Pole-Arina: A Privacy-Preserving Dataset and Benchmark for Static Pole Tricks

Anonymous CVsports submission

Paper ID 7

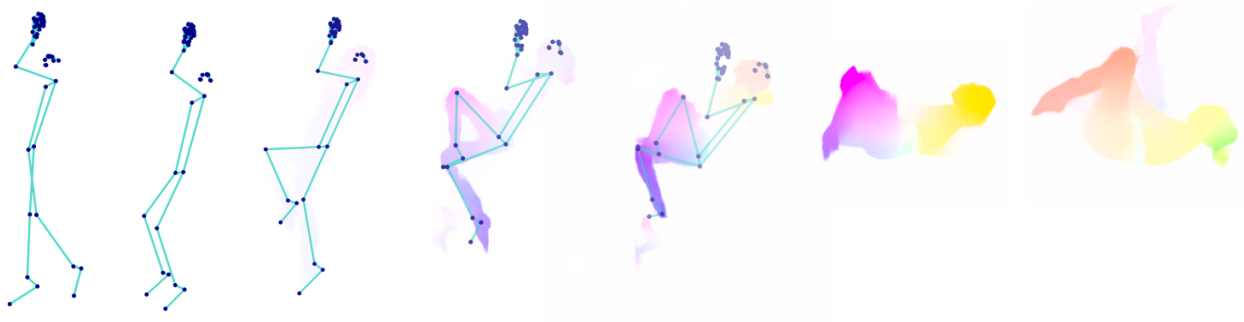


Figure 1. *Pole-Arina* replaces identifiable RGB with privacy-preserving motion representations. We extract 2D pose skeletons and dense optical flow, enabling clip-level trick recognition, frame-wise phase understanding, and coaching-oriented analysis for static pole tricks under strong privacy constraints.

Abstract

Computer vision for sports lacks datasets and models that target pole sports, an activity marked by self-occlusion, body inversions, and sustained contact with the pole. We address this gap by introducing ***Pole-Arina****, a curated, privacy-preserving dataset for markerless analysis of static pole tricks. The collection comprises 836 videos from 58 participants and provides two appearance-suppressing modalities derived from the recordings: pose skeleton sequences (2D joints) and dense optical flow. We annotate each clip with both per-video trick labels and per-frame labels that capture temporal structure (e.g.,* `floor`*,* `on_pole`*, and trick-specific pose states), enabling evaluation of clip-level and frame-wise recognition under privacy constraints. As reference baselines, we benchmark lightweight temporal models across all benchmark settings (skeleton/flow* $\times$ *video/frame) and analyze common confusions and imbalance effects. In addition, we provide a geometry-aware analysis module that measures trick-specific body orientation, joint alignments, and proximities to produce interpretable overlays and actionable feedback. To the best of our knowledge, this is the first dataset and benchmark targeting recognition and analysis of pole tricks from privacy-preserving motion representations. We will release the dataset, annotations, and code upon acceptance.*

1. Introduction

Video-based coaching systems increasingly combine pose estimation with temporal sequence models to analyze human motion and deliver corrective feedback across fitness, sports, and dance [**?** **?** **?** **?** **?**]. Markerless approaches are especially compelling because they avoid wearables and preserve natural movement [**?**]. For fine-grained technique analysis, models must support both clip-level recognition and temporally localized, per-frame understanding under occlusions, viewpoint changes, and rapid transitions, conditions that motivate curated datasets and learning pipelines built on privacy-preserving representations. One such unexplored example is pole dancing.

As a sport, pole dancing has grown into a recognized athletic discipline with standardized competition rules (e.g., within the International Pole and Aerial Sports Federation since 2009) that emphasize technique and objective evaluation [**?**]; while debates about hypersexualization persist, a sport-centered framing prioritizes inclusive, skill-based assessment [**?**]. Static pole tricks demand strength, flexibility, precise alignment, and sustained skin-pole contact (rendering wearables impractical); they are practiced in studios and at home, where dancers commonly self-record. Brief, discriminative end-poses are preceded by visually similar approach phases, occlusions are frequent, viewpoints vary, and naming conventions differ across studios, complicating both recognition and temporal localization. To the best of our knowledge, no prior work provides a public dataset for

pole trick recognition.

We address this gap with *Pole-Arina*, a new privacy-preserving dataset and benchmark for static pole tricks. The dataset contains 836 clips spanning six foundational tricks, performed by multiple volunteers with varied skill levels and including both successful and failed attempts, totaling 212,574 labeled frames. To mitigate privacy concerns while retaining kinematic fidelity, we do not release RGB video; instead, we provide two derived, appearance-suppressing modalities: (i) pose skeleton sequences (2D joints with confidence) and (ii) dense optical flow computed from the standardized clips - Figure 1. We further provide two annotation granularities: per-video trick labels for clip-level action recognition and per-frame labels capturing temporal structure (e.g., `floor`, `on_pole`, and trick-specific pose states), enabling temporally localized evaluation.

On top of this dataset, we benchmark lightweight temporal models across both modalities and both supervision settings (per-video and per-frame). Clip-level recognition is strong for skeletons (91.26% accuracy), while optical flow is more challenging (66.67%), reflecting the fact that pose extraction supplies a strong inductive bias whereas flow must recover discriminative structure from motion alone. For frame-wise recognition, overall accuracies reach 86.99% (skeleton) and 81.23% (flow), but performance is affected by severe label imbalance in the per-frame taxonomy, highlighting open challenges for fine-grained, privacy-preserving recognition. Beyond recognition, we include an interpretable geometric scoring module that encodes technique cues (angles, alignments, distances) on the skeleton to produce actionable, per-rule feedback, and we evaluate a coaching-oriented prototype in a controlled user study.

Our contributions are as follows:

- **Pole Dancing Multimodal Privacy-Preserving Dataset.** We introduce a dataset of 836 clips spanning six fundamental static pole tricks, totaling 212,574 labeled frames, and release privacy-preserving representations (pose skeleton sequences and dense optical flow) together with per-video and per-frame labels.
- **Unified Benchmark Across Modalities and Label Granularities.** We define and benchmark four recognition settings (skeleton/flow $\times$ video/frame), enabling systematic evaluation of clip-level trick recognition and temporally localized per-frame classification under appearance suppression.
- **Baseline Models, Analysis, and Coaching-Oriented Feedback.** We establish lightweight baseline results for all settings, analyze confusions and imbalance effects, and provide an interpretable geometric scoring module with a user-study evaluation in a coaching prototype.

Together, the dataset, benchmark, and baselines establish pole dance as a new computer vision problem setting for privacy-aware, fine-grained athletic technique understanding, leaving clear headroom for improved temporal modeling, robustness, and minority-class recognition.

2. Related Work

Motion capture for athletic analysis. Optical marker-based mocap remains the gold standard for accuracy but is costly, intrusive, and impractical outside controlled labs [**?** **?**]. Wearable inertial systems improve portability yet suffer from drift and magnetic disturbances and can impede natural movement in contact sports [**?**]. Markerless, vision-based pose estimation offers the best trade-off for in-situ training, preserving unconstrained motion while requiring only commodity cameras [**?**].

Pose estimation. Modern methods estimate body keypoints directly from RGB, from DeepPose regression [**?**] and multi-person Part Affinity Fields in OpenPose [**?**] to mobile or edge pipelines such as BlazePose or MediaPipe Pose with monocular estimation [**?** **?**]. Complementary open-source implementations and toolkits, including OpenPifPaf (a composite-field, bottom-up formulation) and MMPose, provide strong baselines and practical training and inference pipelines across a range of backbones and deployment settings [**?** **?** **?**]. These approaches enable non-invasive capture in studios and at home, but remain sensitive to occlusions, fast motion, and viewpoint changes, conditions common in technical sports and dance.

Optical flow and privacy. Optical flow encodes motion as per-pixel displacement fields and suppresses most appearance cues, making it a common input for action recognition when texture or identity should be minimized [**?**]. Recent work explicitly studies appearance-free (motion-only) recognition, showing that competitive performance is possible even when static visual cues are removed, and motivating architectures that recover and leverage explicit flow [**?**]. In privacy-preserving action recognition more broadly, motion-centric encodings are often used as pragmatic compromises: they reduce identifiability while retaining discriminative temporal structure [**?**].

Sports recognition and temporal modeling. For sequence understanding, recurrent models remain strong on modest datasets [**?** **?**], while Transformers capture long-range dependencies given sufficient data [**?** **?**]. Literature reports a common pipeline: pose estimation followed by temporal modeling or rule-based assessment for technique analysis, counting, phase segmentation, and feedback [**?**]. Beyond recognition, pose-based scoring correlates with expert judgments in judged sports [**?**], and computer vision supports biomechanical analysis in both individual and

team settings [**?** **?**]. In dance, pose-based evaluation attains high agreement with expert ratings despite occlusions and rapid motion [**?**].

Sports datasets. Large-scale action sets (e.g., Sports-1M [**?**], Kinetics [**?**]) initiated video recognition but are coarse-grained and RGB-centric, offering limited pose or phase supervision. Fine-grained, judged-sport benchmarks add temporal structure and skill labels, e.g., Diving48 [**?**] and FineGym [**?**], for subtle technique discrimination and phase localization. Pose-annotated sets such as Penn Action [**?**] and PoseTrack [**?**] enable skeleton-based analysis and tracking, while domain datasets target tactics (SoccerNet [**?**], VNL-STES [**?**]) to study multi-person dynamics. Despite this breadth, we find no public dataset for *pole dancing*. Existing resources neither capture its contact-rich setting, frequent self-occlusions, and brief discriminative end-poses, nor address the privacy constraints that make appearance-based RGB release problematic. In contrast to sports benchmarks centered on RGB appearance, PoleArina emphasizes privacy-preserving motion representations for contact-heavy static pole poses, where sustained body-pole interaction and fine-grained end-pose structure are central.

3. Dataset

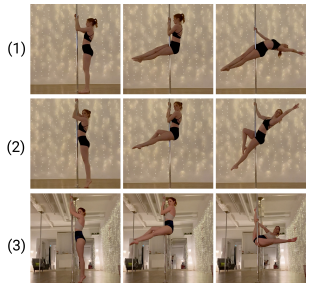
We introduce a privacy-preserving, skeleton and optical flow dataset for fine-grained recognition of *static* pole tricks. Our data consists of per-frame pose sequences (obtained as MediaPipe poses [**?**], having 75 joints with confidence values), computed optical flow using RAFT, and phase-aware annotations suitable for trick recognition and end-pose evaluation.

3.1. Data Collection & Labeling

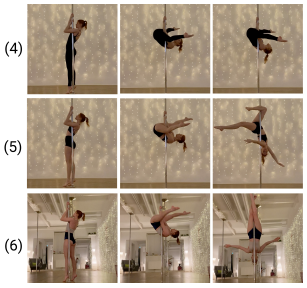
The data collection goal was to create an ethical, realistic representation of fundamental static pole tricks. To that end, the dataset targets six foundational tricks and balances feasibility for novices with sufficient discriminability for modeling. Annotations structure each clip into semantically meaningful temporal states.

Scope and terminology. Pole dance comprises *dance moves* (at least one foot stays on the floor), *spins* (airborne rotations), *floor work* (near-ground movement), and *tricks* (static/semi-static shapes). Poles operate in *static* or *spinning* mode. We adopt terminology from Spin City's Pole Bible and IPSF [**?** **?**]. We target *static pole tricks* to obtain consistent camera-facing orientation, fewer self-occlusions, and reduced motion blur, conditions that improve pose estimation and geometric measurement on commodity video. Each trick follows three phases: *entry* (approach/contact), *transition* (movement on pole), and *end pose* (held target shape).

Selection criteria. To balance feasibility, safety, and discriminability, we included six foundational tricks across two posture types: upright (*Layout*, *Pin-Up*, *Wrist Seat*) and inverted (*Straddle Invert*, *Gemini*, *Crucifix*) - Figure 2. Upright tricks share similar entries but end in distinct shapes; Layout vs. Pin-Up form intentional near-neighbors (Figure 2a). Inverted tricks begin from a basic invert grip and diverge to clearly different end poses (Figure 2b), adding biomechanical complexity while remaining achievable for athletic beginners to lower-intermediate dancers.



(a) 1: Layout, 2: Pin-Up, 3: Wrist Seat



(b) 4: Straddle Invert, 5: Gemini, 6: Crucifix

Figure 2. Progression of each trick, highlighting similar entries and transitions before the final pose.

3.1.1. Data Acquisition

Collecting high-quality and representative data was a crucial step, as together with the annotations, it provides the ground truth for our dataset generation and model evaluation. The aim was to assemble a large enough collection of clips, covering a balanced distribution of the six selected tricks and including both successful and failed attempts. We sought balanced sampling across tricks, experience levels, gender, and age; nevertheless, residual demographic skew and estimator-specific biases may persist. Beyond the scale and diversity of the data, particular attention was given to the ethical aspects of data collection.

We acquired videos via two routes: (i) *online submissions* through a form with a clearly-defined consent protocol, standardized filming guidance (smartphone, frontal full-body framing, even lighting, minimal background), and direct upload; and (ii) *in-class recordings* in dedicated sessions where an instructor demonstrated and supervised each trick with a fixed camera, ensuring consistent viewpoints, safer spotting for beginners, and better balance across the six tricks. Participants were encouraged to submit multiple attempts (including incomplete or failed tries) to capture natural variability.

The final dataset comprises 836 clips from $N = 58$

participants. The protocol deliberately encouraged multiple attempts per participant to capture natural variability in the progression of a trick. Section 3.3 presents a detailed breakdown of the final dataset. Table 3 compiles concise, syllabus-aligned instructions and categories for the six selected tricks.

Ethical considerations. All participants were fully informed about the project, data handling, and withdrawal rights, and provided written consent; collection and processing complied with local data-protection policies. In pole sports, participants typically wear short shorts and a sports bra/tank top to maximize friction with the pole, so identity cues extend to the full body. Combined with mirrors and distinctive studio backgrounds, credible de-identification would require near full-body obfuscation and extensive background handling, substantially reducing video utility. Hence, to protect privacy, we only use and release skeleton joint coordinates, optical flow, and per-video and per-frame labels; raw videos are used solely for creating the dataset and are not shared. No personal identifiers, faces, audio, or background content are stored.

3.1.2. Annotation Protocol

Our dataset provides two annotation granularities: per-video labels and per-frame labels. The per-video labels capture the performed trick (`L`, `P`, `W`, `S`, `G`, `I`), while per-frame labels capture the broad temporal structure of each clip through states such as `floor`, `on_pole`, and trick-specific pose labels (e.g., `*_pose`), enabling supervision of temporal progression from setup to execution. As expected for unconstrained performance videos, the frame-level label distribution is imbalanced, with support states (floor and on_pole) occupying a larger fraction of frames than terminal trick poses; to support transparent benchmarking, we provide the full label definitions and class frequencies with the dataset. The annotation protocol could be extended in future work toward finer phase-aware per-frame labels for the support states, e.g., by separating entry types and intermediate transition states.

Frame labels were annotated with expert guidance at frame-accurate temporal boundaries. We do not provide subjective per-phase quality scores, as these were found to be less consistent across clips; instead, form assessment is handled separately through a geometric rule-based module. Annotation efficiency and consistency across the full dataset were supported by a lightweight custom annotation tool. The dataset further includes per-video metadata such as the trick label and the performer's experience level.

3.2. Dataset Generation

Our pipeline converts raw training videos into stable, privacy-preserving representations for learning and evaluation. Instead of operating on pixels, we extract body skeletons and optical flow vectors of all the recorded videos.

Method	Missed Frames	Avg. # Joints	Avg. Conf with Face	Avg. Conf. w/o Face
MediaPipe	0.12%	60/75	0.76	**0.76**
MMPose	**0.09%**	129/133	**0.77**	0.69
OpenPifPaf	38.9%	52/133	0.27	0.15
OpenPose	2.9%	37/67	0.39	0.39

Table 1. Quantitative results of multiple skeleton extractors run on the entire video collection.

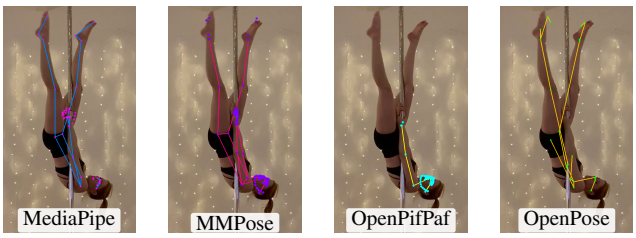


Figure 3. Qualitative example of all extractors on one single frame.

Skeleton Extraction. Accurate assessment of pole tricks requires reliable detection of both fingers and toes. We therefore evaluate four full-body pose estimators (Figure 3): OpenPose BODY_25 with hand keypoints [**?**], MediaPipe Holistic (model complexity 2, confidence threshold 0.3) [**?**], MMPose (RTMPose-L, 384 × 288 [**?**]) with COCO WholeBody keypoints [**?**], and OpenPifPaf (ShuffleNetV2K30-WholeBody [**?**]) with COCO WholeBody keypoints [**?**]. Some methods predict more keypoints than needed; for example, MMPose and OpenPifPaf include 68 facial landmarks that are not relevant to our use case.

For all four methods, we compute extraction statistics over the full dataset, including the percentage of missed frames (defined as all joints having confidence below 0.3), the average number of detected joints per frame, and the average confidence score (Table 1). MMPose yields the lowest number of missed frames and the highest overall confidence on the full keypoint set. However, after excluding facial keypoints, which are not useful in our setting, its average confidence drops significantly, and the number of missed frames increases by 31. MediaPipe, by contrast, remains only slightly behind in total missed frames (262 vs. 201) and shows a more favorable error pattern: most of its missed frames (156) occur at the end of videos, when no person is visible, whereas MMPose failures are more often scattered throughout the central parts of clips, including contiguous intervals. For this reason, we select MediaPipe both as the released skeleton representation and as input to our example action-recognition pipeline.

Trick	Layout	Pin-Up	Wrist Seat	Straddle Invert	Gemini	Inverted Crucifix
#Videos	188	167	131	144	88	94

Table 2. Number of videos per trick end-pose.

Optical Flow Extraction. We extracted optical flow with RAFT [**?**] using the official PyTorch implementation with the `RAFT-Large` model and default inference settings. All videos were first standardized to 30 fps and resized by setting the shorter image edge to 300 px while preserving aspect ratio to reduce computation and ease data sharing. We then computed dense forward flow between consecutive frame pairs only ($t \rightarrow t{+}1$).

3.3. Final Dataset Statistics

Per-video class balance. Table 2 reports per-trick video counts with at least one end-pose. Upright tricks (Layout, Pin-Up) are more frequent than inverted ones, reflecting their lower difficulty. In total, 812/836 clips include one end-pose, 9 of which contain a second attempt, and 24/836 are failed attempts.

Phase distribution. Regarding the per-frame phase distribution, phase `on_pole` immediately stands out, as it holds the majority share of 60.0% of all frames. It is followed by `floor` with 20.5%, while the trick-specific labels contribute only a small fraction individually, with around 2-4% and a combined value of 19.5%. By design, each recording typically shows a single target trick but always includes background, idle, and transition segments. As a result, `on_pole` and `floor` dominate frame counts.

Demographics. To reflect real studio populations while maximizing variability from failed to perfect attempts, we prioritized novice-friendly, beginner–intermediate tricks and encouraged multiple tries. The dataset includes $N = 58$ participants, deliberately emphasizing limited prior experience to diversify entries, transitions, and final shapes: 0=total beginner, 1=intermediate, 2=advanced. Gender distribution mirrors typical class composition: Female n=43 (74.1%), Male n=15 (25.9%), and ages span 20–58 years (median 28). We provide the experience labels as part of our dataset.

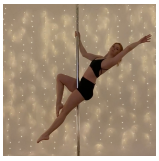
Trick	Description	Category / Level
 Layout	1. Stand on the right side of the pole 2. Pull up, cross legs at ankles 3. Lean back, arch, push hips up	Upright Beginner
 Pin-Up	1. Stand on the right side of the pole 2. Pull up, right toes to left knee 3. Lean slightly back and arch	Upright Beginner
 Wrist Seat	1. Stand on the right side of the pole 2. Pull up, right toes to left knee 3. Place left hand underneath the thigh 4. Lean back, open legs into a V shape	Upright Beginner
 Straddle Invert	1. Stand behind the pole, facing left 2. Stronghold grip, pull up, lean back 3. Open legs into a V shape	Inverted Intermediate
 Gemini	1. Start with Straddle Invert 2. Hook the outside leg, other leg down 3. Chest up, arch and release hands	Inverted Intermediate
 Inverted Crucifix	1. Start with Straddle Invert 2. Place legs into crucifix hold 3. Upper body low, release arms	Inverted Intermediate

Table 3. Compact summary of the selected tricks; terminology aligned with IPSF and Spin City [**? ?**].

4. Trick Recognition & Analysis

Our technical contributions demonstrate the value of the dataset while exposing the remaining challenges of privacy-preserving pole-trick understanding. Specifically, we benchmark skeleton-based and optical-flow-based action recognition under both per-video and per-frame labeling. As a unifying design choice, all action recognition models use a lightweight GRU-based architecture to provide a consistent baseline across modalities and to emphasize dataset difficulty rather than architecture-specific gains. All models use the same fixed train/validation/test split (70/15/15) to ensure fair comparison across modalities and supervision settings. The coaching application enabled by this pipeline

is evaluated separately in Section 5.

4.1. Skeleton-Based Trick Recognition

Data preprocessing. For skeleton-based recognition, each clip is represented as a variable-length sequence of frames, each consisting of 75 joints with (x, y, conf) values. Missing detections at sequence boundaries are trimmed, short internal missing frames are linearly interpolated, and a light temporal smoothing is applied to reduce jitter. Training-only augmentation includes mild spatial perturbations (rotations and Gaussian noise) and temporal warping. For the per-frame setting, the same preprocessing is used, while frame-wise labels are aligned to the processed sequence, and unlabeled or padded positions are masked out during training. To mitigate the strong class imbalance in frame-wise labels, we additionally apply class-aware augmentation during training, with stronger augmentation for underrepresented classes.

Model architecture. We use a lightweight 2-layer bidirectional GRU backbone (hidden size 64, dropout 0.2) for both skeleton-based settings, taking the flattened per-frame joint representation as input. For video-level classification, the final forward and backward hidden states are concatenated and passed to a linear classifier. For per-frame prediction, the recurrent output at each timestep is decoded by a shared linear head to produce frame-wise logits. Models are trained with cross-entropy, using masking for padded or unlabeled frames in the per-frame setting.

4.2. Optical Flow-Based Trick Recognition

Data preprocessing. Optical-flow recognition operates on precomputed dense flow fields, represented as variable-length sequences of two-channel motion frames. Flow sequences are temporally subsampled (take 16, skip 5), resized to a common processing resolution (maximum size 192), normalized using statistics computed on the training split, and augmented only during training. For the per-frame setting, frame-wise labels are assigned from the annotations, and padded or unlabeled positions are masked out.

Model architecture. For optical flow, each frame is first encoded by a 2-layer compact CNN and then aggregated temporally with a 2-layer GRU. In the video-level setting, the final GRU state is used for clip classification. In the per-frame setting, the GRU output at each timestep is decoded by a shared linear layer to obtain frame-wise predictions. This design keeps the temporal backbone comparable to the skeleton-based models while introducing only a modality-specific spatial encoder.

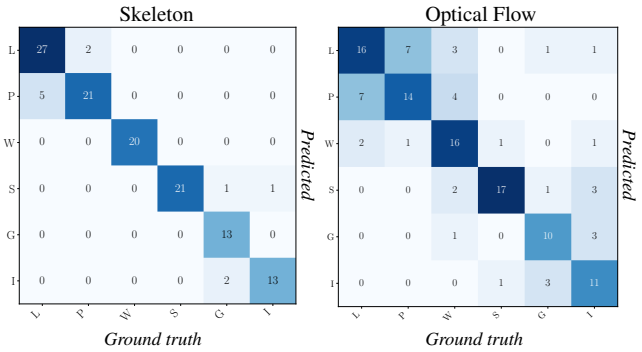


Figure 4. Confusion matrices for our per-video models on both modalities.

4.3. Results and Discussion

We report results for two supervision granularities: (i) *per-video* action recognition and (ii) *per-frame* state/trick recognition. For each task, we evaluate the two privacy-preserving modalities introduced in Pole-Arina: skeleton sequences and dense optical flow. Figures 4 and 5 show the corresponding confusion matrices, while Table 5 provides an overview of our results.

Per-video recognition (clip labels). On the shared test split, the skeleton-based video model achieves **91.26%** accuracy, while the optical-flow model reaches **66.67%** accuracy (Fig. 4). This gap is expected: skeleton extraction already provides a structured representation of body pose and motion, supplying a strong inductive bias for classification. In contrast, optical flow is a weaker and noisier proxy for articulation, particularly in unconstrained recordings with camera motion, background clutter, mirrors, and self-occlusions, leaving greater headroom for improved modeling on the flow stream.

Across both modalities, the dominant confusions occur between `Layout` and `Pin-Up`, which are visually and kinematically similar and thus represent a natural ambiguity in the label space. Additionally, errors tend to remain within the same broad regime: upright tricks are more commonly confused with other upright tricks, and inverted tricks with other inverted tricks. Mixing upright with inverted does not happen for the skeleton model, indicating that the pose representation strongly disambiguates inversion state; it is less rare (though still uncommon) for optical flow, consistent with flow providing weaker explicit cues about body orientation and inversion.

Per-frame recognition (frame labels). For frame-wise classification, the skeleton model achieves **86.99%** accuracy with a **77.71%** macro F1-score, whereas optical flow reaches **81.23%** accuracy but only **45.84%** macro F1 (Fig. 5). This discrepancy confirms that overall frame accu-

Modality	floor	on_pole	L	W	P	S	I	G
Skeleton	87.98	90.35	77.19	81.54	76.43	76.07	86.24	45.89
Optical Flow	88.08	88.03	52.33	68.22	7.99	62.02	0.00	0.00

Table 4. Per-class F1-scores (%) for per-frame recognition, highlighting the stronger effect of class imbalance on optical flow than on skeleton-based recognition.

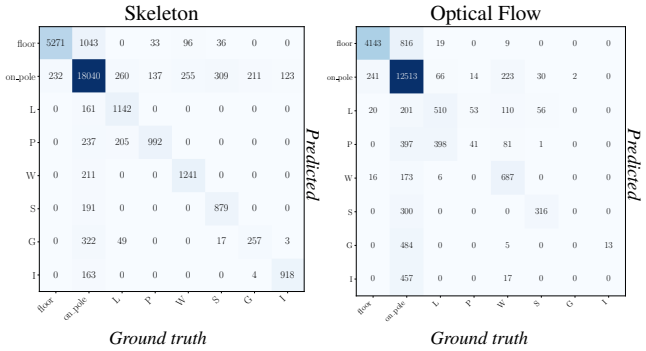


Figure 5. Confusion matrices for our per-frame models on both modalities.

Modality	Task	Acc. (%)	Macro F1 (%)
Skeleton	Per-video	91.26	91.50
Optical Flow	Per-video	66.67	67.11
Skeleton	Per-frame	86.99	77.71
Optical Flow	Per-frame	81.23	45.84

Table 5. Summary of benchmark results across modalities and supervision settings. For the imbalanced per-frame task, macro F1 is more informative than overall accuracy alone.

racy is strongly inflated by class imbalance: the dominant `on_pole` and `floor` states account for most frames and are recognized comparatively well in both modalities, while minority trick classes remain substantially harder.

The class-wise skeleton results show that several trick classes remain meaningfully distinguishable despite this imbalance, including `I` (86.24% F1), `W` (81.54% F1), `L` (77.19% F1), `P` (76.43% F1), and `S` (76.07% F1). The main exception is `G` (`gemini`), which is substantially harder at only 45.89% F1. In contrast, the optical-flow model performs reasonably on the dominant states (`floor`: 88.08% F1, `on_pole`: 88.03% F1) and on some trick classes such as `W` (68.22% F1) and `S` (62.02% F1), but fails on `G` and `I`, both of which obtain 0.00% F1. This indicates that flow alone is much less reliable for fine-grained pose discrimination under severe imbalance, as reflected in Table 4.

Overall, skeletons provide a strong and compact representation for frame-level recognition, with substantially better class-sensitive performance across most trick classes. Optical flow remains useful for dominant motion states, but currently underperforms on fine-grained minority poses. These results support the dual-modality design of Pole-Arina: skeletons offer a strong baseline for privacy-preserving pose recognition, while optical flow defines a more challenging appearance-suppressed setting with clear room for methodological improvement.

4.4. Limitations

We intentionally adopt compact GRU-based models to support low-latency coaching use, and therefore present them as practical baselines rather than capacity-saturating architectures. Skeleton-based results are also *conditioned on pose estimation*: extracting skeletons already introduces a strong inductive bias, and errors or biases of the selected estimator, especially under occlusion and sustained pole contact, directly affect both recognition and geometric scoring. While optical flow reduces appearance leakage, it remains sensitive to camera motion and mirror artifacts. These dependencies are part of the challenge captured by Pole-Arina and point to clear directions for future work in more robust privacy-preserving sports analysis.

5. Coaching Feedback Case Study and User Study

Beyond benchmark recognition, we use Pole-Arina to study a practical downstream application: coaching-oriented feedback for static pole tricks. Our goal is not to replace expert instruction, but to examine whether privacy-preserving motion representations and interpretable pose analysis can support a useful feedback workflow. To this end, we first define a transparent geometric scoring system for end-pose assessment and then evaluate its integration into a prototype coaching application through a user study.

5.1. Geometric Scoring System

We design an interpretable scoring module for skeleton-based end-pose assessment. We considered an autoencoder-based approach that would learn an *ideal* pose and score deviations through reconstruction error [**?**], but such a formulation would require many near-perfect exemplars and would be highly sensitive to dataset bias. Instead, we adopt a geometric, rule-based scorer: it is transparent, aligns with instructor practice through angles, alignments, and relative joint positions, and avoids the subjectivity often discussed in dance evaluation [**?**].

Each trick is scored by **5–7** geometric rules that capture its defining cues: (i) *body orientation* with respect to the pole, (ii) *limb alignment* through joint angles, and (iii) *joint proximity* through characteristic distances, such as ankles being close together or toe–knee contact.

Each rule specifies a target value and tolerance for the required joint angle or distance. A rule is evaluated only if the corresponding joint confidences satisfy $\min(v_i) \geq \tau$,

with $\tau = 0.75$. We then compute the relevant angles and distances from the joint coordinates and assign a *pass* or *fail* decision according to the target and tolerance. The end-pose score is defined as the fraction of passed evaluable rules,

$$\text{score} = \frac{\#\text{passed}}{\#\text{evaluable}}.$$

For interpretability, the system reports per-rule feedback together with the measured values and target ranges. We will provide the complete rule catalog for all six tricks, including targets and tolerances, as part of the benchmark.

5.2. User Study

To assess the practical utility of our pipeline, we integrated the per-frame skeleton-based recognition model and geometric scorer into an end-to-end coaching-guide web application. Given a recorded training video, the system extracts privacy-preserving pose skeletons, predicts the trick label, and computes a geometric end-pose score, returning structured feedback. Computation is centralized on a FastAPI+PyTorch server, while a React/Next.js client runs on consumer laptops and supports a coach-like workflow: upload → trick gallery → per-rule feedback (Figure 6).

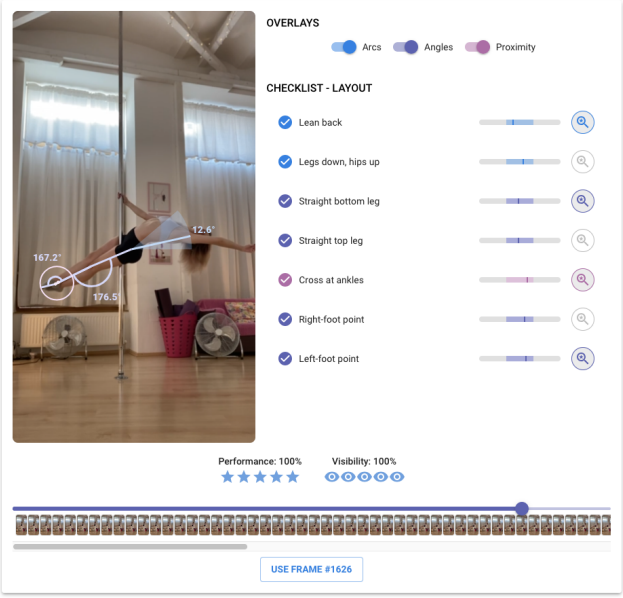


Figure 6. Application interface showing the frame viewer, overlay controls, and per-rule feedback for the predicted trick.

We conducted a between-subjects study with 33 participants comparing a *Control* condition (smartphone video self-review) against an *Experimental* condition (Pole-Arina feedback). Participants were randomly assigned to one condition, performed five practice trials with review after each trial using experience-matched tricks (`Layout` and `Pin-Up`), and then completed post-session Likert ratings and the System Usability Scale (SUS) [**?**], together with optional qualitative feedback and debriefing.

Results. Normality tests (Kolmogorov–Smirnov and Shapiro–Wilk) indicated non-normal distributions, so we used Mann–Whitney U tests for between-condition comparisons. In our between-subjects study ($N = 33$; 17 Experimental, 16 Control), the Pole-Arina condition yielded higher ratings for trust/adoption, clarity, and usability, including a higher SUS score (95.44 ± 4.07 vs. 86.41 ± 11.14; $U = 72.00, p = .020$), while efficiency-related measures showed no significant difference ($p = .485$). Participants' self-ratings generally trended below the system scores. Qualitative feedback highlighted the usefulness of the overlays and expressed interest in future additions such as reference poses and real-time feedback.

These results suggest that interpretable geometry-based feedback can be useful in practice as a downstream application of our benchmark. While this user study does not constitute a benchmark evaluation in itself, it indicates that combining accurate skeleton-based recognition with transparent rule-based explanations can support a coaching-oriented workflow.

6. Conclusion

We introduce the first privacy-preserving benchmark for fine-grained pole dance analysis, an underrepresented yet increasingly popular sport that poses distinctive challenges (contact constraints, self-occlusions, mirrors, and appearance obfuscation). We provide skeleton and optical-flow modalities with expert annotations, enabling learning without releasing identifiable RGB. Strong baseline results from our lightweight skeleton- and flow-based models demonstrate that meaningful recognition is possible from these representations, while leaving substantial headroom for improved temporal modeling, robustness, and feedback quality. Finally, we show how these components can be integrated into a coaching-oriented prototype and evaluated with a user study, connecting benchmark performance to practical feedback. Together, the dataset, baselines, and evaluation establish pole dance as a new, well-defined computer vision problem setting for privacy-aware, fine-grained athletic technique understanding.

References

[] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. PoseTrack: A benchmark for Human Pose Estimation and Tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5167–5176, 2018.

[] Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann.

BlazePose: On-device Real-time Body Pose tracking. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[] John Brooke. SUS-A quick and dirty usability scale. *Usability Evaluation in Industry*, 189(194):4–7, 1996.

[] Yuanhao Cai, Zhicheng Wang, Zhengxiong Luo, Binyi Yin, Angang Du, Haoqian Wang, Xiangyu Zhang, Xinyu Zhou, Erjin Zhou, and Jian Sun. Learning Delicate Local Representations for Multi-person Pose Estimation. In *Computer Vision – ECCV 2020*, pages 455–472, Cham, 2020. Springer International Publishing.

[] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, 2019.

[] 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*, 2017.

[] Spin City. *The Ultimate Pole Bible*. Spin City Aerial Fitness Ltd, 2025.

[] Steffi L Colyer, Murray Evans, Darren P Cosker, and Aki IT Salo. A Review of the Evolution of Vision-Based Motion Analysis and the Integration of Advanced Computer Vision Methods Towards Developing a Markerless System. *Sports Medicine - Open*, 4:1–15, 2018.

[] MMPose Contributors. OpenMMLab Pose Estimation Toolbox and Benchmark. https://github.com/open-mmlab/mmpose, 2020.

[] Adrien Deliège, Anthony Cioppa, Silvio Giancola, Meisam J. Seikavandi, Jacob V. Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B. Moeslund, and Marc Van Droogenbroeck. SoccerNet-v2: A Dataset and Benchmarks for Holistic Understanding of Broadcast Soccer Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2021.

[] Zhao Du, Shan Wang, Ziyan Deng, and Fang Wang. Unveiling the Power of AI Fitness Apps: A Uses and Gratifications Perspective. *Journal of Global Information Management (JGIM)*, 33(1):1–28, 2025.

[] Aysu Ezen-Can. A Comparison of LSTM and BERT for Small Corpus. *arXiv preprint arXiv:2009.05451*, 2020.

[] International Pole Sports Federation. Code of Points 2025 – 2027. https://ipsfsports.org/downloads/Uncategorised/ipsf_pole_sports_code_of_points_2025-2027_final_070120240.pdf, 2025. Accessed: 2025-08-16.

[] Indrajeet Ghosh, Sreenivasan Ramasamy Ramamurthy, Avijoy Chakma, and Nirmalya Roy. Sports analytics review: Artificial intelligence applications, emerging technologies, and algorithmic perspective. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13(5):e1496, 2023.

[] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997.

[] Filip Ilic, Thomas Pock, and Richard P. Wildes. Is Appearance Free Action Recognition Possible? In *European Conference on Computer Vision*, 2022.

[] Filip Ilic, He Zhao, Thomas Pock, and Richard P. Wildes. Selective, Interpretable and Motion Consistent Privacy Attribute Obfuscation for Action Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024.

[] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-Body Human Pose Estimation in the Wild. In *Computer Vision – ECCV 2020*, pages 196–214, Cham, 2020. Springer International Publishing.

[] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-Scale Video Classification with Convolutional Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[] Yeonho Kim and Daijin Kim. Real-time dance evaluation by markerless human pose estimation. *Multimedia Tools and Applications*, 77:31199–31220, 2018.

[] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. PifPaf: Composite Fields for Human Pose Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. OpenPifPaf: Composite Fields for Semantic Keypoint Detection and Spatio-Temporal Association. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–14, 2021.

[] Yingwei Li, Yi Li, and Nuno Vasconcelos. RESOUND: Towards Action Recognition without Representation Bias. In *Proceedings of the European Conference on Computer Vision*, pages 513–528, 2018.

[] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for perceiving and processing reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition*, 2019.

[] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design, 2018.

[] Marina Mikami and Noriyuki Kida. Categorizing Rhythmic Jumping Motion Using Motion Capture without Markers. *Advances in Physical Education*, 13(2):93–105, 2023.

[] Hoang Nguyen, Ankhzaya Jamsrandorj, Vanyi Chao, Yin May Oo, Muhammad Amrulloh Robbani, Kyung-Ryoul Mun, and Jinwook Kim. VNL-STES: A Benchmark Dataset and Model for Spatiotemporal Event Spotting in Volleyball Analytics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5862–5871, 2025.

[] Paritosh Parmar and Brendan Tran Morris. Learning to Score Olympic Events. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–28, 2017.

[] Zhiqiang Pu, Yi Pan, Shijie Wang, Boyin Liu, Min Chen, Hao Ma, and Yixiong Cui. Orientation and Decision-Making for Soccer Based on Sports Analytics and AI: A Systematic Review. *IEEE/CAA Journal of Automatica Sinica*, 11(1):37–57, 2024.

[] Jiping Qu. A dance movement quality evaluation model using transformer encoder and convolutional neural network. *Scientific Reports*, 14(1):32058, 2024.

[] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. FineGym: A Hierarchical Video Dataset for Fine-grained Action Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[] Karen Simonyan and Andrew Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, page 568–576, Cambridge, MA, USA, 2014. MIT Press.

[] Xiang Suo, Weidi Tang, and Zhen Li. Motion Capture Technology in Sports Scenarios: A Survey. *Sensors*, 24(9):2947, 2024.

[] Zachary Teed and Jia Deng. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In *European Conference on Computer Vision*, pages 402–419. Springer, 2020.

[] Atima Tharatipyakul, Thanawat Srikaewsiew, and Suporn Pongnumkul. Deep learning-based pose estimation in providing feedback for physical movement: A review. *Heliyon*, 10(17):e36589, 2024.

[] Alexander Toshev and Christian Szegedy. DeepPose: Human Pose Estimation via Deep Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1653–1660, 2014.

[] Michael Tschannen, Olivier Bachem, and Mario Lucic. Recent Advances in Autoencoder-Based Representation Learning. *arXiv preprint arXiv:1812.05069*, 2018.

[] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All You Need. *Advances in Neural Information Processing Systems*, 30, 2017.

[] Charlene Weaving. Sliding Up and Down a Golden Glory Pole: Pole Dancing and the Olympic Games. *Sport, Ethics and Philosophy*, 14(4):525–536, 2020.

[] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From Actemes to Action: A Strongly-supervised Representation for Detailed Action Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2248–2255, 2013.