HT-Bench: Benchmarking and Learning Dexterous Full-Hand Tactile Representations with Egocentric Vision

Yuzhe Huang[1,2⋆], **Jiaping Wu**[2,3⋆], **Jiaming Jiang**[4], **Hezhe Lin**[2,5], **Aikebaier Aierken**[2,6], **Yunlong Wang**[2], **Kun Cheng**[3], **Wanlin Li**[7], **Chenxi Xiao**[4], **Ziyuan Jiao**[1†], **Yuanxin Zhong**[2†,‡]
[1]Beihang University [2]Rimbot [3]BUPT [4]ShanghaiTech University [5]Tsinghua University [6]CAS [6]Bigai
⋆Equal contributors †Corresponding authors ‡Project Lead

arXiv:2606.19161v2 [cs.RO] 20 Aug 2026

Abstract

Establishing a universal benchmark for tactile representation learning in robotic manipulation remains challenging due to the diversity of tactile sensor designs, data formats, and robot embodiments. Rather than seeking to establish such, we explore a scalable and promising direction for future development: egocentric vision paired with full-hand tactile data. To this end, we introduce **HT-Bench**, a large-scale multi-task benchmark containing 10M RGB frames and 7.8M tactile frames across 226 tasks. HT-Bench evaluates contact-geometry modeling, vision–touch alignment, and generalization to unseen tasks through fine-grained tactile retrieval, masked tactile inpainting, vision-to-tactile synthesis, and multimodal tactile frame prediction. We further propose **HandTouch**, a vector-quantized vision–tactile encoder trained progressively with spatial, cross-modal, and temporal objectives. HandTouch consistently outperforms representative baselines on HT-Bench, improving Recall@5 from 74.65% to 85.23%, reducing masked inpainting RMSE from 0.022 to 0.009, and increasing vision-to-tactile test cIoU from 0.628 to 0.689. Moreover, in real-world downstream manipulation, HandTouch achieves the best average success rate of 68.3% across four contact-rich tasks, outperforming the strongest baseline by 18.3 percentage points. These results show that HandTouch learns transferable tactile representations and that egocentric full-hand tactile data offers a scalable basis for dexterous tactile learning.

1 Introduction

Tactile sensing has attracted increasing attention as a key modality for building robust multimodal foundation models for robotic manipulation, as it captures direct physical interactions—including contact forces, pressure distributions, and slip detection—that vision alone cannot reliably estimate (Yang et al., 2024a; Luo et al., 2026; Huang et al., 2026). With the paradigm shift from vision-centric perception to multimodal robotic policies, tactile sensing has become increasingly important for systems requiring contact-rich reasoning (Chen et al., 2026b; Li et al., 2026) and precise dexterous manipulation (Lin et al., 2025; Pei et al., 2026). Yet, unlike visual perception, which benefits from relatively standardized data formats, scalable architectures, and mature evaluation benchmarks (Dosovitskiy et al., 2021; Radford et al., 2021; Zhai et al., 2023; Russakovsky et al., 2015; Lin et al., 2015), tactile representation learning remains constrained by heterogeneous datasets, sensor-specific processing pipelines, and task-dependent evaluation protocols (Chen et al., 2026a; Cao et al., 2026).

This challenge stems from the intrinsic heterogeneity of tactile information across sensor designs and robotic embodiments. Tactile sensors vary in hardware principles, spatial layouts, signal modalities, and mounting configurations, while robot embodiments introduce different end-effector morphologies and contact patterns (Schneider et al., 2025; Cao et al., 2026; Chen et al., 2026a). Thus, building a universal benchmark that fully accommodates these discrepancies remains impractical. Instead, we focus on a scalable direction: pairing egocentric vision with full-hand tactile data. Egocentric observations provide interaction-centric visual context, while full-hand tactile sensing captures distributed contact during dexterous manipulation. However, existing evaluations still leave open a key question: what encoder can serve as an effective representation backbone for dexterous full-hand tactile perception?

To address this gap, we introduce **HT-Bench** (as shown in fig. 1left), a large-scale multi-task benchmark for evaluating tactile representation learning in dexterous full-hand sensing. HT-Bench aggregates synchronized egocentric visual observations and full-hand tactile sequences, and evaluates encoders under four tasks: fine-grained tactile similarity retrieval, masked tactile inpainting, vision-to-tactile synthesis, and multimodal tactile frame prediction. Instead of relying on a single downstream objective, HT-Bench examines tactile representations from three perspectives: whether they encode meaningful contact geometry, whether they align tactile observations with visual information, and whether they generalize to unseen interaction tasks.

Building on HT-Bench, we propose **HandTouch**, a vector-quantized vision–tactile encoder for general dexterous tactile representation learning. As shown in fig. 1middle, HandTouch follows a progressive training pipeline: it first captures tactile spatial topology via vector-quantized reconstruction, mapping continuous tactile observations into a shared discrete token space; then learns vision–tactile alignment through cross-modal masked tactile inpainting; and finally models temporal contact evolution with multimodal tactile frame prediction.

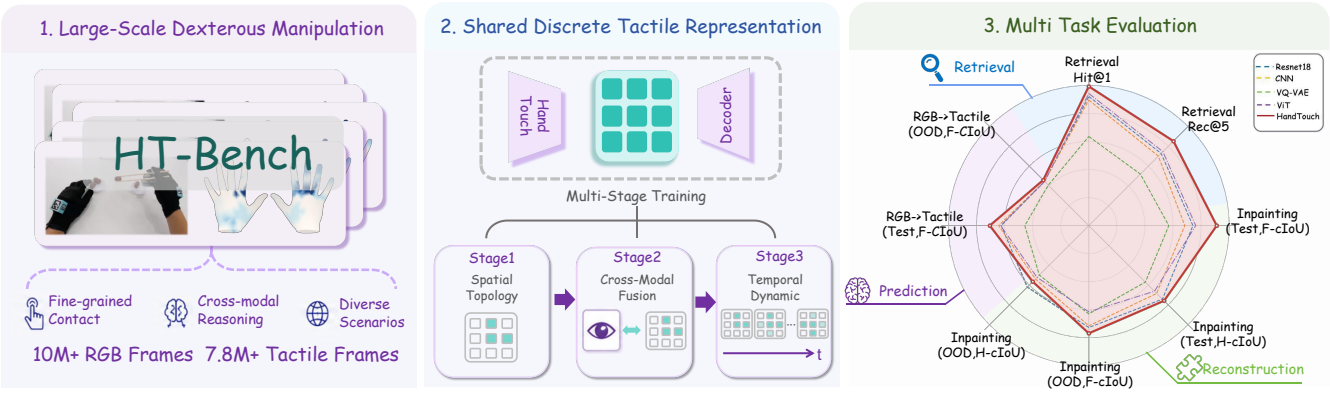

Figure 1: Overview of HT-Bench. 1. HT-Bench pairs egocentric vision with full-hand tactile data to provide a scalable benchmark for dexterous tactile representation learning. It contains 10M RGB frames and 7.8M tactile frames collected from diverse manipulation tasks. 2. HandTouch learns a shared discrete tactile representation through progressive spatial, cross-modal, and temporal training, and 3. is evaluated on fine-grained tactile similarity retrieval, masked tactile inpainting, vision-to-tactile synthesis, and multimodal tactile frame prediction under task-level out-of-distribution splits.

Experiments on HT-Bench show that HandTouch achieves stronger performance than representative tactile encoder baselines. On fine-grained tactile similarity retrieval, HandTouch improves Recall@5 from 74.65% to 85.23% compared with the strongest baseline. On masked tactile inpainting, it reduces full-image RMSE from 0.022 to 0.009 and improves full-image cIoU from 0.762 to 0.912 on the standard test split. For vision-to-tactile synthesis, HandTouch improves OOD cIoU from 0.628 to 0.689, indicating stronger cross-modal tactile generation and better generalization to unseen tasks. Moreover, in real-world downstream manipulation, HandTouch achieves the best average success rate of 68.3% across four contact-rich tasks, outperforming the strongest baseline by 18.3 percentage points.

In summary, our contributions are as follows:

- We introduce **HT-Bench**, a large-scale multi-task benchmark for evaluating dexterous full-hand tactile representation learning based on egocentric vision paired with full-hand tactile data.
- We propose **HandTouch**, a vector-quantized vision–tactile encoder that learns tactile representations through progressive spatial, cross-modal, and temporal training.
- HandTouch achieves consistently stronger performance than representative tactile encoder baselines on HT-Bench and downstream task, demonstrating improved structural retrieval, masked reconstruction, cross-modal synthesis, and out-of-distribution generalization.

We will release all data, evaluation protocols, pretrained weights, and training/testing scripts to facilitate future research in general tactile representation learning.

2 Related Work

Tactile and visuo-tactile representation learning. Tactile representation learning has been studied for object recognition (Yang et al., 2024b; Huang et al., 2026), contact understanding (Xie et al., 2025), manipulation feedback (Li et al., 2025), and visuo-tactile prediction (Chen et al., 2026a). Recent methods show that tactile signals provide complementary physical cues for contact-rich dexterous manipulation. However, most approaches learn representations in task-specific pipelines, where features are tightly coupled with particular sensor configurations, interaction domains, or downstream objectives (Liu et al., 2024; Zorin et al., 2026). Thus, it remains unclear whether these features are transferable tactile representations or task-specific adaptations. This motivates HT-Bench, which evaluates tactile encoders beyond a single objective by assessing contact geometry modeling, visual alignment, temporal contact dynamics, and generalization to unseen tasks. Based on this framework, HandTouch learns general dexterous tactile representations through progressive spatial, cross-modal, and temporal objectives.

Tactile datasets and benchmarks. Recent tactile datasets and benchmarks have advanced tactile learning but also reveal its intrinsic heterogeneity across sensors, embodiments, and tasks. Resources such as Sparsh (Higuera et al., 2024), AnyTouch (Feng et al., 2025, 2026), OpenTouch (Song et al., 2025), and TouchAnything (Zhou et al., 2026) support tactile perception, visuo-tactile learning, or touch-centric manipulation across diverse platforms and scenarios. Other datasets, including VT-DexManip (Liu et al., 2025), multimodal tactile datasets (Chi et al., 2024), and UniVTac (Chen et al., 2026a), further study dexterous manipulation, synchronized sensory streams, or unified visuo-tactile modeling under specific settings. However, variations in sensing principles, spatial layouts, signal formats, and robot embodiments make a universal benchmark difficult to define.

HT-Bench therefore does not aim to resolve all tactile heterogeneity. Instead, we focus on a scalable regime for future embodied AI (Zheng et al., 2026; Wang et al., 2026): egocentric vision paired with full-hand tactile sensing. Egocentric observations provide hand-centric interaction con-

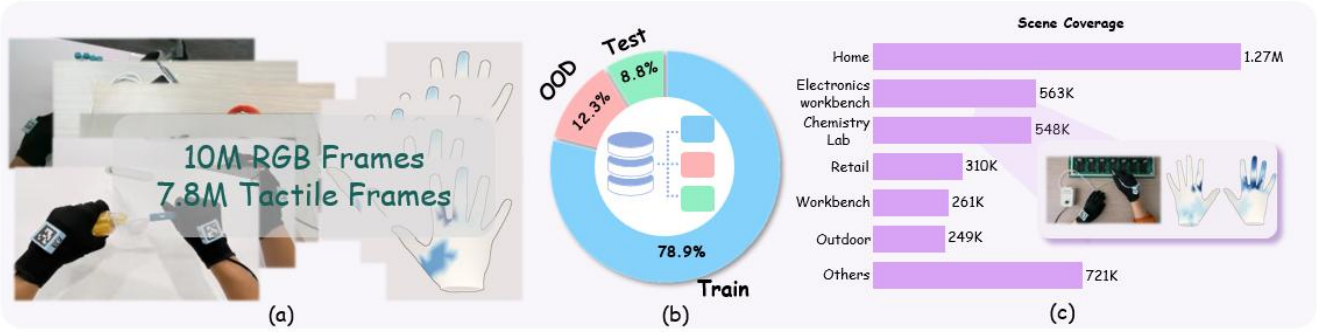

Figure 2: **Statistics and coverage of HT-Bench.** HT-Bench contains 10M RGB frames and 7.8M full-hand tactile frames from dexterous manipulation, with training, test, and task-level OOD splits. It covers diverse scenes such as home, electronics workbench, chemistry lab, retail, workbench, outdoor, and other environments.

Table 1: Comparison with existing tactile benchmarks. HT-Bench jointly supports dexterous full-hand sensing, multi-task evaluation, multi-scene data, and task-level OOD splits.

Benchmark	Full-Hand	Tasks	Scenes	OOD Split
Sparsh (Higuera et al., 2024)	×	Multi-task	Single-scene	×
AnyTouch 2 (Feng et al., 2026)	×	Multi-task	Multi-scene	×
OpenTouch (Song et al., 2025)	✓	Multi-task	Multi-scene	×
TouchAnything (Zhou et al., 2026)	✓	Single-task	Multi-scene	✓
HT-Bench (Ours)	✓	**Multi-task**	**Multi-scene**	✓

text (Fang et al., 2026), while full-hand tactile sensing captures distributed contact patterns in dexterous manipulation. This pairing enables evaluation of contact-geometry modeling, visual alignment, and unseen-task generalization. HT-Bench complements existing datasets with a standardized four-track multi-task protocol in this setting, as compared in table 1.

Masked and discrete representation learning. Masked modeling and vector-quantized representation learning are effective for compact visual and multimodal representations (Razavi, Van den Oord, and Vinyals, 2019; Bao et al., 2021; Van Den Oord, Vinyals, and others, 2017). However, applying them to dexterous hand tactile signals is challenging because tactile observations are sparse, locally structured, and coupled with interaction dynamics (Wu et al., 2026; Xie et al., 2026). HandTouch adapts these ideas to full-hand tactile learning through three progressive stages: vector-quantized reconstruction for spatial topology, cross-modal masked inpainting for vision–touch alignment, and multimodal tactile frame prediction for temporal reasoning. These objectives encourage structurally discriminative and visually grounded tactile representations aligned with the core capabilities evaluated by HT-Bench.

3 HT-Bench: A Multi-Task Tactile Evaluation Benchmark

Extending existing open-source tactile and visuo-tactile datasets (Song et al., 2025; Zhou et al., 2026) with newly collected real-world full-hand tactile sequences, we construct **HT-Bench** for evaluating dexterous full-hand tactile representation learning. HT-Bench contains approximately 10M RGB frames and 7.8M tactile frames, as shown in fig. 2(a). To assess out-of-distribution (OOD) generalization, we adopt a task-level split: one interaction task is held out for OOD evaluation, while the remaining tasks are divided into training and in-distribution test sets with a 9:1 ratio; see fig. 2(b). The data volume and episode distribution across fine-grained scene categories are summarized in fig. 2(c). Based on this dataset, we define four tasks that evaluate complementary capabilities: contact structure understanding, vision–touch alignment, spatial reasoning, and temporal contact modeling.

Fine-Grained Tactile Similarity Retrieval. We evaluate whether tactile embeddings preserve fine-grained contact characteristics with a 1-vs-20 retrieval task. For each query tactile map, candidates are ranked by Structural Similarity Index Measure (SSIM) to obtain tactile-structure reference rankings, which are compared with cosine-similarity rankings induced by learned embeddings.

Masked Tactile Inpainting. We evaluate spatial reasoning under incomplete tactile observations by masking regions of the tactile map and requiring the model to reconstruct missing tactile responses from remaining tactile signals and visual cues. This setting mimics partial tactile information loss from sensor failures or degraded regions, and tests whether representations capture spatially coherent contact patterns.

Vision-to-Tactile Synthesis (RGB-to-Tactile). We evaluate vision–touch alignment by predicting tactile pressure distributions from a single RGB observation. Motivated by humans' ability to form tactile expectations from visual perception, related to cross-modal sensory integration and synesthetic associations (Cytowic, 2002), this task examines

whether learned representations capture the relationship between visual observations and tactile responses induced by object geometry and physical interaction.

Multimodal Tactile Frame Prediction. We evaluate temporal contact modeling by predicting the tactile distribution $\mathbf{t}_T$ from recent visual observations $\mathbf{v}_{T-2}$ and past tactile trajectories $\mathbf{t}_{T-2}$. This task tests whether the representation can integrate visual context and tactile history to anticipate future contact states and capture the temporal evolution of tactile interactions.

4 HandTouch: Vector-Quantized Vision–Tactile Representation Learning

Given the evaluation requirements defined by HT-Bench, we propose **HandTouch**, a vector-quantized vision–tactile encoder for learning transferable full-hand tactile representations. The overall pipeline is illustrated in fig. 3. HandTouch is trained progressively through three stages, each corresponding to a core capability evaluated by HT-Bench. First, it learns unimodal tactile reconstruction to capture spatial contact topology (section 4.1). Second, it incorporates visual cues through cross-modal masked tactile inpainting to learn vision-tactile alignment (section 4.2). Finally, it models contact dynamics through multimodal tactile frame prediction (section 4.3).

4.1 Vector-Quantized Tactile Reconstruction

As illustrated in fig. 3 (upper-left), Stage 1 learns a unimodal tactile reconstruction objective to capture the spatial topology of full-hand tactile maps. Given a normalized tactile map $\mathbf{t} \in [0,1]^{1\times 224\times 224}$, a convolutional projection layer tokenizes it into non-overlapping patches. After adding learnable positional embeddings, the patch tokens are processed by an 8-layer Vision Transformer (ViT) encoder, producing continuous latent features $\mathbf{Z}_e \in \mathbb{R}^{N\times D}$.

To construct a compact and robust discrete feature space, we employ a factorized vector quantizer. Let $\mathcal{C} = \{\mathbf{e}_i\}_{i=1}^{K} \subset \mathbb{R}^d$ denote the shared codebook of size $K = 2048$ with a lower bottleneck dimension $d \ll D$. For the j-th patch embedding $\mathbf{z}_e^{(j)} \in \mathbb{R}^D$, we first project it into the codebook space using an input projection $\mathbf{W}_{\text{in}} \in \mathbb{R}^{d\times D}$:

$$\mathbf{z}_q^{(j)} = \mathbf{e}_k, \quad \text{where } k = \arg\min_i \|\mathbf{W}_{\text{in}}\mathbf{z}_e^{(j)} - \mathbf{e}_i\|_2^2 \tag{1}$$

The quantized token is mapped back to the encoder embedding space by an output projection $\mathbf{W}_{\text{out}} \in \mathbb{R}^{D\times d}$ before being passed to the decoder.

A common issue in vector quantization is codebook collapse, where only a small subset of codebook entries is frequently selected while many entries remain inactive. To mitigate this issue, we track the usage frequency of each codebook entry with an exponential moving average during training. Codebook entries whose cumulative usage falls below a restart threshold τ are reinitialized using randomly sampled active projected features from $\mathbf{W}_{\text{in}}\mathbf{Z}_e$ in the current batch, with small isotropic Gaussian noise added for exploration.

The decoder reconstructs the input tactile map $\hat{\mathbf{t}}$ from the quantized tokens using attention blocks and convolutional upsampling. The Stage 1 objective is defined as:

$$\begin{aligned}\mathcal{L}_{\text{stage1}} = \|\mathbf{t} - \hat{\mathbf{t}}\|_2^2 + \|\mathbf{Z}_q - \text{sg}[\mathbf{W}_{\text{in}}\mathbf{Z}_e]\|_2^2 \\ + \beta\|\,\text{sg}[\mathbf{Z}_q] - \mathbf{W}_{\text{in}}\mathbf{Z}_e\|_2^2\end{aligned} \tag{2}$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operator and β is the commitment loss weight. All modules in this stage are optimized jointly, as indicated by the flame icons in fig. 3.

4.2 Cross-Modal Masked Tactile Inpainting

Stage 2 extends HandTouch from unimodal tactile reconstruction to cross-modal vision–tactile alignment through *Cross-Modal Masked Tactile Inpainting*. As illustrated in fig. 3 (bottom-left), the objective is to reconstruct corrupted tactile maps by using both the remaining tactile observations and synchronized visual cues.

Given a tactile map $\mathbf{t}$, we generate a masked tactile input $\tilde{\mathbf{t}}$ using a curriculum-based dual masking strategy. The first masking mode is *Regional Random Masking*, where tactile regions corresponding to anatomical hand parts, such as the thumb, index finger, middle finger, or palm, are masked to simulate partial tactile information loss. The second mode is *Complete Masking*, where the tactile map is entirely masked, and the model must infer the tactile response primarily from visual context. To gradually increase task difficulty, the probability of applying Complete Masking increases over training. Let $\gamma \in [0,1]$ denote the normalized training progress. We define:

$$P_{\text{full}}(\gamma) = p_{\min} + \frac{(p_{\max} - p_{\min})}{1 + \exp[-12(\gamma - 0.5)]}, \tag{3}$$

where $p_{\min}$ and $p_{\max}$ denote the minimum and maximum probabilities of Complete Masking, respectively. The remaining probability, $1 - P_{\text{full}}(\gamma)$, is assigned to Regional Random Masking. This curriculum gradually shifts the training objective from local tactile inpainting to more challenging vision-conditioned tactile synthesis.

The synchronized visual frame $\mathbf{v}$ is encoded by a frozen pre-trained ViT to extract visual context features $\mathbf{F}_v$. To inject visual information into the tactile reconstruction process, we apply a cross-attention layer followed by a multilayer perceptron. The tactile tokens from the masked tactile input serve as queries, while visual features serve as keys and values. This design allows the tactile decoder to use visual context when reconstructing missing tactile regions. Since Stage 2 processes tactile observations from both left and right hands, we additionally introduce a learnable hand-specific token $\mathbf{t}_{\text{hand}} \in \{\mathbf{t}_{\text{left}}, \mathbf{t}_{\text{right}}\}$ to reduce geometric ambiguity caused by lateral mirroring.

The Stage 2 loss combines reconstruction terms with the vector-quantization losses inherited from Stage 1. Let $\mathbf{M} \in \{0,1\}^{1\times 224\times 224}$ denote a binary occlusion mask, where $\mathbf{M}_{c,h,w} = 1$ indicates a masked pixel.

$$\begin{aligned}\mathcal{L}_{\text{stage2}} = \lambda_{\text{vis}}\|(\mathbf{1} - \mathbf{M}) \odot (\mathbf{t} - \hat{\mathbf{t}}_{\text{cm}})\|_2^2 \\ + \lambda_{\text{mask}}\|\mathbf{M} \odot (\mathbf{t} - \hat{\mathbf{t}}_{\text{cm}})\|_2^2 \\ + \|\mathbf{Z}_q - \text{sg}[\mathbf{W}_{\text{in}}\mathbf{Z}_e]\|_2^2 \\ + \beta\|\,\text{sg}[\mathbf{Z}_q] - \mathbf{W}_{\text{in}}\mathbf{Z}_e\|_2^2\end{aligned} \tag{4}$$

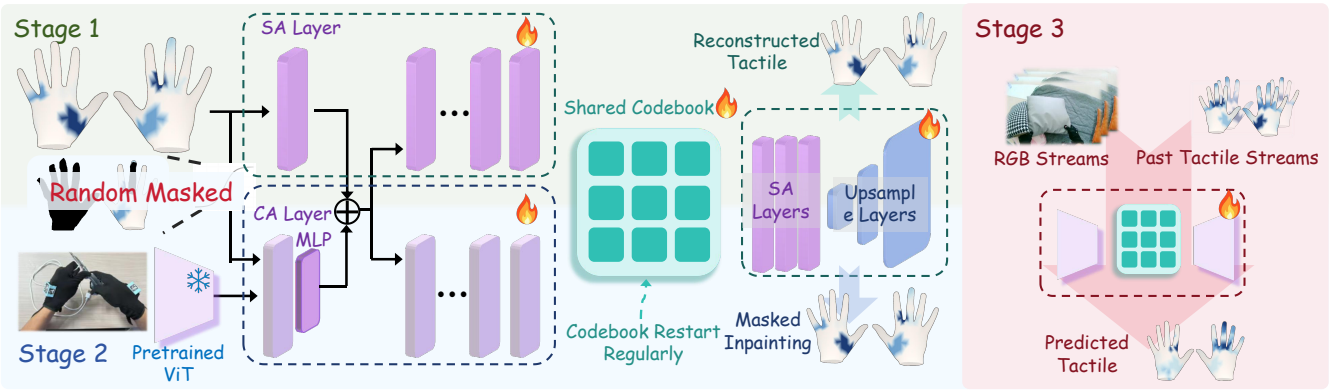

Figure 3: **Training pipeline of HandTouch. Stage 1:** Learning spatial topologies of tactile graphics via unimodal self-attention reconstruction and vector quantization with a shared codebook. **Stage 2:** Reconstructing highly corrupted tactile images under a dynamic regional/complete masking scheme, guided by visual priors injected through cross-attention. **Stage 3:** Forecasting the current tactile distribution $\mathbf{t}_T$ based on sequential visual context $\mathbf{v}_{T-2:T}$ and past tactile histories $\mathbf{t}_{T-2:T-1}$. Modules with flame icons are actively trained in each phase.

where $\odot$ denotes the Hadamard product, $\hat{\mathbf{t}}_{\text{cm}}$ is the cross-modally reconstructed tactile output, and λ_{vis} and λ_{mask} balance the visible-region and masked-region reconstruction losses. We set $\lambda_{\text{mask}} > \lambda_{\text{vis}}$ to emphasize reconstruction quality in the missing tactile regions. The codebook restart mechanism from section 4.1 remains active in this stage to prevent codebook underutilization. All non-frozen modules are optimized jointly.

4.3 Multimodal Tactile Frame Prediction

In the final stage, HandTouch learns temporal contact modeling through *Multimodal Tactile Frame Prediction*. As illustrated in fig. 3 (right), this stage predicts the tactile frame at time step T by integrating recent visual observations and past tactile history.

Formally, the model takes recent visual observations $\mathbf{v}_{T-2}$ and past tactile trajectories $\mathbf{t}_{T-2}$ as input. These multimodal sequences are temporally aggregated and projected into the shared discrete latent space learned in the previous stages. The decoder then predicts the tactile distribution $\hat{\mathbf{t}}_T$ at time step T. This objective requires the model to combine visual context with tactile history, thereby encouraging temporally aware tactile representations.

The Stage 3 objective is defined as:

$$\mathcal{L}_{\text{stage3}} = |\mathbf{t}_T - \hat{\mathbf{t}}_T|_2^2. \tag{5}$$

During this stage, the prediction modules, shared codebook, and decoder are fine-tuned to improve multimodal tactile prediction, as indicated by the flame icons in fig. 3.

5 Experiment

We evaluate the proposed **HandTouch** framework on HT-Bench. To assess its representation capability, we compare it with representative tactile encoder baselines that are commonly used in tactile perception and manipulation tasks, including a CNN-based encoder (Lee, Grimaldi, and Yu, 2026), ResNet-18 (Calandra et al., 2018), a VQ-VAE-based encoder (Xu et al., 2025), and a ViT-based encoder (Zhao et al., 2024). For a fair comparison, all baselines are pretrained on the same training split and evaluated under the same HT-Bench protocols.

5.1 Comparison with Baseline Encoders

We first compare HandTouch with representative tactile representation baselines on fine-grained tactile similarity retrieval, masked tactile inpainting, and vision-to-tactile synthesis. The results are summarized in table 2. Overall, HandTouch achieves the best performance on most evaluation metrics, demonstrating stronger tactile structural modeling, cross-modal alignment, and generalization under both in-distribution and OOD settings.

Fine-Grained Tactile Similarity Retrieval. For fine-grained tactile similarity retrieval, HandTouch outperforms all baseline encoders. Hit@1 measures whether the SSIM-nearest tactile candidate is ranked first by the learned embedding similarity, while Recall@5 measures the fraction of SSIM top-5 candidates recovered in the top-5 retrieved candidates. Compared with the strongest baseline, the ViT-based encoder, HandTouch improves Hit@1 from 94.27% to 99.27% and Recall@5 from 74.65% to 85.23%. This indicates that the embedding space learned by HandTouch better preserves fine-grained structural similarity among tactile observations. In contrast, CNN- and ResNet-based models can capture local spatial patterns but are less effective in organizing tactile samples according to global structural correspondence. VQ-VAE performs notably worse on retrieval, suggesting that reconstruction-oriented discrete latent representations alone may lose subtle structural cues that are important for fine-grained tactile matching.

Spatially Masked Tactile Inpainting. For masked tactile inpainting, HandTouch achieves the lowest reconstruction error and the highest contact overlap on the standard test split. Specifically, it obtains a full-map RMSE of 0.010 and a full-map cIoU of 0.911, substantially outperforming

Table 2: **Quantitative comparison on HT-Bench.** We compare HandTouch with representative tactile encoder baselines on fine-grained tactile similarity retrieval, masked tactile inpainting, and vision-to-tactile synthesis (RGB→Tac). Retrieval performance is evaluated using Hit@1 and Recall@5, while inpainting and synthesis are evaluated on the standard test and task-level OOD splits using RMSE and contact IoU (cIoU). 'F-' and 'H-' denote metrics computed over the full tactile map and masked hole regions, respectively. ↑ and ↓ indicate that higher and lower is better, and the best results are highlighted in **bold**.

Model	Retrieval		Inpainting (Test)				Inpainting (OOD)				RGB→Tac (Test)		RGB→Tac (OOD)	
	Hit@1↑	Rec@5↑	F-RMSE↓	F-cIoU↑	H-RMSE↓	H-cIoU↑	F-RMSE↓	F-cIoU↑	H-RMSE↓	H-cIoU↑	F-RMSE↓	F-cIoU↑	F-RMSE↓	F-cIoU↑
ResNet-18	92.13	72.85	0.025	0.742	0.025	0.742	0.041	0.727	**0.056**	**0.620**	0.038	0.624	0.084	0.445
CNN-based	89.61	70.09	0.030	0.684	0.030	0.684	0.047	0.715	0.068	0.538	0.036	0.642	0.083	0.447
VQ-VAE	63.60	51.98	0.042	0.570	0.042	0.570	0.053	0.630	0.068	0.499	0.060	0.456	0.081	0.408
ViT-based	94.27	74.65	0.022	0.762	0.033	0.662	0.056	0.615	0.065	0.522	0.038	0.628	0.083	0.446
Ours	**99.27**	**85.23**	**0.009**	**0.912**	**0.024**	**0.761**	**0.041**	**0.800**	0.063	0.581	**0.032**	**0.689**	**0.080**	**0.457**

all baselines. We define contact IoU (cIoU) as:

$$\mathrm{cIoU} = \frac{\sum_{i,j} \min(P_{i,j}, \hat{P}_{i,j})}{\sum_{i,j} \max(P_{i,j}, \hat{P}_{i,j})}$$

where $P_{i,j}$ and $\hat{P}_{i,j}$ denote the ground-truth and predicted pressure values at pixel (i, j), respectively. HandTouch also performs well in masked hole regions, achieving 0.024 RMSE and 0.761 cIoU, which demonstrates its ability to infer missing tactile responses from visible tactile context and learned spatial priors. Under the OOD setting, HandTouch maintains the best full-map reconstruction performance, with 0.041 RMSE and 0.800 cIoU. However, ResNet-18 performs better on OOD hole-region metrics. This suggests that although HandTouch generalizes well at the global tactile-map level, precise reconstruction of severely corrupted local regions in unseen tasks remains challenging.

Vision-to-Tactile Synthesis. For vision-to-tactile synthesis, HandTouch demonstrates strong cross-modal predictive capability. On the standard test split, HandTouch reduces the full-map RMSE to 0.032 and improves cIoU to 0.689, outperforming CNN-, ResNet-, VQ-VAE-, and ViT-based baselines. These results indicate that HandTouch learns effective vision–tactile correspondence, allowing it to infer tactile pressure distributions from visual observations. Under the OOD setting, HandTouch achieves the highest cIoU of 0.457 and 0.080 RMSE, indicating better cross-modal generalization to unseen tasks.

Multimodal Tactile Frame Prediction. Unlike the static evaluation tracks, multimodal tactile frame prediction requires temporal modeling from recent visual observations $\mathbf{v}T-2$ and past tactile trajectories $\mathbf{t}T-2$. Since the baselines are designed as single-frame tactile encoders, we report this temporal prediction track separately to evaluate the predictive capability of HandTouch. As shown in fig. 4, HandTouch predicts tactile distributions $\mathbf{t}_T$ that closely match the ground truth. It achieves accurate prediction on the test split (RMSE: 0.031, cIoU: 0.677) and maintains reasonable performance under OOD tasks, demonstrating its ability to model temporal contact dynamics during continuous interaction.

5.2 Ablation Study

We conduct ablation studies to evaluate the contribution of key components in HandTouch, including the drop-based training strategy, the vector-quantized codebook, and visual conditioning. As shown in Table 3, the full model achieves the best OOD performance on both masked tactile inpainting and RGB→Tac, demonstrating stronger generalization to unseen tasks.

Table 3: Ablation study of HandTouch on masked tactile inpainting and vision-to-tactile synthesis. We report cIoU on the standard test and task-level OOD splits. H-cIoU is computed over masked hole regions for inpainting, while F-cIoU is computed over the full tactile map for RGB→Tac. Higher is better, and the best result in each column is highlighted in bold.

Variant	Inpainting		RGB→Tac	
	H-cIoU Test↑	H-cIoU OOD↑	F-cIoU Test↑	F-cIoU OOD↑
w/o drop	0.712	0.567	0.678	0.453
w/o codebook	**0.765**	0.529	**0.708**	0.454
w/o vision	0.698	0.556	0.467	0.415
Ours	0.761	**0.581**	0.689	**0.457**

Removing the drop-based strategy leads to consistent performance degradation across all metrics, indicating that tactile corruption regularization helps the model better infer missing contact patterns from incomplete observations. Removing the codebook slightly improves in-distribution performance, but significantly reduces OOD inpainting H-cIoU from 0.581 to 0.529. This suggests that the discrete codebook serves as an effective representation bottleneck, improving the transferability of tactile features rather than merely fitting the training distribution.

Visual conditioning is also essential. Without visual input, performance drops substantially, especially for RGB→Tac, where test F-cIoU decreases from 0.689 to 0.467. This confirms that egocentric visual cues provide critical cross-modal information for tactile synthesis and also benefit masked tactile reconstruction. Overall, these results validate the proposed design: drop-based training improves robustness, the codebook enhances OOD generalization, and visual conditioning strengthens vision–touch alignment.

5.3 Real-World Downstream Evaluation

To further evaluate whether the learned tactile representations can transfer to real robotic manipulation(seen in fig.

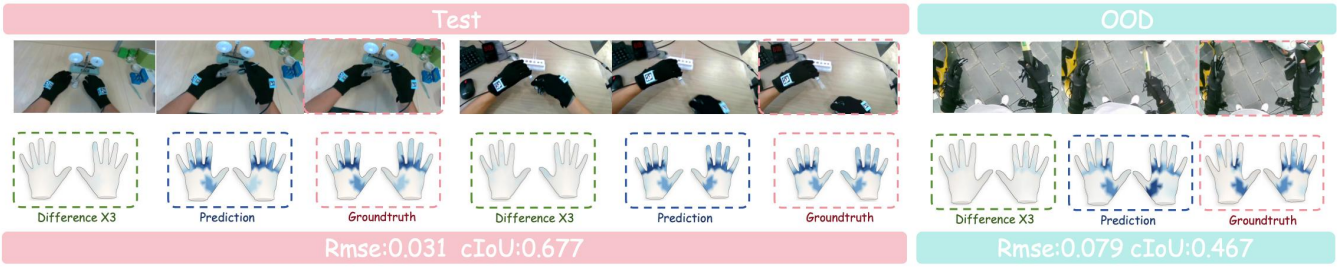

Figure 4: **Qualitative results of multimodal tactile frame prediction.** HandTouch predicts tactile distributions that closely match the ground truth under both in-distribution and OOD settings. Prediction errors are magnified by a factor of 3 for better visualization. These results illustrate the temporal contact modeling capability of HandTouch.

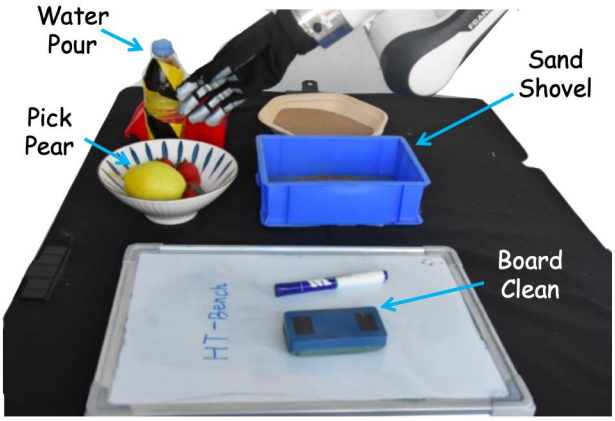

Figure 5: Real-world downstream evaluation setup. We evaluate the learned tactile representation on four contact-rich manipulation tasks, including water pouring, pear picking, sand shoveling, and board cleaning.

Table 4: Success rates (%) on real-world downstream manipulation tasks. Each task is evaluated over 15 trials.

Method	Board Clean	Pear Pick	Water Pour	Sand Shovel	Avg.
ResNet-Scratch	20.0	53.3	13.3	20.0	26.7
ResNet-Trained	46.7	73.3	33.3	46.7	50.0
ViT	40.0	73.3	33.3	33.3	45.0
HandTouch	**66.7**	**86.7**	**53.3**	**66.7**	**68.3**

5), we conduct downstream experiments on four contact-rich tasks: board cleaning, pear picking, water pouring, and sand shoveling. Each task is evaluated over 15 trials, and we report the success rate of each method in Table 4.

Overall, HandTouch achieves the best performance across all tasks, obtaining an average success rate of 68.3%, compared with 50.0% for the trained ResNet encoder and 45.0% for the ViT encoder. The improvement is consistent across different interaction types. On board cleaning and sand shoveling, where maintaining stable and distributed contact is critical, HandTouch improves the success rate to 66.7%, outperforming the strongest baseline by 20.0 percentage points. On pear picking, HandTouch reaches 86.7%, suggesting that the learned representation better captures fine-grained contact cues for stable grasping. Water pouring remains the most challenging task for all methods, as it requires precise contact and motion coordination; nevertheless, HandTouch still achieves the highest success rate of 53.3%.

These results indicate that the proposed representation is not only effective on offline HT-Bench evaluations, but also provides practical benefits for real-world manipulation. In particular, the consistent gains over both convolutional and transformer-based baselines suggest that the progressive spatial, cross-modal, and temporal training design improves the transferability of tactile features to downstream contact-rich robotic tasks.

6 Conclusion and Limitation

In this paper, we introduce **HT-Bench**, a large-scale multi-task benchmark for dexterous tactile representation learning with egocentric vision and full-hand tactile data. HT-Bench evaluates tactile encoders on fine-grained tactile retrieval, masked tactile inpainting, vision-to-tactile synthesis, and multimodal tactile frame prediction, with task-level OOD splits for generalization assessment. We further propose **HandTouch**, a progressive vector-quantized vision–tactile encoder trained with spatial, cross-modal, and temporal objectives.

HandTouch consistently outperforms representative baselines on HT-Bench, demonstrating stronger contact-structure modeling, vision–touch alignment, temporal reasoning, and OOD generalization. Real-world downstream experiments on four contact-rich tasks further validate its transferability, where HandTouch achieves the best average success rate of 68.3%, surpassing the strongest baseline by 18.3 percentage points. These results show that egocentric full-hand tactile data offers a scalable foundation for general tactile representation learning in dexterous manipulation.

A limitation of HT-Bench is that it focuses on egocentric vision paired with full-hand tactile sensing, without yet covering broader sensing paradigms such as fingertip optical tactile sensors, force/torque sensors, skin-like taxel arrays, or non-hand embodiments. Extending the benchmark to more diverse hardware platforms, sensing modalities, and robotic embodiments is an important future direction.

References

Bao, H.; Dong, L.; Piao, S.; and Wei, F. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Calandra, R.; Owens, A.; Jayaraman, D.; Lin, J.; Yuan, W.; Malik, J.; Adelson, E. H.; and Levine, S. 2018. More than a feeling: Learning to grasp and regrasp using vision and touch. *IEEE Robotics and Automation Letters* 3(4):3300–3307.

Cao, Z.; Tian, D.; Guan, R.; Mu, Y.; Sun, X.; Liang, S.; Liu, D.; Huang, T.; Yue, Y.; Ding, H.; Fang, B.; Zhou, A.; Han, Q.-L.; and Xiong, H. 2026. Tactile-based multimodal fusion in embodied intelligence: A survey of vision, language, and contact-driven paradigms.

Chen, B.; Wan, W.; Chen, T.; Guo, X.; Xu, C.; Qi, Y.; Zhang, H.; Wu, L.; Xu, T.; Li, Z.; Wu, Y.; Li, R.; Yang, X.; Luo, P.; Sui, W.; and Mu, Y. 2026a. Univtac: A unified simulation platform for visuo-tactile manipulation data generation, learning, and benchmarking.

Chen, H.; Xu, J.; Chen, H.; Hong, K.; Huang, B.; Liu, C.; Mao, J.; Li, Y.; Du, Y.; and Driggs-Campbell, K. 2026b. Multi-modal manipulation via multi-modal policy consensus.

Chi, H.-G.; Barreiros, J.; Mercat, J.; Ramani, K.; and Kollar, T. 2024. Multi-modal representation learning with tactile data. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 9660–9667. IEEE.

Cytowic, R. E. 2002. *Synesthesia: A union of the senses*. MIT press.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Fang, D.; Wu, Y.; Zhong, Y.; Zhang, R.; Wang, Y.; Jia, X.; and Jiang, Y.-G. 2026. Unidextok: A unified dexterous hand tokenizer from real data. *arXiv preprint arXiv:2606.10683*.

Feng, R.; Hu, J.; Xia, W.; Gao, T.; Shen, A.; Sun, Y.; Fang, B.; and Hu, D. 2025. Anytouch: Learning unified static-dynamic representation across multiple visuo-tactile sensors. *arXiv preprint arXiv:2502.12191*.

Feng, R.; Zhou, Y.; Mei, S.; Zhou, D.; Wang, P.; Cui, S.; Fang, B.; Yao, G.; and Hu, D. 2026. Anytouch 2: General optical tactile representation learning for dynamic tactile perception. *arXiv preprint arXiv:2602.09617*.

Higuera, C.; Sharma, A.; Bodduluri, C. K.; Fan, T.; Lancaster, P.; Kalakrishnan, M.; Kaess, M.; Boots, B.; Lambeta, M.; Wu, T.; and Mukadam, M. 2024. Sparsh: Self-supervised touch representations for vision-based tactile sensing.

Huang, Y.; Lin, P.; Li, W.; Li, D.; Li, J.; Jiang, J.; Xiao, C.; and Jiao, Z. 2026. Tactile-force alignment in vision-language-action models for force-aware manipulation. *arXiv preprint arXiv:2601.20321*.

Lee, W.; Grimaldi, M.; and Yu, T. 2026. Symmetry-aware fusion of vision and tactile sensing via bilateral force priors for robotic manipulation. *arXiv preprint arXiv:2602.13689*.

Li, Y.; Jin, Z.; Liu, J.; and Ma, D. 2025. Visuo-tactile feedback policies for terminal assembly facilitated by reinforcement learning. *Frontiers in Robotics and AI* 12:1660244.

Li, Y.; Chen, Y.; Zhao, Z.; Li, P.; Liu, T.; Huang, S.; and Zhu, Y. 2026. Simultaneous tactile-visual perception for learning multimodal robot manipulation.

Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C. L.; and Dollár, P. 2015. Microsoft coco: Common objects in context.

Lin, P.; Huang, Y.; Li, W.; Ma, J.; Xiao, C.; and Jiao, Z. 2025. Pp-tac: Paper picking using tactile feedback in dexterous robotic hands. *arXiv preprint arXiv:2504.16649*.

Liu, Q.; Ye, Q.; Sun, Z.; Cui, Y.; Li, G.; and Chen, J. 2024. Masked visual-tactile pre-training for robot manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 13859–13875. IEEE.

Liu, Q.; Cui, Y.; Sun, Z.; Li, G.; Chen, J.; and Ye, Q. 2025. Vtdexmanip: A dataset and benchmark for visual-tactile pretraining and dexterous manipulation with reinforcement learning. In *The Thirteenth International Conference on Learning Representations*.

Luo, S.; Li, Y.; Hu, Y.; Yu, C.; Xu, C.; Zhang, J.; Yao, G.; Huang, T.; He, R.; and Wang, Z. 2026. Omniumi: Towards physically grounded robot learning via human-aligned multimodal interaction.

Pei, L.; Yuzhe, H.; Wanlin, L.; Chenxi, X.; and Ziyuan, J. 2026. Dexmove: Learning tactile-guided non-prehensile manipulation with dexterous hands. In *The Fourteenth International Conference on Learning Representations*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning transferable visual models from natural language supervision.

Razavi, A.; Van den Oord, A.; and Vinyals, O. 2019. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems* 32.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. Imagenet large scale visual recognition challenge.

Schneider, T.; Duret, G.; de Farias, C.; Calandra, R.; Chen, L.; and Peters, J. 2025. Tactile mnist: Benchmarking active tactile perception.

Song, Y. R.; Li, J.; Fu, R.; Murphy, D.; Zhou, K.; Shiv, R.; Li, Y.; Xiong, H.; Owens, C. E.; Du, Y.; Luo, Y.; Cheng, X.; Torralba, A.; Matusik, W.; and Liang, P. P. 2025. Opentouch: Bringing full-hand touch to real-world interaction.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems* 30.

Wang, Z.; He, B.; Yu, K.; Lee, S.; Gao, R.; Huang, F.; and Aloimonos, Y. 2026. Humanego: Zero-shot robot learning from minutes of human egocentric videos.

Wu, Y.; Lin, Y.; Lao, W.; Lin, Y.; Wei, Y.-L.; Zheng, W.-S.; and Wu, A. 2026. Dexgrasp-zero: A morphology-aligned policy for zero-shot cross-embodiment dexterous grasping. *arXiv preprint arXiv:2603.16806*.

Xie, Y.; Li, M.; Li, S.; Li, X.; Chen, G.; Ma, F.; Yu, F. R.; and Ding, W. 2025. Universal visuo-tactile video understanding for embodied interaction.

Xie, Y.; Li, M.; Li, S.; Li, X.; Chen, G.; Ma, F.; Yu, F.; and Ding, W. 2026. Universal visuo-tactile video understanding for embodied interaction. *Advances in Neural Information Processing Systems* 38:127864–127883.

Xu, Z.; Uppuluri, R.; Zhang, X.; Fitch, C.; Crandall, P. G.; Shou, W.; Wang, D.; and She, Y. 2025. Unit: Data efficient tactile representation with generalization to unseen objects.

Yang, F.; Feng, C.; Chen, Z.; Park, H.; Wang, D.; Dou, Y.; Zeng, Z.; Chen, X.; Gangopadhyay, R.; Owens, A.; and Wong, A. 2024a. Binding touch to everything: Learning unified multimodal tactile representations.

Yang, F.; Feng, C.; Chen, Z.; Park, H.; Wang, D.; Dou, Y.; Zeng, Z.; Chen, X.; Gangopadhyay, R.; Owens, A.; and Wong, A. 2024b. Binding touch to everything: Learning unified multimodal tactile representations. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 26330–26343.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training.

Zhao, J.; Ma, Y.; Wang, L.; and Adelson, E. H. 2024. Transferable tactile transformers for representation learning across diverse sensors and tasks. *arXiv preprint arXiv:2406.13640*.

Zheng, R.; Niu, D.; Xie, Y.; Wang, J.; Xu, M.; Jiang, Y.; Castañeda, F.; Hu, F.; Tan, Y. L.; Fu, L.; Darrell, T.; Huang, F.; Zhu, Y.; Xu, D.; and Fan, L. 2026. Egoscale: Scaling dexterous manipulation with diverse egocentric human data.

Zhou, J.; Gao, Z.; Hong, F.; Liu, Z.; Zhang, G.; Dai, W.; Zhen, R.; Lyu, C.; Wu, H.; Mao, Y.; Wang, X.; Jiang, Y.; Ding, W.; and Yang, S. 2026. Touchanything: A dataset and framework for bimanual tactile estimation from egocentric video.

Zorin, A.; Si, Z.; Park, M.; Park, J.; Buynitsky, A.; Bhadang, S.; Park, T.; Yoon, S. J.; Park, Y.-L.; Kroemer, O.; et al. 2026. Taco: Benchmarking tactile sensors for object manipulation. *arXiv preprint arXiv:2605.21976*.