standard deviation 1.5, constants $C_1 = (0.01 \cdot 255)^2$, $C_2 = (0.03 \cdot 255)^2$, and

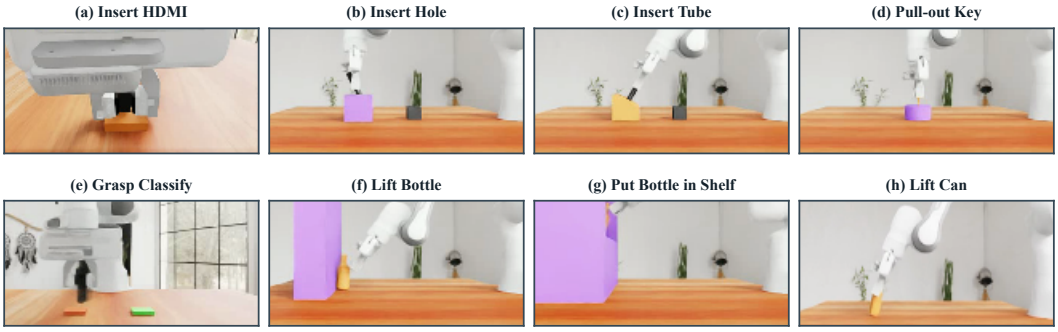

Figure 6: **Illustrative configurations across eight UniVTAC tasks.** Cropped terminal views show insertion, extraction and object-handling scenarios.

Table 6: Image-space agreement with GT. Each task value averages two sensor images; the final row equally averages all eight tasks. MAE uses the 0–255 intensity scale; PSNR is in dB. Lower MAE and higher PSNR/SSIM indicate closer image agreement.

Task	Vision			Tactile		
	MAE ↓	PSNR ↑	SSIM ↑	MAE ↓	PSNR ↑	SSIM ↑
Clutter picking	5.343	30.62	0.753	4.148	32.25	0.794
Object sorting	3.003	34.09	0.927	3.441	33.59	0.846
Dropper manipulation	5.406	30.21	0.745	4.111	32.69	0.836
Bottle uncapping	2.717	36.41	0.913	4.334	32.10	0.862
Lever manipulation	4.936	31.49	0.763	2.842	35.80	0.912
Package manipulation	3.687	33.06	0.881	2.012	38.85	0.957
Lid manipulation	2.597	36.65	0.940	3.137	34.53	0.895
Surface wiping	3.606	27.23	0.920	3.805	31.40	0.756
Mean	3.912	32.47	0.855	3.479	33.90	0.857

population-weighted local moments. The local index is

$$\frac{(2\mu_I\mu_{\hat{I}} + C_1)(2\sigma_{I\hat{I}} + C_2)}{(\mu_I^2 + \mu_{\hat{I}}^2 + C_1)(\sigma_I^2 + \sigma_{\hat{I}}^2 + C_2)}.$$

It is averaged over RGB channels and valid window centers, excluding the five-pixel border. This fully specifies the reported SSIM convention.

The mean visual/tactile MAE is 3.912/3.479, PSNR is 32.47/33.90 dB, and SSIM is 0.855/0.857. These values quantify overall agreement while retaining visible local differences, including tactile marker patterns. The assets do not specify forecast horizon or sampling protocol, and they are selected examples rather than a complete held-out set. Without matched baseline predictions, these image metrics cannot establish relative prediction superiority or separate foreground geometry from background agreement. Table 3 separately reports task success.

F Frozen-checkpoint visualization audit

The paired diagnostic uses post-trained Insert HDMI checkpoints at 1000 updates, not the raw initialization and not our internal base branch. The three dataset positions were fixed at 25, 50 and 75. Both arms use the same first three latent slots, action normalization, target arrays and noise seed; their target and noised-input hashes agree. Checkpoint identities and arrays are archived.

This diagnostic follows the models' native training attention rules, which are part of the algorithm and are not replaced by one shared mask. Thus it is a matched-input reconstruction comparison, not a claim that all internal information routes are identical. Teacher forcing and training-set inputs limit this diagnostic; held-out and paired closed-loop validation remain necessary.

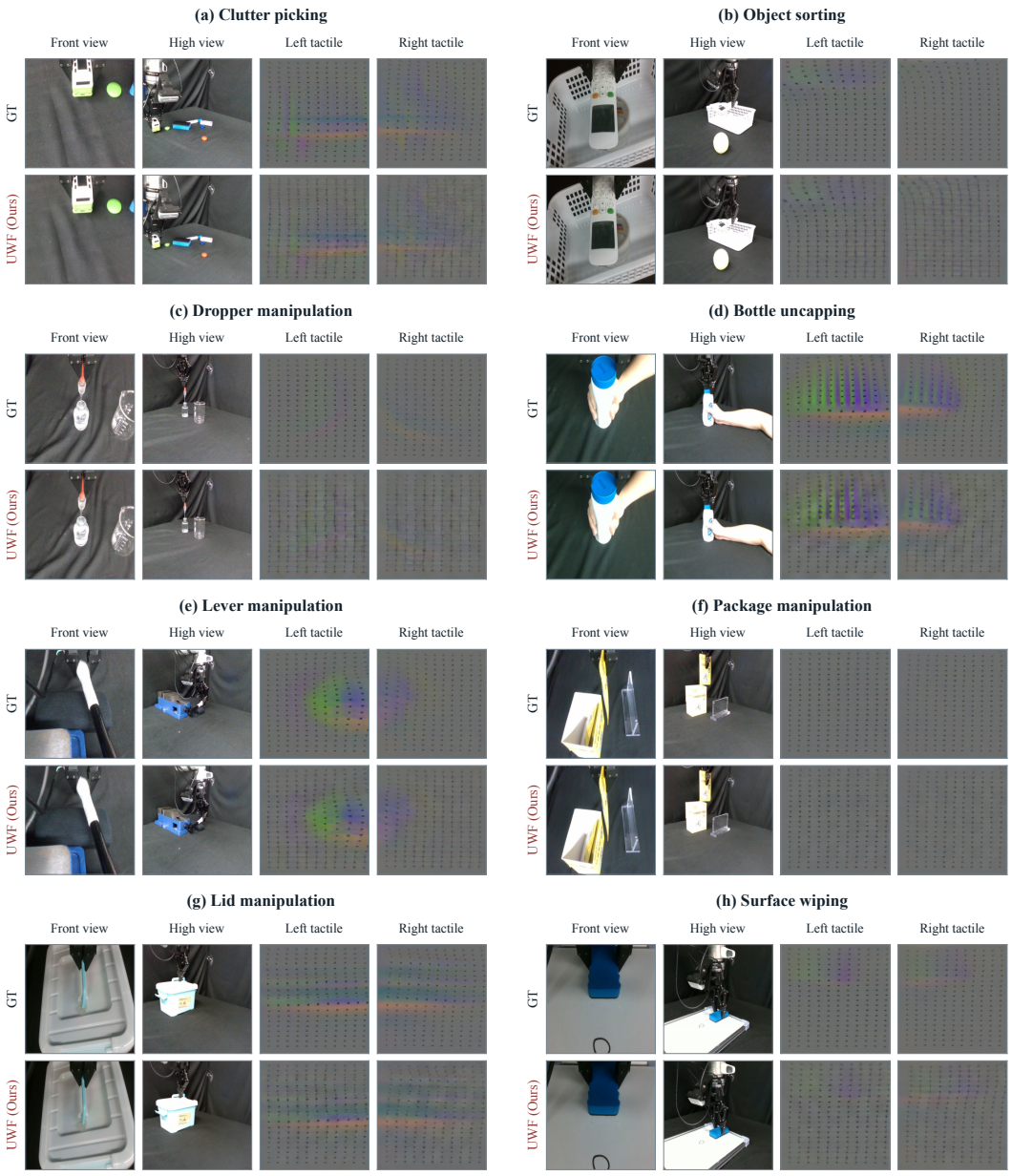

Figure 7: **Ground truth and UWF predictions across eight real-robot tasks.** Each task shows GT above UWF (Ours), with front/high camera views and left/right tactile difference visualizations. All 64 supplied images are retained at their original aspect ratios and colors, with no additional cropping, alignment or enhancement. Quantitative image agreement is reported in Table 6. These selected pairs illustrate prediction fidelity; they do not constitute a comparison with another prediction model.

For episode 50, visual/tactile latent MSE is 0.010047/0.022625 for our checkpoint and 0.018018/0.048711 for the post-trained reference. However, unweighted masked action-velocity MSE is 0.016838 versus 0.002639 in the same diagnostic. The result therefore supports a local reconstruction improvement, not dominance across objectives. All six gates in the separate noise-started attention probe are zero. We retain these findings rather than attributing the reconstruction difference to accepted feedback.

For the qualitative comparison in Figure 2(a), we select window 75 at decoded index 4, corresponding to the first future latent slot, using the same shared crops. Among the six window–endpoint pairs

(three windows, decoded indices 4 and 8), this pair maximizes the smaller of the visual and tactile relative decoded-crop MAE reductions. Visual MAE is 2.638957 versus 1.610949 (39.0% lower), and tactile MAE is 1.316650 versus 1.176799 (10.6% lower), for baseline versus Ours. Figure 2(a) shows the MAE reductions for all six window–endpoint pairs in a matrix, followed by channel-mean visual absolute-error maps for the selected pair; no raw scene comparison is displayed. All maps use the same square-root color normalization from 0 to 79 without clipping or per-model rescaling. The labels measure decoded-crop MAE, not two-slot latent MSE, and do not establish accurate gripper width or contact geometry. This is a selected favorable example, not a random sample or uniformly superior prediction. All three fixed windows have lower decoded-crop MAE at this horizon; the endpoint limitation below is retained.

At the decoded endpoint (frame index 8, window 50, using the original shared crops), Ours has 11.4% higher visual MAE and 3.2% higher tactile MAE than the baseline. The two-future-slot latent-MSE gains therefore do not imply uniformly better decoded-image reconstruction.

Attention is reconstructed in FP32 from the feedback head's actual Q/K, coalition mask and log-prior, leaving the original scaled-dot-product attention output unchanged. It has four heads, 24 action queries and 32 evidence patches. This is feedback-head attention, not backbone-wide saliency or a causal attribution score. The probe uses the joint mixed-clock diagnostic sampler and does not establish equivalence to a sequential world-then-action adapter.

F.1 Change-fidelity definitions and prefix intervention

For each sensor and future slot, let z_0 be its observed context latent, z its GT future and $\widehat{z}$ the reconstructed future. Set $d = z - z_0$ and $e = \widehat{z} - z$. We define change-projection gain $\gamma = \langle \widehat{z} - z_0, d\rangle / \langle d, d\rangle$, with ideal value one. All measured fields have nonzero change energy. Sensor–slot gains are averaged within each window, then across the three windows.

The high-change mask selects exactly the top quarter of spatial positions by channel-mean squared d, independently per sensor and future slot; the same GT-derived mask is used for both models. It is not a contact label. The temporal metric is the MSE of $(\widehat{z}_2 - \widehat{z}_1) - (z_2 - z_1)$, not physical trajectory error. Plotting the error metrics relative to the N_0-TWAM mean does not change their ordering.

The exact decomposition $e = e_{\parallel} + e_{\perp}$, where $e_{\parallel} = (\gamma - 1)d$, separates change-direction error from orthogonal error. Orthogonal-error reduction accounts for 82.3% of the total MSE reduction in vision and 66.1% in tactile sensing. Change-direction error also decreases, by 79.6% and 88.1%, respectively. These correlated summaries describe the same three reconstruction pairs; they are not independent statistical confirmations and do not assign semantic irrelevance to orthogonal directions.

For the frozen-prefix intervention, donor indices follow the fixed cycle $25 \to 50, 50 \to 75, 75 \to 25$. Only the duplicated action-memory hidden states are replaced immediately before the action expert's input projection. The recipient's policy tokens and all other backbone inputs are left intact. Re-inserting the recipient's own memory gives zero maximum absolute difference in both visual and tactile reconstructions for every window. The reported 19.4%/5.2% increases average the three per-window ratios; ratios of mean MSE instead give 19.3%/5.0%. This tests compatibility of the action condition at the current checkpoint, not an ablation of its training history or an immutable-versus-mutable architecture comparison.

Window	Vision intact	Vision swapped	Tactile intact	Tactile swapped
25	0.010630	0.012496	0.014394	0.015566
50	0.010047	0.012278	0.022625	0.023694
75	0.010530	0.012462	0.017351	0.017807

F.2 Spatial localization and temporal direction

Figure 3 uses the same three fixed training windows. For each sensor and future latent slot, we compute channel-mean squared change from the actual GT context, separately for GT and each prediction. This is an activity measure, not squared prediction error. Each map independently selects its top $\mathrm{round}(qHW)$ locations, using stable ordering for ties, for $q \in \{0.10, 0.20, 0.25, 0.30, 0.40, 0.50\}$. The intersection-over-union of predicted and GT sets measures spatial localization. Both sets have

equal cardinality, so missed and extra counts are equal within a field. The four sensor/slot IoUs of each modality are averaged within each window, then equally across the three windows. At 25%, mean IoU is 0.788/0.849 for baseline/ours in vision and 0.681/0.767 in tactile sensing. Mean IoU improves at every evaluated fraction for both modalities; individual fields may still degrade.

Figure 3(a) uses window 50, whereas the decoded-error example in Figure 2(a) uses window 75; all sensors and both future slots are shown without cropping. Visual fields are 16×16, tactile fields 8×8. Blue-gray denotes the GT set; muted green, gold and red denote overlap, missed GT locations and extra predicted locations. Background denotes locations outside both sets. These latent activity regions are not annotated contact regions, object masks or physical displacement fields. The actual context is shared by both models; no prediction-specific context is substituted.

Figure 2(c), fourth panel, compares $\delta z = z_2 - z_1$ with $\delta \hat{z} = \hat{z}_2 - \hat{z}_1$ through their cosine similarity, flattening channels and spatial coordinates separately for each sensor. All norms are nonzero. Values first average the two sensors, then the three windows; connected points show window-level pairs and horizontal marks show means. Vision improves from 0.916 to 0.956 and tactile sensing from 0.887 to 0.946; all three window-level means improve in each modality. The axis starts at 0.78, and 1 denotes aligned directions. This metric omits change amplitude, is geometrically related to normalized squared distance, and does not establish full temporal dynamics or control gains. Thin curves in Figure 2(c), rightmost panel, likewise represent actual windows, not confidence intervals. These training-input diagnostics imply neither spatial independence nor module-level causality.

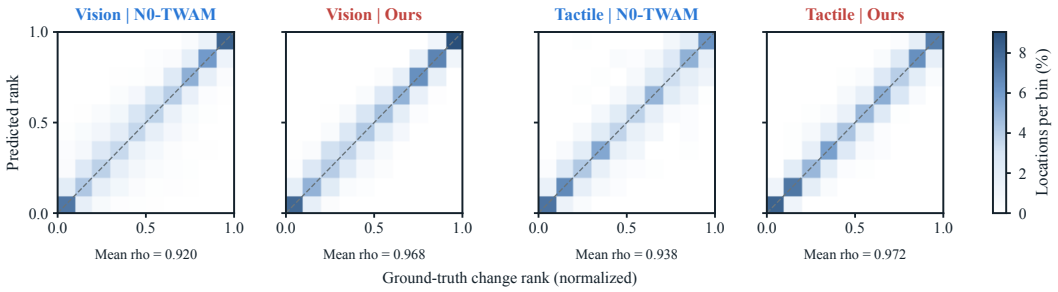

Figure 8: **Spatial change-rank distributions.** GT and predicted ranks for all three fixed training windows, sensors and future slots. Each panel sums to 100% and shares the same bins and linear color scale. The diagonal denotes matching ranks; labels give mean field-wise Spearman correlations. Locations are correlated, not independent trials.

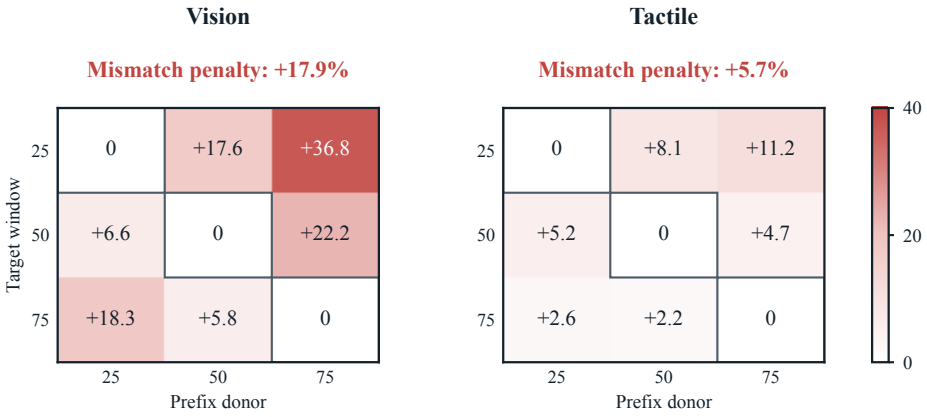

Figure 9: **Conditioning specificity within Ours.** Rows: target windows; columns: prefix donors. Cells give latent-MSE increases relative to the row's matched prefix. The headline penalty averages all six off-diagonal cells; both modalities share a 0–40% scale. Outlined diagonal cells reproduce the original output exactly. This frozen-checkpoint intervention is separate from the baseline comparison: it tests dependence on matching noised-demonstration prefixes, not deployment superiority.

F.3 RANK, SPATIAL GRADIENTS AND COMPLETE PREFIX INTERVENTIONS

Ranking and density. For each sensor and future slot, activity is the channel-mean squared difference between future and actual context latents, computed independently for GT and each prediction. Average ranks resolve ties. Spearman correlation is the Pearson correlation between the two rank vectors; correlations are first averaged over the four sensor/slot fields of each modality, then equally across the three windows. For the density plots in Figure 8, rank $r \in [0, N-1]$ becomes $(r+1/2)/N$. The four joint histograms use identical 10×10 bins on $[0, 1]^2$, without smoothing; each sums to 100%. Vision contributes 3072 locations per method and tactile sensing 768. The shared linear color scale covers every bin without clipping. Spatial locations are correlated, and neither bin counts nor window repetitions constitute independent training seeds.

Spatial-gradient agreement. We take horizontal and vertical first differences of the full channel-wise future-minus-context change field, concatenate them, and compute cosine similarity against the GT gradient vector. Each field has nonzero norm. Aggregation matches the rank statistic. This compares latent spatial structure, not decoded-image edges, physical contact boundaries or action accuracy. Figure 3(b–e) compares all 12 sensor/slot fields on identical horizontal and vertical axes spanning 0.55–1.01. Filled/open circles identify the first/second future slot and the diamond gives the equally weighted overall mean. All first-slot pairs improve in both modalities and metrics; second-slot differences are small and mixed. The following table retains every window-level mean; N0 denotes the post-trained N_0-TWAM baseline, and Grad. denotes gradient cosine similarity. These exploratory statistics share data with Figures 2 and 3; they are correlated descriptions, not independent confirmations or significance tests.

Window	Modality	Rank: N0	Rank: Ours	Grad.: N0	Grad.: Ours
25	Vision	0.9171	0.9695	0.9177	0.9510
25	Tactile	0.9572	0.9651	0.9296	0.9477
50	Vision	0.9262	0.9728	0.9024	0.9449
50	Tactile	0.8977	0.9693	0.8533	0.9122
75	Vision	0.9161	0.9604	0.9083	0.9423
75	Tactile	0.9577	0.9801	0.9230	0.9428

Complete prefix matrix. Slurm job 11774 completed on one H100 with frozen weights and no training. It reused the same checkpoint SHA256, model source hashes, diagnostic seed, dataset positions and native masks as the original cyclic probe. Three intact forwards establish references; nine additional forwards evaluate the Cartesian product of target and donor indices $\{25, 50, 75\}$. Only the copied prefix hidden states are replaced before the action expert's input projection. Assertions preserve all other hidden states and verify that the recipient prefix before replacement equals its original value. Re-inserting the matching prefix gives exactly zero maximum output difference for both modalities at each window. All three cyclic pairs reproduce their recorded errors.

In Figure 9, each matrix cell is $100(M_{ij}/M_{ii} - 1)$, with target i, prefix donor j, and the same two-future-slot latent MSE as the original probe. The aggregate averages all six off-diagonal relative increases equally: 17.8891% for vision and 5.6739% for tactile sensing. Ranges are 5.8228–36.7848% and 2.2031–11.1576%. These differ from the earlier three-pair cycle averages of 19.4%/5.2% because the donor set is larger, not because the original results changed. Both matrices share a 0–40% scale.

Target	Prefix donor	Visual MSE	Tactile MSE
25	25	0.010630	0.014394
25	50	0.012496	0.015566
25	75	0.014540	0.016000
50	25	0.010712	0.023797
50	50	0.010047	0.022625
50	75	0.012278	0.023694
75	25	0.012462	0.017807
75	50	0.011143	0.017733
75	75	0.010530	0.017351

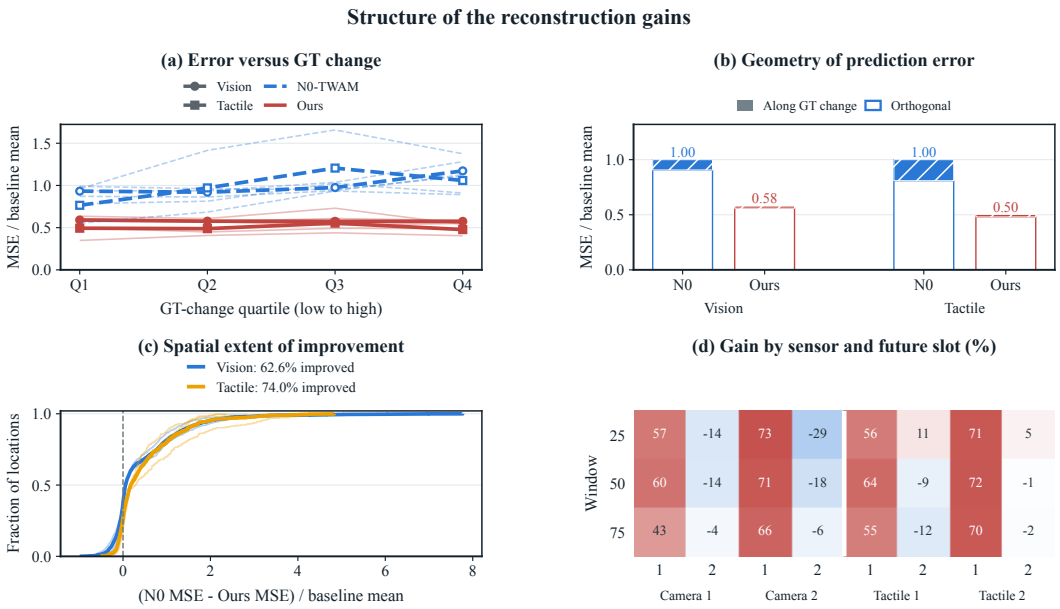

Figure 10: **Where reconstruction improves.** The same three fixed training windows, with no additional samples. (a) Errors across GT-change quartiles; thin curves are individual windows and thick curves are means. Dashed: N_0-TWAM; solid: ours. (b) Exact error decomposition along and orthogonal to GT change. (c) Empirical distributions of local error differences; positive values favor ours. Thin curves are windows, thick curves pool locations. (d) All sensor–future-slot pairs, with relative MSE reductions in percent; blue cells indicate degradation. Panels (a)–(c) share one baseline-MSE denominator within each modality. These are descriptive latent-space diagnostics, not independent-location statistical tests.

The intervention changes an internal action condition containing noised demonstration actions. It establishes within-checkpoint sensitivity to condition matching under this diagnostic, not the benefit of immutable prefixes over separately trained mutable prefixes. It neither compares intervention effects against the baseline architecture nor establishes deployment behavior, feedback adoption or physical safety. The earlier zero-gate and endpoint-error findings remain applicable.

F.4 Definitions for the structure analysis

Figure 10 reuses exactly the paired arrays and three windows above. A field is one sensor at one future latent slot. For each field, let d be GT future minus GT context and e be predicted future minus GT future. Spatial activity is the channel mean of d^2. We sort locations by GT activity independently in each field, use a stable ordering for ties, and partition into four equal-count groups. The same groups evaluate both models. Each quartile value averages channel-mean squared error over its locations and the four fields of a modality; thick lines average the three windows equally. Quartiles express within-field ranks, not common physical displacement thresholds.

For each modality, a single scale M_0 is the baseline MSE averaged over all future coordinates and the three windows. Both models, all quartiles, and both error components share this denominator. Panel (b) decomposes each field by $e_\parallel = \langle e, d\rangle d/\|d\|^2$ and $e_\perp = e - e_\parallel$, giving the exact identity $\|e\|^2 = \|e_\parallel\|^2 + \|e_\perp\|^2$. All fields have nonzero GT-change energy. Components are averaged before division by M_0. The orthogonal term describes latent geometry, not physical noise.

Panel (c) uses $z = (\mathrm{mean}_c\, e_{0,c}^2 - \mathrm{mean}_c\, e_{1,c}^2)/M_0$ at every location, with 0 denoting the baseline and 1 ours. Each window contains four fields per modality: 16×16 for vision and 8×8 for tactile sensing, giving 3072 and 768 locations over three windows, respectively. Thick curves are pooled empirical CDFs, thin curves are window-specific CDFs. Positive fractions are 62.6%/74.0%; these are fractions of correlated latent locations, not trial success rates or independent samples. Full tails and negative differences are retained. Panel (d) instead uses each cell's own baseline

denominator: $100(1 - M_1/M_0)$ for one window/sensor/slot. All 24 cells are shown with a zero-centered symmetric color scale and untruncated values; labels are rounded to whole percentages. Across both modalities, the first future slot improves in 12/12 cells and the second in 2/12. No confidence interval, significance claim or causal module attribution is inferred from these three training windows.

F.5 Direct visualization of latent predictions

Figure 4(a,b) visualizes latent predictions directly. For each modality, the PCA basis is fitted only to GT future latents pooled over the same three fixed training windows and four sensor/slot fields. Channels are centered with their pooled GT means, without per-model fitting or whitening. The leading three eigenvectors form the RGB channels; each eigenvector's largest-magnitude entry is oriented positively. A single per-component affine color range, spanning GT and both predictions across all three windows, is shared by all panels of that modality. No spatial interpolation or model-specific rescaling is applied. The projection retains 53.0% of visual and 49.4% of tactile GT variance, so it is a partial view of the representation, not a substitute for the full-dimensional error.

In the shared three-dimensional PCA subspace, visual MSE decreases from 0.022229 to 0.010355 (53.4%), and tactile MSE from 0.144196 to 0.040309 (72.0%). These values average squared differences of the three orthonormal projection scores over all displayed sensor/slot fields and locations, before the RGB affine color mapping. They quantify the projected representation only and must not be substituted for full-feature MSE.

The error maps use the unprojected 48-channel latents, not PCA colors or decoded RGB. Their means reproduce the original window-50 MSE: 0.018018 versus 0.010047 for vision, and 0.048711 versus 0.022625 for tactile sensing. Figure 2(b) retains the three-window mean reductions of 42.1%/49.7%. This direct feature-space view explains what those metrics measure while keeping them distinct from the decoded-image MAE in Figure 2(a). The improvement remains horizon-dependent; it does not establish uniform decoded-image or closed-loop superiority.

F.6 Multiscale structure, frequency bands and improvement coverage

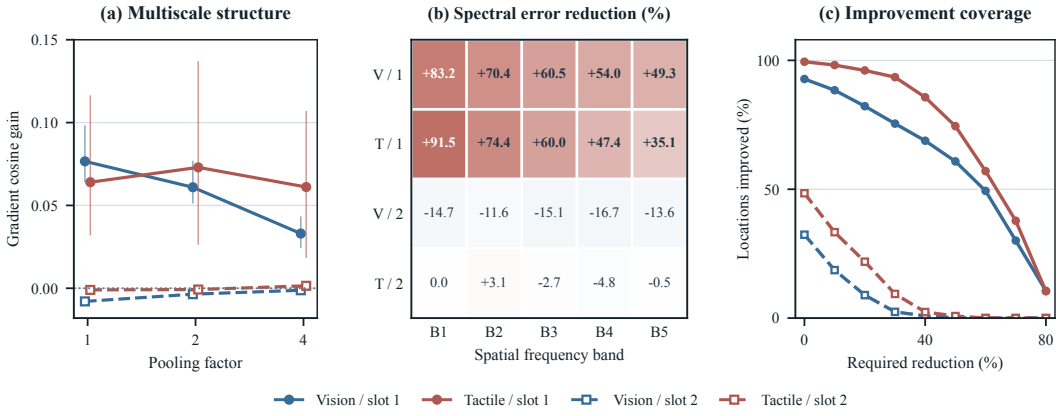

Figure 11: **Where the latent reconstruction gains are distributed.** All three fixed training windows, two sensors per modality and both future slots. V/T denote visual/tactile modalities. (a) Ours minus baseline gradient cosine after spatial average pooling; thin ranges span the three window means for slot 1. (b) Relative spectral error reductions in five radial frequency bands, with positive values favoring Ours. (c) Fraction of spatial locations whose channel-mean squared error decreases by more than the horizontal-axis threshold. Slot 1 improves at 92.8% of visual and 99.5% of tactile locations at threshold zero. Slot 2 is retained to show the horizon dependence. These are descriptive latent-space diagnostics, not decoded-image geometry or independent test trials.

Definitions and aggregation. Figure 11 compares multiscale structure, frequency-band error and the spatial coverage of improvements. We reuse exactly the arrays and matched-input checks from

Appendix F, with no additional training or sample selection. Each modality has two sensors, two future slots and three windows. For pooling factors $k \in \{1, 2, 4\}$, nonoverlapping $k \times k$ spatial blocks are averaged channel-wise. Pooling is applied to the future latent tensors themselves, not to their changes from context. Horizontal and vertical spatial differences are concatenated across channels and space; their cosine with the corresponding GT spatial differences measures structural agreement. This differs from the future-minus-context gradient statistic in Appendix F.3; even at $k = 1$, the two statistics need not agree. Panel (a) averages the paired cosine gains equally over sensors and windows for each future slot. The thin ranges are minimum–maximum window means, not confidence intervals.

For panel (b), an orthonormal two-dimensional discrete Fourier transform is applied to each channel of the prediction residual, separately within each sensor and future slot. Squared coefficient magnitudes are averaged within radial frequency bands with boundaries $0, 1/8, 1/4, 3/8, 1/2, \sqrt{1/2}$ cycles per latent-grid cell; the final band includes its upper endpoint. Thus B1 includes DC and B5 includes diagonal high-frequency coefficients. With $E_{p,b}$ the band-mean error energy, the displayed reduction is $100(1 - \overline{E}_{\mathrm{ours},b}/\overline{E}_{\mathrm{base},b})$, where the bar equally averages sensors and windows. Parseval's identity is checked against the spatial residual energy. No per-model normalization or spectral filtering is applied.

For panel (c), let $e_p(x)$ be squared residual averaged over latent channels at spatial location x. We compute $r(x) = 100(e_{\mathrm{base}}(x) - e_{\mathrm{ours}}(x))/\max(e_{\mathrm{base}}(x), 10^{-15})$ and plot the fraction with $r(x) > \tau$, equally averaging sensors and windows. Spatial locations are correlated; their count is not an independent sample size.

Concentration and extent of the gains. At the first future slot, gradient-cosine gains remain positive at all three pooling scales: $+0.033$ to $+0.077$ for vision and $+0.061$ to $+0.073$ for tactile sensing. Spectral error falls in every evaluated band, by 49.3–83.2% for vision and 35.1–91.5% for tactile sensing. Moreover, 60.8% of visual and 74.5% of tactile locations exhibit more than 50% lower local error. This supports broad first-slot latent improvement across spatial scales and frequency bands. The second slot does not show the same pattern: only 32.4%/48.4% of visual/tactile locations improve at threshold zero, and its spectral changes are mostly unfavorable or small. These results do not establish accurate decoded gripper geometry or attribute the gains to accepted feedback.

G Extended related work and attribution

Action generation and world–action modeling. ACT learns action chunks from demonstrations (Zhao et al., 2023), Diffusion Policy models action distributions by denoising (Chi et al., 2025), and $\pi_{0.5}$ extends vision–language–action learning toward open-world generalization (Intelligence et al., 2025). StarVLA-α, InternVLA-A1 and Xiaomi-Robotics-0 provide further VLA references (Ye et al., 2026b; Cai et al., 2026a;b). LingBot-VA jointly models visual dynamics and actions (Li et al., 2026); GigaWorld-Policy studies an action-centered WAM (Ye et al., 2026a). Fast-WAM explicitly tests whether future generation is needed at inference (Yuan et al., 2026b). Our intervention is at the forecast-to-action interface: we retain prediction but separately expose evidence contribution, residual correction and its acceptance. This distinction is architectural, not evidence that retaining forecasts necessarily improves control.

Tactile representation, prediction and feedback. VITaL and UniVTAC learn visuotactile representations (George et al., 2025; Chen et al., 2026); Tactile-VLA, FTP-1 and N_0-VTLA extend tactile-conditioned policy learning (Huang et al., 2025; Yuan et al., 2026a; NeoteAI Team & Fudan TEAI Team, 2026b). Visuo-Tactile World Models, Dream-Tac and N_0-TWAM study predictive visual and tactile information (Higuera et al., 2026; Lou et al., 2026; NeoteAI Team & Fudan TEAI Team, 2026a). RDP uses a slow–fast policy organization (Xue et al., 2025), whereas TacForcing incorporates execution-time tactile observations during streaming action generation (Zhou et al., 2026). Our predicted-evidence correction is different from refreshing tactile observations; it does not replace that requirement. Oracle-guided interface studies (Yao et al., 2026) are also relevant, but oracle-future results cannot be compared directly with ordinary deployed policies.

Utility, attribution and conservative acceptance. UPN links representations to differentiable planning (Srinivas et al., 2018), TaskSense focuses world models through task-related supervision (Islam & Huber, 2026), and MOPO penalizes model uncertainty in offline policy optimization (Yu et al., 2020). We instead measure demonstration-target error changes relative to a shared base proposal. The coalitional construction inherits the marginal-contribution principle of Shapley attribution (Shapley et al., 1953; Lundberg & Lee, 2017). Its sampled labels are neither exact Shapley values nor physical causal effects. The acceptance head uses classical quantile regression (Koenker & Bassett Jr, 1978); a fitted low quantile alone does not certify joint coverage, safety or closed-loop improvement.

Generative objectives and temporal scheduling. Flow matching supplies our velocity-regression formulation (Lipman et al., 2022). Diffusion Forcing establishes sequence generation with variable token noise levels (Chen et al., 2024). Our mixed-clock sampler operates on world and action streams within that broader scheduling design space; differing clocks are not themselves claimed as a new generative principle. The proposed contribution is their integration with explicit utility and commitment.