2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Tactile Hide and Seek: Bimanual Object Blind Search and Retrieval via Tactile-Only Feedback

Xiangyu Fu[1], Hao Xing[1], Simon Armleder[1], Wenlan Shen[1], Fengyi Wang[1], Julio Rogelio Guadarrama-Olvera[1] and Gordon Cheng[1]

***Abstract*— Locating and identifying objects in vision-denied environments is a critical challenge for intelligent robot systems. To address the limitation of vision, we present a tactile-only method for object search and recognition using custom tactile skin sensors on robot hands. The method involves searching an object in a vision-denied environment with a tactile "hide and seek" strategy. Upon contact, the system employs a novel two-phase classification process: an initial single-handed classification by pushing the object, followed by a two-handed verification stage that incorporates size measurement to confirm the object's identity and reduce critical errors. To support this approach, we introduce the HAS (Hide-and-Seek) dataset, a large-scale, multimodal tactile dataset of 1.1 million samples collected on a custom sensor hardware. Our system achieves an object classification accuracy of 91.1% and a weight classification accuracy of 83.1% on the HAS dataset, with a strict joint accuracy of 79.6%. The full online pipeline attains a 61.4% success rate in real-world identification, with the bimanual verification stage further correcting up to 17.6% of single-hand errors. Comprehensive ablation studies validate the contribution of individual sensor modalities and demonstrate the effectiveness of our tactile-only method for autonomous operation in a non-vision environment. Our project page is available at https://tactile-hide-and-seek.github.io/.**

I. Introduction

The ability to autonomously locate and identify objects is fundamental for intelligent robots operating in real-world environments, from search missions to domestic assistance. However, vision-based systems fail in non-visual settings such as darkness, smoke, or occluded spaces. Such conditions frequently arise in disaster response (e.g., searching through debris or smoke-filled environments), industrial inspection (e.g., inside pipelines or confined machinery), and domestic assistance (e.g., retrieving items from drawers or cabinets). These scenarios highlight the need for tactile-only capabilities, where robots must rely solely on touch to perceive and act.

Humans excel at using touch to search and identify objects when vision is unavailable, as in the blindfolded example of Fig. 1, where tactile feedback alone guides recognition in a cluttered tabletop scene. This human capability directly inspires our *Tactile Hide-and-Seek* (THAS) strategy, which mirrors the same exploration–verification pattern in a robotic system. Our approach begins with exploratory sweeping motions to locate objects. Upon contact, the system executes a novel two-phase classification process: initial single-handed recognition through pushing, followed by bimanual verification incorporating size estimation to confirm object identity and minimize errors.

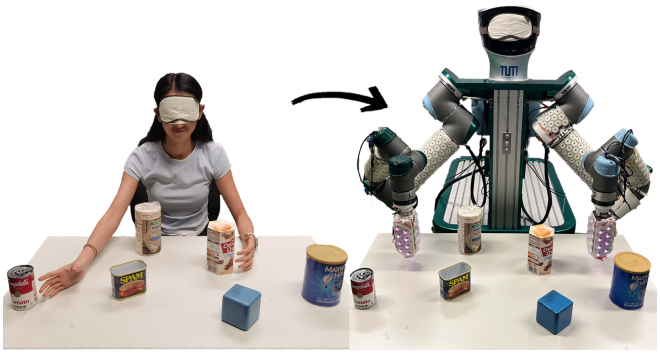

Fig. 1: Human-Inspired Tactile Hide-and-Seek

Despite this inspiration, tactile-only recognition remains highly challenging: feedback is inherently *local and partial*, sensors suffer from *noise and drift*, and single-hand interactions often yield *ambiguous cues*. These factors highlight the need for large-scale data, multimodal fusion, and bimanual exploration to enable robust tactile perception.

Prior research in tactile hardware and control has largely centered on fingertip sensing, single-hand grasp stability, or small-scale recognition datasets [1], [2]. Only a few works have explored the broader problem of *tactile-only object search and retrieval*, and even fewer have addressed *bimanual tactile grasping* [3], [4]. A key barrier has been the lack of large-scale, multimodal datasets to support modern learning-based approaches. Motivated by this gap, we introduce *Hide-and-Seek* (HAS) Dataset: a large-scale multimodal tactile dataset designed as a foundation for advancing tactile-only perception and manipulation.

Overall, our contributions are threefold: **(i)** a complete robotic pipeline that integrates tactile exploration, push-based classification, and compliant bimanual grasping in cluttered, vision-denied environments. **(ii)** a dual-head multimodal classifier that jointly predicts object identity and weight; and **(iii)** a large-scale multimodal tactile dataset covering 61 categories with weight annotations.

The remainder of this paper is organized as follows: Section II reviews related work, Section III details Tactile Hide-and-Seek strategy, Section IV introduces the collected tactile dataset, Section V presents experimental results, and Section VI concludes with future directions.

[1]Authors are with the Institute for Cognitive Systems, Technical University of Munich, Arcisstrae 21, 80333 Munich, Germany { xiangyu.fu, hao.xing, simon.armleder, wenlan.shen, fengyi.wang, gordon } @tum.de

II. Related Work

We review related work in two main areas: (i) tactile perception and multimodal classification, and (ii) bimanual manipulation with compliance control.

A. Tactile Perception and Multimodal Classification

Early tactile research centered on fingertip sensors such as GelSight and BioTac for high-resolution local feedback on texture and slip [5], [6]. More recently, large-area skins and multimodal arrays have enabled whole-hand perception [7], capturing distributed force and proximity. A key bottleneck remains dataset scale: existing corpora are typically small and limited in diversity [8], [9]. To address recognition, recent works explore RNNs [10], multimodal fusion [11], and tactile transformers [12], [13]. Active strategies such as contour following [14], sliding or tapping [15], and probabilistic mapping [16] further highlight the role of exploration. More advanced frameworks, e.g., DexTouch [17], MimicTouch [18], and TactoFind [3], extend these ideas, but remain vision-assisted or single-hand. In contrast, our HAS dataset provides over one million multimodal frames across five sensing streams, enabling large-scale tactile-only classification and setting a foundation for reliable bimanual verification.

B. Bimanual Manipulation and Compliance Control

Dual-arm systems have been studied for coordinated manipulation and assembly [19]. Recent platforms such as ALOHA [20] and UMI [21] demonstrate large-scale imitation learning but rely primarily on vision. More recent work advances bimanual foundations: RDT-1B [22] introduces a diffusion-based policy model, DA-VIL [23] combines RL with variable impedance control, and VTAO-BiManip [24] leverages masked pretraining across visual, tactile, and action modalities. Tactile-enabled systems such as Bi-Touch [4] and dual-arm HRI frameworks [25] incorporate touch, but focus on policy learning rather than recognition. Our framework builds on this line of work by exploiting tactile sensing as the *sole* feedback channel, combining bimanual contact cues with admittance control to achieve stable and adaptive grasping in vision-denied environments.

III. Method

A. Task Formulation

We study tactile-based object recognition and weight estimation in a vision-denied workspace. The task is to predict both the object class $o \in \mathcal{O}$ and the weight label $w \in \mathcal{W}$, where $\mathcal{O}$ contains 33 objects plus a negative class, and $\mathcal{W} = \{\text{none}, \text{light}, \text{medium}, \text{heavy}\}$.

At each time step t, the robot collects multimodal observations:

$$\mathbf{x}_t = \{\mathbf{P}_t,\ \mathbf{s}_t,\ \mathbf{W}_t^{\text{skin}},\ \mathbf{q}_t,\ \mathbf{W}_t^{\text{ft}}\}, \tag{1}$$

where $\mathbf{P}_t \in \mathbb{R}^{N\times 6}$ is the tactile point cloud from $N{=}88$ skin cells $[x, y, z, r, g, b]$; $\mathbf{s}_t \in \mathbb{R}^{3\times 88}$ are raw per-cell prox/force/dist signals; $\mathbf{W}_t^{\text{skin}} \in \mathbb{R}^{2\times 6}$ are virtual wrenches (force- and proximity-based); $\mathbf{q}_t \in \mathbb{R}^7$ is the end-effector pose; and $\mathbf{W}_t^{\text{ft}} \in \mathbb{R}^6$ is the wrist force/torque wrench.

A trajectory $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ is formed by recording T steps of tactile exploration. The recognition network maps a fixed-length time window (e.g., 100 frames with a stride of 40) to the object and weight predictions,

$$f_\theta : \mathbf{X} \mapsto (\hat{o}, \hat{w}), \tag{2}$$

where $\hat{o}$ and $\hat{w}$ denote the predicted object and weight labels, and o, w are the corresponding ground-truth labels. If the predicted class matches the specified search target with confidence $\geq \tau_{\text{cls}}$, the system proceeds to bimanual grasping. The grasp pose $\hat{\mathbf{g}} \in SE(3)^2$ is then computed *deterministically* from the tactile point cloud geometry.

B. Tactile Hide-and-Seek Framework

The overall Tactile Hide-and-Seek (THAS) framework, illustrated in Figure 2, follows a hide and seek strategy composed of four main phases: bimanual tactile exploration, single-arm classification, bimanual verification, and grasping. The system is given a predefined *target object specification* (object category and weight attribute) and aims to autonomously locate and retrieve this object in a vision-denied environment using only tactile sensing.

a) Bimanual tactile exploration: The robot explores the workspace purely through touch while incrementally constructing a 3D tactile map. The space is discretized into a voxel grid of 2 cm resolution, where each voxel is labeled as unknown, free, or contact. A predefined zig-zag trajectory ensures full coverage, with each arm assigned to a distinct, non-overlapping region. Tactile point clouds from the skin sensors are accumulated into this map to provide a consistent representation of explored areas.

b) Push and classify: When the number of contact points $\mathbf{p}_{\text{contact}}$ exceed a threshold, the system considers an object encountered and initiates classification. Upon contact, the robot executes a short push motion (≈ 5 cm) along the estimated object surface normal. The contact normal $\mathbf{n}$ is computed from the relative vector between the hand position $\mathbf{p}_{\text{hand}}$ and the mean position of all contact points $\bar{\mathbf{p}}_{\text{contact}}$:

$$\mathbf{n} = \frac{\mathbf{p}_{\text{hand}} - \bar{\mathbf{p}}_{\text{contact}}}{\|\mathbf{p}_{\text{hand}} - \bar{\mathbf{p}}_{\text{contact}}\|}, \quad \bar{\mathbf{p}}_{\text{contact}} = \frac{1}{N_c}\sum_{i=1}^{N_c} \mathbf{p}_i, \tag{3}$$

where N_c denotes the number of points in contact and $\mathbf{p}_i$ their positions. Multimodal signals from each trajectory are processed by the dual-head classifier to predict both object category $\hat{o}$ and weight label $\hat{w}$. Since a trajectory lasts about two seconds, we classify multiple overlapping windows and select the final output via majority voting.

If confidence is low (defined as the maximum softmax probability, with threshold $\tau_{\text{cls}} = 0.9$ in our experiments), the robot performs up to three additional pushes by first translating the hand tangentially around the estimated object center and then pushing along the estimated surface normal to collect additional evidence; the prediction with the highest confidence is retained. Once the target object is identified with high confidence, exploration with the other hand is terminated; otherwise, the system continues collision-aware searching until all voxels are explored. To avoid redundant

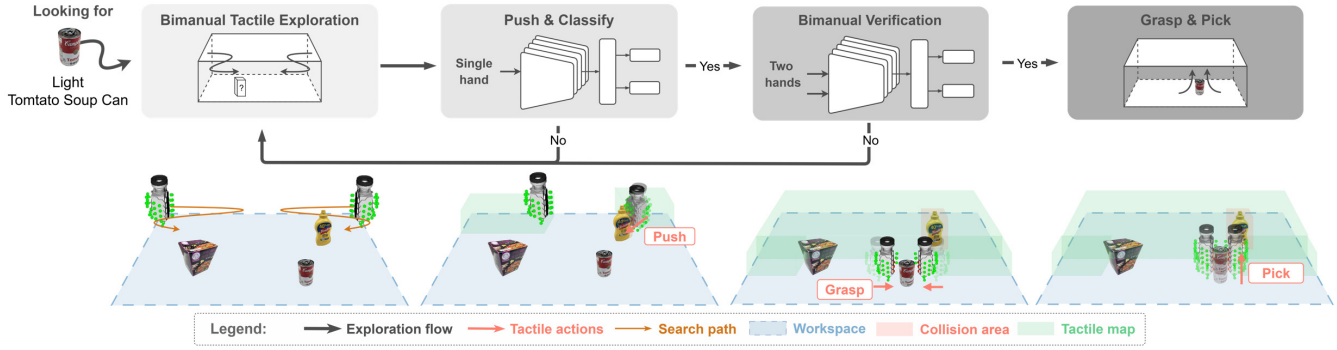

Fig. 2: Overview of the *Hide and Seek* tactile-only framework for target object retrieval. The system is given a predefined target (e.g., a light tomato soup can) and operates without vision. Arrows indicate the decision flow, where low-confidence or non-target results trigger continued exploration.

contacts, non-target objects are marked as collision areas in the tactile map. The obstacle height is estimated from the accumulated tactile point cloud by taking the maximum observed contact height within the corresponding voxel column, and the planner lifts the end-effector above this height before resuming the predefined zig-zag path.

c) Bimanual Verification: Once a candidate object has been identified, the robot performs a two-handed grasp to collect richer tactile data. Both hands' point clouds are passed into the same classifier. This step provides a reliable verification of the object's identity by incorporating shape and size cues that are only available through bimanual contact. If the verification succeeds, the object is accepted as the target.

d) Grasp and Pick: Finally, the confirmed object is grasped with both hands and lifted from the environment. At this point the search is terminated, and the robot has successfully located and retrieved the target in a purely tactile manner.

C. Sensing and Multimodal Representations

During tactile exploration, the robot collects multimodal signals from distributed skin sensors, wrist-mounted force/torque sensors, and the end-effector pose estimated by the robot controller. To make these heterogeneous measurements suitable for learning, the raw signals are transformed into stable, model-ready representations, which then serve as inputs to the multimodal encoder described in Sec. III-D.

a) Windowing: Tactile contacts are brief and noisy; single frames miss transients such as micro-bounces. We therefore use sliding windows of length $T{=}100$ frames with stride $S{=}40$. This aggregates short-term dynamics, improves SNR by temporal pooling, and yields consistent clips for both training and online inference. All signals inside a window are time-synchronized.

b) Robot Skin: We deploy a grid of $N{=}88$ skin cells with known 3D locations. Each cell provides (i) a proximity reading with a finite detection range and (ii) a scalar force obtained from the three sub-electrodes. We use two implementation-friendly representations:

(1) Tactile point cloud (prox+force). We construct a point cloud $\mathbf{P}_t \in \mathbb{R}^{N\times 6}$ at each time step t, anchored to the current end-effector pose:

$$\mathbf{P}_t[k] = \big[x_{k,t},\, y_{k,t},\, z_{k,t},\, r_{k,t},\, g_{k,t},\, b_{k,t}\big], \tag{4}$$

where the 3D position is displaced along the local surface normal by a proximity-derived radius

$$\rho_{k,t} = \min\big(d_{k,t},\, d_{\max}\big), \quad d_{\max} = 0.03 \text{ m}, \tag{5}$$

with $d_{k,t}$ the calibrated distance from the k-th proximity sensor (*no detection* $\Rightarrow \rho_{k,t}{=}d_{\max}$).

The RGB channels encode normalized sensor intensities:

$$r_{k,t} = \widetilde{p}_{k,t}, \qquad g_{k,t} = 1 - \rho_{k,t}/d_{\max}, \qquad b_{k,t} = \widetilde{f}_{k,t}, \tag{6}$$

where $\widetilde{p}_{k,t}$ and $\widetilde{f}_{k,t}$ denote normalized proximity and force signals, respectively. Thus, proximity influences both the point displacement and the R/G channels, while force controls B.

(2) Virtual wrenches from distributed skin The tactile skin offers two complementary sensing modes: *proximity*, which detects objects before contact, and *force*, which measures normal pressure during contact. To unify these signals, we treat proximity as a "virtual force" aligned with the local normal, so that both sensing modes can be aggregated under the same wrench formulation.

For the k-th skin cell in its local frame:

$${}^{\text{cell}_k}P = \begin{bmatrix} 0 \\ 0 \\ w_p\, p_{k,t} \end{bmatrix}, \qquad {}^{\text{cell}_k}F = \begin{bmatrix} 0 \\ 0 \\ w_f \sum_{i=1}^{3} f_{k,i,t} \end{bmatrix}, \tag{7}$$

where $p_{k,t}$ is the proximity measurement, $f_{k,i,t}$ the three capacitive force channels, and w_p, w_f are scalar gains. Both ${}^{\text{cell}_k}P$ and ${}^{\text{cell}_k}F$ are transformed to the hand frame and accumulated across all cells, yielding the virtual wrenches $\mathbf{W}_{\text{skin,prox}}$ and $\mathbf{W}_{\text{skin,force}}$.

For force sensing, using the calibrated transforms $\big({}^{\text{hand}}R_{\text{cell}_k}, {}^{\text{hand}}t_{\text{cell}_k}\big)$, all cell contributions are expressed in the hand frame and aggregated:

$$F_\Sigma = \sum_{k=1}^{N} {}^{\text{hand}}R_{\text{cell}_k}\, {}^{\text{cell}_k}F, \tag{8}$$

$$\tau_\Sigma = \sum_{k=1}^{N} \Big({}^{\text{hand}}t_{\text{cell}_k}\Big) \times \Big({}^{\text{hand}}R_{\text{cell}_k}\, {}^{\text{cell}_k}F\Big). \tag{9}$$

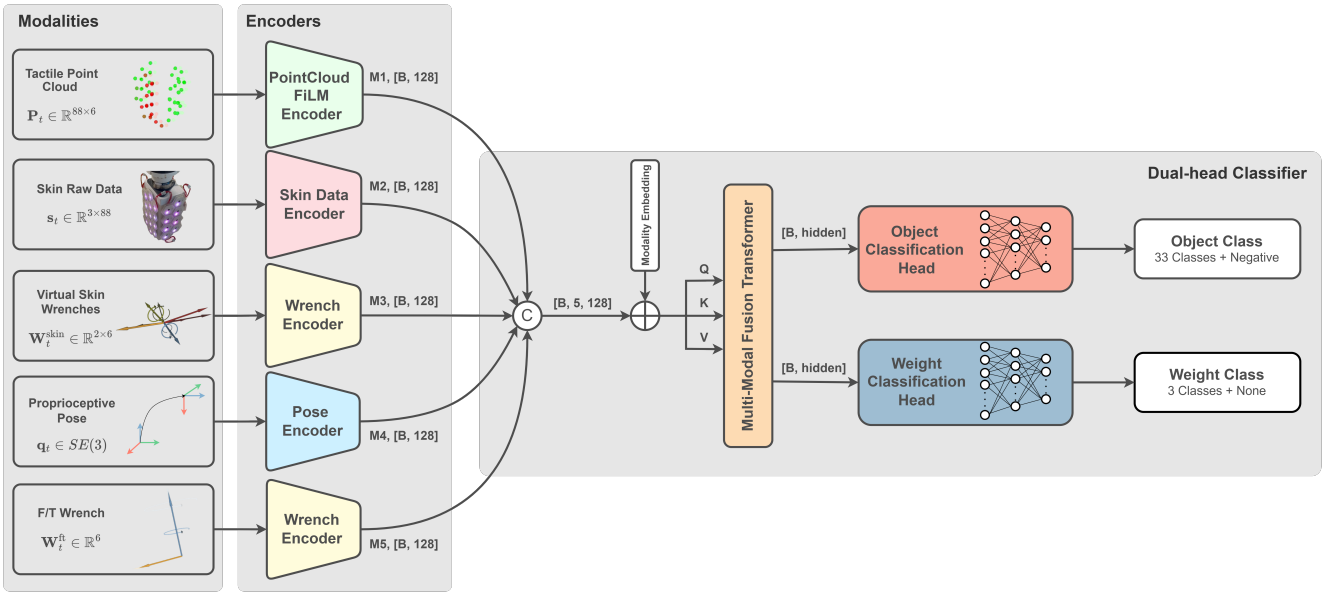

Fig. 3: **Dual-head multimodal classification framework.** Five synchronized modalities are used as input: (M1) tactile point cloud, (M2) raw skin signals, (M3) virtual skin wrenches, (M4) end-effector pose, and (M5) wrist force/torque wrench. Each modality is processed by a dedicated encoder to yield a 128-dim representation. The encoded features are concatenated, augmented with modality embeddings, and integrated by a multi-modal fusion transformer. The fused representation is forwarded to a dual-head classifier that jointly predicts the object class (33 categories + negative) and weight class (3 categories + none).

The same procedure is applied to proximity readings, yielding $\mathbf{W}_{\text{skin,prox}}$. Finally, we form two aggregated 6-D wrenches:

$$\mathbf{W}_{\text{skin,force}} = \left[F_\Sigma^\top, \tau_\Sigma^\top\right]^\top, \quad \mathbf{W}_{\text{skin,prox}} = \left[P_\Sigma^\top, \tau_\Sigma^\top\right]^\top. \tag{10}$$

Stacking both yields $\mathbf{W}_t^{\text{skin}} \in \mathbb{R}^{2\times6}$, which compactly summarizes distributed tactile feedback across the hand.

c) Proprioceptive Pose: Since tactile point clouds are anchored to the end–effector (EE), the same contact pattern in the hand frame may correspond to different world–frame hypotheses. Adding proprioception resolves such ambiguities (e.g., which side of the object is touched) and stabilizes temporal aggregation. We use a 7D pose $\mathbf{q}_t = [x, y, z, q_x, q_y, q_z, q_w]$, with the quaternion renormalized after augmentation.

d) Wrist Force/Torque: Each wrist hosts a 6-axis F/T sensor, providing wrenches $\mathbf{W}_t^{\text{ft}} = [F_\Sigma^\top,\ \tau_\Sigma^\top]^\top \in \mathbb{R}^6$. This modality is useful for distinguishing object weight but lacks the geometric detail captured by the skin.

e) Shapes and Normalization: All modalities are preprocessed consistently. Tactile point clouds (88×6) use per-channel min–max normalization. Raw skin signals (prox/force/dist, 44-D per hand) suffer from noise and drift; we apply LayerNorm to reduce gain variations and session-dependent offsets. Virtual skin wrenches and wrist F/T wrenches (both 6-D) are mean–std normalized using training statistics. End-effector poses (7-D) use min–max normalization for translation with quaternions constrained to unit norm.

f) Data Augmentation: We apply lightweight, modality-aware augmentations to improve robustness. Point clouds are perturbed in geometry (XYZ jitter, point dropout) and appearance (RGB noise). Wrench signals receive Gaussian noise, offsets, and occasional channel dropout. Raw skin signals are corrupted with noise, patch masking, and slight temporal shifts. Poses are perturbed with small translations/rotations and re-normalized quaternions. These augmentations mimic realistic sensor imperfections and mitigate overfitting.

D. Dual-head Multimodal Classifier

The recognition module is designed to jointly predict both the object category and its weight attribute. This dual-head design leverages two key advantages of haptics: (i) weight is an intrinsic tactile property that vision alone cannot access, and (ii) predicted weight further informs the admittance controller by guiding grasp force selection (e.g., heavier objects require stronger but still safe grips).

a) Modality-specific Encoders.: Each observation window $\tilde{\mathbf{X}} \in \mathbb{R}^{T\times D}$ is mapped into a d-dimensional latent embedding (d=128) by a modality-specific encoder. All non-point-cloud modalities operate on the full stream, while the tactile point cloud uses a mask to distinguish between single- and dual-hand exploration.

(1) Skin raw signals and wrench-like streams. We encode each sequence with 1D convolutions for local dynamics, multi-head self-attention for long-range dependencies, and temporal mean pooling to obtain the modality embedding:

$$\mathbf{z}^{(m)} = \tfrac{1}{T}\sum_{t=1}^{T} \mathrm{MHSA}\Big(\mathrm{Conv1D}(\tilde{\mathbf{X}})\Big)_t \in \mathbb{R}^d. \tag{11}$$

(2) Tactile point cloud. We adopt a FiLM encoder where

intensity channels modulate geometry features:

$$\mathbf{G}' = \mathrm{ReLU}\big(\gamma(\mathbf{C}) \odot \mathbf{G} + \beta(\mathbf{C})\big), \tag{12}$$

with modulation parameters $\gamma(\cdot), \beta(\cdot)$ predicted by an MLP from intensity features $\mathbf{C}$. Masked pooling excludes inactive hand points:

$$\mathbf{z}^{(\mathrm{pc})} = \Big[\max_{n:M_n=1} \mathbf{G}'_n,\ \tfrac{1}{\sum_n M_n} \sum_n M_n \mathbf{G}'_n\Big]. \tag{13}$$

(3) Proprioceptive pose. Each pose $(\mathbf{p}_t, \mathbf{q}_t)$ is passed through a lightweight temporal encoder. Quaternion normalization is enforced at every step to prevent drift, producing $\mathbf{z}^{(\mathrm{pose})} \in \mathbb{R}^d$.

b) Transformer-based Fusion: Per-modality embeddings $\{\mathbf{z}^{(m)}\}_{m=1}^{M}$ are treated as tokens. After adding learnable modality embeddings $\mathbf{e}^{(m)}$, an L-layer Transformer encoder integrates information across modalities:

$$\mathbf{Z}_0[m] = \mathbf{z}^{(m)} + \mathbf{e}^{(m)}, \qquad \mathbf{Z}_L = \mathrm{TFEnc}^{(L)}(\mathbf{Z}_0). \tag{14}$$

Masked mean pooling then aggregates the fused tokens:

$$\mathbf{f} = \mathrm{LayerNorm}\left(\tfrac{1}{\sum_m s_m} \sum_{m=1}^{M} s_m\, \mathbf{Z}_L[m]\right) \in \mathbb{R}^d, \tag{15}$$

where $s_m \in \{0,1\}$ indicates modality availability. This formulation naturally handles missing or inactive inputs.

c) Heads and Objective.: A shared MLP backbone maps the fused feature $\mathbf{f}$ into a compact representation $\mathbf{g} \in \mathbb{R}^d$. Two linear heads then output logits for object and weight prediction:

$$\mathbf{o} = \mathbf{W}_{\mathrm{o}}\mathbf{g} + \mathbf{b}_{\mathrm{o}} \in \mathbb{R}^{C_{\mathrm{obj}}}, \quad \mathbf{w} = \mathbf{W}_{\mathrm{w}}\mathbf{g} + \mathbf{b}_{\mathrm{w}} \in \mathbb{R}^{C_{\mathrm{w}}}, \tag{16}$$

where $C_{\mathrm{obj}} = 34$ (33 objects + negative) and $C_{\mathrm{w}} = 4$ (light, medium, heavy, none).

The training loss is a weighted sum of cross-entropy terms:

$$\mathcal{L}_{\mathrm{total}} = \lambda_{\mathrm{obj}}\, \mathcal{L}_{\mathrm{obj}}(\hat{o}, o) + \lambda_{\mathrm{w}}\, \mathcal{L}_{\mathrm{weight}}(\hat{w}, w), \tag{17}$$

where $\mathcal{L}_{\mathrm{obj}}$ and $\mathcal{L}_{\mathrm{weight}}$ are standard cross-entropy losses. We set λ_{obj}=1.0 and λ_{w}=0.5 to prioritize object recognition, while still exploiting weight as a complementary cue, which empirically improves object accuracy.

IV. Hide-and-Seek Dataset

A. Overview

We collect a large-scale Hide-and-Seek (HAS) tactile dataset using Tactile Omnidirectional Mobile Manipulator (TOMM) [26]. The platform is equipped with hand-mounted skins in a controlled, vision-denied tabletop workspace. Because our pipeline relies on bimanual grasps, objects are sized for two-hand manipulation while avoiding inter-arm collisions, yet remain representative of everyday items. Starting from a YCB subset [27], we add boxes, cups, cans, foams, and bags to span diverse shapes, textures, and compliances. Several items appear in multiple weight variants (via ballast or liquid), defining three classes: light (0–150 g), medium (150–300 g), and heavy (>300 g). Weight cues aid recognition and guide admittance-controlled grasping. HAS is designed as a benchmark for tactile-only recognition in vision-denied settings. Dataset and code are available via the project page.

This results in a dataset of **34 objects** (33 objects plus a negative class), with up to three weight variants per object. In total, we obtain **61 object–weight categories** and **1.12M frames** grouped into **5.4k trajectories** ($\approx$ 2 s each), as shown in Fig. 4. For evaluation, we report results at three granularities: 61 raw categories, 34 object classes, and 4 weight classes (light, medium, heavy, none).

To mitigate domain shift, the training and validation/test sets were recorded on different days, with sensor resetting before each session. The official split allocates **69.4%** of trajectories to training, **19.7%** to validation, and **10.9%** to test. Each trajectory is downsampled to 100 Hz and stored frame-wise with five synchronized modalities (see Sec. III-C for details):

- **Tactile point cloud**: XYZ anchored to the end-effector pose with RGB channels encoding proximity and force.
- **Skin raw data**: per-cell proximity, force, and calibrated distance.
- **Virtual skin wrenches**: aggregated proximity- and force-based wrenches (F, τ) in the hand frame.
- **Proprioceptive pose**: 7D pose $(x, y, z, q_x, q_y, q_z, q_w)$.
- **F/T wrench**: 6-axis force/torque wrench.

A binary mask is also provided for the point clouds to distinguish between single-hand and bimanual data (e.g., if only one hand is active, the other 44 points are masked out).

B. Collection Protocol

Data collection consists of two parts: *single-hand pushing* and *bimanual grasping*. For the single-hand setting, objects are placed at random positions in front of the hand to simulate diverse contact conditions. The hand moves forward along a fixed direction for a constant distance (5 cm), followed by a small retraction. This cycle is repeated several times per episode, during which objects may rotate or shift, naturally enriching the observation space. After each sequence, the other hand resets the object to its initial position. For bimanual grasping, the hands start from fixed initial poses and apply a contact force threshold of 0.3 N per hand. This prevents both damage to the robot and deformation of the objects. Objects are placed with random offsets between the two hands to mimic realistic grasping scenarios.

All data are recorded as `rosbag` files, ensuring consistent storage for both *offline training* and *online evaluation*. The same preprocessing pipeline and dataloaders are used in both cases, unifying the data flow. To further mitigate domain shift, data were collected over multiple days with sensor resets before each session. The open-source dataset is provided via Hugging Face in `parquet` format.

C. Class Distribution

Figure 4 (top-right) shows the sample distribution over 61 categories. Overall, the dataset is fairly balanced, but a moderate long-tail effect remains: the most frequent classes (e.g., *Tomato Soup*, *Tape*) contain nearly twice as many samples as the least represented ones (e.g., *Mustard Medium*, *Youcook Light*). The bottom-right histogram further illustrates that

Fig. 4: Overview of the Hide-And-Seek dataset. **Left:** Gallery of 33 objects spanning diverse shapes, materials, and compliances, many with up to three weight variants. **Top-right:** Class distribution of the training set across 61 categories (33 objects $\times$ weights + 1 negative). **Bottom-right:** Histogram of per-class sample counts, with the red dashed line indicating the mean ($\sim$12.7k). In total, the dataset comprises **1.12M frames** grouped into **5.4k trajectories**.

most classes cluster around the mean ($\sim$12.7k samples), with only a few outliers at the extremes.

This imbalance mainly results from the data collection process. In practice, certain recording sessions suffered from missing topics or corrupted signals, and we deliberately removed these low-quality trajectories to ensure dataset reliability. While this filtering improves overall quality, it reduces the number of usable samples for some categories, leading to the observed imbalance. Nevertheless, since the majority of classes remain close to the mean, the dataset still provides a sufficiently balanced basis for robust model training.

V. Experiments

A. Experimental Setup

Figure 5 shows the TOMM Robot, a dual-arm mobile robot designed for whole-body tactile interaction. Our experiments focus on the hands, which are covered by hexagonal skin cells providing force and proximity sensing [7], [28]. Each hand carries 44 cells (88 in total), registered in the kinematic chain, allowing contacts to be represented as *tactile point clouds*. In addition, skin signals are aggregated into virtual wrenches and integrated into an admittance control framework [29], ensuring safe and compliant interaction during pushing and bimanual grasping in a vision-denied setting.

We report three metrics: (i) **Object accuracy**, the ratio of correctly classified objects; (ii) **Weight accuracy**, the ratio of correct weight predictions; and (iii) **Joint accuracy**, which requires both object and weight to be correct. For online trials, we also measure the proportion of errors corrected by the bimanual verification stage.

B. Offline Classification Results

As shown in Table I, under the strict joint metric, multimodal fusion is essential for robust tactile recognition. The five-stream model **M12345** achieves the best overall performance (**79.6%**), closely followed by **M123** (79.0%)

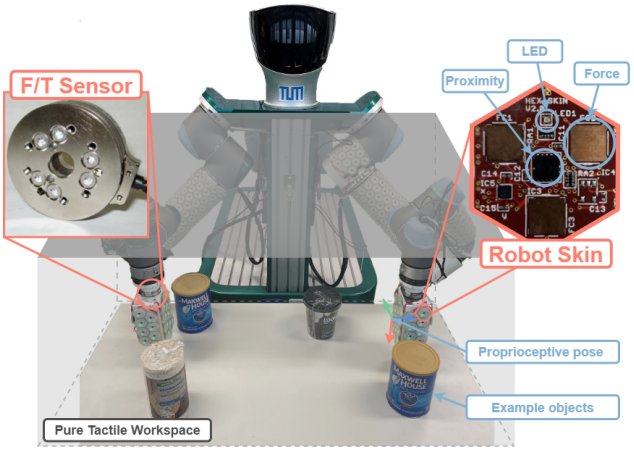

Fig. 5: Experimental platform: the Tactile Omnidirectional Mobile Manipulator (TOMM) equipped with hand-mounted tactile skin. **Left inset:** wrist-mounted F/T sensor. **Right inset:** hexagonal skin cell integrating proximity and force sensing with LEDs for visualization. The workspace is vision-denied (*pure tactile workspace*) and populated with example objects, including multiple weight variants of the same object (e.g., Maxwell Coffee Can).

and **M1234** (78.6%). Interestingly, excluding pose (**M123**) slightly improves over the four-stream baseline, suggesting that pose contributes marginal or noisy information. Among pairs, **M23** (76.1%) shows that raw skin signals (M2) and virtual wrenches (M3) are already strongly complementary.

Ablation insights: Removing skin raw (M2) produces the largest drop (-11.2 pts from **M1234** to **M134**), confirming its role as the dominant modality for object identity. Tactile point cloud (M1) and virtual wrenches (M3) are complementary: removing either (**M124**, **M234**) yields moderate declines. Wrist F/T (M5) contributes little in isolation (15.1% joint) but provides a small consistent gain when fused. Single-modality performance highlights the limitations of isolated cues: M2 or M1 alone reach only $\sim$ 60%, while M4-

TABLE I: Offline ablations on the HAS validation set. **Joint Accuracy** requires both object and weight correct.

Modalities	Object Acc (%)	Weight Acc (%)	Joint Acc (%)
M4	5.6	43.3	1.8
M5	22.2	67.0	15.1
M3	45.3	64.9	35.1
M14	72.8	72.9	55.6
M1	75.5	75.6	60.6
M2	87.4	65.4	59.9
M13	82.5	76.6	67.8
M134	81.2	78.0	67.4
M12	87.8	79.8	73.8
M124	89.6	79.4	75.3
M234	89.4	79.4	74.3
M23	89.7	81.4	76.1
M1234	90.7	82.8	78.6
M123	90.8	**83.2**	79.0
M12345	**91.1**	83.1	**79.6**

* Modalities: M1= Tactile Point Cloud, M2=Skin raw data, M3=Virtual skin wrenches, M4=Proprioceptive pose, M5=Wrist F/T wrench.

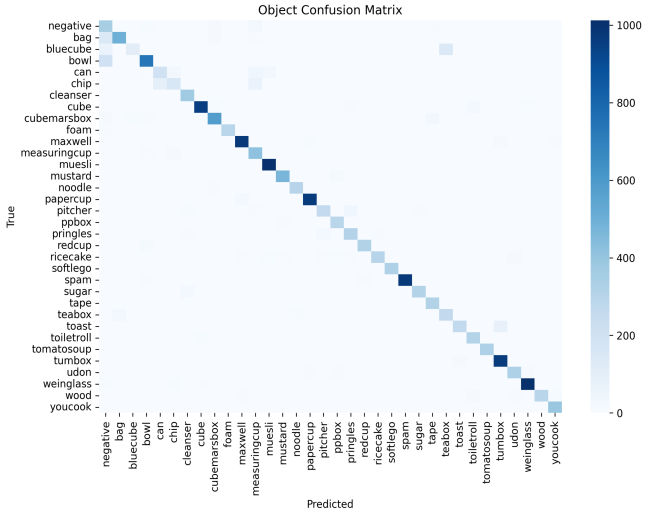

Fig. 6: Confusion matrix over 34 object classes (object-only accuracy).

only (pose) collapses to 1.8% due to lack of direct contact.

The confusion matrix for the 34 object classes (Fig. 6) reveals that most errors occur between objects with similar geometries, such as bowls and cups. This is primarily attributed to the low spatial resolution of the tactile sensors ($\sim$ 2 cm), which limits their ability to discern fine-grained morphological details.

The per-class weight classification results in Table II reveal a clear performance hierarchy. The light class achieves excellent precision (92.76%), indicating highly reliable detection of lightweight objects. In contrast, the "none" class demonstrates significantly lower precision (44.30%), with a high false positive rate indicating frequent confusion between an absence of weight and light tactile contact, often due to sensor noise. The heavy and medium classes show similar, moderate precision levels (79.4%), reflecting the expected difficulty in disambiguating these categories based on tactile cues alone. The overall precision of 83.20% confirms the model's general effectiveness, though performance varies considerably based on semantic and tactile distinctness.

TABLE II: Weight classification per-class metrics

Class	TP	FP	TN	FN	Precision (%)
light	6,701	523	9,480	940	92.76
medium	4,115	1,073	11,079	1,377	79.32
heavy	3,499	909	12,636	600	79.38
none	365	459	16,773	47	44.30
Overall	**14,680**	**2,964**	**–**	**2,964**	**83.20**

* TP, FP, TN, FN denote True Positive, False Positive, True Negative, and False Negative.

TABLE III: Online classification on TOMM with paired trials (same trajectory per push and grasp, n=70).

Setting	Object		Weight		Joint	
	Num	P(%)	Num	P(%)	Num	P(%)
Success Case	55	78.6	53	75.7	43	61.4
Failure Case	15	21.4	17	24.3	27	38.6
Corrected Case	2	13.3	3	17.6	3	11.1

* **Num** = number of trials, **P(%)** = percentage over total trials.
** **Corrected Case** is a false positive predicted by single hand and corrected by the bimanual verification stage.

C. Online Evaluation

Each online trial executes *one continuous trajectory* that first performs a single-hand **push** and then a bimanual **grasp** on the same object. We evaluate the *same* trained classifier twice per trial: (i) immediately after the push, using mask-aware fusion with only the contacting hand's tokens active; and (ii) after the grasp, using both hands' tokens. This paired design controls for object instance, scene layout, and contact location, enabling a fair within-trial comparison.

As shown in Table III, the system achieved correct object recognition in 78.6% of trials (55/70) during the bimanual verification phase. Weight classification proved slightly less accurate, with 75.7% of trials (53/70) correctly identified. The most demanding metric—joint object and weight recognition—was achieved in 61.4% of trials (43/70). Despite incorporating multiple mechanisms to mitigate domain shift, tactile sensor noise and contact variability remain significant limiting factors, which constrain the achievable real-world success rate.

Notably, the bimanual verification stage corrected classification errors from the initial push phase in a significant number of cases. Specifically, it resolved 13.3% of object misclassifications (2/15), 17.6% of weight errors (3/17), and 11.1% of joint recognition failures (8/27). These corrections frequently involved distinguishing between geometrically similar objects such as boxes and cups, where the additional size and shape information obtained through bimanual grasping proved particularly valuable. The results demonstrate that while single-hand pushing provides preliminary tactile data, bimanual verification adds critical information through enveloping grasp and symmetrical contact patterns, enhancing overall reliability in tactile-only recognition tasks.

VI. Summary

We present a tactile-only pipeline for object search in vision-denied settings. Inspired by human behavior, our *Tactile Hide-and-Seek* strategy enables effective exploration

and localization through structured sweeping and contact-driven interaction. To support learning in this setting, we introduce the *HAS dataset*, a large-scale multimodal tactile dataset with weight annotations, designed to capture material and physical cues and to facilitate reproducible evaluation. Using this dataset, we show that multimodal tactile sensing supports accurate object recognition and property estimation via a dual-head classifier. To improve reliability in deployment, we further introduce a bimanual tactile verification step, which reduces critical misclassifications and increases overall success. Together, these components form a tactile-only system that runs end-to-end—from initial search and first contact to compliant bimanual grasping and verification.

Nevertheless, the broader challenge of achieving generalized tactile search in highly unstructured, large-scale, or dynamic real-world environments remains open. Future work must address scalability, environmental variability, and further sensor fusion to extend these promising results to more complex and diverse settings.

ACKNOWLEDGMENT

Wenlan Shen and Simon Armleder were funded by Deutsche Forschungsgemeinschaft (DFG) under grant 505597051 and grant 502086040. Fengyi Wang was funded by the Federal Ministry of Education and Research (BMBF) under grant number 01GQ2108.

REFERENCES

[1] J. Xu, H. Lin, S. Song, *et al.*, "Tandem3d: Active tactile exploration for 3d object recognition," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 10 401–10 407.

[2] M. Kaboli, K. Yao, D. Feng, *et al.*, "Tactile-based active object discrimination and target object search in an unknown workspace," *Auton Robot*, vol. 43, no. 1, pp. 123–152, Jan. 2019.

[3] S. Pai, T. Chen, M. Tippur, *et al.*, "TactoFind: A Tactile Only System for Object Retrieval," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, May 2023, pp. 8025–8032.

[4] Y. Lin, A. Church, M. Yang, *et al.*, "Bi-touch: Bimanual tactile manipulation with sim-to-real deep reinforcement learning," *IEEE Robotics and Automation Letters*, vol. 8, no. 9, pp. 5472–5479, 2023.

[5] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, 2017.

[6] J. A. Fishel and G. E. Loeb, "Sensing tactile microvibrations with the biotac—comparison with human sensitivity," in *2012 4th IEEE RAS & EMBS international conference on biomedical robotics and biomechatronics (BioRob)*, IEEE, 2012, pp. 1122–1127.

[7] G. Cheng, E. Dean-Leon, F. Bergner, *et al.*, "A Comprehensive Realization of Robot Skin: Sensors, Sensing, Control, and Applications," *Proceedings of the IEEE*, vol. 107, no. 10, pp. 2034–2051, Oct. 2019.

[8] R. Gao, Y.-Y. Chang, S. Mall, *et al.*, "Objectfolder: A dataset of objects with implicit visual, auditory, and tactile representations," in *5th Annual Conference on Robot Learning*, 2021.

[9] B. M. R. Lima, V. N. S. S. Danyamraju, T. E. A. d. Oliveira, *et al.*, "A multimodal tactile dataset for dynamic texture classification," *Data in Brief*, vol. 50, p. 109 590, 2023.

[10] F. Wang, X. Fu, N. Thakor, *et al.*, "Human-inspired soft anthropomorphic hand system for neuromorphic object and pose recognition using multimodal signals," *arXiv preprint arXiv:2509.02275*, 2025.

[11] H. Xing, K. Z. Boey, Y. Wu, *et al.*, "Multi-modal graph convolutional network with sinusoidal encoding for robust human action segmentation," *arXiv preprint arXiv:2507.00752*, 2025.

[12] S. Qiu, B. Li, X. Wang, *et al.*, "DT-Transformer: A Text-Tactile Fusion Network for Object Recognition," *IEEE Transactions on Haptics*, vol. 18, no. 1, pp. 164–174, Jan. 2025.

[13] J. Zhao, Y. Ma, L. Wang, *et al.*, "Transferable tactile transformers for representation learning across diverse sensors and tasks," in *8th Annual Conference on Robot Learning*, 2024.

[14] S. Jiang and L. L. Wong, "Active Tactile Exploration using Shape-Dependent Reinforcement Learning," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2022, pp. 8995–9002.

[15] J. Zhao and E. H. Adelson, "Gelsight svelte: A human finger-shaped single-camera tactile robot finger with large sensing coverage and proprioceptive sensing," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2023, pp. 8979–8984.

[16] A.-H. Shahidzadeh, S. J. Yoo, P. Mantripragada, *et al.*, "Actexplore: Active tactile exploration on unknown objects," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 3411–3418.

[17] K.-W. Lee, Y. Qin, X. Wang, *et al.*, "DexTouch: Learning to Seek and Manipulate Objects With Tactile Dexterity," *IEEE Robotics and Automation Letters*, vol. 9, no. 12, pp. 10 772–10 779, Dec. 2024.

[18] K. Yu, Y. Han, Q. Wang, *et al.*, "Mimictouch: Leveraging multi-modal human tactile demonstrations for contact-rich manipulation," in *8th Annual Conference on Robot Learning*, 2024.

[19] C. Smith, Y. Karayiannidis, L. Nalpantidis, *et al.*, "Dual arm manipulation—A survey," *Robotics and Autonomous Systems*, vol. 60, no. 10, pp. 1340–1353, Oct. 2012.

[20] T. Z. Zhao, J. Tompson, D. Driess, *et al.*, "ALOHA unleashed: A simple recipe for robot dexterity," in *8th Annual Conference on Robot Learning*, 2024.

[21] C. Chi, Z. Xu, C. Pan, *et al.*, "Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots," *arXiv preprint arXiv:2402.10329*, 2024.

[22] S. Liu, L. Wu, B. Li, *et al.*, "RDT-1b: A diffusion foundation model for bimanual manipulation," in *The Thirteenth International Conference on Learning Representations*, 2025.

[23] M. F. Karim, S. Bollimuntha, M. S. Hashmi, *et al.*, "Da-vil: Adaptive dual-arm manipulation with reinforcement learning and variable impedance control," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2025, pp. 11 896–11 903.

[24] Z. Sun, Z. Shi, J. Chen, *et al.*, "Vtao-bimanip: Masked visual-tactile-action pre-training with object understanding for bimanual dexterous manipulation," *arXiv preprint arXiv:2501.03606*, 2025.

[25] W. Shen, S. Armleder, and G. Cheng, "Tactile-based dual-arm manipulation with physical human-robot interaction," in *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, 2025, pp. 14–22.

[26] E. Dean-Leon, F. Bergner, K. Ramirez-Amaro, *et al.*, "From multi-modal tactile signals to a compliant control," in *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, Nov. 2016, pp. 892–898.

[27] B. Calli, A. Singh, J. Bruce, *et al.*, "Yale-cmu-berkeley dataset for robotic manipulation research," *The International Journal of Robotics Research*, vol. 36, no. 3, pp. 261–268, 2017.

[28] P. Mittendorfer and G. Cheng, "Humanoid Multimodal Tactile-Sensing Modules," *IEEE Transactions on Robotics*, vol. 27, no. 3, pp. 401–410, Jun. 2011.

[29] S. Armleder, E. Dean-Leon, F. Bergner, *et al.*, "Interactive Force Control Based on Multimodal Robot Skin for Physical Human-Robot Collaboration," *Advanced Intelligent Systems*, vol. 4, no. 2, p. 2 100 047, 2022.