SkillSight: Efficient First-Person Skill Assessment with Gaze

Supplementary Material

A. Supplementary video

We provide a supplementary video that shows an overview of the paper. It also shows qualitative video examples with ego video and gaze patterns.

B. Ablations

SkillSight-T. In Sec. 3.2, we present the three components of SkillSight-T: a visual encoder with gaze attention, a cropped image encoder, and a trajectory encoder. We perform an ablation study on these designs using EgoExo4D [31]. As shown in Table 3, each component contributes a clear gain in accuracy, indicating that these designs are essential for capturing the interaction between ego visual input and gaze in skill assessment.

SkillSight-S. In Sec. 3.3, we introduce the distillation strategy used for training SkillSight-S. We evaluate the influence of each loss through an ablation study, with results summarized in Table 4. The results demonstrate that both the distillation loss and the action recognition loss contribute positively to the performance of SkillSight-S. This highlights the importance of linking gaze patterns with specific actions for effective skill assessment, and video cues for skill assessment can be effectively embedded into the gaze signal.

Input modalities. While SkillSight-S leverages both gaze and head motion as inputs, we further examine the contribution of each modality by separating gaze direction and head rotation. Using the same distillation training, SkillSight-S achieves 41.4 with gaze-only, 41.8 with head-motion-only, and 44.4 with both head-motion and gaze, demonstrating that both modalities are necessary.

TimeSformer	Crop Encoder	Traj. Encoder	Acc (%)
✗	✗	✓	37.0
✗	✓	✗	40.6
✗	✓	✓	44.9
w/o gaze att.	✗	✗	45.5
w/o gaze att.	✗	✓	46.4
w/o gaze att.	✓	✗	47.6
w/o gaze att.	✓	✓	47.9
w gaze att.	✗	✗	47.2
w gaze att.	✗	✓	47.5
w gaze att.	✓	✗	48.6
w gaze att.	✓	✓	**50.1**

Table 3. **Ablation study of SkillSight-T.** We conduct ablation study of the three components in SkillSight-T. A check indicates inclusion.

Method	Acc (%)
Gaze-Only	37.0
SkillSight-S	**44.4**
w/o distillation	40.0
w/o action recognition	40.7

Table 4. **Ablation study of training SkillSight-S.** We compare the performance of SkillSight-S when training under different loss configurations.

C. Gaze normalization process

In Sec. 3.1, we describe the gaze modalities derived from the three-dimensional gaze vector of each eye. We also provide details on how each modality is normalized to remove bias in the recordings.

- **3D fixation points.** For each frame, we calculate the intersection of the left and right gaze rays in the world coordinate. Centered by the segment mean and rotated horizontally so the first direction gaze point has y equal to zero.
- **3D gaze direction.** For each frame, this is a unit vector representing the gaze direction expressed in the cpf coordinate as defined by the subject's perspective [58].
- **2D gaze point projection.** We project the 3D gaze to the 2D ego camera view, value in the range zero to one.
- **Gaze depth.** Distance between the head and the intersection point of the left and right gaze rays.
- **Glass rotation.** For the first frame, adjust the yaw to face forward while keeping pitch and roll. For other frames, compute relative rotation. Convert the rotation representation from pitch, yaw, roll to quaternion.
- **Glass translation.** Center the translation and define the first horizontal movement as the positive x direction.

A modality is included in an experiment if the corresponding dataset provides it. In EgoExo4D [31], the glasses translation is obtained using visual-inertial-odometry, which relies on a low-power SLAM camera. Regardless, a SLAM camera is not a strict requirement; trajectories can alternatively be inferred using IMU alone [62]. We also evaluate a variant of SkillSight-S that uses only head rotation and 3D gaze, and observe a performance of 44.4% on EgoExo4D (a 0.4% drop compared to when trajectory is enabled).

D. Power consumption calculation details

In Sec. 4, we demonstrate the power efficiency analysis. Following the energy computation in [31], the overall en-

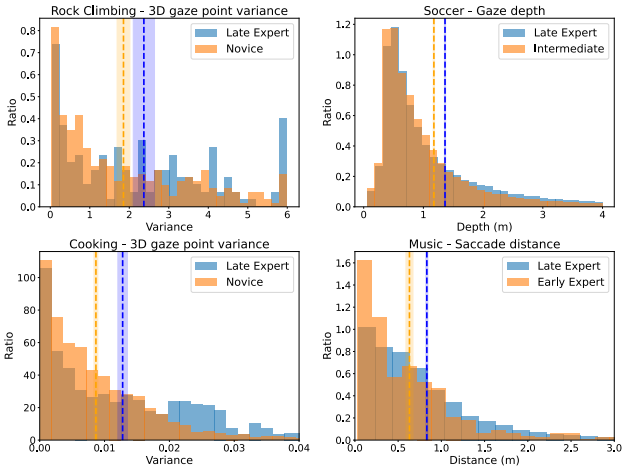

Figure 7. **Distinct gaze pattern analysis.** We present more distinct gaze patterns that SkillSight-S reveals between subjects at different skill levels.

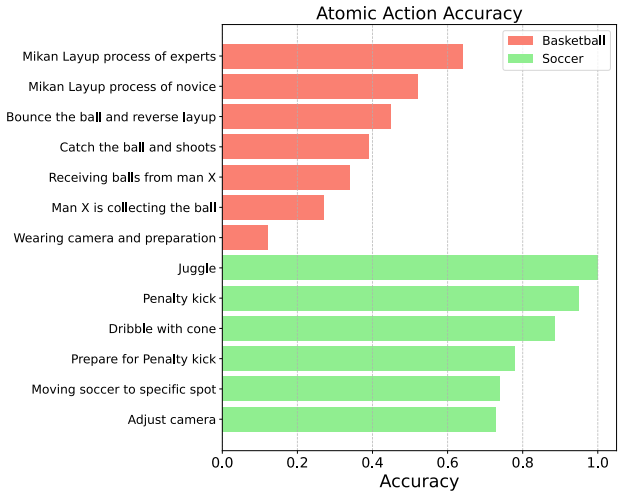

Figure 8. **Atomic action analysis for SkillSight-S.** We present the performance of SkillSight-S across different atomic actions.

ergy consumption rate consists of the following:

- **Compute operations (MACs).** We estimate computational cost by using the PyTorch FLOP counter to measure the total FLOPs in a forward pass, and then convert this value to MACs using the approximation that one MAC equals two FLOPs.
- **Memory transfer (bytes).** We quantify GPU memory movement with the PyTorch memory profiler, which records all operations in the forward pass along with their memory usage. The total memory transfer is the sum of the memory costs of all logged operations.
- **Sensor capture.** For each sensing modality, we measure the period during which it is active by counting the number of samples that include that modality. We require at least one second of data, since energy consumption cannot be defined for a single instantaneous reading.

E. Behavior-level interpretation of gaze

In Figure 6, we present the distinct gaze patterns uncovered from the proficiency groups predicted by SkillSight-S. Figure 7 further illustrates additional gaze behavior insights captured by SkillSight-S. In rock climbing, the variance of 3D gaze points is notably larger for the predicted late expert than for the novice, a result of frequent gaze shifts to gather information from the wall. In soccer, predicted late experts tend to fixate on farther depths, reflecting their attention to broader surroundings and potential targets. In cooking, predicted experienced chefs exhibit a more diverse 3D gaze points while monitoring food. Finally, in music, the model-predicted late experts switch more flexibly between the sheet music and the instrument, resulting in longer saccade distances. These findings enable deeper investigation of how gaze patterns vary across skill levels, allowing a more data-driven understanding of expertise.

F. Analysis of performance across actions

In Figure 8, we analyze the performance of SkillSight-S across different actions by clustering Ego-Exo4D atomic action captions. We observe that gaze alone preserves most skill cues in perception-driven actions, such as basketball layups, soccer dribbling, and penalty kicks, where motion intent is strongly reflected in gaze. In contrast, performance degrades for actions requiring subtle motor execution, such as basketball shooting.