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

Bhargav Srinivasan[1,2*], William Lamousin[1*], Jason Liu[1†], Charles Phan[1†], Ofure Osunbor[2], Shubh Sharma[1], Cade McGeehan[2], Tahir Haroon[2], Vijay Jayasuriya[1], Coltan Parker[2], Scott Juntti[2], Pulkit Kumar[1], and Abhinav Shrivastava[1]

[1] Department of Computer Science, University of Maryland, College Park, USA
[2] Department of Biology, University of Maryland, College Park, USA
bds062@umd.edu

Abstract. Understanding the brain requires studying its observable output: behavior. Such behavioral experiments in cichlid fish require manual annotations of aquarium video at real time playback speed. These annotations are time consuming to generate and isolated to single time points, which lack temporal context of the underlying behavior. We propose FSH-TAL, a temporal localization model that learns from sparse point annotations and predicts behavior-specific temporal segments on long-form video. On 17 hours of footage, FSH-TAL reaches state-of-the-art mAP (42.8% averaged across IoU 0.1-0.7).

Keywords: Behavior Classification · Temporal Action Localization

1 Introduction

Behavior is the observable output of the brain, whose development and function are shaped by genetics [19]. Cichlids, a family of fish, are a model system for studying social behaviors due to their distinct behavioral patterns [9] and social systems [14]. Furthermore, cichlids exhibit high rates of speciation, with more than 1,500 different species, providing an ideal system for comparing social behaviors amongst closely related species, with the potential to understand how genetics influences behavioral outputs [4, 6]. Quantifying behaviors is therefore essential for understanding drivers of variation in behavioral patterns [16].

In many animal behavioral experiments, experts manually annotate hours of recordings using tools such as BORIS [5], marking each behavior as a single point annotation. Automating the annotations of these recordings would save countless hours of human labor, accelerating ethological experiments [3].

Current approaches use pose estimation techniques [15, 18] to train behavior classifiers [21, 22] to automate annotations. However, pose estimation requires high-quality video and human labels for anatomical segments, while standard filming techniques in aquatic environments introduce constraints such as frequent

*Equal contribution. †Equal contribution.

occlusion, low contrast, and small subjects. More generic action classification models [11, 12] can be paired with temporal action localization (TAL) models [7,13,23,24] to address these constraints and increase automation. These models use video transformers that extract features from raw video signal, sidestepping the aforementioned limitations of pose estimation. Still, most TAL models are tested on common databases [1,8] that contain hundreds of hours of videos and hundreds of action classes. Behavioral studies may have as few as ten behaviors of interest and less than 20 hours of footage, introducing data limitations.

In this paper, we combine the fields of animal behavior and computer vision to create a tool that will accelerate discovery in behavioral experiments. Our contributions include: **(1)** a cichlid dataset spanning 8,202 point annotated behaviors across 17 ∼1 hour long recordings (Sec. 2); and **(2)** FSH-TAL, a model that can jointly localize and classify behaviors from this dataset with 42.8% average mAP, at least 15.5% better than previous methods with varied supervision. Our code and data can be found at `https://github.com/bds062/FSH-TAL`.

2 Dataset

We gathered data from behavioral experiments on the cichlid species *Astatotilapia burtoni*. We use 17 videos, each ∼1 hour long, containing two fish, and recorded at 25 FPS on the Raspberry Pi Camera Module 2 from a side profile of the tank. Pairs of expert observers annotated 8,202 time-stamped point events across a pair-behavior ethogram (Bite, Lead, Peck, Quiver, Chase, Tilt) in BORIS [5], with accuracy verified via inter-annotator agreement. We create 5 folds, training on 13-14 videos and testing on the remaining 3-4.

3 Method

Method Overview. Our model (Figure 1) has three key stages: **A)** First, we track points on the fish over time from a given input clip or video. **B)** Second, we train a Multiple Instance Learning (MIL) model using few-shot learning on the visual and point extraction features. **C)** Third, we use the frozen embeddings from the MIL classifier to train FSH-TAL, a temporal localization model.

Point Tracking. We first run Segment Anything Model 3 (SAM3) [2] with the text prompt "fish" and record the top two objects (i.e., male and female fish) ranked by temporal persistence. We extract 9 points from a 3×3 grid over each fish's predicted mask, yielding 18 points per clip distributed over both fish in the video. Point trajectories are then generated by tracking points across frames.

MIL Classification. Our MIL model builds on Trokens [11], a classifier architecture that utilizes DINOv2 [17] features and tracked points fed into a video transformer. Our method improves on Trokens in two ways: 1) We generate trajectory points with SAM3 instead of the native CoTracker [10] to improve object segmentation in a noisy aquarium and 2) We adapt multi-label few-shot classification to address the limitations of our scarce dataset and courtship behaviors that co-occur (*e.g.*, *Male Tilt* and *Female Peck*). Specifically, we train our

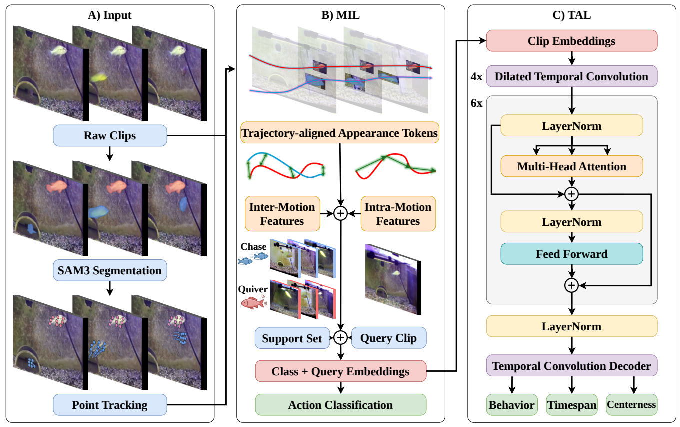

Fig. 1: FSH-TAL architecture. A) Points are generated from raw clips. **B)** We train a classification model using few-shot learning. Attention modules are run over all modules in orange. **C)** We train FSH-TAL using frozen embeddings from the MIL.

models in 5-way 3-shot episodes using the same loss as the original model [11]. Multi-labeled examples contribute to each of their target class' prototypes.

Temporal Localization. Deriving the full behavioral context from video footage requires temporal localization in addition to classification. FSH-TAL achieves both while still being trained on point annotations. Our model reinterprets each annotated point as a ± 2 s timespan, which serves as a weak temporal boundary label for observed behaviors. This time span is consistent with observed durations found across all videos. FSH-TAL then utilizes the frozen embeddings from the MIL, sampled at 4 FPS. Training uses a focal classification loss, Intersection over Union (IoU) regression loss, and centerness binary cross entropy loss on a candidate instance. Inference is done on rolling 90 s windows with 45 s overlap throughout the full recording. Predictions are decoded with a greedy temporal Non-Maximum Suppression (NMS) over each behavior.

4 Results

We evaluated all models on mAP@IoU averaged across all behaviors. The MIL model was evaluated on 4 s rolling window clips with 2 s overlap, after merging adjacent windows with the same behavior prediction (*MIL*). We then post-processed the MIL predictions with Outer–Inner Contrast scoring and temporal NMS (*MIL + post-proc*) [20] as a more thorough baseline. Table 1 shows that both models degraded at tighter thresholds, which is expected as MIL models cannot refine temporal boundaries of predictions beyond their input clip's length.

We compare our methods with previously defined SOTA models for this task. *ASM-Loc* [7] is a weakly supervised TAL model which shows worse performance than our own *MIL + post-proc*. *HR-Pro* [24], which generates temporal

Table 1: Comparison of state-of-the-art methods on our dataset. †Not including parameters used to generate embeddings. σ_{fold} is the spread across folds.

		mAP@IoU (%)					
Method	**#Params**	0.1	0.3	0.5	0.7	Avg	$\sigma_{\textbf{fold}}$
MIL classifier (baseline)	16.0M	25.7	12.0	4.6	0.5	10.5	2.1
MIL + post-proc	16.0M	39.1	20.7	3.1	0.1	14.9	2.3
ASM-Loc [7]	1.8M†	26.8	16.9	6.1	1.2	12.4	2.9
HR-Pro [24]	23.2M	33.6	21.7	7.8	1.2	15.6	4.8
ActionFormer [23]	27.3M†	35.1	32.2	25.4	14.0	27.3	5.1
FSH-TAL (Ours)	15.8M†	**53.3**	**48.7**	**41.3**	**24.7**	**42.8**	7.0

Table 2: Comparison of input features.

		mAP@IoU (%)		
MIL	**SAM3**	0.1	0.7	Avg
✗	✗	41.2	14.7	30.5
✓	✗	44.6	17.0	33.2
✓	✓	**53.3**	**24.7**	**42.8**

Table 3: Comparison of output heads.

		mAP@IoU (%)		
Centerness	**Timespan**	0.1	0.7	Avg
✓	✗	52.2	**24.8**	42.2
✗	✓	**54.3**	24.5	**43.1**
✓	✓	53.3	24.7	42.8

segments from point annotations, and *ActionFormer* [23], a fully supervised method, both show better results at tighter IoU, but fall short of *MIL+post-proc* at IoU=0.1. *ASM-Loc*, *ActionFormer*, and *FSH-TAL* all use classification-aware features which are extracted from the frozen MIL. Table 1 shows that *FSH-TAL* reaches **42.8% mAP** (average IoU 0.1–0.7), outperforming all tested methods.

Tables 2 and 3 show the impact of different architecture decisions within FSH-TAL. First, we compare the results of training FSH-TAL with raw DINOv2 [17] features, MIL features without points, and MIL features with SAM3 [2] points. The increased performance of SAM3 classification-aware embeddings (+9.6%) highlights the effectiveness of point tracking in the noisy environment. Additionally, we test FSH-TAL without the centerness and timespan heads, providing a fixed $\pm 2\,$s window for each prediction. Table 3 shows that removing the centerness head increased average mAP by 0.3%, but decreased mAP at high IoU. These two model heads may not be vital to our problem setting, but may still benefit animal behavior problems with variable timespans and ethograms.

Acknowledgements

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 Abhinav Shrivastava.

References

1. Caba Heilbron, F., Escorcia, V., Ghanem, B., Niebles, J.C.: ActivityNet: A large-scale video benchmark for human activity understanding. In: CVPR. pp. 961–970 (2015). https://doi.org/10.1109/CVPR.2015.7298698
2. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.H., Zhou, Y., Momeni, L., Hazra, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H.K., Dollár, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C.: Sam 3: Segment anything with concepts (2025), `https://arxiv.org/abs/2511.16719`
3. Datta, S.R., Anderson, D.J., Branson, K., Perona, P., Leifer, A.: Computational neuroethology: A call to action. Neuron **104**(1), 11–24 (2019). https://doi.org/10.1016/j.neuron.2019.09.038
4. Escobar-Camacho, D., Carleton, K.L.: Sensory modalities in cichlid fish behavior. Curr. Opin. Behav. Sci. **6**, 115–124 (2015). https://doi.org/10.1016/j.cobeha.2015.11.002
5. Friard, O., Gamba, M.: BORIS: A free, versatile open-source event-logging software for video/audio coding and live observations. Methods Ecol. Evol. **7**(11), 1325–1330 (2016). https://doi.org/10.1111/2041-210X.12584
6. Gante, H.F., Salzburger, W.: Evolution: Cichlid models on the runaway to speciation. Curr. Biol. **22**(22), R956–R958 (2012). https://doi.org/10.1016/j.cub.2012.09.045
7. He, B., Yang, X., Kang, L., Cheng, Z., Zhou, X., Shrivastava, A.: ASM-Loc: Action-aware segment modeling for weakly-supervised temporal action localization. In: CVPR. pp. 13925–13935 (2022). https://doi.org/10.48550/arXiv.2203.15187
8. Idrees, H., Zamir, A.R., Jiang, Y.G., Gorban, A., Laptev, I., Sukthankar, R., Shah, M.: The THUMOS challenge on action recognition for videos in the wild. Comput. Vis. Image Underst. **155**, 1–23 (2017). https://doi.org/10.48550/arXiv.1604.06182
9. Jackson, L.R., Dumitrascu, M., Alward, B.A.: Sex differences in aggression and its neural substrate in a cichlid fish. Sci. Rep. **15**(1), 84 (2025). https://doi.org/10.1038/s41598-024-84188-4
10. Karaev, N., Makarov, I., Wang, J., Neverova, N., Vedaldi, A., Rupprecht, C.: CoTracker3: Simpler and better point tracking by pseudo-labelling real videos. arXiv preprint arXiv:2410.11831 (2024). https://doi.org/10.48550/arXiv.2410.11831
11. Kumar, P., Huang, S., Walmer, M., Rambhatla, S.S., Shrivastava, A.: Trokens: Semantic-aware relational trajectory tokens for few-shot action recognition. In: ICCV (2025). https://doi.org/10.48550/arXiv.2508.03695
12. Kumar, P., Padmanabhan, N., Luo, L., Rambhatla, S.S., Shrivastava, A.: Trajectory-aligned space-time tokens for few-shot action recognition. In: ECCV. pp. 474–493 (2024). https://doi.org/10.48550/arXiv.2407.18249
13. Ma, F., Zhu, L., Yang, Y., Zha, S., Kundu, G., Feiszli, M., Shou, Z.: SF-Net: Single-frame supervision for temporal action localization. In: ECCV. pp. 420–437 (2020). https://doi.org/10.48550/arXiv.2003.06845
14. Maruska, K.P., Fernald, R.D.: Astatotilapia burtoni: A model system for analyzing the neurobiology of behavior. ACS Chem. Neurosci. **9**(8), 1951–1962 (2018). https://doi.org/10.1021/acschemneuro.7b00496
15. Mathis, A., Mamidanna, P., Cury, K.M., Abe, T., Murthy, V.N., Mathis, M.W., Bethge, M.: DeepLabCut: Markerless pose estimation of user-defined

body parts with deep learning. Nat. Neurosci. **21**(9), 1281–1289 (2018). https://doi.org/10.1038/s41593-018-0209-y
16. Nikonov, A.A., Butler, J.M., Field, K.E., Caprio, J., Maruska, K.P.: 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). https://doi.org/10.1242/jeb.157925
17. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. TMLR (2024). https://doi.org/10.48550/arXiv.2304.07193
18. Pereira, T.D., Tabris, N., Matsliah, A., Turner, D.M., Li, J., Ravindranath, S., Papadoyannis, E.S., Normand, E., Deutsch, D.S., Wang, Z.Y., McKenzie-Smith, G.C., Mitelut, C.C., Castro, M.D., D'Uva, J., Kislin, M., Sanes, D.H., Kocher, S.D., Wang, S.S.H., Falkner, A.L., Shaevitz, J.W., Murthy, M.: SLEAP: A deep learning system for multi-animal pose tracking. Nat. Methods **19**(4), 486–495 (2022). https://doi.org/10.1038/s41592-022-01426-1
19. Robinson, G.E., Fernald, R.D., Clayton, D.F.: Genes and social behavior. Science **322**(5903), 896–900 (2008). https://doi.org/10.1126/science.1159277
20. Shou, Z., Gao, H., Zhang, L., Miyazawa, K., Chang, S.F.: AutoLoc: Weakly-supervised temporal action localization. In: ECCV. pp. 154–171 (2018). https://doi.org/10.48550/arXiv.1807.08333
21. Sun, J.J., Karigo, T., Chakraborty, D., Mohanty, S.P., Wild, B., Sun, Q., Chen, C., Anderson, D.J., Perona, P., Yue, Y., Kennedy, A.: The multi-agent behavior dataset: Mouse dyadic social interactions. In: NeurIPS Datasets and Benchmarks Track (2021). https://doi.org/10.48550/arXiv.2104.02710
22. Sun, J.J., Kennedy, A., Zhan, E., Anderson, D.J., Yue, Y., Perona, P.: Task programming: Learning data efficient behavior representations. In: CVPR. pp. 2876–2885 (2021). https://doi.org/10.48550/arXiv.2011.13917
23. Zhang, C.L., Wu, J., Li, Y.: ActionFormer: Localizing moments of actions with transformers. In: ECCV. pp. 492–510 (2022). https://doi.org/10.48550/arXiv.2202.07925
24. Zhang, H., Wang, X., Xu, X., Qing, Z., Gao, C., Sang, N.: HR-Pro: Point-supervised temporal action localization via hierarchical reliability propagation. In: AAAI. vol. 38, pp. 7115–7123 (2024). https://doi.org/10.1609/aaai.v38i7.28539