SkillSpotter: Pose-Aware Multi-View Skilled Action Detection and Grading in Ego-Exo Videos

Björn Braun and Christian Holz

Department of Computer Science, ETH Zürich, Switzerland

https://siplab.org/projects/SkillSpotter

Abstract. To enable personalized, real-time coaching using Augmented Reality glasses or fixed camera setups in domains such as sports, cooking, or music, a system must understand not just *what* a person does, but *how well* they execute an activity. In an ego-exo video setting, this requires simultaneously detecting individual skilled actions and classifying each as correct or needing improvement, which Ego-Exo4D's proficiency demonstration benchmark formalized. We first adapt seven state-of-the-art temporal action detection architectures to this task, extend the evaluation protocol to disentangle detection from grading, and show that existing methods grade near-randomly. We then introduce *SkillSpotter*, a pose-aware multi-view architecture that jointly detects and grades skilled actions through three task-specific modules: (1) adaptive temporal suppression to handle the varying density of skilled actions across diverse activities, (2) gated 3D body pose fusion to leverage body kinematics as a complementary signal to visual features, and (3) bidirectional cross-view attention to combine ego and exo views effectively. *SkillSpotter* improves class-specific mAP from 12.40 to 21.82 (+76%) and balanced accuracy from 55.99% to 60.40% over the best baseline. *SkillSpotter*'s modules transfer to other temporal action detection models with consistent gains and our method generalizes beyond Ego-Exo4D to HoloAssist.

Code: `https://github.com/eth-siplab/SkillSpotter`

Keywords: Skill assessment · Action detection · Egocentric vision

1 Introduction

Automatically assessing how well a person performs a skilled activity could enable personalized, real-time coaching through Augmented Reality (AR) glasses or fixed camera setups in domains such as sports, medical training, and music [48, 61]. This requires detecting *when* skilled actions occur and determining *how well* they are performed. Temporal action detection (TAD) methods localize *when* actions occur in untrimmed recordings by predicting temporal segments with start and end times [37, 38, 40, 45, 59], but do not assess execution quality. Action quality assessment (AQA) methods score *how well* a person performs an

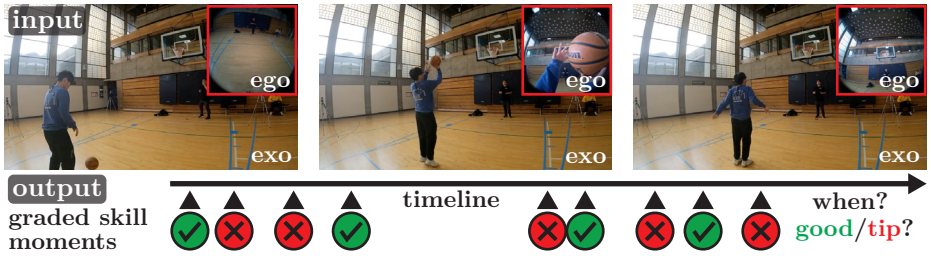

Fig. 1: The *proficiency demonstration* task requires detecting individual skilled actions from untrimmed synchronized ego-exo videos and classifying each action as *good execution* (green) or *tip for improvement* (red). *SkillSpotter* improves class-specific mean average precision from 12.40 to 21.82 (+76%) and balanced accuracy from 55.99% to 60.40% over the best baseline.

action from clips pre-cut around a single execution [53,58,61]. However, neither of these tasks jointly detects and grades individual skill moments at the timestamp level in continuous recordings—a prerequisite for moment-by-moment feedback.

Ego-Exo4D [21] introduced this task in the *proficiency demonstration* benchmark: detecting timestamps of skilled actions in ego-exo videos and additionally classifying each timestamp as good execution or needing improvement (see Fig. 1). Only a minimal baseline exists for this task so far [21]. We adapt seven state-of-the-art TAD architectures to establish a comprehensive baseline. To separately evaluate detection and grading, we extend the evaluation protocol with class-agnostic mAP and balanced accuracy. This reveals that current TAD architectures achieve low detection and grading quality near the random baseline.

In this paper, we introduce *SkillSpotter*, a pose-aware multi-view framework comprising three components that each address specific challenges of timestamp-level skill detection and grading. *Adaptive temporal suppression* addresses the large variation in action density across activities—from sub-second spacing in basketball to several seconds in music—by learning activity-specific suppression radii instead of using a fixed threshold. *Gated 3D body pose fusion* adds body kinematics to better assess execution quality. *Bidirectional cross-view attention* combines ego and exo views to prevent grading accuracy loss observed when naively concatenating these views. We show that our modules transfer to other TAD architectures, and that *SkillSpotter* generalizes beyond Ego-Exo4D to HoloAssist.

We summarize our key contributions as follows:

1. ***SkillSpotter***, a pose-aware multi-view framework for timestamp-level skill detection and grading on the Ego-Exo4D proficiency benchmark, introducing three modules: adaptive temporal suppression, gated 3D body pose fusion, and bidirectional cross-view attention. Our method improves class-specific mAP from 12.40 to 21.82 (+76%) and balanced accuracy from 55.99% to 60.40% over the strongest baseline, with consistent gains across all view settings and activities, generalizing beyond Ego-Exo4D to HoloAssist.

2. **A comprehensive evaluation** of seven state-of-the-art TAD architectures on this task with an extended protocol that separately measures detection and quality grading, revealing that current architectures achieve low detection and near-random grading quality. Applying our proposed modules to these TAD architectures substantially improves their performance as well.
3. **The first empirical inter-annotator agreement ceiling** for the Ego-Exo4D proficiency demonstration benchmark (BA: 64.6, Cohen's κ=0.29), based on the benchmark's multi-annotator data aligned with the evaluation protocol. The ceiling provides a calibration point for future work and shows that *SkillSpotter* reaches 94% of it, suggesting that further BA gains are increasingly limited by label subjectivity rather than model capacity.

2 Related Work

Skill Assessment in Ego-Exo Video. The Ego-Exo4D dataset [21] provides synchronized egocentric and exocentric video of skilled activities across eight scenarios with two proficiency benchmarks: *demonstrator* proficiency (classifying a person's skill level) and *demonstration* proficiency (timestamp-level detection and grading). Recent works have focused on demonstrator proficiency estimation [3, 4, 50] and language feedback generation [2, 5]. Beyond visual features, egoPPG [7] uses heart rate as a complementary cue for proficiency estimation. The EgoExo-Fitness dataset [32] addresses a related problem in the fitness domain, providing segment-level proficiency annotations. Unlike previous work that classifies demonstrator proficiency or generates language feedback, *SkillSpotter* jointly localizes and grades individual skilled actions by combining adaptive temporal suppression, 3D body pose, and ego-exo cross-view attention.

Action Quality Assessment (AQA) evaluates how well a person performs an action, typically by regressing a quality score from a pre-segmented clip [42, 47, 48, 53, 58, 61]. These methods assume known temporal boundaries, whereas Ego-Exo4D demonstration proficiency requires untrimmed timestamp localization and per-timestamp good-versus-tip classification.

Temporal Action Detection and Action Spotting. Temporal action detection (TAD) localizes action segments in untrimmed video, evolving from two-stage proposal pipelines [35, 36, 54, 60] to one-stage anchor-free architectures [34, 45, 46, 57, 59] and DETR-style set prediction [40] or hybrid Mamba–attention backbones [14, 37], unified under the OpenTAD toolbox [38]. These methods target segment localization with intersection-over-union (IoU)-based evaluation. Action spotting also predicts single timestamps with tolerance-based matching [15, 19]. Recent methods use dense per-token classification with displacement heads [24, 51]. The Ego-Exo4D demonstration proficiency benchmark shares this single-timestamp formulation but differs in three key aspects: it requires quality grading of each detected moment, operates on multi-view ego-exo input, and exhibits dense co-occurring events of different classes.

Suppression and Fusion for Dense Events. Standard TAD pipelines apply Soft-NMS [6] to remove duplicate detections, with learned alternatives including density-aware thresholds [39], re-scoring networks [25], and DETR-style set prediction [11] that removes NMS entirely. None of these address timestamp-level, class-aware suppression for co-occurring feedback types. For pose fusion, two-stream RGB–pose architectures are effective for action recognition [18], with 3D pose recoverable from ego-view head tracking [28, 29] or multi-view triangulation [26, 55]; gated fusion [1] and feature-wise modulation [43] offer lightweight integration. Cross-attention across parallel view streams has been explored in representation learning [13, 56] but not for ego-exo skill assessment.

3 Method

3.1 Problem Formulation

Given an untrimmed egocentric and exocentric video of a person performing a skilled activity, such as basketball, cooking, or music, the goal is to detect individual moments of correct or incorrect execution and classify each accordingly. We formalize this following the *demonstration proficiency estimation* task in Ego-Exo4D [21], which jointly requires (i) temporally localizing feedback timestamps and (ii) classifying each as *good execution* or *tip for improvement*.

Input. The input consists of synchronized egocentric and exocentric video features for a single task demonstration: $\mathbf{F}_{\text{ego}} \in \mathbb{R}^{C\times T}$ and $\{\mathbf{F}^m_{\text{exo}}\}_{m=1}^{M}$ with $\mathbf{F}^m_{\text{exo}} \in \mathbb{R}^{C\times T}$, where C is the feature dimension and T the number of temporal tokens. Optionally, 3D body pose sequences $\mathbf{P} \in \mathbb{R}^{D_p\times T}$, where D_p is the pose feature dimension, provide complementary skeletal information.

Output. The model predicts a set of detections $\hat{\mathcal{D}} = \{(\hat{t}_k, \hat{y}_k, \hat{s}_k)\}_{k=1}^{K}$, where $\hat{t}_k \in \mathbb{R}^+$ is a predicted timestamp in seconds, $\hat{y}_k \in \{\texttt{good}, \texttt{tip}\}$ is the quality, and $\hat{s}_k \in [0, 1]$ is the confidence score. Ground-truth annotations $\mathcal{G} = \{(t_j, y_j)\}_{j=1}^{J}$ consist of expert-annotated timestamps each labeled as a *good execution* or a *tip for improvement*. Multiple annotations can share the same timestamp.

Evaluation. For each radius $\delta \in \{0.25, 0.5, 1.0\}$ s, the evaluator sorts predictions by confidence and greedily matches each prediction to the closest unmatched ground-truth timestamp if $|\hat{t}_k - t_j| \le \delta$. The original benchmark reports class-specific mAP, which requires both correct localization and correct label assignment. To decompose detection from classification, we **extend the evaluation protocol** with three complementary metrics: (i) *class-agnostic mAP* (mAP_A), which ignores the quality label and measures pure timestamp detection; (ii) *class-specific mAP* (mAP_S), equivalent to the original metric; and (iii) *balanced accuracy* (BA) and *macro-F1* (F1), computed on matched prediction–ground-truth pairs, which partially decouple classification quality from localization but still depend on matching coverage.

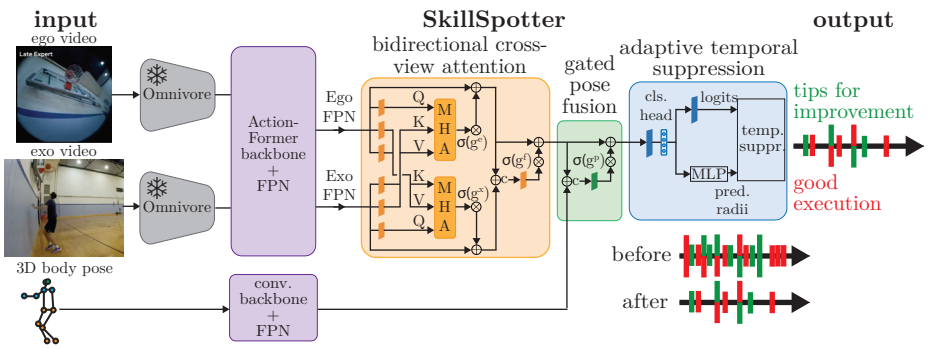

Fig. 2: *SkillSpotter* architecture for skilled action detection and quality grading (Ego+Exos setting shown). *SkillSpotter* takes pre-extracted Omnivore features from ego and exo videos and estimated 3D body pose as input. Video features pass through an ActionFormer backbone and pose is encoded by a separate convolutional backbone. The output is a set of timestamped detections, each classified as good execution or tip for improvement. For single-view settings, the cross-view module is removed.

3.2 Overall Architecture

SkillSpotter extracts Omnivore [20] video features from synchronized ego and exo videos, processes them with an ActionFormer [59] temporal backbone, and finally, predicts per-timestamp quality logits (see Fig. 2). We introduce three modules that each target a specific challenge of timestamp-level skill detection. *Bidirectional cross-view attention* (Sec. 3.2) allows information exchange between ego and exo features. *Gated pose fusion* (Sec. 3.2) incorporates 3D body kinematics as a complementary signal for assessing execution quality. *Adaptive temporal suppression* (Sec. 3.2) replaces fixed-radius non-maximum suppression (NMS) with learned, activity-specific suppression radii to address the large variation in skilled action density across activities. Formally, the ActionFormer backbone processes the video features $\mathbf{F} \in \mathbb{R}^{C \times T}$ into a multi-scale temporal pyramid of L=8 levels $\{\mathbf{f}_l \in \mathbb{R}^{D \times T_l}\}_{l=0}^{L-1}$ with feature dimension D=512 and $T_l = T/2^l$. For single-view settings (Ego and Exos), cross-view attention is removed.

Bidirectional Cross-View Attention. Naively concatenating ego and exo features degrades grading performance for ActionFormer (see Sec. 4). We introduce bidirectional cross-view attention to enable information exchange between views. The ego and exo feature streams are each processed through a shared-weight backbone into separate pyramids $\{\mathbf{f}_{\text{ego},l}\}$ and $\{\mathbf{f}_{\text{exo},l}\}$. At each pyramid level l, both streams attend to each other via multi-head cross-attention (N_h=4 heads) with separate query, key, and value projections per direction:

$$\mathbf{f}'_{\text{ego},l} = \mathbf{f}_{\text{ego},l} + \sigma(g_l^{\text{e}}) \cdot \text{MHA}(\mathbf{f}_{\text{ego},l},\, \mathbf{f}_{\text{exo},l}), \tag{1}$$

$$\mathbf{f}'_{\text{exo},l} = \mathbf{f}_{\text{exo},l} + \sigma(g_l^{\text{x}}) \cdot \text{MHA}(\mathbf{f}_{\text{exo},l},\, \mathbf{f}_{\text{ego},l}), \tag{2}$$

where $\mathrm{MHA}(\mathbf{Q}, \mathbf{KV})$ denotes multi-head cross-attention with the first argument as queries and the second as keys and values. Cross-view attention operates over the full temporal extent, unlike the backbone's windowed self-attention. The enhanced features are merged into a unified representation:

$$\mathbf{f}_l = \mathbf{f}'_{\text{ego},l} + \sigma(g_l^f) \cdot \mathrm{Proj}([\mathbf{f}'_{\text{ego},l}; \mathbf{f}'_{\text{exo},l}]), \tag{3}$$

where the fusion gate g_l^f is initialized to -1.0 and the residual path is from Ego, making the default model ego-centric. Pose fusion (Eq. (4)) is applied after cross-view fusion to integrate skeletal information into the unified representation.

Gated Pose Fusion. Skill assessment is inherently tied to body mechanics. We incorporate 3D body pose as a complementary signal via gated late fusion at each pyramid level. Our default pose input $\mathbf{P} \in \mathbb{R}^{D_p \times T}$ with D_p=118 consists of 51 raw 3D keypoint coordinates (17 COCO joints $\times$ 3) and 67 kinematic features (joint angles, pairwise distances, and inter-frame velocities). We train with the manually annotated GT Ego-Exo4D body pose [21] and evaluate with the pose that we predict ourselves, ensuring no ground-truth leakage at test time. We report the GT-vs-predicted pose comparison in Sec. G (suppl.). For the egocentric view, we predict body pose from the Aria camera trajectory using the official Ego-Exo4D baseline [12,21,28,29] (PA-MPJPE: 10.70 cm). For exocentric views, we detect the actor, estimate 2D keypoints with ViTPose [55], and recover 3D pose via calibrated multi-view triangulation with RANSAC-DLT (PA-MPJPE: 17.98 cm). We encode pose $\mathbf{P}$ with a convolutional backbone to 256 channels and map it to a temporal feature pyramid with output dimension D=512 per level. At each level l, we fuse video and pose features via a gated residual:

$$\mathbf{f}'_l = \mathbf{f}_l + \sigma(g_l^p) \cdot \mathrm{Proj}([\mathbf{f}_l; \mathbf{p}_l]) \tag{4}$$

where $\mathrm{Proj} : \mathbb{R}^{2D} \to \mathbb{R}^D$ is a 1$\times$1 masked convolution followed by group normalization, ReLU, and dropout. g_l^p is a learnable scalar gate initialized to -2.0.

Adaptive Temporal Suppression. Standard post-processing in TAD applies fixed-radius NMS to remove duplicate detections. However, in skill assessment, events of different classes can co-occur at near or identical timestamps, and event density varies dramatically across scenarios: in Basketball, over 80% of events have a neighbor within 0.5 s, whereas in Music fewer than 20% do (see Fig. 3). A single suppression threshold either erases valid co-occurring detections in dense scenarios or under-suppresses in sparse ones.

We replace fixed NMS with a learnable suppression mechanism that predicts per-detection suppression radii conditioned on the detection's features and scenario. Unlike Adaptive NMS [39], our mechanism operates in the temporal domain with continuous exponential decay and class-aware constraints that allow co-occurring event types to coexist. For each candidate detection $\mathbf{h}_k \in \mathbb{R}^D$ from the classification head, a two-layer MLP predicts a suppression radius:

$$r_k = \mathrm{Softplus}(\mathrm{MLP}(\mathbf{h}_k)) + \mathbf{e}_s, \tag{5}$$

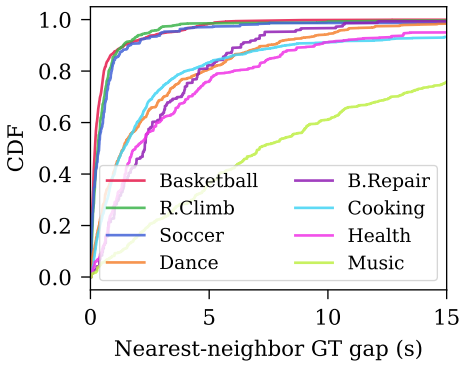

Fig. 3: Cumulative distribution of temporal distances between consecutive ground-truth events, shown per scenario. We see that the event density varies strongly across scenarios: in Basketball, over 80% of events have a neighbor within 0.5 s, whereas in Music fewer than 20% do. This variation means no single fixed suppression radius can suit all scenarios, motivating our learned adaptive suppression.

where MLP : $\mathbb{R}^D \to \mathbb{R}^{128} \to \mathbb{R}^1$ with ReLU activation, and $\mathbf{e}_s \in \mathbb{R}$ is a per-scenario offset from a learnable embedding table (S=8 scenarios, indexed by the scenario label from the dataset metadata).

Given K candidate detections sorted by confidence score s_k, each higher-scoring candidate i suppresses lower-scoring candidate j via exponential decay:

$$w_{ij} = \exp\Big(-\frac{|t_i - t_j|}{r_i}\Big) \cdot \mathbb{1}[s_i > s_j] \cdot \mathbb{1}[y_i = y_j], \tag{6}$$

where the last indicator enforces *class-aware* suppression: detections of different quality classes are not directly suppressed by this class-aware term, preserving co-occurring good/tip feedback at the same timestamp. The adjusted score is $\tilde{s}_j = s_j \cdot \exp\big(-\sum_i w_{ij} \,/\, \tau\big)$, where τ is a learnable temperature initialized to 1.0.

To train the suppression module end-to-end, we introduce an auxiliary loss $\mathcal{L}_{\text{sup}}$ as the suppression module only operates at inference time and receives no gradient from the detection objective. For each training video, we apply the suppressor to the classification scores to obtain adjusted scores $\tilde{s}_k$. Gaussian targets centered on ground-truth timestamps define the supervision signal: $\hat{y}_k = \max_j \exp\big(-(t_k - t_j^*)^2 / 2\sigma_s^2\big)$ with σ_s=1.0 grid units. The auxiliary loss is the binary cross-entropy between the adjusted scores and the Gaussian targets: $\mathcal{L}_{\text{sup}} = \text{BCE}(\text{logit}(\tilde{s}_k), \hat{y}_k)$. The indicator $[\hat{s}_i > \hat{s}_j]$ and candidate selection act as fixed masks that determine suppression topology. Gradients flow through the smooth exponential decay $\exp(-d_{ij}/r_i)$ to the radius predictor and temperature.

Training Objective For classification, we use sigmoid focal loss [59]. For temporal score alignment, we adopt the cumulative loss formulation of Kwak et al. [30]

used by the Ego-Exo4D baseline [21]:

$$\mathcal{L}_{\text{reg}} = \frac{1}{N_{\text{pos}}} \sum_{c=1}^{N_c} \sum_{t} \left(\text{CDF}^c_{\text{pred}}(t) - \text{CDF}^c_{\text{gt}}(t) \right)^2, \tag{7}$$

where N_{pos} is the loss normalizer, N_c is the number of classes, and $\text{CDF}^c_{\text{pred}}(t)$ and $\text{CDF}^c_{\text{gt}}(t)$ denote cumulative class-wise temporal score distributions. The final loss combines all terms with $\lambda_{\text{reg}} = \lambda_{\text{sup}} = 1.0$:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}} + \lambda_{\text{sup}} \mathcal{L}_{\text{sup}}. \tag{8}$$

3.3 Training

We train all models for 15 epochs (5 warmup epochs) with AdamW [41] (learning rate of 1×10^{-3}, weight decay of 0.05) and a batch size of 8. Training takes less than 15 minutes on one NVIDIA H200 GPU for all view settings. Our full Ego+Exos model has 67.37M parameters and runs at 19.70 FPS on one H200 GPU. See Sec. A (suppl.) for a full performance analysis.

4 Experiments

4.1 Dataset and Evaluation Protocol

Following egoPPG [7], we use 90% of the official training set of the *proficiency demonstration* task for training, 10% for validation, and the official validation set for testing. As input, we use frozen Omnivore features (C=1536) to ensure fair comparison across all architectures and avoid overfitting the video backbone on the moderately sized training set (556 takes). Takes without the required pre-extracted Omnivore features are excluded (see Tabs. 17 and 18 for statistics, suppl.). The scenario sizes are highly imbalanced with 3–46 takes but classes near-balanced (49.5% good).

We evaluate three view settings: *Ego* (egocentric only), *Exos* (four exocentric views), and *Ego+Exos* (all five synchronized). Following Ego-Exo4D [21], we concatenate all feature streams for the Exos and Ego+Exos settings. For our cross-view attention model (see Sec. 3.2), we keep ego and mean-pooled exo features as separate streams. Single-view results without concatenation are in Sec. C (suppl.). All metrics are averaged over timestamp matching radii $\{0.25, 0.5, 1.0\}$ s (see Sec. 3.1). For fair comparison, all methods share the same Omnivore features and training protocol, and each baseline uses its default Soft NMS configuration as implemented in OpenTAD [38].

4.2 Evaluation Results

In Tab. 1, we compare *SkillSpotter* to seven state-of-the-art TAD architectures [14, 37, 40, 45, 46, 57, 59]. All re-implemented baselines substantially outperform the

original benchmark result [21] (mAP_S: 3.27 vs. 7.63–12.40) but remain below 12.40 class-specific mAP, with detection limited to 8.65–17.11 mAP_A and quality classification within 6 percentage points of the random baseline (BA 49.51–55.99% vs. 50.9% random). The naive baselines follow the Ego-Exo4D protocol [21] (see Tab. 1). F1 tracks BA closely throughout all experiments; we focus on BA hereafter. To reduce the chance that weak baseline performance is caused by under-tuning, we sweep TadTR learning rates/query counts (see Sec. B, suppl.) and compare single-view vs. concatenated inputs for four of the implemented baselines (see Sec. C, suppl.). These sweeps do not close the gap to *SkillSpotter*. We additionally evaluate if AQA heads are better suited for this task than our classification head by substituting three AQA heads into our pipeline (see Sec. F, suppl.). Our default head remains strongest in grading quality, while detection stays comparable across heads.

SkillSpotter improves class-specific mAP from 12.40 to 21.82 (+76%) and balanced accuracy from 55.99% to 60.40% on Ego (see Tab. 1). These gains generalize across all view settings: Ego (21.82 mAP_S, 60.40 BA), Exos (21.12, 60.59), and Ego+Exos (21.34, 60.39), outperforming all baselines by a large margin. Notably, some baselines degrade under naive Ego+Exos concatenation (e.g., ActionFormer BA 55.99→50.34) compared to only using a single view as input. We analyze this instability in Sec. 4.3.

Table 1: Timestamp-level skill assessment on Ego-Exo4D demonstration proficiency. We report class-specific mean average precision (mAP_S), class-agnostic mAP (mAP_A), balanced accuracy (BA), and macro-F1 across three view settings. All metrics are averaged over matching radii $\{0.25, 0.5, 1.0\}$ s. Baselines use Soft NMS. Our method uses learned adaptive suppression. † Original benchmark result [21], for which only mAP_S was reported. * Uses cross-view attention for Ego+Exos.

	Ego				**Exos**				**Ego+Exos**			
Model	mAP_S	mAP_A	BA	F1	mAP_S	mAP_A	BA	F1	mAP_S	mAP_A	BA	F1
Random	0.73	1.49	50.90	50.03	0.70	1.47	50.44	49.59	0.70	1.46	50.15	49.39
Uniform tips	0.71	1.49	50.00	27.15	0.71	1.47	50.00	27.15	0.72	1.46	50.00	27.15
Uniform good	0.70	1.52	50.00	38.55	0.68	1.47	50.00	38.55	0.67	1.46	50.00	38.55
Baseline†	3.27	–	–	–	3.84	–	–	–	3.57	–	–	–
VideoMambaSuite [14]	7.63	8.65	49.51	49.35	7.06	8.17	46.03	43.88	3.69	3.91	52.70	51.96
TadTR [40]	7.79	10.79	49.68	33.62	6.37	10.48	52.31	34.63	4.12	6.74	52.81	48.23
DyFADet [57]	10.09	12.53	48.89	46.75	3.57	4.48	49.52	47.72	3.18	5.64	47.63	35.45
TriDet [45]	10.35	14.79	49.17	40.10	8.99	12.16	50.06	36.78	8.23	11.78	48.92	48.77
CausalTAD [37]	11.42	16.07	52.86	52.07	11.82	14.05	50.78	50.41	13.16	17.21	54.98	54.95
TemporalMaxer [46]	12.34	16.69	54.34	53.90	10.38	15.17	52.18	52.16	11.27	15.75	50.09	49.58
ActionFormer [59]	12.40	17.11	55.99	55.91	13.18	18.25	55.03	55.03	13.82	17.85	50.34	50.17
SkillSpotter*	**21.82**	**27.89**	**60.40**	**60.02**	**21.12**	**27.47**	**60.59**	**60.55**	**21.34**	**28.01**	**60.39**	**59.37**
Δ best baseline	**+9.42**	**+10.78**	**+4.41**	**+4.11**	**+7.94**	**+9.22**	**+5.56**	**+5.52**	**+7.52**	**+10.16**	**+5.41**	**+4.42**

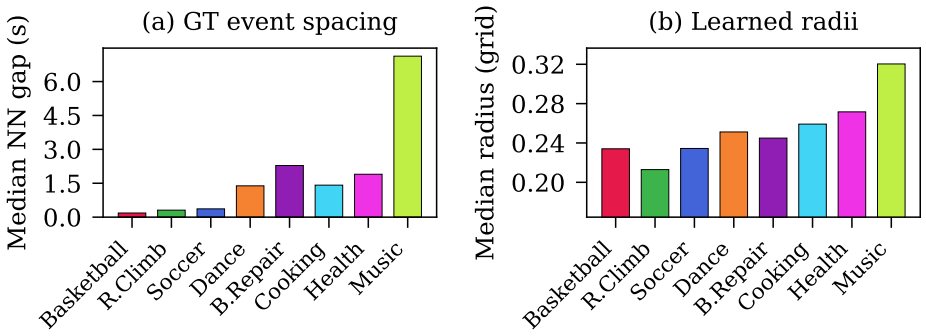

Fig. 4: (a) On Ego-Exo4D, the temporal distance between ground-truth events varies strongly across scenarios. (b) *SkillSpotter* learns to match this variation through our proposed adaptive temporal suppression module, which learns to predict scenario-specific suppression radii. Dense scenarios, such as Basketball, receive smaller radii, sparse ones, such as Music, larger radii (Spearman $\rho = 0.83$, $p = 0.010$).

4.3 Ablation Study

Suppression Strategy. Removing Soft NMS entirely (No NMS) from the ActionFormer baseline [59] increases mAP_A substantially (Ego: 16.85→22.50) but mAP_S only modestly (12.12→13.96), while BA decreases slightly (55.34→54.81). This means that the additional detections recovered without suppression are poorly classified, indicating a trade-off between detection and grading precision.

Our introduced temporal adaptive suppression outperforms both fixed strategies (No NMS and Soft NMS) by a large margin: mAP_S rises from 13.96 to 18.82, mAP_A from 22.50 to 25.87, and BA from 54.81 to 59.74—jointly improving detection and classification. Class-aware decay restricts suppression to detections of the same class (see Eq. (6)), allowing temporally overlapping good and tip events to coexist. We sweep Soft NMS σ in Sec. D (suppl.), showing that no single fixed σ jointly optimizes detection and classification in our setting, and that the best fixed setting remains below adaptive suppression.

Fig. 4 shows the ground-truth event density (left) and the median learned suppression radii (right) per scenario of our temporal suppression module. The median learned radius per scenario correlates with the ground-truth event spacing (Spearman $\rho = 0.83$, $p = 0.010$): without any explicit density supervision, the model learns smaller radii for densely labeled scenarios such as Basketball (median nearest-neighbor gap 0.05 s, radius 0.23 grid units) and larger radii for sparser ones such as Music (7.05 s, 0.32 grid units). Per-scenario radius distributions are provided in Sec. H (suppl.).

Pose Fusion. Adding 3D body pose further improves both detection and classification. On Ego, mAP_S increases from 18.82 to 21.82 (+3.00), mAP_A from 25.87 to 27.89 (+2.02), and BA from 59.74 to 60.40. On Exos, the BA gain is even larger (58.29→60.59, +2.30), suggesting that pose is particularly complemen-

tary to exo-view appearance features. We also evaluate whether predicted pose compared to ground truth during inference affects performance and find only a marginal decrease when using our predicted pose from the ego and exo views (see Sec. G, suppl.). Adding hand-pose (using Ego-Exo4D's POTTER baseline) to our gated mechanism gives mAP_S 21.82→20.92, BA 60.40→59.56 on Ego. We therefore did not include hand pose further.

Cross-View Attention. While adaptive suppression and pose fusion improve all metrics in the single-view settings, the Ego+Exos setting exhibits a classification collapse: BA drops from 51.79 to 45.77 with adaptive suppression and pose fusion.

Cross-view attention addresses this by maintaining view-specific feature streams with explicit bidirectional information exchange before gated fusion, rather than relying on the backbone to separate concatenated ego-exo representations. BA recovers to 60.39 (+14.62 over the concatenated), which is close to the Ego/Exos single-stream results (60.40/60.59 in Tab. 2) and higher than the Ego+Exos per-camera single-view variant (55.93 in Tab. 8, suppl.). We observe a slight mAP_S decrease (21.99→21.34) relative to the pose-only concatenated row, indicating a minor detection-grading tradeoff. These results are consistent with the benefit of view-specific processing with explicit bidirectional exchange before fusion.

Table 2: Component ablation of our model. All rows use the ActionFormer backbone. Each row progressively adds one component. * Cross-view attention applies to Ego+Exos only; Ego and Exos results are unchanged from the row above.

	Ego				**Exos**				**Ego+Exos**			
Configuration	mAP_S	mAP_A	BA	F1	mAP_S	mAP_A	BA	F1	mAP_S	mAP_A	BA	F1
ActionFormer (Soft NMS)	12.12	16.85	55.34	55.24	13.15	18.13	54.75	54.75	13.71	17.86	50.24	50.01
ActionFormer (No NMS)	13.96	22.50	54.81	54.26	15.42	24.47	54.85	54.65	17.30	24.12	51.79	51.40
+ Adaptive Supp.	18.82	25.87	59.74	59.47	21.03	27.14	58.29	57.65	21.18	27.94	43.06	39.39
+ Pose Fusion	**21.82**	**27.89**	**60.40**	**60.02**	**21.12**	**27.47**	**60.59**	**60.55**	**21.99**	**28.04**	45.77	43.57
+ Cross-View Attn* (Ours)	–	–	–	–	–	–	–	–	21.34	28.01	**60.39**	**59.37**

Cross-Backbone Evaluation. We also apply adaptive suppression and pose fusion to all implement backbones, except TadTR [40] (as it lacks a suppression stage), to test whether our modules transfer beyond ActionFormer (see Tab. 3 and Sec. E, suppl.). Adaptive suppression and pose fusion improve detection (mAP_S, mAP_A) across all backbones except VideoMambaSuite, for which only Adaptive Suppression improves performance. Pose fusion's benefit to grading (BA, F1) is less consistent: it improves BA across all view settings only for TriDet and CausalTAD, while the other three each show a BA regression in at least one view setting.

Table 3: *SkillSpotter*'s adaptive suppression and pose fusion transfer to five further TAD backbones on Ego-Exo4D demonstration proficiency. TriDet [45], CausalTAD [37], VideoMambaSuite [14], and TemporalMaxer [46] improve substantially across all view settings; DyFADet [57] improves detection throughout but suffers a grading regression on Exos (†). We exclude TadTR [40] because its DETR-style set prediction removes the suppression stage.

		Ego				**Exos**				**Ego+Exos**			
Backbone	**Configuration**	mAP_S	mAP_A	BA	F1	mAP_S	mAP_A	BA	F1	mAP_S	mAP_A	BA	F1
TriDet [45]	Soft NMS	10.35	14.79	49.17	40.10	8.99	12.16	50.06	36.78	8.23	11.78	48.92	48.77
	No NMS	11.80	17.80	52.03	50.56	11.47	15.45	48.53	48.49	9.46	14.71	48.83	48.25
	+ Our modules	**20.47**	**26.91**	**59.98**	**58.77**	**19.21**	**24.32**	**60.56**	**59.93**	**19.33**	**25.56**	**57.85**	**57.84**
CausalTAD [37]	Soft NMS	11.42	16.07	52.86	52.07	11.82	14.05	50.78	50.41	13.16	17.21	54.98	54.95
	No NMS	12.77	22.71	51.96	50.20	12.57	18.91	53.16	52.87	12.69	21.45	54.63	54.47
	+ Our modules	**21.36**	**27.46**	**59.34**	**59.23**	**20.76**	**25.81**	**60.53**	**60.36**	**20.02**	**26.83**	**60.01**	**59.51**
VideoMambaSuite [14]	Soft NMS	7.63	8.65	49.51	49.35	7.06	8.17	46.03	43.88	3.69	3.91	52.70	51.96
	No NMS	13.60	16.22	53.55	53.23	11.44	15.70	47.79	40.48	6.91	9.22	52.02	47.96
	+ Our modules	**21.39**	**27.11**	**55.01**	**53.50**	**18.66**	**26.33**	**54.46**	**54.06**	**19.91**	**26.03**	**56.92**	**56.90**
TemporalMaxer [46]	Soft NMS	12.34	16.69	54.34	53.90	10.38	15.17	52.18	52.16	11.27	15.75	50.09	49.58
	No NMS	14.76	22.83	53.89	53.12	11.56	20.53	52.57	52.55	13.20	22.42	52.07	50.46
	+ Our modules	**20.15**	**27.85**	**58.66**	**57.48**	**20.65**	**27.01**	**59.66**	**59.66**	**20.55**	**27.38**	**59.42**	**59.34**
DyFADet [57]	Soft NMS	10.09	12.53	48.89	46.75	3.57	4.48	**49.52**	**47.72**	3.18	5.64	47.63	35.45
	No NMS	12.74	16.26	50.14	46.48	3.68	4.69	47.23	45.00	3.69	5.57	47.54	40.41
	+ Our modules	**21.56**	**28.68**	**56.77**	**56.51**	**20.16**	**25.22**	41.83†	36.71†	**20.94**	**27.25**	**58.86**	**58.49**

4.4 Analysis

Per-Scenario Breakdown. Tab. 4 (top) reports Ego results per scenario for our method and ActionFormer. We compare against ActionFormer because it provides the strongest baseline trade-off in the Ego setting. Per-scenario trends are consistent across view settings. Our method improves mAP_S and mAP_A across all eight scenarios, with the largest mAP_S gains on Basketball (+17.89) and Rock Climbing (+10.02). BA improves by up to +9.68 (Dance), with the exception of Cooking, Health, and Soccer where ActionFormer retains a higher BA. The largest gains appear in Basketball, Rock Climbing, and Dance, which are also among the denser scenarios (see Fig. 4). We also observe that scenarios with coarse full-body motion benefit more than scenarios dominated by subtle manipulation cues. This pattern is consistent with adaptive suppression helping dense-event regimes and with the current feature set (appearance + body pose) capturing gross kinematics better than fine hand-object interactions. Cooking and Music remain challenging for detection (mAP_S: 3.08 and 1.57 for our method), indicating that current appearance-plus-body-pose cues still miss part of the quality signal. In particular, fine-grained hand-object interactions likely require stronger hand-level or object-state representations than we currently use. Soccer (BA: 49.48) and Cooking (F1: 43.22) remain difficult grading regimes. We interpret Soccer cautiously because the evaluation split contains only three videos (see Tab. 18, suppl.). We additionally report detection recall (R), defined as the fraction of ground-truth events matched by any prediction within the matching radius. Our method consistently achieves higher recall than ActionFormer across all scenarios (e.g., 72.89% vs. 49.69% on Basketball), confirming

Table 4: Per-scenario and per-radius breakdown of *SkillSpotter* vs. ActionFormer (Ego setting). *SkillSpotter* improves detection (mAP_S, mAP_A) and recall (R) across all eight scenarios, with the largest gains in dense, full-body scenarios such as Basketball and Rock Climbing. Per-radius results show that *SkillSpotter* is especially effective at the strictest matching threshold (0.25 s).

	***SkillSpotter* (Ours)**					**ActionFormer**				
	mAP_S	mAP_A	BA	F1	R	mAP_S	mAP_A	BA	F1	R
Per Scenario										
Basketball	**40.76**	**49.02**	**59.78**	**59.49**	**72.89**	22.87	29.60	52.69	52.66	49.69
Bike Repair	**7.65**	**10.98**	**63.09**	**62.13**	**51.13**	3.01	4.45	53.56	52.10	33.68
Cooking	**3.08**	**6.75**	51.64	43.22	**27.32**	1.94	4.34	**52.02**	**48.02**	21.37
Dance	**8.02**	**11.58**	**61.27**	**59.69**	**52.75**	5.79	9.58	51.59	46.99	37.92
Health	**2.67**	**3.96**	56.24	56.18	**20.05**	1.70	2.75	**59.97**	**57.91**	18.65
Music	**1.57**	**1.91**	**58.94**	**58.90**	**23.03**	0.41	0.89	58.19	54.70	15.74
Rock Climbing	**28.87**	**42.47**	**58.84**	**55.94**	**73.86**	18.85	26.79	52.90	51.59	53.05
Soccer	**12.99**	**24.54**	49.48	36.57	**37.02**	7.60	12.20	**58.37**	**57.76**	27.11
Per Radius										
0.25s	**13.49**	**17.06**	**64.75**	**64.89**	**38.92**	4.21	6.89	54.70	54.67	21.28
0.5s	**21.87**	**28.15**	**59.79**	**59.28**	**52.21**	10.21	14.62	55.76	55.69	34.90
1.0s	**30.09**	**38.47**	**56.65**	**55.89**	**63.09**	21.93	29.05	55.55	55.37	53.25
Average	**21.82**	**27.89**	**60.40**	**60.02**	**51.41**	12.12	16.85	55.34	55.24	36.48

that the BA and F1 improvements are not artifacts of selective matching but reflect genuinely better coverage of annotated events.

Per-Radius Breakdown. Tab. 4 (bottom) reports results at individual matching radii for *SkillSpotter* and ActionFormer on Ego. *SkillSpotter* improves mAP_S over ActionFormer at every radius, with the largest relative gain at the strictest threshold (13.49 vs. 4.21 at 0.25 s). For both methods, BA decreases at looser radii while mAP_S increases: stricter matching retains temporally precise detections, while looser matching admits predictions farther from annotated timestamps. Detection recall (R) confirms that this BA drop is not an artifact of selective matching: at 0.25 s only 38.92% of GT events are matched, rising to 63.09% at 1.0 s, so the lower BA at loose radii reflects genuinely harder cases entering the evaluation pool. *SkillSpotter* consistently matches more GT events than ActionFormer at every radius (e.g., 38.92% vs. 21.28% at 0.25 s), indicating that its mAP and BA gains are not due to cherry-picking easy timestamps. The mAP_A–mAP_S gap widens at looser radii for both methods (e.g., Ours: 3.57 at 0.25 s vs. 8.38 at 1.0 s).

Table 5: Mistake detection on HoloAssist [49]. *SkillSpotter* performs joint action detection and classification on *untrimmed* video, while baselines classify on *pre-segmented* clips. PC/RC and PM/RM denote precision/recall for the Correct and Mistake classes, respectively.

Model	F1	PC	RC	PM	RM
Random	27.7	60.9	10.2	**15.0**	46.2
HoloAssist [49]	36.2	85.5	43.1	9.7	11.5
DR-MoE [22]	**57.0**	**97.0**	60.0	8.0	**63.0**
SkillSpotter (ours)	53.3	95.5	**90.7**	10.6	20.7

Inter-Annotator Agreement. Ego-Exo4D provides 2–5 expert annotators per take, giving a first empirical ceiling for grading quality on this benchmark. For each pair of experts annotating the same take, we match their timestamps using the same matching radii as our evaluation protocol ($\delta \in \{0.25, 0.5, 1.0\}$ s). Only 13–27% of expert annotations have an inter-expert counterpart across these radii. On the co-located subset, averaged across radii, inter-expert balanced accuracy is 64.6 with a Cohen's κ of 0.29 (fair agreement). Our method's BA of 60.40 reaches 94% of this empirical ceiling, suggesting that further gains in balanced accuracy on this benchmark are increasingly limited by label subjectivity rather than model capacity, and that mAP_S remains the more discriminative metric for measuring future progress.

4.5 Generalization to HoloAssist

To test generalization beyond Ego-Exo4D, we evaluate *SkillSpotter* on HoloAssist [49], an egocentric mistake-detection benchmark. While HoloAssist's official protocol performs classification on *pre-segmented* clips, *SkillSpotter* extends this protocol by *jointly* performing action detection and classification on *untrimmed* video. On HoloAssist's official classification metric, *SkillSpotter* reaches F1=53.3, close to the SOTA (DR-MoE [22], F1=57.0; Tab. 5). We additionally extend HoloAssist to action localization, which has not been previously evaluated on this dataset: our modules more than double ActionFormer's detection (mAP_S: 3.33→7.24), with absolute performance consistent with the hand-manipulation scenarios (Cooking, Music) in Ego-Exo4D.

4.6 Discussion and Limitations

Performance varies substantially across scenarios: mAP_S ranges from 1.57 (Music) to 40.76 (Basketball), indicating that the current method is not uniformly robust across activity types. We interpret this variation together with strong scenario-size imbalance (3–46 videos and 191–2,409 annotations in our filtered evaluation split; see Tab. 18, suppl.). Accordingly, we treat Soccer (3 videos) as a low-sample regime with high metric variance and avoid strong scenario-specific conclusions, similar to the findings in egoPPG [7]. At the same time, persistent

errors in high-volume scenarios such as Cooking and Music indicate that data scarcity alone does not explain the remaining failure cases. We observe larger gains in scenarios with coarse full-body motion (e.g., Basketball and Rock Climbing) than in scenarios dominated by subtle manipulation cues (e.g., Cooking). Cross-view attention recovers Ego+Exos BA to single-view levels (60.39) but does not surpass them, indicating that effective multi-view fusion for skill grading remains an open challenge [31]. However, resolving the classification collapse is a necessary prerequisite for future multi-view improvements. Our HoloAssist results confirm that *SkillSpotter* and its modules generalize beyond Ego-Exo4D, though broader cross-dataset evaluation remains an open direction. Furthermore, our inter-annotator agreement analysis yields an empirical BA ceiling of 64.6 for this benchmark, which our method's 60.40 BA approaches (94%). This indicates that skill assessment is inherently subjective and that mAP_S remains the more discriminative metric for measuring future progress on EgoExo4D. Promising directions for future work include incorporating physiological signals from egocentric vision, such as heart rate [7, 10], heart-rate variability [16, 17], electrodermal activity [8, 9], emotional states [27], multi-view recordings of motions and activities [23, 44], and pose-object interaction features to address the remaining failure scenarios. End-to-end training from raw video could further jointly optimize feature extraction and temporal modeling.

5 Conclusion

We have introduced *SkillSpotter*, a pose-aware multi-view framework that jointly localizes and grades skilled actions in untrimmed ego-exo video through three components specifically designed for this task: adaptive temporal suppression, gated 3D body pose fusion, and bidirectional cross-view attention. On the Ego-Exo4D demonstration proficiency benchmark, *SkillSpotter* improves class-specific mAP from 12.40 to 21.82 and balanced accuracy from 55.99% to 60.40% compared to the best baseline, with consistent gains across all viewpoints. Our adaptive suppression and pose fusion modules transfer to other backbones with similarly substantial improvements and *SkillSpotter* itself generalizes beyond Ego-Exo4D to HoloAssist. The gains are thus not architecture- or dataset-specific.

Performance remains uneven across scenarios, with the strongest results in activities dominated by coarse full-body motion and weaker results in scenarios driven by subtle manipulation cues (Cooking and Music), indicating that the current appearance-plus-body-pose representation is insufficient for some tasks. We additionally establish the first empirical inter-annotator agreement ceiling for this benchmark (BA: 64.6, Cohen's κ=0.29). *SkillSpotter* reaches 94% of this ceiling, which suggests that further BA gains are increasingly limited by label subjectivity rather than model capacity. Incorporating pose-object interaction features and extending the framework toward natural-language feedback are promising directions for future work.

References

1. Arevalo, J., Solorio, T., Montes-y Gómez, M., González, F.A.: Gated multimodal units for information fusion. arXiv preprint arXiv:1702.01992 (2017)
2. Ashutosh, K., Nagarajan, T., Pavlakos, G., Kitani, K., Grauman, K.: Expertaf: Expert actionable feedback from video. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13582–13594 (2025)
3. Bianchi, E., Liotta, A.: Pats: Proficiency-aware temporal sampling for multi-view sports skill assessment. In: 2025 IEEE International Workshop on Sport, Technology and Research (STAR). pp. 1–6. IEEE (2025)
4. Bianchi, E., Liotta, A.: Skillformer: Unified multi-view video understanding for proficiency estimation. arXiv preprint arXiv:2505.08665 (2025)
5. Bianchi, E., Staiano, J., Liotta, A.: Profvlm: A lightweight video-language model for multi-view proficiency estimation. arXiv preprint arXiv:2509.26278 (2025)
6. Bodla, N., Singh, B., Chellappa, R., Davis, L.S.: Soft-nms–improving object detection with one line of code. In: Proceedings of the IEEE international conference on computer vision. pp. 5561–5569 (2017)
7. Braun, B., Armani, R., Meier, M., Moebus, M., Holz, C.: egoPPG: Heart rate estimation from eye-tracking cameras in egocentric systems to benefit downstream vision tasks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5579–5590 (2025)
8. Braun, B., McDuff, D., Baltrusaitis, T., Holz, C.: Video-based sympathetic arousal assessment via peripheral blood flow estimation. Biomedical Optics Express **14**(12), 6607–6628 (2023)
9. Braun, B., McDuff, D., Baltrusaitis, T., Streli, P., Moebus, M., Holz, C.: Sympcam: Remote optical measurement of sympathetic arousal. In: 2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI). pp. 1–8. IEEE (2024)
10. Braun, B., McDuff, D., Holz, C.: How suboptimal is training rppg models with videos and targets from different body sites? In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition. pp. 410–418 (2024)
11. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
12. Castillo, A., Escobar, M., Jeanneret, G., Pumarola, A., Arbeláez, P., Thabet, A., Sanakoyeu, A.: Bodiffusion: Diffusing sparse observations for full-body human motion synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4221–4231 (2023)
13. Chen, C.F.R., Fan, Q., Panda, R.: Crossvit: Cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 357–366 (2021)
14. Chen, G., Huang, Y., Xu, J., Pei, B., Wang, J., Chen, Z., Li, Z., Lu, T., Wang, L.: Video mamba suite: State space model as a versatile alternative for video understanding. International Journal of Computer Vision **134**(1), 20 (2026)
15. Deliege, A., Cioppa, A., Giancola, S., Seikavandi, M.J., Dueholm, J.V., Nasrollahi, K., Ghanem, B., Moeslund, T.B., Van Droogenbroeck, M.: Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4508–4519 (2021)

16. Demirel, B.U., Holz, C.: Continuous heart rate variability estimation from ppg via state-space modeling. IEEE Transactions on Biomedical Engineering pp. 1–8 (2026).
17. Demirel, B.U., Holz, C.: EgoHRV: Continuous heart rate variability estimation from egocentric systems for autonomic response and skill assessment. In: European Conference on Computer Vision (ECCV) (2026)
18. Duan, H., Zhao, Y., Chen, K., Lin, D., Dai, B.: Revisiting skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2969–2978 (2022)
19. Giancola, S., Amine, M., Dghaily, T., Ghanem, B.: Soccernet: A scalable dataset for action spotting in soccer videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 1711–1721 (2018)
20. Girdhar, R., Singh, M., Ravi, N., Van Der Maaten, L., Joulin, A., Misra, I.: Omnivore: A single model for many visual modalities. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16102–16112 (2022)
21. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., et al.: Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19383–19400 (2024)
22. Han, B., Xu, Q., Bao, S., Yang, Z., Li, S., Huang, Q.: Dual-stage reweighted moe for long-tailed egocentric mistake detection. arXiv preprint arXiv:2509.12990 (2025)
23. Hollidt, D., Streli, P., Jiang, J., Haghighi, Y., Qian, C., Liu, X., Holz, C.: EgoSim: An egocentric multi-view simulator and real dataset for body-worn cameras during motion and activity. In: Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track (2024)
24. Hong, J., Zhang, H., Gharbi, M., Fisher, M., Fatahalian, K.: Spotting temporally precise, fine-grained events in video. In: European Conference on Computer Vision. pp. 33–51. Springer (2022)
25. Hosang, J., Benenson, R., Schiele, B.: Learning non-maximum suppression. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4507–4515 (2017)
26. Iskakov, K., Burkov, E., Lempitsky, V., Malkov, Y.: Learnable triangulation of human pose. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7718–7727 (2019)
27. Jammot, M., Braun, B., Streli, P., Wampfler, R., Holz, C.: egoEMOTION: Egocentric vision and physiological signals for emotion and personality recognition in real-world tasks. arXiv preprint arXiv:2510.22129 (2025)
28. Jiang, J., Streli, P., Meier, M., Holz, C.: EgoPoser: Robust real-time egocentric pose estimation from sparse and intermittent observations everywhere. In: European Conference on Computer Vision (ECCV) (2024)
29. Jiang, J., Streli, P., Qiu, H., Fender, A., Laich, L., Snape, P., Holz, C.: Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In: European conference on computer vision. pp. 443–460. Springer (2022)
30. Kwak, I., Guo, J.Z., Hantman, A., Kriegman, D., Branson, K.: Detecting the starting frame of actions in video. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 489–497 (2020)
31. Li, Y., Nagarajan, T., Xiong, B., Grauman, K.: Ego-exo: Transferring visual representations from third-person to first-person videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6943–6953 (2021)

32. Li, Y.M., Huang, W.J., Wang, A.L., Zeng, L.A., Meng, J.K., Zheng, W.S.: Egoexo-fitness: Towards egocentric and exocentric full-body action understanding. In: European conference on computer vision. pp. 363–382. Springer (2024)
33. Li, Y.M., Wang, A.L., Zeng, L.A., Lin, K.Y., Tang, Y.M., Zheng, W.: Techcoach: Towards technical-point-aware descriptive action coaching. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 6699–6707 (2026)
34. Lin, C., Xu, C., Luo, D., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., Fu, Y.: Learning salient boundary feature for anchor-free temporal action localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3320–3329 (2021)
35. Lin, T., Liu, X., Li, X., Ding, E., Wen, S.: Bmn: Boundary-matching network for temporal action proposal generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3889–3898 (2019)
36. Lin, T., Zhao, X., Su, H., Wang, C., Yang, M.: Bsn: Boundary sensitive network for temporal action proposal generation. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
37. Liu, S., Sui, L., Zhang, C.L., Mu, F., Zhao, C., Ghanem, B.: Harnessing temporal causality for advanced temporal action detection. arXiv preprint arXiv:2407.17792 (2024)
38. Liu, S., Zhao, C., Zohra, F., Soldan, M., Pardo, A., Xu, M., Alssum, L., Ramazanova, M., Alcázar, J.L., Cioppa, A., et al.: Opentad: A unified framework and comprehensive study of temporal action detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2625–2635 (2025)
39. Liu, S., Huang, D., Wang, Y.: Adaptive nms: Refining pedestrian detection in a crowd. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6459–6468 (2019)
40. Liu, X., Wang, Q., Hu, Y., Tang, X., Zhang, S., Bai, S., Bai, X.: End-to-end temporal action detection with transformer. IEEE Transactions on Image Processing **31**, 5427–5441 (2022)
41. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
42. Parmar, P., Morris, B.T.: What and how well you performed? a multitask learning approach to action quality assessment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 304–313 (2019)
43. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
44. Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhania, D., Wang, R., Yao, A.: Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21096–21106 (2022)
45. Shi, D., Zhong, Y., Cao, Q., Ma, L., Li, J., Tao, D.: Tridet: Temporal action detection with relative boundary modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18857–18866 (2023)
46. Tang, T.N., Kim, K., Sohn, K.: Temporalmaxer: Maximize temporal context with only max pooling for temporal action localization. arXiv preprint arXiv:2303.09055 (2023)
47. Tang, Y., Ni, Z., Zhou, J., Zhang, D., Lu, J., Wu, Y., Zhou, J.: Uncertainty-aware score distribution learning for action quality assessment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9839–9848 (2020)

48. Wang, S., Yang, D., Zhai, P., Yu, Q., Suo, T., Sun, Z., Li, K., Zhang, L.: A survey of video-based action quality assessment. In: 2021 International conference on networking systems of AI (INSAI). pp. 1–9. IEEE (2021)
49. Wang, X., Kwon, T., Rad, M., Pan, B., Chakraborty, I., Andrist, S., Bohus, D., Feniello, A., Tekin, B., Frujeri, F.V., et al.: Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20270–20281 (2023)
50. Wu, C.H., Ashutosh, K., Grauman, K.: Skillsight: Efficient first-person skill assessment with gaze. arXiv preprint arXiv:2511.19629 (2025)
51. Xarles, A., Escalera, S., Moeslund, T.B., Clapés, A.: T-deed: Temporal-discriminability enhancer encoder-decoder for precise event spotting in sports videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3410–3419 (2024)
52. Xu, A., Zeng, L.A., Zheng, W.S.: Likert scoring with grade decoupling for long-term action assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3232–3241 (2022)
53. Xu, J., Yin, S., Zhao, G., Wang, Z., Peng, Y.: Fineparser: A fine-grained spatio-temporal action parser for human-centric action quality assessment. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 14628–14637 (2024)
54. Xu, M., Zhao, C., Rojas, D.S., Thabet, A., Ghanem, B.: G-tad: Sub-graph localization for temporal action detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10156–10165 (2020)
55. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: Vitpose: Simple vision transformer baselines for human pose estimation. Advances in neural information processing systems **35**, 38571–38584 (2022)
56. Yan, S., Xiong, X., Arnab, A., Lu, Z., Zhang, M., Sun, C., Schmid, C.: Multiview transformers for video recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3333–3343 (2022)
57. Yang, L., Zheng, Z., Han, Y., Cheng, H., Song, S., Huang, G., Li, F.: Dyfadet: Dynamic feature aggregation for temporal action detection. In: European conference on computer vision. pp. 305–322. Springer (2024)
58. Yu, X., Rao, Y., Zhao, W., Lu, J., Zhou, J.: Group-aware contrastive regression for action quality assessment. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7919–7928 (2021)
59. Zhang, C.L., Wu, J., Li, Y.: Actionformer: Localizing moments of actions with transformers. In: European Conference on Computer Vision. pp. 492–510. Springer (2022)
60. Zhao, C., Thabet, A.K., Ghanem, B.: Video self-stitching graph network for temporal action localization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13658–13667 (2021)
61. Zhou, K., Cai, R., Wang, L., Shum, H.P., Liang, X.: A comprehensive survey of action quality assessment: Method and benchmark. arXiv preprint arXiv:2412.11149 (2024)

A Computational Cost

Tab. 6 reports model size, FLOPs, throughput, and latency for each ablation configuration. Adaptive suppression adds negligible overhead (+0.07M parameters, <1 GFLOPs). Pose fusion introduces a separate pose backbone, increasing parameters by ∼13M and reducing FPS from 46.80 to 36.90 (Ego). Cross-view attention is the most expensive addition, bringing the full Ego+Exos model to 67.37M parameters and 19.70 FPS. FPS/latency reflect model inference only. They do not include video decoding, Omnivore feature extraction, and pose estimation.

Table 6: Computational cost per ablation configuration. Params in millions (M), latency (Lat.) in milliseconds. FPS and latency measured on a single NVIDIA H200 GPU with batch size 1.

	Ego				**Exo**				**Ego+Exos**			
Configuration	Params	GFLOPs	FPS	Lat.	Params	GFLOPs	FPS	Lat.	Params	GFLOPs	FPS	Lat.
ActionFormer	**33.18**	**32.18**	**49.50**	**20.20**	**40.26**	**46.67**	**49.30**	**20.30**	**42.62**	**51.50**	**49.30**	**20.30**
+ Adaptive Supp.	33.25	32.72	46.80	21.40	40.33	47.21	46.70	21.40	42.69	52.04	46.90	21.30
+ Pose Fusion	46.35	40.94	36.90	27.10	53.43	55.43	37.30	26.80	55.79	60.26	31.30	32.00
+ Cross-View Attn	–	–	–	–	–	–	–	–	67.37	77.38	19.70	50.80

B TadTR Learning Rate and Query Count

In contrast to feature pyramid methods (ActionFormer, TriDet, CausalTAD, *SkillSpotter*) where every temporal position serves as a candidate detection, DETR-based detectors use a fixed set of learned queries, each predicting one event, imposing a hard upper bound on the number of detections. Ego-Exo4D takes contain 42.3 annotations on average (median: 38, max: 266). The default query count of TadTR [40] is 40, which covers only 54.1% of takes. 100 queries cover 96.1% and 200 queries 99.7%. Tab. 7 ablates both the query count (40, 100, 200) and the learning rate (10^{-3} as used by all anchor-free models in our benchmark, and 10^{-4} following the standard DETR protocol).

The best detection result (lr=10^{-3}, 200 queries, Ego mAP_S: 7.79) remains far below ActionFormer (12.40) and our method (21.82), reported in Tab. 1. At lr=10^{-4}, performance degrades across all query counts, with 200 queries collapsing entirely (Ego mAP_A: 0.66).

C Single-View vs. Concatenated Features

The Ego-Exo4D baseline implementation [21] concatenates all available camera features into a single representation for the Exo and Ego+Exos views. For evaluation completeness, Tab. 8 compares this concatenation strategy against single-view processing, where each camera view is treated as an independent sample.

Table 7: TadTR [40] ablation over two different learning rates and three different number of queries. Takes contain 42.3 annotations on average; 40/100/200 queries cover 54.1%/96.1%/99.7% of takes.

		Ego				**Exo**				**Ego+Exos**			
lr	**Queries**	mAP_S	mAP_A	BA	F1	mAP_S	mAP_A	BA	F1	mAP_S	mAP_A	BA	F1
10^{-3}	40	5.02	6.59	49.84	49.60	6.10	8.61	50.53	50.51	**6.51**	**8.89**	51.03	50.80
10^{-4}	40	3.47	5.60	52.52	52.34	1.74	2.91	49.18	48.47	5.46	7.79	51.29	51.28
10^{-3}	100	6.12	8.57	52.22	51.74	**7.17**	**11.09**	50.67	50.38	3.53	5.33	49.59	46.95
10^{-4}	100	2.77	5.14	53.63	52.60	4.28	7.41	52.43	51.51	2.63	5.58	51.15	51.11
10^{-3}	200	**7.79**	**10.79**	49.68	48.03	6.37	10.48	52.31	51.25	4.12	6.74	52.81	48.23
10^{-4}	200	0.35	0.66	**55.73**	**55.64**	2.51	4.97	**55.03**	**54.92**	1.79	3.89	**54.69**	**54.67**

Concatenation improves detection mAP for most baselines but does not consistently improve classification. For our method, concatenation yields the highest Ego+Exos mAP_S (21.99) but BA drops (45.77), similar to TriDet [45] and ActionFormer [59]. Our introduced cross-view attention recovers BA to 60.39 with minimal detection loss.

Table 8: Single-view vs. concatenated features for Exo and Ego+Exos settings. Ego results are identical across configurations and omitted. In the single-view setting, each exo camera is treated as an independent sample.

		Exo				**Ego+Exos**			
Model	**View**	mAP_S	mAP_A	BA	F1	mAP_S	mAP_A	BA	F1
TadTR [40]	Single	9.09	13.57	51.07	50.89	8.26	12.69	50.72	50.11
TadTR [40]	Concat.	6.37	10.48	52.31	34.63	4.12	6.74	52.81	48.23
TriDet [45]	Single	10.42	13.93	53.33	52.37	12.38	16.10	55.32	55.15
TriDet [45]	Concat.	8.99	12.16	50.06	36.78	8.23	11.78	48.92	48.77
CausalTAD [37]	Single	8.08	10.14	53.05	52.32	8.15	9.87	52.04	51.41
CausalTAD [37]	Concat.	11.82	14.05	50.78	50.41	13.16	17.21	54.98	54.95
ActionFormer [59]	Single	9.49	11.77	51.31	50.91	8.15	10.21	53.33	53.00
ActionFormer [59]	Concat.	13.18	18.25	55.03	55.03	13.82	17.85	50.34	50.17
Ours	Single	20.33	**27.63**	56.78	56.76	20.48	27.91	55.93	55.86
Ours	Concat.	**21.12**	27.47	**60.59**	**60.55**	**21.99**	**28.04**	45.77	43.57
Ours	Cross-view	–	–	–	–	21.34	28.01	**60.39**	**59.37**

D Soft NMS Sigma Sweep

Tab. 9 reports ActionFormer [59] with Soft NMS at varying σ values. Detection metrics (mAP_A, mAP_S) decrease monotonically with increasing σ. No fixed σ

value can reach the performance when not using any NMS (No NMS). BA peaks at σ=0.5 (Ego: 55.34) but the margin over No NMS (54.81) is small.

Table 9: Soft NMS σ sweep on ActionFormer. All rows use the same trained model with different post-processing.

	Ego				**Exo**				**Ego+Exos**			
σ	mAP_S	mAP_A	BA	F1	mAP_S	mAP_A	BA	F1	mAP_S	mAP_A	BA	F1
No NMS	**13.96**	**22.50**	54.81	54.26	**15.42**	**24.47**	54.85	54.65	**17.30**	**24.12**	**51.79**	**51.40**
0.25	13.18	19.30	54.92	54.64	14.10	20.65	**54.96**	**54.91**	15.52	20.46	51.51	51.29
0.5	12.12	16.85	**55.34**	**55.24**	13.15	18.13	54.75	54.75	13.71	17.86	50.24	50.01
1.0	11.53	15.99	52.84	52.81	12.35	17.47	53.22	53.18	11.72	15.48	48.48	48.12
2.0	7.40	11.00	52.01	51.97	8.22	12.48	51.57	51.50	8.57	12.03	49.69	49.33
3.0	6.64	10.37	52.15	52.11	7.46	11.60	51.28	51.28	7.32	11.41	49.79	49.53

E Cross-Backbone Evaluation

We apply adaptive temporal suppression and gated pose fusion to all evaluated architectures except TadTR [40] to evaluate if our modules are specific to the ActionFormer backbone (see Tab. 3). We exclude TadTR because its DETR-style set prediction removes the suppression stage entirely, making adaptive temporal suppression inapplicable. All models use the same pre-extracted Omnivore features, training protocol, and evaluation settings as in the main paper.

Adaptive Suppression Consistently Improves Detection. Across all backbones and all view settings, adaptive suppression substantially improves mAP_S over both the Soft NMS and No NMS baselines. These gains are consistent with those observed on ActionFormer (see Tab. 2, main paper).

Pose Fusion's Effect on Grading Is Backbone- and View-Dependent. Unlike adaptive suppression, gated pose fusion does not uniformly improve balanced accuracy (BA) across all backbones. On CausalTAD and TriDet, pose fusion improves BA in every view setting (e.g., CausalTAD Ego: 54.01→59.34, +5.33; TriDet Exos: 57.54→60.56, +3.02). On the remaining three backbones, pose fusion slightly degrades single-view grading: VideoMambaSuite (Ego: 56.89→55.01; Exos: 57.59→54.46) and TemporalMaxer (Ego: 59.95→58.66; Exos: 61.27→59.66) both lose BA when pose is added on Ego and Exos, while gaining mAP_S/mAP_A. In Ego+Exos settings, however, pose fusion consistently recovers or improves grading even where it hurts single-view BA (e.g., TemporalMaxer Ego+Exos: 46.86→59.42).

Cross-View Attention Is Architecture-Dependent. Unlike ActionFormer, which suffers severe BA collapse under naive Ego+Exos concatenation (see Tab. 2, main paper), all of the evaluated backbones achieve reasonable Ego+Exos BA with adaptive suppression and pose fusion alone. Adding cross-view attention does not improve and can degrade performance in these cases. We attribute this to the fact that these architectures do not exhibit the classification collapse under multi-view concatenation that we observe with ActionFormer (see Tab. 2, Ego+Exos BA: 45.77 vs. 57.85 and 60.01). Since cross-view attention specifically addresses this collapse, we apply it only to ActionFormer, where the failure mode is present.

Table 10: Component ablation on the VideoMambaSuite [14] backbone. Each row progressively adds one module. * Uses cross-view attention for Ego+Exos.

	Ego				**Exos**				**Ego+Exos**			
Configuration	mAP_S	mAP_A	BA	F1	mAP_S	mAP_A	BA	F1	mAP_S	mAP_A	BA	F1
VideoMambaSuite (Soft NMS)	7.63	8.65	49.51	49.35	7.06	8.17	46.03	43.88	3.69	3.91	52.70	51.96
VideoMambaSuite (No NMS)	13.60	16.22	53.55	53.23	11.44	15.70	47.79	40.48	6.91	9.22	52.02	47.96
+ Adaptive Supp.	17.78	23.38	**56.89**	**56.87**	**20.08**	**26.62**	**57.59**	**57.57**	20.67	26.72	56.77	56.05
+ Pose	**21.39**	**27.11**	55.01	53.50	18.66	26.33	54.46	54.06	19.91	26.03	**56.92**	**56.90**
+ Cross-View Attn*	–	–	–	–	–	–	–	–	**21.51**	**28.48**	55.52	55.05

Table 11: Component ablation on the TriDet [45] backbone. Each row progressively adds one module. * Uses cross-view attention for Ego+Exos.

	Ego				**Exos**				**Ego+Exos**			
Configuration	mAP_S	mAP_A	BA	F1	mAP_S	mAP_A	BA	F1	mAP_S	mAP_A	BA	F1
TriDet (Soft NMS)	10.35	14.79	49.17	40.10	8.99	12.16	50.06	36.78	8.23	11.78	48.92	48.77
TriDet (No NMS)	11.80	17.80	52.03	50.56	11.47	15.45	48.53	48.49	9.46	14.71	48.83	48.25
+ Adaptive Supp.	18.84	25.98	58.89	57.30	19.03	23.85	57.54	57.15	**19.74**	23.94	53.37	53.10
+ Pose	**20.47**	**26.91**	**59.98**	**58.77**	**19.21**	**24.32**	**60.56**	**59.93**	19.33	**25.56**	**57.85**	**57.84**
+ Cross-View Attn*	–	–	–	–	–	–	–	–	18.70	23.63	54.08	53.61

F AQA Drop-In Classification Heads

Action quality assessment (AQA) methods regress a scalar quality score from pre-segmented clips but provide no localization. To test whether AQA heads perform better on skill action grading than our classification head, we replace *SkillSpotter*'s classification head with three AQA heads, adapting only the components required by our binary good/tip classification task: TechCoach [33], GDLT [52], and USDL [47]. Our default head wins on grading quality (BA, F1)

Table 12: Component ablation on the TemporalMaxer [46] backbone. Each row progressively adds one module. * Uses cross-view attention for Ego+Exos.

	Ego				Exos				Ego+Exos			
Configuration	mAP_S	mAP_A	BA	F1	mAP_S	mAP_A	BA	F1	mAP_S	mAP_A	BA	F1
TemporalMaxer (Soft NMS)	12.34	16.69	54.34	53.90	10.38	15.17	52.18	52.16	11.27	15.75	50.09	49.58
TemporalMaxer (No NMS)	14.76	22.83	53.89	53.12	11.56	20.53	52.57	52.55	13.20	22.42	52.07	50.46
+ Adaptive Supp.	**20.51**	25.90	**59.95**	**59.65**	**21.27**	25.82	**61.27**	**61.27**	19.56	26.36	46.86	44.73
+ Pose	20.15	**27.85**	58.66	57.48	20.65	**27.01**	59.66	59.66	**20.55**	**27.38**	**59.42**	**59.34**
+ Cross-View Attn*	–	–	–	–	–	–	–	–	19.61	27.46	58.38	56.57

Table 13: Component ablation on the CausalTAD [37] backbone. Each row progressively adds one module. * Uses cross-view attention for Ego+Exos.

	Ego				Exos				Ego+Exos			
Configuration	mAP_S	mAP_A	BA	F1	mAP_S	mAP_A	BA	F1	mAP_S	mAP_A	BA	F1
CausalTAD (Soft NMS)	11.42	16.07	52.86	52.07	11.82	14.05	50.78	50.41	13.16	17.21	54.98	54.95
CausalTAD (No NMS)	12.77	22.71	51.96	50.20	12.57	18.91	53.16	52.87	12.69	21.45	54.63	54.47
+ Adaptive Supp.	19.64	25.78	54.01	52.79	17.68	22.55	55.67	54.96	18.90	26.42	58.15	55.85
+ Pose	**21.36**	**27.46**	**59.34**	**59.23**	**20.76**	**25.81**	**60.53**	**60.36**	20.02	26.83	**60.01**	**59.51**
+ Cross-View Attn*	–	–	–	–	–	–	–	–	**23.29**	**28.41**	59.11	59.45

Table 14: Component ablation on the DyFADet [57] backbone. Each row progressively adds one module. * Uses cross-view attention for Ego+Exos.

	Ego				Exos				Ego+Exos			
Configuration	mAP_S	mAP_A	BA	F1	mAP_S	mAP_A	BA	F1	mAP_S	mAP_A	BA	F1
DyFADet (Soft NMS)	10.09	12.53	48.89	46.75	3.57	4.48	49.52	47.72	3.18	5.64	47.63	35.45
DyFADet (No NMS)	12.74	16.26	50.14	46.48	3.68	4.69	47.23	45.00	3.69	5.57	47.54	40.41
+ Adaptive Supp.	21.14	26.15	55.33	54.75	16.08	21.38	**54.36**	**52.17**	18.79	23.41	55.94	55.48
+ Pose	**21.56**	**28.68**	**56.77**	**56.51**	**20.16**	**25.22**	41.83	36.71	20.94	27.25	**58.86**	**58.49**
+ Cross-View Attn*	–	–	–	–	–	–	–	–	**23.18**	**28.62**	57.50	57.46

in all three view settings by a wide margin, while detection quality is slightly better using USDL in most settings. TechCoach's performance on grading is closest to ours on Ego (BA: 58.87), but is unstable across views, collapsing on Exos (BA: 43.11).

Table 15: *SkillSpotter* with drop-in AQA classification heads across all three view settings. Our default head achieves the best grading quality (BA, F1) in every setting; USDL achieves the best detection (mAP_S, mAP_A) in most settings despite trailing substantially in grading.

	Ego				**Exos**				**Ego+Exos**			
Head Type	mAP_S	mAP_A	BA	F1	mAP_S	mAP_A	BA	F1	mAP_S	mAP_A	BA	F1
TechCoach [33]	20.10	27.15	58.87	57.60	19.66	27.72	43.11	37.77	18.84	26.16	58.90	58.38
GDLT [52]	21.52	27.28	42.48	38.53	19.37	26.33	46.77	46.17	20.82	26.28	40.68	36.10
USDL [47]	21.44	**29.12**	49.99	49.14	**21.63**	**28.47**	50.87	50.03	**21.67**	**28.11**	54.78	54.60
Ours (default)	**21.82**	27.89	**60.40**	**60.02**	21.12	27.47	**60.59**	**60.55**	21.34	28.01	**60.39**	**59.37**

G Predicted vs. Ground Truth Pose

Tab. 16 compares three pose configurations: training and testing on predicted pose, training on ground truth with predicted pose at test time (our default), and training and testing on ground truth pose (oracle). Training on ground truth pose outperforms training on predicted pose, even when both test on predicted input (e.g., Ego mAP_S: 21.82 vs. 19.96), indicating that the model learns cleaner biomechanical representations from ground truth annotations. The gap between our default setting and the oracle is small across all configurations (e.g., Ego BA: 60.40 vs. 61.39, Δ=0.99), confirming that predicted pose quality is sufficient for effective skill grading during inference.

Table 16: Impact of pose quality during training and inference. Predicted ego pose via the official Ego-Exo4D baseline [21] (PA-MPJPE: 10.70 cm); predicted exo pose via ViTPose [55] with multi-view triangulation (PA-MPJPE: 17.98 cm). We report Procrustes-aligned MPJPE as our pose features (joint angles, distances, velocities) are invariant to the global coordinate frame.

		Ego				**Exo**				**Ego+Exos**			
Training	**Testing**	mAP_S	mAP_A	BA	F1	mAP_S	mAP_A	BA	F1	mAP_S	mAP_A	BA	F1
Predicted	Predicted	19.96	27.26	60.31	60.01	20.36	27.13	59.61	58.92	20.74	27.69	58.71	57.77
Ground truth	Predicted	21.82	27.89	60.40	60.02	21.12	27.47	60.59	60.55	21.34	28.01	60.39	59.37
Ground truth	Ground truth	**22.09**	**28.00**	**61.39**	**61.04**	**21.74**	**27.70**	**60.87**	**60.82**	**21.60**	**28.03**	**60.59**	**59.72**

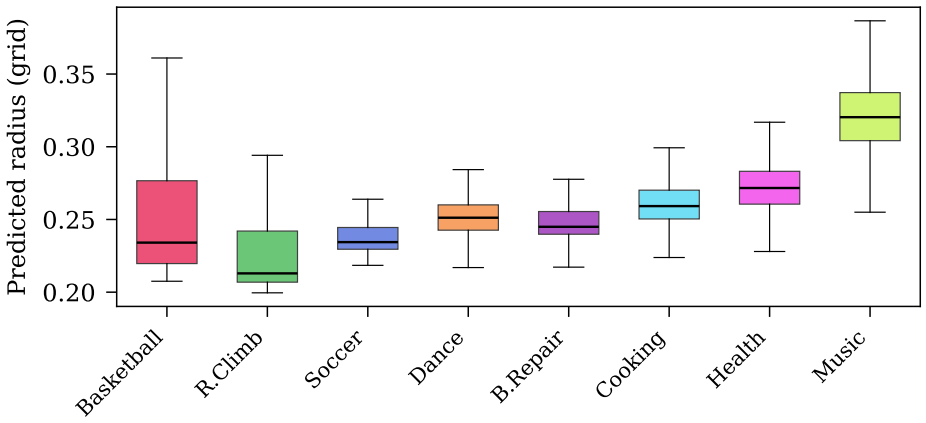

Fig. 5: Distribution of learned suppression radii per scenario. Scenarios with denser event distributions (Basketball, Rock Climbing) produce smaller radii, while sparse scenarios (Music, Health) produce larger radii. Box plots show median (orange line), interquartile range (box), and 1.5×IQR whiskers.

H Per-Scenario Radius Distributions

Fig. 5 shows the distribution of learned suppression radii across all eight scenarios. The suppression head learns scenario-specific radii without explicit supervision on the radius values—only the auxiliary suppression loss (Sec. 3.2) provides a training signal. Scenarios with sparse, well-separated events (Music, Health) produce larger median radii (0.32 and 0.27 grid units, respectively), allowing broader suppression of isolated false positives. Conversely, dense-event scenarios (Rock Climbing, Basketball) yield smaller median radii (0.21 and 0.23) but exhibit wider spread, reflecting the need for context-dependent radii: detections near temporal clusters require tight suppression to preserve neighbors, while isolated detections permit larger radii. This adaptive behavior explains the failure mode of fixed-radius NMS: a single radius cannot simultaneously avoid over-suppression in dense scenarios and under-suppression in sparse ones (see Fig. 3).

I Dataset Statistics

We report annotation statistics at two levels: the official Ego-Exo4D proficiency annotation files and the filtered evaluation database that our models are actually trained and evaluated on.

Official Annotations. Tab. 17 shows the per-scenario annotation counts directly from the official annotation files released by the Ego-Exo4D benchmark. The training split contains 556 takes with 11,942 annotations, and the test split contains 177 takes with 3,707 annotations. Good and bad labels are near-perfectly balanced across all scenarios (approximately 50/50).

Table 17: Per-scenario annotation statistics from the official Ego-Exo4D proficiency annotation files. Good/bad labels are near-perfectly balanced overall, with minor per-scenario deviations.

	Train				Test			
Scenario	Total	Good	Bad	Takes	Total	Good	Bad	Takes
Basketball	7,984	4,083	3,901	146	2,492	1,293	1,199	47
Bike Repair	1,448	704	744	41	191	107	84	9
Cooking	5,212	2,689	2,523	80	1,109	602	507	24
Dance	1,942	957	985	80	843	405	438	35
Health	1,804	764	1,040	42	547	237	310	12
Music	2,516	1,034	1,482	94	703	281	422	32
Rock Climbing	2,702	1,364	1,338	65	732	395	337	15
Soccer	685	261	424	8	380	137	243	3
All	24,293	11,856	12,437	556	6,997	3,457	3,540	177

Evaluation Database. Our preprocessing pipeline converts the official annotations into a unified database while applying several necessary filters: (1) takes whose take_uid is absent from the Ego-Exo4D metadata are dropped; (2) takes with anomalous frame rates (outside $[10, 100]$ fps) are excluded; and (3) views for which the pre-extracted Omnivore feature files are missing are skipped. These filters reduce the training set from 556 to 492 takes and the test set from 177 to 176 takes (Tab. 18). The small reduction confirms that the vast majority of officially annotated takes have complete metadata, valid frame rates, and available features. The dropped takes are predominantly missing pre-extracted Omnivore features for one or more views.

Table 18: Per-scenario annotation statistics as used in our evaluation pipeline. Takes without all required views or pre-extracted features are excluded from the official counts (Tab. 17).

	Train				Test			
Scenario	Total	Good	Bad	Takes	Total	Good	Bad	Takes
Basketball	6,899	3,577	3,322	129	2,409	1,260	1,149	46
Bike Repair	1,311	643	668	34	191	107	84	9
Cooking	4,605	2,368	2,237	69	1,109	602	507	24
Dance	1,813	892	921	75	843	405	438	35
Health	1,473	633	840	35	547	237	310	12
Music	2,080	826	1,254	83	703	281	422	32
Rock Climbing	2,445	1,217	1,228	61	732	395	337	15
Soccer	527	221	306	6	380	137	243	3
All	21,153	10,377	10,776	492	6,914	3,424	3,490	176