Automated Temporal Localization of Behavior in Cichlid Fish with Few-Shot Trajectory Tokens

Bhargav Srinivasan*, William Lamousin*, Jason Liu†, Charles Phan†, Ofure Osunbor, Shubh Sharma, Cade McGeehan, Tahir Haroon, Vijay Jayasuriya, Coltan Parker, Scott Juntti, Pulkit Kumar, Abhinav Shrivastava
University of Maryland, College Park, USA
bds062@umd.edu

Abstract

Understanding the brain requires studying its observable output: behavior. Cichlid fish provide a model system for changing brain function and studying its effect on well-defined behaviors. Such experiments require manual annotation of aquarium video at real-time playback speed, which is time-consuming to generate and produces single time points. However, these annotations lack the temporal context of the underlying behavior. We provide a benchmarking dataset for animal behavior containing ∼17 hours of annotated footage of cichlid fish. We then propose FSH-TAL, a temporal localization model that learns from point annotations and predicts behavior-specific temporal segments in full-length videos. On our dataset, FSH-TAL reaches state-of-the-art mAP (41.0% averaged across IoU 0.1–0.7, 17.4 points above the strongest baseline). Our project page is available at https://fsh-tal-wacv27.github.io/.

1. Introduction

Behavior is the observable output of the brain, whose development and function are shaped by genetics [29]. Cichlids, a family of fish, are a model system for studying social behaviors due to their distinct behavioral patterns [16] and social systems [22]. They also exhibit high rates of speciation, with more than 1,500 species, providing an ideal system for comparing social behaviors among closely related species. This makes them ideal for understanding how genetics influences behavioral output [8, 11]. Experiments using these fish have demonstrated various connections between brain development, genetics, and evolution, making them highly valuable to the scientific community [3]. Quantifying distinct behaviors is therefore essential for understanding the drivers of variation in behavioral patterns [26]. However, extracting this information from the volume of footage produced by behavioral experiments remains a major bottleneck in the study of animal behavior (ethology) [35].

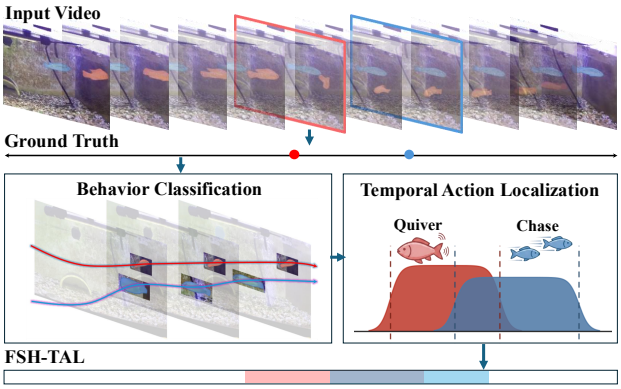


Figure 1. **FSH-TAL motivation.** We propose a method to jointly classify and localize behaviors of cichlid fish from long-form video to accelerate animal behavioral studies.

In many animal behavioral experiments, experts manually annotate hours of recordings using tools such as BORIS [10] to generate behavior counts and determine the effects of a treatment. BORIS allows annotators to record point annotations using hotkeys while watching a video and provides tools for organizing experiments and projects. Although more efficient than recording annotations in a spreadsheet, BORIS still requires annotators to watch each recording in real time. An hour-long video may contain hundreds of behavior occurrences, often clustered together, requiring multiple passes to annotate each event accurately.

Human annotations can also be subjective, particularly when behaviors occur in rapid succession or are difficult to identify from a single camera angle. Thus, multiple researchers must annotate the same video and compare their annotations using inter-rater reliability. Even with shared guidelines, poor inter-rater reliability scores are possible,

*Equal contribution. †Equal contribution.

which mandates recurrent annotation of the same video. Automating this process could provide more consistent annotations, save substantial human effort, and accelerate behavioral experiments for scientific discovery [7].

Current automated approaches often use pose-estimation methods [23, 28] to convert video into trajectories of anatomically defined keypoints, which are then used by downstream behavior classifiers [32, 33]. These trajectories provide an intuitive representation of animal motion but require manual annotation of anatomical landmarks and depend on those landmarks being reliably tracked. However, in animal behavior experiments, weakly-supervised point annotations are more readily available than pose labels.

An alternative is to bypass the explicit pose representation and classify behaviors directly from video. Action-classification models can learn appearance and motion cues without requiring anatomical keypoints, but typically depend on large labeled datasets [6, 13]. Few-shot action-classification methods reduce this requirement by learning from only a small number of labeled examples [18, 19]. However, both settings generally assume trimmed clips, whereas behavioral studies can produce long recordings containing many behaviors with unlabeled boundaries. Temporal action localization (TAL) addresses this problem by identifying each behavior and its temporal extent [14, 21, 36, 37], but most TAL methods require more video, distinct temporal boundaries, and higher quality annotations than behavioral studies can provide [4, 15].

Together, these limitations highlight the need for both benchmarks and methods designed for long behavioral recordings, where training data is limited and only point annotations are available. In this paper, we address this setting through the following contributions:

1. We introduce a benchmark for long-form cichlid behavior localization, comprising 8,202 expert point annotations across 17 recordings of $\sim$1 hour each (Sec. 3).
2. We propose FSH-TAL, a point-supervised TAL model that incorporates a few-shot action-classification objective to jointly localize and classify behaviors with limited training data. It reaches 41.0% average mAP, outperforming prior methods with varied supervision by at least 17.4 points (Sec. 5).
3. We report ablation tests on model components, pinpointing a 13.1 mAP improvement from object-grounded trajectory tokens and a further 8.5 mAP improvement from the few-shot objective.

2. Related Work

2.1. Computational Ethology

Video-based behavior classification is a fundamental component of computational ethology [1, 7]. Current methods for this task utilize pose-estimation systems, such as DeepLabCut [23] and SLEAP [28], that track anatomical points for downstream classifiers such as SimBA [12]. SimBA uses random forest models trained on pose estimation features to accurately recognize behaviors. Such models benefit from large benchmarking datasets such as CalMS21 [32], which proposes a behavior recognition task using labeled mice assays. While pose estimation is feasible in CalMS21 due to the size of the subjects relative to the enclosure and clarity of the footage, producing the same accurate pose estimations for cichlids is inconsistent due to frequent occlusions, low contrast, and the noisy aquarium environment. Our approach replaces anatomical keypoints with generic tracked points derived from SAM3 [5], which requires neither a species-specific skeleton nor pose labels.

2.2. Temporal Action Localization

TAL tasks require a model to predict the start, end, and class of an action, with different types of TAL depending on the strength of labels given in training. ActionFormer [36] is fully supervised and requires start/end labels; ASM-Loc [14] is weakly supervised from video-level labels; SF-Net [21] and HR-Pro [37] are point-supervised. The metrics of these models are reported on the ActivityNet [4] and THUMOS14 [15] datasets, whose action classes are visually distinct (e.g., swimming vs. playing guitar) and span average durations of 4 to 50 seconds. Cichlid action classes differ significantly: behaviors of interest are visually similar to one another and background noise, and many behaviors are near-instantaneous rather than spanning several seconds. As a result, benchmark performance on ActivityNet and THUMOS14 is unlikely to be a reliable predictor of performance on cichlid annotation tasks.

2.3. Few-Shot and Trajectory-Token Recognition

Trajectory-token methods [18, 19] collect tokens along the spatiotemporal path of tracked points, providing a representation that follows the subject through video for action classification. This method provides both motion and visual signal, which is integral for understanding video data [9, 20]. Episodic few-shot training is a popular method for tasks involving scarce data, which allows models to learn from fewer examples by understanding the similarities and differences between examples in each class [2, 25]. Together, these methods are highly suitable for capturing actions and form the foundation of our behavior classifier.

3. Dataset

We gather our data from behavioral experiments on the cichlid species *Astatotilapia burtoni*. We use 17 videos, each approximately 1 hour long, for our training and testing. In each video, a female injected with prostaglandin F2α is introduced to an 80-liter tank containing a male and a terracotta pot acting as a spawning site. The injection of

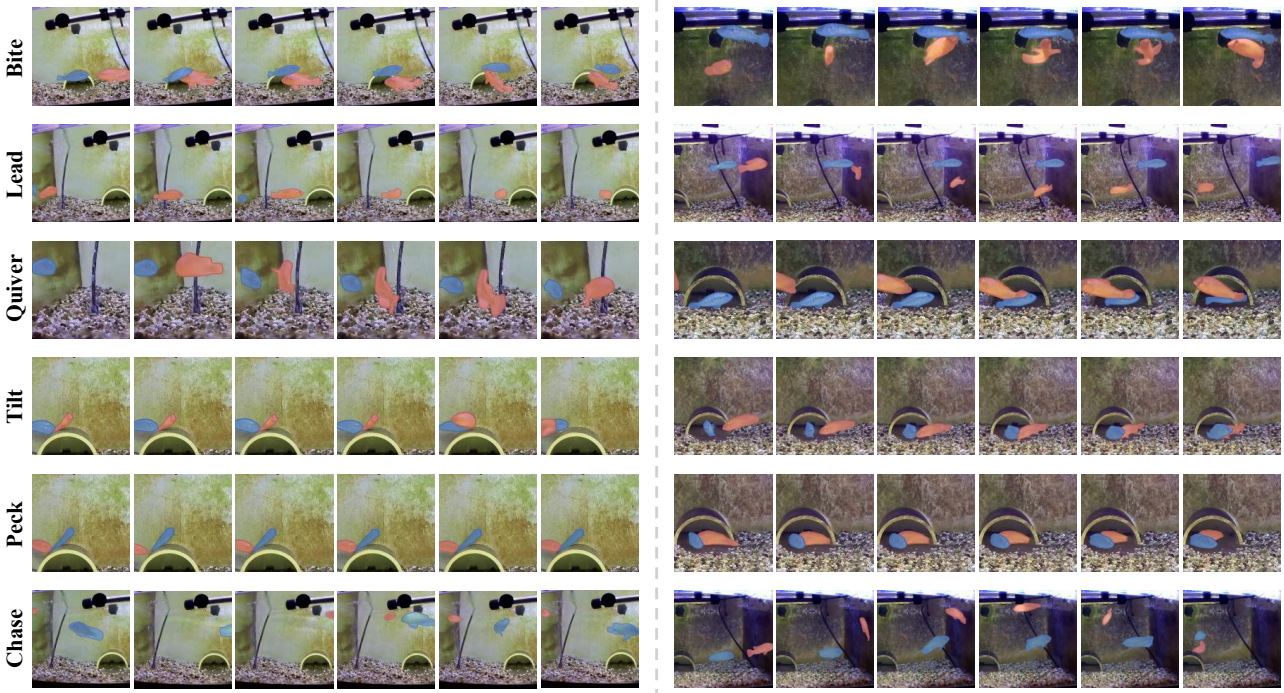


Figure 2. **Behavior examples.** For each of the six behaviors of interest, two 4 s clips from different videos including the behavior are shown. Blue and red masks generated by SAM3 are overlaid on both subjects. A raw input version can be found in the supplemental.

prostaglandin F2α increases sexual receptivity in the female, mimicking her fertile window. To determine if olfaction plays a role in female reproductive behaviors, half of the female subjects had a noseblock placed over their nostrils. All videos were recorded on a Raspberry Pi Camera Module 2 at 25 FPS from a side profile of the tank. Each video consists of exactly two cichlid fish in one tank, but the specific individuals and tank differ between recordings.

Expert observers annotated each video in BORIS [10], creating point annotations for each behavior. Each point annotation consists of a behavior label and a timestamp. Since the biological studies rely primarily on the counts of these behaviors rather than the times they occur, these annotations indicate a behavior occurred in the surrounding time ($\pm$ 2 seconds), and are not necessarily centered or even within the duration of the behavior. Each video in the dataset was independently annotated by one of four observers trained to identify the reproductive behaviors of the *A. burtoni*. Observers followed a standardized annotation guideline to identify and timestamp behaviors using BORIS.

In total, 8,202 point annotations are included in the dataset. Fig. 2 displays examples of the six behaviors of interest: Bite, Lead, Peck, Quiver, Chase, and Tilt. These experiments focus on social interactions and reproductive behaviors. Although each annotation marks the occurrence of a behavior from an individual, most behaviors require participation from the other subject to be recorded. Descriptions of each behavior in the ethogram are as follows:

- **Quiver**: The male swims up to female, displays the side of his body to the female and shakes.
- **Lead**: The male leads the female by swimming away from her slowly with conspicuous fin wiggling, often toward the nest. Often occurs after a quiver.
- **Peck**: Either fish pecks at the other fish's anal fin (bottom of tail) during circling behavior (circling occurs when both fish slowly orbit each other, usually around the nest, tilting and pecking at one another), while the receiving fish is often tilted slightly onto its side.
- **Tilt**: Either fish tilts or leans on its side to display its anal fin for the other fish to peck, usually while rotating during circling, even if the other fish does not peck.
- **Bite**: The male nips at the female outside the context of circling, typically when she is ignoring or fleeing him.
- **Chase**: During a chase, the male chases the female over a distance of at least one body length outside of a mating context. Bites often occur during Chases.

Aside from the quantity of data present, this dataset differs from the most popular TAL datasets ActivityNet and THUMOS14 [4, 15] in two key ways: **1)** The duration of behaviors is much shorter and **2)** Behaviors can co-occur. Together, these two factors lead to many annotations clustering closely together and overlapping, while multiple minutes may have no behaviors occurring at all. These clustered segments can occur from one individual quickly re-

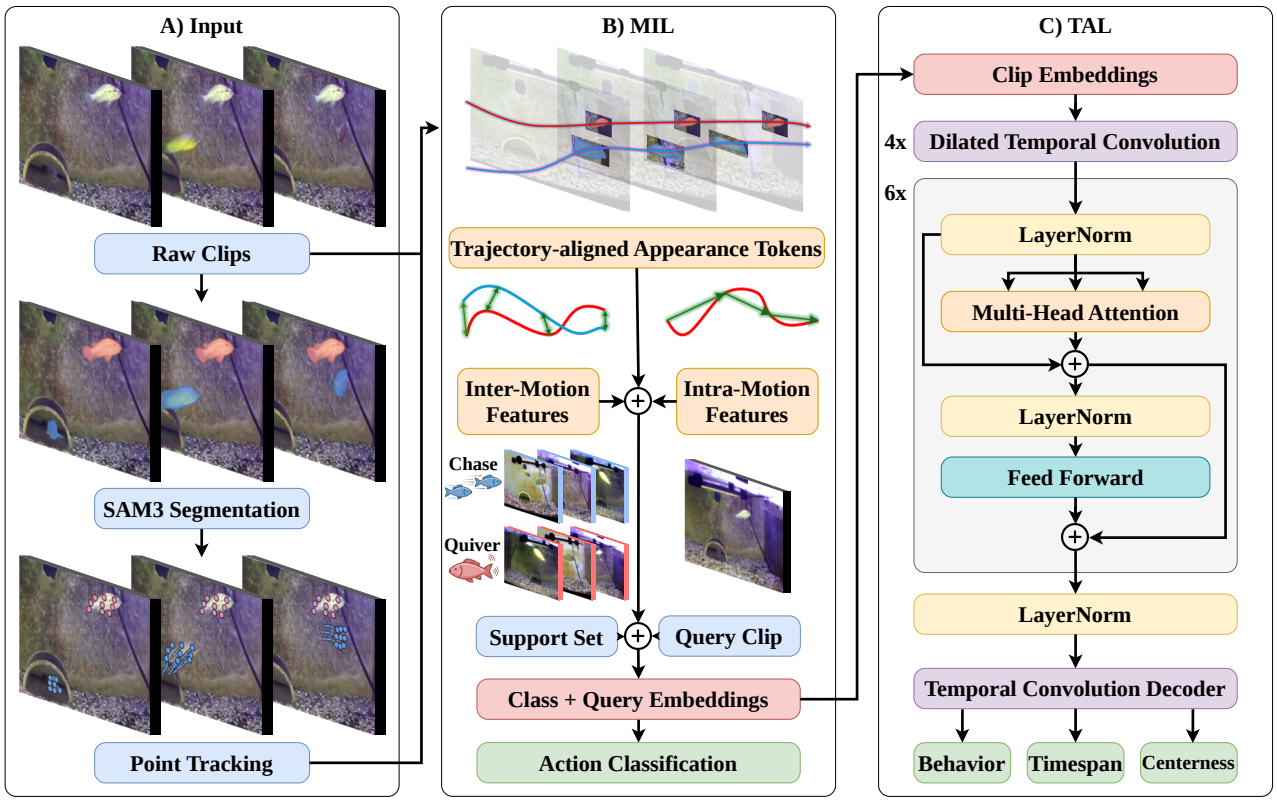


Figure 3. **FSH-TAL architecture. A)** Point trajectories are extracted from segmentation masks produced by SAM3. **B)** Using the videos and point trajectories, we train our classification model in few-shot episodes to predict behaviors. **C)** The localizer is trained on frozen embeddings from the MIL model to predict a starting and ending boundary, a centerness score, and the behavior itself.

peating similar behaviors over and over (*e.g.*, *Male Bite* then *Chase* then *Bite*) or from the male and female performing courtship behaviors at the same time (*e.g.*, *Male Tilt* and *Female Peck*). 67.7% of annotations occur within 2 s of another annotation, as explained in the supplemental. These annotation patterns suggest that multi-labeled data analysis and modeling techniques are necessary.

4. Method

4.1. Overview

Our model (Fig. 3) has three key stages: **A)** we run a text-prompted object segmentation model to generate object masks propagated throughout the target video. Points are extracted from these masks and tracked over time to create point trajectories. **B)** we train a Multiple Instance Learning (MIL) recognition model, which classifies short, 4 s clips from our dataset. It uses few-shot learning to produce multi-labeled predictions. **C)** we use frozen embeddings from the MIL to train a temporal localization model consisting of a dilated convolutional stack, pre-norm transformer blocks, and convolutional decoder heads.

4.2. Problem Formulation

A recording V is an hour of continuous video of a breeding pair. Annotations are *point events* $\{(t_i, c_i)\}$: a timestamp and a behavior class, with no start or end. The task is to emit segments $(s_j, e_j, c_j, \alpha_j)$ where s_j, e_j are the start and end time of the j-th predicted segment, c_j is the predicted class, and α_j is the confidence. Since the extent of a behavior is unknown given our labeled training data, we require a weakly supervised localization model.

4.3. Point Tracking

First, we run Segment Anything Model 3.1 (SAM3) [5, 24] with the text prompt "fish" to identify and mask the fish from the background of the tank. SAM3 is run over sequential 24-second chunks, and the centroids of the top two objects (*i.e.,* male and female fish) are passed between chunks, allowing for consistent labels. Stationary objects over the 24 seconds are discarded to prevent confusing background objects as fish. We then extract 9 points from a 3×3 grid over each fish's mask, yielding 18 point trajectories per clip.

We emphasize that the 3×3 layout is deliberately not anatomical. Keypoint detectors for fish require extensive

per-species human annotation, whereas sampling SAM3 masks does not require labels, yet still describes the subject's motion. Sec. 6 shows that most of the benefit of SAM3 points comes from locating each fish: a single point per fish recovers 74% of the gain over a uniform grid.

4.4. MIL Classification

Point trajectories and video footage are then used as input for the classification backbone. Our classification model builds on Trokens [19], which feeds DINOv2 [27] video features and point trajectories into a video transformer to classify behavior. We improve on it in two ways: (1) trajectory points are sampled from SAM3 masks of the two fish rather than from Trokens' CoTracker tracks [17], improving object grounding in a noisy aquarium. (2) We adapt multi-label classification to handle co-occurring courtship behaviors. We train our MIL with 4 s clips taken from the full length videos, labeled with all behaviors present over the clip's duration. We also include clips where no behavior occurs (60% of the 25,349 clips) to enhance the model.

4.4.1. Trajectory Tokens

Each 4 s clip is subsampled to 16 frames and encoded by a frozen DINOv2 [27] ViT-B/14 at 224×224 pixels. Trokens samples the patch feature map at the 18 points we track. Within Trokens, a Relational Motion Modeling module describes motion within each trajectory as well as motion between trajectories, capturing interactions between tracked objects. Their outputs are added to the trajectory tokens before the temporal transformer that classifies the clip.

4.4.2. Multi-Label Episodes

The default Trokens model uses few-shot methods for single-labeled data, which we adapt for our data. Since 29% of our labeled clips carry two or more behaviors (see supplementary), using multi-labeled examples becomes necessary. We therefore construct closed episodes for few-shot learning: given $N{=}5$ sampled classes, we select the support set for each class from $K{=}3$ examples containing the target class and $Q{=}6$ examples for the query set. Support and query examples need only contain the target class, so multi-labeled examples are available for both. We use the same combination of contrastive and classification losses as Trokens [19], and refer readers there for further details.

4.4.3. Feature Bank

After training, we freeze the MIL backbone and run it once over each video in inference mode. Each 4 s window yields 16 per-frame embeddings (the mean of its 18 point tokens). Windows advance by 2 s and keep only their 8 new frames, giving one 768-d vector per 0.25 s, stored in a bank of shape $(T, 768)$ spanning the recording (14,400 steps per hour). This bank is the only interface between the MIL Classification and Temporal Localization aspects of FSH-TAL.

4.5. Temporal Localization

Completing the behavior annotation task requires localization in addition to classification. The localizer interprets each annotated point as a ±2 s timespan, which serves as a weak boundary label inspired by the annotation guidelines for all recordings. This time span is consistent with empirical observations found across all videos. Starting with the feature bank, we normalize and linearly project to 384 dimensions with a learned matrix, summed with a learned position embedding. Then, for each 90 s window (45 s overlap) within the full-length video, these features are passed through a dilated convolutional stack (dilations 1, 2, 4, 8) that provides further temporal context to the embeddings that feed into 6 pre-norm transformer blocks with 8 attention heads, for 15.8 M trainable parameters in total.

Next, three convolutional heads decode classification (per-class sigmoid over six behaviors), boundary regression (distances to both localized boundaries), and centerness (how centrally a timestep sits inside the event it is assigned to [34]). For each candidate timestep retained by the class head, we predict a distribution over possible left and right boundary offsets, following TriDet [30]: the timespan head emits 2×16 logits per timestep, softmaxed over 16 bins spanning $[0, 64]$ timesteps. Offsets are then predicted by computing the expected value over each distribution, scaled by a learned positive factor.

4.6. Target Assignment

Finally, segments are predicted based on outputs from the convolutional heads. A timestep (every 0.25 s) is positive for every event whose ±2 s window contains it. Timesteps can thus be claimed by several, overlapping events. Classification is multi-hot because the head emits independent per-class sigmoids and evaluation scores per-class AP. Regression and centerness heads are assigned the event whose center is nearest, an adaptation of FCOS's min-area rule [34].

4.7. Implementation Details

Training uses a focal classification loss, an Intersection over Union (IoU) regression loss, and a centerness binary cross-entropy loss. Inference runs on rolling 90 s windows with 45 s overlap over the full recording, scoring each timestep by $\sigma(\text{class}) \cdot \sigma(\text{centerness})$ and decoding with greedy temporal non-maximum suppression per behavior. Training details can be found in the supplemental.

5. Experiments

5.1. Protocol

We evaluate all models with mean Average Precision at IoU (mAP@IoU) averaged over the behaviors present in each recording to assess the accuracy of temporal localization predictions. All results are averaged over the held-out

Table 1. **Results on our dataset.** Average is over IoU 0.1–0.7. All methods are averaged over the 17 held-out recordings. Variance and further results are shown in the supplemental.

	mAP@IoU (%)				
Method	0.1	0.3	0.5	0.7	Avg
MIL classifier	29.9	14.7	5.4	0.2	12.2
MIL + post-proc	44.3	20.7	2.6	0.2	16.0
ASM-Loc [14]	18.2	10.9	4.1	1.2	8.3
HR-Pro [37]	30.6	19.2	6.7	1.2	14.1
ActionFormer [36]	29.2	26.6	22.7	14.4	23.6
FSH-TAL (Ours)	**51.7**	**47.0**	**38.7**	**24.1**	**41.0**

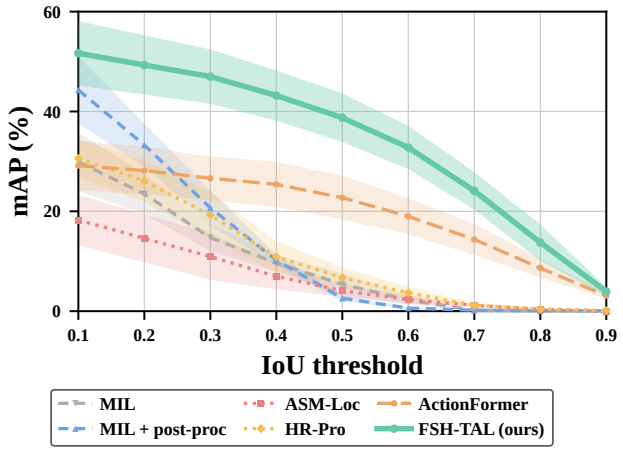


Figure 4. **mAP@IoU results.** Average is over IoU 0.1–0.7. Lines follow Tab. 1, bands show one standard deviation across folds.

recordings from 5 distinct folds. Each fold holds out 3 or 4 entire recordings, and both the MIL backbone and the localizer are retrained for every fold, so no stage is trained on a recording it is evaluated on. For every method, the ground truth is a 4 s window centered on each annotated point. The MIL model is evaluated on 4 s rolling windows with 2 s overlap, merging adjacent windows with the same prediction (*MIL*). Additionally, we post-process its predictions with Outer–Inner Contrast scoring and temporal NMS (*MIL + post-proc*) [31] as a stronger baseline.

5.2. Baselines

We compare our method with previously defined SOTA models for this task. *ASM-Loc* [14] is weakly supervised, *HR-Pro* [37] is point-supervised, and *ActionFormer* [36] is fully supervised on the same ± 2 s pseudo-segments as FSH-TAL. ASM-Loc, ActionFormer, and our localizer all use an identical feature bank derived from the frozen MIL embeddings. HR-Pro operates on its own 1024-d features but the same folds and ground truth, as a point-supervised comparison. Further details are included in the supplemental.

Table 2. **Per-behavior results.** Average is over IoU 0.1–0.7, and over held-out recordings containing that behavior. # Recordings shows how many of the 17 recordings contain the behavior and # Annotations the total annotated events of that behavior. σ_{fold} is the standard deviation of the 5 fold means.

Behavior	**# Recordings**	**# Annotations**	**AP (%)**	σ_{fold}
Bite	15	3516	35.9	7.5
Chase	12	846	35.2	16.9
Lead	16	1455	43.4	11.5
Peck	11	529	48.3	19.9
Quiver	17	1262	42.6	4.2
Tilt	11	594	49.4	16.4

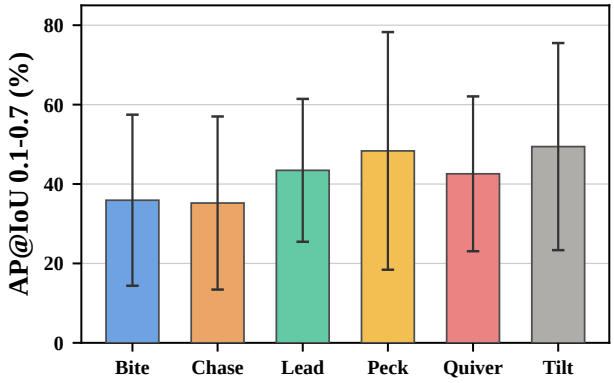


Figure 5. **Per-behavior AP.** Bars show the mean over held-out recordings containing the behavior, whiskers show one standard deviation across those recordings.

5.3. Results

Tab. 1 and Fig. 4 show mAP@IoU for all baselines compared to FSH-TAL, evaluated on our dataset. Both baseline MIL variants degrade at tighter thresholds, which is expected as a classifier cannot refine boundaries beyond its input clip length. FSH-TAL reaches **41.0% mAP** (average IoU 0.1–0.7), outperforming all tested methods, including fully-supervised ActionFormer (23.6%). The margin widens at looser IoU thresholds (+22.5 points at IoU=0.1, but +9.7 at IoU=0.7 relative to ActionFormer), indicating FSH-TAL's gain is driven by recognition quality rather than by boundary precision, further verified in Sec. 6. ActionFormer shares our feature bank, but in its default THUMOS14 configuration our fixed 4 s targets place 98% of its positives on one of six pyramid levels, leaving the other five to train almost only on background (see supplemental). FSH-TAL performs better than all other tested methods at every IoU threshold from 0.1 to 0.9. It also beats every baseline on each of the 17 held-out recordings, by at least 4.8 points. MIL + post-proc and HR-Pro come closest at IoU 0.1 (44.3% and 30.6%) but fall to 1.2% or below by IoU 0.7, where FSH-TAL still reaches 24.1%.

Table 3. **Training cost for one fold.** Peak GPU memory is sampled at 2 s intervals. †Trains on the frozen MIL bank.

Method	#Params (M)	Train (min)	Peak GPU (GB)
MIL classifier	16.0	296	17.2
ASM-Loc [14]	1.8†	9.1	0.3
HR-Pro [37]	23.2	33.0	10.2
ActionFormer [36]	27.3†	15.7	4.1
Ours (localizer)	**15.8**†	**42.7**	**2.6**

Table 4. **Inference cost on an RTX A5000 for one hour of video.**

Stage	Time	GPU (GB)
1. SAM3 tracking	1.0 h	13.5
2. MIL features	3.1 min	7.4
3. Localizer decode	7.6 s	0.08
Full pipeline	1.1 h	13.5
Stages 2–3 only	3.2 min	7.4

5.4. Per-Behavior Classification

Tab. 2 reports AP per behavior averaged across all videos in which the behavior occurs. Since each behavior is averaged over a different subset of recordings, the mean of the six APs (42.5%) differs from the overall 41.0%. *Bite* is the most common behavior by more than a factor of two and among the worst localized, while *Peck* is 6.6× rarer and among the best localized. Well-defined courtship behaviors such as *Tilt* and *Peck* score higher than *Bite* and *Chase*, whose definitions are less exact. Fig. 5 shows the spread of per-behavior AP across held-out videos. The spread is widest for the three behaviors present in 12 or fewer recordings (*Peck*, *Chase*, *Tilt*), while *Lead* varies least. Further results can be found in the supplemental.

5.5. Runtime and Memory

All runtime and memory costs are measured on one 24 GB NVIDIA RTX A5000. FSH-TAL was designed to fit training and inference on one GPU. Training is dominated by the MIL backbone (296 min and 17.2 GB per fold, Tab. 3), which ASM-Loc, ActionFormer, and our localizer all share as a frozen feature bank. On top of it, the localizer converges in 42.7 min on 2.6 GB, less memory than both ActionFormer and HR-Pro. At inference (Tab. 4), streaming SAM3 at 4 FPS tracks both fish in real time (1.0 h per hour of video on 13.5 GB). Downstream steps are cheap: MIL features take 3.1 min (19.5× faster than real time) and the localizer decoding step takes 7.6 s per video-hour (474×). The full pipeline processes an hour of video in 1.1 h, so our entire ∼17-hour dataset can be re-analyzed in about 19 h. Manual annotation required a trained observer to watch every video at roughly real time playback speed and yielded only point annotations, whereas FSH-TAL predicts temporal segments. Since tracks are computed once per recording, re-scoring with a retrained detector, a different ethogram, or a new operating threshold reruns only stages 2–3, which take 3.2 min per video-hour (19× faster than real time).

6. Ablations

Every ablation is a complete 5-fold run over all 17 held-out recordings, evaluated against the same fixed 4 s ground truth window. We also show additional ablations, including object selection and the SAM3 prompt, in the supplemental.

6.1. Classification Stage

6.1.1. Input Features

Tab. 5 varies the frozen embeddings passed to the detector. Replacing the frozen MIL embeddings with raw DINOv2 [27] embeddings at the same 4 FPS, removing the classification model entirely, costs **14.3** mAP and is worse on 16 of 17 recordings. Therefore, FSH-TAL's performance relies heavily on classification-aware embeddings. Retraining the MIL backbone on video features only (a uniform 16×16 patch grid, trained with 5 instead of 6 queries per class to fit in GPU memory) without tracked keypoints costs **13.1** mAP, worse on 17 of 17 recordings. Similarly, retraining the MIL backbone on point trajectory features only without appearance features (only a black video is provided) costs **5.6** mAP, worse on 14 of 17 recordings. Both object-grounded trajectory tokens and appearance features are therefore valuable, though tracking matters more.

Tab. 6 compares how many points are used to generate the point trajectory features: 8 points keeps the four corners of each fish's 3×3 grid, 2 points keeps only the grid centers, while the baseline uses 18 points (9 points for each fish across the 3×3 grid). Reducing to a single point per fish costs **3.4** mAP, which shows that most of the benefit comes from knowing where the subjects are, not from identifying their shape, pose, or relation between the two.

6.1.2. Few-shot Objective

Training the MIL with standard supervised classification instead of the few-shot objective (random batches and the classification loss only, without episodes or the contrastive loss) costs **8.5** mAP, proving to be worse on all 17 recordings. Due to the scarcity of our data, a few-shot objective is beneficial for model results. Details are in the supplemental.

6.2. Detector Components

6.2.1. Output Heads

Tab. 7 removes one head at a time. Without the timespan head, each prediction is a fixed ±2 s window. Without the centerness head, timesteps are scored by class probability alone. Removing the centerness head increased average

Table 5. **Comparison of input features.**

MIL	SAM3	Appearance	mAP@IoU (%) 0.1	0.7	Avg
✗	✗	✓	36.9	13.1	26.7
✓	✗	✓	37.6	13.7	27.9
✓	✓	✗	46.4	19.0	35.4
✓	✓	✓	**51.7**	**24.1**	**41.0**

Table 6. **Comparison of SAM3 point density.**

Points	mAP@IoU (%) 0.1	0.5	0.7	Avg
256 (uniform grid)	37.6	25.4	13.7	27.9
2 (1/fish)	47.8	35.7	21.8	37.6
8 (4/fish)	49.5	37.6	22.4	39.2
18 (9/fish)	**51.7**	**38.7**	**24.1**	**41.0**

Table 7. **Comparison of output heads.**

Centerness	Timespan	mAP@IoU (%) 0.1	0.7	Avg
✓	✗	49.8	22.9	40.0
✗	✓	51.5	23.2	**41.3**
✓	✓	**51.7**	**24.1**	41.0

mAP by **0.3**, but decreased mAP at high IoU. Removing one head costs at most 1.0 mAP, but removing both costs **3.7** mAP (supplemental). Both heads may still benefit animal behavior problems with variable timespans and ethograms.

6.2.2. MLP Expansion

Halving the transformer MLP expansion removes 3.55 M parameters (15.79 M to 12.24 M) for a **1.2** mAP drop, worse on only 9 of 17 recordings. The full MLP width is not clearly needed for a 6-behavior ethogram at this corpus size.

7. Discussion and Limitations

The lack of ground truth intervals provides a substantial challenge for this task. Since the ground-truth is derived from point annotations using an empirically derived ± 2 s window, mAP at high IoU measures agreement with the convention rather than with a true behavior interval. Addressing this requires annotation of true intervals on our dataset. This would both remove the fixed-width paradigm that flattens our detector ablations and provide an evaluation reference independent of the supervision convention.

A complementary limitation raised from the nature of our dataset is the absence of a pose-based baseline. Existing pose estimation methods require pose-labeled training data to fit their downstream classifiers, but no such labels exist for our dataset. Producing a pose-labeled subset of our dataset would require a sustained manual annotation effort that our method is designed to reduce. We note, however, that this gap means we cannot compare our TAL results against pose-estimation based pipelines.

Fine-tuning SAM3 and DINOv2 could improve mask quality, point placement, and tracking through occlusion, since every later stage inherits SAM3's tracking features.

We note that a large advantage of our model is that the majority of our dataset specialization occurs only from the SAM3 prompt. While the point extraction, number of subjects, and timing scheme would also need to be changed for most setups, this pipeline can be adapted to other subjects in new environments. However, broader testing across different tanks, lighting, annotators, and species is required for validation of this claim. Overall, we see this method as a starting point for generalized automation trained off point annotations in behavioral experiments.

8. Conclusion

In this paper, we present an improved method for converting point annotations into temporally localized behavior segments, assessed on long-form cichlid recordings. We contribute a cichlid behavior dataset comprising 17 annotated recordings, with a total of 8,202 point events across the hour-long videos. We further contribute FSH-TAL, a 3-stage model combining object masks from SAM3, an action classification backbone trained on the extracted point information, and a temporal localizer that outputs behavior-labeled segments. Evaluation on our dataset demonstrates higher performance (41.0% mAP across IoU 0.1–0.7) than all tested TAL models on every one of the 17 recordings, including a fully supervised detector trained on identical pseudo-segments. We also perform a controlled attribution of that result: object-grounded trajectory tokens are worth 13.1 mAP and the few-shot objective 8.5, while removing one detector head inherited from mainstream TAL costs at most 1.0 mAP under fixed-width point supervision, though removing both the centerness and timespan heads costs 3.7 mAP. Our work provides a practical approach for converting routine point annotations into temporal behavior measurements, enabling streamlined data collection for broad genetic experiments. The full pipeline processes an hour of video in 1.1 h on a single GPU, close to real time.

9. Acknowledgments

The authors are supported by the University of Maryland's Gemstone Honors Program, the University of Maryland Institute for Advanced Computer Studies (UMIACS), and the Juntti Lab. This work was partially supported by NSF CAREER Award (#2238769) to AS. All biological studies were approved under IACUC Protocol #2133862.

References

[1] David J. Anderson and Pietro Perona. Toward a science of computational ethology. *Neuron*, 84(1):18–31, 2014. 2

[2] Mina Bishay, Georgios Zoumpourlis, and Ioannis Patras. TARN: Temporal attentive relation network for few-shot and zero-shot action recognition. In *BMVC*, 2019. 2

[3] David Brawand, Catherine E Wagner, Yang I Li, Milan Malinsky, Irene Keller, Shaohua Fan, Oleg Simakov, Alvin Y Ng, Zhi Wei Lim, Etienne Bezault, Jason Turner-Maier, Jeremy Johnson, Rosa Alcazar, Hyun Ji Noh, Pamela Russell, Bronwen Aken, Jessica Alföldi, Chris Amemiya, Naoual Azzouzi, Jean-François Baroiller, Frederique Barloy-Hubler, Aaron Berlin, Ryan Bloomquist, Karen L Carleton, Matthew A Conte, Helena D'Cotta, Orly Eshel, Leslie Gaffney, Francis Galibert, Hugo F Gante, Sante Gnerre, Lucie Greuter, Richard Guyon, Natalie S Haddad, Wilfried Haerty, Rayna M Harris, Hans A Hofmann, Thibaut Hourlier, Gideon Hulata, David B Jaffe, Marcia Lara, Alison P Lee, Iain MacCallum, Salome Mwaiko, Masato Nikaido, Hidenori Nishihara, Catherine Ozouf-Costaz, David J Penman, Dariusz Przybylski, Michaelle Rakotomanga, Suzy C P Renn, Filipe J Ribeiro, Micha Ron, Walter Salzburger, Luis Sanchez-Pulido, M Emilia Santos, Steve Searle, Ted Sharpe, Ross Swofford, Frederick J Tan, Louise Williams, Sarah Young, Shuangye Yin, Norihiro Okada, Thomas D Kocher, Eric A Miska, Eric S Lander, Byrappa Venkatesh, Russell D Fernald, Axel Meyer, Chris P Ponting, J Todd Streelman, Kerstin Lindblad-Toh, Ole Seehausen, and Federica Di Palma. The genomic substrate for adaptive radiation in African cichlid fish. *Nature*, 513(7518): 375–381, 2014. 1

[4] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015. 2, 3

[5] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *ICLR*, 2026. 2, 4

[6] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the Kinetics dataset. In *CVPR*, pages 4724–4733, 2017. 2

[7] Sandeep Robert Datta, David J. Anderson, Kristin Branson, Pietro Perona, and Andrew Leifer. Computational neuroethology: A call to action. *Neuron*, 104(1):11–24, 2019. 2

[8] Daniel Escobar-Camacho and Karen L. Carleton. Sensory modalities in cichlid fish behavior. *Curr. Opin. Behav. Sci.*, 6:115–124, 2015. 1

[9] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast networks for video recognition. In *ICCV*, pages 6202–6211, 2019. 2

[10] Olivier Friard and Marco Gamba. BORIS: A free, versatile open-source event-logging software for video/audio coding and live observations. *Methods Ecol. Evol.*, 7(11):1325–1330, 2016. 1, 3

[11] Hugo F. Gante and Walter Salzburger. Evolution: Cichlid models on the runaway to speciation. *Curr. Biol.*, 22(22): R956–R958, 2012. 1

[12] Nastacia L. Goodwin, Jia J. Choong, Sophia Hwang, Kayla Pitts, Liana Bloom, Aasiya Islam, Yizhe Y. Zhang, Eric R. Szelenyi, Xiaoyu Tong, Emily L. Newman, Klaus Miczek, Hayden R. Wright, Ryan J. McLaughlin, Zane C. Norville, Neir Eshel, Mitra Heshmati, Simon R. O. Nilsson, and Sam A. Golden. Simple behavioral analysis (SimBA) as a platform for explainable machine learning in behavioral neuroscience. *Nat. Neurosci.*, 27(7):1411–1424, 2024. 2

[13] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter N. Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "Something Something" video database for learning and evaluating visual common sense. In *ICCV*, pages 5843–5851, 2017. 2

[14] Bo He, Xitong Yang, Le Kang, Zhiyu Cheng, Xin Zhou, and Abhinav Shrivastava. ASM-Loc: Action-aware segment modeling for weakly-supervised temporal action localization. In *CVPR*, pages 13925–13935, 2022. 2, 6, 7

[15] Haroon Idrees, Amir R. Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The THUMOS challenge on action recognition for videos "in the wild". *Comput. Vis. Image Underst.*, 155:1–23, 2017. 2, 3

[16] Lillian R. Jackson, Mariam Dumitrascu, and Beau A. Alward. Sex differences in aggression and its neural substrate in a cichlid fish. *Sci. Rep.*, 15(1):84, 2025. 1

[17] Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker3: Simpler and better point tracking by pseudo-labelling real videos. In *ICCV*, pages 6013–6022, 2025. 5

[18] Pulkit Kumar, Namitha Padmanabhan, Luke Luo, Sai Saketh Rambhatla, and Abhinav Shrivastava. Trajectory-aligned space-time tokens for few-shot action recognition. In *ECCV*, pages 474–493, 2024. 2

[19] Pulkit Kumar, Shuaiyi Huang, Matthew Walmer, Sai Saketh Rambhatla, and Abhinav Shrivastava. Trokens: Semantic-aware relational trajectory tokens for few-shot action recognition. In *ICCV*, pages 13544–13556, 2025. 2, 5

[20] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video Swin transformer. In *CVPR*, pages 3202–3211, 2022. 2

[21] Fan Ma, Linchao Zhu, Yi Yang, Shengxin Zha, Gourab Kundu, Matt Feiszli, and Zheng Shou. SF-Net: Single-frame supervision for temporal action localization. In *ECCV*, pages 420–437, 2020. 2

[22] Karen P. Maruska and Russell D. Fernald. Astatotilapia burtoni: A model system for analyzing the neurobiology of behavior. *ACS Chem. Neurosci.*, 9(8):1951–1962, 2018. 1

[23] Alexander Mathis, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.*, 21(9):1281–1289, 2018. 2

[24] Meta AI. SAM 3.1: Object Multiplex. https://github.com/facebookresearch/sam3, 2026. Release notes, March 2026. 4

[25] Khoi D. Nguyen, Quoc-Huy Tran, Khoi Nguyen, Binh-Son Hua, and Rang Nguyen. Inductive and transductive few-shot video classification via appearance and temporal alignments. In *ECCV*, pages 471–487, 2022. 2

[26] Alexandre A. Nikonov, Julie M. Butler, Karen E. Field, John Caprio, and Karen P. Maruska. Reproductive and metabolic state differences in olfactory responses to amino acids in a mouth brooding African cichlid fish. *J. Exp. Biol.*, 220(16): 2980–2992, 2017. 1

[27] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024. 5, 7

[28] Talmo D. Pereira, Nathaniel Tabris, Arie Matsliah, David M. Turner, Junyu Li, Shruthi Ravindranath, Eleni S. Papadoyannis, Edna Normand, David S. Deutsch, Z. Yan Wang, Grace C. McKenzie-Smith, Catalin C. Mitelut, Marielisa Diez Castro, John D'Uva, Mikhail Kislin, Dan H. Sanes, Sarah D. Kocher, Samuel S.-H. Wang, Annegret L. Falkner, Joshua W. Shaevitz, and Mala Murthy. SLEAP: A deep learning system for multi-animal pose tracking. *Nat. Methods*, 19(4):486–495, 2022. 2

[29] Gene E. Robinson, Russell D. Fernald, and David F. Clayton. Genes and social behavior. *Science*, 322(5903):896–900, 2008. 1

[30] Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. TriDet: Temporal action detection with relative boundary modeling. In *CVPR*, pages 18857–18866, 2023. 5

[31] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. AutoLoc: Weakly-supervised temporal action localization in untrimmed videos. In *ECCV*, pages 154–171, 2018. 6

[32] Jennifer J. Sun, Tomomi Karigo, Dipam Chakraborty, Sharada P. Mohanty, Benjamin Wild, Quan Sun, Chen Chen, David J. Anderson, Pietro Perona, Yisong Yue, and Ann Kennedy. The multi-agent behavior dataset: Mouse dyadic social interactions. In *NeurIPS Datasets and Benchmarks Track*, 2021. 2

[33] Jennifer J. Sun, Ann Kennedy, Eric Zhan, David J. Anderson, Yisong Yue, and Pietro Perona. Task programming: Learning data efficient behavior representations. In *CVPR*, pages 2876–2885, 2021. 2

[34] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *ICCV*, pages 9627–9636, 2019. 5

[35] Ben G. Weinstein. A computer vision for animal ecology. *J. Anim. Ecol.*, 87(3):533–545, 2018. 1

[36] Chen-Lin Zhang, Jianxin Wu, and Yin Li. ActionFormer: Localizing moments of actions with transformers. In *ECCV*, pages 492–510, 2022. 2, 6, 7

[37] Huaxin Zhang, Xiang Wang, Xiaohao Xu, Zhiwu Qing, Changxin Gao, and Nong Sang. HR-Pro: Point-supervised temporal action localization via hierarchical reliability propagation. In *AAAI*, pages 7115–7123, 2024. 2, 6, 7