IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 4 December 2025, accepted 29 December 2025, date of publication 12 January 2026, date of current version 16 January 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3652803

RESEARCH ARTICLE

Egocentric Hand Activity Video Dataset and Bidirectional Motion-Priors for Hand Action Recognition

JIYOUNG SEO[1], DONG IN LEE[1], PILHYEON LEE[2], JIWOO LEE[1], YOUNHEE GIL[3], KARTHIK RAMANI[4,5], (Member, IEEE), AND SANGPIL KIM[1]

[1]Department of Artificial Intelligence, Korea University, Seoul 02841, Republic of Korea
[2]Department of Artificial Intelligence, Inha University, Incheon 22212, Republic of Korea
[3]Electronics and Telecommunications Research Institute (ETRI), Daejeon 34129, Republic of Korea
[4]Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA
[5]School of Mechanical Engineering, Purdue University, West Lafayette, IN 47907, USA

Corresponding author: Sangpil Kim (spk7@korea.ac.kr)

This work was supported in part by the Electronics and Telecommunications Research Institute (ETRI) grant funded by Korean Government (Development of Spatial Media Technology and Interaction Technology for Convergence of the Real and Virtual World, 90%) under Grant 26ZC1100, and in part by the National Research Foundation of Korea (NRF) grant funded by Korean Government [Ministry of Science and ICT (MSIT)] (10%) under Grant RS-2025-00521602.

ABSTRACT Recognizing tool-based hand activities from a first-person view is a critical yet challenging task in computer vision, due to the complexity of hand-object interactions and often subtle, ambiguous motion patterns. In real-world manufacturing scenarios, these challenges are exacerbated by bidirectional action pairs whose visual cues are almost identical, with differences revealed only through subtle motion dynamics. However, existing datasets rarely capture these direction-sensitive interactions at scale, particularly in realistic tool-use contexts, limiting the ability of current models to learn fine-grained motion dynamics essential for accurate recognition. We introduce **Ego-Bi** (Egocentric-Bidirectional dataset), a large-scale, real-world egocentric RGB video dataset comprising 1,223 video sequences and 622,737 frames that cover diverse tool-use activities in unconstrained environments. Ego-Bi provides an extended 38-category hand type taxonomy, detailed object–tool labels, and challenging bidirectional action pairs, offering rich semantic and temporal cues for modeling complex hand–object interactions. In addition, to address the ambiguity in motion dynamics, we propose a **BMP** (Bidirectional Motion Prior module) that derives rotation and directional cues from predicted 3D hand poses to improve class separability of visually similar actions. Experimental results on Ego-Bi demonstrate that our approach improves bidirectional action recognition accuracy by +8.96% over the baseline, while also yielding consistent gains across general action classes without requiring costly 3D pose annotations. Furthermore, the proposed motion priors generalize effectively to other egocentric benchmarks, underscoring their robustness in handling visually similar, direction-sensitive actions.

INDEX TERMS Hand action recognition, hand-object interaction, dynamic motion cue, hand type taxonomy, hand pose estimation.

I. INTRODUCTION

Egocentric hand action recognition has become an active area of research in AR, XR, and computer vision, with a large range of applications in human-computer interaction, task automation, and wearable devices. Therefore, understanding hands with computer vision systems has been deeply explored through hand pose estimation [1], [2], [3], hand gesture recognition [4], [5], [6], and hand action recognition [7], [8], [9]. Recent approaches in hand action recognition commonly

The associate editor coordinating the review of this manuscript and approving it for publication was Meryem Erbilek.

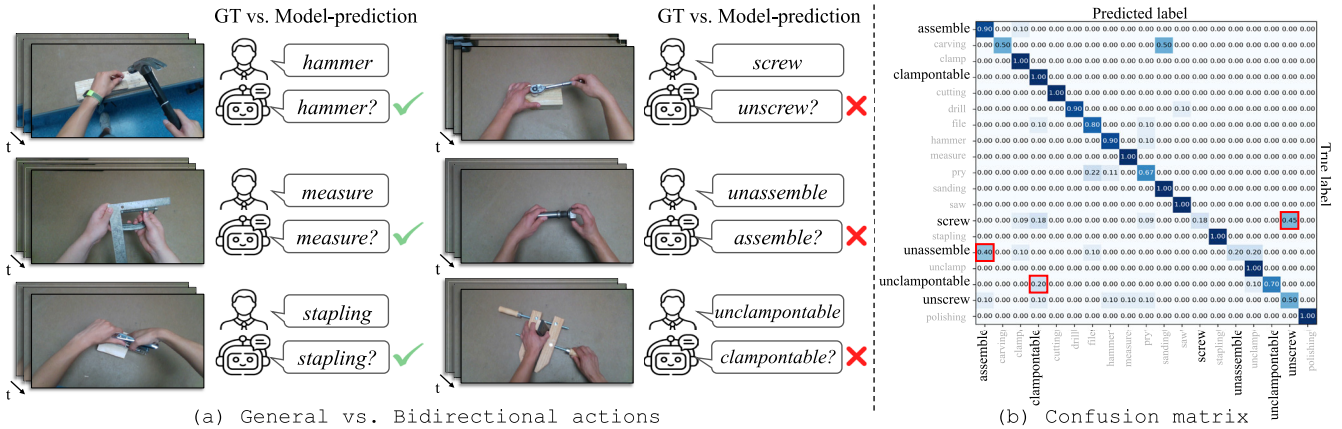

FIGURE 1. **Examples of action recognition in our dataset: `(a) General vs. Bidirectional actions`, and `(b) Confusion matrix` with a conventional baseline model. While general actions are correctly recognized by conventional models, bidirectional actions with opposite motion dynamics are frequently misclassified, as further reflected in the confusion matrix results.**

integrate RGB appearance features, which provide semantic context of hand–object interactions, with pose-based representations that capture fine-grained motion dynamics difficult to infer from appearance alone. For example, HandFormer [10] achieves state-of-the-art performance on H2O [11] by fusing sparsely sampled RGB with densely sampled pose modalities using its proposed pose encoder. Similar trends appear in full-body action recognition [12], where skeleton poses provide strong contextual cues for temporal modeling. These studies highlight that relying solely on appearance features is insufficient, as explicit motion cues play a critical role in improving recognition, particularly when dealing with subtle hand dynamics.

However, for local body parts like the hands, where motion is subtle and structurally complex, recognizing fine-grained tool-use actions remains challenging due to (a) the lack of large-scale datasets that capture diverse, temporally rich hand–object interactions, and (b) the difficulty of extracting dynamic motion cues, particularly for bidirectional action pairs (e.g., screw vs. unscrew, assemble vs. unassemble), where actions share almost identical appearance and differ only in motion direction. In realistic tool-use manufacturing scenarios, actions with opposite motion dynamics but otherwise similar visual and contextual conditions are frequently misclassified by conventional models. This problem becomes more severe when temporal motion patterns are weakly represented, as shown by the action-class confusion matrix analysis in Figure 1(b). In addition, typical misclassification cases for general actions and bidirectional actions are shown in Figure 1(a), further highlighting the challenge of distinguishing bidirectional actions. These findings motivate the design of modules capable of explicitly capturing fine-grained, orientation- and direction-aware motion dynamics from hand pose sequences.

Previous works have explored extracting temporal features from hand pose data, for instance, De Smedt et al. [13], which computed direction and rotation histograms by tracking the palm joint over time. While effective for general gesture recognition, such histogram-based representations aggregate motion directions over time and may discard sequential ordering of joint movements. As a result, bidirectional actions with opposite translational directions can produce nearly identical histogram signatures, making them difficult to distinguish in realistic scenarios. In addition, due to the fine-grained and articulated nature of hand movements, previous research [7], [10], [14] often depend on datasets that include high-precision labels such as 3D hand pose or 6D object pose ground truth. However, acquiring such labels typically requires costly pipelines, such as depth sensors, marker-based capture systems, or labor-intensive multi-view annotation processes, which pose significant scalability challenges.

To address this gap, we propose a large-scale egocentric hand action recognition benchmark, *Ego-Bi*, constructed from real-world videos in manufacturing environments. Ego-Bi consists of 1,223 video sequences and 622,737 frames, covering a diverse range of tool-based interaction scenarios rarely explored in previous works [11], [15], [16]. In particular, it includes challenging bidirectional action pairs such as screw/unscrew, clamp/unclamp, and assemble/unassemble, which are difficult to distinguish without fine-grained temporal geometric cues. Beyond providing diverse action scenarios, Ego-Bi includes a fine-grained extension of the hand type taxonomy from previous work [17] and frame-wise annotations for all sequences. This enables more precise action understanding by leveraging morphological and functional cues, and also serves as a rich semantic signal for training models.

Furthermore, we introduce a novel Bidirectional Motion-Prior Module *(BMP)*, designed to enhance recognition of bidirectional actions without requiring high-cost annotations. BMP analyzes the geometric patterns of joint-level temporal dynamics from predicted 3D hand poses, enabling the model to capture subtle differences between visually similar actions.

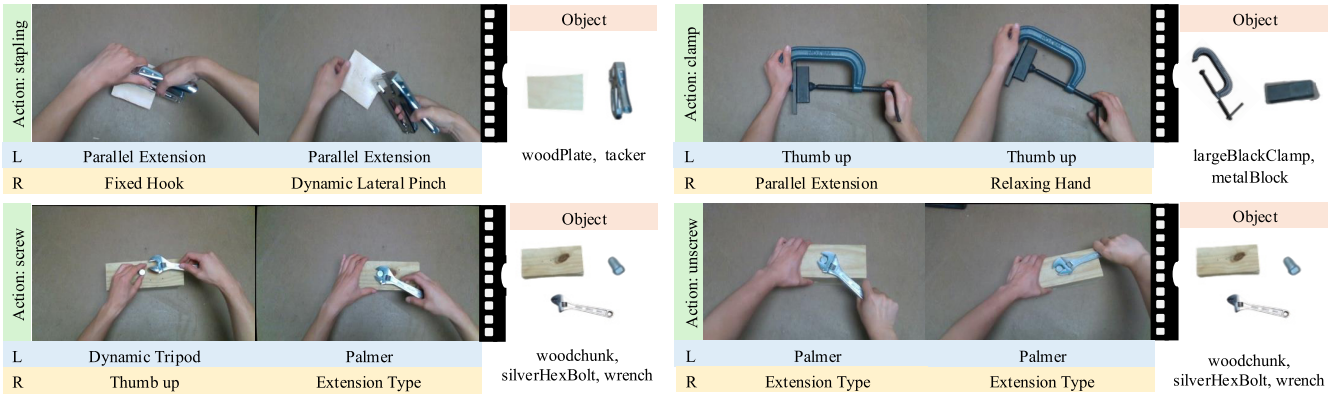

FIGURE 2. Examples of annotation samples in Ego-Bi. The top two examples show general action classes involving hand-tool interactions, while the bottom ones depict bidirectional action class pairs (e.g., screw vs. unscrew), which are visually similar but semantically opposite.

When integrated into an RGB with pose fusion pipeline and trained end-to-end, it achieves robust performance even in the absence of object or pose ground-truth labels.

Our main contributions are listed as follows:

- We propose a new large-scale real-world egocentric hand activity video dataset, Ego-Bi, comprising 1,223 sequences and 622,737 frames with detailed annotations including an extended 38-category fine-grained hand type taxonomy, object labels, and bidirectional action classes.
- We design a Bidirectional Motion-Prior module, BMP, that significantly enhances the classification of challenging action pairs, and integrate it into a unified action recognition framework.
- Experimental results show that our newly proposed pipeline significantly improves recognition accuracy by +8.96% over the baseline, even when using only predicted 3D hand poses.

II. RELATED WORK

A. HAND ACTION RECOGNITION

Previous studies [18], [19], [20] relied on static spatial cues such as hand/object bounding boxes, inter-object distances, and relative positions to localize hands and manipulated objects. As temporal modeling became crucial, ConvNet-based approaches like C2D [21], I3D [22], and SlowFast [23] were widely adopted to capture spatio-temporal features. Recurrent models such as LSTA [24] further introduced attention into LSTMs, while self-attention frameworks like EgoACO [25] integrated action, object, and context cues for enhanced temporal aggregation. More recently, transformer-based models were introduced to directly leverage temporal coherence across frames via attention mechanisms. Some studies [8], [26] fused hand joint coordinates with derived features such as velocity and inter-joint distance changes to capture both spatial and temporal context to improve action recognition accuracy. More recent efforts include TREAR [27], which employs inter-frame attention and mutual-attentional fusion for RGB-D inputs, and HTT [14], Hierarchical Temporal Transformer adopts a two-stage design, where a local transformer extracts short-term temporal features and a global transformer aggregates them into clip-level representations. Furthermore, high-performing models [7], [8], [10] for egocentric hand action recognition combine pose-centric or geometry-aware representations with efficient transformer pipelines, delivering competitive results on challenging benchmarks. However, despite these advances, existing approaches still face challenges in modeling bidirectional actions and leveraging semantic priors, motivating our work on integrating motion and hand-type cues into transformer-based recognition.

B. HAND TYPE TAXONOMY

In the context of computer vision, several works [17], [28], [29], [30], [31], [32], [33] have leveraged hand taxonomies or grasp-type annotations to provide semantic cues for hand action recognition. Early works such as Napier [28] and Cutkosky et al. [29] established foundational taxonomies: Napier introduced the distinction between power and precision grasp, while Cutkosky developed hierarchical grasp categories. Subsequently, Feix et al. [32] provided a systematic taxonomy of human hand types, serving as a foundation for subsequent research in hand action recognition. Building upon such refined taxonomies, Cai et al. [34] combined grasp types and object attributes to better characterize manipulation, while Bullock et al. [30] and Stival et al. [31] highlighted how taxonomy-based representations and hand synergies could guide sementic perception. More recently, Cho et al. [33] proposed a hand type taxonomy consisting of 33 categories; however, it is limited to grasp classification, neglecting other functional hand types. In comparison, Roh et al. [17] introduced a taxonomy that categorizes both grasp and control types, offering a more comprehensive classification. However, their control taxonomy only differentiates between non-deformable and deformable object interactions, failing to fully capture the complexity of real-world hand-object interactions. To address these limitations, we propose an extended hand type taxonomy based on Roh et al. [17], refining the classification of control types into four distinct categories: touch, force, rotate, and straight.

TABLE 1. Overview of hand–object interaction datasets. The table summarizes dataset statistics (frames, sequences, action classes, objects, subjects). The columns under *Labels* show the available ground-truth annotation types. *Scenario* specifies the activity domain, and *Tasks* indicates the main application, such as AR (action recognition), PE (pose estimation), GC (grasp classification), or HS (hand segmentation).

Dataset	#Frames	#Seq.	#Classes	#Objects	#Subj.	Labels				Scenario	Tasks
						Action	Pose	Object	Grasp/Type		
H2O [11]	571K	-	36	8	4	✓	✓	✓	✗	Daily-activities	AR+PE
FPHA [15]	105K	1,175	45	26	6	✓	✓	✓	✗	Daily-activities	AR+PE
Cho et al. [33]	1.49M	-	-	30	99	✗	✓	✓	✓	General	GC+PE
EgoHands [20]	4.8K	48	4	-	4	✓	✗	✗	✗	Game-play	HS
GTEA [35]	31K	28	7	16	4	✓	✗	✓	✗	Daily-activities	AR+HS
Assembly101 [36]	111M	362	1,380	90	53	✓	✓	✓	✗	Toy-assembly	AR
EGTEA Gaze+ [37]	2.5M	86	106	53	32	✓	✗	✓	✗	Kitchen	AR+HS
Kim et al. [38]	401K	790	15	23	5	✓	✗	✓	✗	Industrial	HS
Ego-Bi (Ours)	622K	1,223	19	63	5	✓	✗	✓	✓	Industrial	AR

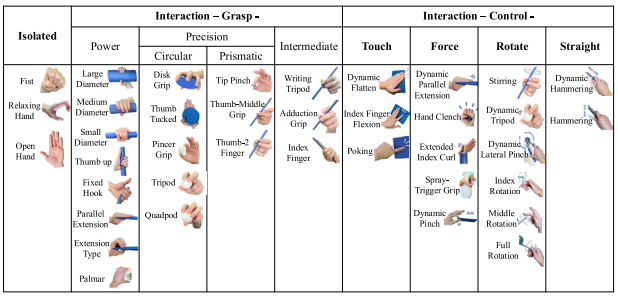

FIGURE 3. Extended hand type taxonomy (expansion of Roh et al. [17]). Our taxonomy introduces additional categories, including a new Rotate class, enabling finer annotation of dynamic hand–object interactions.

C. HAND POSE DYNAMICS IN ACTION RECOGNITION

3D hand pose is one of the most informative cues for hand-action recognition: the 3D coordinates of hand joints alone can reveal a user's intention and the nature of hand–object interaction. Most previous studies [15], [20], [34], [39] treat hand pose as a static, per-frame signal. They use 3D joint coordinates to capture the hand's geometric shape or to assign semantic meanings to specific poses. To overcome this limitation of treating hand pose only as a static signal, recent works [40], [41], [42] extend hand-pose analysis along the time axis. By modeling joint trajectories, these methods capture subtle changes in hand position and configuration, enabling finer segmentation of action steps. Typical approaches represent pose sequences as spatio-temporal graphs [41], [42] or feed them into transformer-based encoders [9], [43] for recognition. Nevertheless, most of these methods still rely on plain trajectories and do not explicitly encode fine-grained temporal cues such as motion direction or cumulative rotation. An exception is the work of De Smedt et al. [13], who introduced HoHD (Histogram of Hand Directions) and HoWR (Histogram of Wrist Rotations) to summarize spatial displacements and wrist-rotation patterns. Yet, because these features are built by frequency histograms, they inevitably lose detailed bidirectional information, leaving room for more precise rotation-aware representations.

III. DATASET CURATION

We provide a hand action video dataset with various annotations, including hand action, objects, and hand type per frame. Our dataset includes 622,737 frames with RGB images with 5 subjects and has a large number of sequences and frames compared with the other datasets, as shown in Table 1. Each subject interacts with various objects and tools in each video. The example annotation samples of our dataset are shown in Figure 2. Detailed statistics and visualization of our dataset are provided in Appendix 10 and 11.

A. COMPARISON TO PRIOR DATASETS

The proposed dataset builds upon our previous work [38], which consists of 790 sequences and over 400K frames of tool-use hand activities captured via synchronized thermal and RGB-D cameras with hand segmentation annotations. While the prior dataset primarily focused on hand segmentation performance and modality fusion (e.g., LWIR + RGB-D), our newly proposed dataset significantly expands its scope to enable richer semantic understanding of hand-object interactions to recognize the hand activity. Specifically, we introduce (1) fine-grained hand type annotations tailored to functional taxonomy, (2) additional action-pair labels capturing bidirectional or semantically opposed interactions, and (3) an expanded object and tool categories. These enhancements make our dataset well-suited for both low-level perception and high-level activity recognition in egocentric, real-world manufacturing scenarios.

B. HAND TYPE ANNOTATION

Figure 3 shows an extended version of the hand type taxonomy used in our dataset. Building upon the 31-class functional hand type taxonomy from prior work [17], we expand it to 38 categories by introducing seven additional hand types tailored to the complex motion dynamics observed in real-world manufacturing scenes. For the original 31 categories, we employed the pretrained hand type classifier from [17] to perform pseudo labeling on our video sequences. Furthermore, the newly added hand types were manually annotated to better capture interaction-specific motion patterns involving tools and objects.

Hand types are first categorized into Isolated or Interaction forms, depending on whether the hand is in contact with an object. Within the Grasp category, we adopt four subtypes to describe how an object is held. Notably,

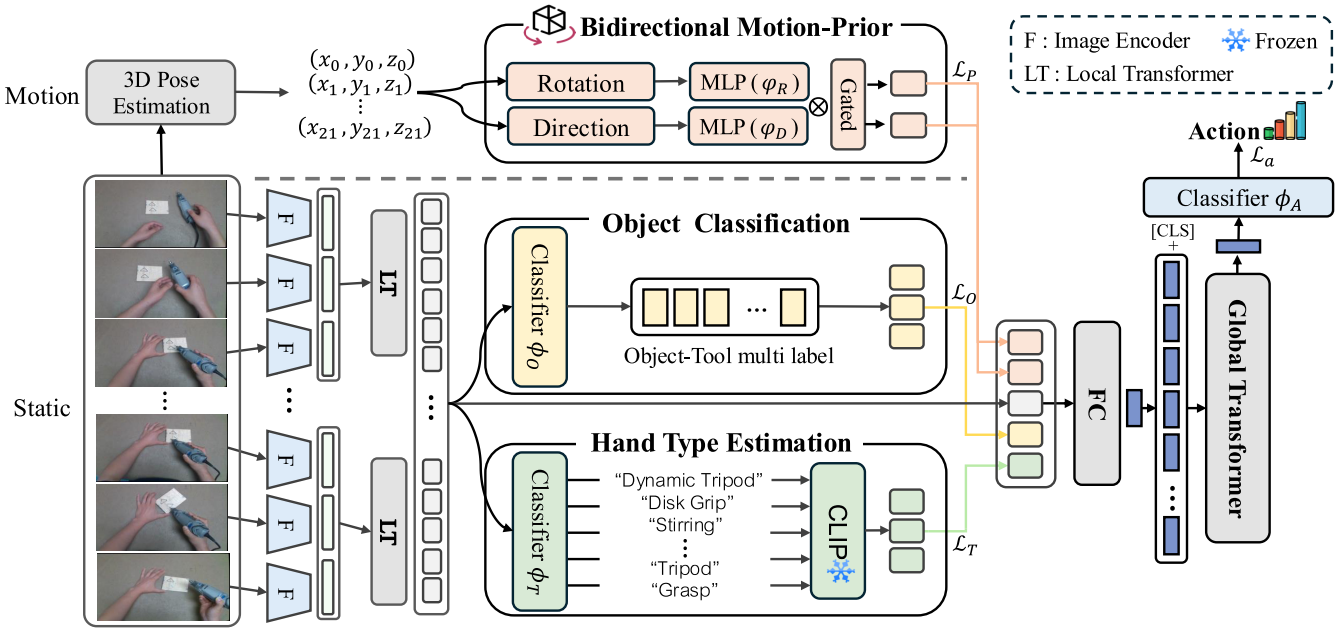

FIGURE 4. **Overall pipeline of the hand action recognition framework using our dataset, Ego-Bi.RGB frames are used for object classification and hand type estimation to capture semantic features. In parallel, 3D hand poses are estimated, and the dominant hand is identified by a simple rule-based method before entering the bidirectional motion-prior module (BMP). Finally, semantic and motion features are fused within a global transformer for action prediction.**

we add the new *Palmar* type under the Circular group to better represent palm-based contact. Then we reorganize the Control category into four refined systems—Touch, Force, Rotate, and Straight—to more accurately describe motion intent. In particular, the *Rotate and Straight* subtypes are expanded with four and two newly introduced types, respectively, to better support the recognition of bidirectional actions in our dataset. These additions capture fine-grained motion cues like rotational dynamics, palm-based contacts, and direction-specific gestures, which are essential for recognizing bidirectional manipulations. The seven newly added hand types do not have counterparts in existing datasets such as H2O or FPHA because those datasets mainly contain low-complexity, unidirectional manipulation tasks that lack torque-dependent or direction-sensitive variations. Our hand type annotations instead describe functional interaction semantics—capturing rotational dynamics, torque polarity changes, and direction-specific control states—which cannot be inferred from pose-only labels. Accordingly, the extended 38-category taxonomy serves as a semantically consistent refinement of the original 31-type scheme, tailored to represent the manipulation states that arise in real-world bidirectional tasks.

C. TEMPORAL ACTIONS AND OBJECTS ANNOTATION

As shown in Table 2, our data consists of a total of 19 action labels, which exist as follows: *assemble, carving, clamp, clampontable, cutting, drill, file, hammer, measure, pry, sanding, saw, screw, stapling, unassemble, unclamp, unclampontable, unscrew, and polishing*. Among them are four pairs of bidirectional classes in which only the hand's movement is opposed under similar conditions of the background, the tool used, and the object. For instance, tightening and loosening constitute a bidirectional pair, differing exclusively in motion dynamics while the tool, object, and background remain constant. Furthermore, our dataset consists of 34 and 29 object and tool categories that the hand interacts with, respectively, and there are things that are meaningfully similar between the object and the tool due to the special situation in which the hand interacts with the tool. This overlap reflects the inherent complexity of hand–object interactions, where similar object–tool combinations can correspond to different actions depending on the hand's temporal dynamics. We manually curated labels for objects, tools, and action classes by visually inspecting each sequence, ensuring semantic accuracy and consistency across the dataset. The details of the object

TABLE 2. **Action class labels used in our dataset. Bidirectional action classes are highlighted. Object and tool class labels are provided in the appendix.**

Action	↔	Action
1: assemble	✓	15: unassemble
3: clamp	✓	16: unclamp
4: clampontable	✓	17: unclampontable
13: screw	✓	18: unscrew
2: carving	✗	5: cutting
6: drill	✗	7: file
8: hammer	✗	9: measure
10: pry	✗	11: sanding
12: saw	✗	14: stapling
19: polishing	✗	

and tool categories are provided in Appendix 10 and Appendix 11.

IV. METHOD

A. OVERVIEW

Figure 4 illustrates the overall pipeline of our egocentric hand action recognition framework using our dataset. The input video clip $\mathcal{S} = \{I_S^{(i)} \in \mathbb{R}^{3\times H\times W} \mid i = 1,\ldots,T\}$ which consists of T frames, we built on HTT [14] core design which utilizes the local transformer block and the global temporal transformer to aggregate the predicted features over $\mathcal{S}$ for final action recognition. Our model consists of an object-tools multi-label classification module and a hand type estimation module based on RGB images, as well as a novel bidirectional motion-prior module we propose. Each component outputs the predicted features and then is fused into a global action transformer, which feeds forward to predict the final action label. Our proposed architecture improves the recognition performance not only for bidirectional action classes included in our dataset, but also generalizes to other real-world datasets with complex motion patterns, without requiring additional supervised learning based on pose ground truth.

We first divide the entire video clip $\mathcal{S}$ into $n_t = \lceil T/t \rceil$ numbers of temporal segments, denoted as $\text{seg}_t(\mathcal{S}) = (\bar{\mathcal{S}}_1,\ldots,\bar{\mathcal{S}}_{n_t})$. Each segment is processed into local transformer block, which takes as input a sequence of per-frame feature vectors F_I extracted by the image encoder F (ResNet [44]) and outputs the corresponding temporal representation $g_s(I)$.

B. HAND TYPE ESTIMATION

To estimate the hand type, we use the temporal representation $g_s(I)$ extracted from the input sequence as input to a hand type classification branch. Given ground-truth annotations of hand type from the preceding data labeling process, the target probability is represented as a one-hot vector w_t. We also define the predicted class as a one-hot vector $\hat{w}_t$ by taking the argmax over the predicted probability distribution, i.e., $\hat{w}_t[r] = \mathbb{I}\{r = \arg\max_j \phi_T(t)[j]\}$, where $\mathbb{I}\{\cdot\}$ denotes the indicator function. To incorporate rich semantic information, we map each hand type class into a descriptive vector using a pretrained CLIP [45] model. The predicted hand type is represented as a softmax probability vector $\phi_T(t)$ over n_T categories, which is supervised using the cross-entropy loss:

$$\mathcal{L}_{\text{T}} = -\sum_{r=1}^{n_T} w_t[r]\cdot\log\phi_T(t)[r], \tag{1}$$

where n_T denotes the set of all possible hand type category classes, $w_t[r]$ is the one-hot encoded ground-truth vector, $\phi_T(t)[r]$ represents the predicted probability for hand type class r obtained via the softmax output of the classifier ϕ_T.

C. OBJECT MULTI-LABEL CLASSIFICATION

To recognize the objects and tools present in each frame, we apply an object classification branch that also takes the temporal representation $g_s(I)$ as input. Since multiple objects and tools may appear simultaneously, the classifier ϕ_O is designed to output a multi-label prediction vector $\hat{\mathbf{y}} \in [0,1]^{n_C}$, where each element is computed as $\hat{y}_r = \sigma(\phi_O(t)[r])$, and $\sigma(\cdot)$ denotes the sigmoid activation function. The object classifier is supervised using the binary cross-entropy (BCE) loss $\mathcal{L}_O$:

$$\mathcal{L}_O = -\frac{1}{n_C}\sum_{r=1}^{n_C}\left[y_r\cdot\log(\hat{y}_r) + (1-y_r)\cdot\log(1-\hat{y}_r)\right], \tag{2}$$

where n_C denotes the number of object/tool categories. For each class r, the ground-truth label $y_r \in \{0,1\}$ indicates the presence of the object in the current frame, and $\hat{y}_r \in [0,1]$ is the predicted confidence score from the sigmoid output of the classifier ϕ_O.

D. BIDIRECTIONAL MOTION-PRIOR MODULE

Unlike previous methods [10], [14] which directly utilize ground-truth 3D pose annotations as input or the supervised module, our model instead leverages 3D keypoints estimated by the 3D pose estimation model [46]. Our dataset includes a diverse set of bidirectional action classes, which can be broadly categorized into two types: (1) *tightening-type* actions, and (2) *loosening-type* actions. Recognizing these actions is particularly challenging since both directions share nearly identical visual context—objects, and backgrounds often appear indistinguishable—making temporal motion the only reliable cue for classification. To explicitly capture these subtle yet crucial motion patterns, we design a *Bidirectional Motion-Prior module* that incorporates two complementary motion cues. The first component, detailed in Section IV-D1, utilizes the signed cumulative rotation of grasp-related keypoints in 3D space, effectively capturing fine-grained rotational dynamics during manipulation. In addition, Section IV-D2 describes how we compute directional changes in the 2D wrist trajectory over time, which reflect the direction of the hand in a more global sense. To the best of our knowledge, no prior work has directly tackled the recognition of bidirectional action classes by explicitly modeling directional motion cues in complex hand-object interaction scenarios. Our framework sets a new direction for motion-aware action recognition in visually ambiguous, real-world hand-object interaction environments.

1) ROTATION PRIOR

Let $\mathbf{P} \in \mathbb{R}^{2\times 21\times 3}$ denote the full set of estimated 3D keypoints for both hands, where the first dimension indexes the left and right hand respectively, each with 21 joints in 3D space. Assuming that the tool is primarily manipulated using one hand, we designate this as the *dominant hand* based on a simple rule: for each frame, we select the hand showing the larger motion change across keypoints as the dominant one. Once the dominant hand is selected, we retain only its 21 keypoints for subsequent computation. Given that grasping small tools such as screwdrivers or hexbolts

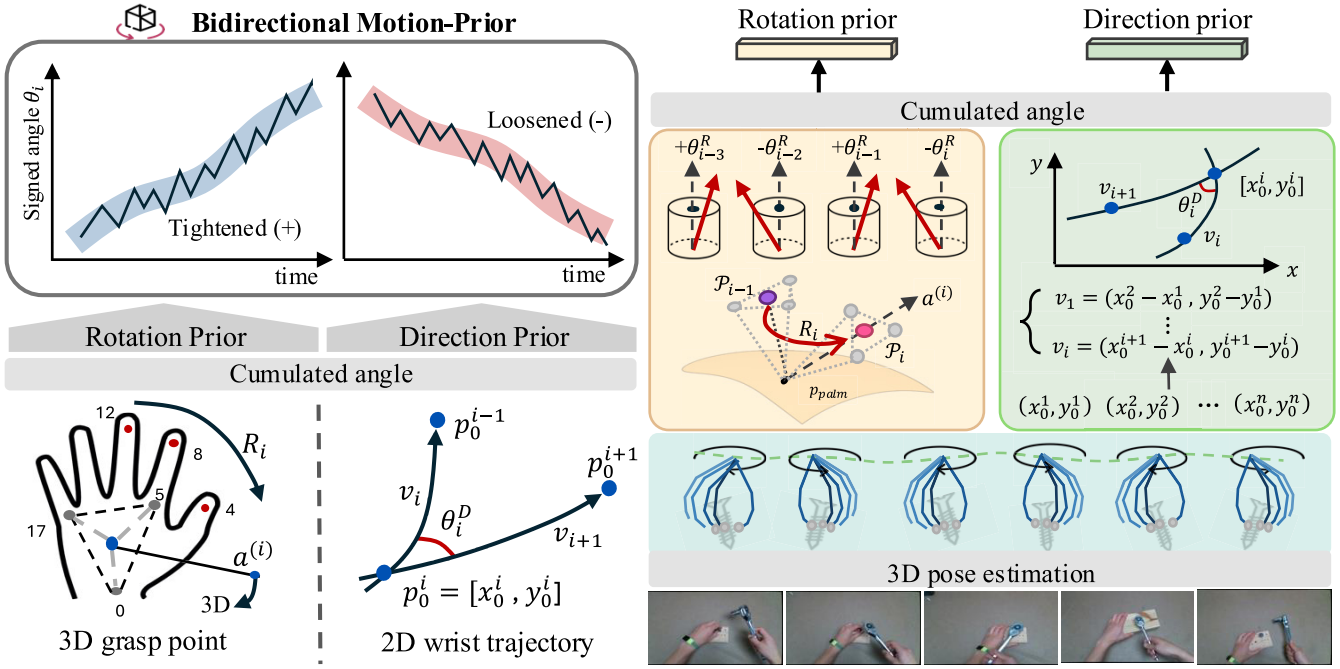

FIGURE 5. **Illustration of our proposed BMP for hand action recognition.We compute the signed rotation of 3D finger grasp contact points and the 2D wrist movement angles over time, based on the estimated 3D hand pose, to generate rotation and direction priors. These priors are accumulated using sliding motion buffers and used to distinguish fine-grained bidirectional actions.**

typically relies on a limited subset of fingers rather than the entire hand, we compute the centroid of their fingertip joints from the dominant hand to define the grasp contact point $\mathbf{p}_{\text{grasp}}$:

$$\mathbf{p}_{\text{grasp}} = \frac{1}{|\mathcal{F}|} \sum_{j \in \mathcal{F}} \mathbf{p}_j, \tag{3}$$

where $\mathbf{p}_i \in \mathbb{R}^3$ denotes the 3D coordinate of keypoint i from the dominant hand as estimated by the hand pose model [46], $\mathcal{F}$ denotes the set of selected grasp points. In our experiments, we set $|\mathcal{F}| = 3$ and chose the fingertip joints of the thumb ($\mathbf{p}_4$), index ($\mathbf{p}_8$), and middle finger ($\mathbf{p}_{12}$).

In addition to the grasp point, we approximate the palm center $\mathbf{p}_{\text{palm}}$ to capture the global position of the hand. Specifically, we compute the centroid of a selected set of base joints $\mathcal{P}$:

$$\mathbf{p}_{\text{palm}} = \frac{1}{|\mathcal{P}|} \sum_{j \in \mathcal{P}} \mathbf{p}_j, \tag{4}$$

where $\mathcal{P}$ denotes the set of selected palm-related keypoints. For the palm center, we used $|\mathcal{P}| = 3$ for base joints, namely the wrist ($\mathbf{p}_0$), index base joint ($\mathbf{p}_5$), and pinky base joint ($\mathbf{p}_{17}$). Using these two centroid points, we define the grasp axis vector $a^{(i)} \in \mathbb{R}^3$ at each frame i as

$$\mathbf{a}^{(i)} = \mathbf{p}^{(i)}_{\text{grasp}} - \mathbf{p}^{(i)}_{\text{palm}}. \tag{5}$$

This grasp axis vector serves as a proxy for the interaction axis of the object being manipulated, even without access to the object's explicit pose ground-truth. To capture the frame-to-frame grasp points rotation, we extract three fingertip points $\mathbf{P}_i = \{\mathbf{p}^{(i)}_4, \mathbf{p}^{(i)}_8, \mathbf{p}^{(i)}_{12}\}$ and compute the best-fit rotation matrix $\mathbf{R}_i$ between consecutive frames using the Kabsch algorithm [47]. A detailed description of the Kabsch-based rotation estimation algorithm is provided in Appendix 1. A 3D rotation matrix $\mathbf{R}_i \in \text{SO}(3)$, where

$$\text{SO}(3) = \left\{ \mathbf{R} \in \mathbb{R}^{3\times 3} \;\middle|\; \mathbf{R}^\top \mathbf{R} = I,\ \det(\mathbf{R}) = 1 \right\}, \tag{6}$$

can also be expressed in an axis–angle format, which represents the rotation as an angle around a specific axis. Using Rodrigues' formula [48], we convert the rotation matrix $\mathbf{R}_i$ into this format by first computing the rotation angle α from its trace $\text{Tr}(\mathbf{R}_i)$, i.e., the sum of its diagonal elements. The corresponding rotation axis $\mathbf{u}$ is then derived from the skew-symmetric part of $\mathbf{R}_i - \mathbf{R}_i^\top$. This yields a compact rotation vector $\mathbf{r}_i = \alpha \cdot \mathbf{u}$ that encodes the same rotation as $\mathbf{R}_i$ in a more interpretable form.

$$\alpha = \cos^{-1}\left(\frac{\text{Tr}(\mathbf{R}_i)-1}{2}\right), \quad \mathbf{u} = \frac{1}{2\sin\alpha}\begin{bmatrix} R_{32} - R_{23} \\ R_{13} - R_{31} \\ R_{21} - R_{12} \end{bmatrix},$$
$$\mathbf{r}_i = \alpha\, \mathbf{u}. \tag{7}$$

Here, $\alpha \in \mathbb{R}$ is the rotation angle, $\mathbf{u} \in \mathbb{R}^3$ is the unit rotation axis, and $\mathbf{r}_i \in \mathbb{R}^3$ is the compact rotation vector. We define the signed scalar rotation angle for frame i as

$$\theta_i^R = \text{sign}(\mathbf{r}_i \cdot \mathbf{a}^{(i)}) \cdot \|\mathbf{r}_i\|_2, \tag{8}$$

where $\mathbf{a}^{(i)}$ is the predefined reference axis at frame i, $\|\cdot\|_2$ is the Euclidean norm, and $\text{sign}(\cdot)$ indicates the alignment of $\mathbf{r}_i$ with the reference axis (positive when aligned, negative when anti-aligned). This value represents the instantaneous

signed magnitude of grasp motion, serving as a key feature for capturing fine-grained rotational patterns. Then, the cumulative rotation angle up to frame i is obtained by summing the signed scalar rotation angles θ_i^R:

$$\Theta_i = \sum_{k=1}^{T} \theta_k^R \cdot \mathbb{I}(k \le i). \tag{9}$$

Due to the inherently oscillating nature of back-and-forth hand motions, the cumulative rotation angle can become unstable or fluctuate over time. Rather than simply smoothing out noise, our goal is to generate a consistent directional prior that remains stable even across reversed but semantically related actions. To this end, we apply a moving average filter to stabilize the rotation signal and produce the final accumulated rotation angle, which is used as our rotation-prior motion buffer:

$$\hat{\Theta}_i = \frac{1}{\mathcal{K}} \sum_{k=-\lfloor \mathcal{K}/2 \rfloor}^{\lfloor \mathcal{K}/2 \rfloor} \Theta_{i+k}, \tag{10}$$

where $\mathcal{K}$ is the smoothing window size of the filter, $\lfloor \cdot \rfloor$ is the floor operation, and $\hat{\Theta}_i$ is the smoothed rotation feature. This formulation provides a rotation-aware prior based solely on predicted hand poses, without requiring any explicit 6D object annotations, and significantly improves the recognition of bidirectional hand actions in our pipeline.

2) DIRECTION PRIOR

In many hand-object interaction scenarios, the key factor distinguishing an action pair lies not in appearance, but in the directional trajectory pattern of the wrist over time. To incorporate prior knowledge of wrist orientation and movement direction, we design a dedicated module that encodes orientation cues (e.g., clockwise or counterclockwise) from the wrist's temporal trajectory, enabling more reliable discrimination of bidirectional actions.

Given the wrist 2D keypoints from the pose estimation results, the displacement vector $\mathbf{v}_i$ is computed based on the wrist coordinates $\mathbf{p}_0^{(i)} = (x_0^i, y_0^i) \in \mathbb{R}^2$ from two consecutive frames i and $i+1$:

$$\mathbf{v}_i = \left[x_0^{i+1} - x_0^i,\ y_0^{i+1} - y_0^i \right]. \tag{11}$$

To quantify the change in wrist movement direction between successive frames, we compute the signed angle θ_i^D between consecutive displacement vectors as

$$\theta_i^D = \text{atan2}(\|\mathbf{v}_i \times \mathbf{v}_{i+1}\|_2,\ \mathbf{v}_i \cdot \mathbf{v}_{i+1}), \tag{12}$$

where the dot product $\mathbf{v}_i \cdot \mathbf{v}_{i+1}$ captures the directional similarity, while the magnitude of the cross product $\|\mathbf{v}_i \times \mathbf{v}_{i+1}\|_2$ determines the orientation of the rotation. Here, $\text{atan2}(y, x)$ returns the signed angle whose tangent is y/x, while accounting for the correct quadrant, and $\|\cdot\|_2$ denotes the Euclidean (L2) norm. This formulation preserves both the magnitude and sign of the angular change, enabling accurate representation of directional changes in motion. Finally, following the same procedure as in Eq. 9, we accumulate the signed angles θ_i^D over time to compute the cumulative direction angle up to frame i, which serves as a temporally grounded motion prior for recognizing bidirectional hand actions:

$$\Phi_i = \sum_{k=1}^{T} \theta_k^D \cdot \mathbb{I}(k \le i). \tag{13}$$

The cumulative signed angle θ_i^D provides a directional pattern of wrist movement: it tends to accumulate positive values under clockwise wrist motion and negative values under counterclockwise motion. To suppress noise in wrist motion trajectories and reinforce temporal cues, we apply the same moving average smoothing procedure used in the rotation prior computation (Eq. 10) to the directional trend. Formally, this process is defined as

$$\hat{\Phi}_i = \frac{1}{\mathcal{K}} \sum_{k=-\lfloor \mathcal{K}/2 \rfloor}^{\lfloor \mathcal{K}/2 \rfloor} \Phi_{i+k}, \tag{14}$$

where $\mathcal{K}$ denotes the temporal window size. This yields a more stable directionality signal over time, enabling more reliable discrimination between action pairs with opposite wrist movement patterns. Unlike directly inputting raw wrist coordinates, this temporally grounded directionality prior enhances the model's ability to distinguish between action pairs.

Figure 5 illustrates how our priors enable the model to distinguish between opposing actions in practice. For instance, in *loosened-type* actions like unscrew or unassemble, the cumulative rotation and direction trends move in the opposite direction compared to their *tightened-type* counterparts. These trends are visualized using sample action segments, demonstrating how our priors encode motion asymmetry that is otherwise lost in static appearance features.

3) GATED MECHANISM

To leverage the smoothed motion priors defined above, we first construct motion buffers that represent temporal sequences of rotation and direction signals:

$$B_R = [\hat{\Theta}_1, \ldots, \hat{\Theta}_T] \in \mathbb{R}^T, \quad B_D = [\hat{\Phi}_1, \ldots, \hat{\Phi}_T] \in \mathbb{R}^T. \tag{15}$$

Each motion buffer derived from the bidirectional motion-prior modules—grasp-point rotation and wrist direction—is passed through an MLP [49] network, ϕ_R and ϕ_D respectively, to extract temporal embeddings:

$$\mathbf{h}_R = \phi_R(B_R), \quad \mathbf{h}_D = \phi_D(B_D). \tag{16}$$

To adaptively weigh these motion cues based on the action context, we compute a global summary feature $\mathbf{h}_{\text{ctx}}$ by averaging temporal features from RGB, object, and hand type embeddings. This summary feature represents a high-level context of the current action sequence and is used as input to two separate gating modules, each consisting of a linear

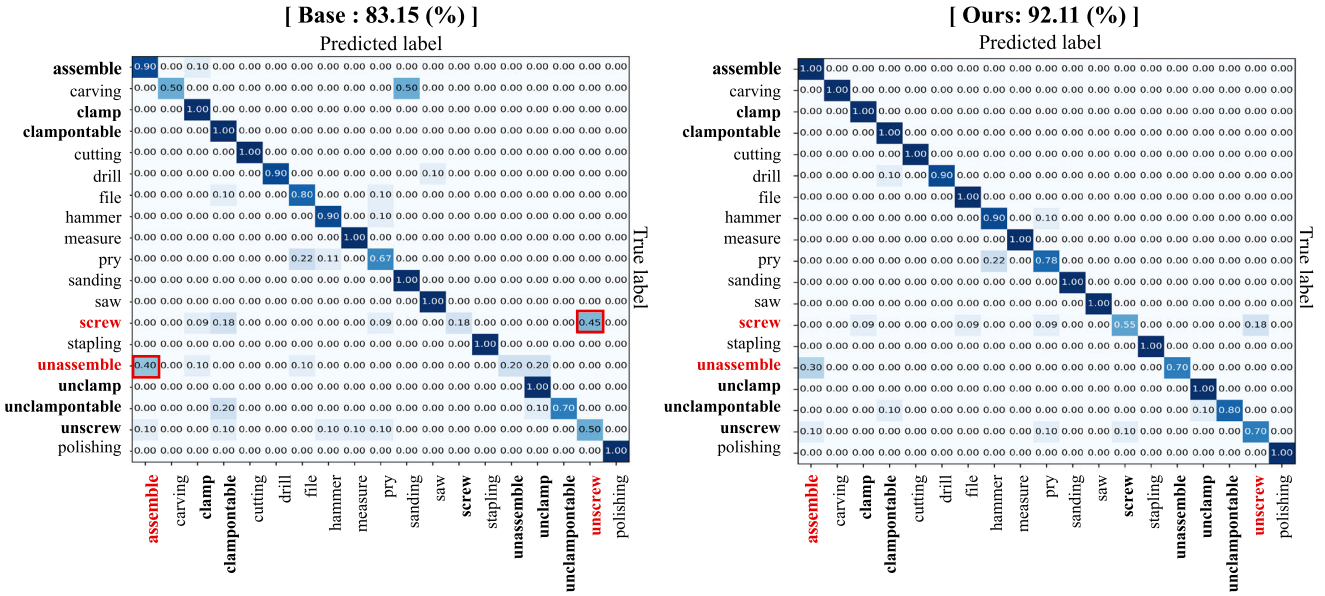

FIGURE 6. Confusion matrices comparing the baseline and our model with hand type and motion prior modules. Bidirectional action pairs are shown in bold, and those frequently confused by the model are highlighted in red.

layer followed by a sigmoid activation. The gating modules predict soft weights $\alpha_R = \sigma(W_R \mathbf{h}_{\text{ctx}} + b_R)$ and $\alpha_D = \sigma(W_D \mathbf{h}_{\text{ctx}} + b_D)$, which modulate the rotation and direction embeddings as $\tilde{\mathbf{h}}_R = \alpha_R \odot \mathbf{h}_R$ and $\tilde{\mathbf{h}}_D = \alpha_D \odot \mathbf{h}_D$, where $\odot$ denotes element-wise multiplication. This mechanism allows the model to adaptively weight the contribution of each motion prior, giving emphasis to rotational or directional dynamics in actions that involve movement, while reducing their influence for static or non-dynamic. As a result, the gated rotation and direction features are concatenated into a joint embedding, represented as $\mathbf{z} = [\tilde{\mathbf{h}}_R \,\|\, \tilde{\mathbf{h}}_D]$, where $\|$ denotes vector concatenation.

Finally, this joint representation $\mathbf{z}$ is optimized with weighted supervision using the Proxy-NCA loss [50], denoted as L_P:

$$\mathcal{L}_P = \frac{1}{N} \sum_{i=1}^{N} \left(\|\mathbf{z}_i - \mathbf{p}_{y_i}\|_2^2 + \log \sum_{r \in \mathcal{N}_a \setminus \{y_i\}} \exp\left(-\|\mathbf{z}_i - \mathbf{p}_r\|_2^2 \right) \right), \tag{17}$$

where N is the number of samples in a batch, y_i is the ground-truth action class of sample i, $\mathcal{N}_a$ is the set of all action classes, and $\mathbf{p}_r$ denotes the learnable proxy vector for class r. Each proxy $\mathbf{p}_r$ is optimized jointly with the network parameters during training, encouraging each sample embedding $\mathbf{z}_i$ to be pulled closer to its corresponding proxy $\mathbf{p}_{y_i}$ while being pushed away from the proxies of all other classes $r \neq y_i$.

E. FUSED TO GLOBAL TRANSFORMER

For each frame I belonging to a temporal segment $\bar{\mathcal{S}} \subset \mathcal{S}$, where $\bar{\mathcal{S}} \in \text{seg}_t(\mathcal{S})$, the representation $h(I)$ is computed by fusing the gated motion-prior embeddings, object prediction feature, hand type prediction feature, and the image feature $g_{\bar{\mathcal{S}}}(I)$ extracted by the local transformer:

$$h(I) = FC_1\left(g_{\bar{\mathcal{S}}}(I) \| FC_2(O_I) \| FC_3(H_I) \| \phi_R(\hat{\Theta}_I) \| \phi_D(\hat{\Phi}_I) \right), \tag{18}$$

where $FC_1(\cdot)$, $FC_2(\cdot)$, and $FC_3(\cdot)$ are fully connected layers that map their inputs into d-dimensional embeddings. Here, I denotes the 2D input image feature, O_I the object classification feature, H_I the hand type feature, and $\hat{\Theta}_I$ and $\hat{\Phi}_I$ the rotation and direction priors from the BMP. The operator $\|$ indicates vector concatenation.

F. LEARNING OBJECTIVES

After that, the fused feature $h(I)$ is then combined with a [CLS] token and fed into a global transformer with positional encoding, and outputs the predicted action label for final action recognition. The final loss in the end-to-end classification of action recognition is defined as $\mathcal{L}_{\text{a}} = -a_i \log \hat{a}_i$, where a_i is the ground-truth label for the sample i-th and $\hat{a}_i$ is the corresponding predicted action probability passed through the action classifier ϕ_A, represented as $\hat{a}_i = [p(a_1|\mathcal{S}), \ldots, p(a_{n_a}|\mathcal{S})] = \text{softmax}(\phi_A(\alpha_{\text{out}}))$. Hence, given input video clip $\mathcal{S}$, the final total loss is represented as

$$\mathcal{L}_{\text{tot}} = \lambda_{\text{a}} \mathcal{L}_{\text{a}} + \frac{1}{T} \sum_{\bar{\mathcal{S}} \in \text{seg}_t(\mathcal{S})} \sum_{I \in \bar{\mathcal{S}}} \left(\lambda_{\text{P}} \mathcal{L}_{\text{P}}(I) + \lambda_{\text{O}} \mathcal{L}_{\text{O}}(I) + \lambda_{\text{T}} \mathcal{L}_{\text{T}}(I) \right), \tag{19}$$

where λ_a, λ_P, λ_O, and λ_T are hyperparameters that control the relative contribution of each loss term.

TABLE 3. Ablation study on input feature combinations for the global transformer. Each column indicates whether a specific module is included: Object feature (Obj), hand type estimation (Typ), rotation prior (Rotation), direction prior (Direction).

Method	Input Features for Global Transformer					Test Accuracy (%)		
	RGB	Obj	Typ	Rotation	Direction	FPHA [15]	H2O [11]	Ego-Bi
Base	✓	✓	✗	✗	✗	91.65	85.12	83.15
Base + T	✓	✓	✓	✗	✗	94.61	86.78	86.75
Base + R	✓	✓	✗	✓	✗	91.83	87.70	86.02
Base + D	✓	✓	✗	✗	✓	92.52	89.34	88.17
Base + T + R	✓	✓	✓	✓	✗	94.79	90.16	88.53
Base + T + D	✓	✓	✓	✗	✓	95.01	90.98	89.96
Base + T + R + D (Ours)	✓	✓	✓	✓	✓	**95.13**	**91.80**	**92.11**

TABLE 4. Comparison of hand action recognition performance when using the original 31-category hand type taxonomy versus our extended 38-category taxonomy.

Hand Type Taxonomy	Overall Accuracy (%)	Bidirectional Accuracy (%)
31 Categories [17]	89.96%	68.15%
38 Categories (Ours)	**92.11%**	**79.39%**

TABLE 5. Ablation on 3D hand pose estimation models. Performance remains consistent across Wilor3D, MediaPipe, and WildHand inputs, demonstrating the robustness to variations in pose estimation quality.

Methods	Wilor3D [46]	Mediapipe [51]	WildHand [52]	Mean
Accuracy (%)	92.11	90.83	91.62	**91.52**

V. EXPERIMENT

A. IMPLEMENTATION DETAILS

1) TRAINING SETUP

Following [11], we use a subject-wise split, holding out one subject for testing while training on the other four. Each video clip $\mathcal{S}$ is divided into temporal segments $\text{seg}_t(\mathcal{S})$, and local action cues are encoded within a hierarchical temporal transformer structure, where each token corresponds to a segment of consecutive frames. Based on the configuration of HTT [14], we set the maximum input sequence lengths to $T = 128$ for the global action transformer and $t = 16$ for the local pose transformer. All input images are resized to $H = 270$, $W = 480$, and models are trained for 45 epochs using the Adam optimizer with an initial learning rate of 3e-05 and a batch size of 2. The implementation is based on PyTorch, and experiments are conducted on an NVIDIA RTX 3090 GPU. We applied a step decay schedule, reducing the learning rate by a factor of 0.5 every 15 epochs. The complete training objective is described in Section IV-F, and the weighting coefficients are set as follows: $\lambda_a = 0.3$, while λ_P, λ_O, and λ_T are all set to 1.0.

2) EVALUATION PROTOCOL

We evaluate our model on the hand action recognition task, with particular emphasis on bidirectional action pairs. As a performance metric, we report the Top-1 classification accuracy on the test set, which serves as a standard metric for action recognition benchmarks. Predictions are made at the video level by aggregating temporally segmented features, and the results are evaluated against the ground-truth action labels. First, we compare our dataset Ego-Bi with other benchmarks such as FPHA [38] and H2O [11], as shown in Table 3. These other datasets are collected in indoor settings and provide ground-truth labels for action, object category, and hand pose. In contrast, our dataset does not include ground-truth hand pose annotations, therefore, we report ablation study results using a model built on the same HTT [14] baseline, but with the pose-supervised module removed. We also compare with the performance of prior work [17] using hand type labels based on the original taxonomy before our extension. This allows us to assess the impact of our newly annotated, fine-grained hand type taxonomy on recognition performance. For all datasets, we adopt a modality-consistent evaluation setting where only RGB inputs are used, even for FPHA and H2O that provide RGB-D data. This follows the standard protocol used in strong performance baselines such as HTT and HandFormer, and ensures fair comparison with our RGB-only Ego-Bi pipeline without introducing modality imbalance.

To further disentangle the contributions of each proposed component, we also compare several model variants: (1) an RGB-based model with spatially guided object priors, (2) a model leveraging our extended hand type taxonomy via concatenated prediction features, and (3) variants enhanced with the proposed rotation and wrist direction priors. As shown in Table 6, we quantitatively evaluate the effectiveness of our BMP module by measuring the separability of bidirectional action classes using the motion priors generated from each module. Specifically, we compute AUC and d' scores based on the priors produced by the rotation prior and direction prior modules, in order to assess how well the proposed motion buffers distinguish between opposite-direction action pairs across the entire dataset. In addition, the corresponding ROC curves are presented in Figure 7, providing a visual comparison of the discriminative ability of each prior. We also evaluate the performance of various baseline models on our dataset in comparison with

TABLE 6. Quantitative separability of our motion priors. Higher AUC and d' indicate stronger discrimination between opposite-rotation and opposite-direction action pairs.

Ours (w/ Motion priors)	Metrics	
	AUC ↑	d′ ↑
Rotation prior	**0.822**	**1.084**
Direction prior	**0.840**	**1.453**

other benchmarks, as shown in Table 7, to illustrate the relative difficulty of our dataset and highlight where models perform well or struggle. To account for the distribution of action categories in our dataset, we further evaluate model robustness using complementary metrics such as macro-level precision and F1-score, as presented in Table 8. These metrics are particularly useful in our dataset, where the distribution of action classes is inherently unbalanced, and allow for a more reliable interpretation of model performance across underrepresented or visually similar classes. Furthermore, to evaluate the generalization ability of our model across different subjects, we additionally report leave-one-subject-out (LOSO) cross validation results in Table 9. Specifically, we perform subject-wise k-fold validation by rotating the held-out subject across all five participants. In addition to quantitative evaluation, we also analyze the model's ability to separate ambiguous action pairs via confusion matrices as shown in Figure 6.

B. IMPACT OF HAND TYPE AS A SEMANTIC PRIOR

As shown in Table 3, applying the original hand type taxonomy built on the previous research [17] to FPHA [15] and H2O [11] datasets led to performance improvements of approximately 2.96% and 1.66%, respectively. However, on our dataset Ego-Bi, which incorporates a more fine-grained and extended hand type semantic cue, the performance gain was even more significant (+3.5%). Notably, our dataset involves various bidirectional hand–object interactions, so the larger performance boost on our dataset highlights the added value of our extended taxonomy in scenarios where nuanced hand semantics are crucial. This result suggests that the richer hand type annotations in our dataset serve as more informative semantic cues, thereby contributing more effectively to the final hand action recognition. It demonstrates the importance of detailed morphological and functional hand labeling in complex interaction scenarios.

Furthermore, Table 4 reports a controlled comparison between the previous 31-taxonomy [17] and our extended 38-category scheme in our Ego-Bi dataset. For experiment, the seven additional categories were collapsed into their most semantically compatible parent classes within the original taxonomy, allowing the two systems to be evaluated under identical label granularity. Bidirectional Accuracy is computed over inverse action pairs to evaluate the taxonomy's ability to disambiguate direction-sensitive motions. Our extended taxonomy yields higher accuracy and substantially reduces errors on the bidirectional actions, illustrating that the added torque- and direction-sensitive types provide essential semantic distinctions.

C. EFFECT OF MOTION PRIORS ON BIDIRECTIONAL PAIRS

To investigate the contribution of our bidirectional motion priors in fine-grained hand action recognition, we conduct an ablation study by selectively enabling each prior component. Table 3 reports the action recognition accuracy across these variant model settings. Injecting either the rotation or direction prior leads to performance gains of +2.87% and +5.02% over the baseline, respectively, demonstrating that both priors provide complementary cues that enhance recognition accuracy. Similar to the improvement obtained by adding the semantic cue of hand type, these results confirm that motion priors also contribute positively to the final performance. Our full model, which combines all semantic inputs with gated rotation and direction priors, achieves the highest performance (92.11%). These results support that motion priors are not only complementary to semantic tokens like object and hand type, but also essential for precise modeling of bidirectional actions. We further evaluate the quantitative separability of our motion priors specifically on bidirectional action pairs in our dataset, as presented in Table 6. For the rotation prior, we obtain an AUC of 0.822 and a d' score of 1.084, while the direction prior achieves an AUC of 0.840 and a d' score of 1.453. These results indicate that both priors provide meaningful discriminative power for distinguishing opposite-bidirectional action pairs.

1) ROBUSTNESS TO POSE ESTIMATION NOISE

It is important to note that the effectiveness of our motion priors does not rely on highly accurate hand pose estimation. Unlike approaches that directly consume absolute 3D keypoint coordinates—where estimation errors caused by occlusion, motion blur, or rapid hand deformation would propagate to the model—our priors are derived from *relative* temporal changes accumulated over motion segments. This design makes the priors inherently invariant to global shifts, scale changes, or moderate jitter in predicted keypoints. As long as the estimated hand trajectory preserves the coarse directional trend of the movement, our rotation and direction priors remain stable. This property explains why our framework maintains nearly identical action recognition performance even when replacing Wilor3D [46] with lower-accuracy pose estimators such as MediaPipe [51] or WildHands [52] (see Table 5). Thus, the motion prior acts as a denoised and noise-tolerant representation of hand dynamics rather than a direct reflection of raw pose accuracy.

D. COMPREHENSIVE EVALUATION ON EGO-BI DATASET

As reported in Table 7, we benchmark various hand action recognition baselines, including C2D [21], SlowFast [23], HandFormer [10], and HTT, on our dataset, with the HTT framework achieving the highest performance. In contrast, reproducing other models such as HandFormer [10] leads to

TABLE 7. Comparison of action recognition performance (%) across baselines on FPHA, H2O, Assembly101, and our Ego-Bi datasets ('-' indicates not reported; HTT* denotes the HTT variant trained without pose ground-truth supervision).

Dataset	C2D [21]	SlowFast [23]	HandFormer [10]	HTT* [14]
FPHA [15]	-	-	-	91.65
H2O [11]	70.66	77.69	93.39	85.12
Assembly101 [36]	-	-	41.06	-
Ego-Bi (Ours)	**58.64**	**67.14**	**80.31**	**83.15**

TABLE 8. Extended evaluation results using both micro- and macro-averaged metrics to account for class imbalance. The consistent gains across micro-accuracy, macro-precision, and Macro-F1 indicate that our method improves performance for both frequent and infrequent action classes.

Method	Input Features for Global Transformer					Evaluation Metric (%)		
	RGB	Obj	Typ	Rotation	Direction	Micro-Accuracy	Macro-Precision	Macro-F1 score
Base	✓	✓	✗	✗	✗	83.15	82.15	80.13
Base + T	✓	✓	✓	✗	✗	86.75	82.68	80.21
Base + R	✓	✓	✗	✓	✗	86.02	82.26	81.08
Base + D	✓	✓	✗	✗	✓	88.17	83.15	83.14
Base + T + R	✓	✓	✓	✓	✗	88.53	86.02	83.48
Base + T + D	✓	✓	✓	✗	✓	89.96	86.93	86.26
Base + T + R + D (Ours)	✓	✓	✓	✓	✓	**92.11**	**88.17**	**87.55**

TABLE 9. k-fold leave-one-subject-out (LOSO) validation results. Each fold uses a different subject for testing.

Fold	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean
Test Subject	Subject A	Subject B	Subject C	Subject D	Subject E	–
Accuracy (%)	92.11	91.62	92.62	91.81	87.56	**91.14**

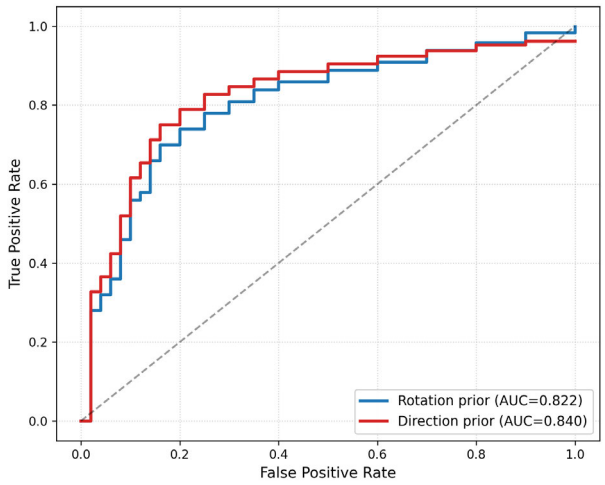

FIGURE 7. ROC curves comparing the classification performance of the rotation prior and direction prior. The curves show the true positive rate (TPR) versus false positive rate (FPR) for each prior, with their corresponding AUC values indicating the discriminative ability between bidirectional action pairs.

lower scores, largely due to the fact that the original research implementations rely on ground-truth pose inputs, which creates a performance gap when compared with the state-of-the-art results reported on datasets like H2O [11](93.39%).

Despite this limitation, our approach—by integrating semantic type priors and the proposed motion priors—achieves robust recognition without access to pose ground truth, reaching competitive performance levels that are comparable to, and in some cases approach, those obtained with pose-supervised baselines.

Furthermore, we compare the confusion matrices of the baseline model, which excludes both hand type semantic cues and motion priors, with those of our full model. As shown in Figure 6 (left), the baseline model frequently confuses bidirectional pairs such as screw vs. unscrew, assemble vs. unassemble, highlighted in red bounding boxes, reflecting its limited ability to capture directional motion. In contrast, our model (right) not only significantly reduces cross-pair confusion but also achieves a much clearer separation between opposite actions, underscoring the crucial role of motion priors and semantic feature fusion in learning discriminative temporal motion dynamics.

Lastly, in Table 9, the results obtained from Leave-One-Subject-Out (LOSO) k-fold validation—where each subject is held out in turn for testing while the remaining participants are used for training—provide a more rigorous evaluation of model stability and subject-agnostic generalization across different individuals in real-world egocentric scenarios. Taken together, the experiments demonstrate that our proposed method substantially improves robustness against noisy variations and enhances generalization under cross-subject settings, establishing Ego-Bi as a challenging yet valuable benchmark characterized by fine-grained hand-type taxonomy and diverse bidirectional action classes.

VI. CONCLUSION

In this work, we presented a new benchmark dataset *Ego-Bi*, for egocentric hand action recognition that provides

TABLE 10. Object class labels used in our dataset.

Object	Object
0: acrylicboard	1: bigBolt
2: blackHexBolt	3: bolt
4: countersunk	5: curvedMetalPlate
6: hexBolt	7: hexNut
8: metalBlock	9: metalPlate
10: metalPlateBolt	11: metalRoundFinPlate
12: metalSpring	13: metalboard
14: nail	15: nut
16: plasticBottle	17: pvc
18: roundHeadScrew	19: silverHexBolt
20: silverHugeHexBolt	21: smallWood
22: smallWoodBlock	23: socketCapScrew
24: spacer	25: styrofoam
26: uBolt	27: vHandleScrewLong
28: vHandleScrewShort	29: washer
30: woodBlock	31: woodChunk
32: woodPlate	33: woodboard

TABLE 11. Tool class labels used in our dataset.

Tool	Tool
34: allenKey	35: ballPeenHammer
36: channelLockPliers	37: clawHammer
38: drill	39: flatFile
40: foundFile	41: grinder
42: hacksaw	43: hand
44: largeBlackClamp	45: lifter
46: longNosePliers	47: measuringStick
48: measuringTape	49: miniHacksaw
50: powerSaw	51: roundFile
52: sandpaper	53: screwdriver
54: smallBlackClamp	55: smallGreyClamp
56: socketDriver	57: socketWrench
58: squareFile	59: tacker
60: vise	61: woodenClamp
62: wrench	

fine-grained hand type annotations and includes challenging bidirectional action pairs. Ego-Bi comprises 1,223 sequences and 622,737 frames, with annotations covering 38 hand types and 19 actions, including bidirectional pairs that challenge appearance-based recognition. To address the inherent ambiguities in recognizing visually similar but semantically opposite actions, we introduced the Bidirectional Motion Prior module *(BMP)*. By leveraging predicted 3D hand poses, our model extracts meaningful temporal motion cues—such as rotation direction and cumulative angular change—without requiring additional manual supervision. Experimental results demonstrate that our method significantly improves recognition accuracy on bidirectional actions, even when relying solely on predicted hand poses. These findings highlight the importance of temporal dynamics and directional priors in egocentric activity understanding. We believe Ego-Bi and BMP together provide a strong foundation for advancing research on fine-grained hand–object interaction, and expect future work to build upon them for more robust, real-world applications in manufacturing and assistive robotics.

VII. LIMITATIONS AND FUTURE WORK

First, our dataset does not provide explicit pose annotations. While our motion prior method achieves strong performance using predicted poses, further improvements could be expected with access to accurate pose ground truth. Second, we do not utilize additional modalities such as depth, thermal, or inertial signals, which could enhance the robustness of motion understanding. We leave these directions—leveraging high-fidelity annotations and multimodal inputs—as promising avenues for future work.

TABLE 12. Duration-robustness analysis using different temporal window sizes for pose-token and action-token generation. Rows correspond to pose token window size, and columns correspond to action token window size.

Window size	64	128 (default)	256	Mean
8	91.74	92.11	91.62	91.82
16 (default)	92.03	92.11	91.95	92.03
32	91.88	92.04	91.73	91.88
Mean	91.88	92.09	91.77	**91.91**

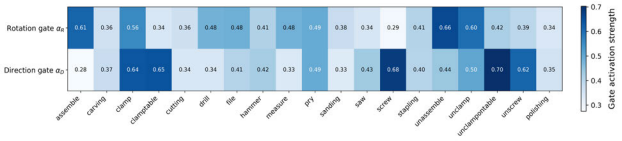

FIGURE 8. Heatmap visualization of the learned adaptive gating weights.

Algorithm 1 Rotation Estimation via Kabsch Algorithm

Input: Fingertip positions at frame i:
$\mathbf{P}_i = \{\mathbf{p}_4^{(i)}, \mathbf{p}_8^{(i)}, \mathbf{p}_{12}^{(i)}\}$
Fingertip positions at frame $t+1$:
$\mathbf{P}_{i+1} = \{\mathbf{p}_4^{(i+1)}, \mathbf{p}_8^{(i+1)}, \mathbf{p}_{12}^{(i+1)}\}$
Output: Estimated rotation matrix $\mathbf{R} \in SO(3)$
Center both point sets by subtracting their centroids;
$\bar{\mathbf{P}}_i \leftarrow \mathbf{P}_i - \text{mean}(\mathbf{P}_i)$;
$\bar{\mathbf{P}}_{i+1} \leftarrow \mathbf{P}_{i+1} - \text{mean}(\mathbf{P}_{i+1})$;
Compute covariance matrix: $\mathbf{H} \leftarrow \bar{\mathbf{P}}_i^\top \bar{\mathbf{P}}_{i+1}$;
Compute SVD: $\mathbf{H} = \mathbf{U}\Sigma\mathbf{V}^\top$;
Compute rotation matrix: $\mathbf{R} \leftarrow \mathbf{V}\mathbf{U}^\top$;
if $\det(\mathbf{R}) < 0$ **then**
  Flip last column of $\mathbf{V}$: $\mathbf{V}[:, 3] \leftarrow -\mathbf{V}[:, 3]$;
  Recompute: $\mathbf{R}_i \leftarrow \mathbf{V}\mathbf{U}^\top$;
return $\mathbf{R}_i$

APPENDIX A
DATASET STATISTICS

Appendix A presents detailed statistics of our curated dataset. Our dataset contains 622,737 frames with 1,223 sequences from 5 subjects, and following previous work [11], we split it into training and test sets by leaving out all samples from one subject.

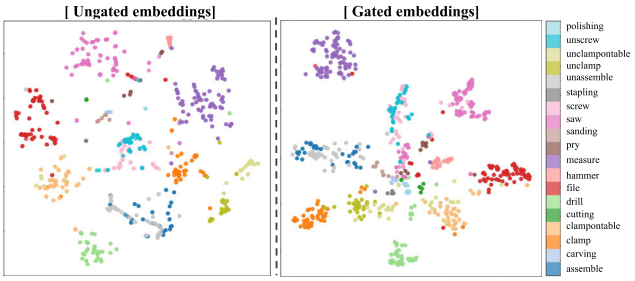

FIGURE 9. t-SNE visualization of integrated gated motion embeddings. The proposed gating mechanism yields clearer inter-class separation, especially for bidirectional action pairs.

As shown in Figure 10, we visualize our dataset statistics which contains (1) per frame and sequence in action classes, (2) object and tool categories distribution, and (3) hand types per action distribution. Bidirectional action classes often exhibit highly similar hand type distributions. These sequences naturally share semantic cues due to their similar functional hand types.

APPENDIX B
DATASET CURATION DETAILS

Our dataset builds upon the hand segmentation dataset introduced by Kim and Chi [38], with newly added annotations and an extended set of frames and sequences to create a more comprehensive benchmark on hand action recognition.

To extend the original dataset, we collected additional sequences using the same Intel D435i camera setup, capturing synchronized RGB and depth images. Building upon the data acquisition setup introduced in previous work [38], where RGB-D images were captured using helmet-mounted cameras enclosed in 3D printed cases, we construct an extended dataset with additional sequences and annotations. The input resolution of all experiments in the main paper is resized to 480×270 as used in previous research [14], [17].

APPENDIX C
IMPLEMENTATION DETAILS OF THE BIDIRECTIONAL MOTION-PRIOR (BMP) MODULE

This section provides a reproducibility-oriented description of the Bidirectional Motion-Prior (BMP) module, summarizing the processing pipeline from raw pose trajectories to the final motion-prior embeddings integrated into the action-recognition backbone.

As defined in Sections IV-D1 and IV-D2, the BMP module produces two temporally accumulated scalar trajectories: the rotation prior $\{\hat{\Theta}_i\}$ and the direction prior $\{\hat{\Phi}_i\}$. For implementation, both trajectories are first smoothed using a small moving-average filter (window size 3) and uniformly resampled to a fixed length of 512, forming two one-dimensional motion buffers.

These fixed-length buffers are then embedded by a shared lightweight MLP with four layers (512–256–128–128) and LeakyReLU activations in the hidden layers, yielding two 128-dimensional motion-prior embeddings for rotation and direction, respectively. To adaptively modulate the contribution of each prior given the current action context, we compute a global context vector by temporally averaging the Transformer token outputs, and pass this through a linear layer followed by a sigmoid nonlinearity to obtain 128-dimensional gate vectors. The gates are applied elementwise to the motion-prior embeddings, producing gated rotation and direction features.

At each timestamp, these gated motion-prior features are concatenated with the pose, object, and hand-type token features, resulting in fused representation, which is projected to the Transformer dimension D via a linear layer and fed into the global action Transformer. In addition, the two motion-prior embeddings are concatenated into a 256-dimensional vector and regularized with a ProxyNCA loss, using one proxy per action class and a loss weight of $\lambda = 0.3$, to enhance the separability of direction-sensitive action pairs.

APPENDIX D
DATASET ANNOTATION EXAMPLES

We provide more annotation samples of our proposed datasets on various action classes in Figure 11. Figure 11 presents additional examples of our dataset annotations across various action classes. Each row illustrates a specific action instance, showing both left (L) and right (R) hand types, the associated manipulated objects, and the corresponding grasp type. The visual diversity highlights the richness of our dataset in terms of hand-object interactions, grasp types, and tool usage patterns, which are essential for fine-grained egocentric hand action understanding.

A. OBJECT AND TOOLS CATEGORIES

As shown in Table 2, we provide the action classes included in our dataset. In this section, we present a table showing the object and tool category labels included in the dataset beyond action labels.

APPENDIX E
HAND TYPE TAXONOMY

Figure 15 shows our extended function-based hand type taxonomy. Compared to the original version consisting of 31 types used in previous research [17], we expand the taxonomy to include 38 fine-grained categories. The 7 newly added hand types are primarily designed to better capture rotational dynamics, palmar configurations, and other nuanced motion cues commonly observed in egocentric manipulation tasks. Specifically, we introduce *Dynamic Lateral Pinch, Index Rotation, Middle Rotation, and Full Rotation* under the Rotate category to distinguish repetitive turning motions in different directions. In the Straight motion category, we add *Hammering* and *Dynamic Diameter*, which represent linear force-intensive gestures. To better model palm-surface interactions that are common in tool handling and pressing actions, we incorporate the *Palmer* hand type

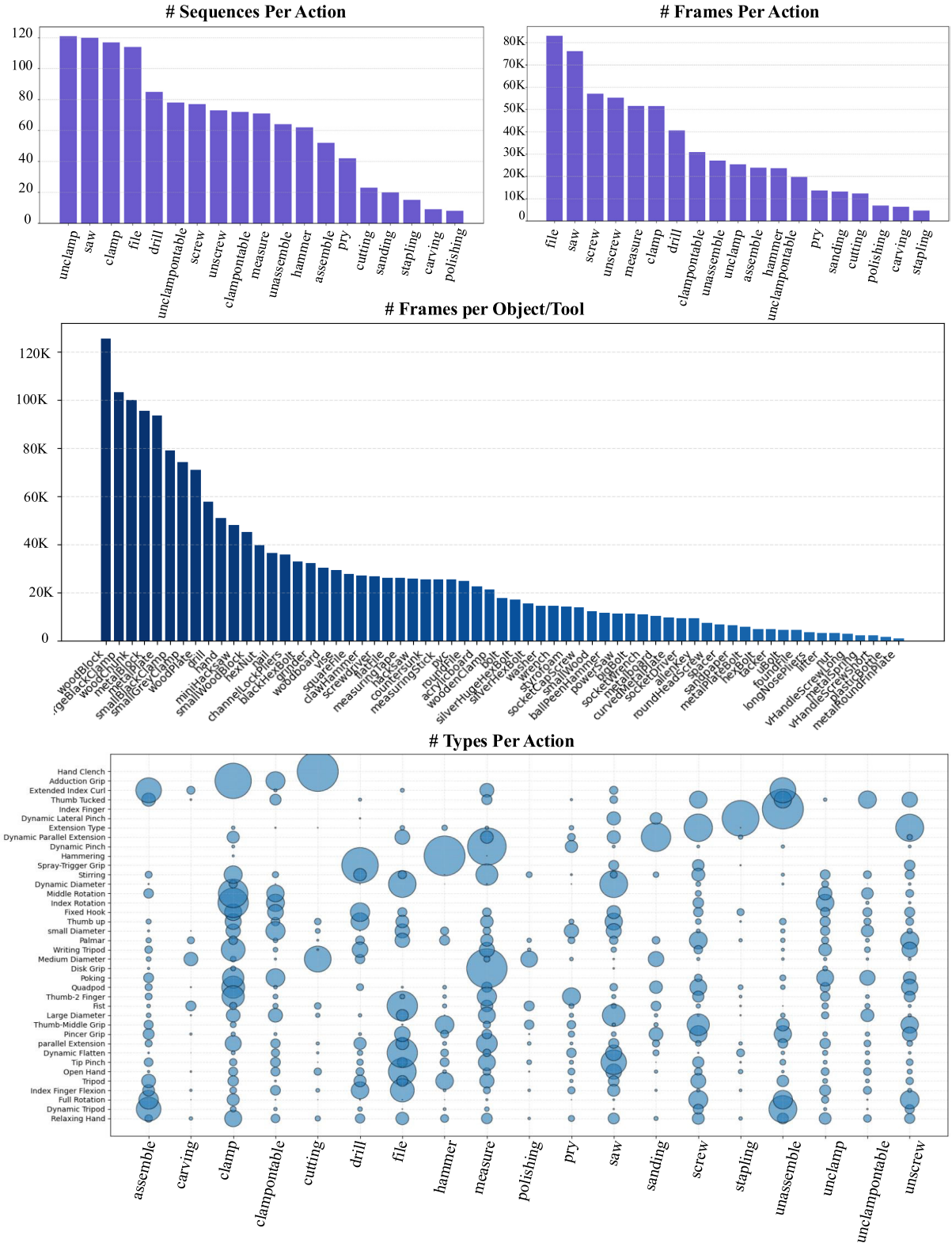

FIGURE 10. **Overview of dataset statistics. This figure presents a visualization of the dataset distribution across action classes, object categories, and hand types.**

under the Grasp group. These additional types provide more expressive cues for recognizing subtle differences in action directionality and intent, particularly for bidirectional actions.

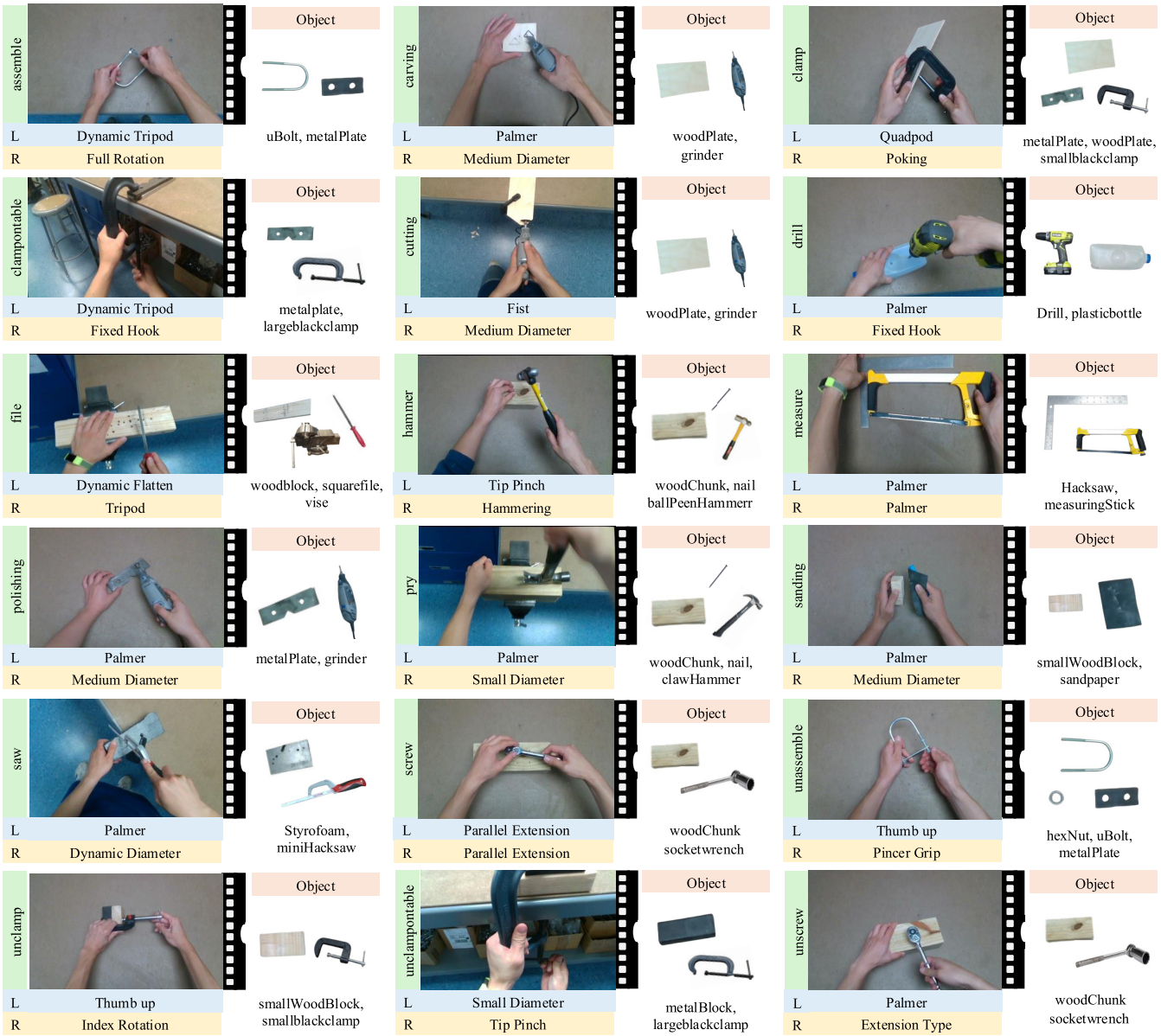

FIGURE 11. Examples of dataset annotations. This figure provides additional samples from our dataset, showing diverse action classes, hand types, and manipulated objects along with their corresponding annotations.

APPENDIX F
DURATION-ROBUSTNESS ANALYSIS

To assess whether our framework is robust to variations in action duration, we conducted a controlled experiment by modifying the temporal window sizes used for local and global transformer in our proposed pipeline. These hyperparameters implicitly govern the level of temporal compression and therefore simulate actions performed at different speeds. We varied the pose-token window size across {8, 16 (default), 32} and the action-token window size across {64, 128 (default), 256}, resulting in nine temporal configurations. As shown in Table 12, recognition accuracy remains highly stable across all settings, with fluctuations within ±0.3%. This indicates that the hierarchical temporal downsampling and window-based tokenization in the HTT-based backbone provide strong duration normalization, and that the proposed semantic–motion fusion pipeline is insensitive to absolute sequence length.

APPENDIX G
KABSCH ALGORITHM

In section IV-D1, we compute the best-fit rotation $\mathbf{R}_t$ between consecutive frames with three grasp points $\mathbf{P}_i = \{\mathbf{p}_4^{(i)}, \mathbf{p}_8^{(i)}, \mathbf{p}_{12}^{(i)}\}$.

A detailed pseudo-algorithm is written in Algorithm 1, which describes the steps for estimating the rotation

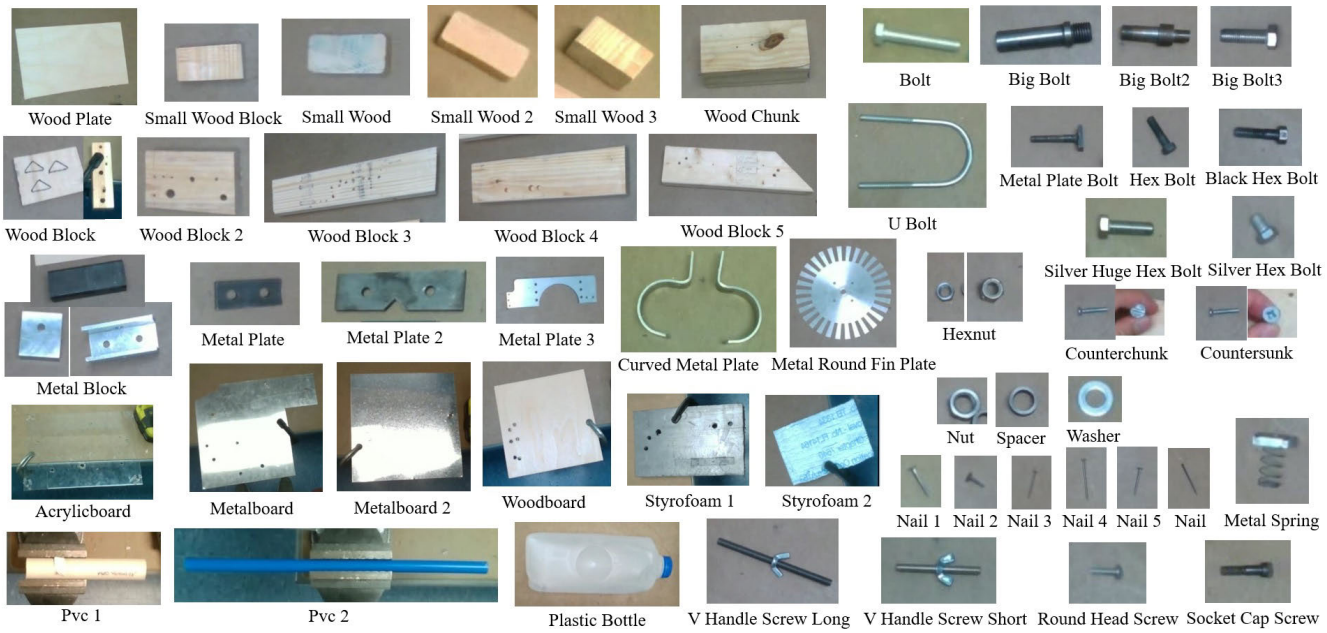

FIGURE 12. Example images of object categories included in our dataset.

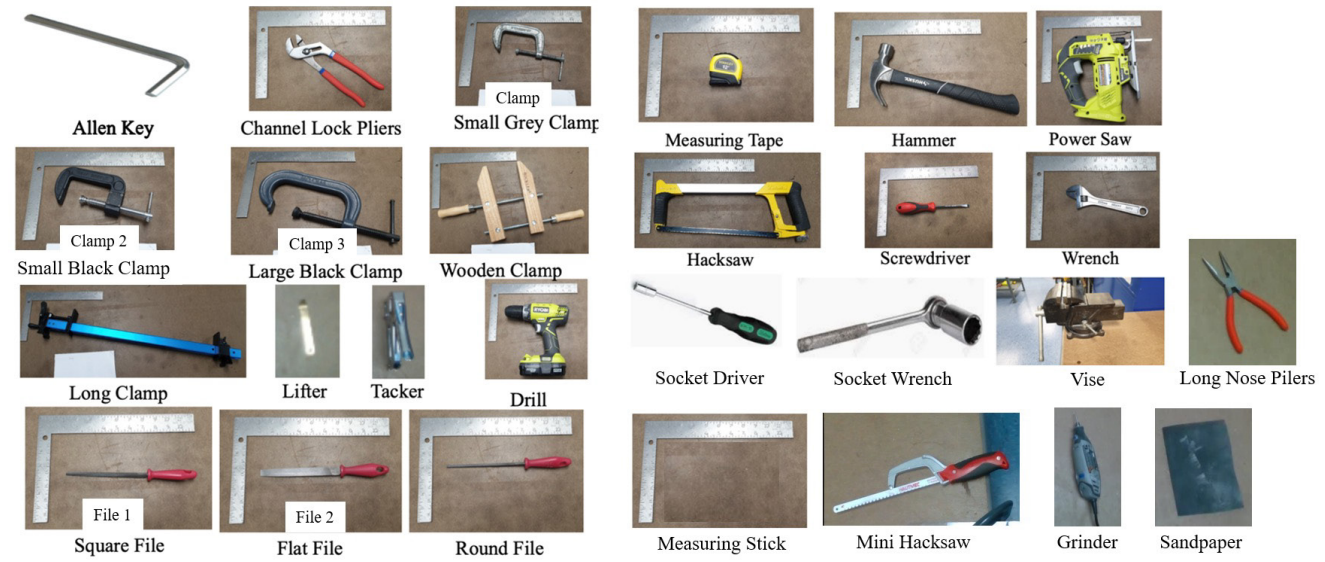

FIGURE 13. Example images of tool categories included in our dataset.

matrix between two consecutive frames using the Kabsch algorithm [47]. This procedure serves as the foundation for constructing our rotation prior, as it transforms local finger joint movements into stable frame-to-frame rotational cues that can be directly leveraged for action recognition.

APPENDIX H
ADAPTIVE GATING MECHANISM AND REPRESENTATION ANALYSIS

To assess whether the proposed gating mechanism meaningfully modulates motion features rather than operating as a uniform scaling factor, we first examine the learned gate activations across action categories. As shown in Figure 8, the rotation and direction gates exhibit distinct activation patterns depending on the manipulation type. Rotation-heavy actions such as *assemble* and *unassemble* preferentially activate the rotation gate, while direction-sensitive actions such as *screw* and *unscrew* show stronger activation in the direction gate. Hybrid actions like *clamp* and *unclamp* engage both gates, suggesting that the module adapts its weighting based on task semantics.

To further validate whether this behavior leads to meaningful representational changes, we compare the embedding space before and after applying the gating module. We extract

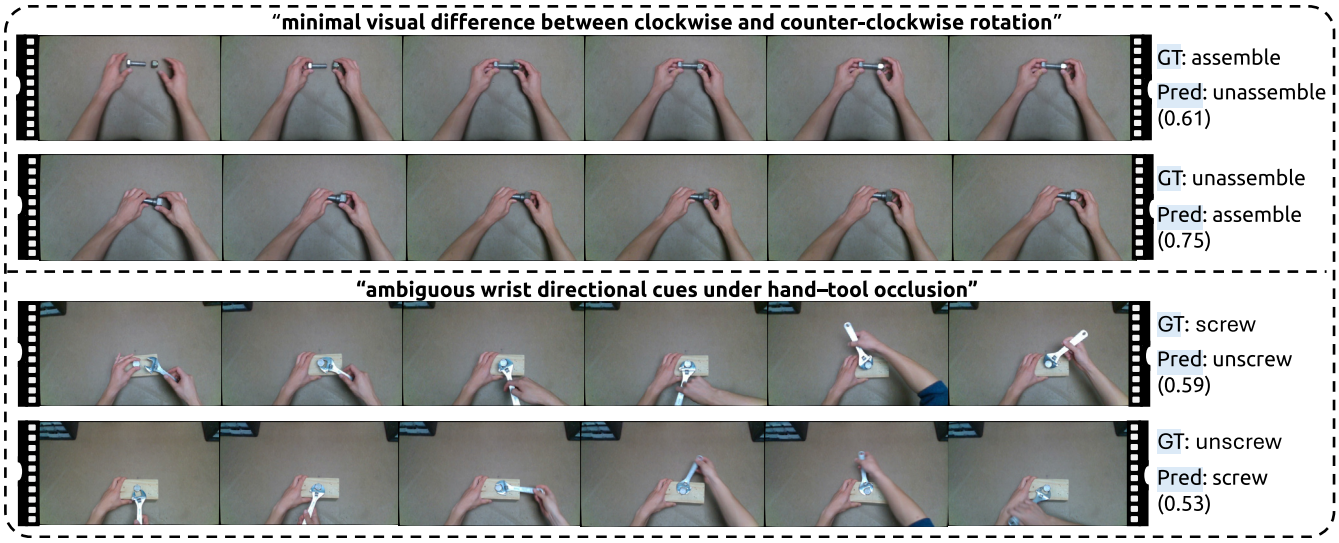

FIGURE 14. Additional qualitative misclassified examples of bidirectional action pairs.

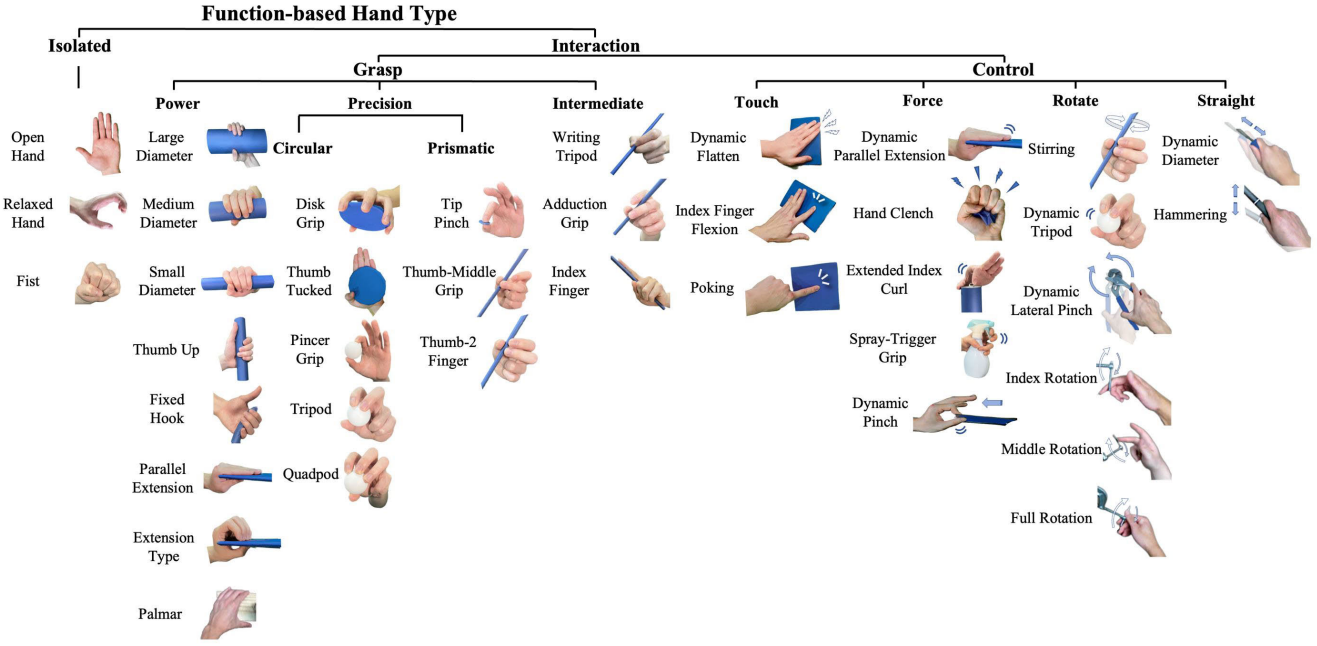

FIGURE 15. Extended hand type taxonomy.

(i) the baseline representation obtained without gating and (ii) the gated joint representation $\mathbf{z} = [\,\tilde{\mathbf{h}}_R \,\|\, \tilde{\mathbf{h}}_D\,]$, which is used as the input to the global transformer. Both representations are projected into a two-dimensional space using PCA followed by t-SNE. As shown in Figure 9, the gated embeddings form more coherent and separable clusters compared to the ungated baseline. In particular, bidirectional action pairs (e.g., *screw* vs. *unscrew*) become more distinguishable, indicating that the gating mechanism enhances discriminability and organizes the latent space in a semantically meaningful way.

APPENDIX I
ADDITIONAL QUALITATIVE ANALYSIS OF MISCLASSIFIED BIDIRECTIONAL ACTIONS

To complement the quantitative evaluation presented in the main manuscript, we include additional qualitative examples of misclassification cases for bidirectional action pairs. As discussed earlier, these actions are inherently challenging due to subtle directional reversals, limited visual cues, and frequent hand–object occlusions under RGB-only observation settings.

Figure 14 illustrates two representative cases in which the proposed method incorrectly classified *screw* as *unscrew* and *assemble* as *unassemble*. Each example includes: (i) the temporal frame sequence, (ii) the predicted probability, and (iii) a short analysis describing factors that contributed to the misclassification. These examples reveal that failure often arises during transitional micro-actions where motion cues are weak or ambiguous, making direction inference difficult. We also observe that noisy hand pose estimation and partial occlusion of the manipulated object occasionally amplify prediction uncertainty. Notably, the probability curves indicate hesitation between candidate classes, confirming that the confusion is not random but occurs in structurally ambiguous motion states. This evidence reinforces the challenges of recognizing direction-sensitive actions in real-world tool interaction environments without explicit 3D pose or object supervision.

APPENDIX J
DATA IMBALANCED PROBLEM

As shown in Figure 10, our dataset exhibits the difference in distribution across action and object classes, indicating the presence of an imbalanced problem.

We observe that certain categories, such as *polishing and carving*, as well as some tool/object classes, contain only a small number of sequences and frames. This imbalance can introduce misleading interpretations of model performance when using standard evaluation metrics such as Top-1 Accuracy. The average accuracy across all classes may not faithfully reflect the model's actual capabilities, especially if high-frequency classes dominate the performance as analyzed in previous research [53]. Therefore, caution is required when interpreting mean accuracy in the presence of such imbalance. Although we adopt Top-1 accuracy for consistency with prior works [7], [11], [14], [19] in action recognition, we acknowledge its limitations in reflecting performance on rare classes. To complement this, we further report per-class accuracy and confusion matrices, offering a more comprehensive view of model behavior under class imbalance. This analysis demonstrates that our approach maintains consistent performance across frequent and rare classes, highlighting its robustness beyond aggregate accuracy metrics.

REFERENCES

[1] L. Ge, Y. Cai, J. Weng, and J. Yuan, "Hand PointNet: 3D hand pose estimation using point sets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8417–8426.
[2] C. Zimmermann and T. Brox, "Learning to estimate 3D hand pose from single RGB images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4913–4921.
[3] M. Oberweger and V. Lepetit, "DeepPrior++: Improving fast and accurate 3D hand pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 585–594.
[4] G. Li, H. Tang, Y. Sun, J. Kong, G. Jiang, D. Jiang, B. Tao, S. Xu, and H. Liu, "Hand gesture recognition based on convolution neural network," *Cluster Comput.*, vol. 22, no. S2, pp. 2719–2729, Mar. 2019.
[5] E. Fertl, E. Castillo, G. Stettinger, M. P. Cuéllar, and D. P. Morales, "Hand gesture recognition on edge devices: Sensor technologies, algorithms, and processing hardware," *Sensors*, vol. 25, no. 6, p. 1687, Mar. 2025.
[6] O. Kopuklu, A. Gunduz, N. Kose, and G. Rigoll, "Real-time hand gesture detection and classification using convolutional neural networks," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–8.
[7] W. Mucha, M. Wray, and M. Kampel, "SHARP: Segmentation of hands and arms by range using pseudo-depth for enhanced egocentric 3D hand pose estimation and action recognition," in *Proc. Int. Conf. Pattern Recognit.*, 2024, pp. 178–193.
[8] W. Mucha and M. Kampel, "In my perspective, in my hands: Accurate egocentric 2D hand pose and action recognition," in *Proc. IEEE 18th Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2024, pp. 1–9.
[9] H. Cho, C. Kim, J. Kim, S. Lee, E. Ismayilzada, and S. Baek, "Transformer-based unified recognition of two hands manipulating objects," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, Jun. 2023, pp. 4769–4778.
[10] M. S. Shamil, D. Chatterjee, F. Sener, S. Ma, and A. Yao, "On the utility of 3D hand poses for action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 436–454.
[11] T. Kwon, B. Tekin, J. Stühmer, F. Bogo, and M. Pollefeys, "H2O: Two hands manipulating objects for first person interaction recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 10118–10128.
[12] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3192–3199.
[13] Q. D. Smedt, H. Wannous, and J. Vandeborre, "Skeleton-based dynamic hand gesture recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2016, pp. 1206–1214.
[14] Y. Wen, H. Pan, L. Yang, J. Pan, T. Komura, and W. Wang, "Hierarchical temporal transformer for 3D hand pose estimation and action recognition from egocentric RGB videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 21243–21253.
[15] G. Garcia-Hernando, S. Yuan, S. Baek, and T.-K. Kim, "First-person hand action benchmark with RGB-D videos and 3D hand pose annotations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 409–419.
[16] C. Zimmermann, D. Ceylan, S. Yan, B. Russell, M. Argus, and T. Brox, "FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 813–822.
[17] W. Roh, S. H. Lee, W. J. Ryoo, J. Lee, G. Oh, S. Hwang, H.-G. Chi, and S. Kim, "Functional hand type prior for 3D hand pose estimation and action recognition from egocentric view monocular videos," in *Proc. BMVC*, 2023, p. 193.
[18] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2847–2854.
[19] Y. Li, Z. Ye, and J. M. Rehg, "Delving into egocentric actions," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Jun. 2015, pp. 287–295.
[20] S. Bambach, S. Lee, D. Crandall, and Y. Chen, "Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1949–1957.
[21] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.
[22] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6299–6308.
[23] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 6202–6211.
[24] S. Sudhakaran, S. Escalera, and O. Lanz, "LSTA: Long short-term attention for egocentric action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9946–9955.
[25] S. Sudhakaran, S. Escalera, and O. Lanz, "Learning to recognize actions on objects in egocentric video with attention dictionaries," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 6674–6687, Jun. 2023.
[26] V.-D. Le, V.-N. Hoang, T.-T. Nguyen, V.-H. Le, T.-H. Tran, H. Vu, and T.-L. Le, "Hand activity recognition from automatic estimated egocentric skeletons combining slow fast and graphical neural networks," *Vietnam J. Comput. Sci.*, vol. 10, no. 1, pp. 75–100, Feb. 2023.
[27] X. Li, Y. Hou, P. Wang, Z. Gao, M. Xu, and W. Li, "Trear: Transformer-based RGB-D egocentric action recognition," *IEEE Trans. Cognit. Develop. Syst.*, vol. 14, no. 1, pp. 246–252, Mar. 2022.

[28] J. R. Napier, "The prehensile movements of the human hand," *J. Bone Joint Surgery. Brit.*, vol. 38, no. 4, pp. 902–913, Nov. 1956.
[29] M. R. Cutkosky, "On grasp choice, grasp models, and the design of hands for manufacturing tasks," *IEEE Trans. Robot. Autom.*, vol. 5, no. 3, pp. 269–279, Jun. 1989.
[30] I. M. Bullock, R. R. Ma, and A. M. Dollar, "A hand-centric classification of human and robot dexterous manipulation," *IEEE Trans. Haptics*, vol. 6, no. 2, pp. 129–144, Apr. 2013.
[31] F. Stival, S. Michieletto, M. Cognolato, E. Pagello, H. Müller, and M. Atzori, "A quantitative taxonomy of human hand grasps," *J. NeuroEngineering Rehabil.*, vol. 16, no. 1, p. 28, Dec. 2019.
[32] T. Feix, J. Romero, H.-B. Schmiedmayer, A. M. Dollar, and D. Kragic, "The GRASP taxonomy of human grasp types," *IEEE Trans. Hum.-Mach. Syst.*, vol. 46, no. 1, pp. 66–77, Feb. 2016.
[33] W. Cho, J. Lee, M. Yi, M. Kim, T. Woo, D. Kim, T. Ha, H. Lee, J. H. Ryu, W. Woo, and T. K. Kim, "Dense hand-object (ho) graspnet with full grasping taxonomy and dynamics," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2024, pp. 284–303.
[34] M. Cai, K. Kitani, and Y. Sato, "Understanding hand-object manipulation with grasp types and object attributes," in *Proc. Robotics, Sci. Syst.*, 2016, pp. 1–10.
[35] A. Fathi, X. Ren, and J. M. Rehg, "Learning to recognize objects in egocentric activities," in *Proc. CVPR*, Jun. 2011, pp. 3281–3288.
[36] F. Sener, D. Chatterjee, D. Shelepov, K. He, D. Singhania, R. Wang, and A. Yao, "Assembly101: A large-scale multi-view video dataset for understanding procedural activities," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 21064–21074.
[37] Y. Li, M. Liu, and J. M. Rehg, "In the eye of the beholder: Gaze and actions in first person video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 6731–6747, Jun. 2023.
[38] S. Kim, H. Chi, X. Hu, A. Vegesana, and K. Ramani, "First-person view hand segmentation of multi-modal hand activity video dataset," in *Proc. BMVC*, 2020.
[39] G. Rogez, J. S. Supancic, and D. Ramanan, "Understanding everyday hands in action from RGB-D images," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3889–3897.
[40] S. Das, S. Sharma, R. Dai, F. Brémond, and M. Thonnat, "VPN: Learning video-pose embedding for activities of daily living," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 72–90.
[41] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 7444–7452.
[42] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 140–149.
[43] F. Yang, D. Li, and G. Wang, "Spatial temporal block transformer network for skeleton-based action recognition," in *Proc. China Autom. Congr. (CAC)*, Nov. 2022, pp. 1259–1264.
[44] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 630–645.
[45] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
[46] R. A. Potamias, J. Zhang, J. Deng, and S. Zafeiriou, "WiLoR: End-to-end 3D hand localization and reconstruction in-the-wild," in *Proc. Comput. Vis. Pattern Recognit. Conf.*, 2025, pp. 12242–12254.
[47] W. Kabsch, "Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants," *J. Appl. Crystallogr.*, vol. 26, no. 6, pp. 795–800, Dec. 1993.
[48] J. S. Dai, "Euler–Rodrigues formula variations, quaternion conjugation and intrinsic connections," *Mechanism Mach. Theory*, vol. 92, pp. 144–152, Oct. 2015.
[49] H. Taud and J. Mas, "Multilayer perceptron (MLP)," in *Geomatic Approaches for Modeling Land Change Scenarios*. Cham, Switzerland: Springer, 2017, pp. 451–455.
[50] E. W. Teh, T. DeVries, and G. W. Taylor, "ProxyNCA++: Revisiting and revitalizing proxy neighborhood component analysis," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 448–464.
[51] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, "MediaPipe hands: On-device real-time hand tracking," 2020, *arXiv:2006.10214*.
[52] A. Prakash, R. Tu, M. Chang, and S. Gupta, "3D hand pose estimation in everyday egocentric images," in *Proc. Eur. Conf. Comput. Vis.*, 2023, pp. 183–202.
[53] L. A. Jeni, J. F. Cohn, and F. D. L. Torre, "Facing imbalanced data-recommendations for the use of performance metrics," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, 2013, pp. 245–251.

JIYOUNG SEO was born in South Korea. She received the B.S. degree in statistics and artificial intelligence from the University of Seoul, South Korea, in February 2025. She is currently pursuing the M.S. degree in artificial intelligence with Korea University. She is a member of the Computer Vision Laboratory, Korea University. She has contributed to the development of large-scale datasets and practical systems that integrate video, audio, and language understanding. Her technical expertise includes deep learning-based vision models, transformer architectures, and multimodal reasoning frameworks. During her undergraduate studies, she gained experience in pattern recognition, 3D graphics, and computer vision projects. Her expertise also spans deep learning frameworks, such as PyTorch and TensorFlow, and the design of large-scale multimodal datasets. Her long-term research goal is to advance multimodal AI systems that can perceive, reason, and act in complex real-world environments. Her research interests include 3D computer vision, physical AI, multimodal learning, retrieval-augmented generation, and action recognition.

DONG IN LEE received the B.S. degree in computer and electronic system engineering from the Hankuk University of Foreign Studies, South Korea, in 2023. He is currently pursuing the Ph.D. degree with the Department of Artificial Intelligence, Korea University, Seoul, South Korea, and a Visiting Researcher with the Convergence Design Laboratory, Purdue University, West Lafayette, IN, USA. His research interests include 3D computer vision, 3D generation and editing, 3D human–object reconstruction, and AI watermarking.

PILHYEON LEE received the B.S. degree in computer science and engineering from Chung-Ang University, Seoul, South Korea, in 2018, and the Ph.D. degree in computer science from Yonsei University, Seoul, in 2023. From 2023 to 2024, he was a Postdoctoral Researcher with Yonsei University. Since 2024, he has been an Assistant Professor with the Department of Artificial Intelligence, Inha University, Incheon, South Korea. His research interests include multimodal learning, video understanding, data-efficient training, generative models, and representation learning.

JIWOO LEE was born in South Korea. He is currently pursuing the B.S. degree in computer science and engineering with Hongik University, South Korea. Since February 2025, he has been an Undergraduate Researcher with the Computer Vision Laboratory, Korea University. During his undergraduate studies, he gained practical experience in computer vision through projects in areas, including low-light image enhancement. His long-term research goal is to develop foundational computer vision technologies that can seamlessly fuse the physical and digital worlds, creating a truly immersive and interactive reality. His research interests include 3D computer vision, AR/VR, and action recognition.

YOUNHEE GIL received the B.S. and M.S. degrees in computer engineering from Pusan National University, South Korea, in 1999 and 2001, respectively. She has been a Principal Researcher with the Content Research Division, Electronics and Telecommunications Research Institute, Daejeon, South Korea. Her research interests include NUI/NUX, VR/AR/MR, and accessibility.

KARTHIK RAMANI (Member, IEEE) received the B.Tech. degree in mechanical engineering from Indian Institute of Technology Madras, in 1985, the M.S. degree in mechanical engineering from Ohio State University, in 1987, and the Ph.D. degree in mechanical engineering from Stanford University, in 1991. He is currently a Donald W. Feddersen Professor with the School of Mechanical Engineering, Purdue University, with courtesy appointments in electrical and computer engineering and the College of Education. He has published in ACM (CHI & UIST), IEEE (CVPR, ICCV, ICRA), ECCV, ICLR, *Scientific Reports*, and ASME JMD. His research interests include collaborative intelligence, human–machine interactions, spatial interfaces, deep shape learning, and manufacturing productivity.

SANGPIL KIM received the B.Sc. degree in computer science from Korea University, South Korea, and the Ph.D. degree in electrical and computer engineering from Purdue University. He is currently an Assistant Professor with the Department of Computer Science and Engineering, Korea University. His current research focuses on the intersection of computer vision and deep learning with an emphasis on applications of multi-modal fusion for developing generative models.