Look Closer: Action-Guided Dense Visual Dynamics for Proficiency Estimation

Deokhyun Ahn[1,3], Hyukjin Kim[2,3], Jae-Ho Han[1], Bumsub Ham[2], Heeseung Choi[3], Ig-Jae Kim[3], Haksub Kim[3]
[1]Korea University [2]Yonsei University [3]Korea Institute of Science and Technology (KIST)

***Abstract*—Human proficiency estimation is challenging because skill is a high-level and often subjective property that is not captured by action identity alone. Existing approaches leverage multimodal observations such as ego-exo video, gaze, pose, and inertial signals, but the gap between sensor-level evidence and abstract proficiency labels remains difficult to bridge. We address this gap by using language annotations as temporal guidance rather than direct semantic supervision. Specifically, atomic action descriptions decompose each activity into proficiency-relevant action windows, where we extract dense visual dynamics from frame-level features. We then summarize each action window into an action-dynamics feature and compare it against high- and low-proficiency prototypes constructed from expert and novice executions within the same activity-action category, yielding a high-low prototype margin that measures relative execution quality.**

***Index Terms*—Multi-modality, Ego-Exo4D [1], Proficiency Estimation**

I. Introduction

Proficiency estimation remains difficult even when ego video, exo video, and gaze are available. The challenge is not simply to recognize which action is being performed, but to judge how well it is executed. A video-gaze baseline can capture global activity context, but sparse full-take sampling may dilute the brief execution moments where proficiency differences emerge.

As illustrated in Fig. 3, we use atomic action descriptions as temporal guidance to identify where proficiency-relevant execution cues occur. An atomic action description is a short language annotation associated with a local action segment, such as *shoot*, *pass*, or *cut*. We do not use these descriptions as direct semantic supervision; instead, we use them to localize action-specific windows and extract dense visual dynamics.

To our knowledge, prior proficiency-estimation frameworks have not explicitly accounted for the interplay among activity-dependent proficiency evidence, action-level execution moments, relative high-low proficiency references, and the ordinal structure of skill labels. We address this gap by structuring proficiency estimation around where execution quality should be measured, what proficiency reference it should be compared against, and how predictions should be calibrated under ordered skill levels. We summarize our contributions as follows:

- *Activity-aware aspect expert routing.* We formulate proficiency estimation as activity-dependent evidence aggregation. Ego, exo, and gaze features are organized into aspect-oriented expert branches aligned with modality-specific evidence, such as gaze-related signals from Aria eye-gaze and ego video or interaction-related cues from ego-centric visual features. An activity-aware router combines a fixed activity-aspect prior with learned sample-conditioned gates to re-weight expert evidence for each input.
- *Atomic-action-guided high-low prototype margin.* We use atomic action descriptions to define where execution quality should be measured. Each localized action window is summarized into an action-dynamics feature and compared with high- and low-proficiency prototypes from the same activity-action category, producing relative evidence of expert-like versus novice-like execution.
- *Ordinal proficiency calibration.* We treat proficiency labels as ordered levels rather than independent classes. Activity-normalized prototype margins are fused with the global baseline prediction through a CORAL [2] ordinal head, enabling calibration under the natural Novice-to-Late-Expert progression.

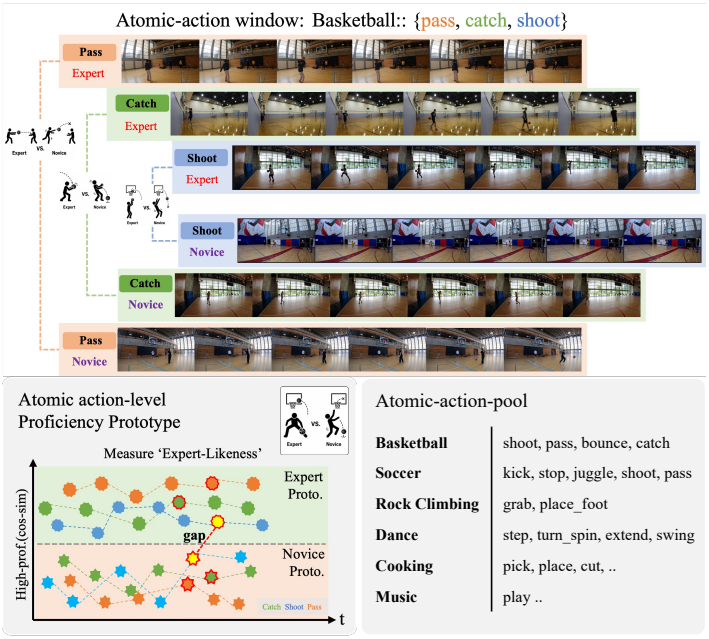

Fig. 1: **Where should we look to assess skill?** We use atomic-action windows to compare dense visual dynamics against expert and novice prototypes, yielding an action-level proficiency margin.

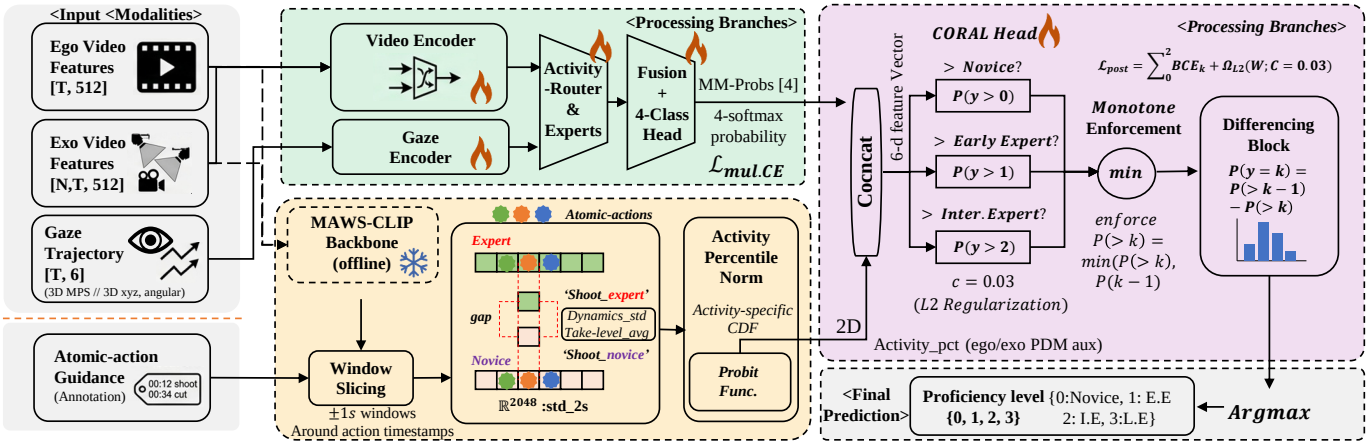

Fig. 2: **Overview of the proposed pipeline.** A frozen video-gaze classifier provides multimodal class probabilities, while atomic-action guidance localizes dense MAWS windows to compute an activity-normalized proficiency dynamics score. We concatenate both signals and use a CORAL ordinal head to predict the final proficiency level.

II. Methods

A. Multi-modal Baseline with Activity-Aware Aspect Routing

Before calibration, we train a Stage-1 multi-modal baseline that maps ego, exo, and gaze features into a 4D proficiency probability vector. Let $H^m \in \mathbb{R}^{T\times d}$ denote the temporal feature sequence of modality $m \in \{\text{ego}, \text{exo}, \text{gaze}\}$, and let $\mathcal{A}$ denote a set of predefined aspect-oriented expert branches. These branches organize modality-specific evidence around activity-relevant factors, such as gaze-related evidence from Aria eye-gaze and ego video, interaction-related evidence from ego-centric visual features, and timing/procedure-related evidence from ego-exo video context. They are not intended to explicitly extract semantic labels; instead, they provide an inductive structure for aggregating modality-specific evidence.

For each aspect branch $a \in \mathcal{A}$, an expert receives a designated subset of modality features $\mathcal{S}_a$ and produces an aspect-oriented evidence vector e_a and an auxiliary aspect-polarity prediction:

$$e_a = \text{AttnPool}\left(\phi_a\left([H^m]_{m\in\mathcal{S}_a}\right)\right), \qquad p_a = \text{softmax}(W_a e_a),$$

where ϕ_a is an aspect-specific MLP and $p_a \in \mathbb{R}^3$ represents novice-like, unknown, and expert-like polarity.

The activity-aware router re-weights aspect expert evidence by combining a fixed activity-aspect relevance prior with learned sample-conditioned gates. Given the activity identity and the global multi-modal context c, the router computes

$$w = \text{softmax}\left(f_{\text{rtr}}([E(\alpha); c]) + \log(\pi_\alpha + \epsilon)\right),$$

where $E(\alpha)$ is the activity embedding, π_α is a fixed activity-aspect relevance prior, and ϵ avoids numerical instability. The prior provides an activity-level bias, while the learned gate f_{rtr} adapts the expert weights for each input sample. The router-weighted evidence is then fused into the baseline prediction:

$$\bar{e} = \sum_{a\in\mathcal{A}} w_a e_a, \qquad p_{\text{base}} = \text{softmax}(W_o \bar{e}) \in \mathbb{R}^4.$$

The Stage-1 baseline is trained with a multi-task classification loss:

$$\mathcal{L}_{\text{mul.CE}} = \mathcal{L}_{\text{cls}} + \lambda_{\text{asp}}\mathcal{L}_{\text{asp}} + \lambda_{\text{rtr}}\mathcal{L}_{\text{rtr}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}.$$

Here, $\mathcal{L}_{\text{cls}}$ is the overall proficiency classification loss, $\mathcal{L}_{\text{asp}}$ weakly supervises aspect-polarity predictions when commentary-derived pseudo-labels are available, $\mathcal{L}_{\text{rtr}}$ regularizes the routing weights, and $\mathcal{L}_{\text{con}}$ encourages consistency between aspect-level evidence and the overall proficiency prediction. Commentary-derived aspect labels are used only during training and are not required at inference.

B. Dense Action Windows and High-Low Prototype Dynamics Margin (PDM)

The Stage-1 baseline produces a global proficiency estimate, but sparse video features can miss short execution dynamics that distinguish proficiency levels. We therefore extract dense local features around atomic action timestamps. For each stream $s \in \{\text{ego}, \text{exo}\}$ and timestamp t, we extract dense MAWS features $X_t^{(s)} = \{x_i^{(s)}\}_{i=1}^N$ within a bounded $[t-1\,\text{s}, t+1\,\text{s}]$ window, where $N \approx 60$ frames at 30 fps. We summarize local visual dispersion by the temporal standard deviation feature

$$z_t^{(s)} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i^{(s)} - \bar{x}_t^{(s)}\right)^2},$$

where $\bar{x}_t^{(s)}$ is the mean feature within the window and the square and square root are applied element-wise. This action-dynamics feature captures the temporal variation of dense visual features within a local action window without requiring additional frame-level supervision.

Measuring raw similarity to expert prototypes can be biased by activity-specific visual context shared across proficiency levels. Thus, we use a relative high-low prototype margin. Using only the train set, we aggregate action-dynamics features into activity-action buckets $b = (\alpha, \beta)$, where α is the activity

and β is the atomic action type, such as *Basketball:shoot*. For each stream s and bucket b, we define a high prototype $\mu_{\text{high}}^{(s,b)}$ as the mean feature of Intermediate/Late experts, and a low prototype $\mu_{\text{low}}^{(s,b)}$ as the mean feature of Novice/Early performers. The collection of high-low prototypes over all activity-action buckets forms the train-time prototype bank used for PDM scoring. For a test action-dynamics feature $z_t^{(s)}$, the window-level prototype dynamics margin is

$$m_t^{(s)} = \cos\left(z_t^{(s)}, \mu_{\text{high}}^{(s,b)}\right) - \cos\left(z_t^{(s)}, \mu_{\text{low}}^{(s,b)}\right).$$

A positive value indicates greater similarity to high-proficiency executions within the same activity-action bucket. For each take and stream, we average the window-level margins over all available atomic-action windows to obtain a take-level stream-specific PDM score $m^{(s)}$.

In the main action-annotated setting, $z_t^{(s)}$ is extracted from windows centered at atomic-action timestamps. The same PDM computation can also be applied to *annotation-free* temporal proposals by scoring each proposal against all train-time high-low prototypes associated with the same activity and taking the maximum margin before activity-level normalization.

C. Ordinal Calibration via CORAL Head

Since motion scales vary across activities, we normalize each take-level stream-specific PDM score using the empirical CDF of the train set within the same activity:

$$u^{(s)} = \Phi^{-1}\left(F_{\alpha}^{\text{train}}\left(m^{(s)}\right)\right), \qquad s \in \{\text{ego}, \text{exo}\}.$$

We concatenate the normalized ego/exo PDM scores with the Stage-1 baseline probability vector:

$$v = [p_{\text{base}};\, u^{(\text{ego})};\, u^{(\text{exo})}].$$

Missing exo PDM values are zero-filled in our experiments.

To leverage the inherent order of proficiency labels, we fit a Stage-2 CORAL ordinal head on the frozen features. This reformulates the 4-class problem into three cumulative binary boundaries,

$$q_k = P(y > k \mid v), \qquad k \in \{0, 1, 2\}.$$

The post-hoc objective is

$$\mathcal{L}_{\text{post}} = \sum_{k=0}^{2} \text{BCE}\left(\mathbf{1}\{y > k\}, q_k\right) + \Omega_{\text{L2}}(W; C = 0.03),$$

where Ω_{L2} denotes the L2 penalty on the logistic regression weights with $C = 0.03$. We implement the CORAL head as an L2-regularized logistic regression head with $C = 0.03$. After enforcing $q_0 \geq q_1 \geq q_2$, the final class probabilities are obtained by differencing the cumulative probabilities.

Table I. Proficiency benchmark results on Ego-Exo4D.

Method	Pretrained	Acc	mF1	MAE	Spearman
TimeSFormer [3]	HowTo100M	0.468	-	-	-
EgoPulseFormer [4]	EgoPPG-DB	0.453	-	-	-
SkillFormer [5]	K600	0.459	-	-	-
ProfVLM (AGP) [6]	K600 + SmoLM2	0.442	-	-	-
SkillSight [7]	EgoVLPv2	0.501	-	-	-
Full (proposed)	MAWS-CLIP-2B	**0.542**	**0.517**	**0.654**	**0.468**

Table II. Per-proficiency level performance comparison.

	F1		Precision	
Class	Baseline	Proposed	Baseline	Proposed
Novice	0.295	**0.418**	0.275	**0.341**
Early Expert	0.524	**0.582**	0.514	**0.550**
Intermediate Expert	0.423	**0.430**	0.511	**0.553**
Late Expert	0.494	**0.537**	0.495	**0.536**

Table III. Per-activity performance comparison.

Activity	Full	w/o gaze	w/o PDM	w/o CORAL	w/o router	w/o aspect
Basketball	0.659	0.565	0.585	0.630	0.711	0.585
Cooking	0.433	0.467	0.517	0.500	0.350	0.433
Dance	0.572	0.501	0.580	0.570	0.549	0.493
Music	0.685	0.556	0.778	0.741	0.481	0.537
Rock Climbing	0.439	0.408	0.357	0.366	0.399	0.441
Soccer	0.576	0.636	0.606	0.606	0.636	0.606
Overall	0.538	0.484	0.505	0.511	0.515	0.494

III. Experiments

A. Experimental Setup

Dataset and metrics. We evaluate our method on the Ego-Exo4D proficiency estimation task following the standard validation protocol [5]–[7]. The model predicts four ordered proficiency levels: Novice, Early Expert, Intermediate Expert, and Late Expert. While prior work mainly reports accuracy, accuracy alone can obscure class imbalance and ordinal prediction errors. We therefore additionally report macro-F1 for class-balanced performance, MAE for ordinal error, and Spearman correlation for ranking consistency.

Implementation details. We use frozen 16-frame ego/exo video features and gaze trajectories, with temporal video Transformers, a GRU-based gaze encoder, and activity-aware aspect routing. Stage-1 training uses AdamW (lr $= 2\times10^{-4}$, weight decay 10^{-4}), batch size 8, gradient clipping 1.0, mixed precision, and $(\lambda_{\text{asp}}, \lambda_{\text{rtr}}, \lambda_{\text{con}}) = (0.5, 0.01, 0.1)$. The CORAL head is fit on frozen $[p_{\text{base}}; u^{(\text{ego})}; u^{(\text{exo})}]$ features with $C = 0.03$, and all results are averaged over three seeds.

B. Results and Analysis

Main benchmark. Table I compares the proposed model with prior Ego-Exo4D proficiency estimation baselines. Our full model achieves the best accuracy among the compared methods. Because most previous methods do not report macro-F1, MAE, or Spearman correlation, we report these additional metrics for our model to better reflect class balance, ordinal error, and ranking consistency.

Per-level and per-activity behavior. Table II shows that the proposed model improves both F1 and precision across all four proficiency levels. Table III further shows that each component has activity-dependent strengths rather than uniformly

Table IV. Ablation on the Ego-Exo4D, with 3-seed mean.

	Stage	V	G	Rtr	Asp	PDM	CORAL	Acc	mF1	MAE	Sp
1	Baseline (video only)	✓	–	–	–	–	–	0.430	0.383	0.796	0.382
2	+ gaze	✓	✓	–	–	–	–	0.430	0.395	0.801	0.355
3	+ activity router	✓	✓	✓	–	–	–	0.442	0.400	0.785	0.403
4	+ aspect experts	✓	✓	✓	✓	–	–	0.498	0.459	0.727	0.415
5	+ PDM (w/o CORAL)	✓	✓	✓	✓	✓	–	0.529	0.504	0.681	0.460
6	+ PDM + CORAL (Full)	✓	✓	✓	✓	✓	✓	**0.542**	**0.517**	**0.654**	**0.468**
	w/o gaze from Full	✓	–	✓	✓	✓	✓	0.490	0.463	0.676	0.466

Notation. V: ego/exo video, G: gaze, Rtr: activity router, Asp: aspect experts, PDM: Prototype Dynamics Margin, Sp: Spearman correlation.

Table V. PDM robustness and control analysis.

Setting	Acc	MAE	Acc. gain	MAE red.
Main baseline	0.4976	0.7268	–	–
Action-timestamp dependency				
Proposed w/ action timestamps	0.5420	0.6540	+4.4 pp	10.0%
Proposed w/o action timestamps	**0.5537**	**0.6114**	**+5.6 pp**	**15.9%**
High-low prototype validity				
Full proposed	**0.5420**	**0.6540**	**+4.4 pp**	**10.0%**
w/ shuffled proficiency labels	0.5008	0.6837	+0.3 pp	5.9%
w/ random prototypes	0.4976	0.7309	+0.0 pp	-0.6%
w/ activity-average score	0.5098	0.7081	+1.2 pp	2.6%

improving every activity. For example, gaze contributes clearly in Basketball, Dance, Music, and Rock Climbing, but is less beneficial in Cooking and Soccer, suggesting that our current gaze representation may require more structured modeling for activities with different attention patterns or rapid target switches. PDM is particularly useful in Basketball and Rock Climbing, where short execution dynamics are important, while some activities can rely more on global context or baseline cues. Similarly, router, aspect, and CORAL ablations occasionally perform better on individual activities, but none of them provides the best balance across all activities. The main finding is therefore not that a single modality or component dominates every scenario, but that combining global ego-exo-gaze evidence, localized PDM signals, and ordinal calibration yields the strongest overall performance.

Component ablation. Table IV decomposes the proposed model from a video-only baseline to the full system under a unified evaluation setting. The activity router and aspect experts strengthen the global backbone, while PDM adds localized high-low prototype evidence from dense action windows. The CORAL head further calibrates the prediction under the ordered proficiency structure. Overall, the full system improves accuracy by +11.2 percentage points over the video-only baseline and reduces MAE from 0.796 to 0.654. The leave-one-out gaze ablation suggests that gaze contributes most clearly when combined with routing, aspect experts, PDM, and ordinal calibration.

PDM robustness and controls. Table V evaluates whether PDM depends on action timestamps and whether its high-low prototype signal captures meaningful proficiency structure. Replacing action-annotated windows with uniform 2s temporal proposals preserves the gain over the main baseline, suggesting that PDM can recover high-low evidence without manually annotated action boundaries at inference. We further test three controls: shuffled proficiency labels

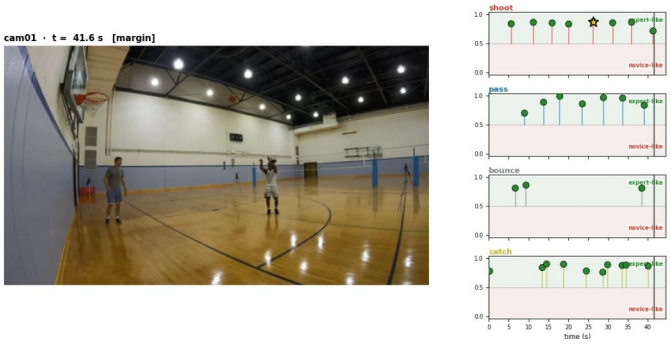

Fig. 3: **Where should we look to assess skill?** We use atomic-action windows to compare dense visual dynamics against expert and novice prototypes, yielding an action-level proficiency margin.

before prototype construction, random prototype references, and an activity-average score. The activity-average control replaces action-specific high-low comparisons with a single activity-level score, removing the ability to compare each local action against its corresponding high/low prototypes. The gains largely collapse under these controls, indicating that PDM improves performance through proficiency-aware and action-specific high-low evidence rather than simply adding an auxiliary score.

Computational cost. Our method contains 5.58M parameters and adds only a post-hoc CORAL calibration head on top of frozen baseline probabilities and two ego/exo PDM auxiliary features.

Qualitative Result

IV. Conclusion

This paper shows that proficiency estimation is not only a problem of recognizing what action is performed, but also of identifying where execution quality is expressed and how it differs across ordered skill levels. The proposed framework integrates activity-aware ego-exo-gaze evidence, dense action-window PDM, and CORAL-based ordinal calibration to combine global context with local high-low execution cues. Our results suggest that balanced integration of complementary evidence is more important than relying on any single modality or component.

REFERENCES

[1] K. Grauman, A. Westbury, L. Torresani, K. Kitani, J. Malik, T. Afouras, K. Ashutosh, V. Baiyya, S. Bansal, B. 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*, 2024, pp. 19 383–19 400.

[2] W. Cao, V. Mirjalili, and S. Raschka, "Rank consistent ordinal regression for neural networks with application to age estimation," *Pattern Recognition Letters*, vol. 140, pp. 325–331, 2020.

[3] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *Icml*, vol. 2, no. 3, 2021, p. 4.

[4] B. Braun, R. Armani, M. Meier, M. Moebus, and C. Holz, "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*, 2025, pp. 5579–5590.

[5] E. Bianchi and A. Liotta, "Skillformer: unified multiview video understanding for proficiency estimation," in *Eighteenth International Conference on Machine Vision (ICMV 2025)*, vol. 14114. SPIE, 2026, pp. 685–692.

[6] E. Bianchi, J. Staiano, and A. Liotta, "Profvlm: A lightweight video-language model for multi-view proficiency estimation," *Computer Vision and Image Understanding*, p. 104749, 2026.

[7] C. H. Wu, K. Ashutosh, and K. Grauman, "Skillsight: Efficient first-person skill assessment with gaze," *arXiv preprint arXiv:2511.19629*, 2025.