TacBootstrap: A Tactile-Informed Copilot for Contact-Rich Bimanual Manipulation from UMI Demonstrations

Anonymous authors
Paper under double-blind review

Abstract

Collecting successful bimanual demonstrations remains difficult when visual teleoperation cannot resolve local contact. We introduce TacBootstrap, a tactile-informed contact copilot trained exclusively on Universal Manipulation Interface (UMI) demonstrations. Both UMI and robot grippers use fingertip Daimon vision-based tactile sensors, supplying shear and depth observations. The operator controls non-contact motion through virtual reality (VR) and manually transfers authority near contact; VR commands then pause while the copilot controls the contact phase. A Slow-Fast encoder combines wrist-image context with three frames of tactile and state observations. Current-state retrieval supplies a demonstrated trajectory anchor, and an anchor-based flow head predicts contact-conditioned residuals to generate bimanual local-pose and gripper action chunks. Receding-horizon execution closes the loop with measured observations. Robot recordings are used only for final evaluation. Across paper-cup separation, cap tightening, and network-cable insertion, TacBootstrap achieves 77.7% task-macro success versus 55.4% for robot VR, with higher success in every task; direct UMI achieves 100%. The macro average of task-level attempts per success decreases from 1.81 to 1.29. Recorded durations are lower but have nonuniform recording boundaries. These system-level results support using portable visual-tactile demonstrations for robot-side contact assistance.

1 Introduction

Robot learning depends on diverse, high-quality demonstrations (Khazatsky et al., 2024; Open X-Embodiment Collaboration, 2023; Mandlekar et al., 2022). In contact-rich bimanual tasks, recording motion is not enough: each collection attempt must achieve physical task success. VR teleoperation lets operators guide both arms during free motion, but near contact, occlusion, delay, and limited haptic feedback make grasping, alignment, and constrained motion difficult. Wrist images provide object and scene context, whereas fingertip vision-based tactile sensing provides complementary local contact information (Dong et al., 2017; Lambeta et al., 2020). This distinction motivates assistance that uses touch where visual teleoperation is most fragile.

UMI decouples demonstration collection from repeated robot operation (Chi et al., 2024); tactile and multimodal extensions enrich portable recordings (Xu et al., 2025; Luo et al., 2026). The resulting question is whether portable visual-tactile demonstrations can train a robot contact copilot without robot training data. We use Daimon vision-based tactile sensors at the fingertips of both the UMI interface and robot grippers, supplying shear and depth inputs in both settings. This provides corresponding tactile modalities across collection and deployment, while leaving embodiment and sensor-instance differences to be addressed explicitly.

TacBootstrap addresses this question through phase-gated contact copiloting. The operator controls non-contact motion through VR, then manually triggers a hard handoff near contact. VR commands pause while the policy controls both arms' local SE(3) motion and gripper adjustment. The human retains task choice, approach, and handoff timing; the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

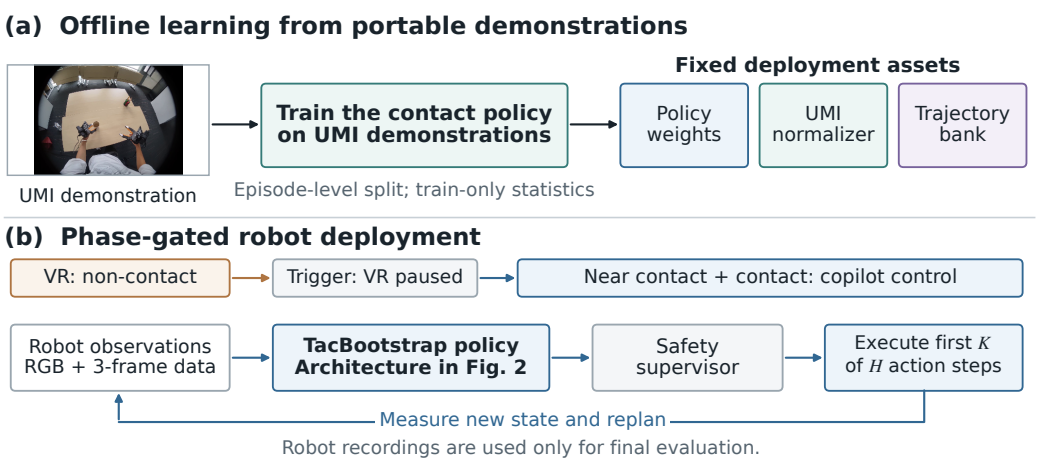


Figure 1: TacBootstrap learning and deployment. (a) Portable UMI demonstrations supply policy weights, normalization statistics, and the trajectory bank; the photograph is a recorded UMI scene. (b) A manual trigger pauses VR near contact and transfers authority to the copilot. Predicted chunks pass through the safety supervisor, the robot executes the first K of H steps, and measured observations close the replanning loop. Robot recordings are used only for final evaluation.

policy handles the bounded contact phase. This is traded control, not continuous action blending or full-task autonomy.

The policy separates scene context from local contact feedback. Its Slow branch encodes bilateral wrist images; its Fast branch combines three frames of Daimon tactile shear/depth with wrench, TCP, and gripper state, without persistent recurrent memory. Current-state retrieval selects a UMI reference trajectory as a motion prior, while anchor-based flow predicts residuals conditioned on the observed contact state. Each chunk contains H bimanual pose-and-gripper steps; the controller executes the first K and replans from measured observations. Policy weights, normalizers, and the trajectory bank use UMI training data only. Robot recordings are reserved for final evaluation, not adaptation or parameter selection.

We compare robot VR, direct UMI, and TacBootstrap on paper-cup separation, cap tightening, and network-cable insertion. The primary endpoint is success per collection attempt. TacBootstrap achieves 77.7% task-macro success versus 55.4% for robot VR, with the same direction in every task; task-level attempts per success decrease from 1.81 to 1.29. Recorded durations are reported descriptively because their boundaries differ. These results evaluate the complete interface, not the causal contribution of tactile sensing alone.

Our contributions are threefold:

- Tactile-informed contact copiloting. We combine fingertip Daimon sensing on UMI and robot grippers with a phase-gated interface that preserves human-controlled approach and assigns learned bimanual control to the near-contact and contact phases.
- Reference-conditioned visual-tactile policy. UMI-only learning combines wrist-image context and three-frame tactile feedback with state-retrieved trajectory anchors and flow-based residuals, generating TCP-local SE(3)-and-gripper action chunks from UMI data alone.
- Collection-oriented evaluation. Three-task comparisons with robot VR and direct UMI show a descriptive 22.3-percentage-point increase in task-macro robot collection success, with explicit collection-yield metrics and a separation between robot evaluation and policy training.

2 Related Work

Portable and scalable demonstrations. Large-scale robot corpora broaden scene and embodiment coverage (Khazatsky et al., 2024; Open X-Embodiment Collaboration, 2023), but do not remove the cost of each failed collection attempt. UMI uses a portable handheld interface, relative trajectories, and latency matching to collect in-the-wild robot demonstrations without operating the robot in every scene (Chi et al., 2024). exUMI adds modular tactile sensing and an action-aware tactile representation (Xu et al., 2025); OmniUMI further records RGB, depth, tactile, grasp-force, trajectory, and wrench signals in a human-aligned multimodal interface (Luo et al., 2026). ACT shows how embodiment-matched low-cost interfaces support precise bimanual collection (Zhao et al., 2023). Our focus is neither a new handheld device nor full-task autonomous imitation. We use portable data to train a robot-side copilot whose authority is restricted to near-contact and contact phases.

Visual-tactile reactive policies. Optical touch sensors provide local contact geometry and motion information (Dong et al., 2017; Lambeta et al., 2020); contact-correlated outputs are not automatically calibrated forces. This distinction matters when portable and robot-mounted sensors share a modality but not a verified physical scale. Reactive Diffusion Policy separates slow visual context from a fast tactile feedback path (Xue et al., 2025). UniReflex and TactileReflex similarly emphasize fast force or tactile correction (Huang et al., 2026; Feng et al., 2026), while TouchWorld combines tactile prediction with reactive refinement (Zhou et al., 2026). ViTaL decomposes reaching and local contact-rich interaction, supporting the broader idea that global visual context and local contact control operate at different scales (Zhao et al., 2025). Building on this separation, TacBootstrap combines UMI-only training, phase-gated robot authority, state-retrieved trajectory anchors, and demonstration-collection endpoints. The contribution is this integrated contact-copiloting formulation, rather than the Slow-Fast decomposition alone. MimicTouch learns from human tactile demonstrations and uses online residual reinforcement learning to bridge embodiment gaps (Yu et al., 2025). Our setting instead excludes robot interaction from parameter learning.

Shared autonomy for data collection. Goal-inference assistance reasons over operator intent (Javdani et al., 2015). Shared autonomy has combined user inputs with learned latent actions or autonomous corrections (Jeon et al., 2020; Yoneda et al., 2023). PATO assists repetitive portions of teleoperation and requests human input when uncertain, improving data-collection scalability (Dass et al., 2023). Recent work also learns corrective copilots for fine-grained contact-rich teleoperation (Sha et al., 2026). Our interface makes a different control choice: the human dominates the approach, then manually initiates a hard handoff for a bounded contact phase. Unlike interactive imitation approaches such as DAgger (Ross et al., 2011), new robot demonstrations are not aggregated into training in our evaluation.

Generative action chunks. ACT and Diffusion Policy model coherent action sequences rather than isolated commands (Zhao et al., 2023; Chi et al., 2023). Rectified flow and flow matching learn transport vector fields (Liu et al., 2023; Lipman et al., 2023); conditional and consistency flow policies extend this idea to manipulation (Chisari et al., 2025; Yan et al., 2025). Our flow decoder operates around a retrieved reference rather than generating an unconstrained chunk. The retrieved UMI trajectory supplies a local anchor, and flow predicts a state-conditioned residual distribution around it.

Retrieval and residual control. VINN predicts actions nonparametrically from nearest demonstrations (Pari et al., 2021). Residual control combines nominal control with learned corrections (Johannink et al., 2018). Our reference is a state-selected UMI action chunk, not a human action stream or handcrafted nominal controller; the residual is learned offline, not through reward-driven robot interaction.

3 Problem Formulation

Let t denote a deployment decision and $m_t \in \{\text{VR}, \text{copilot}\}$ the active authority. Before handoff, $m_t = \text{VR}$ and the robot follows operator commands. At manual trigger time t_s, VR actions are paused and $m_t = \text{copilot}$ until completion, abort, or safety stop.

The copilot observes bilateral wrist images $I_t^{L,R}$ and a causal fast window

$$F_t = \{X_\tau, D_\tau, W_\tau, P_\tau, G_\tau, M_\tau\}_{\tau=t-2}^{t}, \tag{1}$$

where X, D are Daimon fingertip shear and depth channels, W is wrench, P contains both TCP states, G contains gripper states, and M records validity and source masks. TCP denotes the tool center point. No persistent hidden state crosses inference calls.

For future step h, the action is

$$a_{t,h} = \left[\Delta p^L, \Delta e^L, \Delta g^L, \Delta p^R, \Delta e^R, \Delta g^R\right]_{t,h} \in \mathbb{R}^{14}, \tag{2}$$

where translation Δp, Euler-angle rotation Δe, and gripper increment Δg are relative to the current state. Each arm's SE(3) increment is represented in that arm's current TCP frame. All H targets share the measured decision-time reference; they are not successive deltas accumulated across the chunk. The policy predicts $A_t = a_{t,1:H}$, executes $a_{t,1:K}$, and replans. H and K are configurable.

The primary outcome for condition c and task j is attempt-level success $p_{c,j}$. Attempts per success, $1/p_{c,j}$, captures the collection yield. Successful-demonstration duration is the intended efficiency endpoint, excluding failed attempts, reset, and setup. The available recorded durations have different boundaries and are analyzed descriptively, not as end-to-end time-to-success.

4 Method

Figure 1 separates offline learning from phase-gated robot execution. Figure 2 expands the policy into its observation encoder, current-state reference retrieval, and residual flow decoder.

4.1 Fingertip tactile sensing and representation

Both the handheld UMI grippers and robot grippers carry Daimon vision-based tactile sensors at their fingertips. The sensors supply spatial shear and depth channels to the Fast branch; tactile depth is distinct from scene depth in an external camera. Wrench and kinematic states form separate vector inputs. Across interfaces, observations share semantic fields and validity masks, not assumed physical units. Sensor-native tactile values pass through the deployment adapter and use UMI training statistics for normalization; missing channels remain explicit through M_t. A shared sensor family does not establish identical axis conventions, numerical scales, or cross-instance calibration.

Actions use current-state increments rather than absolute poses. This removes dependence on a global world frame. Gripper supervision and deployment commands are separate interfaces: a fixed monotone actuator mapping and hardware bounds convert the normalized model increment to the low-level controller.

Table 1 summarizes the input and control interface.

4.2 Slow-Fast three-frame encoder

The Slow branch applies a shared visual backbone to both wrist images and projects the features into per-hand context vectors $s_t^{L,R}$. DINOv2 (Oquab et al., 2023) provides the visual representation, leaving the fast branch to resolve short-horizon contact changes. For each of the latest three Fast frames, shared tactile encoders process per-finger Daimon shear/depth, while vector encoders process wrench, TCP, gripper, timing, and masks. An order-aware

Table 1: Observation, action, and authority contract. "Training statistics" always means UMI training episodes; robot-test moments are never used.

Component	Temporal role	Representation	Deployment handling
Bilateral wrist RGB	Slow	visual context	causal latest-valid frame
Daimon tactile	Fast, 3 frames	native shear/depth	source mask + UMI norm
Wrench	Fast, 3 frames	per-hand vector	validity + UMI norm
TCP/gripper state	Fast, 3 frames	current local state	explicit timestamps
Trajectory anchor	Reference	retrieved UMI chunk	never a VR command
Policy action	Output	local SE(3)+gripper	H predicted, first K executed
Control authority	Phase state	VR or copilot	manual hard handoff

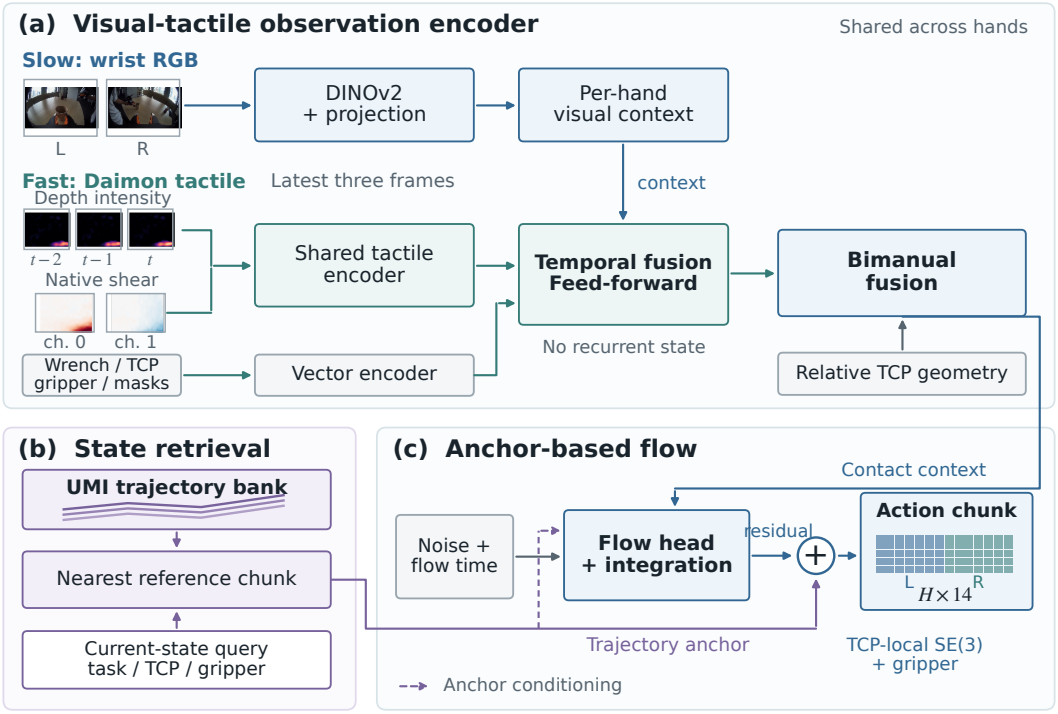


Figure 2: TacBootstrap visual-tactile architecture. (a) Slow wrist-image context conditions an ordered, feed-forward fusion of three Fast frames containing Daimon tactile and state observations; relative TCP geometry supports bimanual fusion. (b) Current task, TCP, and gripper state retrieve a reference chunk from the UMI training bank. (c) Conditional flow integration generates a residual that is added to the same anchor, yielding an $H \times 14$ action chunk. The purple bypass shows anchor addition; the dashed arrow shows anchor conditioning. RGB and tactile thumbnails are recorded UMI inputs; bank curves and the action grid are schematic. Native shear visualization does not establish physical units or axes (Appendix H).

feed-forward block aggregates the encoded frames. The fast representation is conditioned on Slow context and hand identity, then fused across hands with relative TCP geometry. The three-frame representation has no persistent recurrent state: each prediction depends on an explicit causal window, so handoff and restart do not require initializing a recurrent history.

4.3 Current-state retrieval

For each eligible UMI training window, we store a normalized state descriptor r_i and future action chunk A_i. The descriptor contains task identity, bilateral TCP geometry, and gripper

state available at the decision time. Deployment retrieves

$$i_{\text{ret}} = \arg\min_i d(r_t, r_i), \qquad A_t^{\text{anc}} = A_{\text{ret}}, \tag{3}$$

where d is a weighted distance in UMI-normalized state space. Retrieval is restricted to the UMI bank. A state-specific anchor retains a demonstrated action pattern without requiring the operator to supply a concurrent VR command.

The bank contains training episodes only, never validation or robot-test trajectories. The retrieval protocol excludes the target episode during training to prevent copying its supervised future, and applies task compatibility before ranking geometry and gripper descriptors. Each candidate future is re-expressed relative to its own query TCP state, yielding an action pattern rather than an absolute pose in the recorded scene. Distance weights and bank construction are fixed before robot evaluation.

4.4 Anchor-based flow action head

Let Y_t be the demonstrated chunk in normalized local action coordinates and $R_t = Y_t - A_t^{\text{anc}}$ its residual in that coordinate chart. For $\epsilon \sim \mathcal{N}(0, I)$ and $\tau \sim \mathcal{U}(0, 1)$, define $z_\tau = (1-\tau)\epsilon + \tau R_t$ and optimize

$$\mathcal{L}_{\text{flow}} = \mathbb{E}\left[\sum_{h=1}^{H} w_h \left\| v_\theta(z_\tau, \tau, c_t, A_t^{\text{anc}})_h - (R_t - \epsilon)_h \right\|_2^2\right], \tag{4}$$

where c_t is the Slow-Fast encoding. Numerical integration yields $\widehat{R}_t$ and $\widehat{A}_t = A_t^{\text{anc}} + \widehat{R}_t$. The loss covers all H valid steps with nonnegative weights w_h emphasizing the earliest executed portion; padded steps are masked. Coordinate-chart addition is not exact SO(3) composition. Denormalized local rotation coordinates are converted to rotation matrices before composing with the measured TCP pose. The Euler convention and angular range belong to the deployment contract. Anchor retrieval provides a demonstrated continuation, while the residual decoder conditions its correction on the measured contact state; their separate effects require ablation.

4.5 Training and deployment boundary

Training uses processed UMI clips containing aligned observations and future action targets. Data checks preserve source streams and verify numeric conversion; missing shear remains explicit rather than being treated as a measured zero. Splits are at episode level, so overlapping windows from one recording cannot enter different splits. Normalizers are fitted on training episodes only. Robot recordings are excluded from training, adaptation, selection, and threshold setting. The trajectory bank uses the same training partition, making the learning and retrieval boundaries consistent.

At deployment, a supervisor accepts commands only after handoff. An external safety layer clips workspace, incremental motion, gripper, velocity, and force commands; checks packet freshness; and exposes hold, abort, watchdog, and emergency-stop paths. These safeguards are not learned and do not establish safety outside the tested envelope.

4.6 Phase transition and receding-horizon execution

The runtime separates VR, copilot enabled, and hold/abort. A manual trigger pauses VR, snapshots bilateral TCP and gripper states, and validates the three-frame observation window. Failed freshness, validity, workspace, or actuator checks resolve to hold. Each copilot cycle encodes the latest causal observations, retrieves a UMI anchor, integrates the flow residual, and passes the first K of H predicted steps through the safety layer. Replanning uses the measured post-execution state rather than the previous prediction, allowing observed slip, compliance, or saturation to enter the next decision. Completion, operator abort, stale packets, or a safety event terminates copilot authority. Only the active authority can issue robot commands.

5 Experiments

5.1 Research questions

The evaluation addresses three questions: whether contact copiloting improves robot collection success over VR, whether the improvement is consistent across tasks, and how collection yield and recorded duration compare with direct UMI. Success is the primary endpoint; duration is a descriptive endpoint with the recording-boundary qualifications below.

5.2 Tasks and collection conditions

We evaluate three contact-rich tasks. Paper-cup separation requires separating nested cups without crushing or dropping them. Cap tightening requires aligning and tightening a cap to the task-defined terminal state. Network-cable insertion requires aligning and fully seating a connector without damaging the cable or port. The same operator, task objects, and task-level criteria are used across interfaces; attempt counts may differ.

Robot VR uses bilateral VR teleoperation throughout. UMI records the human task with the portable interface. TacBootstrap uses VR for non-contact motion and a manual hard handoff near contact; VR commands remain paused until termination. Hardware photographs appear in Appendix I.

5.3 Attempt protocol and independence

An attempt begins when the operator starts the task-specific motion and ends at task success, declared failure, abort, or the task timeout. The success criteria are interface independent: the resulting physical task state, rather than trajectory similarity or model confidence, determines the label. The operator and objects are held fixed across interfaces. We allow unequal attempt counts because the estimand is a condition-specific probability, but do not treat control ticks, flow samples, or overlapping observation windows as independent replicates.

The phase-gated protocol defines trigger time and copilot termination as separate events, distinguishing the human-controlled approach from the learned contact phase. The attempt remains the sampling unit: a failed handoff, safety stop, or operator abort is not removed as a policy warm-up error. This inclusion rule avoids conditioning success on the copilot having reached a favorable state.

Success labels cover all attempts, including the UMI condition, for which every attempt succeeded. Recorded durations require a separate qualification: the VR summary includes successful and failed executions, and UMI segments use partially different start and end events. We retain these measured summaries without relabeling them as a harmonized successful-demonstration endpoint. Setup and reset are not measured by these summaries, so neither end-to-end throughput nor total operator time per successful demonstration is inferred.

5.4 Metrics and statistical analysis

We report attempt-level success and compute attempts per success as its reciprocal. Table 2 reports point estimates uniformly across conditions; available uncertainty and duration distributions are provided in the appendix. Cross-task results are equal-task macro averages rather than pooled attempts, so a task with more records does not dominate the comparison. The macro average of reciprocal success rates is not the reciprocal of macro-average success.

Available robot VR denominators permit Wilson 95% intervals (Wilson, 1927), reported in Appendix E. UMI and copilot denominators and copilot duration variability are unavailable, precluding comparable uncertainty estimates. We report descriptive percentage-point differences and duration ratios, without significance claims or denominators inferred from rounded rates.

Table 2: Three-task collection outcomes, reported uniformly as point estimates. Success covers all attempts, including UMI; duration is the recorded mean in seconds. Recording boundaries differ, so duration comparisons are descriptive. Uncertainty reporting is discussed in Appendix E.

Task	Interface	Success (%)	Mean duration (s)	Attempts/success
Paper-cup separation	Robot VR	60.0	58.97	1.67
	TacBootstrap	80.0	34.97	1.25
	UMI	100.0	10.96	1.00
Cap tightening	Robot VR	56.3	82.30	1.78
	TacBootstrap	78.1	60.82	1.28
	UMI	100.0	39.34	1.00
Network-cable insertion	Robot VR	50.0	50.43	2.00
	TacBootstrap	75.0	28.51	1.33
	UMI	100.0	6.59	1.00

Robot VR versus TacBootstrap is the primary robot-side comparison. Direct UMI is a portable human-execution reference, not an autonomous robot baseline. The comparison evaluates the collection interface, not isolated model components.

6 Results

6.1 Task-level outcomes

Table 2 reports the three-task comparison. TacBootstrap achieves 80.0%, 78.1%, and 75.0% success for paper-cup separation, cap tightening, and network-cable insertion, respectively. Corresponding robot VR rates are 60.0%, 56.3%, and 50.0%. Thus the observed differences are 20.0, 21.8, and 25.0 percentage points, with the same direction in every task. UMI succeeds in all collection attempts for each task.

The largest observed difference occurs in network-cable insertion, which also has the lowest robot VR success. This pattern is consistent with the motivation for assistance during constrained contact, but does not identify a failure mechanism or isolate a policy component. The statistical reporting scope in Section 5.4 limits inferential comparisons.

6.2 Recorded task durations

Recorded mean durations for TacBootstrap are 34.97, 60.82, and 28.51 s, compared with 58.97, 82.30, and 50.43 s for robot VR. These are 40.7%, 26.1%, and 43.5% lower, respectively. Direct UMI has the shortest recorded duration in each task. Because recording boundaries differ, this ordering describes the logs rather than establishing an equivalent reduction in successful-demonstration collection time.

Reference duration distributions are provided in Table 4 in the appendix. Without a copilot duration distribution, the comparison cannot determine whether its lower mean also reduces variability or the long-duration tail.

6.3 Cross-task summary

Table 3 reports equal-task macro averages. TacBootstrap reaches 77.7% success compared with 55.4% for robot VR, a difference of 22.3 percentage points. Task-level attempts per success decrease from 1.81 to 1.29, while UMI requires 1.00. This inverse-rate measure summarizes collection yield; it is derived from success and is not independent corroborating evidence. Recorded mean duration decreases from 63.90 to 41.43 s, a descriptive 35.2% difference. UMI's corresponding mean is 18.97 s.

The outcomes evaluate the complete collection interface rather than isolated policy components. Task-specific mechanisms remain hypotheses (Appendix G).

Table 3: Equal-task macro averages. Each task contributes equally; attempts per success is averaged after taking the task-level reciprocal. Duration comparisons retain the recording-boundary qualification in Table 2.

Interface	Success (%)	Duration (s)	Attempts/success
Robot VR	55.4	63.90	1.81
TacBootstrap	77.7	41.43	1.29
UMI	100.0	18.97	1.00

7 Discussion and Limitations

Phase gating. The design assigns distinct roles to the operator and policy. VR preserves task choice and long-range adaptability; the learned policy concentrates on local contact, where fingertip observations complement ambiguous visual feedback. A hard handoff gives every executed command one source, but an early, late, or poorly initialized transition can invalidate a competent policy.

Retrieval and flow. Retrieval supplies a demonstrated trajectory near the current state, reducing the burden of generating a bimanual chunk from scratch. Daimon tactile history supplies contact observations for the flow residual, complementing the geometric anchor. The components also create separable failure modes: retrieval can return an irrelevant anchor, and flow can over-correct a good one.

Cross-device touch. Using Daimon sensors on both interfaces supplies a common tactile modality, not proof of matching physical units or axis conventions. Sensor mounting and instance-specific response still require a verified adapter; unpaired marginal distributions alone cannot establish calibration.

Limitations. The comparison uses one operator and three tasks, limiting conclusions about other operators, objects, and contact conditions. Copilot and UMI denominators are absent from the available summaries, so uncertainty in their success rates and between-method differences is not quantified. Duration boundaries differ and do not measure end-to-end throughput or workload. The reported architecture also requires a version-matched deployment record for independent replication (Appendix C). Matched ablations are needed to isolate tactile, history, retrieval, and decoder effects (Appendix D). Manual triggering and a bounded safety envelope limit autonomy; unknown tactile units limit claims of calibrated cross-device transfer.

8 Conclusion

TacBootstrap uses portable visual-tactile UMI demonstrations to train a tactile-informed copilot for robot-side contact assistance. Daimon fingertip observations complement wrist-image context, while state-retrieved anchors and flow residuals generate bimanual pose-and-gripper chunks after a manual VR handoff. Across three tasks, task-macro success increases from 55.4% with robot VR to 77.7%, and the macro average of task-level attempts per success decreases from 1.81 to 1.29. The evaluated system supports a division of labor between human-guided approach and learned contact control without robot-data policy training. Broader conclusions require quantified uncertainty, harmonized timing, matched component ablations, and verified cross-device calibration.

References

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In Robotics: Science and Systems, 2023. URL https://arxiv.org/abs/2303.04137.

Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. In Robotics: Science and Systems, 2024. URL https://arxiv.org/abs/2402.10329.

Eugenio Chisari, Nick Heppert, Max Argus, Tim Welschehold, Thomas Brox, and Abhinav Valada. Learning robotic manipulation policies from point clouds with conditional flow matching. In Proceedings of the 8th Conference on Robot Learning, volume 270 of Proceedings of Machine Learning Research, pp. 982–993, 2025. URL https://proceedings.mlr.press/v270/chisari25a.html.

Shivin Dass, Karl Pertsch, Hejia Zhang, Youngwoon Lee, Joseph J. Lim, and Stefanos Nikolaidis. PATO: Policy assisted teleoperation for scalable robot data collection. In Robotics: Science and Systems, 2023. doi: 10.15607/RSS.2023.XIX.013. URL https://roboticsproceedings.org/rss19/p013.html.

Siyuan Dong, Wenzhen Yuan, and Edward Adelson. Improved GelSight tactile sensor for measuring geometry and slip. In IEEE/RSJ International Conference on Intelligent Robots and Systems, 2017. doi: 10.1109/IROS.2017.8202149. URL https://arxiv.org/abs/1708.00922.

Ziyan Feng, Yulong Fu, Zheng Li, Yuxin He, Jieji Ren, Yudong Zhong, Lujia Wang, Jinni Zhou, and Qiang Nie. TactileReflex: Noise-statistics-driven vision-tactile reflex control for force-sensitive manipulation. In IEEE/RSJ International Conference on Intelligent Robots and Systems, 2026. URL https://arxiv.org/abs/2605.23568.

Yan Huang, Shoujie Li, Ziwu Song, and Wenbo Ding. UniReflex: Plug-and-play force control for pretrained generative policies via fast-slow reflex. arXiv preprint arXiv:2608.17432, 2026. URL https://arxiv.org/abs/2608.17432.

Shervin Javdani, Siddhartha S. Srinivasa, and J. Andrew Bagnell. Shared autonomy via hindsight optimization. arXiv preprint arXiv:1503.07619, 2015. URL https://arxiv.org/abs/1503.07619.

Hong Jun Jeon, Dylan P. Losey, and Dorsa Sadigh. Shared autonomy with learned latent actions. In Robotics: Science and Systems, 2020. doi: 10.15607/RSS.2020.XVI.011. URL https://roboticsproceedings.org/rss16/p011.html.

Tobias Johannink, Shikhar Bahl, Ashvin Nair, Jianlan Luo, Avinash Kumar, Matthias Loskyll, Juan Aparicio Ojea, Eugen Solowjow, and Sergey Levine. Residual reinforcement learning for robot control. arXiv preprint arXiv:1812.03201, 2018. URL https://arxiv.org/abs/1812.03201.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, et al. DROID: A large-scale in-the-wild robot manipulation dataset. In Robotics: Science and Systems, 2024. doi: 10.15607/RSS.2024.XX.120. URL https://roboticsproceedings.org/rss20/p120.html.

Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, Dinesh Jayaraman, and Roberto Calandra. DIGIT: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. IEEE Robotics and Automation Letters, 2020. doi: 10.1109/LRA.2020.2977257. URL https://arxiv.org/abs/2005.14679.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In International Conference on Learning Representations, 2023. URL https://arxiv.org/abs/2210.02747.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In International Conference on Learning Representations, 2023. URL https://arxiv.org/abs/2209.03003.

Shaqi Luo, Yuanyuan Li, Youhao Hu, Chenhao Yu, Chaoran Xu, Jiachen Zhang, Guocai Yao, Tiejun Huang, Ran He, and Zhongyuan Wang. OmniUMI: Towards physically grounded robot learning via human-aligned multimodal interaction. arXiv preprint arXiv:2604.10647, 2026. URL https://arxiv.org/abs/2604.10647.

Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In Proceedings of the 5th Conference on Robot Learning, volume 164 of Proceedings of Machine Learning Research, pp. 1678–1690, 2022. URL https://proceedings.mlr.press/v164/mandlekar22a.html.

Open X-Embodiment Collaboration. Open X-embodiment: Robotic learning datasets and RT-X models. arXiv preprint arXiv:2310.08864, 2023. URL https://arxiv.org/abs/2310.08864.

Maxime Oquab, Timothee Darcet, Theo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaa El-Nouby, et al. DINOv2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193, 2023. URL https://arxiv.org/abs/2304.07193.

Jyothish Pari, Nur Muhammad Shafiullah, Sridhar Pandian Arunachalam, and Lerrel Pinto. The surprising effectiveness of representation learning for visual imitation. arXiv preprint arXiv:2112.01511, 2021. URL https://arxiv.org/abs/2112.01511.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, volume 15 of Proceedings of Machine Learning Research, pp. 627–635, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

Shuo Sha, Yixuan Wang, Binghao Huang, Antonio Loquercio, and Yunzhu Li. Efficient and reliable teleoperation through real-to-sim-to-real shared autonomy. arXiv preprint arXiv:2603.17016, 2026. URL https://arxiv.org/abs/2603.17016.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953. URL https://www.jstor.org/stable/2276774.

Yue Xu, Litao Wei, Pengyu An, Qingyu Zhang, and Yong-Lu Li. exUMI: Extensible robot teaching system with action-aware task-agnostic tactile representation. In Proceedings of the 9th Conference on Robot Learning, volume 305 of Proceedings of Machine Learning Research, pp. 2536–2554, 2025. URL https://proceedings.mlr.press/v305/xu25e.html.

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. In Robotics: Science and Systems, 2025. URL https://arxiv.org/abs/2503.02881.

Ge Yan, Jiyue Zhu, Yuquan Deng, Shiqi Yang, Ri-Zhao Qiu, Xuxin Cheng, Marius Memmel, Ranjay Krishna, Ankit Goyal, Xiaolong Wang, and Dieter Fox. ManiFlow: A general robot manipulation policy via consistency flow training. In Proceedings of the 9th Conference on Robot Learning, volume 305 of Proceedings of Machine Learning Research, pp. 2268–2293, 2025. URL https://proceedings.mlr.press/v305/yan25a.html.

Takuma Yoneda, Luzhe Sun, Ge Yang, Bradly C. Stadie, and Matthew R. Walter. To the noise and back: Diffusion for shared autonomy. In Robotics: Science and Systems, 2023. doi: 10.15607/RSS.2023.XIX.014. URL https://roboticsproceedings.org/rss19/p014.html.

Kelin Yu, Yunhai Han, Qixian Wang, Vaibhav Saxena, Danfei Xu, and Ye Zhao. MimicTouch: Leveraging multi-modal human tactile demonstrations for contact-rich manipulation. In Proceedings of the 8th Conference on Robot Learning, volume 270 of Proceedings of Machine Learning Research, pp. 4844–4865, 2025. URL https://proceedings.mlr.press/v270/yu25c.html.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In Robotics: Science and Systems, 2023. URL https://arxiv.org/abs/2304.13705.

Zifan Zhao, Siddhant Haldar, Jinda Cui, Lerrel Pinto, and Raunaq Bhirangi. Touch begins where vision ends: Generalizable policies for contact-rich manipulation. arXiv preprint arXiv:2506.13762, 2025. URL https://arxiv.org/abs/2506.13762.

Jianyi Zhou, Feiyang Hong, Yunhao Li, Yicheng Zhao, Yongjue Cen, Zirui Liu, Jiakang Huang, Zirui Chen, Ruiyang Zhang, Weizhuo Zhu, Xuhua Song, and Shuo Yang. TouchWorld: A predictive and reactive tactile foundation model for dexterous manipulation. arXiv preprint arXiv:2607.07287, 2026. URL https://arxiv.org/abs/2607.07287.

AI use statement

Generative AI tools assisted manuscript organization, language editing, code review, figure layout, and summary calculations during drafting. The authors confirmed the reported experimental summaries and remain responsible for checking all claims, citations, statistical analyses, and artifacts.

Ethics statement

Robot experiments use independent command limits, watchdog behavior, manual enable, and emergency-stop controls. The learned policy is not a safety controller. The present study evaluates one collection interface with one operator rather than drawing population-level conclusions; an expanded user study will require the applicable consent and institutional review process.

Reproducibility statement

We provide detailed descriptions of our method. We will release the source code, dataset upon publication, subject to any applicable dataset and licensing restrictions.

A Metric definitions

For task j and condition c, let $S_{c,j}$ and $N_{c,j}$ denote successful and total attempts. The reported rate is $\hat{p}_{c,j} = S_{c,j}/N_{c,j}$. Wilson intervals use the attempt as sampling unit. The task-macro rate is $J^{-1}\sum_j \hat{p}_{c,j}$ rather than $\sum_j S_{c,j}/\sum_j N_{c,j}$. Attempts per success is $1/\hat{p}_{c,j}$ and is undefined when labels are absent. Successful-demonstration duration is measured only for successful attempts from the agreed start event to task success. Attempts per success is a derived inverse rate, not an independent endpoint. Its interpretation as a mean waiting time assumes stable, independent Bernoulli attempts; clustered sessions instead require a session-aware analysis. A 100% sample rate does not imply a zero-width confidence interval.

B Operational failure taxonomy

Perception failures include missing or stale sensor packets. Retrieval failures select a reference incompatible with current contact geometry. Generation failures produce an unsuitable correction despite a reasonable anchor. Execution failures arise from saturation, tracking error, or the external safety layer. Authority-transition failures occur when the operator triggers too early or too late. These categories are a reporting protocol, not measured

failure frequencies. They distinguish improvements to sensing, retrieval, policy, low-level control, and the interface.

C Deployment configuration and implementation scope

Figures 1 and 2 describe the three-frame, state-retrieval method. A reproducible deployment record associates robot trials with the model weights, architecture revision, normalizer, and trajectory bank. It also specifies H, K, flow integration steps, retrieval distance weights, Euler convention, actuator mapping, and safety limits. These settings are configuration parameters rather than universal constants of the method.

The audited local prototype is a different variant: it uses a temporal GRU, a global medoid anchor, and rotation-vector actions. Its offline evaluations are not used as evidence for the three-frame method. A version-matched checkpoint and robot-run manifest are not included in the present evidence package, which limits independent verification of the architecture–result link. The available Task1-S2 data audit covers 217 recordings and 17,648 frames; these numbers describe that subset, not the full three-task training corpus.

D Mechanism and cross-device validation

Removing tactile input tests information beyond vision and proprioception. Reducing three frames to one tests short temporal context without recurrent state. Replacing state retrieval with a task-level anchor tests geometric conditioning. Replacing flow with a deterministic residual head tests generative decoding. Training data, safety limits, handoff protocol, and robot trial allocation must remain matched. The current interface comparison does not include these component ablations.

Unpaired sensor distributions cannot identify shear-axis rotation, gain, offset, or depth calibration. A verified mapping requires paired raw frames or known physical stimuli, separate from final-test outcome recordings. Unitless visual normalization is not evidence of a shared physical normalizer.

E Statistical reporting scope

Success rates and copilot mean durations are reported from the authors' experimental summaries. UMI success covers all attempts, not a filtered set of successful demonstrations. Reference duration distributions are available for VR and UMI, whereas copilot duration variance and the copilot/UMI task-level denominators are not provided. We consequently report neither a copilot success interval nor a significance test for its difference from VR. Rounded success rates cannot uniquely determine an integer attempt denominator.

For robot VR, the two-sided Wilson 95% success intervals are [38.7, 78.1]%, [33.2, 76.9]%, and [29.9, 70.1]% for paper-cup separation, cap tightening, and network-cable insertion, respectively. Recorded duration standard deviations are 19.70, 20.90, and 12.07 s for robot VR and 3.41, 9.03, and 3.79 s for UMI, in the same task order. Table 2 shows only point estimates to maintain a consistent presentation across conditions.

For a matched-duration analysis, all interfaces require a common start event and task-success event. Existing VR durations mix successful and failed attempts, and UMI recording segments use different boundaries. Multiplying duration by attempts per success would therefore not recover wall-clock time to one successful demonstration. Session-level timing, failed-attempt durations, setup, and reset would be needed for that separate endpoint.

F Reference duration distributions

Table 4 gives distributional summaries behind the reference means. Robot VR interquartile ranges are above the corresponding UMI ranges in all tasks. The UMI network-cable maxi-

Table 4: Reference duration distributions in seconds. Values are source-defined valid-record durations, not a harmonized successful-attempt endpoint.

Task	Interface	Mean	Median	Q1	Q3	P95	Range
Paper-cup separation	UMI	10.96	9.83	8.48	13.16	17.71	5.72–23.59
	Robot VR	58.97	58.85	47.94	70.04	83.72	21.94–103.66
Cap tightening	UMI	39.34	39.16	34.46	44.17	54.36	5.54–74.18
	Robot VR	82.30	84.96	73.67	92.24	112.04	43.36–120.73
Network-cable insertion	UMI	6.59	5.90	4.94	7.10	9.72	3.56–44.53
	Robot VR	50.43	48.35	40.37	60.28	67.72	33.93–74.76

mum is much larger than its 95th percentile, motivating robust summaries in addition to the mean. These distributional observations do not resolve the recording-boundary differences.

G Task-specific interpretation

Paper-cup separation. Robot VR has the highest observed success among the three VR task references, but its Wilson interval remains broad. Copilot success is 20.0 percentage points higher, and recorded mean duration is 40.7% below the VR mean. The task probes compliant-object handling and bilateral coordination. Tactile feedback may help avoid over-grasping during separation, but crushing, dropping, and incomplete separation were not independently quantified; this mechanism remains a hypothesis.

Cap tightening. This task has the longest VR and UMI durations and the widest observed robot spread. The copilot recorded duration lies 26.1% below the VR mean. Because the operation contains alignment followed by sustained rotational contact, a single success label can hide whether failure occurred before thread engagement or during tightening. Phase-specific failure tags and trigger timing are therefore particularly important for distinguishing failure mechanisms.

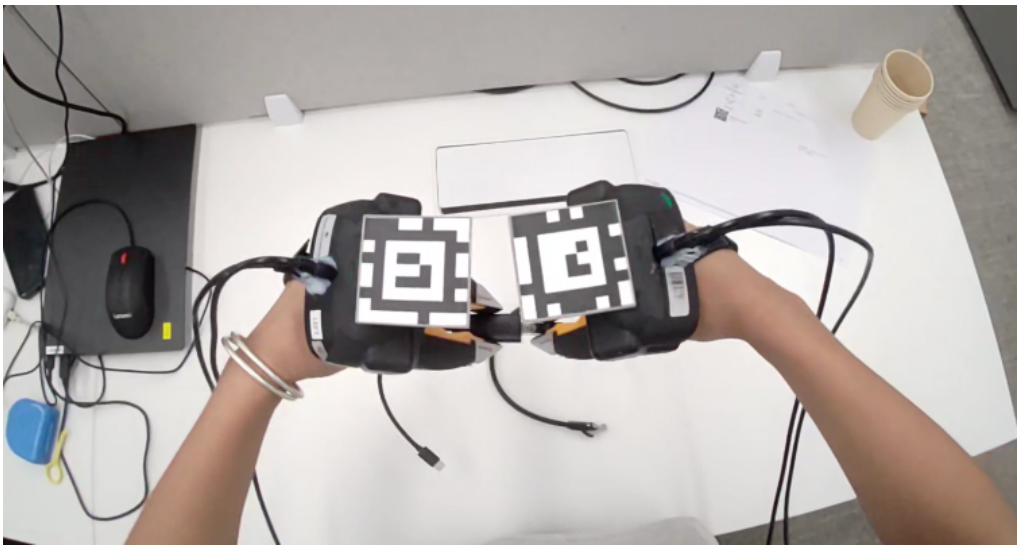
Network-cable insertion. Robot VR has a 50.0% observed success rate, while UMI records are shortest in this task. The copilot recorded duration is 43.5% below the VR mean. Insertion can fail through position error, orientation error, obstruction, or premature force; these mechanisms place different demands on retrieval and fast tactile feedback. The success criterion distinguishes a cleanly seated connector from partial insertion rather than relying only on motion completion.

H Figure data provenance

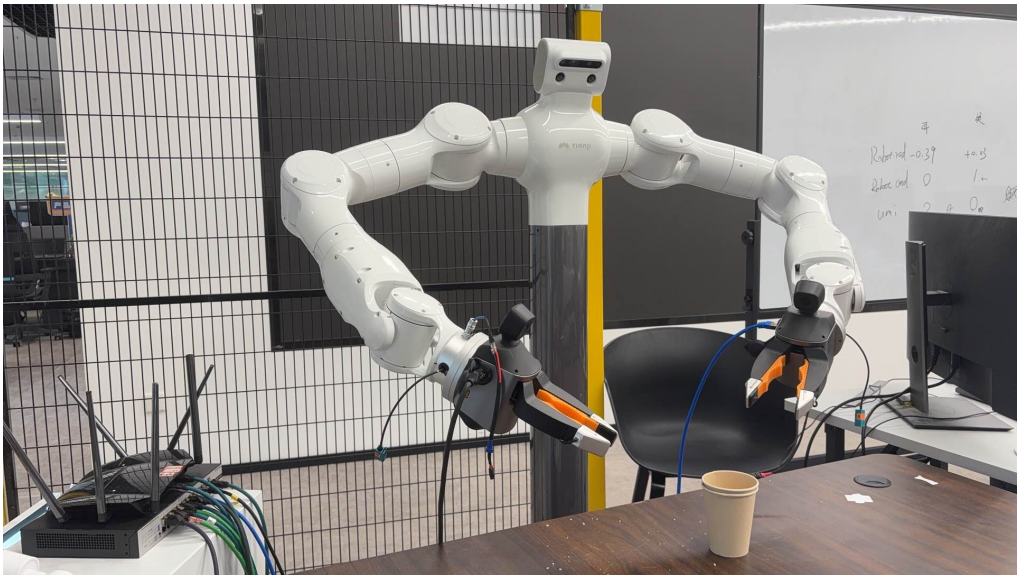
The recorded illustrations in Figures 1 and 2 use UMI episode 000003, with current clip frame 74. The tactile history uses clip frames 71, 73, and 74, selected by the existing causal 50-Hz cache from the 60-Hz export. The ego image displays the left view of the recorded side-by-side pair. Image aspect ratios are retained. Depth thumbnails show cached video intensity with one shared display range; the two shear thumbnails show native channels from the reference-inferred decoding cache, with a common symmetric display range. These display choices do not fit model normalizers or identify tactile axes or physical units. The bank curves and action grid are conceptual illustrations rather than measured retrieval or policy outputs. The figure manifest records source frames, file hashes, display ranges, and decoding status.

I Experimental hardware

Figure 3 shows the two experimental setups. The portable bilateral UMI interface is used for human demonstration collection, which supplies the policy's training data. The bimanual robot platform is used for robot VR and TacBootstrap evaluation. Both interfaces use Daimon vision-based tactile sensors at the gripper fingertips to supply shear/depth inputs; robot recordings are not used for training.



(a) Portable bilateral UMI demonstration interface



(b) Bimanual robot evaluation platform

Figure 3: Experimental hardware. (a) Handheld UMI demonstration interface. (b) Bimanual robot platform for VR and TacBootstrap evaluation. Both use Daimon vision-based tactile sensors at the gripper fingertips as tactile inputs. Author-provided photographs retain their original aspect ratios without cropping or content modification.