Detecting Precise Hand Touch Moments in Egocentric Video

Huy Anh Nguyen Feras Dayoub Minh Hoai
Australian Institute for Machine Learning, Adelaide University, Australia
{huyanh.nguyen, feras.dayoub, minhhoai.nguyen}@adelaide.edu.au

Abstract

We address the challenging task of detecting the precise moment when hands make contact with objects in egocentric videos. This frame-level detection is crucial for augmented reality, human-computer interaction, assistive technologies, and robot learning applications, where contact onset signals action initiation or completion. Temporally precise detection is particularly challenging due to subtle hand motion variations near contact, frequent occlusions, fine-grained manipulation patterns, and the inherent motion dynamics of first-person perspectives.

To tackle these challenges, we propose a Hand-informed Context Enhanced module (HiCE; pronounced 'high-see') that leverages spatiotemporal features from hand regions and their surrounding context through cross-attention mechanisms, learning to identify potential contact patterns. Our approach is further refined with a grasp-aware loss and soft label that emphasizes hand pose patterns and movement dynamics characteristic of touch events, enabling the model to distinguish between near-contact and actual contact frames. We also introduce TouchMoment, an egocentric dataset containing 4,021 videos and 8,456 annotated contact moments spanning over one million frames. Experiments on TouchMoment show that, under a strict evaluation criterion that counts a prediction as correct only if it falls within a two-frame tolerance of the ground-truth moment, our method achieves substantial gains and outperforms state-of-the-art event-spotting baselines by 16.91% average precision. Code is available at https://github.com/bbvisual/hice.

1. Introduction

This paper studies the problem of detecting the precise moment when a hand makes contact with an object in egocentric video. Touch moments mark critical temporal boundaries in manipulation—indicating when an action begins, transitions, or completes—and thus provide rich cues for understanding human behavior, assessing motor capability, analyzing skilled performance, and transferring dexterous

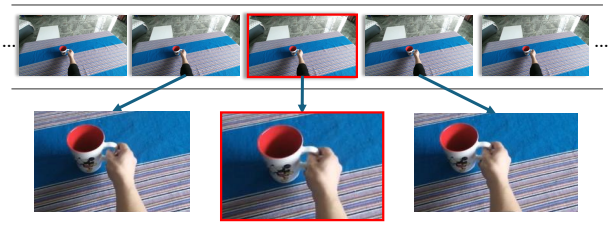

Figure 1. We develop a model to detect the precise touch moment—the exact frame when the hand first makes contact with an object in egocentric video. As illustrated, distinguishing the true touch moment (the frame with a red border) is challenging due to subtle near-touch hand motions, rapid camera movement, and fine-grained grasp changes that make near-touch frames visually similar to the actual moment of contact.

demonstrations to robots. Egocentric video, collected from head-mounted or wearable cameras, is a natural modality for this task because it directly captures hand motions and the user's visual perspective during manipulation. Reliable identification of touch moments in such recordings can therefore support a wide range of applications that depend on fine-grained temporal detection.

In this work, we focus on detecting the first touch moment—the single frame at which contact between the hand and an object is first established. Touch moment is treated as an instantaneous event rather than a temporal segment. When a video contains multiple interactions, our goal is to detect each of these first-contact moments. We consider only intentional touch events that contribute to the manipulation process, excluding incidental brushes or accidental collisions. Formally, given an input video sequence, the task is to identify a sparse set of frames marking these transitions from non-contact to contact across the sequence.

Detecting the exact touch moment in egocentric video, however, is extremely challenging, as illustrated in Fig. 1. Rapid head motion introduces strong egomotion, while approaching hands often create severe occlusions and perspective distortions at close range. Subtle micro-motions immediately before contact, fine-grained pre-touch hand shaping, and motion blur during dynamic activities further obscure the true moment of contact. These factors make the touch

moment difficult to distinguish from near-touch frames or other ambiguous transitional states, requiring models that can reliably extract discriminative hand cues despite noise, occlusion, and continuously shifting viewpoints inherent to first-person video.

Precise moment detection is not new in computer vision, and prior work has addressed onset prediction and fine-grained action spotting across several domains. However, these methods are typically developed under assumptions that differ substantially from egocentric hand–object interaction. For example, techniques for detecting the onset of facial action units rely on static or near-static camera setups, making them unsuitable for the strong egomotion intrinsic to first-person video. Likewise, action-spotting methods in sports—such as those used for Olympic events—operate on videos where the target activity dominates the entire frame and the subject remains consistently visible, allowing key moments to be identified without localizing fine-scale regions. In contrast, touch moments in egocentric video occur within compact spatial areas around the fingertips and depend on subtle grasp configurations leading up to contact.

To address these challenges, we introduce the Hand-informed Context Enhancement module (HiCE; pronounced 'high-see'), designed to augment frame-level feature extractors with dedicated hand-centric spatiotemporal modeling. Because the cues that signal an impending touch moment are brief and easily diluted in full-frame representations, HiCE explicitly encodes fine-grained hand motion patterns and injects these signals back into the global feature stream to preserve subtle temporal transitions. The module first extracts hand regions using off-the-shelf detectors (or ground truth boxes during training when available) and expands them to include relevant local context. These regions are then processed by a backbone that captures both spatial structure and short-term temporal dynamics from neighboring frames. The resulting hand-specific features are fused with global frame features through cross-attention, guiding the model to attend to contact-critical regions. To further strengthen hand representation learning, we incorporate a grasp loss that leverages grasp predictions from [4].

To support research on frame-accurate touch detection, we also introduce TouchMoment, an egocentric dataset comprising 4,021 videos and 8,456 annotated touch moments across diverse objects, surfaces, and environments. Experiments on TouchMoment show that our HiCE module delivers substantial gains over strong baselines, including under strict evaluation settings where predictions must fall within a two-frame tolerance of the ground-truth touch moment. In short, our contributions are: **(1)** We study and formalize frame-level touch moment detection as a fine-grained egocentric event-spotting task—an important problem for understanding manipulation, skill analysis, and human–robot learning that has received little dedicated attention in prior work. **(2)** We propose HiCE, a hand-centric enhancement module that injects specialized spatiotemporal hand features into event-spotting architectures, leading to substantial improvements in detecting precise touch moments. **(3)** We introduce TouchMoment, a dataset for egocentric touch detection, enabling systematic investigation, comparison, and benchmarking of methods for this emerging task.

2. Related Works

Temporal action localization and event spotting. Action spotting, introduced by [7, 10], differs from traditional temporal action localization (TAL) by identifying the precise onset frame of an action rather than predicting start and end boundaries of action segments. While TAL methods such as [2, 26, 37, 39, 40], typically evaluate using temporal IoU over multi-second intervals, action spotting employs frame-level mAP metrics at tolerance windows ranging from loose (5–60 seconds) to tight (1-5 seconds). One way to adapt TAL models to spotting is to force dense per-frame predictions through dense anchors or sliding-window proposals. This is inefficient for two reasons. First, achieving frame-level precision requires generating proposals at extremely fine temporal strides, leading to a large number of overlapping windows and substantial computational cost. Second, these proposal mechanisms are designed to regress start and end times of multi-frame temporal segments, which makes them less effective for capturing the subtle, instantaneous transitions characteristic of spotting tasks. As noted in the survey [36], proposal-based TAL formulations are therefore not ideal for high-precision, frame-level event detection. As applications demand finer temporal understanding, action spotting has evolved toward tighter tolerances. [13] demonstrated that tolerances below 4 frames are necessary for tasks requiring precise temporal boundaries—such as sports performance analysis where millisecond differences matter, skill assessment where action phases must be clearly delineated, and robot learning from human demonstrations where contact timing determines manipulation success. At these sub-second scales, distinguishing the exact moment of ball contact, foot landing, or hand-object touch becomes critical, particularly in egocentric scenarios where rapid, subtle events occur within minimal temporal windows. Current spotting approaches fall into two categories. Two-phase methods first extract features, then perform temporal localization. Zhou et al. [43] pioneered this paradigm using an ensemble of five models [9, 12, 22, 30, 38] trained on SoccerNet, with subsequent works [27, 34] adopting their pre-extracted features for temporal detection, or using features to distill in [8]. While effective for precise spotting, this reliance on fixed features limits adaptability to new domains. End-to-end methods emerged to address this limitation. E2E-Spot and

T-DEED [13, 35] employ ResNet-Y backbones with temporal gates (GSM[28]/GSF[29]) and sequence modeling GRU [5] in E2E-Spot, Scalable-Granularity Perception (SGP-Layer) [26] encoder-decoder in T-DEED. Beyond standard classification heads, T-DEED and ASTRA [34, 35] introduced displacement heads that refine predictions along the temporal axis, enabling sub-frame-level localization. However, these methods process frames uniformly without spatial reasoning about salient regions. UGLF [31] addresses this by leveraging vision-language models [15, 42] to localize objects and fuse their features via attention [32]. This approach has two limitations: (1) it requires pre-defining object categories for grounding, and (2) it treats all objects equally. For egocentric touch detection, where contact cues are localized to hand regions, uniform spatial attention is suboptimal. Nevertheless, incorporating spatial object features demonstrates substantial improvements over global frame representations.

Hand-object contact analysis. Hand analysis plays a central role in understanding human behavior and interaction in computer vision. Beyond a large body of work on detecting and tracking the hands themselves (e.g., [14, 18]), a growing line of research has focused on modeling how hands interact with objects, particularly through contact. A rich set of image-based methods has explored frame-level hand–object contact recognition, showing that identifying whether a hand is in contact provides strong cues for action recognition, affordance reasoning, and grasp understanding. Early works such as 100DOH [25], ContactHands [19], and more recent methods like Hands23 [4] classify contact states or categorize interaction types from single frames, relying primarily on local hand appearance and object context. Narasimhaswamy et al. [20] focus on hand detection and hand–body association, using spatial overlap between hand and body regions to infer hand–body contact, demonstrating how contact cues can be derived indirectly from static imagery.

Beyond predicting binary contact, several works aim to localize where contact occurs on objects. HOT [3] detects pixel-level human–object contact regions by training a segmentation-based model to predict dense contact heatmaps and body-part labels from a single RGB image, using manually annotated 2D polygons as supervision. This provides a richer spatial description of contact than binary classification but remains a purely appearance-based formulation without temporal reasoning. At a higher level of abstraction, Goyal et al. [11] leverage contact maps and grasp types to infer object functionality and affordances, but requires rich supervision and is computationally expensive.

More recently, video-based interaction models such as HOISTFormer [21] address hand–object detection, segmentation, and tracking in video. While such models can in principle be adapted for touch spotting, they face two challenges: (1) they are trained on either sparsely annotated datasets such as VISOR [6] or dominated by third-person views, where hand–object motion is more pronounced and contact regions are stable; and (2) in egocentric video, hand motion is subtle, interactions occur at close range, and high-frame-rate videos introduce additional difficulty in distinguishing the exact onset of contact.

While the above approaches highlight the semantic and spatial importance of contact, they operate either on isolated frames or on interaction segments that span many frames. As a result, they lack the temporal granularity needed to discriminate the subtle micro-movements that immediately precede contact, particularly in egocentric video where occlusion, hand jitter, rapid approach motion, and strong ego-motion make individual frames highly ambiguous. Consequently, existing contact-understanding methods cannot determine when contact first occurs. This motivates the need for temporally informed, hand-centric models capable of capturing short-term motion cues and distinguishing near-touch states from the precise moment of first contact—the focus of our work.

3. TouchMoment: An Egocentric Touch Dataset

Despite the importance of identifying precise touch moments in egocentric video, no existing dataset provides large-scale, frame-level annotations of hand touches suitable for developing and benchmarking methods for this task. To fill this gap, we introduce TouchMoment, a new egocentric video dataset that captures a wide range of everyday manipulation scenarios and supplies exact touch-moment annotations for each interaction. The dataset comprises 4,021 videos sourced from diverse environments, object categories, and interaction contexts, and includes 8,456 manually annotated touch moments recorded at the frame level (Tab. 1). Alongside touch annotations, TouchMoment provides temporal segmentation of interaction sequences and localized hand regions to support hand-centric modeling. The following sections detail the data sources, annotation protocol, and an analysis of the dataset's characteristics.

3.1. Data Sources

To build TouchMoment, we source egocentric interaction sequences from two publicly available datasets: HOI4D and the egocentric subset of TACO, both of which contain rich recordings of hand–object manipulation suitable for annotating frame-level touch moments.

HOI4D is a large-scale benchmark for category-level human–object interaction understanding. It provides egocentric RGB-D video sequences captured across diverse indoor environments, featuring a broad range of object categories

and interaction types. The dataset includes dense annotations such as 3D hand pose, object pose, panoptic segmentation, and motion segmentation, making it a valuable source of fine-grained hand–object interaction footage.

We further incorporate sequences from the egocentric portion of the TACO dataset, which contains recordings of bimanual tool and object manipulation captured using a head-mounted RGB camera at 30 fps. These videos include continuous interaction patterns such as grasping, re-grasping, and coordinated motion of both hands, offering temporal variety and diverse manipulation contexts that complement the HOI4D clips. Compared to HOI4D, TACO contains longer and more structured interaction sequences, often involving closely coordinated bimanual activity.

From these data sources, we select segments where the hand–object interface is visible, the transition into contact is discernible, and temporal continuity is preserved. These criteria ensure that annotators can reliably identify and label the precise frame corresponding to each touch moment.

3.2. Touch Annotation

We define a touch event as the moment when any part of the hand makes physical contact with an object. Annotators determine this moment based on observable changes in hand motion, object motion, or object deformation. We include only intentional and visually unambiguous interactions, and exclude accidental contact or cases where the contact moment cannot be reliably observed. After the moment of touch, the hand remains in contact with the object for a short duration, ensuring the event reflects a meaningful interaction rather than a single-frame alignment.

To ensure consistency, we cross-check annotations and resolve disagreements through direct comparison. For HOI4D, which provides action segment annotations, we manually annotate all touch events in the validation and test splits. For the training split, we manually annotate 10% of the segments and use these annotations to develop an automatic annotation tool that assigns touch frames to the remaining of the training data. On a held-out subset with full manual annotation, this automatic tool differs from manual labels by an average of 1.94 frames, demonstrating sufficient accuracy for large-scale training data. For TACO, we manually annotate all selected segments. Unlike HOI4D, TACO does not provide action segmentation, and the hand–object contact patterns are more varied in timing and subtlety. As a result, direct manual annotation is required to determine consistent touch boundaries.

4. Proposed Methodology

In this section, we formalize the frame-level touch detection task and present our approach. We first introduce the problem formulation, then describe the Hand-informed Context Enhancer (HiCE) for augmenting frame features with localized hand cues. We detail the temporal modeling block and prediction heads that operate on these enriched features, and finally present the training supervision used to optimize the full architecture.

	HOI4D			TACO	
	train	val	test	train	test
# Frames	686K	41K	127K	156K	35K
# Touch events	4979	324	930	1689	364
# Clips (0 touches)	1	0	0	5	0
# Clips (1 touch)	42	0	10	198	89
# Clips (2 touches)	1937	106	349	738	133
# Clips (3 touches)	180	16	38	5	3
# Clips ($\geq$4 touches)	128	16	27	0	0

Table 1. **TouchMoment dataset statistics.**

4.1. Problem definition

We follow the precise event spotting formulation introduced in E2E-Spot [13], temporally precise hand touch event detection takes a sequence of L frames $\mathcal{X} = \{x_i\}_{i=1}^{L}$ as input and predicts a sparse set of touch events $\{(t, \hat{y}_t)\} \in \mathbb{N} \times \{0, 1\}$. A prediction is considered correct if it falls within tolerance δ frames of the ground truth label and has the correct class. Since our task falls under Precise Event Spotting (PES), evaluation uses small temporal tolerances. We adopt $\delta \leq 2$, which requires models to localize touch events with near frame-level precision. Our approach builds upon T-DEED [35] by incorporating explicit hand-object context modeling for egocentric perspective interaction scenarios.

4.2. Hand-informed Context Enhanced module

In standard end-to-end precise action spotting architectures, the model consists of a feature extractor, a temporal reasoning module, and prediction heads, as illustrated in Fig. 2. Our work focuses on strengthening the feature extractor by enriching it with explicit hand–context cues. To achieve this, we introduce a cross-attention mechanism that allows the global image features to attend selectively to localized hand regions. Given the global feature map $\mathcal{F} \in \mathbb{R}^{H \times W \times C}$ and the hand patch features $\mathcal{F}_{\text{lh}}, \mathcal{F}_{\text{rh}} \in \mathbb{R}^{H_p \times W_p \times C}$, we first augment them with 2D sinusoidal positional embeddings and hand-identity embeddings (to distinguish left and right hands), then flatten them into token sequences. Queries, keys, and values are constructed as

$$Q = f_Q(\mathcal{F} + E_{\text{pos}}), \tag{1}$$

$$K = f_K\big([\mathcal{F}_{\text{lh}}, \mathcal{F}_{\text{rh}}] + E_{\text{pos}} + E_{\text{id}}\big), \tag{2}$$

$$V = f_V([\mathcal{F}_{\text{lh}}, \mathcal{F}_{\text{rh}}]), \tag{3}$$

where $[\mathcal{F}_{\text{lh}}, \mathcal{F}_{\text{rh}}]$ denotes the stacked left and right hand features, E_{pos} are positional embeddings, E_{id} are left/right

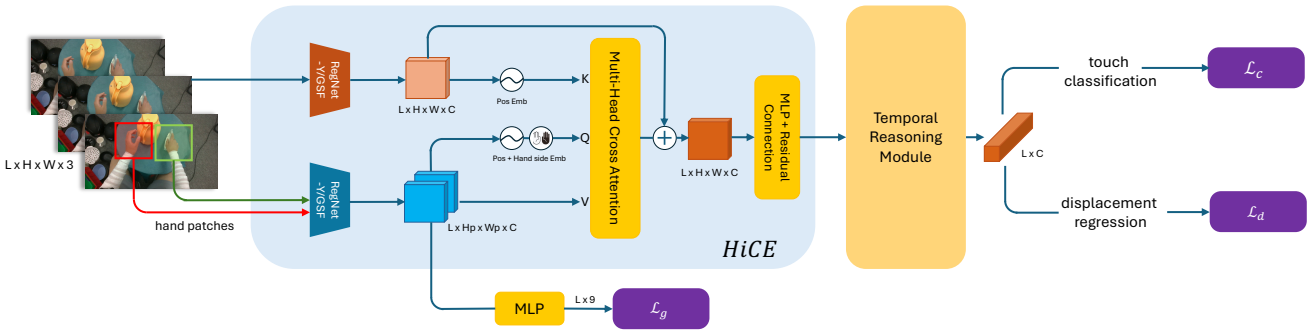

Figure 2. **The Hand-informed Context Enhanced (HiCE) module** augments the feature extractor with a parallel hand-patch branch that processes left and right hand crops alongside global frame features using RegNet-Y backbones. Hand patches are expanded and encoded with positional and identity embeddings, then used as keys and values in a multi-head cross-attention block where global tokens act as queries. The resulting hand-aware global features are passed into the temporal reasoning module for touch classification and displacement prediction, while the hand features are also used by an auxiliary grasp-prediction head.

hand identity embeddings, and f_Q, f_K, f_V are learned linear projections. We then apply cross-attention in the standard form:

$$\text{CrossAttn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \quad (4)$$

and update the global features via residual connections:

$$\mathcal{F}' = \text{CrossAttn}(Q, K, V) + \mathcal{F}, \quad (5)$$
$$\mathcal{F}_{\text{enhanced}} = \text{FFN}(\mathcal{F}') + \mathcal{F}'. \quad (6)$$

Here d is the feature dimension. As illustrated in Fig. 2, this design allows the global representation to selectively integrate contact-relevant information from hand regions, producing enhanced frame-level features enriched with explicit hand context.

4.3. Temporal Reasoning Module

For temporal modeling, we adopt the encoder-decoder architecture from T-DEED, which has demonstrated superior performance over sequential approaches like GRU used in E2E-Spot. The T-DEED architecture employs Scalable-Granularity Perception (SGP) layers [26] that process features at multiple temporal scales while maintaining high token discriminability—a critical property for precise event localization. The encoder-decoder structure with SGP-Mixer layers enables the model to capture both local and global temporal dependencies while preserving frame-level prediction precision through skip connections that restore the original temporal resolution.

4.4. Prediction Heads

We employ a dual-head architecture commonly used in action spotting methods [8, 27, 34, 35]. The classification head predicts the probability of touch events at each frame using a linear layer followed by softmax activation, generating predictions $\hat{y}^c \in \mathbb{R}^{L\times 2}$ for binary touch classification. The displacement regression head refines temporal localization by predicting frame-level offsets $\hat{y}^d \in \mathbb{R}^{L\times 1}$ to ground truth event timestamps. This dual-head approach enables more precise event localization compared to the label dilation technique employed in E2E-Spot, as it can detect events within a wider temporal window while maintaining high localization accuracy through learned displacement offsets.

Additionally, we integrate a grasp classification head [23] to provide auxiliary supervision for the hand patch encoder. This module consists of a four-layer MLP that processes concatenated hand patch features $[\mathcal{F}_{\text{lh}}, \mathcal{F}_{\text{rh}}] \in \mathbb{R}^{2\times H_p \times W_p \times C}$ and predicts hand grasp categories $\hat{y}^g \in \mathbb{R}^9$ based on the Cutkosky taxonomy [4], which divides grasps into prehensile and non-prehensile categories. This auxiliary task encourages the hand encoder to learn more discriminative hand representations that are beneficial for touch event detection.

4.5. Training Supervision

Given an input clip of L frames, the model produces predictions $(\hat{y}^c, \hat{y}^d, \hat{y}^g)$, where $\hat{y}^c \in \mathbb{R}^{L\times 2}$ denotes the classification probabilities for touch events, $\hat{y}^d \in \mathbb{R}$ represents the regressed displacement offset, and $\hat{y}^g \in \mathbb{R}^9$ corresponds to the 9-class grasp predictions for both hands. Pseudo-labels for grasp classification are obtained from [4]. When one or both hands are absent in a frame, the corresponding grasp loss is masked to zero to prevent penalizing missing annotations.

The model is trained with a multi-task objective that jointly optimizes frame-level classification, displacement regression, and grasp classification. The overall training loss is defined as:

$$\mathcal{L} = \mathcal{L}_c + \mathcal{L}_d + \lambda_g \mathcal{L}_g \tag{7}$$
$$= \frac{1}{L} \sum_{l=1}^{L} \Big(\mathcal{L}_c(y_l^c, \hat{y}_l^c) + \text{MSE}(y_l^d, \hat{y}_l^d) + \lambda_g \, \text{CE}(y_l^g, \hat{y}_l^g) \Big),$$

where λ_g controls the contribution of the grasp loss, and $\mathcal{L}_c$ denotes the classification loss, implemented as either Cross Entropy or Focal Loss depending on the displacement window and the characteristics of the dataset.

To further improve temporal discrimination around the touch moment, we modify the supervision within the displacement window. In the standard formulation, all frames within this window are given a hard label of 1 and treated equally as ground truth. Instead, we adopt a Gaussian soft-label, where the target value decreases with distance from the annotated touch frame. This provides a smoother supervisory signal and encourages the model to concentrate its confidence near the true moment of contact rather than spreading it uniformly across neighboring frames. During temporal offset refinement, where predicted displacement offsets are applied to adjust the temporal location of each event, we use bilinear interpolation to distribute fractional offsets across adjacent frames, preventing quantization artifacts. We further modulate the offset with a Gaussian attenuation term to ensure that large displacement values do not introduce unstable shifts in the refined score distribution. For convenience, we refer to this refinement procedure as Gauss-TOR.

5. Experimental Analysis

This section describes the experimental setup on TouchMoment, including implementation details, evaluation metrics, and the baseline methods used for comparison. We then present quantitative results followed by an ablation study to assess the contribution of each component in our approach.

5.1. Implementation Details

For consistency with prior baselines and to accommodate hardware constraints, we train all models using clips of length $L = 40$ frames, a batch size of 6, and 5000 clips per epoch for a total of 50 epochs. We use the AdamW optimizer [17] with an initial learning rate of $4{\times}10^{-4}$, along with a three-epoch linear warm-up followed by cosine annealing. The backbone is RegNet-Y 800MF [24], initialized with ImageNet-pretrained weights from the `timm` library [33]. Hand patches are enlarged by a factor of 1.2 relative to the detected bounding box, padded to a square, and resized to 224×224. The feature dimension is set to $C = 768$ with a downscaling factor of two.

Dataset-specific configurations are as follows. HOI4D contains temporally sparse touch events, so we adopt a displacement window of four and use Focal Loss [16] with $\alpha = 0.9$ and $\gamma = 2$. TACO, in contrast, includes frequent two-handed interactions where left- and right-hand touches occur in close temporal proximity. This requires a smaller displacement window of one and a weighted cross-entropy loss (weight 5.0) to reduce ambiguous supervision. The grasp loss weight is fixed to $\lambda_g = 0.2$ for all experiments.

In our experiments, we found that applying MixUp [41] significantly degraded performance, reducing mAP across thresholds by 7.37%. We attribute this drop to the additional noise introduced in hand–object regions when two sequences are blended. Consequently, MixUp is excluded from our final model, and all remaining model settings and data augmentation strategies follow [35].

5.2. Evaluation Metrics

We follow the standard precise event spotting protocol and report AP for the touch class at tolerance thresholds $\delta \in \{0, 1, 2\}$. A prediction is counted as correct if it falls within δ frames of the annotated touch moment. Unlike sports-based spotting benchmarks, where thresholds of $\delta \leq 4$ are commonly used, touch events in egocentric video often occur in close temporal proximity. These low tolerance values emphasize the model's ability to localize touch events at near frame-level precision.

5.3. Baseline Method

We compare our method against three end-to-end baselines: E2E-Spot, T-DEED, and UGLF [13, 31, 35]. Other action-spotting models such as ASTRA, COMEDIAN, and Soares [8, 27, 34] are excluded because they are designed specifically for the SoccerNetv2 benchmark and depend on Baidu's pre-extracted features [43], which are not applicable in our setting. All baselines in our evaluation are trained end-to-end from raw frames.

For fair comparison, we apply the Soft Non-Maximum Suppression (SNMS) [1] used in T-DEED to all baseline models. We report AP at tolerance thresholds $\delta \in \{0, 1, 2\}$, as well as the mean AP averaged over these thresholds (mAP). For completeness, we provide results both with and without applying NMS/SNMS.

5.4. Experiment Results

Tab. 2 presents touch spotting performance on the HOI4D and TACO subsets of TouchMoment. We report results both without and with NMS/SNMS to distinguish raw temporal precision from post-processed detection quality. The No NMS setting evaluates models directly on their frame-level outputs, reflecting their intrinsic ability to localize the touch moment. In contrast, NMS/SNMS suppresses redundant detections and typically improves overall mAP by reducing false positives, though it may slightly reduce AP@δ=0 when closely spaced peaks are merged or shifted, an important consideration since touch events in egocentric video

(a) HOI4D result.

	Without using NMS				With NMS/SNMS			
	mAP	δ=0	δ=1	δ=2	mAP	δ=0	δ=1	δ=2
E2E Spot	8.71	6.44	8.95	10.73	15.37	2.01	18.10	26.01
UGLF	19.65	<u>13.27</u>	20.51	25.16	<u>31.85</u>	<u>6.17</u>	<u>37.75</u>	<u>51.63</u>
T-DEED	<u>20.32</u>	12.49	<u>21.43</u>	<u>27.04</u>	29.67	4.80	33.91	50.31
Ours	**32.89**(+12.57)	**14.25**(+0.98)	**35.94**(+14.51)	**48.47**(+21.43)	**36.47**(+4.62)	**6.55**(+0.38)	**42.89**(+5.14)	**59.99**(+8.36)

(b) TACO result.

	Without using NMS				With NMS/SNMS			
	mAP	δ=0	δ=1	δ=2	mAP	δ=0	δ=1	δ=2
E2E Spot	16.03	11.85	16.11	20.14	27.2	5.64	32.38	43.59
UGLF	<u>19.83</u>	<u>12.95</u>	<u>20.65</u>	<u>25.90</u>	<u>31.26</u>	6.33	<u>36.39</u>	<u>51.05</u>
T-DEED	17.25	12.56	15.97	20.60	27.95	<u>7.56</u>	33.59	42.69
Ours	**41.08**(+21.25)	**24.25**(+11.3)	**43.47**(+22.82)	**55.51**(+29.61)	**48.38**(+16.12)	**16.78**(+9.22)	**56.18**(+19.79)	**72.18**(+21.13)

Table 2. Quantitative comparison with end-to-end baselines on touch event detection on HOI4D and TACO. The best performing measures are highlighted in **bold**, and second best measures are in <u>underline</u>.

often occur only a few frames apart.

Among the baselines, UGLF achieves the highest AP@δ=0. This outcome is expected because UGLF leverages an external vision–language detector to identify scene objects and extract object-centric features, providing strong spatial priors that help the model attend to interaction-relevant regions. Such cues are particularly beneficial for the strict δ=0 setting, where correct localization depends on capturing subtle spatial differences between adjacent frames. T-DEED, by contrast, improves substantially over E2E-Spot through its displacement heads, which encourage dense candidate predictions followed by temporal offset refinement. However, without explicit instance-level information or structured knowledge of scene entities, T-DEED cannot match UGLF in most cases where precise spatial grounding is essential. This comparison underscores the benefit of incorporating structured spatial cues in general sporting-action spotters and further motivates our hand-centric formulation for egocentric touch detection.

Across most evaluation conditions, our method achieves the strongest performance. On HOI4D, we observe substantial gains at δ=1 and δ=2, without NMS, our method outperforms the best baseline by 14.51 AP@δ=1 and 21.43 AP@δ=2, yielding a 12.57 mAP improvement. With NMS/SNMS, the gains remain pronounced, reaching 36.47 mAP and 59.99 AP@δ=2, the highest among all methods.

On TACO, which involves denser and faster hand–object interactions, our method again delivers consistent improvements. Without NMS, it reaches 41.08 mAP, surpassing the best baseline by 21.25, with gains at all thresholds: 11.3 on AP@δ=0, 22.82 on AP@δ=1, and 29.61 on AP@δ=2. With NMS/SNMS applied, performance further increases to 48.38 mAP and 72.18 AP@δ=2, exceeding the strongest baseline by 16.12 and 21.13, respectively.

5.5. Ablation Studies and Qualitative Results

Tab. 3 summarizes ablations on HOI4D and TACO, evaluating the effect of individual model components, clip length, and context size. Using HiCE alone provides a strong baseline, but the full model yields the best performance across both datasets. On HOI4D, removing soft labels results in a clear performance drop. HOI4D is trained with a large displacement window (size 4), meaning many frames surrounding the true touch moment are annotated as positives. Without soft labels, these near-touch frames become indistinguishable from the true contact frame. Soft labels help resolve this ambiguity by assigning smoothly decaying supervision around the ground truth, enabling the model to better isolate the precise moment of contact. HOI4D also benefits considerably from grasp supervision: the dataset is larger, more diverse, and contains richer hand–object configurations, allowing the grasp branch to regularize the hand-centric representation rather than overfit. On TACO, however, the contributions of each constituent component are different. TACO is only about one quarter of the size of HOI4D and uses a much smaller displacement window (Size 1), resulting in far fewer positive frames and less variation. In this more limited setting, the grasp branch tends to distribute predictions across neighboring frames to model the grasp pattern, which improves robustness across thresholds but slightly dilutes exact-frame precision at $\delta = 0$. Soft labels also provide less benefit on TACO because near-

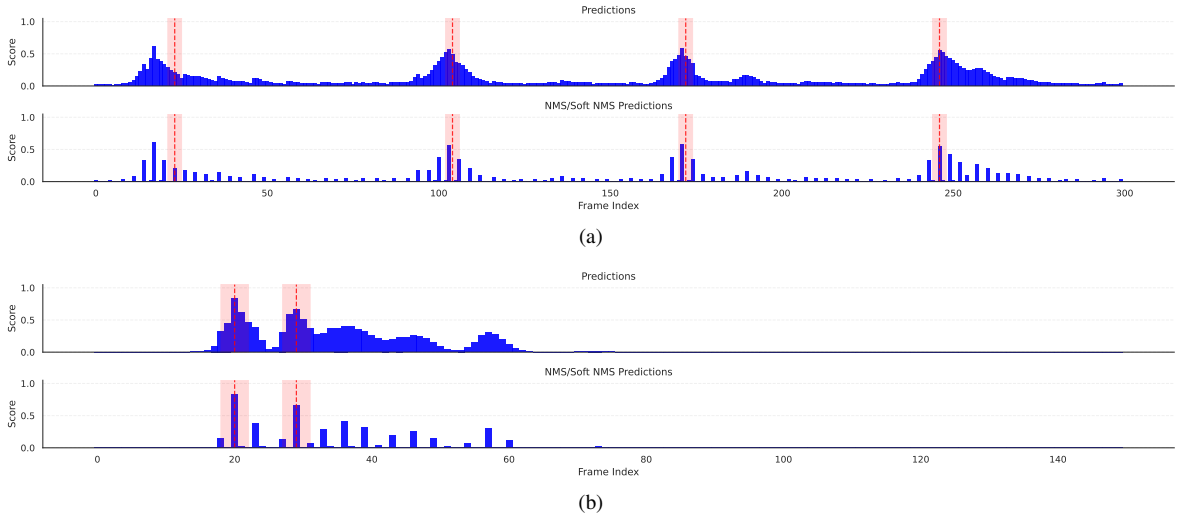

Figure 3. **Qualitative examples of T-DEED with HiCE on HOI4D (a) and TACO (b)**. For each example, the top plot shows raw prediction score and the bottom plot shows predictions after NMS/SNMS. Ground-truth touch frames are indicated by red dashed lines, with a tolerance window of $\delta = 2$ frames.

touch frames are tightly clustered and the narrow displacement window already enforces sharp supervision. Temporal offset refinement (Gauss-TOR) consistently improves accuracy on both datasets, confirming its role in stabilizing offset predictions and producing sharper detection peaks. For clip length and context size, $L = 40$ and $\times 1.2$ yield the best overall balance, as insufficient context harms detection while excessive context dilutes the relevant signal.

	HOI4D		TACO	
	mAP	δ=0	mAP	δ=0
Model components				
Proposed	**32.89**	**14.25**	41.08	**24.25**
w/o Grasp Loss	26.96	11.70	**44.13**	21.87
w/o Gauss-TOR	30.3	14.23	33.09	22.74
w/o Soft Label	23.38	11.66	41.09	18.84
only HiCE	21.10	11.37	30.79	15.97
Clip length				
$L = 25$	17.16	7.99	**44.96**	21.70
$L = 40$ (proposed)	32.89	**14.25**	41.08	**24.25**
$L = 50$	33.80	13.20	40.54	16.33
$L = 80$	**34.93**	12.13	37.63	19.16
Hand context size				
$\times 1.0$	31.40	13.94	32.15	14.95
$\times 1.2$ (proposed)	**32.89**	**14.25**	**41.08**	**24.25**
$\times 1.5$	32.22	12.81	33.41	19.57

Table 3. **Ablation study for TouchMoment**. We evaluate the effect of individual model components, clip length L, and context size on HOI4D and TACO subsets. Bold indicates best results.

Fig. 3 presents qualitative results for T-DEED with HiCE on HOI4D (top) and TACO (bottom). Each example contains two plots: the upper plot displays the raw prediction scores, while the lower plot shows the predictions after NMS/SNMS. In each plot, blue bars represent the predicted touch scores (ranging from 0 to 1). The red dashed line marks the ground-truth touch frame, and the shaded red region denotes the tolerance window of $\delta = 2$ frames. Predictions that fall within this region are counted as true positives. In the HOI4D example, training with a displacement window of 4 generates multiple nearby candidate peaks that consolidate tightly around the ground-truth after applying displacement offset and post-processing. In the TACO example, despite two touch events occurring only nine frames apart (300 ms), the model produces two distinct peaks, demonstrating HiCE's ability to resolve closely-spaced temporal events.

6. Conclusions and Future Work

In this paper, we tackled the task of detecting the precise moment when a hand makes contact with an object in egocentric video. We introduced HiCE, a hand-informed context enhancement module that augments frame-level features with specialized spatiotemporal hand representations. This approach leverages cross-attention to fuse hand-centric cues with global scene context, enabling models to discriminate true touch moments from visually similar near-contact frames. Alongside the architectural innovation, we presented TouchMoment, a dataset comprising 4,021 videos with 8,456 annotated touch moments across diverse manipulation scenarios. Our experiments on TouchMoment and existing egocentric datasets demonstrate HiCE's effectiveness in achieving frame-accurate touch detection under strict temporal tolerance.

Acknowledgements: This work was funded by the Australian Institute for Machine Learning (Adelaide University) and the Centre for Augmented Reasoning, an initiative by the Department of Education, Australian Government.

References

[1] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis. Soft-nms — improving object detection with one line of code. In *Proceedings of the International Conference on Computer Vision*, 2017. 6

[2] S. Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, and Juan Carlos Niebles. SST: Single-stream temporal action proposals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[3] Yixin Chen, Sai Kumar Dwivedi, Michael J. Black, and Dimitrios Tzionas. Detecting human-object contact in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 3

[4] Tianyi Cheng, Dandan Shan, Ayda Sultan, Richard E. L. Higgins, and David F. Fouhey. Towards a richer 2d understanding of hands at scale. In *Advances in Neural Information Processing Systems*, 2023. 2, 3, 5

[5] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *ArXiv*, 2014. 3

[6] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. EPIC-KITCHENS VISOR Benchmark: Video segmentations and object relations. In *Advances in Neural Information Processing Systems*, 2022. 3

[7] Adrien Deliège, Anthony Cioppa, Silvio Giancola, Meisam J. Seikavandi, Jacob V. Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B. Moeslund, and Marc Van Droogenbroeck. SoccerNet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[8] Julien Denize, Mykola Liashuha, Jaonary Rabarisoa, Astrid Orcesi, and Romain Hérault. COMEDIAN: Self-supervised learning and knowledge distillation for action spotting using transformers. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision Workshops*, 2024. 2, 5, 6

[9] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the International Conference on Computer Vision*, 2019. 2

[10] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. SoccerNet: A scalable dataset for action spotting in soccer videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018. 2

[11] Mohit Goyal, Sahil Modi, Rishabh Goyal, and Saurabh Gupta. Human hands as probes for interactive object understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 3

[12] Bo He, Xitong Yang, Zuxuan Wu, Hao Chen, Ser-Nam Lim, and Abhinav Shrivastava. GTA: Global Temporal Attention for Video Action Understanding. In *Proceedings of the British Machine Vision Conference*, 2020. 2

[13] James Hong, Haotian Zhang, Michaël Gharbi, Matthew Fisher, and Kayvon Fatahalian. Spotting temporally precise, fine-grained events in video. In *Proceedings of the European Conference on Computer Vision*, 2022. 2, 3, 4, 6

[14] Mingzhen Huang, Supreeth Narasimhaswamy, Saif Vazir, Haibin Ling, and Minh Hoai. Forward propagation, backward regression, and pose association for hand tracking in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 3

[15] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 3

[16] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020. 6

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of International Conference on Learning and Representation*, 2019. 6

[18] Supreeth Narasimhaswamy, Zhengwei Wei, Yang Wang, Justin Zhang, and Minh Hoai. Contextual attention for hand detection in the wild. In *Proceedings of the International Conference on Computer Vision*, 2019. 3

[19] Supreeth Narasimhaswamy, Trung Nguyen, and Minh Hoai. Detecting hands and recognizing physical contact in the wild. In *Advances in Neural Information Processing Systems*, 2020. 3

[20] Supreeth Narasimhaswamy, Thanh Nguyen, Mingzhen Huang, and Minh Hoai. Whose hands are these? hand detection and hand-body association in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 3

[21] Supreeth Narasimhaswamy, Huy Anh Nguyen, Lihan Huang, and Minh Hoai. HOIST-Former: Hand-held objects identification segmentation and tracking in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 3

[22] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In *Proceedings of the International Conference on Computer Vision Workshops*, 2021. 2

[23] Aditya Prakash, Ruisen Tu, Matthew Chang, and Saurabh Gupta. 3d hand pose estimation in everyday egocentric images. In *Proceedings of the European Conference on Computer Vision*, 2024. 5

[24] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 6

[25] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F. Fouhey. Understanding Human Hands in Contact at Internet

Scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3

[26] Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. TriDet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2, 3, 5

[27] João V. B. Soares and Avijit Shah. Action spotting using dense detection anchors revisited: Submission to the soccernet challenge 2022. *ArXiv*, 2022. 2, 5, 6

[28] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Gate-Shift Networks for Video Action Recognition . In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3

[29] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Gate-shift-fuse for video action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 10913–10928, 2023. 3

[30] Du Tran, Heng Wang, Matt Feiszli, and Lorenzo Torresani. Video Classification With Channel-Separated Convolutional Networks . In *Proceedings of the International Conference on Computer Vision*, 2019. 2

[31] Kim Hoang Tran, Phuc Vuong Do, Ngoc Quoc Ly, and Ngan Le. Unifying Global and Local Scene Entities Modelling for Precise Action Spotting. In *Proceedings of the International Joint Conference on Neural Networks*, 2024. 3, 6

[32] Khoa Vo, Sang Truong, Kashu Yamazaki, Bhiksha Raj, Minh-Triet Tran, and Ngan Le. Aoe-net: Entities interactions modeling with adaptive attention mechanism for temporal action proposals generation. *International Journal of Computer Vision*, 131(1):302–323, 2022. 3

[33] Ross Wightman, Hugo Touvron, and Hervé Jegou. Resnet strikes back: An improved training procedure in timm. *ArXiv*, 2021. 6

[34] Artur Xarles, Sergio Escalera, Thomas B. Moeslund, and Albert Clapés. Astra: An action spotting transformer for soccer videos. In *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, 2023. 2, 3, 5, 6

[35] Artur Xarles, Sergio Escalera, Thomas B. Moeslund, and Albert Clapés. T-DEED: Temporal-Discriminability Enhancer Encoder-Decoder for Precise Event Spotting in Sports Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2024. 3, 4, 5, 6

[36] Hao Xu, Arbind Agrahari Baniya, Sam Well, Mohamed Reda Bouadjenek, Richard Dazeley, and Sunil Aryal. Deep learning for sports video event detection: Tasks, datasets, methods, and challenges. *ArXiv*, 2025. 2

[37] Mengmeng Xu, Chen Zhao, David S. Rojas, Ali Thabet, and Bernard Ghanem. G-TAD: Sub-Graph Localization for Temporal Action Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[38] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal Pyramid Network for Action Recognition . In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[39] Fangqiu Yi, Hongyu Wen, and Tingting Jiang. Asformer: Transformer for action segmentation. In *Proceedings of the British Machine Vision Conference*, 2021. 2

[40] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *Proceedings of the European Conference on Computer Vision*, 2022. 2

[41] Hongyi Zhang, Moustapha Cissé, Yann Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proceedings of International Conference on Learning and Representation*, 2018. 6

[42] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. GLIPv2: Unifying localization and vision-language understanding. In *Advances in Neural Information Processing Systems*, 2022. 3

[43] Xin Zhou, Le Kang, Zhiyu Cheng, Bo He, and Jingyu Xin. Feature combination meets attention: Baidu soccer embeddings and transformer based temporal detection. *ArXiv*, 2021. 2, 6