SkillSight: Efficient First-Person Skill Assessment with Gaze

Chi Hsuan Wu, Kumar Ashutosh, Kristen Grauman
Univeristy of Texas at Austin

Abstract

Egocentric perception on smart glasses could transform how we learn new skills in the physical world, but automatic skill assessment remains a fundamental technical challenge. We introduce SkillSight for power-efficient skill assessment from first-person data. Central to our approach is the hypothesis that skill level is evident not only in how a person performs an activity (video), but also in how they direct their attention when doing so (gaze). Our two-stage framework first learns to jointly model gaze and egocentric video when predicting skill level, then distills a gaze-only student model. At inference, the student model requires only gaze input, drastically reducing power consumption by eliminating continuous video processing. Experiments on three datasets spanning cooking, music, and sports establish, for the first time, the valuable role of gaze in skill understanding across diverse real-world settings. Our SkillSight teacher model achieves state-of-the-art performance, while our gaze-only student variant maintains high accuracy using 73× less power than competing methods. These results pave the way for in-the-wild AI-supported skill learning.

1. Introduction

Egocentric perception is poised to transform AI assistants on smart glasses which, by seeing through the eyes of a user, could provide in-the-moment contextually relevant information and recommendations. Of particular interest are assistants to support learning new skills in various domains such as exercise, sports, cooking, and music [4, 20, 25, 31, 37, 38, 64, 92, 98]. *Skill assessment*—the task of quantifying the degree of skill exhibited in a given execution—plays a crucial role: it would enable timely support [64], tracking of personal progress [22], and identifying areas for improvement [99]. Across these capabilities and more, skill assessment has the potential to personalize learning and enhance user performance in real-world tasks. Meanwhile, the portability of wearable glasses opens up seamless in-the-wild capture even for dynamic physical activities that go well beyond lab environments—e.g., the soccer pitch, the dance floor, or basketball court.

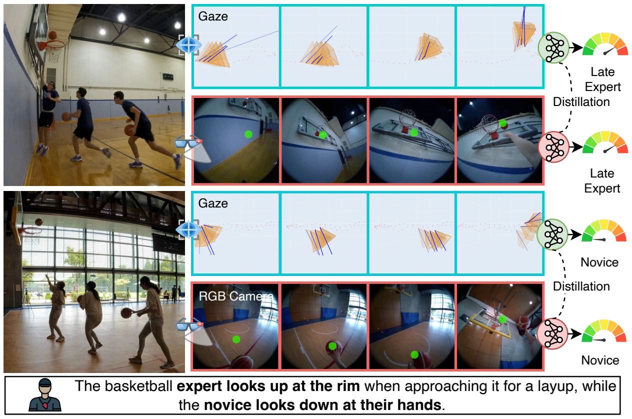


Figure 1. **Skill assessment with gaze.** Experts and novices exhibit distinct attention behaviors, influencing both how they move their head and eyes and what they see, as illustrated here with clips from an expert (top) and novice (bottom) basketball layup from [31]. The proposed method explores the associations between gaze, action, and expertise to achieve accurate and power-efficient skill assessment, using either ego-video and gaze, or gaze alone. The blue ray indicates gaze direction and depth, while shading shows camera motion over past frames. Note: leftmost third-person timelapses and commentary text are for illustration only.

However, prior research on skill assessment primarily relies on third-person visual perspectives of a subject's body poses [4, 12, 63, 65, 102], assuming prior setup of camera(s) in each target environment. Only limited work considers skill assessment from an egocentric perspective [6, 20, 31, 37], and there the low visibility of the camera-wearer's full body remains a critical challenge outside of table-top settings. Furthermore, the high power consumption of continuous video recording is an obstacle for vision-based methods—at odds with application needs for real-time, interactive skill learning.

Among the sensing modalities on smart glasses, we hypothesize that *gaze* is uniquely informative for assessing skill. Gaze complements vision: together, they reveal not only what the user is attending to, but also their intention [51, 104]. This synergy exposes fine-grained execution details that cameras alone cannot capture. In cognitive science, it is well known that people often fixate on objects they intend to manipulate or evaluate [30], while in domains

---

[1] Project page: https://vision.cs.utexas.edu/projects/skillsight/

as broad as sports [48], surgery [27], and music [40], experts display distinctive gaze patterns that enable them to execute complex motor actions more skillfully. For example, volleyball experts fixate earlier on the ball's contact point with their arms compared to novices [48], while skilled soccer players allocate more gaze to their surroundings while handling the ball [1], and the final steady fixation of the *quiet eye* is a signature not only of skilled athletes [85] but also skilled surgeons [15, 87], drivers [86], and musicians [21].

Could incorporating gaze into AI skill assessment provide such access to the cognitive and motor processes underlying an individual's actions, allowing more accurate estimates? To this end, we introduce SkillSight, a two-stage multimodal learning framework for first-person data. First, we train a teacher model SkillSight-T that integrates egocentric video and gaze to capture skill-related features. SkillSight-T generalizes across in-the-wild scenes by modeling interactions between gaze and action, encoding object fixations and transitions from gaze-cropped images, and modeling the dynamic gaze patterns. In the second stage, we train a student model SkillSight-S that relies *only* on gaze as input and keeps the camera off during inference—significantly reducing power consumption, while also increasing user privacy. To connect action, skill, and gaze, we train SkillSight-S via knowledge distillation, transferring visual information from SkillSight-T into gaze. Gaze signals encode spatial and temporal patterns of attention (e.g., fixations, saccades) that correlate closely with visual cues in egocentric video, enabling SkillSight-S to infer skill-related features without RGB input.

We evaluate our method on three datasets (Ego-Exo4D [31], Multisense Badminton [76], Expert-Novice Soccer [2]) spanning cooking, music, and various sports. SkillSight-T outperforms previous video-based methods by 5% (10% relative). SkillSight-S, which relies solely on gaze, performs competitively while consuming 14× to 73× less power, and outperforming existing methods aimed at efficiency [24, 69, 81]. Beyond performance, we provide quantitative and qualitative analyses revealing when and how gaze reflects skill. Together, these results highlight gaze as a powerful cue for scalable skill assessment.

Overall, we pioneer skill assessment using gaze signals across diverse domains and dynamic in-the-wild scenarios involving significant subject motion across the scene (e.g., climbing a boulder or dribbling to the basket for a layup, as opposed to tabletop activities). We are the first to explore power-efficient, privacy-preserving egocentric skill assessment, paving the way for practical deployment on resource-constrained smart glasses. Moreover, our analysis reveals how model predictions align with and even enhance established psychological theories, offering new quantitative, data-driven insights into complex gaze–skill relationships.

2. Related Work

Egocentric video and gaze. Gaze complements egocentric video by revealing attention and intention. Prior work predicts gaze from the ego view to model decision-making [35, 44, 45, 50] and leverages gaze for tasks such as action recognition [51, 101], motion anticipation [3, 61, 104], privacy filtering [77], attended-object detection [17, 55], intention understanding [42, 70], error detection [56], and learning sports play [78]. However, all such work focuses on aligning gaze with actions rather than assessing performance quality. Skill assessment demands recognizing subtle behavioral differences, beyond simply identifying actions. We instead explore how discriminative gaze trajectories reveal expertise across diverse domains, establishing gaze as a reliable and scalable cue for skill.

Relation of gaze and skills in cognitive science. Existing psychology studies investigate the relationship between gaze patterns and everyday tasks [46], decision-making [30], goal-directed behavior [32], task difficulty [16], and anticipation of future procedural steps [79]. As discussed above, cognitive science research links gaze patterns to proficiency: in medicine, gaze helps assess diagnostic and surgery skills [10, 27, 57]; in sports, expert athletes demonstrate distinct gaze strategies [39, 48]. We take inspiration from their findings. Further, building on this foundation, our work enables large-scale, data-driven learning of gaze-skill relations in diverse in-the-wild settings, uncovering subtle patterns beyond controlled psychology studies.

Skill assessment. Prior work on skill assessment focuses on third-person pose analysis in fitness [68], skating [94], and diving [97]. In contrast, first-person perspectives captured by wearables offer cues for real-time feedback in hand-centric tasks [20, 37, 66, 91, 93, 105] or sports [6, 31]. Recent work further incorporates text [29, 53, 95, 103], audio [89, 96, 100], human skeletons [19, 23, 49], PPG [9], and IMU [2, 41]. To our knowledge, [37] is the only vision work estimating skill with gaze, and it is shown only on static tasks (cooking and lab work) where the subject remains stationary. Our approach instead extends gaze-based skill assessment to dynamic activities, shows broad applicability across settings, and introduces technical novelty to explicitly capture the gaze-action interplay.

Efficient methods for wearable devices. Power efficiency is critical in wearable devices. Prior work addresses it through adaptive power management [80], distributed computation [88], selectively sampling clips [11, 43], and lightweight model architectures [24, 52]. More relevant to our work, another direction reduces reliance on power-hungry video by using lighter modalities: audio can suggest when to process video frames [28, 54, 69], and IMU with sparse video frames is sufficient for action recognition [81]. Nevertheless, all these prior methods still depend on periodic visual input, requiring frequent camera toggling

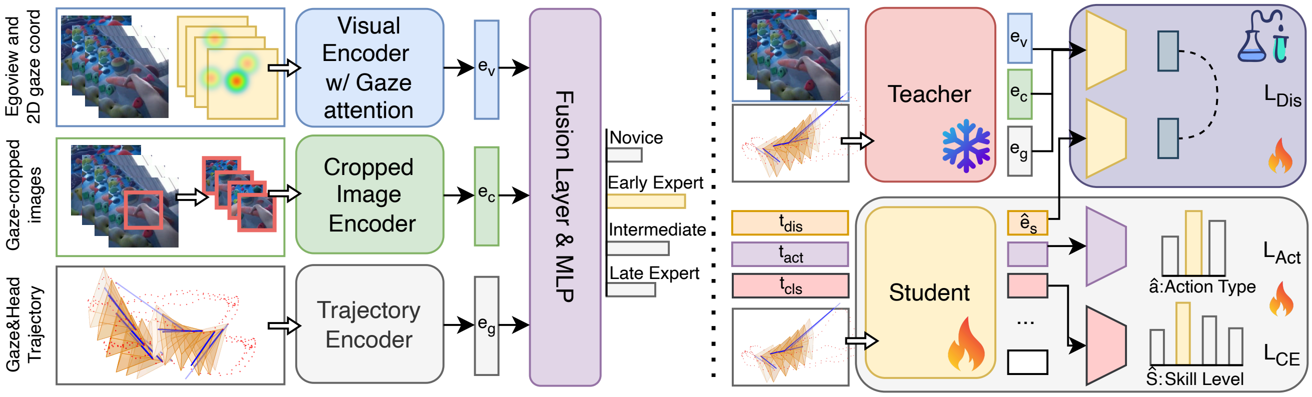


Figure 2. **Left: Overview of SkillSight-Teacher.** We incorporate three components that encode action and gaze correlation, attended object sequence, and gaze trajectory for skill assessment. These features are fused by the fusion layer for prediction. **Right: Overview of distillation method.** SkillSight-Student learns to distill knowledge from the teacher feature $[e_v, e_c, e_g]$ using the distillation token t_{dis}. As guidance for evaluating skill in context, the student model performs subtask recognition with the action recognition token t_{act}.

or low frame rate operation, which undermines both hardware simplicity and power efficiency due to startup latency [58] and transient power spikes when switching on the camera [5, 47]. In contrast, we distill visual supervision during training but use only gaze at inference, removing the need for camera input and substantially lowering sensing and model power, as we will quantify in results.

3. Method

We formally define the problem statement (Sec. 3.1), then introduce our model (Sec. 3.2 and 3.3) and describe data and implementation details (Sec. 3.4).

3.1. Problem statement

Consider a dataset $\mathcal{E} = \{(V, G, S)\}$, where each $V_i = \{v_i^t\}_{t=1}^T$ is the egocentric video demonstration with its frames v_i^t, $G_i = \{g_i^t\}_{t=1}^T$ is the gaze pattern, and S_i is the skill-level of the demonstrator. Although skill is inherently complex, recent studies and datasets have introduced rigorous objective means to quantify skill [2, 31, 76], formalizing this research direction.

Consistent with current hardware, we suppose that the device records the glasses' rotation and translation, as well as the 3D gaze vector of each eye, from which we derive g_i^t, which includes the 3D fixation points, the 3D gaze direction relative to the center of two eyes, the 2D coordinate of the gaze projection on the egoview video $g_{2d} \in R^2$, the depth of the gaze, and the translation and quaternion rotation of the glass. Current devices efficiently estimate gaze with eye cameras, IR, EOG, and/or IMU; we detail data resources [2, 31, 76] in Sec. 3.4 and quantify power load in Sec. 4.

The goal of this work is to classify[1] the skill level S_i using modalities from the smart glasses. We consider two setups: (1) **Video+Gaze**: we leverage both video and gaze during training and inference. Formally, we aim to learn a function $\mathcal{F}_v(V, G) \rightarrow S$, and call this variant of our method SkillSight-T(eacher) (2) **Gaze-only**: Continuous camera recording is power consuming and impractical for long-duration use. To reduce the reliance on camera, we use both video and gaze during training but rely only on gaze during inference. Specifically, we aim to learn $\mathcal{F}_g(G) \rightarrow S$, and call this variant of our method SkillSight-S(tudent).

3.2. Teacher model: Skill from action and attention

First we train a classifier that takes both egocentric video (*what the subject is doing*) and gaze (*how they are attending to their surroundings*) for skill level classification. To ensure robust generalization across dynamic and static scenarios, SkillSight-T integrates gaze and visual signals through three components: (1) the interaction between the subject's actions and gaze regions by applying the gaze attention to the visual encoder; (2) the sequence of subject's attended objects by encoding the gaze-cropped images; and (3) the dynamics of the subject's gaze over time. Fig. 2 shows the overview, and each part is described next.

Action and gaze interaction We leverage g_{2d}^t to identify the gaze-attended region in v^t, and incorporate gaze information into the visual encoder f_V (e.g. TimeSformer [7]). By knowing where the subject is looking, the model learns skill assessment by capturing the correlations between visual focus and actions. Specifically, we introduce a gaze-induced attention map $A_g = \{A_g^t\}_{t=1}^T$ into the first spatial encoder $f_{V,0}$ of f_V. Let $X = \{X^t\}_{t=1}^T$ be the input of $f_{V,0}$. For each timestep t, $f_{V,0}$ spatially divides X^t into p^2 patches with size $L \times L$ and computes an attention map $A_v^t \in R^{p \times p}$. Next, we apply a Gaussian kernel centered at

---

[1] Similarly, one could formulate the task as regression to a real-valued score [82, 89, 97]. We target discrete classes to account for the granularity of expertise differences discernible by human judges [2, 8, 26, 67] and to align with multiple existing annotated datasets [31, 63, 76].

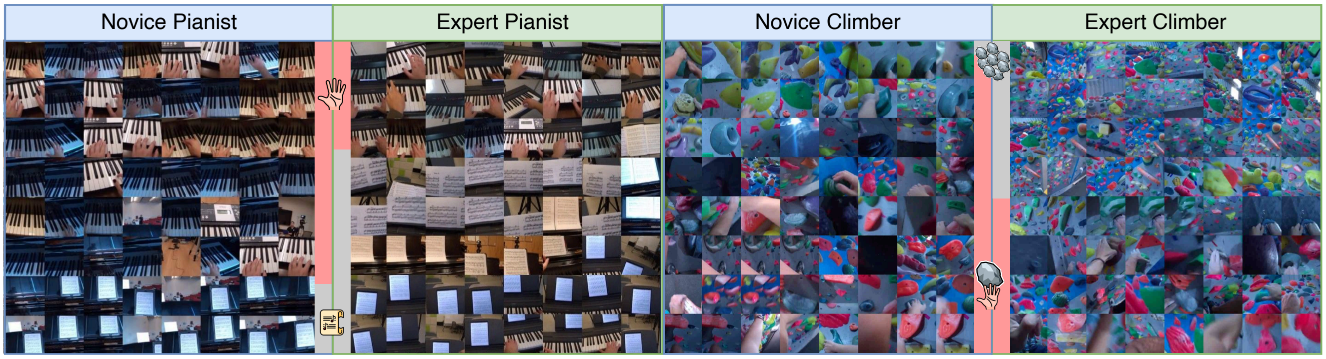


Figure 3. **What does an expert vs. novice tend to see more of?** In these distributions, each patch crops the egocentric frame based on the subject's gaze coordinates. Our representation surfaces interesting patterns, like (left two boxes) how novice pianists fixate on their hands more often than experts do (77% vs. 45%, as quantified with hand detection), or (right two boxes) how bouldering experts exhibit greater gaze depth (1.4 m vs. 1.1 m) as they analyze moves further up the wall, resulting in smaller rocks in the crops. These patterns emerging from in-the-wild video are consistent with and even deepen prior findings from psychology [13].

patch $c^t = \lfloor g_{2d}^t / L \rfloor$ and construct A_g^t with:

$$A_g^t[m,n] = \exp\left(-\frac{d_c^t(m,n)}{2\sigma^2}\right) / \sum_{m',n'} \exp\left(-\frac{d_c^t(m',n')}{2\sigma^2}\right), \tag{1}$$

with $d_c^t(m,n) = ||(m,n) - c^t||^2$. The modified attention map is

$$A_m^t = \sigma(A_v^t + \lambda_c A_g^t), \tag{2}$$

where σ is the softmax operation and λ_c is a learnable parameter for each scenario c, e.g., basketball, soccer. Finally, we obtain the embedding

$$e_v = f_V(V, g_{2d}). \tag{3}$$

Unlike prior gaze-based action recognition methods [36, 51, 59], which pool gaze information at late-stage features, our method emphasizes gaze in the earliest spatial encoder, allowing the model to semantically highlight gaze regions.

Attended object sequence We represent the subject's attended objects by spatially cropping v^t with g_{2d}^t. We observe that the distribution of attended objects for novices and experts differs significantly between tasks (see Fig. 3). For instance, novice pianists fixate on their hands more often than expert pianists, who dwell more on the sheet music. This observation motivates leveraging the sequence of gazed-upon objects to reflect skill.

While the sequence of attended objects is meaningful for skill assessment, we do not treat gaze-cropped image sequences $V_c = \{v_c^t\}_{t=1}^T$ as video [37] since the crops are taken from varying regions and lack spatial alignment across frames. Instead, we first compute semantic embeddings for v_c^t using a pretrained image encoder f_I, and a subsequent temporal encoder f_T models the sequence-level relationships, yielding the gaze-crop encoding:

$$e_c = f_T([f_I(v_c^1), ... f_I(v_c^T)]). \tag{4}$$

Gaze dynamics While 2D gaze and ego-view video [36, 37, 51, 59] highlight what a subject is looking at, they do not explicitly reflect the gaze dynamics such as the fixation frequency, the saccade speed, and the change of gaze location in the 3D environment—which all show significant differences across subjects with different skill levels [39, 48]. To that end, G_i contains rich 3D information about the trajectory of the subject, the gaze direction, and the gaze depth. We encode G_i using a transformer-based encoder f_g. To avoid bias in the gaze signals such as where the subject is facing, we normalize by calculating the gaze signals relative to the signals in the first frame. See Supp. for details and analysis. Formally, this yields our third component to encode the gaze dynamics:

$$e_g = f_g(G). \tag{5}$$

We concatenate the features from the three components and pass the combined feature to the fusion layer f_m for prediction. Specifically, we construct SkillSight-T as

$$\hat{S} = \mathcal{F}_v(V, G) = f_m([e_v, e_c, e_g]), \tag{6}$$

and use standard cross-entropy loss L_{CE} for training. Our modules reason about where and why the user is looking by explicitly modeling the spatial and semantic interaction between gaze and visual, capturing skill-related patterns more effectively than simply inputting raw gaze (see Supp.).

3.3. Student model: Distillation with gaze

Having defined the variant of our model that processes both gaze and video, next we generalize our approach to accommodate gaze alone—reducing power use and increasing privacy—without losing action-specific cues in video.

To this end, we propose SkillSight-S, a lightweight method that relies solely on gaze for skill assessment. With only gaze signals required at inference, the egocentric camera remains deactivated. As already discussed, cognitive science establishes a strong correlation between gaze behavior and skill level [16, 39, 48, 71]. Furthermore, eye-tracking cameras consume far less power [62, 75] than typical RGB cameras and mitigate privacy concerns since they only capture the user's eyes rather than the full environment. These properties make gaze a natural choice for power-efficient skill assessment.

But to what extent can video cues (what the user sees) be embedded *into* the gaze signal? Intuitively, people exhibit consistent gaze patterns when observing certain objects or performing specific actions, making it natural to distill visual information into gaze. This correlation is amplified in the skill assessment setting, where the subject's actions are aligned with the goal of the skilled activity (e.g., cooking a dish, shooting a free throw), take place in skill-conducive environments (e.g., a kitchen, gym), and involve interactions with specific skill-relevant objects (e.g., pot and whisk, basketball and hoop). These properties make our problem amenable to knowledge distillation.

SkillSight-S consists of a transformer-based encoder, f_s that takes G as input. We train f_s using knowledge distillation from the teacher $\mathcal{F}_v$ described in Sec. 3.2. We employ a distillation token, t_{dis} [83], to align the student features with those of the teacher. We also introduce an action recognition token, t_{act}, to classify the subject's subtask, e.g. dribbling, and penalty kick, based on G. This multi-branch architecture improves skill assessment by associating skilled gaze patterns with the subject's action. Specifically,

$$\hat{e}_s, \hat{S}, \hat{a} = f_s([t_{cls}, t_{dis}, t_{act}, G]) \tag{7}$$

where $\hat{a}$ predicts the subtask label, $\hat{S}$ predicts the skill level, and $\hat{e}_s$ is for distillation learning. The training objective of action classification is standard cross-entropy loss L_{act}. The distillation loss is computed as:

$$L_{dis} = ||f_p(\hat{e}_s) - f_t([e_v, e_c, e_g])||_1 \tag{8}$$

where f_p is a projection layer that aligns the features of $\mathcal{F}_g$ and $\mathcal{F}_v$, a common practice in knowledge distillation [74, 90]. We add another projection layer f_t to mitigate the impact of modality-specific teacher signals that the student cannot effectively capture. We set the loss weights λ_{dis} and λ_{act} with validation data and train the student model with:

$$L_{student} = L_{CE} + \lambda_{dis} L_{dis} + \lambda_{act} L_{act}. \tag{9}$$

3.4. Data and implementation details

Method and training details Following the Ego-Exo4D benchmark [31], we segment long videos into 10 equally spaced clips and average segment-level predictions for classification. Note that we use untrimmed videos without making strong assumptions about where the skilled portions of the sequence occur. To better model skill-relevant dynamics, we configure both the teacher and student models to process 16-frame clips at 2 FPS, balancing temporal coverage with computational efficiency. We use TimeSformer [7] pretrained on EgoVLPv2 [73] as f_V, achieving state-of-the-art egocentric video understanding, and Dinov2 [60] as f_I for its strong spatial representation. Both f_s and f_g are 4-layer transformer encoders with a 768-dimensional hidden size, and f_m is a 3-layer MLP. SkillSight-T is trained for 15 epochs using SGD (learning rate 5×10^{-3}, batch size 8), and SkillSight-S for 10 epochs using AdamW (learning rate 1×10^{-4}, batch size 32). All models are trained on 8 NVIDIA Quadro RTX 6000 GPUs. SkillSight-S processes a single sample in 1.6 ms on average using a single GPU.

Data sources and statistics We evaluate our method on three datasets. (1) **Ego-Exo4D** [31] consists of 5,048 videos recorded by 740 participants. We use all the scenarios provided with the demonstrator proficiency estimation benchmark: soccer, basketball, rock climbing, dance, music, and cooking. Following [9], we use 10% from the official training set for validation, and the held-out official validation set for testing. Each subject is annotated with one of four skill levels: novice, early expert, intermediate expert, and late expert. (2) **Multi-Sense Badminton** (MSB) [76] encompasses 7,763 badminton forehand and backhand swing data from 25 players. The skill levels are annotated into beginner, intermediate, and expert. We follow the official cross-validation split. (3) **Expert-Novice Soccer** [2] contains 288 recordings from 8 subjects performing 9 different soccer movements such as kicks, dribbling, and juggling. Subjects are labeled as expert and novice. We follow the official cross-validation.

These datasets were chosen because they have gaze, camera pose, and ground truth skill labels provided by expert annotators (e.g., domain-specific coaches and teachers). In total, the gaze is from 3 distinct wearable devices, reflecting today's good availability of this modality. Ego-Exo4D and Expert-Novice Soccer include 3D gaze, while MSB provides 2D gaze. Expert-Novice Soccer does not contain video; therefore, we train its teacher model using body motion (21 joint positions over time) and gaze. For all datasets, no subject overlaps between the train-test splits.

4. Experiment

We first describe baselines, followed by results and qualitative examples. Finally, we analyze power efficiency and our performance across different scenarios.

Baselines We compare to video action/skill recognition methods [7, 8, 24], methods using diverse modalities from glasses [69, 72, 81], and ego methods using gaze [37, 51]:

- **TimeSformer** [7], **X3D-XS** [24], **Skillformer** [8]: The first two are standard video-classification models. X3D-XS is an efficient architecture suitable for deployment

Method	Modalities	Power (mW)	**EgoExo4D** [31]							**MSB** [76]
			Overall	Soccer	Basketball	Bouldering	Music	Dance	Cooking	Badminton
Majority vote	—	—	32.3	74.4	35.7	0.0	44.0	43.3	50.9	41.1
E2GoMotion [72]	V	329.3	34.9	55.8	49.0	3.0	16.7	50.4	50.9	43.5
Skillformer [8]	V	697.5	42.4	74.4	42.0	27.0	47.2	43.3	**58.5**	44.0
TimeSformer [7]	V	697.5	45.5	76.7	53.2	28.0	36.1	44.8	56.6	50.5
EgoExoLearn [37]	V+G	141.4	42.3	74.4	46.9	25.2	44.4	43.3	50.9	31.7
Beholder [51]	V+G	132.4	34.1	72.1	42.7	21.4	50.0	26.8	24.5	30.6
SkillSight-T	V+G	943	**50.1**	**81.4**	**55.2**	**28.9**	**50.0**	**56.7**	**58.5**	**53.1**
X3D-XS [24]	V	88	34.2	72.1	**45.5**	24.5	38.9	26.8	17.0	42.7
EgoDistill [81]	V+I	16.5	42.6	74.4	35.0	**38.4**	50.0	43.3	43.4	43.4
EgoTrigger [69]	V+A	9.9	34.1	65.1	37.8	22.6	41.7	26.8	5.7	*no audio*
Gaze-only	G	9.5	37.0	76.7	25.2	31.5	44.4	40.2	39.6	42.3
SkillSight-S	G	9.5	**44.4**	**79.1**	42.0	34.6	**52.8**	**44.1**	**47.2**	**47.0**

Table 1. **Results on the Ego-Exo4D [31] (left) and Multi-Sense Badminton (MSB) [76] (right) benchmarks. Top section:** SkillSight-T outperforms all prior methods across all scenarios in terms of accuracy (%). **Bottom section:** SkillSight-S surpasses all power-efficient methods in overall accuracy (44.4%) as well as 5 of the 7 individual scenarios. Even when compared to the more expensive, power-consuming baselines (top section), SkillSight-S still ranks second in overall accuracy, while using 14× to 73× less power (mW). Bold face indicates best accuracy and underline indicates second-best. (V:Visual, G:Gaze, I:IMU, A:Audio).

Expert–Novice Soccer [2]		
Method	Modalities	Overall
Majority vote	—	50.0
Gaze-only	G	66.0
Body-motion-only	M	71.2
SkillSight-S	G	72.6
Body-motion+Gaze	M+G	**73.3**

Table 2. **Results on Expert-Novice Soccer [2].** Since the Expert-Novice Soccer does not include video, we use transformer baselines with full body motion (M) and eye-tracking gaze (G). SkillSight-S outperforms both the gaze-only and motion-only baselines, showing the effectiveness of our distillation technique.

on smart glasses, while TimeSformer represents the Ego-Exo4D baseline for proficiency estimation [31]. Skillformer builds on TimeSformer, fine-tuned via LoRA [34].

- **EgoDistill** [81], **EgoTrigger** [69]: EgoDistill is a power-efficient approach that processes a single RGB frame together with the corresponding sequence of IMU readings from the glasses for action recognition. EgoTrigger, similar to [28], reduces power consumption by leveraging audio cues to decide whether to process the visual stream.
- **E2GoMotion** [72]: The method leverages event-camera data for action recognition. Since no skill dataset contains event-camera recordings, we provide full-frame-rate optical flow to their model as a proxy, identical to the upper bound reported in their study.
- **EgoExoLearn** [37], **Beholder** [51]: EgoExoLearn crops ego-view video around gaze points and uses I3D [14] for skill classification, while Beholder performs gaze-weighted pooling of visual features for action recognition.
- **Gaze-only**: This method only takes gaze as input and shares the same architecture with SkillSight-S. We use cross-entropy loss for training without distillation.

Of all the baselines, only Skillformer [8] and EgoExoLearn [37] are specifically for skill assessment, and only EgoExoLearn utilizes gaze. Other models [7, 24, 51, 69, 72, 81] originally target action recognition; to broaden the pool of baselines, we adapt them for skill assessment by adjusting the output dimension and training on the same datasets. X3D-XS [24], EgoDistill [81], and EgoTrigger [69] are power-efficient methods leveraging less computation or lightweight modalities. We evaluate using standard accuracy metrics and estimated power consumption.

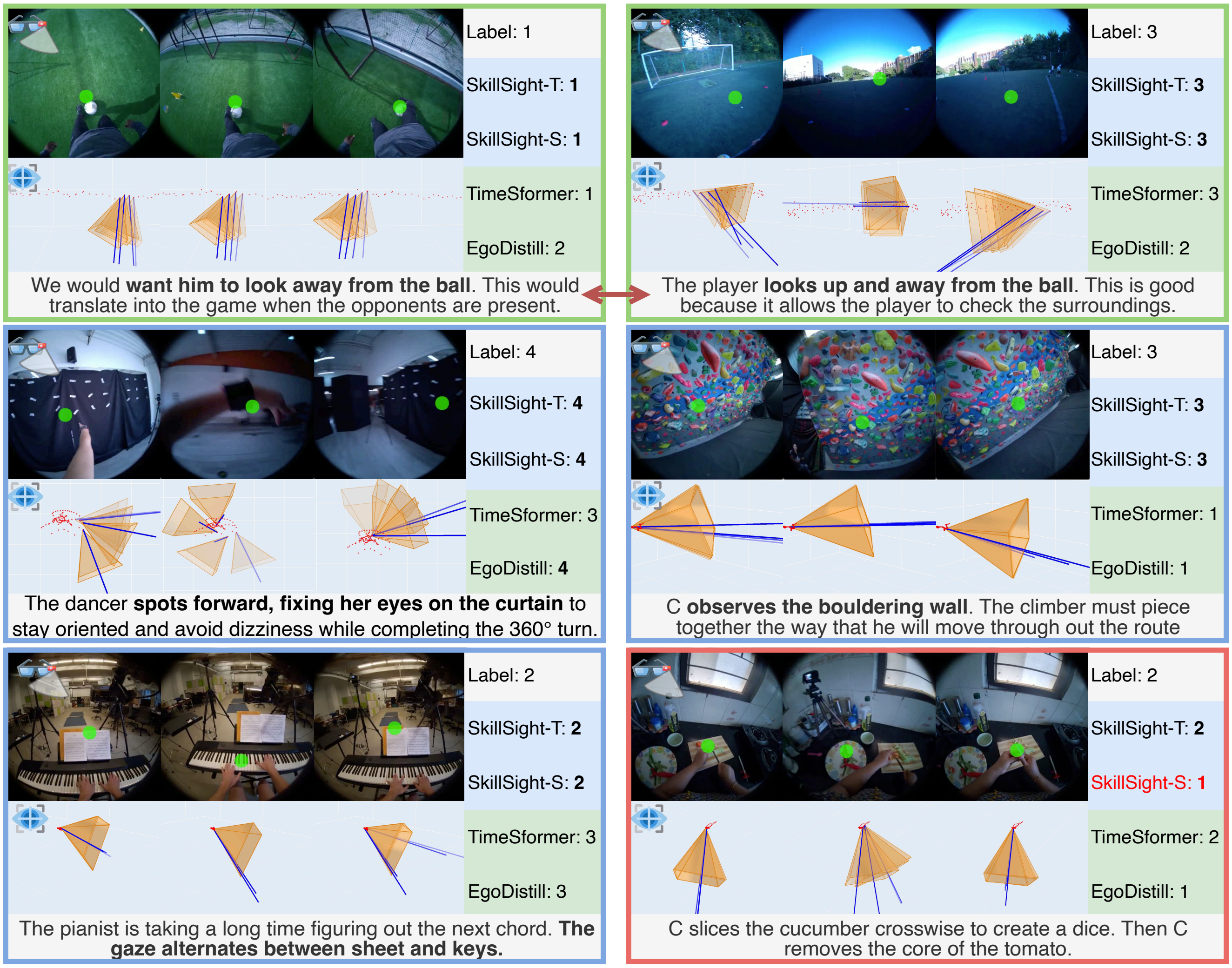
Results Table 1 reports results on Ego-Exo4D [31] and Multisense Badminton [76]. SkillSight-T outperforms all baselines across seven scenarios in both datasets, achieving an average relative gain of 10% over the strongest baseline. Remarkably, SkillSight-S, which uses only gaze as input, outperforms not only all the power-efficient baselines (bottom), but also the majority of the power-hungry baselines—despite using 14× to 73× less power (details below). It also achieves the best performance among power-efficient baselines in five of seven individual scenarios.[2]

Notably, SkillSight-T is superior in both static scenes, i.e. cooking and music, and dynamic sports, i.e. soccer, basketball, dance, rock climbing, dancing, and badminton. We attribute the robust prediction to our designs for incorporating gaze with vision, allowing the model to learn from the attended objects, the actions, and the gaze transition. We show that SkillSight-T outperforms a naive end-to-end model by 8% as well as more **ablations** in Supp.

Despite having the lowest power consumption, SkillSight-S outperforms other power-efficient baselines (Tab. 1, bottom). We see a significant improvement

---

[2] EgoPPG [9] reports its performance on a modified EgoExo4D test set, which is not directly comparable to the results in Tab. 1. On the modified test set, SkillSight-T outperforms EgoPPG by 11% relative. SkillSight-S exceeds EgoPPG by 0.5% relative and uses significantly less power.



We would **want him to look away from the ball**. This would translate into the game when the opponents are present.

The player **looks up and away from the ball**. This is good because it allows the player to check the surroundings.

The dancer **spots forward, fixing her eyes on the curtain** to stay oriented and avoid dizziness while completing the 360° turn.

C **observes the bouldering wall**. The climber must piece together the way that he will move through out the route

The pianist is taking a long time figuring out the next chord. **The gaze alternates between sheet and keys.**

C slices the cucumber crosswise to create a dice. Then C removes the core of the tomato.

Figure 4. **Qualitative results.** Both SkillSight-T and SkillSight-S better predict skill level than prior work. Experts and novices show distinct gaze patterns consistent with Ego-Exo4D [31] expert commentaries, shown for reference but not used by any model. The last example (bottom right) shows a failure case, highlighting the challenge of assessing skill from subtle movements. Blue rays show gaze direction and depth, and frustrum/ray shading indicates recent glasses motion. Ground-truth labels range from 1 (novice) to 4 (late expert).

compared to the Gaze-only baseline. This shows that SkillSight-S effectively learns the knowledge of SkillSight-T through our distillation technique. Models that rely only on first-person visual input, e.g., X3D-XS [24], fail to learn consistent skill patterns across scenarios. EgoDistill [81] and EgoTrigger [69] use a single frame together with head rotation or audio to represent the subject's action; however, these modalities struggle to reveal subtle differences in actions for rating skill. On the other hand, gaze directly captures how subjects actively shift attention to complete tasks. This highlights gaze as a compact, highly informative signal for low-power skill assessment.

We present qualitative results in Fig. 4. We see that across different scenarios, experts and novices demonstrate different gaze patterns. For instance, when dribbling in soccer (first row), the novice looks down on the ball while the expert looks away from the ball to check the surroundings. When expert dancers perform a spin (middle left), they fixate their eyes early to the front to avoid dizziness. These patterns show important cues that our methods leverage to access skills robustly, showing the benefit of explicitly modeling multiple aspects of gaze and skill together. By contrast, TimeSformer [7] and Skillformer [8]—neither of which uses gaze—struggle when subjects exhibit few motion cues. For example, an ego-view clip alone may not reveal that a performer shifts gaze from sheet music to their hands (bottom left), offering limited cues for skill assessment. EgoExoLearn [37] and Beholder [51] restrict processing to visual regions around gaze. While this approach is effective when gaze remains on the hands, it discards valuable contextual information when gaze shifts away from the body. For example, in bouldering (middle right), they may focus on the

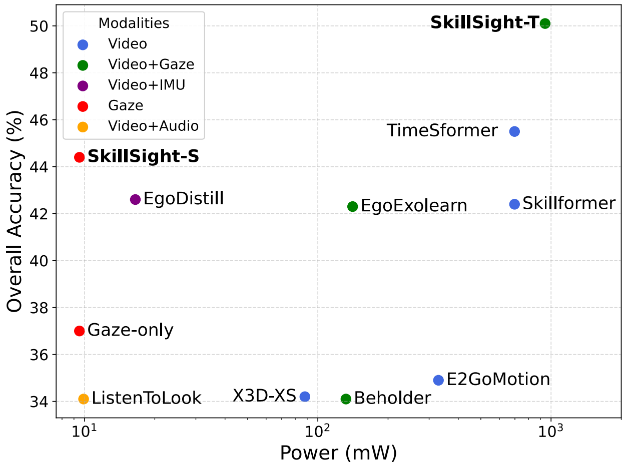


Figure 5. **Power–accuracy tradeoff**. SkillSight-T outperforms all baselines in accuracy, while SkillSight-S achieves the second-best accuracy and consumes the least energy. The optimal method would attain maximal accuracy with minimal power (top left).

wall. Prior approaches that limit attention to the gaze region therefore overlook cues critical for assessing skill. Finally, we show a failure case where gaze does not reflect skill when the subject is slicing vegetables (bottom right), showcasing the limitation of gaze when subtle hand movements are required.

Table 2 reports results on Expert-Novice Soccer [2]. They highlight the effectiveness of our distillation framework. SkillSight-S, using only smart-glasses signals, surpasses both Gaze-only and Body-motion-only baselines, the latter of which requires subjects to wear body-mounted IMUs. Across all datasets, our method enhances gaze-based models and achieves competitive, power-efficient performance suitable for skill assessment on smart glasses.

Efficiency analysis. Accurately measuring power consumption for wearable device applications is crucial. Using well-established measurements [31], the total energy consumption can be divided into three components: sensor triggering energy (γ), compute energy (α), and memory transfer energy (β). See Supp. for full explanation. We employ weighting parameters based on real-world estimates of the power consumption. Specifically, $\alpha = 4.6$ pJ/MAC [18], $\beta = 80$ pJ/byte [33], $\gamma_{\text{rgb}} = 35$ mW, $\gamma_{\text{IMU}} = 1.2$ mW, $\gamma_{\text{audio}} = 0.3$ mW [62], and $\gamma_{\text{eye}} = 7.8$ mW [75]. All values are taken from hardware designed for smart glasses.

The overall energy consumption rate of a model is:

$$P = \frac{\alpha N}{T} + \frac{\beta B}{T} + \sum_{m} \gamma_m \delta_m, \tag{10}$$

where N is the number of MACs in the model forward pass, B is the number of bytes required for read/write operations, m indexes the modalities used by the model, $\delta_m = 1$ when the model uses modality m, and 0 otherwise. T is the time interval between successive inferences.

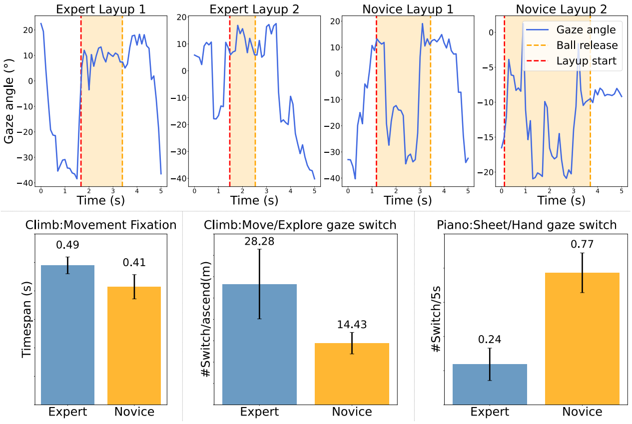


Figure 6. **Gaze pattern analysis.** SkillSight-S reveals distinct gaze patterns between model-predicted experts and novices.

Figure 5 shows that SkillSight-S achieves the best overall trade-off between power consumption and accuracy. It outperforms all power-efficient baselines while reducing the power consumption of the best baseline, i.e. EgoDistill [81], by 43%. Moreover, SkillSight-S demonstrates competitive performance compared to video-based methods, which are power intensive *regardless of the architecture* due to the energy cost of sensing and visual feature encoding. Compared to TimeSformer [7], SkillSight-S achieves over 73× lower energy cost with only a 1.1% drop in accuracy. Our approach provides an efficient foundation for real-time assistance or skill assessment.

Psychology insight from SkillSight. Figures 6 and 3 show gaze behavior insights from SkillSight-S. In basketball layups, model-predicted experts consistently look up toward the rim, while novices look down at the ball (top). In bouldering, our predicted experts show longer movement-related fixations (e.g., grasp or foot placement) (bottom left), consistent with sports science [84]. Beyond that, experts switch more often between movement-related and exploratory fixations when ascending (bottom middle). Figure 3 shows that novice pianists focus more on the hands, aligning with psychology findings [13], while SkillSight further shows more frequent gaze transitions between the sheet and hands (Fig. 6, bottom right). SkillSight not only aligns with established psychological findings, it also facilitates finer exploration of expert-novice gaze strategies.

5. Conclusion

We investigate how gaze behavior reflects skill level across dynamic and static scenarios. Our methods integrate gaze with egocentric visuals to assess skill by modeling attention during task execution. Moreover, our distillation framework enables a lightweight model using only gaze, achieving competitive accuracy while using significantly less power. Our work lays the foundation for future AI-driven instructional and assistive systems on smart glasses.

Acknowledgement

Research supported in part by a gift from Amazon and the UT Austin IFML NSF AI Institute. We thank Zihui Xue for valuable advice on head pose representation and normalization process, and the members of the UT Austin Computer Vision Group for helpful discussions.

References

[1] Visual strategies of young soccer players during a passing test – a pilot study. *Journal of Eye Movement Research*, 15 (1), 2022. 2

[2] Yusuke Akamatsu, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama. Classification of expert-novice level using eye tracking and motion data via conditional multimodal variational autoencoder. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1360–1364. IEEE, 2021. 2, 3, 5, 6, 8

[3] Taravat Anvari, Markus Lappe, and Marc H E de Lussanet. Where does gaze lead? integrating gaze and motion for enhanced 3d pose estimation. In *2025 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 76–83, 2025. 2

[4] Kumar Ashutosh, Tushar Nagarajan, Georgios Pavlakos, Kris Kitani, and Kristen Grauman. Expertaf: Expert actionable feedback from video. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13582–13594, 2025. 1

[5] Alpha Yaya Balde, Emmanuel Bergeret, Denis Cajal, and Jean-Pierre Toumazet. Low power environmental image sensors for remote photogrammetry. *Sensors*, 22(19):7617, 2022. 3

[6] Gedas Bertasius, Hyun Soo Park, Stella X Yu, and Jianbo Shi. Am i a baller? basketball performance assessment from first-person videos. In *Proceedings of the IEEE international conference on computer vision*, pages 2177–2185, 2017. 1, 2

[7] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Icml*, page 4, 2021. 3, 5, 6, 7, 8

[8] Edoardo Bianchi and Antonio Liotta. Skillformer: Unified multi-view video understanding for proficiency estimation, 2025. 3, 5, 6, 7

[9] Björn Braun, Rayan Armani, Manuel Meier, Max Moebus, and Christian Holz. egoppg: Heart rate estimation from eye-tracking cameras in egocentric systems to benefit downstream vision tasks. *arXiv preprint arXiv:2502.20879*, 2025. 2, 5, 6

[10] Tad T Brunyé, Trafton Drew, Donald L Weaver, and Joann G Elmore. A review of eye tracking for understanding and improving diagnostic interpretation. *Cognitive research: principles and implications*, 4(1):7, 2019. 2

[11] Shyamal Buch, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. Flexible frame selection for efficient video reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29071–29082, 2025. 2

[12] James Burgess, Xiaohan Wang, Yuhui Zhang, Anita Rau, Alejandro Lozano, Lisa Dunlap, Trevor Darrell, and Serena Yeung-Levy. Video action differencing. *arXiv preprint arXiv:2503.07860*, 2025. 1

[13] Michel A. Cara. The effect of practice and musical structure on pianists' eye-hand span and visual monitoring. *Journal of Eye Movement Research*, 16(2):1–18, 2023. 4, 8

[14] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 6

[15] Joe Causer, Adam Harvey, Richard Snelgrove, Gary Arsenault, and Oshin Vartanian. Quiet eye training improves surgical performance: A randomized controlled study. *Frontiers in Psychology*, 5:821, 2014. 2

[16] Longfei CHEN, Yuichi NAKAMURA, Kazuaki KONDO, Dima DAMEN, and Walterio MAYOL-CUEVAS. Integration of experts' and beginners' machine operation experiences to obtain a detailed task model. *IEICE TRANSACTIONS on Information*, E104-D(1):152–161, 2021. 2, 5

[17] Dima Damen, Teesid Leelasawassuk, Osian Haines, Andrew Calway, and Walterio W Mayol-Cuevas. You-do, i-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video. In *BMVC*, page 3, 2014. 2

[18] Radosvet Desislavov, Fernando Martínez-Plumed, and José Hernández-Orallo. Trends in ai inference energy consumption: Beyond the performance-vs-parameter laws of deep learning. *Sustainable Computing: Informatics and Systems*, 38:100857, 2023. 8

[19] Linfeng Dong, Wei Wang, Yu Qiao, and Xiao Sun. Lucidaction: A hierarchical and multi-model dataset for comprehensive action quality assessment. *Advances in Neural Information Processing Systems*, 37:96468–96482, 2024. 2

[20] Hazel Doughty, Walterio Mayol-Cuevas, and Dima Damen. The pros and cons: Rank-aware temporal attention for skill determination in long videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7862–7871, 2019. 1, 2

[21] Veronique Drai-Zerbib and Emmanuel Baccino. The influence of expertise in music reading on the detection of temporal violations. *Visual Cognition*, 20(3):267–282, 2012. 2

[22] Tobias Drey, Pascal Jansen, Fabian Fischbach, Julian Frommel, and Enrico Rukzio. Towards progress assessment for adaptive hints in educational virtual reality games. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, page 1–9, New York, NY, USA, 2020. Association for Computing Machinery. 1

[23] Amr Elkholy, Mohamed E Hussein, Walid Gomaa, Dima Damen, and Emmanuel Saba. Efficient and robust skeleton-based quality assessment and abnormality detection in human action performance. *IEEE journal of biomedical and health informatics*, 24(1):280–291, 2019. 2

[24] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the*

IEEE/CVF conference on computer vision and pattern recognition, pages 203–213, 2020. 2, 5, 6, 7

[25] Shijia Feng, Michael Wray, and Walterio Mayol-Cuevas. Evostruggle: A dataset capturing the evolution of struggle across activities and skill levels. *arXiv preprint arXiv:2510.01362*, 2025. 1

[26] Isabel Funke, Sören Torge Mees, Jürgen Weitz, and Stefanie Speidel. Video-based surgical skill assessment using 3d convolutional neural networks. *International Journal of Computer Assisted Radiology and Surgery*, 14(7):1217–1225, 2019. 3

[27] Soline Galuret, Nicolas Vallée, Alexandre Tronchot, Herve Thomazeau, Pierre Jannin, and Arnaud Huaulmé. Gaze behavior is related to objective technical skills assessment during virtual reality simulator-based surgical training: a proof of concept. *International Journal of Computer Assisted Radiology and Surgery*, 18(9):1697–1705, 2023. 2

[28] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10457–10467, 2020. 2, 6

[29] Kumie Gedamu, Yanli Ji, Yang Yang, Jie Shao, and Heng Tao Shen. Visual-semantic alignment temporal parsing for action quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2

[30] Kerstin Gidlöf, Annika Wallin, Richard Dewhurst, and Kenneth Holmqvist. Using eye tracking to trace a cognitive process: Gaze behaviour during decision making in a natural environment. *Journal of eye movement research*, 6 (1), 2013. 1, 2

[31] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, 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*, pages 19383–19400, 2024. 1, 2, 3, 5, 6, 7, 8

[32] Mary M Hayhoe and Jonathan Samir Matthis. Control of gaze in natural environments: effects of rewards and costs, uncertainty and memory in target selection. *Interface focus*, 8(4):20180009, 2018. 2

[33] Mark Horowitz. 1.1 computing's energy problem (and what we can do about it). In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, pages 10–14, 2014. 8

[34] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 6

[35] Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *Proceedings of the European conference on computer vision (ECCV)*, pages 754–769, 2018. 2

[36] Yifei Huang, Minjie Cai, Zhenqiang Li, Feng Lu, and Yoichi Sato. Mutual context network for jointly estimating egocentric gaze and action. *IEEE Transactions on Image Processing*, 29:7795–7806, 2020. 4

[37] Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, et al. Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22072–22086, 2024. 1, 2, 4, 5, 6, 7

[38] Mina Huh, Zihui Xue, Ujjaini Das, Kumar Ashutosh, Kristen Grauman, and Amy Pavel. Vid2coach: Transforming how-to videos into task assistants. *arXiv preprint arXiv:2506.00717*, 2025. 1

[39] Inhyeok Jeong, Kento Nakagawa, Rieko Osu, and Kazuyuki Kanosue. Difference in gaze control ability between low and high skill players of a real-time strategy game in esports. *PloS one*, 17(3):e0265526, 2022. 2, 4, 5

[40] Jakob Karolus, Johannes Sylupp, Albrecht Schmidt, and Paweł W Woźniak. Eyepiano: leveraging gaze for reflective piano learning. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, pages 1209–1223, 2023. 2

[41] Aftab Khan, Sebastian Mellor, Rachel King, Balazs Janko, William Harwin, R Simon Sherratt, Ian Craddock, and Thomas Plötz. Generalized and efficient skill assessment from imu data with applications in gymnastics and medical training. *ACM Transactions on Computing for Healthcare*, 2(1):1–21, 2020. 2

[42] Robert Konrad, Nitish Padmanaban, J Gabriel Buckmaster, Kevin C Boyle, and Gordon Wetzstein. Gazegpt: Augmenting human capabilities using gaze-contingent contextual ai for smart eyewear. *arXiv preprint arXiv:2401.17217*, 2024. 2

[43] Bruno Korbar, Du Tran, and Lorenzo Torresani. Scsampler: Sampling salient clips from video for efficient action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6232–6242, 2019. 2

[44] Bolin Lai, Miao Liu, Fiona Ryan, and James M Rehg. In the eye of transformer: Global-local correlation for egocentric gaze estimation. *arXiv preprint arXiv:2208.04464*, 2022. 2

[45] Bolin Lai, Fiona Ryan, Wenqi Jia, Miao Liu, and James M Rehg. Listen to look into the future: Audio-visual egocentric gaze anticipation. In *European Conference on Computer Vision*, pages 192–210. Springer, 2024. 2

[46] Michael Land, Neil Mennie, and Jennifer Rusted. The roles of vision and eye movements in the control of activities of daily living. *Perception*, 28(11):1311–1328, 1999. 2

[47] Chae Young Lee, Maxwell Fite, Tejus Rao, Sara Achour, Zerina Kapetanovic, et al. Hypercam: Low-power onboard computer vision for iot cameras. *arXiv preprint arXiv:2501.10547*, 2025. 3

[48] Seungmin Lee and Jongseong An. Gaze control and motor performance in motor expertise studies: Focused review of field application research on perceptual skill training. *International Journal of Applied Sports Sciences*, 35(1), 2023. 2, 4, 5

[49] Qing Lei, Huiying Li, Hongbo Zhang, Jixiang Du, and Shangce Gao. Multi-skeleton structures graph convolutional network for action quality assessment in long videos. *Applied Intelligence*, 53(19):21692–21705, 2023. 2

[50] Yin Li, Alireza Fathi, and James M Rehg. Learning to predict gaze in egocentric video. In *Proceedings of the IEEE international conference on computer vision*, pages 3216–3223, 2013. 2

[51] Yin Li, Miao Liu, and James M Rehg. In the eye of the beholder: Gaze and actions in first person video. *IEEE transactions on pattern analysis and machine intelligence*, 45 (6):6731–6747, 2021. 1, 2, 4, 5, 6, 7

[52] Junhua Liao, Haihan Duan, Kanghui Feng, Wanbing Zhao, Yanbing Yang, and Liangyin Chen. A light weight model for active speaker detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22932–22941, 2023. 2

[53] Abrar Majeedi, Viswanatha Reddy Gajjala, Satya Sai Srinath GNVV Namburi, and Yin Li. Rica^2: Rubric-informed, calibrated assessment of actions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2

[54] Sagnik Majumder, Hao Jiang, Pierre Moulon, Ethan Henderson, Paul Calamia, Kristen Grauman, and Vamsi Krishna Ithapu. Chat2map: Efficient scene mapping from multi-ego conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10554–10564, 2023. 2

[55] Michele Mazzamuto*, Francesco Ragusa*, Antonino Furnari*, and Giovanni Maria Farinella*. Learning to detect attended objects in cultural sites with gaze signals and weak object supervision. *ACM Journal on Computing and Cultural Heritage*, 17(3):1–21, 2024. 2

[56] Michele Mazzamuto, Antonino Furnari, Yoichi Sato, and Giovanni Maria Farinella. Gazing into missteps: Leveraging eye-gaze for unsupervised mistake detection in egocentric videos of skilled human activities. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8310–8320, 2025. 2

[57] Rachel Melnyk, Timothy Campbell, Tyler Holler, Katherine Cameron, Patrick Saba, Michael W Witthaus, Jean Joseph, and Ahmed Ghazi. See like an expert: Gaze-augmented training enhances skill acquisition in a virtual reality robotic suturing task. *Journal of Endourology*, 35 (3):376–382, 2021. 2

[58] Meta Platforms, Inc. Project aria glasses user manual. `https://facebookresearch.github.io/projectaria_tools/docs/ARK/glasses_manual/glasses_user_manual`, 2025. Accessed: 2025-10-06. 3, 1

[59] Kyle Min and Jason J Corso. Integrating human gaze into attention for egocentric activity recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1069–1078, 2021. 4

[60] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 5

[61] Süleyman Özdel, Yao Rong, Berat Mert Albaba, Yen-Ling Kuo, Xi Wang, and Enkelejda Kasneci. Gaze-guided graph neural network for action anticipation conditioned on intention. In *Proceedings of the 2024 Symposium on Eye Tracking Research and Applications*, pages 1–9, 2024. 2

[62] Francesca Palermo, Luca Casciano, Lokmane Demagh, Aurelio Teliti, Niccolò Antonello, Giacomo Gervasoni, Hazem Hesham Yousef Shalby, Marco Brando Paracchini, Simone Mentasti, Hao Quan, et al. Advancements in context recognition for edge devices and smart eyewear: Sensors and applications. *IEEE Access*, 2025. 5, 8, 1

[63] Yulu Pan, Ce Zhang, and Gedas Bertasius. Basket: A large-scale video dataset for fine-grained skill estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1, 3

[64] Sunny Panchal, Apratim Bhattacharyya, Guillaume Berger, Antoine Mercier, Cornelius Böhm, Florian Dietrichkeit, Reza Pourreza, Xuanlin Li, Pulkit Madan, Mingu Lee, et al. What to say and when to say it: Live fitness coaching as a testbed for situated interaction. *Advances in Neural Information Processing Systems*, 37:75853–75882, 2024. 1

[65] Paritosh Parmar and Brendan Tran Morris. 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*, pages 304–313, 2019. 1

[66] Paritosh Parmar, Jaiden Reddy, and Brendan Morris. Piano skills assessment. In *2021 IEEE 23rd international workshop on multimedia signal processing (MMSP)*, pages 1–5. IEEE, 2021. 2

[67] Paritosh Parmar, Jaiden Reddy, and Brendan Morris. Piano skills assessment. In *2021 IEEE 23rd international workshop on multimedia signal processing (MMSP)*, pages 1–5. IEEE, 2021. 3

[68] Paritosh Parmar, Amol Gharat, and Helge Rhodin. Domain knowledge-informed self-supervised representations for workout form assessment. In *European Conference on Computer Vision*, pages 105–123. Springer, 2022. 2

[69] Akshay Paruchuri, Sinan Hersek, Lavisha Aggarwal, Qiao Yang, Xin Liu, Achin Kulshrestha, Andrea Colaco, Henry Fuchs, and Ishan Chatterjee. Egotrigger: Toward audio-driven image capture for human memory enhancement in all-day energy-efficient smart glasses. *arXiv preprint arXiv:2508.01915*, 2025. 2, 5, 6, 7

[70] Taiying Peng, Jiacheng Hua, Miao Liu, and Feng Lu. In the eye of mllm: Benchmarking egocentric video intent understanding with gaze-guided prompting. *arXiv preprint arXiv:2509.07447*, 2025. 2

[71] Joris Perra, Bénédicte Poulin-Charronnat, Thierry Baccino, and Véronique Drai-Zerbib. Review on eye-hand span in sight-reading of music. *Journal of eye movement research*, 14(4):10–16910, 2021. 5

[72] Chiara Plizzari, Mirco Planamente, Gabriele Goletto, Marco Cannici, Emanuele Gusso, Matteo Matteucci, and Barbara Caputo. E2 (go) motion: Motion augmented event

stream for egocentric action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19935–19947, 2022. 5, 6

[73] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. *arXiv preprint arXiv:2307.05463*, 2023. 5

[74] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets, 2015. 5

[75] Nicolas Schärer, Federico Villani, Aishwarya Melatur, Steven Peter, Tommaso Polonelli, and Michele Magno. Electrasight: Fully onboard eye tracking for smart glasses with hybrid eog (heog). *IEEE Internet of Things Journal*, 2025. 5, 8

[76] Minwoo Seong, Gwangbin Kim, Dohyeon Yeo, Yumin Kang, Heesan Yang, Joseph DelPreto, Wojciech Matusik, Daniela Rus, and SeungJun Kim. Multisensebadminton: Wearable sensor–based biomechanical dataset for evaluation of badminton performance. *Scientific Data*, 11(1):343, 2024. 2, 3, 5, 6

[77] Julian Steil, Marion Koelle, Wilko Heuten, Susanne Boll, and Andreas Bulling. Privaceye: privacy-preserving head-mounted eye tracking using egocentric scene image and eye movement features. In *Proceedings of the 11th ACM symposium on eye tracking research & applications*, pages 1–10, 2019. 2

[78] Shan Su, Jung Pyo Hong, Jianbo Shi, and Hyun Soo Park. Predicting behaviors of basketball players from first person videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1501–1510, 2017. 2

[79] Brian Sullivan, Casimir JH Ludwig, Dima Damen, Walterio Mayol-Cuevas, and Iain D Gilchrist. Look-ahead fixations during visuomotor behavior: Evidence from assembling a camping tent. *Journal of vision*, 21(3):13–13, 2021. 2

[80] R Sunder, Umesh Kumar Lilhore, Anjani Kumar Rai, Ehab Ghith, Mehdi Tlija, Sarita Simaiya, and Afraz Hussain Majeed. Smartapm framework for adaptive power management in wearable devices using deep reinforcement learning. *Scientific Reports*, 15(1):6911, 2025. 2

[81] Shuhan Tan, Tushar Nagarajan, and Kristen Grauman. Egodistill: Egocentric head motion distillation for efficient video understanding. *Advances in Neural Information Processing Systems*, 36:33485–33498, 2023. 2, 5, 6, 7, 8

[82] Yansong Tang, Zanlin Ni, Jiahuan Zhou, Danyang Zhang, Jiwen Lu, Ying Wu, and Jie Zhou. Uncertainty-aware score distribution learning for action quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9839–9848, 2020. 3

[83] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 5

[84] P Vansteenkiste, L Zeuwts, FJA Deconinck, and M Lenoir. Exploring new heights: Visual behaviour of novice, intermediate, and experienced climbers. In *CONGRESS BOOK*, page 116, 2018. 8

[85] Joan N. Vickers. Visual control when aiming at a far target. *Journal of Experimental Psychology: Human Perception and Performance*, 22(2):342–354, 1996. 2

[86] Joan N. Vickers and D. J. Lew. Quiet eye duration predicts expertise in a simulated driving task. *Cognitive Processing*, 17(3):311–319, 2016. 2

[87] Samuel J. Vine, R. J. Chaytor, J. S. McGrath, and R. S. W. Masters. Gaze training improves laparoscopic surgical performance. *Surgical Endoscopy*, 25(12):3731–3739, 2011. 2

[88] Chaowei Wang, Ziye Wang, Weiwei Guan, Wenjie Wang, Lexi Xu, Lihua Li, Sai Huang, and Weidong Wang. Trustworthy health monitoring based on distributed wearable electronics with edge intelligence. *IEEE Transactions on Consumer Electronics*, 70(1):2333–2341, 2024. 2

[89] Fengshun Wang, Qiurui Wang, and Dan Chen. From beats to scores: A multi-modal framework for comprehensive figure skating assessment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5905–5914, 2025. 2, 3

[90] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4933–4942, 2019. 5

[91] Tianyu Wang, Yijie Wang, and Mian Li. Towards accurate and interpretable surgical skill assessment: A video-based method incorporating recognized surgical gestures and skill levels. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 668–678. Springer, 2020. 2

[92] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, 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*, pages 20270–20281, 2023. 1

[93] Junbin Xiao, Nanxin Huang, Hao Qiu, Zhulin Tao, Xun Yang, Richang Hong, Meng Wang, and Angela Yao. Egoblind: Towards egocentric visual assistance for the blind people. *arXiv preprint arXiv:2503.08221*, 2025. 2

[94] Chengming Xu, Yanwei Fu, Bing Zhang, Zitian Chen, Yu-Gang Jiang, and Xiangyang Xue. Learning to score figure skating sport videos. *IEEE transactions on circuits and systems for video technology*, 30(12):4578–4590, 2019. 2

[95] Huangbiao Xu, Xiao Ke, Yuezhou Li, Rui Xu, Huanqi Wu, Xiaofeng Lin, and Wenzhong Guo. Vision-language action knowledge learning for semantic-aware action quality assessment. In *European Conference on Computer Vision*, pages 423–440. Springer, 2024. 2

[96] Huangbiao Xu, Xiao Ke, Huanqi Wu, Rui Xu, Yuezhou Li, and Wenzhong Guo. Language-guided audio-visual learning for long-term sports assessment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23967–23977, 2025. 2

[97] Jinglin Xu, Yongming Rao, Xumin Yu, Guangyi Chen, Jie Zhou, and Jiwen Lu. Finediving: A fine-grained dataset for procedure-aware action quality assessment. In *CVPR*, pages 2949–2958, 2022. 2, 3

[98] Jinglin Xu, Sibo Yin, Guohao Zhao, Zishuo Wang, and Yuxin Peng. 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*, pages 14628–14637, 2024. 1

[99] Han Yi, Yulu Pan, Feihong He, Xinyu Liu, Benjamin Zhang, Oluwatumininu Oguntola, and Gedas Bertasius. Exact: A video-language benchmark for expert action analysis. *arXiv preprint arXiv:2506.06277*, 2025. 1

[100] Ling-An Zeng and Wei-Shi Zheng. Multimodal action quality assessment. *IEEE Transactions on Image Processing*, 33:1600–1613, 2024. 2

[101] Mengmi Zhang, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Jiashi Feng. Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4372–4381, 2017. 2

[102] Shiyi Zhang, Wenxun Dai, Sujia Wang, Xiangwei Shen, Jiwen Lu, Jie Zhou, and Yansong Tang. Logo: A long-form video dataset for group action quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2405–2414, 2023. 1

[103] Shiyi Zhang, Sule Bai, Guangyi Chen, Lei Chen, Jiwen Lu, Junle Wang, and Yansong Tang. Narrative action evaluation with prompt-guided multimodal interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18430–18439, 2024. 2

[104] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, C Karen Liu, and Leonidas J Guibas. Gimo: Gaze-informed human motion prediction in context. In *European Conference on Computer Vision*, pages 676–694. Springer, 2022. 1, 2

[105] Sheng Zhou, Junbin Xiao, Qingyun Li, Yicong Li, Xun Yang, Dan Guo, Meng Wang, Tat-Seng Chua, and Angela Yao. Egotextvqa: Towards egocentric scene-text aware video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3363–3373, 2025. 2