Proficiency Estimation

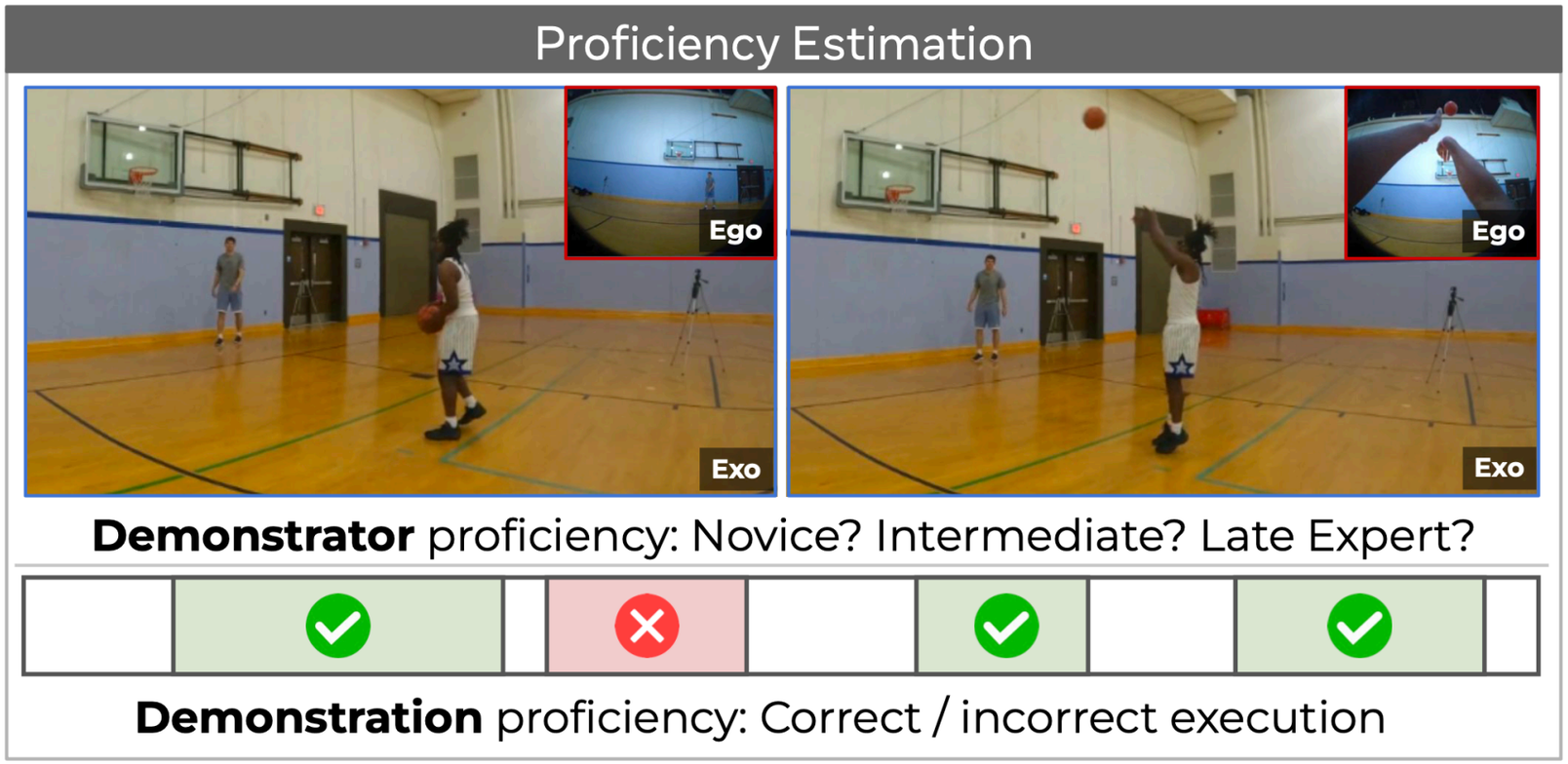
Final Project Update

Noah Frahm, Andrea Dunn Beltran

790: 3D Generative Models Project Update

Understanding Proficiency Estimation

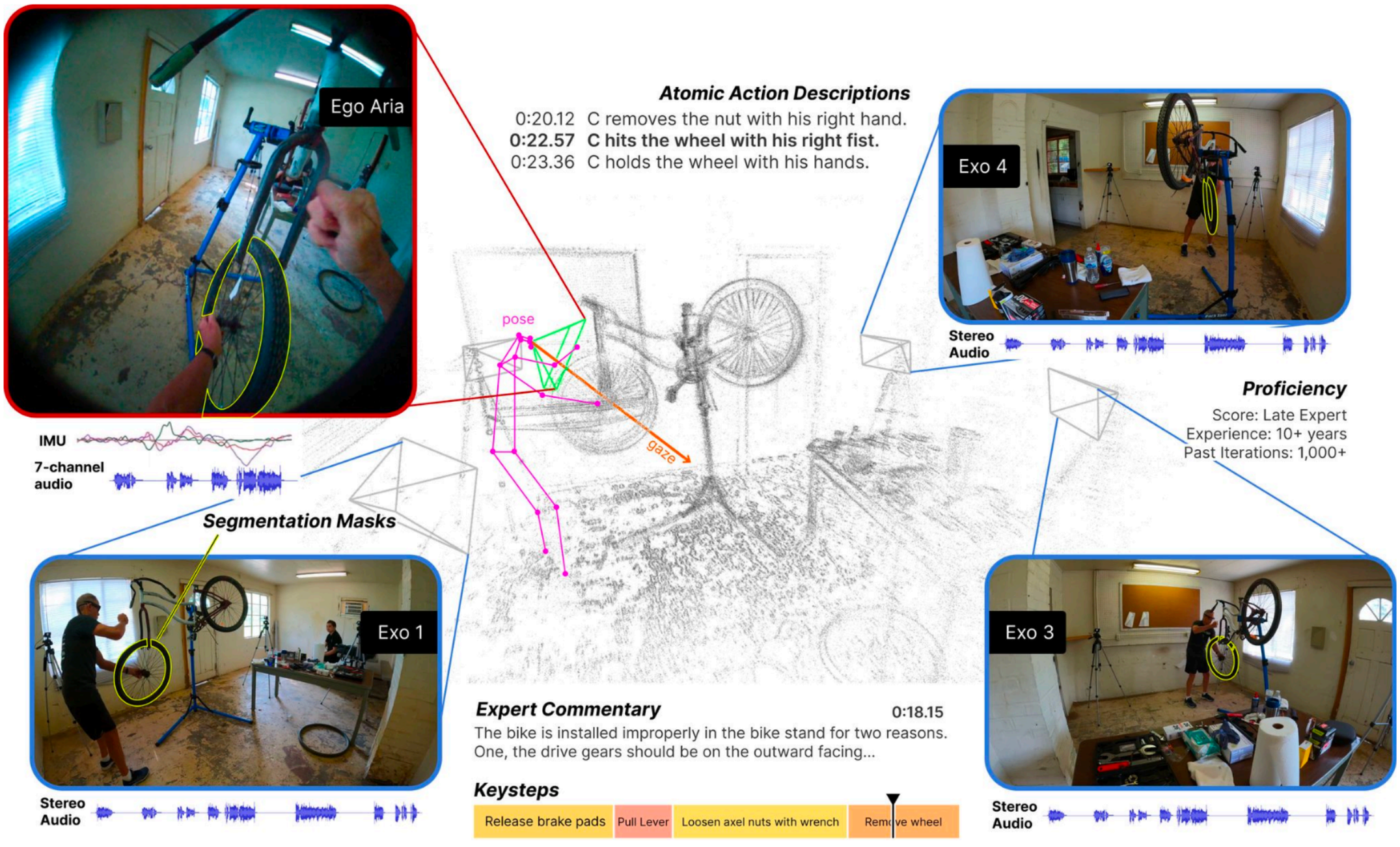
Analyzing Video for Skill Assessment



- Labeling/recognizing key moments for skill level assessment
- Analyzing video footage to determine skill level

Ego-Exo 4D Dataset

Diverse, large-scale multi-modal, multi-view, video dataset



Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives

Building on Baselines

Combining video and 3D features

1

Classification using **Video Features**

2

Classification using **3D Features**

3

Classification using **Video Features** & **3D Features**

Project Progress

Part 1

1 — 2 — 3 — 4

Project Pitch
Invested EgoExo4D dataset and identified missing information

Project Update 1
Recreated TimeSFormer baselines and tested different data splits

Project Update 2
Extracted 3D features using MotionAGFormer with aria validation

Constructed Transformer

Project Progress

Part 2

1 — 2 — 3 — 4

Final Project Update

Ran classification transformer on pose only data

Experimented with different tokenization techniques and tuned hyper parameters

Investigated different proficiency classes splits

EgoExo4D Benchmarks

Demonstrator Proficiency

EgoExo4D

Method	Pretrain	Accuracy		
	-	Ego	Exos	Ego + Exos
Random	-	25.7	25.7	25.7
Majority-class	-	43.0	43.0	43.0
TimeSFormer [13]	-	**43.3**	40.5	40.8
TimeSFormer [13]	K400	41.9	45.2	45.2
TimeSFormer [13]	EgoVLP	42.7	**52.3**	**51.8**
Inference with multiple takes per demonstrator				
TimeSFormer [13]	EgoVLP	46.3	51.0	52.6

Table 20. **Demonstrator proficiency estimation benchmark.** We report top-1 accuracies for various baselines on the demonstrator proficiency estimation task.

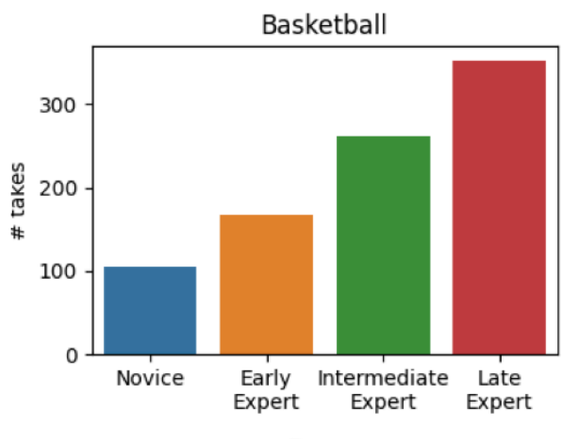
Scenario	Random	Majority-class	TimeSFormer
All Scenarios	25.25	30.25	37.21
Dance	23.96	48.04	43.55
Rock Climbing	25.60	33.06	-
Soccer	29.04	69.12	68.75
Basketball	25.56	42.88	47.29

Dwindling Dataset

Data availability

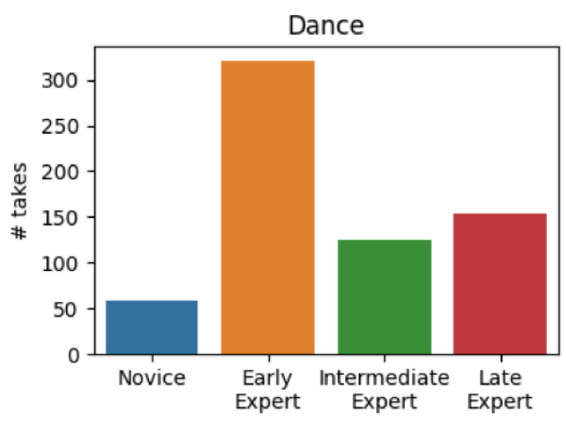
Basketball

- Good proficiency class spread
- Large dataset



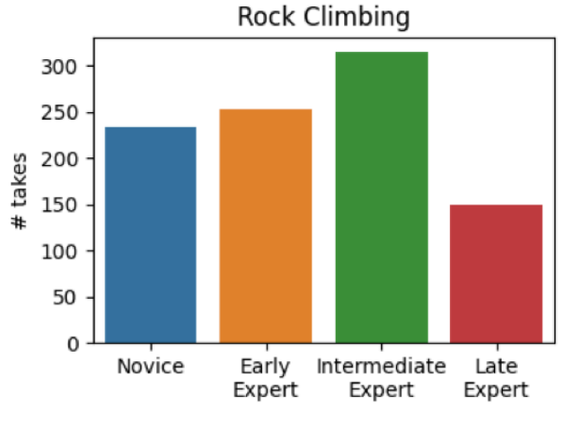
Dance

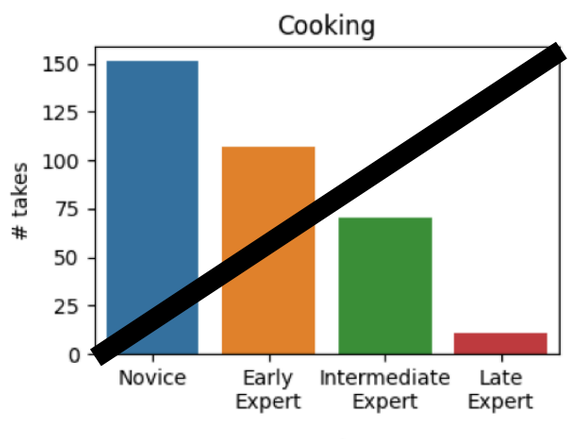
- Good proficiency class spread
- Large dataset



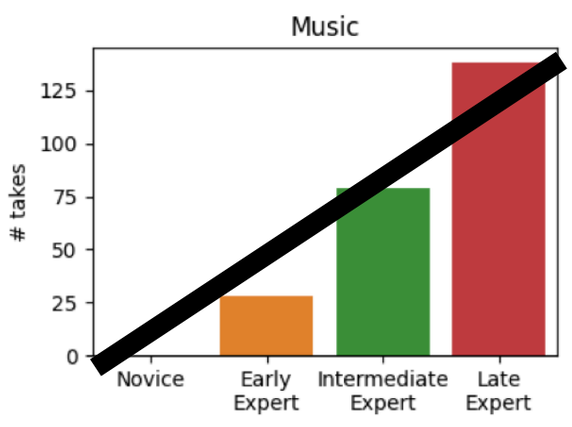
Rock Climbing

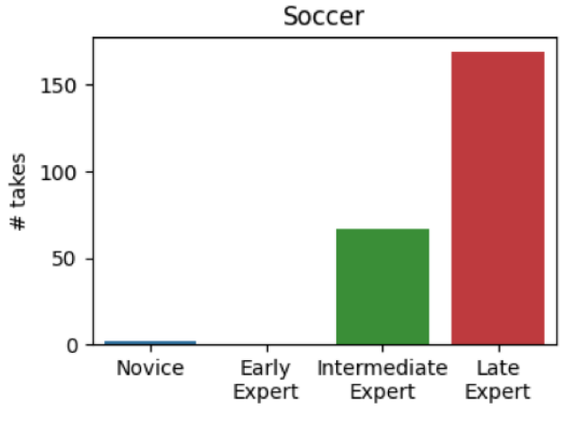
- Poor pose extraction due to heavy occlusions and camera angle





Not Physical Activities





Soccer

- Only has 2 proficiency classes
- Smallest task dataset

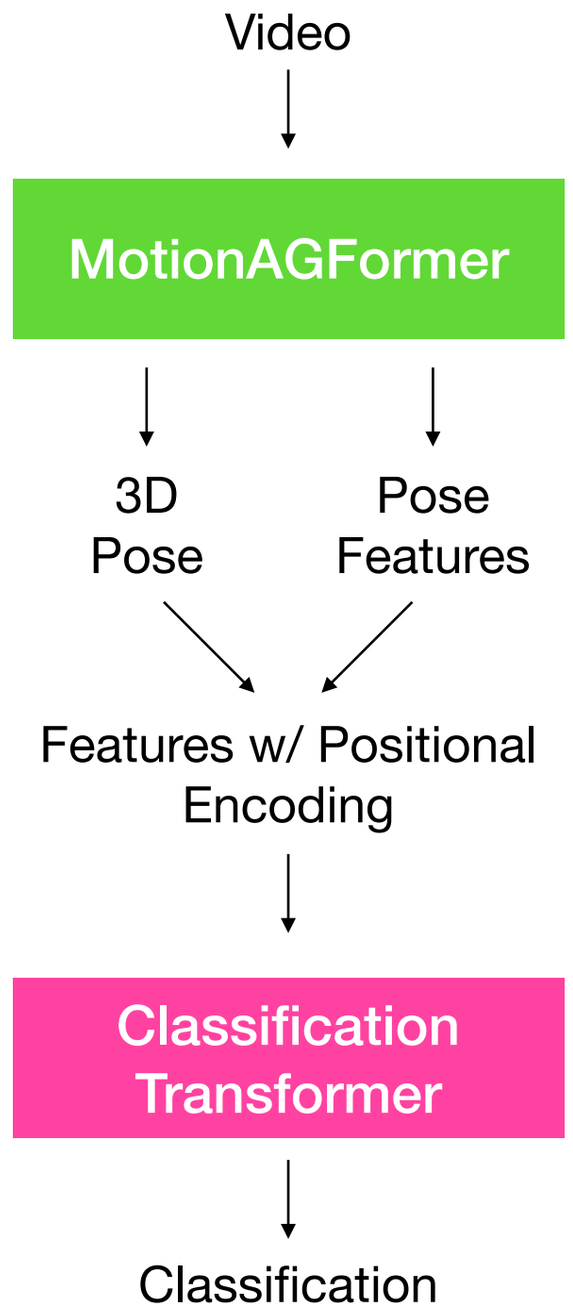
Classification:
3D Features

Old Pipeline

Per Frame:

Pose feature:	17 x 128
3D pose:	17 x 3
Encoded 3D Pose:	17 x 60
Token size:	17 x 188
Flattened:	1 x 3196

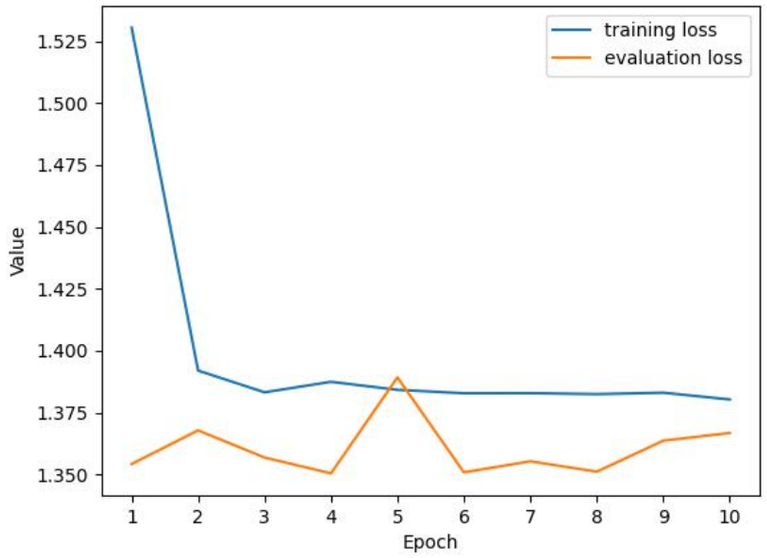
Input: 100 frame/video 100 x 3196



Results

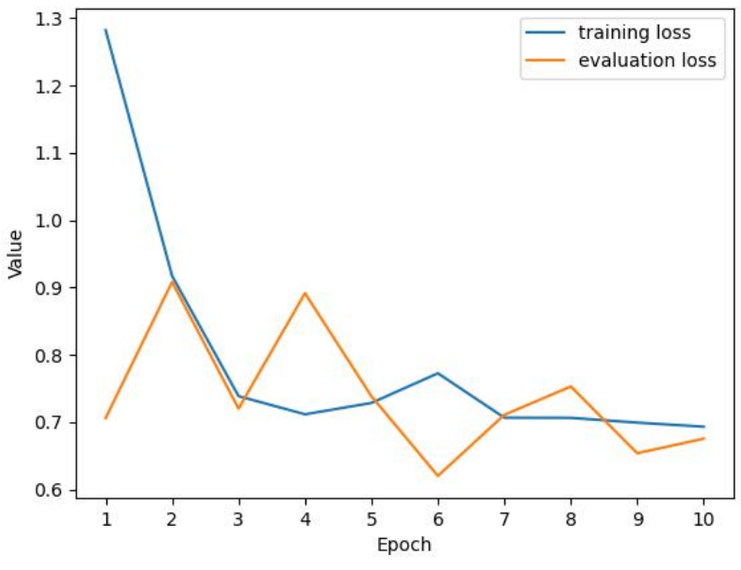
Old Pipeline

Soccer



Top1 Acc: 80.70%
Majority: 80.70%

Dance

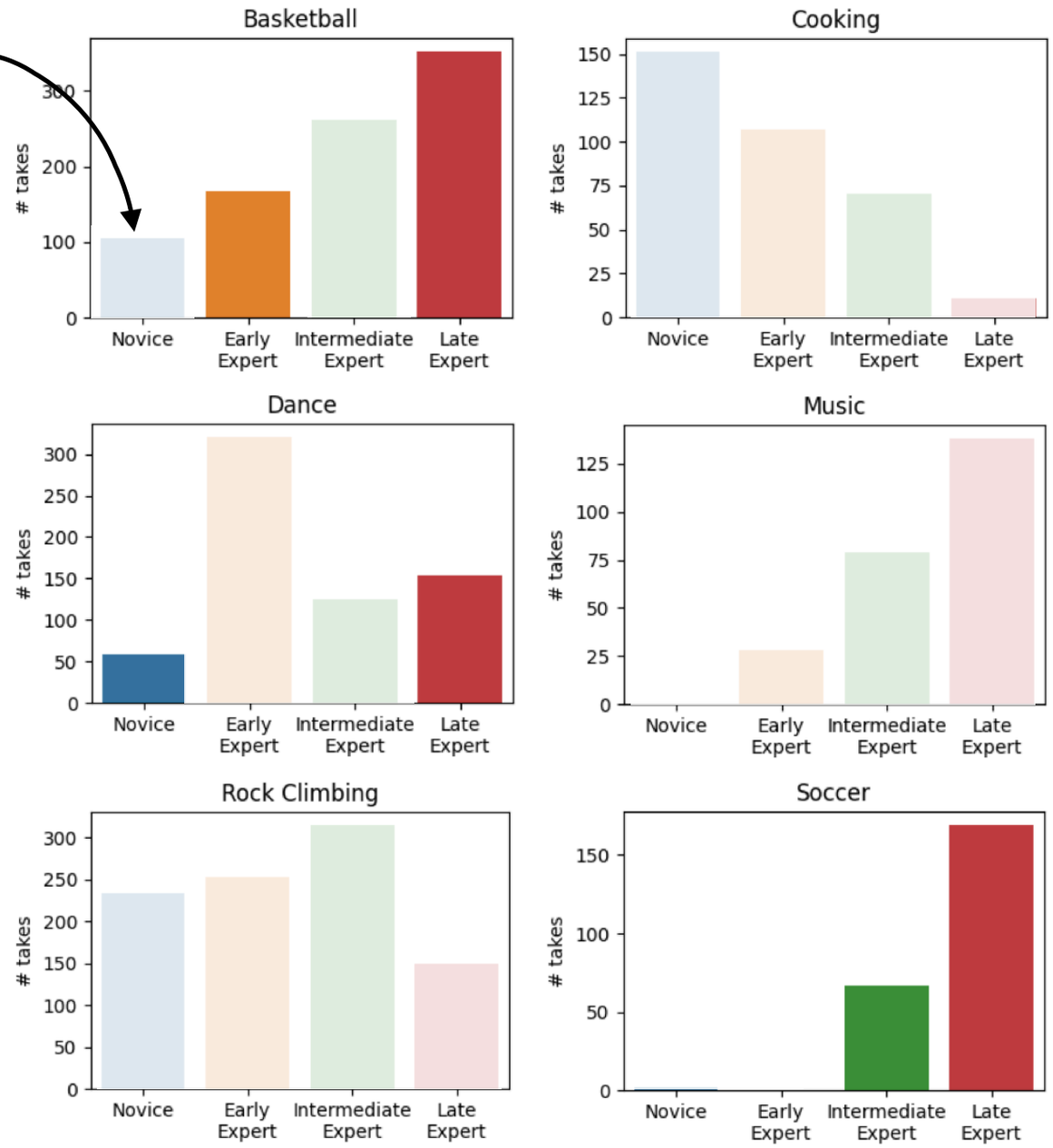


Top1 Acc: 62.20%
Majority: 62.20%

Binary vs Multi-class

Simplifying the Problem

not present in previous release



TimeSformer Results

Majority, Multiclass, Binary

Tasks	Number classes	TimeSFormer (video only)	
		Majority of Dataset	Top_1
Soccer	**2 classes**	0.625	0.6875
Dance	**4 classes**	0.516	0.4355
	2 classes	0.75	0.9688
Basketball	**3 classes**	0.434	0.4729
	2 classes	0.746	0.5479

Experimenting with Tokenization

Changing Transformer Inputs

Old: 100 fs x **(17 kpts x (60 encoded pose + 128 features))** ⟶ 100 x **(17x188)** = 100 x **(3,196)**

Body_Position: 17 kpts x **(100 fs x 3 pose)**
17 x **(300)**

Body_Enposition: 17 kpts x **(100 fs x 60 encoded pose)**
17 x **(600)**

Frame_Position: 100 fs x **(17 kpts x 60 encoded pose)**
100 x **(1021)**

Frame_Features: 100 fs x **(17 kpts x (3 pose + 128 features))**
100 x **(17 x 131)** = 100 x **(2,227)**

Other Ablations

Changing Transformer Inputs and Architecture

- **Linear Layer** vs no linear layer
  - Linear Layer + ReLU -> 512 dim
- **Normalized** vs denormalized pose
- Encoded pose (60 dim) vs **non-encoded** pose (3 dim)
- Multi-class vs **binary**

Metrics we considered: F1, AUC, Top-1

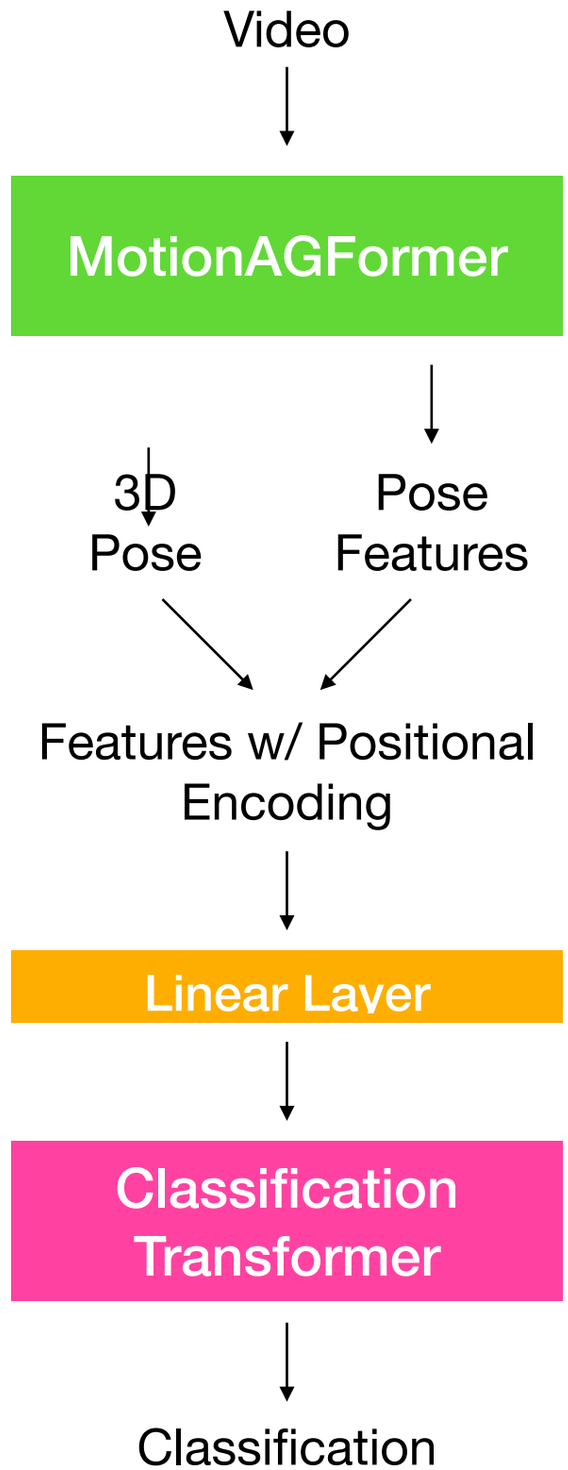
Classification:
3D Features

New Pipeline

Per Frame:

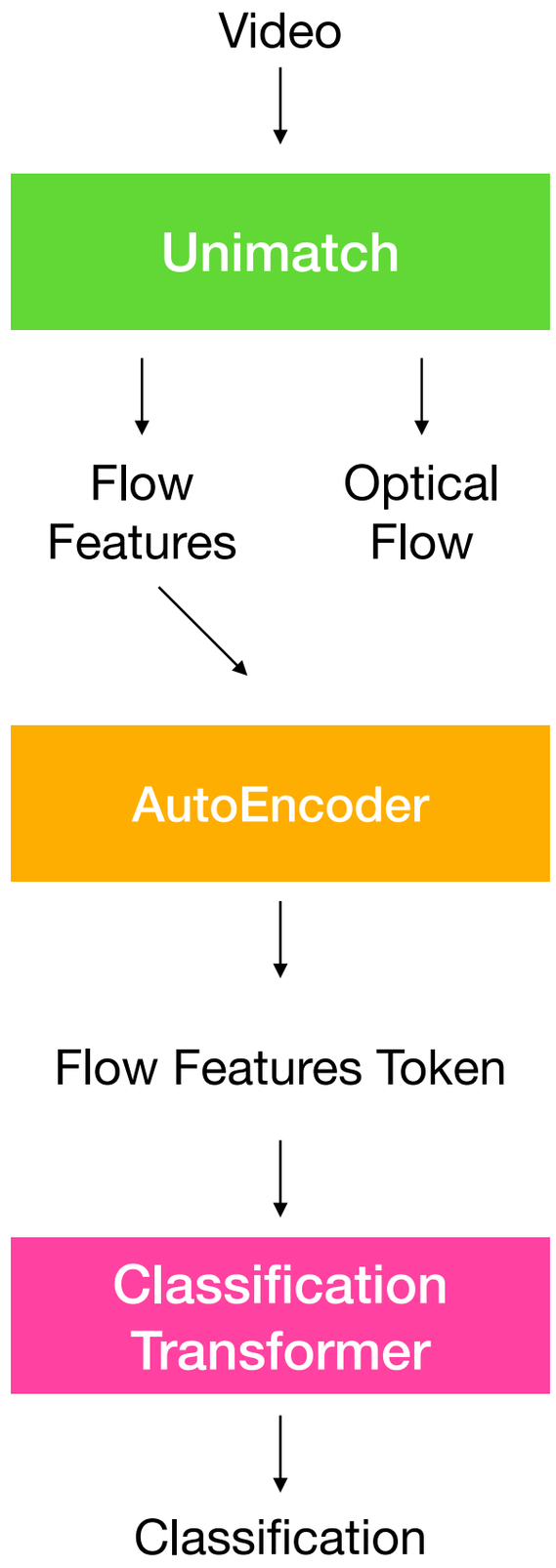
Pose feature:	17 x 128
3D pose:	17 x 3
Token size:	17 x 131
Flattened:	1 x 2227
Projected:	1 x 512

Input: 100 frame/video 100 x 512



Classification:
3D Features

New Pipeline



Per Frame:

Flow Feature:	128 x 64 x 120
Flattened:	128 x 1920
Token size:	1 x 1920

Pose Only Classification

3D feature Classification

Tasks	Number classes	Majority Guess			Pose Only Clasification		
		Top_1	F1	AUC	Top_1	F1	AUC
Soccer	**2 classes**	0.807	0.4464	0.5	0.679	0.413	0.313
Dance	**4 classes**	0.622	0.1916	0.5	0.452	0.453	0.64
	2 classes	0.756	0.4305	0.5	0.806	0.752	0.818
Basketball	**3 classes**	0.43	0.2004	0.5	0.494	0.319	0.654
	2 classes	0.672	0.4018	0.5	0.682	0.637	0.717

Model Details:

- NORMALIZED Features
- NO ENCODING on 3D pose
- Linear Layer

Soccer Results

Majority, Multiclass, Binary

Soccer	Top_1	F1	AUC
Majority Guess	**0.807**	**0.4464**	**0.5**
Pose Only Clasification	0.679	0.413	0.313

Dance Results

Majority, Multiclass, Binary

Dance 2 classes	Top_1	F1	AUC
Majority Guess	0.756	0.4305	0.5
Pose Only Clasification	**0.806**	**0.752**	**0.818**

Dance 4 classes	Top_1	F1	AUC
Majority Guess	**0.622**	0.1916	0.5
Pose Only Clasification	0.452	**0.453**	**0.64**

Basketball Results

Majority, Multiclass, Binary

Basketball 2 classes	Top_1	F1	AUC
Majority Guess	0.672	0.4018	0.5
Pose Only Clasification	**0.682**	**0.637**	**0.717**

Basketball 3 classes	Top_1	F1	AUC
Majority Guess	0.43	0.2004	0.5
Pose Only Clasification	**0.494**	**0.319**	**0.654**

TimeSformer Results

Majority, Multiclass, Binary

Tasks	Number classes	TimeSFormer (video only)	
		Majority of Dataset	Top_1
Soccer	**2 classes**	0.625	0.6875
Dance	**4 classes**	0.516	0.4355
	2 classes	0.75	0.9688
Basketball	**3 classes**	0.434	0.4729
	2 classes	0.746	0.5479

Next Steps

Tasks and Modalities

- Investigate other modalities
  - Optical Flow
  - Monocular Depth
- Fusing different modalities for tasks
- Expand to more tasks