BEYOND GRASPING: A MULTIMODAL HUMAN DEMONSTRATION DATASET AND BENCHMARK FOR DEXTEROUS TOOL USE

Anonymous authors
Paper under double-blind review



Figure 1: **PalmDex** is a large-scale multimodal dataset that focuses on full-hand dexterous tool use.

ABSTRACT

Dexterous tool use requires an embodied agent to understand task procedures, select appropriate tools and grasp types, regulate contact forces, and generate coordinated hand motions, yet these coupled capabilities remain underrepresented in existing manipulation datasets. We introduce **PalmDex**, a robot-free multimodal human demonstration dataset and benchmark covering 80 distinct tools and task-relevant objects, 59 real-world tasks, and approximately 31 hours of synchronized recordings, with more than 200 annotated clips per task on average. PalmDex provides synchronized egocentric and exocentric videos, 3D hand kinematics, wrist trajectories, and full-hand tactile signals. The benchmark assesses four complementary capabilities: tool-use cognition, interaction knowledge, cross-modal alignment, and motion prior learning. It comprises 43,391 question-answer pairs, together with temporal grounding, cross-modal retrieval, and motion prediction tasks. Experiments with more than 20 vision-language models and task-specific baselines show that no single model performs consistently across all four capabilities. Modality ablations reveal that vision primarily supports scene and procedure understanding, while temporally aligned tactile signals provide complementary cues for contact- and force-related reasoning. Furthermore, fine-tuning exclusively on PalmDex yields an average absolute improvement of 12.7 percentage points across four shared tasks on the independently collected EgoTouch dataset, without using any target-domain training data. These results establish PalmDex as a unified benchmark for learning and evaluating transferable interaction knowledge in dexterous tool use. The dataset, benchmark, and code will be publicly released.

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.

1 INTRODUCTION

Tool use is a defining capability of embodied intelligence, requiring agents to integrate perception, semantic reasoning, and dexterous manipulation to achieve task goals. Unlike generic object manipulation, successful tool use depends not only on stable physical interaction but also on understanding functional affordances, selecting task-conditioned grasp configurations, and coordinating manipulation strategies over extended horizons. Developing such capabilities is essential for deploying embodied agents in open-world environments, where tools provide a flexible interface for accomplishing diverse tasks beyond an agent's innate physical abilities.

Recent advances in embodied learning have been driven by large-scale datasets collected from teleoperation systems or robot-free human demonstrations. These resources have significantly improved policy learning for grasping, pick-and-place, navigation, and general household manipulation. However, they provide limited support for learning functional tool use. Existing manipulation datasets primarily emphasize motion imitation and task execution, while recent tool-related benchmarks focus on visual or language-based reasoning without modeling physical interaction. Consequently, current embodied models are rarely evaluated on their ability to jointly understand tool affordances, reason about task-dependent manipulation, and execute dexterous tool-use behaviors. This disconnect has left embodied tool use largely unexplored as a unified learning problem.

To address this gap, we present **PalmDex**, a large-scale robot-free dataset for embodied tool use. The dataset is collected from natural human demonstrations using a portable wearable sensing platform, enabling scalable data acquisition without restricting demonstrations to a specific robotic embodiment. Each demonstration is synchronously annotated with rich multimodal observations, including hand kinematics, wrist poses, multi-view RGB videos, tactile signals, and semantic labels describing tasks, action segments, and tool affordances. By jointly capturing the semantic and physical aspects of interaction, the dataset supports learning representations that bridge perception, reasoning, and manipulation, while facilitating transfer across heterogeneous robotic platforms.

Beyond data collection, we argue that embodied tool use should be evaluated as an integrated capability rather than an isolated manipulation task. We therefore establish four complementary benchmarks: Tool-Use Cognition for high-level procedural understanding, Interaction Knowledge for multimodal inference of grasping, force, and action progression, Cross-Modal Alignment for visual–tactile representation learning, and Motion Prior for wrist-trajectory and 3D hand-pose prediction. We evaluate representative multimodal models and cross-dataset transfer across these tasks. The results expose substantial gaps in reasoning, multimodal interaction understanding, alignment, and dexterous motion prediction, highlighting PalmDex as a comprehensive testbed for tool-use learning, as shown in Figure 1. In summary, our contributions are as follows:

- We introduce a robot-free multimodal dataset for embodied tool use, comprising large-scale human demonstrations with synchronized visual, hand kinematic, tactile, and semantic annotations collected across diverse real-world tasks.
- We establish a unified benchmark for embodied tool use spanning four complementary capabilities: tool-use cognition, interaction knowledge, cross-modal alignment, and motion prior learning, together with standardized evaluation protocols and metrics.
- We conduct comprehensive evaluations with representative multimodal large language models and visual baselines, revealing substantial capability gaps in current models, and further validate the transferability and generalization enabled by PalmDex through zero-shot transfer experiments on existing embodied benchmarks.

2 RELATED WORK

2.1 DEXTEROUS MANIPULATION DATASETS

Dexterous manipulation datasets have evolved from visual geometry toward physical interaction. Early datasets focus on 3D hand–object pose and reconstruction (Hampali et al., 2020; Chao et al., 2021), while recent egocentric datasets add RGB-D, multi-view geometry, and action semantics (Liu et al., 2022; Banerjee et al., 2025; Perrett et al., 2025; Hoque et al., 2026). These resources capture what the hand is doing, but reveal little about how contact is physically established.

Recent datasets address this gap through pressure, force, and tactile sensing. EgoPressure and FEEL model contact regions, force states, and contacted objects (Zhao et al., 2025; Dessalene et al., 2026); OpenTouch aligns vision, hand pose, and full-hand touch (Song et al., 2025); and TouchAnything/EgoTouch and OmniVTA/OmniViTac extend tactile modelling to continuous pressure prediction and contact dynamics (Zhou et al., 2026; Zheng et al., 2026b). Nevertheless, full-hand touch, articulated motion, and temporally structured tool use are rarely captured together, limiting the study of functional interactions that depend jointly on contact and procedure.

2.2 Robot-Free Data Collection

Large-scale robot datasets aggregate vision–action trajectories across platforms and environments (Open X-Embodiment Collaboration, 2024; Walke et al., 2023; Khazatsky et al., 2024; Wu et al., 2025), but their supervision remains tied to specific robots, end effectors, and control spaces. UMI reduces this dependence through portable human-operated collection, although its actions are still defined around a particular gripper (Chi et al., 2024).

Egocentric human datasets instead treat the hand as a general-purpose manipulator (Hoque et al., 2026; Zhou et al., 2026; Jawaid & Xiang, 2025; Punamiya et al., 2026; Zheng et al., 2026a). This robot-free paradigm scales without deploying a target robot and preserves natural approach, contact, adjustment, and correction behaviours. Its central challenge is therefore not merely collecting demonstrations, but extracting interaction knowledge that remains transferable across embodiments. Table 1 compares PalmDex with representative tactile and robot-manipulation datasets.

Table 1: Capability comparison with tactile and robot-manipulation datasets (✓: available, ✗: not; more ⋆indicates richer tool-use diversity).

Dataset	Egocentric	Embodiment-free	Hand pose	Full-hand Tactile	Success Annotation	Tool-use Diversity
PalmDex (ours)	✓	✓	✓	✓	✓	⋆⋆⋆⋆⋆
OpenTouch	✓	✓	✓	✓	✗	⋆⋆
Touch Anything	✓	✓	✓	✓	✗	⋆⋆
OmniVTA / OmniViTac	✗	✗	✗	✗	✗	⋆⋆⋆
Daimon-Infinity	✗	✗	✗	✗	✗	⋆⋆⋆⋆
OpenLET / LET-Dex	✗	✗	✓	✗	✗	⋆⋆
Open X-Embodiment	✗	✗	✗	✗	✗	⋆⋆⋆⋆

Notes. Tool-use diversity is qualitatively rated by the breadth of tool categories and manipulation tasks covered, with more stars indicating greater diversity.

2.3 Benchmarking In Dexterous Manipulation

Dexterous manipulation benchmarks have similarly progressed from geometric parsing to semantic, physical, and predictive reasoning. HOI4D and HOT3D evaluate hand–object pose and tracking (Liu et al., 2022; Banerjee et al., 2025); Ego4D NLQ and HD-EPIC introduce language-based temporal and semantic understanding (Grauman et al., 2022; Perrett et al., 2025); and tactile benchmarks evaluate pressure, contact, retrieval, synthesis, and multimodal prediction (Zhao et al., 2025; Dessalene et al., 2026; Song et al., 2025; Zhou et al., 2026; Huang et al., 2026). Motion-oriented benchmarks further forecast future wrist and fingertip trajectories (Hoque et al., 2026; Jawaid & Xiang, 2025).

However, these capabilities are often evaluated in isolation, allowing a model to recognize tool use without understanding how the interaction is physically executed or how it evolves over time. PalmDex instead treats embodied tool use as an integrated capability, jointly considering semantic understanding, physical interaction, multimodal correspondence, and motion prediction. This formulation moves beyond recognition alone toward assessing whether a model captures the knowledge required to reproduce tool-use behavior.

3 Dataset Design

3.1 Data Collection And Annotation System

To capture the sensory and procedural information involved in embodied tool use, we develop a portable multimodal data collection and annotation system, as shown in Figure 2. Egocentric and

third-person cameras provide complementary visual observations, a UDEXREAL motion-capture glove records 21-joint 3D hand kinematics, and a Juqiao tactile glove captures full-hand contact signals. All streams are synchronized through a unified timestamp pipeline to produce temporally aligned RGB, hand-pose, and tactile demonstrations. Each demonstration is further enriched with hierarchical annotations, including natural-language descriptions, tool and object identities, object categories, grasp types, fine-grained action steps, high-level process stages, and task-success information. These annotations establish explicit correspondences among visual observations, physical interactions, and procedural progress, supporting the four PalmDex benchmark tracks. System is shown in Figure 2.

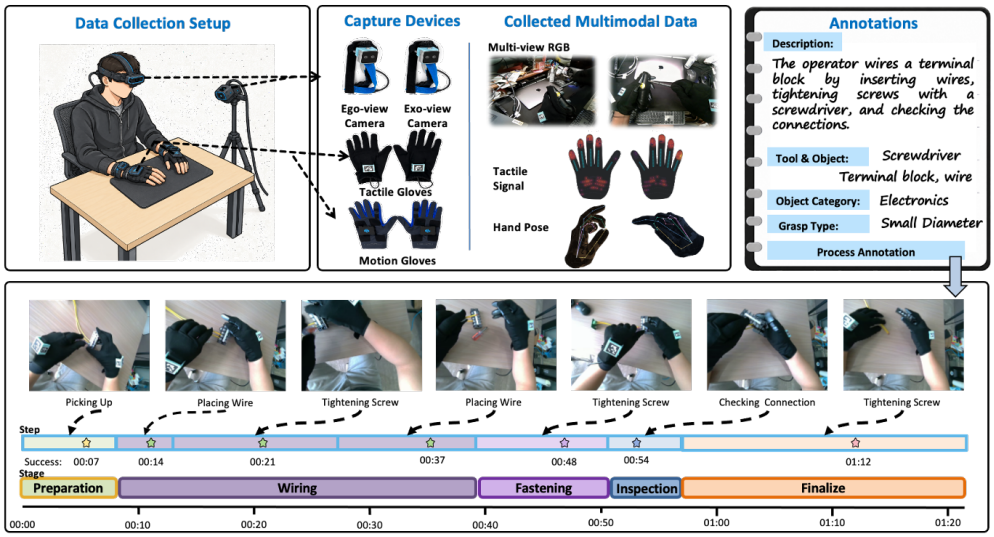

Figure 2: Data collection and annotation system

3.2 TOOL-USE TASKS

Unlike existing datasets that primarily focus on grasping and pick-and-place, PalmDex is designed around *functional tool use*. Each task requires participants to grasp and operate a purpose-specific tool to act on a target object or environment. Successful execution therefore involves not only stable grasping, but also task-oriented grasp adaptation, fine finger coordination, bimanual role allocation, procedural understanding, and controlled force application. PalmDex contains 18 tool-use tasks captured across six professionally instrumented environments, spanning laboratory, electronics, kitchen, office, and general workbench settings. The tasks are grouped into the following categories:

- **Laboratory sciences.** Tasks from chemistry, biology, physics, and pharmaceutical laboratories include grinding materials with a mortar and pestle, transferring crucibles with tongs, pipetting samples between vessels, dispensing liquids with a dropper, and performing precision weighing.
- **Electronics assembly and inspection.** Workbench tasks include installing a memory module into a DIMM slot, testing IMU devices, manipulating surface-mount components with a vacuum pickup pen, and connecting terminal wires.
- **Kitchen, office, and everyday tool use.** Everyday tasks include cleaning a mug with a brush, mixing beverages, operating kettles and microwave ovens, writing with pens and correction tape, and using tools such as hammers, screwdrivers, and soldering irons.

Together, these tasks encompass diverse tool–object interactions and require coordinated transitions between reaching, grasping, reorienting, contacting, and releasing. PalmDex thus emphasizes contact-rich, goal-directed tool use beyond simple object grasping and transport.

3.3 Dataset Statistics

Scale. In total, PalmDex comprises roughly 31 hours of synchronized egocentric recordings and spans 3,112 episodes, including 2,696 episodes collected with our robot-free capture setup and 416 episodes from the EgoTouch dataset. Our capture provides approximately 1.8M frames per view at 30 Hz and is temporally segmented into more than 12,000 labeled tool-use steps. Each frame is paired with egocentric and third-person video, a 256-channel full-hand tactile signal, and 72-dimensional finger-joint kinematics, grounding every clip jointly in vision, touch, and pose.

Manipulation verbs. The top panel of Figure 3 shows the distribution of manipulation verbs on a logarithmic scale, spanning 53 distinct verbs in 14 semantic categories. Beyond grasp-and-transport primitives such as *pick up* and *place*, PalmDex is dominated by tool-mediated operations such as liquid handling, scooping and weighing, grinding and stirring, screw fastening, card insertion and removal, chip suction-picking, connector plugging, and clamping-based transfer, demonstrating coverage of contact-rich dexterous tool use well beyond standard pick-and-place.

Tools and objects. The bottom panel of Figure 3 shows that PalmDex covers 80 distinct tools and objects grouped into 15 categories, ranging from laboratory glassware, pipettes, spatulas, tongs, and brushes to SMT and PC-assembly parts and everyday hand tools. As the taxonomy inset of Figure 3 shows, no single family dominates. Laboratory glassware and SMT electronics are the largest categories, each accounting for 17% of object mentions, followed by containers and bottles at 12%, gripping tools at 11%, and pipettes and scoops at 10%.

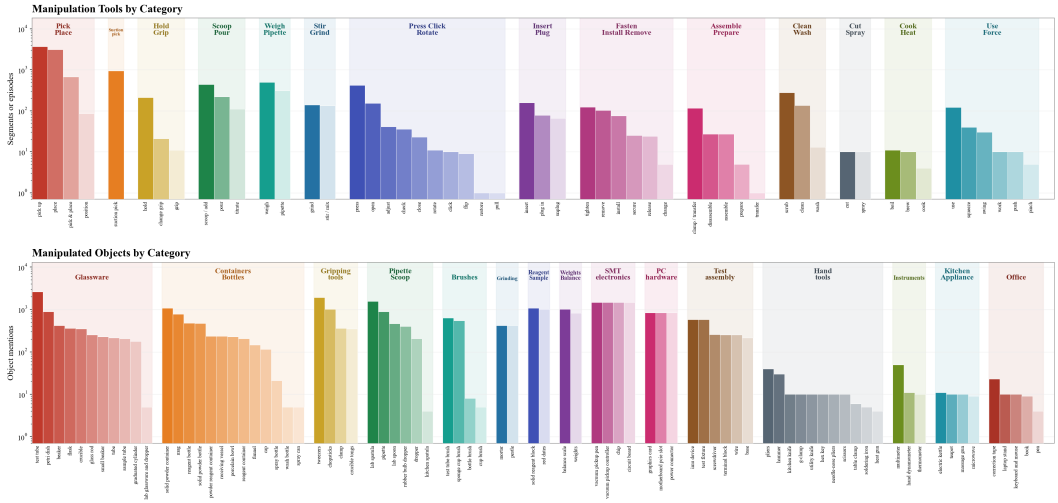

Figure 3: Frequency of verb clusters (top) and object (noun) clusters (bottom) in our manipulation annotations, grouped by category and shown on a logarithmic scale.

4 Benchmarks

4.1 Problem Formulation

We formulate robot-free tool-use learning as acquiring transferable manipulation knowledge from multimodal human demonstrations. Each demonstration is represented as

$$\mathcal{D}_i = \{\mathbf{V}_i^{\text{ego}}, \mathbf{V}_i^{\text{exo}}, \mathbf{T}_i, \mathbf{H}_i, \mathbf{S}_i\}, \tag{1}$$

where $\mathbf{V}^{\text{ego}}$ and $\mathbf{V}^{\text{exo}}$ denote egocentric and exocentric videos, $\mathbf{T}$ denotes tactile observations, $\mathbf{H}$ contains hand and wrist kinematics, and $\mathbf{S}$ contains semantic annotations.

Given a task-specific subset of these observations, each benchmark task is formulated as

$$\hat{\mathbf{Y}}_\tau = f_\theta^\tau\left(\mathcal{X}_\tau, q\right), \qquad \mathcal{X}_\tau \subseteq \mathcal{D}_i, \tag{2}$$

where τ denotes a benchmark task, q is an optional task query, and $\mathbf{Y}_\tau$ is its prediction target. The permitted inputs are defined separately for each task to prevent target or future information leakage.

4.2 BENCHMARK TASKS

We organize the benchmark into four complementary task groups, each targeting a distinct tool-use capability. Beyond conventional scene and action understanding, these tasks evaluate whether multimodal human demonstrations provide actionable and transferable knowledge for teaching different robotic embodiments to use tools. Each task group is evaluated independently as a complementary capability rather than as a stage in a sequential pipeline.

Tool-Use Cognition. Our benchmark evaluates tool-use cognition along seven dimensions: task recognition, tool identification, procedure sequencing, force reasoning, grasp-type classification, key-stage localization, and holistic reasoning. Key-stage localization is formulated as temporal grounding and evaluated by temporal IoU, while the remaining six dimensions are video question-answering tasks. Together, they assess both semantic understanding of tool-use demonstrations and procedural understanding of how actions are organized toward a task goal. This group contains 32,367 question–answer pairs; detailed task statistics are provided in Appendix B.1. **Interaction Knowledge.** Beyond visual recognition, this benchmark evaluates whether a unified model can infer how to use a tool from a high-level instruction and multimodal context. We define seven tasks: current action inference, tool selection, grasp type prediction, force-bearing hand identification, force trend prediction, step completion estimation, and next-action prediction. For current action inference and tool selection, direct visual cues are masked, requiring reasoning from the instruction and preceding context. Together, these tasks assess what action to take, which tool and grasp to use, how to apply force, and how to proceed. Further details are provided in Appendix B.2.1. **Cross-Modal Alignment.** This task evaluates whether visual and tactile observations of the same interaction can be aligned in a shared representation space, capturing complementary appearance and contact information.

Motion Prior. This task predicts task-conditioned wrist and hand motions from human demonstrations, providing coarse motion guidance for downstream retargeting and robotic execution.

4.3 EVALUATION METRICS

We adopt task-specific metrics for the four benchmark groups. The question-answering tasks in Tool-Use Cognition and Interaction Knowledge are evaluated using accuracy, while key-stage localization is measured by temporal intersection over union (tIoU). For Cross-Modal Alignment, we report Recall@1 and Recall@5 for both vision-to-touch and touch-to-vision retrieval. Motion Prior is evaluated using average displacement error (ADE) and final displacement error (FDE) for wrist trajectory prediction, and mean per-joint position error (MJAE) for wrist-root-relative 3D hand-pose prediction. Higher values indicate better performance for accuracy, tIoU, and Recall@K, whereas lower values are preferred for motion errors. Formal definitions are provided in Appendix B.5.

5 EXPERIMENTS

Experimental Setup. All experiments use a shared episode-level 8:1:1 split with no overlap across training, validation, and test sets. Tool-use cognition is evaluated zero-shot, where models receive only the egocentric video clip and textual question, without tactile input, visual markers, or task-specific training. We evaluate representative proprietary MLLMs from the GPT families, general-purpose open-source MLLMs including Qwen-VL, InternVL, Ovis, LLaVA, Gemma, Kimi-VL, DeepSeek-VL, and GLM-V, as well as embodied models such as Magma.

For the remaining tracks, models are trained exclusively on the PalmDex training split and evaluated on the held-out test split. Interaction-knowledge tasks take four egocentric frames and their temporally aligned dual-hand tactile heatmaps as input. Unless otherwise specified, we adopt Qwen2.5-VL-3B as the base model, freeze its vision encoder, and perform parameter-efficient fine-tuning using LoRA. Cross-modal alignment is evaluated through bidirectional vision-to-touch and touch-to-vision retrieval, while motion prediction comprises future wrist-trajectory prediction and wrist-root-relative 3D hand-pose prediction. Full model versions, prompts, preprocessing procedures, training hyperparameters, and implementation details are provided in Appendix B.

5.1 SCENE COGNITION: UNDERSTANDING TOOL USE

Table 2 shows that PalmDex remains challenging, with the best overall accuracy reaching only 39.8%. GPT-6 Astra achieves the strongest overall performance, with 87.9% on task recognition, 79.0% on tool recognition, and 61.1% on action ordering, demonstrating a clear advantage in procedural reasoning. However, temporal grounding remains difficult, with GPT-5.4 achieving the best result of only 0.346 mIoU. Among open-source models, Qwen3-VL-30B-A3B reaches the highest overall accuracy of 36.3%, while different models lead tool recognition, force-trend prediction, and hand-pose prediction. These results show that current models still lack a unified understanding of fine-grained tool-use procedures.

Table 2: Zero-shot tool-use cognition results on PalmDex. All metrics are accuracy (%) except Temp (mIoU). Full results are provided in Appendix C.1.

Model	Task	Tool	Force	Hand pose	Ordering	Temporal	Overall
General-Purpose Proprietary MLLMs (API)							
GPT-6 Astra	**87.9**	**79.0**	70.4	<u>78.2</u>	**61.1**	0.224	**39.8**
GPT-5.5	85.4	69.7	70.1	69.8	20.9	<u>0.336</u>	35.9
GPT-5.4	<u>87.0</u>	72.5	68.9	64.0	<u>21.8</u>	**0.346**	33.5
General-Purpose Open-Source MLLMs:							
Qwen2.5-VL-3B	58.1	56.1	41.0	23.5	1.7	0.169	24.3
Qwen2.5-VL-32B	67.5	61.9	16.8	29.6	5.9	0.143	19.0
Qwen3-VL-2B	64.3	60.2	64.8	75.6	5.9	0.078	22.8
Qwen3-VL-8B	61.7	66.0	66.5	62.7	6.3	0.023	32.0
Qwen3-VL-30B-A3B	71.1	72.8	69.2	75.4	8.8	0.099	<u>36.3</u>
InternVL3-8B	57.8	53.1	30.8	23.0	4.2	0.100	23.8
InternVL3.5-8B	59.7	61.9	61.6	**78.4**	3.8	0.108	19.6
Ovis2.5-9B	78.9	<u>75.2</u>	<u>75.4</u>	69.9	4.6	0.176	31.2
LLaVA-OneVision-1.5-8B	54.6	49.0	**84.7**	75.3	2.5	0.003	28.3
Gemma-3-27B	65.9	61.2	36.0	33.3	1.7	0.166	18.0
Kimi-VL-A3B	65.6	66.0	55.6	46.0	3.3	0.053	31.1
DeepSeek-VL2-small	53.6	24.2	0.0	0.0	3.8	0.000	10.7
Embodied-Specific MLLMs:							
Magma-8B	39.9	39.5	28.1	32.0	1.7	0.021	23.7

Notes. Best and second-best results are marked in bold and underlined.

Figure 4 further reveals imbalanced capability profiles. Models that perform well on task and tool recognition do not necessarily generalize to physical interaction or procedural reasoning, and no single model consistently dominates all dimensions. Together with the absolute scores in Table 2, these profiles identify temporal and procedural reasoning as the most pronounced bottleneck. Figure 5 provides instance-level evidence for these observations. Models can often infer the high-level task or tool from salient visual cues, yet still predict an incorrect force-bearing hand, force trend, hand pose, or action order. Some predictions are only partially correct, indicating that recognizing individual visual cues does not guarantee a coherent understanding of the complete tool-use process.

5.2 BEYOND RECOGNITION: LEARNING INTERACTION KNOWLEDGE

Table 3(a) reveals a clear complementarity between vision and touch. Touch is substantially stronger for grasp reasoning, reaching up to 68.2% on grasp-type prediction compared with the 14.3% chance level, and also performs better on force-related tasks. In contrast, vision dominates semantic and procedural reasoning, reaching up to 86.6% for current-action recognition and 58.6% for next-action prediction. Combining both modalities further raises grasp-type accuracy to 92.8%–100.0% and consistently improves force-bearing-hand prediction, confirming the benefit of complementary multimodal cues. However, fusion does not always improve force-trend prediction, suggesting that effective visual–tactile integration remains challenging. Temporal shuffling mainly degrades dynamic predictions while preserving grasp recognition, further indicating that temporal alignment is critical for interaction dynamics but less so for static contact geometry.

Table 3(b) further demonstrates cross-dataset transfer. Models fine-tuned only on PalmDex improve across all 12 task–backbone combinations when evaluated directly on EgoTouch without using any EgoTouch training data. The gains reach 18.5 points for grasp type, 20.3 for force-bearing hand, and 23.1 for force trend, compared with only 0.8–4.3 points for tool recognition. PalmDex therefore

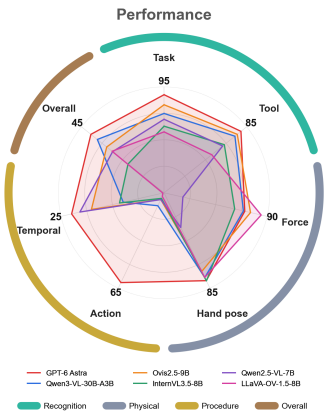

Figure 4: Task-wise capabilities of representative models.

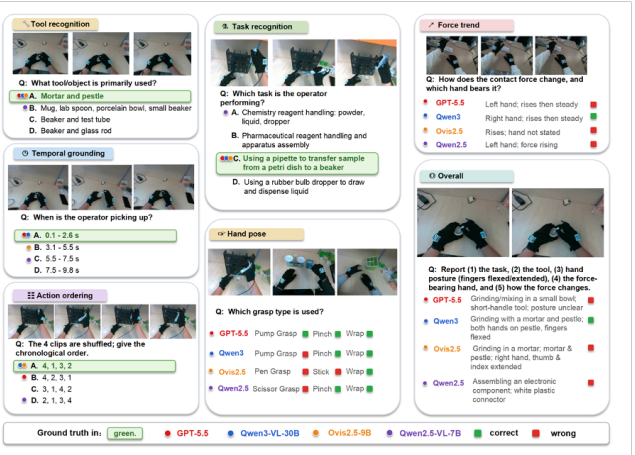

Figure 5: Representative qualitative results on PalmDex. Green and red markers indicate correct and incorrect predictions.

Table 3: Interaction-knowledge accuracy↑ (%) on (a) PalmDex and (b) EgoTouch.

(a) PalmDex in-domain evaluation

		Qwen-3B				Qwen-7B				InternVL-8B			
ID	Task	V	T	V+T	S	V	T	V+T	S	V	T	V+T	S
T1	Tool Recognition	–	**71.4**	–	66.7	–	75.0	–	**76.2**	–	**76.2**	–	67.9
T2	Grasp Type	91.5	68.2	**97.9**	94.1	97.9	41.1	**100.0**	87.7	**94.1**	45.3	92.8	88.1
T3	Force-bearing Hand	57.9	64.0	**65.8**	43.0	55.3	64.0	**76.3**	61.4	57.9	61.4	**63.2**	56.1
T4	Force Trend	41.3	45.3	**48.3**	–	31.7	**41.7**	28.0	–	43.7	**45.7**	42.7	–
T5	Current Action	**67.0**	33.9	**67.0**	61.6	**86.6**	36.6	67.9	70.5	65.2	27.7	67.0	**69.6**
T6	Step Completion	57.5	47.5	**60.0**	55.0	70.0	45.0	67.5	**72.5**	**67.5**	**67.5**	65.0	62.5
T7	Next Action	**58.6**	15.5	**58.6**	17.2	**58.6**	20.7	46.6	43.1	44.8	19.0	**51.7**	46.6

(b) EgoTouch cross-dataset transfer

		Qwen-3B		Qwen-7B		InternVL-8B	
ID	Task	Zero-shot	PalmDex-FT	Zero-shot	PalmDex-FT	Zero-shot	PalmDex-FT
T1	Tool Recognition	11.1	**14.8** ↑ 3.7	25.9	**26.7** ↑ 0.8	24.1	**28.4** ↑ 4.3
T2	Grasp Type	22.2	**40.7** ↑ 18.5	20.1	**32.2** ↑ 12.1	15.2	**29.6** ↑ 14.4
T3	Force-bearing Hand	44.5	**63.7** ↑ 19.2	48.4	**68.7** ↑ 20.3	43.1	**55.6** ↑ 12.5
T4	Force Trend	31.5	**33.3** ↑ 1.8	24.1	**45.2** ↑ 21.1	35.2	**58.3** ↑ 23.1

Notes. V: video; T: touch; V+T: both modalities; S: temporally shuffled touch. Green values indicate absolute improvements over zero-shot performance in percentage points. Qwen-3B/7B denote Qwen2.5-VL-3B/7B, respectively.

transfers primarily through robot-free knowledge of grasp, contact, and force, rather than dataset-specific semantic vocabularies.

5.3 Cross-Modal Knowledge Alignment

We evaluate whether synchronized demonstrations induce shared representations of visual and tactile interaction states. Given a query from one modality, the model retrieves its temporally corresponding sample from the other modality, and is evaluated in both vision-to-touch and touch-to-vision directions. Table 4 reports the cross-modal retrieval results on PalmDex. Compared with CCA Rasiwasia et al. (2010) and PLS-CA Jong et al. (2001) , HandTouchHuang et al. (2026) achieves substantially stronger performance across all retrieval metrics. Furthermore, HandTouch uses head-only adaptation, whereas HandTouch† jointly adapts the tactile encoder and retrieval head; SigLIP remains frozen in both settings. Qualitative retrieval results are presented in Figure 6.

5.4 Motion Prior Learning

Table 14 shows that learned temporal models substantially outperform linear and Kalman baselines, particularly at $H = 90$. CVAE performs best across all metrics, reducing wrist-trajectory errors

Table 4: Bidirectional cross-modal retrieval on PalmDex.

Method	V→T↑(%)				T→V↑(%)				Mean mAP↑(%)
	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	
PLS-CA	0.09	0.40	0.81	0.63	0.07	0.41	0.75	0.49	0.56
CCA	1.79	7.60	13.57	6.13	2.26	10.12	17.46	7.64	6.88
HandTouch	6.35	22.55	34.99	15.27	6.32	20.95	33.02	14.87	15.07
HandTouch†	**14.22**	**39.70**	**53.07**	**26.57**	**14.18**	**40.38**	**53.25**	**26.63**	**26.60**

Notes. V→T and T→V denote video-to-tactile and tactile-to-video retrieval, respectively.

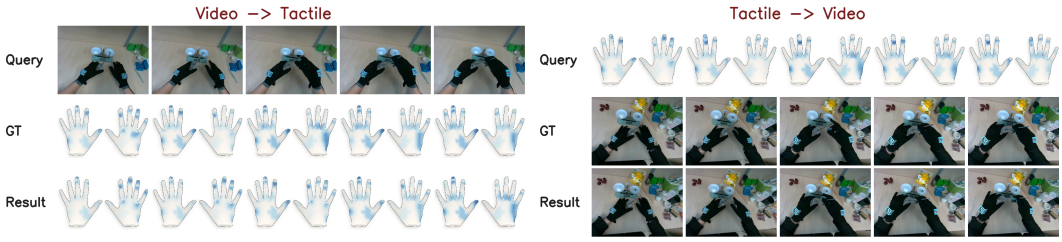

Figure 6: Visualization of cross-modal retrieval results

by 18.0%–22.6% and finger-joint errors by 3.5%–6.3% relative to the strongest competing baselines. Errors increase with the prediction horizon for all methods, reflecting accumulated motion uncertainty. DDPM remains less competitive under the current setting, while CVAE's consistent advantage highlights the importance of modeling multiple plausible tool-use motions. Figure 7shows the motion prediction results.

Table 5: Motion prediction on PalmDex. Full results are provided in Appendix Table 14.

Model	Wrist Trajectory (*px*)↓				Finger Joints (*deg*)↓			
	Avg H_{30}	Final H_{30}	Avg H_{90}	Final H_{90}	Avg H_{30}	Final H_{30}	Avg H_{90}	Final H_{90}
RNN	6.31	10.02	9.44	13.66	4.83	6.36	6.45	7.97
LSTM	6.33	10.02	9.49	13.69	4.45	5.78	6.03	7.47
GRU	6.33	10.00	9.35	13.50	4.50	5.85	6.07	7.51
Seq2Seq	6.25	9.91	9.28	13.33	4.43	5.64	6.03	7.40
iTransformer	6.42	10.09	9.14	13.17	4.55	6.17	6.47	8.32
CVAE	**4.82**	**7.70**	**7.44**	**10.80**	**4.15**	**5.44**	**5.74**	**7.07**

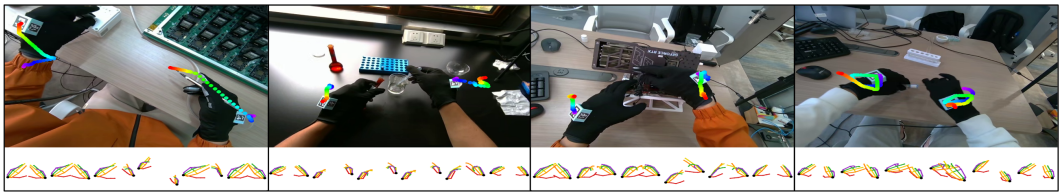

Figure 7: Visualization of hands motion prediction. Top: both wrists' image-plane trajectories, coloured by time from blue to red. Bottom: the corresponding dual-hand finger-skeleton sequence.

6 Conclusion

We presented PalmDex, a robot-free multimodal dataset and benchmark for functional tool use, integrating synchronized vision, full-hand touch, 3D hand motion, and procedural annotations. Our results show that current models remain limited in temporal and physical interaction reasoning, while vision and touch provide complementary cues for understanding procedures, grasping, contact, and force. PalmDex supervision further transfers interaction knowledge across datasets, demonstrating its generalization potential. By jointly capturing semantic, physical, and motion information, PalmDex provides a unified foundation for learning not only what tool use looks like, but how it is performed.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing and for visual refinement of selected elements in Figure 1. We did not use generative AI for generating research ideas, experimental data, or scientific conclusions. Several multimodal large language models were also evaluated as experimental baselines in our benchmark. All AI-assisted revisions were reviewed by the authors, and we take full responsibility for the final content of this work.

REFERENCES

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. HOT3D: Hand and object tracking in 3D from egocentric multi-view videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7061–7071, 2025.

Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

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 *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

Eadom Dessalene, Botao He, Michael Maynord, Yonatan Tussa, Pavan Mantripragada, Yianni Karabati, Nirupam Roy, and Yiannis Aloimonos. FEEL (force-enhanced egocentric learning): A dataset for physical action understanding. *arXiv preprint arXiv:2603.15847*, 2026.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18973–18990, 2022.

Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. HOnnotate: A method for 3D annotation of hand and object poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3193–3203, 2020.

Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. EgoDex: Learning dexterous manipulation from large-scale egocentric video. In *International Conference on Learning Representations (ICLR)*, 2026.

Yuzhe Huang, Jiaping Wu, Jiaming Jiang, Hezhe Lin, Aikebaier Aierken, Yunlong Wang, Kun Cheng, Wanlin Li, Chenxi Xiao, Ziyuan Jiao, and Yuanxin Zhong. HT-Bench: Benchmarking and learning dexterous full-hand tactile representations with egocentric vision. *arXiv preprint arXiv:2606.19161*, 2026. URL https://arxiv.org/abs/2606.19161.

Ahad Jawaid and Yu Xiang. OpenEgo: A large-scale multimodal egocentric dataset for dexterous manipulation. *arXiv preprint arXiv:2509.05513*, 2025.

Sijmen De Jong, Barry M. Wise, and N. Lawrence Ricker. Canonical partial least squares and continuum power regression. *John Wiley & Sons, Ltd.*, 15(2):85–100, 2001.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, et al. DROID: A large-scale in-the-wild robot manipulation dataset. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. HOI4D: A 4D egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Open X-Embodiment Collaboration. Open X-embodiment: Robotic learning datasets and RT-X models. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Kumar Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, Jacob Chalk, Zhifan Zhu, Rhodri Guerrier, Fahd Abdelazim, Bin Zhu, Davide Moltisanti, Michael Wray, Hazel Doughty, and Dima Damen. HD-EPIC: A highly-detailed egocentric video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23901–23913, 2025.

Ryan Punamiya, Simar Kareer, Zeyi Liu, Josh Citron, Ri-Zhao Qiu, Xiongyi Cai, Alexey Gavryushin, Jiaqi Chen, Davide Liconti, Lawrence Y. Zhu, et al. EgoVerse: An egocentric human dataset for robot learning from around the world. *arXiv preprint arXiv:2604.07607*, 2026.

Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R. G. Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM International Conference on Multimedia*, pp. 251–260, 2010. doi: 10.1145/1873951.1873987.

Yuxin Ray Song, Jinzhou Li, Rao Fu, Devin Murphy, Kaichen Zhou, Rishi Shiv, Yaqi Li, Haoyu Xiong, Crystal Elaine Owens, Yilun Du, Yiyue Luo, Xianyi Cheng, Antonio Torralba, Wojciech Matusik, and Paul Pu Liang. OpenTouch: Bringing full-hand touch to real-world interaction. *arXiv preprint arXiv:2512.16842*, 2025.

Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Maximilian Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. BridgeData V2: A dataset for robot learning at scale. In *Conference on Robot Learning (CoRL)*, 2023.

Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, Xiaozhu Ju, Zhuqin Yang, Meng Li, Yinuo Zhao, Zhiyuan Xu, Guang Yang, et al. RoboMIND: Benchmark on multi-embodiment intelligence normative data for robot manipulation. In *Proceedings of Robotics: Science and Systems (RSS)*, 2025.

Yiming Zhao, Taein Kwon, Paul Streli, Marc Pollefeys, and Christian Holz. EgoPressure: A dataset for hand pressure and pose estimation in egocentric vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Ruijie Zheng, Dantong Niu, Yuqi Xie, Jing Wang, Mengda Xu, Yunfan Jiang, Fernando Castañeda, Fengyuan Hu, You Liang Tan, Letian Fu, Trevor Darrell, Furong Huang, Yuke Zhu, Danfei Xu, and Linxi Fan. EgoScale: Scaling dexterous manipulation with diverse egocentric human data. *arXiv preprint arXiv:2602.16710*, 2026a.

Yuhang Zheng, Songen Gu, Weize Li, Yupeng Zheng, Yujie Zang, Shuai Tian, Xiang Li, Ce Hao, Chen Gao, Si Liu, Haoran Li, Yilun Chen, Shuicheng Yan, and Wenchao Ding. OmniVTA: Visuotactile world modeling for contact-rich robotic manipulation. *arXiv preprint arXiv:2603.19201*, 2026b.

Jianyi Zhou, Ziteng Gao, Feiyang Hong, Zirui Liu, Guannan Zhang, Weisheng Dai, Ruichen Zhen, Chuqiao Lyu, Haotian Wu, Yinian Mao, Xushi Wang, Yuxiang Jiang, Wenbo Ding, and Shuo Yang. TouchAnything: A dataset and framework for bimanual tactile estimation from egocentric video. *arXiv preprint arXiv:2605.13083*, 2026.

A Dataset Details

A.1 Task And Data Statistics

PalmDex is recorded as 30 sessions that instantiate 18 tool-use task families across three domains: laboratory sciences, electronics assembly and inspection, and kitchen, office, and everyday use. The families draw on 50 tools and task-relevant objects, exercised through 80 fine-grained manipulation actions that reduce to 53 verbs. Its 2,696 episodes are temporally segmented into 12,399 labeled tool-use steps and total roughly 31 hours, captured as two synchronized 30 Hz video views—about

1.8M frames per view—together with a 256-channel full-hand tactile signal and 72-dimensional finger-joint kinematics. Table 6 breaks the counts down by domain, and Table 7 gives the contents of one record.

Table 6: PalmDex dataset statistics by task domain. Tools & objects and manipulation actions are distinct counts, so domain rows may share entries and need not sum to the total. Sensor modalities and recording scale are summarized in the text.

Domain	Real-world tasks	Tools & objects	Manipulation actions	Episodes	Steps
Laboratory sciences	13	34	65	1,922	8,486
Electronics	3	12	15	580	3,138
Kitchen, office & everyday	43	37	49	569	1150
Total	**59**	**80**	**121**	**3,071**	**12,774**

Table 7: Contents of one record. Each tool-use step is grounded jointly in five synchronized sensor streams and carries human-annotated per-step labels: the sensor streams are recorded automatically, and the step boundaries and semantic labels are human-annotated.

Field	Content	Source
Egocentric video	First-person RGB frames	sensor (ego camera)
Exocentric video	Third-person RGB frames	sensor (exo camera)
Hand kinematics	21-joint / 72-D 3D hand pose	sensor (UDEXREAL glove)
Full-hand tactile	256-channel contact pressure	sensor (Juqiao glove)
Step boundaries	Temporal segmentation into tool-use steps	human-annotated
Task label	Task family / description	human-annotated
Action label	Per-step action verb	human-annotated
Tool / object	Tool and object in contact	human-annotated

A.2 Data Collection

Each demonstration is recorded with five synchronized modalities, summarized in Table 8. All streams share a common clock and are aligned to a single 30 Hz timeline through a unified timestamp pipeline, so every egocentric frame has a corresponding exocentric frame, tactile reading, and hand-kinematics sample.

Table 8: Recording modalities of the PalmDex capture system. All streams are synchronized to a shared 30 Hz timeline.

Modality	Device	Rate	Dimension / Resolution	Output
Egocentric RGB	Intel RealSense D405	30 Hz	640×360	RGB video
Exocentric RGB	Intel RealSense D405	30 Hz	640×360	RGB video
Hand kinematics	UDEXREAL glove	30 Hz	72-D / 21 joints	3D hand pose
Full-hand tactile	Juqiao tactile glove	30 Hz	256 channels	contact-pressure signal

B Benchmark Details

This section details how the questions in each benchmark are formed, sampled, and evaluated. For each benchmark we give its question prototype, describe the input given to the model and how the ground-truth answer is derived, and—for the multiple-choice benchmarks—how the distractors are constructed. Placeholders in a prototype are typeset in `monospace`.

All benchmarks share a single, episode-level 80:10:10 train/validation/test partition. The split is defined once over the global set of episodes and is stratified by data source; every question inherits

the split of the episode it originates from. Consequently an episode assigned to the test set is held out across *all* benchmarks, which precludes cross-benchmark leakage: no episode in the training portion of one benchmark appears in the test portion of another. The split is deterministic under a fixed seed and therefore reproducible.

B.1 TOOL-USE COGNITION

Tool-Use Cognition comprises seven video question-answering benchmarks that probe complementary aspects of tool-use understanding: *task recognition*, *tool recognition*, *temporal grounding*, *force-trend estimation*, *hand-pose description*, *action ordering*, and a comprehensive *overall* task. We give the question prototype and ground-truth construction of each below. All seven benchmarks follow the shared episode-level 80:10:10 partition of Appendix B and generate VQA questions from it at their own granularity, one per episode or one per manipulation step, so the per-benchmark counts differ; they are reported in Table 9.

Table 9: Tool-Use Cognition benchmark scale (question–answer pairs).

Benchmark	Train	Val	Test	Total
Task recognition	2,144	255	297	2,696
Tool recognition	2,156	270	270	2,696
Temporal grounding	6,027	761	714	7,502
Force trend	5,730	759	684	7,173
Hand pose	2,156	270	270	2,696
Action ordering	1,945	247	239	2,431
Overall	5,730	759	684	7,173
Total	**25,888**	**3,321**	**3,158**	**32,367**

B.1.1 QUESTION PROTOTYPES

This section details the generation of QA pairs for each task within the Tool-use cognition benchmark. The 8 shows an "input–question–answer" example for this benchmark.

Task Recognition. As each episode carries out one high-level activity, the question prototype "Which task is the operator performing? A. `<option>` ... D. `<option>`" tests a model's ability to recognise that activity from a full operation video. The model receives one full-episode video and selects a single option. The positive answer is the task description associated with the episode, taken from the dataset metadata. To form negatives, we sample three distractor task descriptions from the task vocabulary; question difficulty is controlled by the number of *strong* (semantically close) distractors, giving easy, medium, and hard splits.

Tool Recognition. To assess whether a model can identify the object being manipulated, the prototype "What tool or object is the operator primarily using? A. `<option>` ... D. `<option>`" is posed over a single manipulation step. The model receives one segment clip and selects a single option. The positive answer is the object annotated for that segment. To form negatives, we sample three tools from other segments; hard distractors are drawn from segments recorded in the same scene domain, so that the alternatives remain visually and functionally plausible.

Temporal Grounding. To evaluate temporal localisation, the prototype "When is the operator `<action>`?" asks the model to return the time span of the queried action as an interval in seconds. The model receives the full episode video, and the positive answer is the annotated start and end time of the segment in which the action occurs. As the task is open-ended localisation, no distractors are provided; predictions are scored by temporal Intersection-over-Union against the ground-truth interval. Questions are stratified by the duration of the target action ($<$3 s, 3–8 s, 8–15 s, $>$15 s).

Force Trend. To probe the understanding of contact dynamics, the prototype "How does the grip force change during this step, and which hand bears the main force?" asks the model to describe the



Figure 8: Representative Examples of the PalmDex Tool-Use Cognition Benchmark.

force evolution with the terms 'rises', 'stays steady', and 'falls', combined in order when the force passes through several phases. The model receives one manipulation segment (a whole episode for EgoTouch). The positive answer is the ordered force-phase sequence together with the dominant hand, derived from the synchronised 256-channel tactile glove: the phase sequence from the temporal profile of the aggregated contact pressure, and the dominant hand from the side bearing the larger cumulative force. The question is open-ended and has no distractors.

Hand Pose. To probe the perception of how a tool is held, the prototype "How is the operator holding the tool?" asks the model to answer the operator's hand pose. The model receives one segment clip—one episode for EgoTouch. The positive answer is the hand pose of the operating hand, derived from the synchronised tactile and kinematic streams: the fingers in contact are those whose force exceeds the per-episode median, and the operating hand is read from the tactile force channels. The question has no multiple-choice distractors.

Action Ordering. To assess temporal reasoning over an entire task, the prototype "The clips are the steps of one task shown in a random order; determine their correct chronological order" presents N segment clips ($N \in [3, 8]$) from a single episode in a randomly permuted order. The model receives the N clips and returns a permutation of their indices. The positive answer is the true chronological order, recovered from the segment timestamps; the random presentation order serves as the confound, and question difficulty grows with the number of steps N.

Overall. To measure joint understanding, the prototype asks the model, from a single clip, to report "(1) the task, (2) the tool used, (3) the hand pose, (4) the hand bearing the main contact force, and (5) how that force changes". The model receives one segment clip. The positive answer is assembled by joining the ground truth of the constituent benchmarks for the same clip: the episode task label, the segment object, the hand pose from the tactile and kinematic streams, and the dominant hand and force-phase sequence from the tactile glove. The question is open-ended and has no distractors.

B.1.2 Prompt Templates and Scoring

Visual Prompts. PalmDex does not use visual prompts: no point, box, or mask marker is overlaid on any frame. The target object and action are referred to purely in the question text, and each model receives the raw clip, uniformly sampled into frames according to its own default frame budget.

Text Prompts. All models are evaluated zero-shot with identical text prompts, each prefixed with the video token(s) and with no in-context examples. The user prompt is instantiated per answer format as follows, where {Question} is the benchmark template of Appendix B.1 and {Options} the options (multiple-choice only):

```
Task recognition
USER: {Question} Options: {Options} Answer directly with only the
letter of the correct option (A, B, C, or D).

Tool recognition
USER: {Question} Options: {Options} Answer directly with only the
letter of the correct option (A, B, C, or D).

Temporal grounding
USER: {Question} Answer with the start and end time in seconds,
wrapped as <answer>(start, end)</answer>.

Force trend
USER: {Question} Answer directly in natural language, using 'rises',
'stays steady', and 'falls', and name the hand bearing the main force.

Hand pose
USER: {Question} Answer directly in natural language, describing the
operator's hand pose.

Action ordering
USER: {Question} Answer with the clip indices in chronological order,
separated by commas, e.g. 3,1,4,2.

Overall
USER: {Question} Answer directly in natural language, following the
structure requested in the question.
```

B.2 Interaction Knowledge

The modality ablation in the main paper recasts a subset of PalmDex into a unified-input, closed-set probing suite. Unlike the Tool-Use Cognition benchmarks above, which take a whole video and are open-ended or multiple-choice, every probe here takes the same fixed-size multimodal input and is answered from a small closed label set, so that a single accuracy is directly comparable across input modalities. This subsection specifies the seven probes, the input, the answer format, and the fine-tuning configuration.

B.2.1 Task Definitions and Label Construction

The seven probes span the four capabilities that tool use requires in turn: recognising the tool, forming the grasp, perceiving the contact force, and following the task procedure. All labels are derived from the synchronised tactile-glove and kinematic streams, never from the video.

T1 – Tool Recognition under masked vision. Given only tactile input, the model names the tool in contact from the closed set of tools that occur in the same scene domain. The label is the object annotated for the segment.

T2 – Grasp Type. The model answers the operator's hand pose. The label is obtained from a tool-to-grasp mapping and refined by a contact filter that retains the fingers whose force exceeds the per-episode median.

T3 – Force-bearing Hand. The model decides which hand bears the main contact force, choosing among *left*, *right*, and *none*. The label is the side with the larger cumulative contact force over the window, and *none* is assigned when both hands stay below a small contact threshold.

T4 – Force Trend. From the past $1.5\,\mathrm{s}$ of input, the model predicts how the contact force will evolve over the next $1\,\mathrm{s}$, choosing among *rising*, *falling*, *flat*, and *rise-then-fall*. The label is read from the slope of the aggregated contact-pressure profile in the future window. As the future is only sometimes determined, we additionally report on the physically-predictable subset, whose test partition contains 300 samples.

T5 – Current Action. The model recognises the action being performed at the reference instant from a ten-way action vocabulary, and the label is the segment action at that time.

T6 – Step Completion. The model decides whether the current step has been completed, answering *yes* or *no*; the label is whether the segment end is reached by the reference instant.

T7 – Next Action. The model predicts the action of the following step from a nine-way vocabulary, and the label is the action of the next segment.

B.2.2 Fine-tuning Configuration

We fine-tune Qwen2.5-VL-3B-Instruct with LoRA of rank $r{=}8$, scaling $\alpha{=}16$, and dropout 0.05, applied to the q, k, v, and o attention projections of the language decoder while the vision encoder is kept frozen. Only about 3.7M parameters are updated, amounting to 0.1% of the model. We optimise the answer-token cross-entropy with AdamW at a learning rate of $1{\times}10^{-4}$ for three epochs in bf16, capping the per-image pixel budget for memory. A separate adapter is trained for each probe–arm combination. All models reuse the episode-level $80{:}10{:}10$ split of Appendix B and are evaluated on the held-out test episodes, so no episode seen in training appears at test time. Figure 9 shows the pipeline, and Table 10 lists the hyper-parameters.

Table 10: Fine-tuning hyper-parameters for the unified-input probes. A separate adapter is trained per probe–arm combination.

Setting	Value
Backbone	Qwen2.5-VL-3B / 7B, InternVL-8B
Input per sample	4 ego frames + 4 dual-hand tactile heatmaps
Vision encoder	frozen
LoRA r / α / dropout	8 / 16 / 0.05
Target modules	q, k, v, o projections
Trainable parameters	3.7 M (0.1%)
Optimizer	AdamW
Learning rate	$1{\times}10^{-4}$
Epochs	3
Precision	bf16
Per-image budget	capped (`max_pixels`)
Tactile representation	dual-hand heatmaps

B.3 Cross-Modal Alignment

This benchmark evaluates whether temporally synchronized video and tactile observations can be aligned in a shared representation space. We perform bidirectional video-to-tactile and tactile-to-video retrieval under single-hand and dual-hand tactile input settings. Both settings use the same data partition, visual observations, retrieval architecture, training protocol, and evaluation procedure.

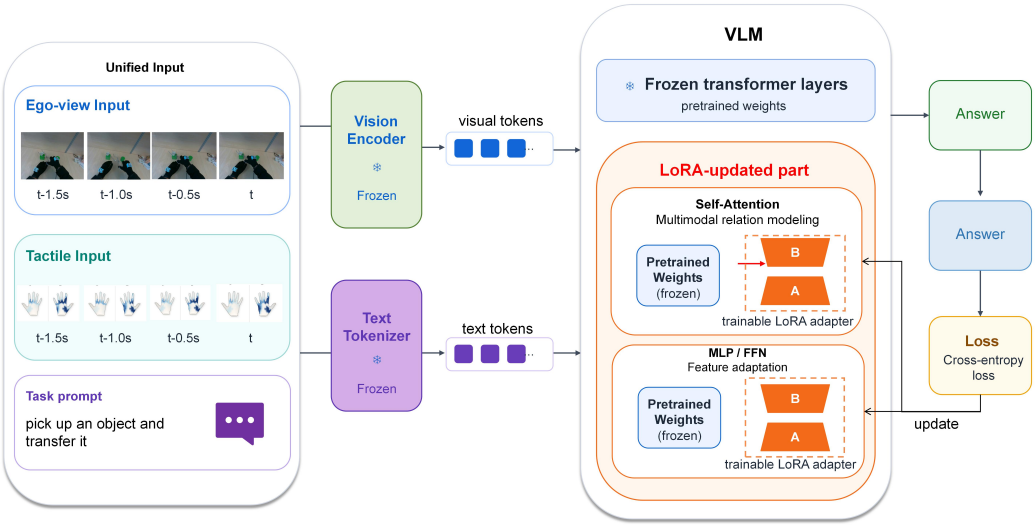

Figure 9: Fine-tuning pipeline for unified-input probes. Egocentric frames and aligned tactile heatmaps are combined into a multimodal prompt, with the vision encoder frozen and LoRA used for adaptation. Video, touch, and video+touch settings differ only in their input modalities.

B.3.1 Task and Input Construction

All experiments use a shared episode-level split that yields 66,980 training windows, 8,639 validation windows, and 8,532 test windows. The split is applied before window construction, so all windows from one episode remain in the same subset. Frame records are grouped by scene, action, and sequence identifier and divided at discontinuous frame indices. We extract non-overlapping 20-frame windows with a stride of 20. Each sample contains a center-camera RGB sequence and its temporally synchronized tactile matrices; windows with incomplete observations are removed.

Each window is identified by its scene, action, sequence identifier, and temporal boundaries. A video window and a tactile window form a positive pair only when they share the same window identifier. All other windows in the retrieval pool are negatives, leaving one positive candidate for each query.

Each RGB frame is resized to 384×384 with bicubic interpolation and normalized by $x/127.5 - 1$. In the single-hand setting, the video sequence is paired with synchronized right-hand tactile observations. In the dual-hand setting, synchronized tactile matrices from the left and right hands are provided jointly as the tactile input. The two settings therefore differ only in whether the tactile observation contains one hand or both hands.

B.3.2 Model and Training Setup

The visual branch uses a frozen SigLIP So400m/patch14-384 encoder. Mean pooling over its 729 patch tokens produces a 1,152-dimensional feature for each video frame. The tactile branch is initialized from the Stage1 encoder of HandTouch Huang et al. (2026) and receives the tactile matrices specified by the hand setting.

The visual retrieval head averages the 20 frame-level features and projects the resulting vector from 1,152 to 64 dimensions. The tactile retrieval head applies a two-layer bidirectional GRU with a hidden size of 120 per direction, concatenates its final forward and backward states, and projects the resulting 240-dimensional vector to 64 dimensions. Both embeddings are L2-normalized, and their dot product gives the cosine similarity used for training and retrieval.

Given a candidate set of $B = 128$ synchronized pairs, Equation 3 defines the symmetric InfoNCE objective used for training:

$$S_{ij} = \frac{(\mathbf{z}_i^v)^\top \mathbf{z}_j^t}{\tau}, \qquad \tau = 0.07,$$
$$\mathcal{L} = -\frac{1}{2B} \sum_{i=1}^{B} \left[\log \frac{\exp(S_{ii})}{\sum_{j=1}^{B} \exp(S_{ij})} + \log \frac{\exp(S_{ii})}{\sum_{j=1}^{B} \exp(S_{ji})} \right]. \tag{3}$$

where the two terms normalize over tactile and video candidates, respectively, and the diagonal entries of S are the synchronized positive pairs.

For HandTouch, both the SigLIP and tactile encoders are frozen, and only the retrieval head is optimized with a learning rate of 10^{-4}. For HandTouch†, the SigLIP encoder remains frozen, while the tactile encoder and retrieval head are jointly fine-tuned with learning rates of 10^{-5} and 10^{-4}, respectively. We use AdamW with a weight decay of 0.01. Both settings are trained for 300 epochs with a five-epoch linear warm-up followed by cosine decay.

B.4 Motion Prior

The motion-prior benchmark predicts task-conditioned hand motion from a short egocentric history. From the past 30 frames—1 s at 30 Hz—the model forecasts the next H frames of the active hand, at horizons H=30 (1 s) and H=90 (3 s). Two targets are evaluated: the wrist trajectory in the image plane, as a 2-D pixel sequence scored by average and final displacement error (ADE / FDE, in pixels); and the finger configuration, as 21 joint angles per hand scored by mean joint-angle error (MJAE, in degrees). Targets are expressed as a per-step delta from the last observed frame and are per-channel normalised. Under one training recipe we benchmark nine predictors: a Kalman filter, an RNN, a Seq2Seq model, a TCN, iTransformer, DLinear, SiMLPe, a DDPM, and a CVAE; the neural models are trained with Adam at 1×10^{-3} for 15 epochs. Table 11 lists the configuration, and Figure 11 presents qualitative predictions on twenty-five held-out tool-use episodes spanning both targets: each panel overlays both wrists' image-plane trajectories, coloured by time, together with the corresponding dual-hand finger-skeleton sequence, showing that the forecast motion tracks the true reach, grasp, and manipulation phases across diverse tools and at both horizons.

Table 11: Motion-prior benchmark configuration. All predictors share the same history, horizons, targets, and split.

Setting	Value
Input history	30 frames (1 s @ 30 Hz), stride 15
Horizons H	30 (1 s) / 90 (3 s)
Targets	wrist uv (2-D, px) ; joint angles (21-D, deg)
Target Encoding	delta-from-last, per-channel normalised
Predictors	Kalman, RNN, Seq2Seq, TCN, iTransformer, DLinear, SiMLPe, DDPM, CVAE
Optimizer	Adam, 1×10^{-3}
Epochs	15

B.5 Metrics Details

Question Answering. For the question-answering tasks in Tool-Use Cognition and Interaction Knowledge, we report accuracy:

$$\mathrm{Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(\hat{y}_i = y_i\right), \tag{4}$$

where y_i and $\hat{y}_i$ denote the ground-truth and predicted answers, respectively.

Temporal Grounding. For key-stage localization, the model predicts a temporal interval $\hat{\mathcal{I}} = [\hat{t}_s, \hat{t}_e]$, which is compared with the ground-truth interval $\mathcal{I} = [t_s, t_e]$ using temporal intersection

over union:

$$\mathrm{tIoU} = \frac{\left|\hat{\mathcal{I}} \cap \mathcal{I}\right|}{\left|\hat{\mathcal{I}} \cup \mathcal{I}\right|}. \tag{5}$$

A higher tIoU indicates more accurate localization of the key tool-use stage.

Cross-Modal Retrieval. We evaluate bidirectional retrieval between visual and tactile observations using Recall@K:

$$\mathrm{R@}K = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left(r_i \le K\right), \tag{6}$$

where r_i is the rank of the ground-truth match for query i. We report R@1 and R@5 in both vision-to-touch and touch-to-vision directions.

Wrist Trajectory Prediction. Given an input image, the model predicts the future trajectory of the wrist end-effector. We measure trajectory accuracy using average displacement error (ADE) and final displacement error (FDE):

$$\mathrm{ADE} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \left\|\hat{\mathbf{p}}_{i,t}^{w} - \mathbf{p}_{i,t}^{w}\right\|_2, \tag{7}$$

$$\mathrm{FDE} = \frac{1}{N}\sum_{i=1}^{N} \left\|\hat{\mathbf{p}}_{i,T}^{w} - \mathbf{p}_{i,T}^{w}\right\|_2, \tag{8}$$

where $\hat{\mathbf{p}}_{i,t}^{w}$ and $\mathbf{p}_{i,t}^{w}$ are the predicted and ground-truth wrist positions at time t. ADE measures the average trajectory deviation, while FDE measures the endpoint error.

Hand Pose Prediction. Given an image, an initial pose, and a predefined 3D hand template, the model predicts the joint configuration in a wrist-centered coordinate system. We report mean joint angle error (MJAE):

$$\mathrm{MJAE} = \frac{1}{NTJ}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{J} \left\|\hat{\boldsymbol{\theta}}_{i,t,j} - \boldsymbol{\theta}_{i,t,j}\right\|_1, \tag{9}$$

where $\boldsymbol{\theta}_{i,t,j}$ denotes the joint angle parameters of joint j. This wrist-centered evaluation removes global hand translation and focuses on the accuracy of the predicted hand articulation.

C Additional Experimental Results

C.1 Full Zero-Shot Leaderboard On Tool-Use Cognition

We evaluate 26 representative MLLMs, including proprietary, general-purpose open-source, and embodied-specific models. Table 12 reports the complete zero-shot leaderboard on the Tool-Use Cognition benchmark, extending the condensed results of Table 2 to the full set of evaluated models. GPT-6 Astra achieves the best overall score of 39.8%, while GPT-5.4 remains strongest on temporal grounding. Among open-source models, Qwen3-VL-30B-A3B obtains the highest overall score of 36.3%, with different models leading force-trend and hand-pose prediction. Overall, the results highlight substantial capability imbalance across models and tasks, suggesting that strong performance on individual dimensions does not yet translate into comprehensive tool-use cognition.

C.2 Cross-Modal Alignment Results

Table 13 reports the complete bidirectional retrieval results for both hand settings. Full fine-tuning performs best under both tactile-input settings. It increases mean mAP from 14.01% to 26.44% with single-hand input and from 15.07% to 26.60% with dual-hand input. CCA performs below the two neural alignment variants, while PLS-CA gives the lowest retrieval scores. Dual-hand input

Table 12: Zero-shot tool-use cognition results on PalmDex. All metrics are accuracy (%) except Temp (mIoU).

Model	Task	Tool	Force	Hand pose	Action	Temporal	Overall
General-Purpose Proprietary MLLMs (API)							
GPT-6 Astra	**87.9**	79.0	70.4	78.2	**61.1**	0.224	39.8
GPT-5.5	85.4	69.7	70.1	69.8	20.9	<u>0.336</u>	<u>35.9</u>
GPT-5.4	<u>87.0</u>	72.5	68.9	64.0	<u>21.8</u>	**0.346**	33.5
GPT-5.4-mini	73.1	64.6	68.8	68.9	12.1	0.266	29.7
General-Purpose Open-Source MLLMs							
Qwen2.5-VL-3B	58.1	56.1	41.0	23.5	1.7	0.169	24.3
Qwen2.5-VL-7B	65.9	60.2	16.4	30.9	4.6	0.204	28.0
Qwen2.5-VL-32B	67.5	61.9	16.8	29.6	5.9	0.143	19.0
Qwen2.5-VL-72B	70.1	69.7	60.7	52.1	5.0	0.176	33.3
Qwen3-VL-2B	64.3	60.2	64.8	<u>75.6</u>	5.9	0.078	22.8
Qwen3-VL-4B	69.5	69.1	73.9	46.2	0.8	0.024	31.0
Qwen3-VL-8B	61.7	66.0	66.5	62.7	6.3	0.023	32.0
Qwen3-VL-30B-A3B	71.1	<u>72.8</u>	69.2	75.4	8.8	0.099	**36.3**
InternVL2.5-8B	48.7	49.3	54.4	64.3	7.1	0.145	29.9
InternVL3-8B	57.8	53.1	30.8	23.0	4.2	0.100	23.8
InternVL3.5-8B	59.7	61.9	61.6	**78.4**	3.8	0.108	19.6
Ovis2-8B	68.2	61.2	63.6	32.7	3.4	0.121	31.1
Ovis2.5-9B	78.9	**75.2**	<u>75.4</u>	69.9	4.6	0.176	31.2
LLaVA-NeXT-Video-7B	34.4	34.7	6.1	11.5	5.9	0.014	11.5
LLaVA-OneVision-7B	50.0	47.6	2.7	20.0	1.7	0.081	22.8
LLaVA-OneVision-1.5-8B	54.6	49.0	**84.7**	75.3	2.5	0.003	28.3
MiniCPM-V-2.6	48.1	38.8	8.2	20.8	4.6	0.120	22.2
Gemma-3-27B	65.9	61.2	36.0	33.3	1.7	0.166	18.0
Kimi-VL-A3B	65.6	66.0	55.6	46.0	3.3	0.053	31.1
DeepSeek-VL2-small	53.6	24.2	0.0	0.0	3.8	0.000	10.7
GLM-4V-9B	65.9	53.7	6.8	78.3	4.8	0.031	24.3
Embodied-Specific MLLMs							
Magma-8B	39.9	39.5	28.1	32.0	1.7	0.021	23.7

Notes. Best and second-best results are marked in bold and underlined.

improves mean mAP by 1.06 points for HandTouch and by 0.16 points after full fine-tuning. Fig 10 shows additional cross-modal retrieval results.

Table 13: Full single- and dual-hand cross-modal retrieval results on PalmDex.

Method	V→T(%)				T→V(%)				Mean
	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	mAP(%)
(b) Single-hand (right) tactile input									
PLS-CA	0.08	0.42	0.82	0.54	0.05	0.20	0.42	0.35	0.45
CCA	1.45	6.60	11.38	5.39	2.37	9.97	17.21	7.61	6.50
HandTouch	5.15	19.55	31.85	13.74	5.78	20.38	31.69	14.28	14.01
HandTouch†	**14.02**	**38.79**	**52.49**	**26.05**	**14.60**	**39.77**	**52.58**	**26.83**	**26.44**
(b) Dual-hand tactile input									
PLS-CA	0.09	0.40	0.81	0.63	0.07	0.41	0.75	0.49	0.56
CCA	1.79	7.60	13.57	6.13	2.26	10.12	17.46	7.64	6.88
HandTouch	6.35	22.55	34.99	15.27	6.32	20.95	33.02	14.87	15.07
HandTouch†	**14.22**	**39.70**	**53.07**	**26.57**	**14.18**	**40.38**	**53.25**	**26.63**	**26.60**

Notes. V→T and T→V denote video-to-tactile and tactile-to-video retrieval, respectively.

C.3 Motion Prior Learning Results

Table 14 compares 12 representative motion-prediction models on wrist trajectories and finger-joint configurations. CVAE performs best across all metrics, with average errors of $4.82/7.44$ px for wrist

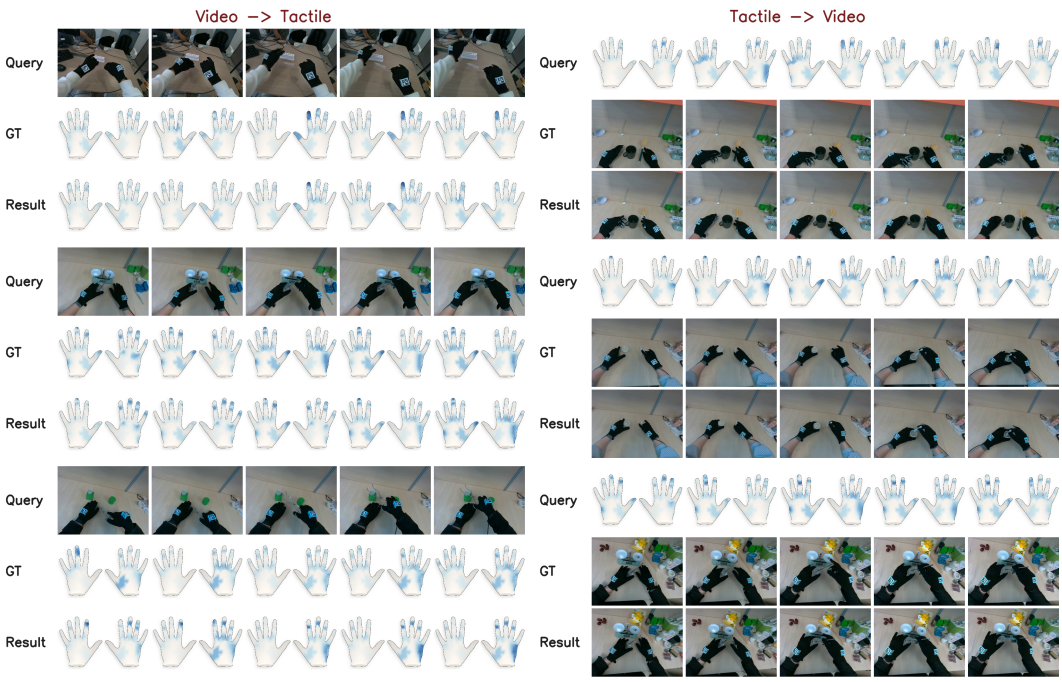

Figure 10: Visualization of cross-modal retrieval results

trajectories and $4.15°/5.74°$ for finger joints at H_{30}/H_{90}. Classical baselines degrade substantially at longer prediction horizons, whereas deep temporal models remain more stable. The superior performance of CVAE further suggests that explicitly modeling the uncertainty of future motion is beneficial for long-horizon wrist and hand-pose prediction. Figure 11 shows additional hands motion prediction results.

Table 14: Motion prediction on PalmDex.

	Wrist Trajectory (px)↓				Finger Joints (deg)↓			
Model	Avg H_{30}	Final H_{30}	Avg H_{90}	Final H_{90}	Avg H_{30}	Final H_{30}	Avg H_{90}	Final H_{90}
Linear	17.34	33.95	47.27	92.83	31.51	59.50	93.40	182.47
Kalman	10.61	19.97	26.03	49.69	9.88	17.39	25.59	48.27
RNN	6.31	10.02	9.44	13.66	4.83	6.36	6.45	7.97
LSTM	6.33	10.02	9.49	13.69	4.45	5.78	6.03	7.47
GRU	6.33	10.00	9.35	13.50	4.50	5.85	6.07	7.51
Seq2Seq	6.25	9.91	9.28	13.33	4.43	5.64	6.03	7.40
TCN	6.23	9.93	9.22	13.34	4.63	6.06	6.30	7.80
DLinear	6.24	9.96	9.38	13.52	5.04	6.69	6.98	8.99
iTransformer	6.42	10.09	9.14	13.17	4.55	6.17	6.47	8.32
siMLPe	6.39	10.08	9.58	13.60	4.65	5.95	6.32	7.75
CVAE	**4.82**	**7.70**	**7.44**	**10.80**	**4.15**	**5.44**	**5.74**	**7.07**
DDPM	7.90	11.21	14.21	17.89	10.29	12.56	18.94	20.85



Figure 11: Visualization of hands motion prediction. Top: both wrists' image-plane trajectories, coloured by time from blue to red. Bottom: the corresponding dual-hand finger-skeleton sequence.