TacClaw-Infinity: A Large-Scale Vision-Tactile-Language-Action Dataset for Dexterous Manipulation

Fucai Zhu[1,3,*], Xidong Zhang[1,3,*], Yichi Zhang[1,*], Yipan Wang[5], Fubo Wang[3], Zibin Wu[6], Kaiqi Li[7], Jianhao Jiao[8], Shihan Lu[9], Di Hu[10], Mengdi Xu[11], Siyu Zhu[12], Xiaoxiao Long[2], Yao Yao[2], Xun Cao[2], Zhuo Xu[13], Koichi Hashimoto[4], Qifeng Chen[14], Zhe Li[7], Jianfei Yang[7], Haofei Lu[15], Danica Kragic[15], Shuicheng Yan[16], Jianghua Duan[1], Michael Yu Wang[1,3], Weihao Yuan[1,2,✉]

[1]Daimon Robotics [2]Nanjing University [3]Great Bay University
[4]Tohoku University [5]Southern University of Science and Technology
[6]The Hong Kong University of Science and Technology (Guangzhou)
[7]Nanyang Technological University [8]University College London [9]Northwest University
[10]Renmin University of China [11]Tsinghua University [12]Fudan University
[13]DeepMind [14]The Hong Kong University of Science and Technology
[15]Royal Institute of Technology [16]National University of Singapore

[*] Equal Contributions, ✉ Corresponding Author

Large-scale robot learning increasingly relies on diverse human demonstrations, yet tactile manipulation data remain difficult to scale, and existing multimodal datasets often lack consistent quality control and structured supervision. We present **TacClaw-Infinity**, a large-scale structured Vision-Tactile-Language-Action (VTLA) dataset for contact-rich manipulation, collected with **TacClaw**, a tactile-enabled UMI-style handheld interface that records synchronized visual observations, fingertip tactile signals, and end-effector trajectories without requiring a target robot during data collection. TacClaw-Infinity is built through distributed collection across homes, laboratories, industrial workplaces, retail stores, and outdoor environments including tasks contributed by academic and industrial collaborators. Together, these efforts yield over **10,000 hours**, 1 million episodes, and **1 billion frames** spanning more than **1,000 task types**. We develop a scalable pipeline for trajectory reconstruction, multimodal quality control, embodiment-aware filtering, and structured annotation, covering contact-centric segmentation, subtask description, and spatial grounding. We validate TacClaw-Infinity at both the data and model levels: At the data level, we assess the collection interface, the action-space compatibility of the curated demonstrations, and the structured annotations, while for the model level, we demonstrate cross-embodiment transfer across three heterogeneous robots without robot-specific demonstrations. The code and dataset will be made public.

Project • **GitHub** • **ModelScope** • **Hugging Face**

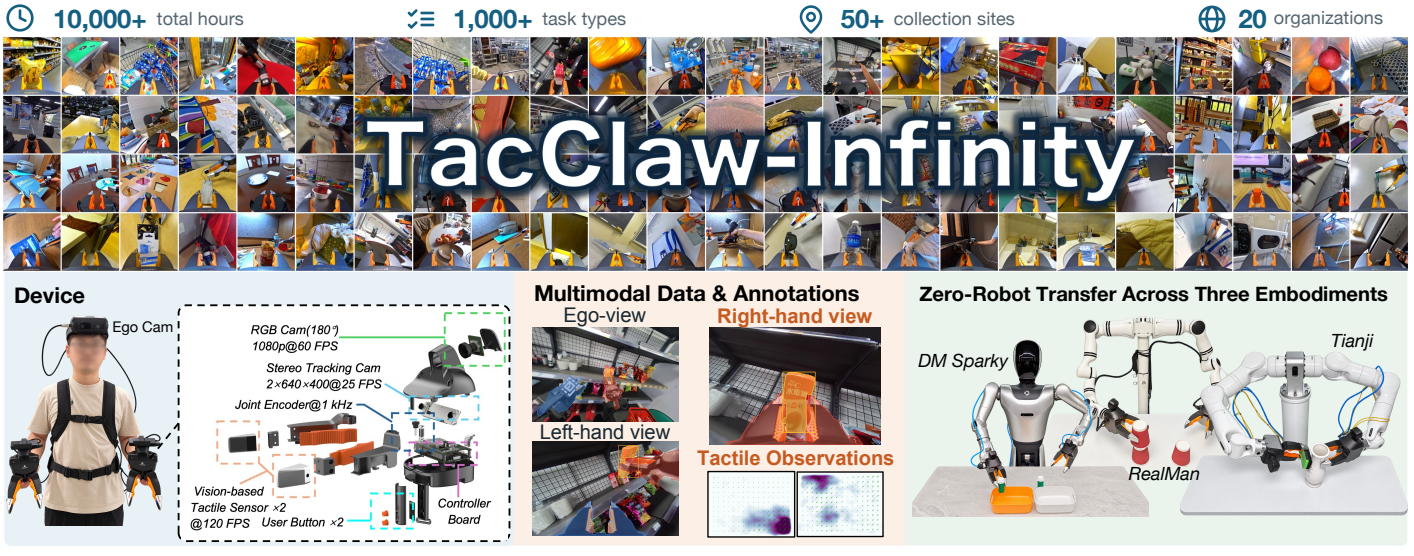


Fig. 1. Overview of TacClaw-Infinity. A tactile-enabled UMI-style interface supports distributed large-scale collection of synchronized visual, tactile, language, and action data across diverse tasks and environments. The resulting dataset enables zero-robot policy transfer across different heterogeneous robot embodiments.

I. Introduction

Large-scale robot learning increasingly relies on diverse human demonstrations to support generalizable manipulation across tasks, objects, and environments. While recent vision-language-action (VLA) models benefit from large-scale multimodal data, contact-rich manipulation additionally requires tactile information that is difficult to infer from vision alone, especially under occlusion, sustained contact, or fine-grained interaction. However, tactile manipulation data remain difficult to scale because collection is often tied to robot-specific sensor integration, calibration, and teleoperation. Moreover, large-scale multimodal demonstrations can vary substantially in quality and typically provide limited structured supervision for contact, task progress, and interaction targets.

To address these challenges, we introduce **TacClaw-Infinity**, a large-scale structured Vision-Tactile-Language-Action (VTLA) dataset for contact-rich manipulation, as illustrated in Fig. 1. The dataset is collected with **TacClaw**, a tactile-enabled UMI-style handheld interface that synchronously records visual observations, fingertip tactile signals, and end-effector trajectories without requiring a target robot during data collection. This robot-independent design enables scalable acquisition across different operators and environments while preserving rich physical interaction signals.

To achieve both scale and diversity, TacClaw-Infinity is collected at more than 50 physical locations spanning homes, laboratories, industrial workplaces, retail stores, outdoor environments, and other everyday settings. Leveraging the portability of TacClaw, we combine protocol-driven recordings of common skills from a shared task catalog with open-ended tasks chosen by collectors. Academic and industrial collaborating organizations further contribute tasks according to their own research and application interests. Together, these efforts yield data of over **10,000 hours**, **1 million episodes**, and **1 billion frames** spanning more than **1,000 task types**.

To make the heterogeneous corpus reliable and learning-ready, we develop a scalable processing pipeline covering trajectory reconstruction, multimodal quality control, and embodiment-aware feasibility filtering based on the reachability and manipulability of multiple robot embodiments. We further provide automatic temporal, semantic, and spatial annotations, including contact-centric subtask segmentation, segment-level language descriptions, and contact-anchored target-object localization and mask tracking. These annotations transform raw multimodal streams into structured supervision for downstream policy learning.

We validate TacClaw-Infinity from both data and model perspectives. At the data level, we evaluate the haptic fidelity of the TacClaw interface, characterize the cross-embodiment action-space compatibility of the collected demonstrations, and assess the quality of the structured annotations. At the model level, we evaluate two representative policy families across three heterogeneous robot embodiments and demonstrate that policies trained solely on TacClaw-Infinity can transfer without embodiment-specific demonstrations. Controlled comparisons further isolate the contributions of tactile observations to downstream policy learning.

Our main contributions are summarized as follows:

- We introduce **TacClaw-Infinity**, a large-scale structured VTLA dataset spanning more than **1,000 task types**, collected in a distributed manner across 50+ locations and 20 organizations.
- We develop a scalable robot-independent collection and processing pipeline with trajectory reconstruction, embodiment-aware feasibility assessment, and automatic temporal, semantic, and spatial annotations. We further evaluate the collection interface, action-space compatibility, and annotation quality.
- We validate the dataset across three robot embodiments and two policy families, demonstrating both cross-embodiment transferability and the effectiveness of tactile sensing.

II. Related Work

A. Large-Scale Robot Demonstration Datasets

Large-scale robot demonstration datasets such as RoboNet, RT-1, BridgeData V2, DROID, and Open X-Embodiment have shown that scaling task, environment, embodiment, and demonstration diversity can substantially improve policy generalization [8], [9], [10], [2], [1]. These resources also support generalist vision-language-action policies through large-scale visual, language, and action supervision [11], [12], [1]. Multimodal datasets such as RH20T further incorporate force, audio, and other sensing modalities [13]. However, existing large-scale datasets remain largely vision-centric or depend on robot-side collection, while synchronized high-resolution tactile data remain scarce. This limits their ability to capture contact-rich interaction states such as slip, insertion, and grasp stability. In contrast, TacClaw-Infinity scales synchronized vision–tactile–language–action demonstrations across diverse environments, providing dense contact information and structured supervision for contact-rich manipulation.

B. UMI-Style and Tactile Data Collection

Embodiment-free collection provides a scalable alternative to robot-side demonstrations. UMI enables in-the-wild data collection with a handheld gripper while retaining observation and action interfaces compatible with downstream robot deployment [3]. Subsequent work improves the scalability of this paradigm: FastUMI simplifies hardware and pose tracking [14], FastUMI-100K scales collection to more than 100,000 multimodal demonstrations [15], and large-scale studies further highlight the importance of object and environment diversity for generalization and enable foundation-model pretraining for open-vocabulary and zero-shot cross-embodiment manipulation [16], [17].

Another line of work focuses on improving the fidelity and deployability of UMI data. VISTA, HiFi-UMI, and UMI-Bench address visual-domain mismatch, trajectory and

TABLE I

COMPARISON WITH REPRESENTATIVE LARGE-SCALE, UMI-STYLE, AND TACTILE MANIPULATION DATASETS. "DISTRIBUTED" INDICATES COLLECTION ACROSS MULTIPLE PHYSICAL LOCATIONS OR CONTRIBUTING INSTITUTIONS; TEMPORAL, SEMANTIC, AND SPATIAL ANNOTATIONS DENOTE WITHIN-EPISODE STRUCTURED SUPERVISION.

Dataset	Collection	Distributed	Episodes	Hours	Tasks	Tactile	Temporal	Semantic	Spatial
Open X-Embodiment [1]	Real-Robot	✓	1M+	–	527	×	×	×	×
DROID [2]	Real-Robot	✓	76K	350	86	×	×	×	×
UMI [3]	UMI	✓	1.4K	–	4	×	×	×	×
ViTaMIn [4]	Tactile UMI	×	747	–	6	✓	×	×	×
exUMI [5]	Tactile UMI	×	1.7K	–	9	✓	×	×	×
Touch in the Wild [6]	Tactile UMI	✓	2.7K+	–	43	✓	×	×	×
FreeTacMan [7]	Tactile UMI	×	10K+	–	50	✓	×	×	×
TacClaw-Infinity (Ours)	**Tactile UMI**	✓	**1M+**	**10K+**	**1K+**	✓	✓	✓	✓

synchronization accuracy, kinematic feasibility, and standardized real-robot evaluation, making robot-free demonstrations more reliable for downstream policy learning [18], [19], [20]. Meanwhile, tactile UMI systems extend the predominantly visuomotor interface with direct measurements of physical interaction. ViTaMIn, FreeTacMan, and Touch in the Wild introduce portable visuo-tactile collection [4], [7], [6], while exUMI, ViTaMIn-B, and TacUMI further explore tactile predictive learning, bimanual collection, and multimodal manipulation [5], [21], [22]. More recent systems broaden tactile UMI toward online feasibility and recovery collection, richer force and multimodal sensing, dexterous hands, and whole-body manipulation [23], [24], [25], [26], [27], [28].

Beyond collection interfaces, tactile-conditioned policies and foundation models have investigated multimodal representation learning, language-conditioned action generation, and visuo-tactile world modeling [29], [30], [31], [32]. Nevertheless, existing work largely focuses on individual interfaces, task-specific datasets, or policy-learning mechanisms. As summarized in Table I, jointly scaling synchronized vision–tactile–language–action data across distributed environments while providing temporal, semantic, and spatial supervision remains underexplored. TacClaw-Infinity addresses this gap by combining large-scale multi-site tactile UMI collection with structured contact-centric annotations for both single-hand and bimanual manipulation.

C. Structured Annotation for Long-Horizon Manipulation

Long-horizon demonstrations contain multiple interaction stages and subtask transitions, motivating temporal segmentation methods such as MS-TCN, MS-TCN++, ASFormer, and ASRF [33], [34], [35], [36], as well as language-conditioned and VLM-based demonstration parsing [37], [38]. However, these approaches are often specialized to particular datasets or task families. TacUMI further demonstrates that multimodal signals can improve tactile event segmentation [22]. In contrast, TacClaw-Infinity adopts a task-agnostic, contact-centric annotation framework that jointly provides temporal stages, segment-level language descriptions, and spatial target annotations, converting heterogeneous demonstrations into structured supervision for large-scale VTLA learning.

III. DATASET

A. Dataset Overview

TacClaw-Infinity is a large-scale structured Vision-Tactile-Language-Action (VTLA) dataset designed for contact-rich manipulation. It combines robot-independent tactile demonstration collection with scalable data curation and structured annotation. The dataset contains over **10,000 hours**, **1 million episodes**, and **1 billion frames** spanning more than **1,000 task types**, as reported in TABLE I. Each episode provides synchronized visual observations, fingertip tactile signals, end-effector trajectories, and language descriptions, together with temporal, semantic, and spatial annotations. The corpus includes both single-hand and bimanual manipulation and covers diverse objects, actions, environments, and physical interaction patterns.

B. TacClaw and Data Collection Protocol

TacClaw is a tactile-enabled UMI-style handheld interface used as the primary data-acquisition platform for TacClaw-Infinity. An exploded view of the device is shown in Fig. 1. The system integrates a wrist-mounted camera, a stereo camera, an inertial measurement unit (IMU), a gripper encoder, and a DM-Tac X visuotactile sensor, which captures three-axis tactile signals from the contact region at up to 120 Hz. Operators can directly perform manipulation demonstrations without requiring a target robot during data collection, while accurately recording end-effector motion, gripper state, visual observations, and tactile feedback.

Gripper mechanics determine how much of the contact reaches the operator's hand. In UMI [3], the trigger-activated push mechanism offsets the fingers from the pinch point; sliding finger-loop designs, such as the wearable parallel-hand exoskeleton [39], introduce a similar offset. Since operators regulate force by feel, this attenuation limits how lightly a demonstration can be performed. TacClaw instead lets the fingers act directly on the jaws through a short, rigid path. In a blindfolded study across three interfaces (Sec. V-A), the mean first-perceived contact force is 11.6 gf with TacClaw, versus 49.8 gf and 204.0 gf for the trigger-driven and finger-loop mechanisms respectively.

Each TacClaw records a monocular wrist RGB stream, three co-registered fingertip tactile modalities—deformation,

shear, and depth—and its end-effector trajectory, all synchronized at 30 Hz; bimanual episodes combine two devices with an egocentric view, together with an episode-level language description.

Pose is estimated by visual-inertial odometry from a stereo module with a hardware-synchronized IMU. To relate the two handheld devices and the egocentric camera in a common frame, each TacClaw carries a fiducial marker on its rear face, visible to the ego camera during collection. Observing both markers lets us register the three local coordinate systems into a shared reference, which improves relative pose accuracy and is what makes the cross-view annotation in Sec. III-C possible.

TacClaw-Infinity is collected across many sites. Exploiting the portability of TacClaw, 200 devices were deployed to 20 organizations and more than 50 physical locations across homes, laboratories, industrial workplaces, retail stores, and outdoor environments. Tasks are either drawn from a shared task catalogue, which yields repeated demonstrations of common skills, or defined by collectors and collaborators according to their own research and application interests, which contributes a long tail of less frequent manipulation behaviors.

C. Data Curation and Annotation Pipeline

Large-scale distributed data collection introduces variations in trajectory quality, sensing reliability, and robot executability. We therefore organize the pipeline into two stages: trajectory curation and data annotation. For trajectory curation, we evaluate each demonstration in the action space of the target embodiment and filter by workspace reachability and manipulability. For data annotation, as shown in Fig. 2, tactile contact events serve as shared physical anchors for temporal, semantic, and spatial annotations.

a) Action-space Analysis: To filter demonstrations that are not reliably executable by target robots, we develop a unified cross-platform action-space evaluation framework for Tianji Marvin M6S Lite, DM-Sparky One, and RealMan RM75-b, considering both reachability and manipulability. We first transform the demonstrated TCP poses into the base frame of each target robot for robot-specific kinematic evaluation.

Reachability. Reachability measures whether a target robot can reproduce a demonstrated pose within its joint limits. For each target pose, we solve inverse kinematics to obtain the joint configuration $\mathbf{q}_t^*$ and reconstruct the corresponding pose through forward kinematics,

$$\hat{\mathbf{T}}_t = \mathrm{FK}(\mathbf{q}_t^*). \tag{1}$$

A frame is considered reachable when the positional and rotational reconstruction errors satisfy $e_{p,t} < 5$ mm and $e_{R,t} < 3^\circ$, respectively. For bimanual trajectories, both arms must satisfy these criteria simultaneously. We report reachability as the proportion of reachable frames.

Manipulability. For each reachable configuration, we further evaluate its local Cartesian motion capability using

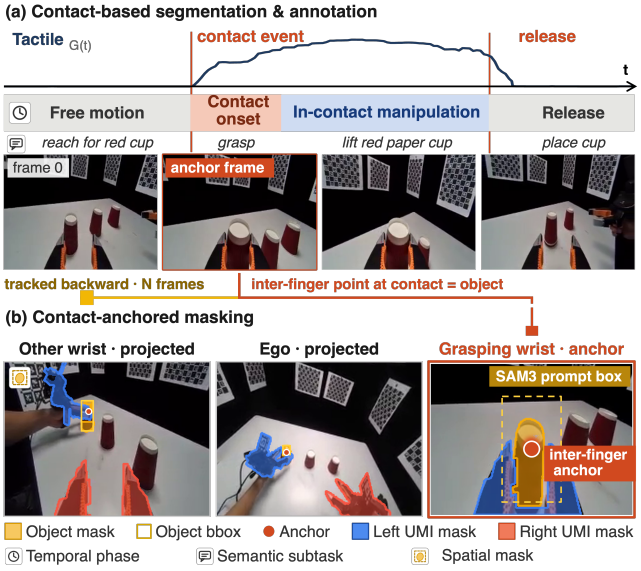


Fig. 2. Contact-anchored structured annotation. (a) Tactile contact and release events partition a demonstration into manipulation stages and provide boundaries for segment-level language annotation. (b) The inter-finger point at contact provides a spatial anchor for identifying the manipulated object. Because the anchor is a single point in space, it projects into every camera of the rig, so each view is segmented independently from the same physical prior. The URDF silhouettes of both devices, rendered per view, serve as negative regions.

the isotropy of the weighted geometric Jacobian $\tilde{\mathbf{J}}_t$, where rotational components are scaled by $0.30\ \mathrm{m\,rad^{-1}}$:

$$\eta_t = \frac{\sigma_{\min}(\tilde{\mathbf{J}}_t)}{\sigma_{\max}(\tilde{\mathbf{J}}_t)}. \tag{2}$$

Higher η_t indicates better-conditioned local motion. Based on perturbation experiments across all evaluated robots, we use $\eta_t = 0.04$ as the low-manipulability threshold. We therefore define the low-manipulability rate for each arm as

$$\text{Low Manip.}_a = \frac{\sum_{t\in\mathcal{R}_a} \mathbb{I}[\eta_{a,t} < 0.04]}{|\mathcal{R}_a|} \times 100\%, \tag{3}$$

where $a \in \{L, R\}$ and $\mathcal{R}_a$ denotes the set of reachable frames for the corresponding arm.

During subsequent policy training, trajectories identified as unreachable or low-manipulability are filtered out. We report the reachability and manipulability statistics of different robot embodiments in Sec. V-B.

b) Contact-aware Temporal Segmentation: As illustrated in Fig. 2(a), we divide each demonstration into four interaction phases: free motion, contact onset, in-contact manipulation, and release.

To automatically identify these phases, we employ a temporal model to learn frame-wise phase transitions. Specifically, we formulate the two fingertip depth maps of each arm into a compact 8-D descriptor to provide a stable and contact-relevant representation for phase recognition, and feed it together with the two-view visual features and proprioceptive states into an ASFormer-based temporal model to predict the interaction phase of each frame. The model also predicts

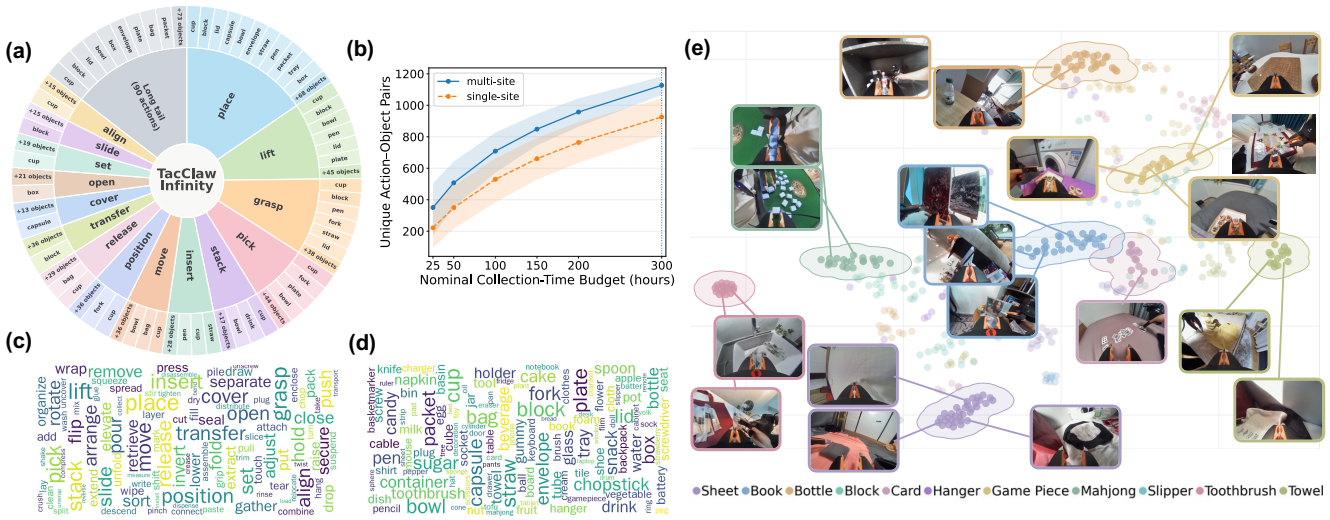


Fig. 3. Statistics of TacClaw-Infinity. (a) Hierarchical distribution of manipulation actions and their associated object categories. (b) Unique action–object pairs under a matched collection-time budget, multi-site versus single-site; shaded bands are 95% resampling intervals over 2,000 draws. (c, d) Action and object vocabularies, with word size proportional to frequency. (e) t-SNE of tactile features colored by object category.

boundary scores, which are used to suppress weak transitions and merge overly short segments.

Finally, Qwen3-VL-Plus generates a subtask description for each segment based on its phase label and frames sampled from each camera.

c) Contact-anchored Spatial Annotation: Fig. 2(b) illustrates the procedure. Each TacClaw is tracked as a rigid body, allowing us to render its URDF silhouette into every synchronized view and obtain per-view masks of both TacClaws.

For the manipulated object, we exploit a physical prior provided by tactile sensing: at the instant contact occurs, the object between the fingertips is directly identified as the manipulation target, without relying on semantic recognition. For each contact event, the URDF-defined inter-finger contact point provides a 3-D anchor, which is projected into every synchronized view to spatially localize the same target across cameras. In each view, we prompt SAM3 using the projected anchor, a local box of fixed physical extent, and an object noun generated by a VLM, while using the rendered TacClaw masks as negative regions. The resulting object mask is then propagated backward to the preceding contact-related event, providing temporally consistent target annotations.

IV. Dataset Statistics and Analysis

We analyze TacClaw-Infinity in terms of scene and task diversity, action and object diversity, and tactile signal diversity to characterize both the breadth of the dataset and the richness of its physical interactions. We further quantify the benefit of distributed acquisition by comparing action–object coverage between single-site and multi-site collection under matched collection-time budgets.

a) Scene and Task Diversity: TacClaw-Infinity spans more than **1,000 task types** collected at more than **50 distinct physical locations** across homes, university laboratories, industrial workplaces, retail stores, and outdoor environments, resulting in substantial variation in backgrounds, layouts, objects, operators, and manipulation conditions. The combination of predefined collection protocols and open-ended distributed acquisition provides both repeated demonstrations of common skills and a long tail of less frequent manipulation behaviors. For example, the *pick-cup* task is instantiated by 192 collectors across 8 organizations and 9 distinct scene settings, illustrating substantial diversity in collection contexts for the same manipulation skill.

b) Action and Object Diversity: As shown in Fig. 3(a), TacClaw-Infinity covers a broad hierarchical distribution of manipulation actions and associated object categories, with the full vocabularies shown in Fig. 3(c, d). Beyond grasping and placing, it includes contact-rich operations such as insertion, alignment, adjustment, wiping, rinsing, and plugging, many of which lie in the long tail of the action vocabulary (Fig. 3(c)). These action–object compositions span diverse geometric and physical interaction constraints, providing broad supervision for transferable manipulation learning. For instance, the *place-block* task spans 56 devices and 6 scenes, with blocks varying in 9 colors, 3 materials, and 5 shapes. It further covers 11 spatial relations and 11 behavior patterns, demonstrating rich variation in both object properties and manipulation behaviors within a single action–object composition.

c) Tactile Signal Diversity: Fig. 3(e) visualizes the tactile feature distribution for selected object categories. The substantial intra-category variation and partial inter-category separation reflect the diversity of contact states induced by different objects and manipulation behaviors.

d) Action–Object Coverage of Multi-site Collection: In this section, we compare the diversity achieved by collecting data across multiple sites with that achieved at a single site, under a fixed collection-time budget. To isolate the effect of site coverage from that of the task protocol, we restrict our analysis to open-ended recordings in which collectors

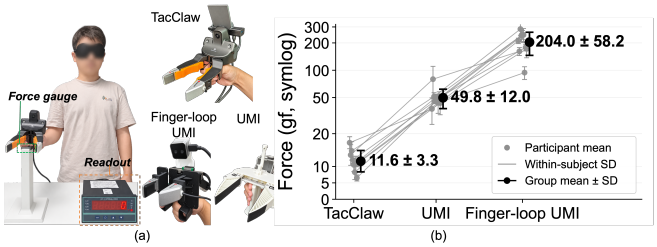


Fig. 4. Haptic subject experiment ($n = 9$). (a) Setup: a blindfolded participant closes the jaws onto a force gauge and says "stop" at first perceived contact; the three interfaces, which differ in how the fingers drive the jaws, are shown at right. (b) Perception threshold per interface. Gray markers denote participant means with within-subject standard deviation; black markers denote group mean $\pm$ standard deviation. The vertical axis uses a symlog scale (linear below 20 gf, logarithmic above). All interfaces were fabricated in-house; UMI follows the authors' released design files.

selected their own tasks. For each budget ranging from 25 to 300 h, we randomly sample, without replacement, episodes whose total duration matches the given budget, drawing either from all sites (multi-site; i.e., the distributed collection described in Sec. III-B, which spans multiple organizations and operators) or from a single site (single-site). We then count the number of distinct action–object pairs covered. This sampling procedure is repeated 2,000 times to estimate confidence intervals.

As reported in Fig. 3(b), multi-site collection consistently covers more action-object pairs than single-site collection, with 201 more pairs (95% confidence interval: $[60, 335]$) at the budget of 300 hours, reflecting the greater diversity of scenes, operators, objects, and task preferences across sites.

V. Experiments

We evaluate TacClaw-Infinity from multiple perspectives, including device-level haptic fidelity, action-space compatibility, annotation quality, and zero-robot cross-embodiment transfer. Together, these experiments assess both the reliability of our data collection and processing pipeline and the utility of the resulting dataset for downstream robot learning.

A. Haptic Subject Experiment

To test the design argument of Sec. III-B, we measure the lightest contact an operator can feel through each interface. Holding the interface with its jaws around the probe of a force gauge (Fig. 4(a)), a blindfolded participant slowly closes the jaws and says "stop" at the first moment contact is felt; the gauge reading at that moment is the *perception threshold*. We compare TacClaw with the original UMI gripper [3] and a sliding-rail finger-loop gripper structurally similar to [39], all shown in Fig. 4(a). Nine participants (mean age 25.6, range 20–33) each completed two practice trials and eight recorded trials per interface; the order of the three interfaces was randomized per participant.

The perception threshold is 11.6 ± 3.3 gram-force (gf) with TacClaw, 49.8 ± 12.0 gf with UMI, and 204.0 ± 58.2 gf with the finger-loop gripper (Fig. 4(b)). All nine participants had the lowest threshold with TacClaw (Friedman $p < 0.001$; pairwise Wilcoxon $p_{\mathrm{Holm}} = 0.012$). The finger-loop gripper is also the least consistent, which may be related to static friction in the rail. Of the three, only direct drive transmits contact to the hand without an intervening mechanism, and it yields the lowest threshold.

B. Action-space Validation

As shown in Table II, the collected demonstrations maintain high bimanual reachability across all three embodiments. DM-Sparky achieves 100% coverage with full-body control, while Tianji and RealMan reach 97.23% and 95.15%, respectively. Meanwhile, low-manipulability rates remain below 5% for all embodiments, indicating that most reachable poses stay away from poorly conditioned or near-singular configurations. Together, these results indicate that the collected demonstrations exhibit strong cross-embodiment kinematic compatibility and are largely executable across heterogeneous robot platforms.

C. Annotation Quality

a) Temporal: We evaluate the contact-phase segmenter of Sec. III-C on a manually annotated benchmark of 137 distinct demonstrations (41 single-arm, 96 bimanual), with 15 bimanual demonstrations designated for testing; each arm is evaluated as a separate sequence, yielding 30 test sequences. Table III reports segmental F1 at 25% overlap (F1@25), edit score, and frame accuracy, plus F1@25 restricted to the contact-onset and release classes (C-F1@25), the boundaries our annotation depends on.

All baselines receive the same visual and proprioceptive input; MS-TCN++ [34], FACT [40], and LTContext [41] are standard segmentation architectures, and ASFormer (no tactile) [35] is our backbone without the tactile descriptor.

The latter comparison isolates what tactile sensing contributes: +6.3 F1@25 and +6.6 C-F1@25, with contact-onset recall rising from 58.6% to 68.0%. Remaining errors concentrate on release, which is confused with free motion as tactile response decays gradually.

b) Spatial: Coverage of the spatial annotation depends on how reliably the object can be tracked. On protocol-driven recordings, where the manipulated entity is a discrete graspable object, 98.5% of frames inside the labelling window carry a mask. Coverage shows that a mask is produced, not that it is exact; mask quality is bounded by SAM3 tracking. Benchmarking mask accuracy and ablating its effect on downstream policies are left for future work.

c) Semantic: A 5% random sample of the VLM-generated subtask descriptions was manually reviewed for correct identification of the action and object. Roughly 83% passed; tasks containing failed samples are held back for manual re-annotation. Since review is sample-based, errors may still remain in other descriptions.

D. Zero-robot Cross-Embodiment Transfer

a) Experiment Setup: We evaluate the dataset and learning framework across three heterogeneous robot embodiments, including Tianji, RealMan, and DM-Sparky, to examine cross-embodiment transferability under different kinematic and hardware configurations. Two representative

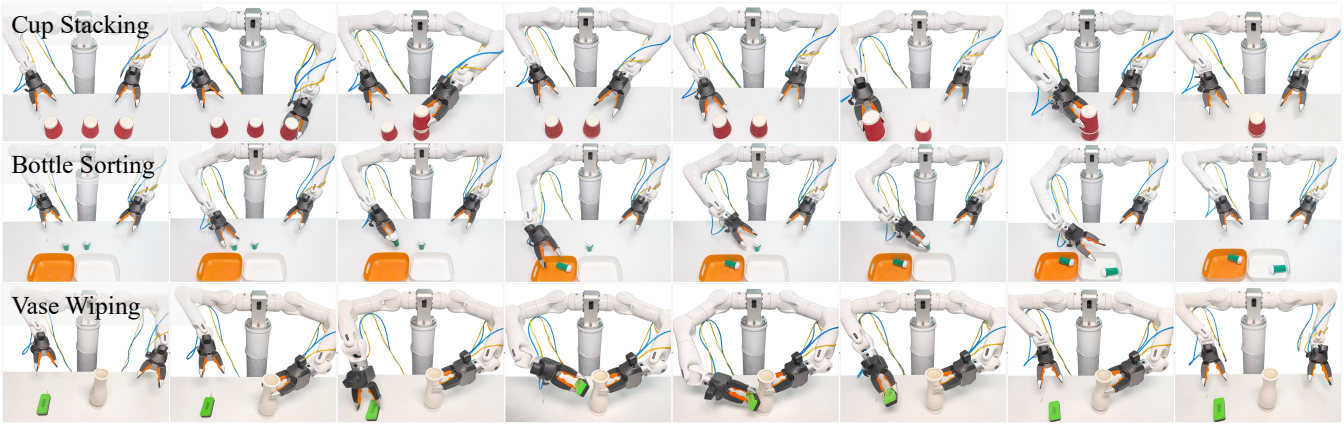


Fig. 5. Representative execution sequences of the three bimanual long-horizon tasks: *Cup Stacking*, *Bottle Sorting*, and *Vase Wiping*. The tasks respectively evaluate pick-and-place coordination, tactile-dependent state discrimination, and dynamic contact regulation.

policy families including $\pi_{0.5}$ and StarVLA are considered. Using both policy families allows us to evaluate whether the collected data generalize across different VLA architectures rather than being specialized to a single policy design.

We consider three representative bimanual long-horizon manipulation tasks with different levels of contact dependence, as shown in Fig. 5. *Cup stacking* mainly evaluates conventional pick-and-place capability and long-horizon coordination. *Bottle sorting* requires the policy to infer object states from relatively static tactile observations during grasping and placement. *Vase wiping* involves sustained and dynamically varying surface contact, requiring continuous tactile feedback for stable interaction throughout the wiping process. Together, these tasks cover free-space manipulation, static tactile perception, and dynamic contact regulation.

For implementation, $\pi_{0.5}$ is trained on 8 NVIDIA A100 GPUs, while StarVLA is trained on 4 NVIDIA A100 GPUs, with the total batch size fixed to 64 for both methods. All tasks are trained for five epochs, and checkpoints from epochs 3, 4, and 5 are evaluated under the same model-selection protocol, with the best-performing checkpoint reported. Unless otherwise specified, all hyperparameters follow the default configurations of $\pi_{0.5}$ and StarVLA. We use 1,600 training episodes for each task and conduct 50 evaluation trials per task. For tactile variants, tactile observations are introduced to both policy families using the same simple image-input formulation.

b) Result analysis: As shown in Table IV, policies trained solely on TacClaw demonstrations can be directly deployed across three heterogeneous robot embodiments without collecting embodiment-specific robot demonstrations, while still achieving strong manipulation performance. The same policy formulation transfers across Tianji Marvin M6S Lite, RealMan RM75-b, and DM-Sparky One despite substantial differences in kinematics and hardware configurations. This cross-embodiment generalization indicates that TacClaw trajectories preserve sufficiently accurate and robot-executable motion information, supporting their use for downstream robot learning.

TABLE II

CROSS-EMBODIMENT REACHABILITY & MANIPULABILITY (%).

Embodiment	**Left Reach.**	**Right Reach.**	**Bimanual Reach.**	**Low Manip. L/R**
Tianji	99.20	98.02	97.23	2.5776 / 3.1611
DM-Sparky (full-body)	100.0	100.0	100.0	0.0 / 0.0
RealMan	97.38	96.79	95.15	4.3127 / 4.3526

TABLE III

CONTACT-PHASE SEGMENTATION ON 30 HELD-OUT PER-ARM SEQUENCES. THE METRICS OF MEAN ± STANDARD DEVIATION OVER 3 SEEDS ARE PRESENTED (%).

Method	**F1@25**	**Edit**	**Acc**	**C-F1@25**
MS-TCN++	48.0 ± 2.6	56.9 ± 3.7	50.2 ± 2.3	43.0 ± 2.4
FACT	57.7 ± 0.7	63.8 ± 4.1	56.8 ± 4.4	45.9 ± 3.1
LTContext	55.4 ± 4.7	63.3 ± 4.8	62.7 ± 4.9	45.4 ± 4.1
ASFormer (no tactile)	57.6 ± 0.5	63.3 ± 1.2	69.5 ± 0.8	47.7 ± 1.0
Ours	$\mathbf{63.9 \pm 0.7}$	$\mathbf{67.3 \pm 2.0}$	$\mathbf{70.4 \pm 0.9}$	$\mathbf{54.3 \pm 0.6}$

The benefit of tactile sensing becomes particularly evident on contact-dependent tasks. Averaged across the three embodiments, adding tactile observations improves $\pi_{0.5}$ from 50.0% to 74.0% overall SR and StarVLA from 46.0% to 72.7%. The gains are concentrated on *Bottle Sorting*, where SR increases by 44.7 and 43.3 percentage points for $\pi_{0.5}$ and StarVLA, respectively, and on *Vase Wiping*, with corresponding gains of 22.0 and 38.7 points. Importantly, tactile observations are introduced only through a simple image-input formulation, without a specialized tactile architecture or controller. These results indicate that TacClaw provides sufficiently informative tactile signals to improve contact-state discrimination and sustained contact regulation across different policy families and robot embodiments.

TABLE IV

ZERO-ROBOT CROSS-EMBODIMENT TRANSFER RESULTS. ALL RESULTS ARE REPORTED AS SUCCESS RATES (SR, %).

Policy	Embodiment	Cup Stacking	Bottle Sorting	Vase Wiping	Avg.
$\pi_{0.5}$	Tianji	90	44	28	54.0
	RealMan	86	42	16	48.0
	DM Sparky	82	44	18	48.0
$\pi_{0.5}$ + Tactile	Tianji	94	96	54	81.3
	RealMan	80	76	38	64.7
	DM Sparky	100	92	36	76.0
StarVLA	Tianji	78	46	12	45.3
	RealMan	78	38	8	41.3
	DM Sparky	80	54	20	51.3
StarVLA + Tactile	Tianji	70	94	60	74.7
	RealMan	76	84	58	72.7
	DM Sparky	84	90	38	70.7

VI. CONCLUSION

We presented **TacClaw-Infinity**, a large-scale structured Vision-Tactile-Language-Action dataset for contact-rich manipulation. By combining robot-independent tactile demonstration collection with distributed acquisition, scalable data curation, and structured annotation, TacClaw-Infinity brings together large-scale multimodal interaction data with temporal, semantic, and spatial supervision. Our evaluations validate the dataset from both data and model perspectives: the TacClaw interface provides sensitive contact feedback, the trajectories are kinematically compatible across embodiments, and the annotations provide contact-aware supervision. Policies trained on the resulting data further demonstrate cross-embodiment transfer across heterogeneous robots without requiring embodiment-specific demonstrations, while controlled comparisons highlight the utility of tactile observations for contact-rich manipulation. We hope TacClaw-Infinity provides a foundation for scalable visuotactile robot learning and encourages further research on learning physical interaction from diverse human demonstrations.

REFERENCES

[1] Open X-Embodiment Collaboration, "Open X-Embodiment: Robotic learning datasets and RT-X models," in *ICRA*, 2024.

[2] A. Khazatsky *et al.*, "Droid: A large-scale in-the-wild robot manipulation dataset," in *RSS*, 2024.

[3] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song, "Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots," in *RSS*, 2024.

[4] F. Liu, C. Li, Y. Qin, A. Shaw, J. Xu, P. Abbeel, and R. Chen, "Vitamin: Learning contact-rich tasks through robot-free visuo-tactile manipulation interface," *arXiv preprint arXiv:2504.06156*, 2025.

[5] Y. Xu, L. Wei, P. An, Q. Zhang, and Y.-L. Li, "exUMI: Extensible robot teaching system with action-aware task-agnostic tactile representation," in *CoRL*, 2025.

[6] X. Zhu, B. Huang, and Y. Li, "Touch in the wild: Learning fine-grained manipulation with a portable visuo-tactile gripper," in *NeurIPS*, 2025.

[7] L. Wu, C. Yu, J. Ren, L. Chen, Y. Jiang, R. Huang, G. Gu, and H. Li, "FreeTacMan: Robot-free visuo-tactile data collection system for contact-rich manipulation," *arXiv preprint arXiv:2506.01941*, 2025.

[8] S. Dasari, F. Ebert, S. Tian, S. Nair, B. Bucher, K. Schmeckpeper, S. Singh, S. Levine, and C. Finn, "Robonet: Large-scale multi-robot learning," in *CoRL*, 2020.

[9] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu *et al.*, "Rt-1: Robotics transformer for real-world control at scale," in *RSS*, 2023.

[10] H. R. Walke, K. Black, A. Lee, M. J. Kim, M. Du, C. Zheng, T. Zhao, P. Hansen-Estruch, Q. Vuong, A. W. He, V. Myers, K. Fang, C. Finn, and S. Levine, "Bridgedata v2: A dataset for robot learning at scale," in *CoRL*, 2023.

[11] A. Brohan *et al.*, "RT-2: Vision-language-action models transfer web knowledge to robotic control," in *CoRL*, 2023.

[12] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. P. Foster, P. R. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn, "OpenVLA: An open-source vision-language-action model," in *CoRL*, 2025.

[13] H.-S. Fang, H. Fang, Z. Tang, J. Liu, C. Wang, J. Wang, H. Zhu, and C. Lu, "Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot," in *ICRA*, 2024.

[14] Zhaxizhuoma, K. Liu, C. Guan, Z. Jia, Z. Wu, X. Liu, T. Wang, S. Liang, P. Chen, P. Zhang, H. Song, D. Qu, D. Wang, Z. Wang, N. Cao, Y. Ding, B. Zhao, and X. Li, "FastUMI: A scalable and hardware-independent universal manipulation interface with dataset," in *CoRL*, 2025.

[15] K. Liu, Z. Jia, Y. Li, Zhaxizhuoma, P. Chen, S. Liu, X. Liu, P. Zhang, H. Song, X. Ye, N. Cao, Z. Wang, J. Zeng, D. Wang, Y. Ding, B. Zhao, and X. Li, "FastUMI-100K: Advancing data-driven robotic manipulation with a large-scale UMI-style dataset," *arXiv preprint arXiv:2510.08022*, 2025.

[16] F. Lin, Y. Hu, P. Sheng, C. Wen, J. You, and Y. Gao, "Data scaling laws in imitation learning for robotic manipulation," in *ICLR*, 2025.

[17] S. Liu, B. Li, K. Ma, L. Wu, H. Tan, X. Ouyang, H. Su, and J. Zhu, "RDT2: Exploring the scaling limit of UMI data towards zero-shot cross-embodiment generalization," *arXiv preprint arXiv:2602.03310*, 2026.

[18] S. Yang, L. Guo, O. Lu, Zhaxizhuoma, D. Zhang, X. Wang, T. Xiao, F. Yan, Z. Chen, Y. Ding, C. Yu, C. Bai, and X. Li, "VISTA: Vision-grounded and physics-validated adaptation of UMI data for VLA training," *arXiv preprint arXiv:2606.04708*, 2026.

[19] Simple AI, Y. Wei, J. Ma, J. Wang, W. Zhou, Y. Zuo, K. Rui, M. Li, J. Zhang, Z. Pan, X. Wang, H. Jia, H. Du, Z. Zeng, J. Ma, G. Qin, D. Zhang, and X. Li, "HiFi-UMI: Learning deployable manipulation policies from high-fidelity UMI data alone," *arXiv preprint arXiv:2607.25895*, 2026.

[20] S. Jin, Y. Wang, Y. Duan, D. Wu, G. Dong, X. Liu, X. Li, H. Jia, Z. Zhang, T. Wang, Z. Jia, Y. Yao, C. Bai, Zhaxizhuoma, S. Liu, N. Cao, J. Wang, C. Yu, and Y. Ding, "UMI-Bench 1.0: An open and reproducible real-world benchmark for tabletop robotic manipulation with UMI data," *arXiv preprint arXiv:2606.10382*, 2026.

[21] C. Li, C. Liu, D. Wang, S. Zhang, L. Li, Z. Zeng, F. Liu, J. Xu, and R. Chen, "Vitamin-b: A reliable and efficient visuo-tactile bimanual manipulation interface," *arXiv preprint arXiv:2511.05858*, 2025.

[22] T. Cheng, K. Chen, L. Chen, L. Zhang, Y. Zhang, Y. Ling, M. Hamad, Z. Bing, F. Wu, K. Sharma, and A. Knoll, "Tacumi: A multi-modal universal manipulation interface for contact-rich tasks," in *ICRA*, 2026.

[23] L. Wu, J. Ren, C. Jiang, J. Zhou, S. Peng, R. Huang, G. Gu, L. Chen, and H. Li, "TAMEn: Tactile-aware manipulation engine for closed-loop data collection in contact-rich tasks," *arXiv preprint arXiv:2604.07335*, 2026.

[24] S. Luo, Y. Li, Y. Hu, C. Yu, C. Xu, J. Zhang, G. Yao, T. Huang, R. He, and Z. Wang, "OmniUMI: Towards physically grounded robot learning via human-aligned multimodal interaction," *arXiv preprint arXiv:2604.10647*, 2026.

[25] H. Li, Y. Xu, Y. Tang, Y. Dong, C. Liu, C. Yu, X. Li, S. Huang, Y. Shen, N. Xue, and Y.-L. Li, "AetheRock: An arm-worn robot teaching system for force-guided vision-tactile learning," *arXiv preprint arXiv:2606.09777*, 2026.

[26] M. Xu, H. Zhang, Y. Hou, Z. Xu, L. Fan, M. Veloso, and S. Song, "DexUMI: Using human hand as the universal manipulation interface for dexterous manipulation," *arXiv preprint arXiv:2505.21864*, 2025.

[27] C. Xu, Y. Jiang, J. Huan, Y. Fu, H. Zhou, W. Yuan, J. Yu, W. Zhang, H. Yuan, and Z. Lu, "RealDexUMI: A wearable universal manipulation interface for dexterous robot learning," *arXiv preprint arXiv:2606.06033*, 2026.

[28] J. Jang, Z. Gu, A. Cueva, Z. Chai, J. Sheng, T. Nguyen, H. Galundia, Y. Wu, H. Xue, I. Legene, O. Mediratta, D. Doan, A. Collins, S. Sadegh, K. Kim, R. Dhalbisoi, Z. Chen, and Y. Zhao, "WT-UMI: Tactile-based whole-body manipulation via force-supervised contact-aware planning," *arXiv preprint arXiv:2606.13232*, 2026.

[29] Q. Liu, Y. Cui, Z. Sun, G. Li, J. Chen, and Q. Ye, "Vtdexmanip: A dataset and benchmark for visual-tactile pretraining and dexterous manipulation with reinforcement learning," in *ICLR*, 2025.

[30] P. Hao, C. Zhang, D. Li, X. Cao, X. Hao, S. Cui, and S. Wang, "Tla: Tactile-language-action model for contact-rich manipulation," *arXiv preprint arXiv:2503.08548*, 2025.

[31] C. Zhang, P. Hao, X. Cao, X. Hao, S. Cui, and S. Wang, "Vtla: Vision-tactile-language-action model with preference learning for insertion manipulation," *arXiv preprint arXiv:2505.09577*, 2025.

[32] Y. Zheng, S. Gu, W. Li, Y. Zheng, Y. Zang, S. Tian, X. Li, R. Wu, C. Hao, C. Gao *et al.*, "Omnivta: Visuo-tactile world modeling for contact-rich robotic manipulation," *arXiv preprint arXiv:2603.19201*, 2026.

[33] Y. Abu Farha and J. Gall, "Ms-tcn: Multi-stage temporal convolutional network for action segmentation," in *CVPR*, 2019.

[34] S.-J. Li, Y. A. Farha, Y. Liu, M.-M. Cheng, and J. Gall, "MS-TCN++: Multi-stage temporal convolutional network for action segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 6647–6658, 2023.

[35] F. Yi, H. Wen, and T. Jiang, "Asformer: Transformer for action segmentation," in *BMVC*, 2021.

[36] Y. Ishikawa, S. Kasai, Y. Aoki, and H. Kataoka, "Alleviating over-segmentation errors by detecting action boundaries," in *WACV*, 2021.

[37] D. Raj, O. Patil, W. Gu, C. Baral, and N. Gopalan, "Learning temporally composable task segmentations with language," in *IROS*, 2024.

[38] B. Wang, J. Zhang, S. Dong, I. Fang, and C. Feng, "VLM See, Robot Do: Human demo video to robot action plan via vision language model," in *IROS*, 2025.

[39] Z. Lu, L. Chen, H. Dai, H. Li, Z. Zhao, B. Zheng, N. F. Lepora, and C. Yang, "Visual-tactile robot grasping based on human skill learning from demonstrations using a wearable parallel hand exoskeleton," *IEEERA-L*, vol. 8, no. 9, pp. 5384–5391, 2023.

[40] Z. Lu and E. Elhamifar, "FACT: Frame-action cross-attention temporal modeling for efficient action segmentation," in *CVPR*, 2024.

[41] E. Bahrami, G. Francesca, and J. Gall, "How much temporal long-term context is needed for action segmentation?" in *ICCV*, 2023.