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

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

In this supplemental, we provide additional details on our dataset, as well as additional results of FSH-TAL. Quantitative details of our dataset are provided in Sec. 1. Sec. 2 provides further implementation details of FSH-TAL. Further testing of our model is provided in Sec. 3 and assorted additional ablations are shown in Sec. 4.

1. Dataset

Mixed-sex groups of adult *A. burtoni* derived from Lake Tanganyika were raised at the University of Maryland. Fish were maintained in 32- to 80-liter community tanks with pH 8.0, 25 °C water temperature, and a 12:12-h light:dark cycle with 15-min dusk and dawn periods. Gravel and terracotta pots provided substrate that facilitates the establishment and maintenance of territories for reproduction. Fish were fed daily with cichlid food pellets (Jehmco) between 0900 h and 1100 h. All animal maintenance and experimental procedures were approved by the Institutional Animal Care and Use Committee of the University of Maryland (protocol 2133862).

Fig. 1 shows how behaviors cluster temporally. Transitions with high probabilities, Chase → Bite, Tilt → Peck, and Quiver → Lead, all agree with the biological explanation of behaviors. Fig. 2 displays the total annotations for each behavior. Fig. 3 includes the number of behaviors that are found within close proximity of another behavior. Fig. 4 shows examples of each behavior of interest without the SAM3 masks. An example of a cluster of annotations is displayed in Fig. 5.

2. Method

2.1. MIL Training Loss

Training the MIL uses two forms of loss: classification and few-shot. Classification loss is binary cross entropy of predicted labels with true labels. With the few-shot paradigm, batches are chosen in the standard N×K method, where a contrastive loss between the query embeddings and support embeddings is calculated. Training the MIL uses the sum of these two losses. For further details please refer to [3].

2.2. SAM3 Tracking

We track with SAM 3.1 [1, 4] in streaming mode. Each recording is sampled at 4 FPS and processed in consecutive 96-frame (24 s) sessions at an input resolution of 1008 px. Object selection then follows the Point Tracking section of the main paper: within each session, short detections and stationary objects are discarded, the two fish slots are carried across sessions by position, free slots are filled by confidence (frames × mean detection probability), and re-

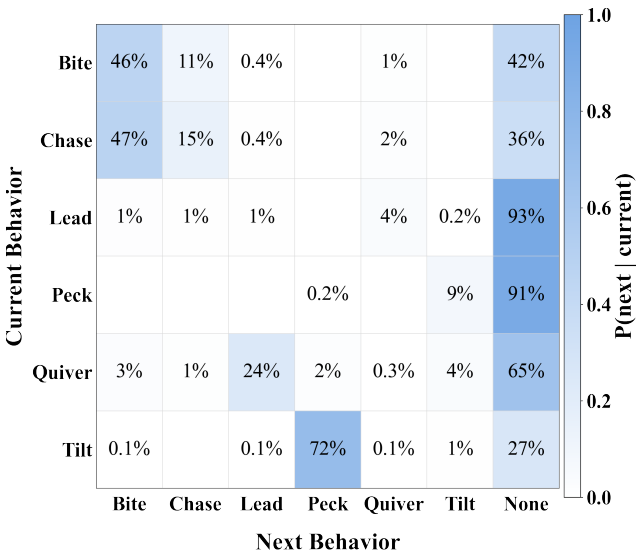


Figure 1. **Transition probabilities.** Probabilities are calculated by counting only the next behavior that occurs in the 2 s after the current behavior. The None column correlates to no behaviors occurring in the two seconds after a behavior.

*Equal contribution. †Equal contribution.

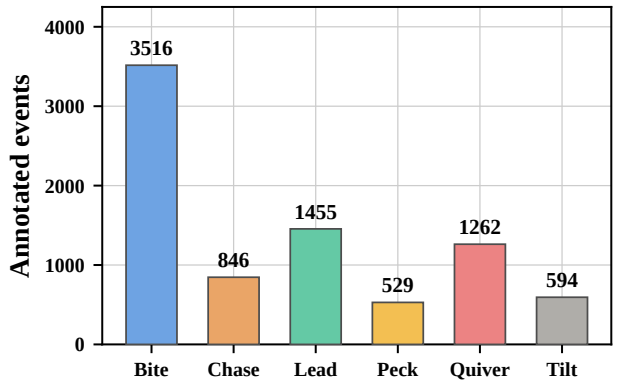


Figure 2. **Behavior counts.** Total annotations over the whole 17 hour dataset.

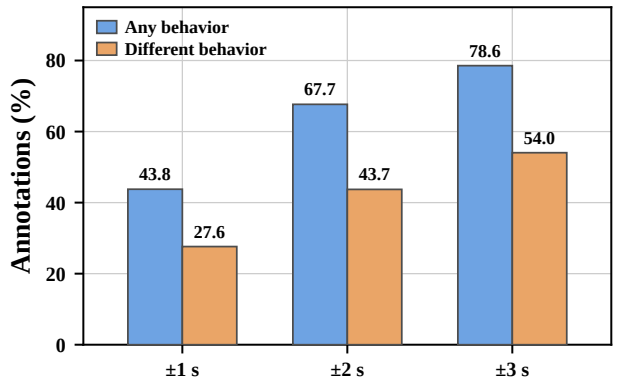


Figure 3. **Annotation co-occurrences.** Counts what percent of annotations have another annotation within 1,2,3 s across any behavior or strictly different behaviors.

detected fragments of the same fish are stitched together. On one RTX A5000, tracking one hour of video takes 1.0 h with 13.5 GB of peak GPU memory. Object selection takes a further 17 s on the CPU.

2.3. Implementation Details

The altered Trokens classification backbone trains for 50 epochs with AdamW at base learning rate 10^{-4}, weight decay 0.5, 5 warmup epochs and step decay at epochs 20 and 40, batch size 32, mixed precision, with the DINOv2 encoder frozen throughout. FSH-TAL trains for 100 epochs with AdamW at 5×10^{-4}, weight decay 0.05, batch size 16, gradient clipping at 1.0, dropout and stochastic depth 0.1, sampling 64 random 90 s windows per recording per epoch. Each localizer trains on the 13 or 14 recordings outside its fold and is evaluated at its final epoch, with no checkpoint selection. Focal loss uses α=0.25, γ=2. Loss weights are 1.0, 1.0 and 0.5 for classification, regression, and centerness, respectively. Decoding keeps proposals above score 0.05 and at most 400 proposals per behavior per recording. All results use seed 0.

3. Experiments

3.1. Results

Tab. 1 shows a more detailed table of mAP@IoU for all baselines compared to FSH-TAL, including standard deviation of values across folds. Tab. 2 shows the same metrics but with means and standard deviations across all 17 videos for all methods. Every method, including the baselines, is scored by the same per-recording evaluator as FSH-TAL, which is what Table 1 of the main paper reports. ASM-Loc is trained and evaluated on 60 s chunks and drops chunks without annotated behavior, so we re-run its selected checkpoint on every chunk of each held-out recording and shift the detections back onto the recording timeline. Their own evaluators, which pool each fold's recordings, give 10.7%, 15.6%, and 31.0%.

ActionFormer's 17.4-point deficit likely results from its architecture and configuration, which are poorly suited for our dataset. It trains on the same feature bank as FSH-TAL, so the gap is not a feature effect. We run its default THUMOS14 configuration, changing only the dataset fields, the number of epochs (60), and the test-time buffer length for hour-long recordings. Its feature pyramid, which is meant to capture actions at varying temporal scales, has six levels (strides 1–32), each regressing a fixed range of distances. Every pseudo-segment spans exactly 16 feature steps (4 s), so on our training data 97.7% of positive assignments fall on the stride-4 level, 2.3% on the two finest levels, and none on the three coarsest. Five of its six levels are therefore trained almost only to emit background. Furthermore, its default configuration is tuned for THUMOS/ActivityNet datasets.

Additionally, the reported ASM-Loc configuration omits its multi-step refinement stage. When training the multi-step refinement stage (6.5 M parameters), the model scored 7.8% against the base model's 8.3%. We therefore report the stronger base configuration.

3.2. Variance

FSH-TAL scores 41.0% averaged over all held-out recordings, with a standard deviation of 12.4 points. Fig. 6 shows per-behavior results colored by video. This large standard deviation across recordings is represented within folds as well ($\sigma_{\text{fold}} = 4.92$). Fig. 7 shows that each fold has a large variance tightly correlated to the per-recording variance. Drawing groups of 3/4 recordings from a corpus with $\sigma = 12.4$ gives an expected fold-mean spread of $12.4/\sqrt{17/5} = 6.72$, against the 4.92 observed. Fold spread is therefore no larger than the recording spread would predict. For this reason, ablations are compared pairwise per recording rather than only by their averages over the 17 videos.

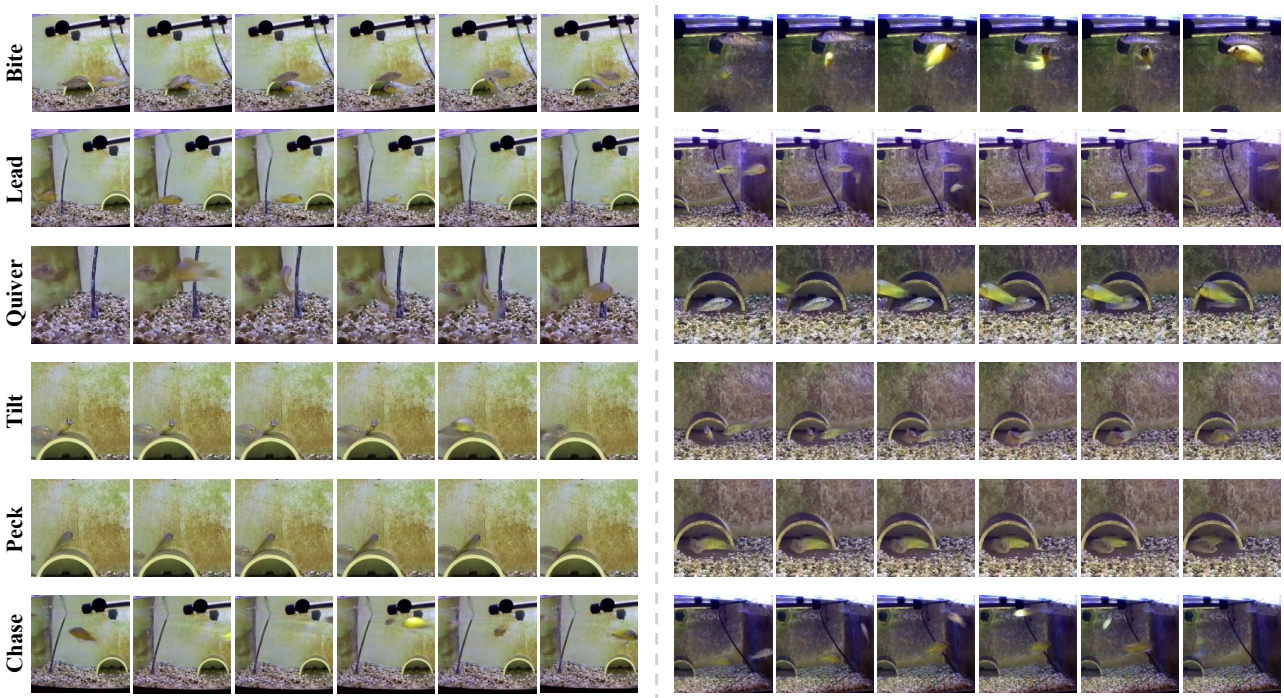


Figure 4. **Behavior examples.** For each of the six behaviors of interest, two four second clips from different videos are shown.

Table 1. **Full mAP@IoU comparison with fold-to-fold spread.** Every entry is the mean over the 5 recording-level cross-validation folds of mAP (%) at that IoU, followed by ± one standard deviation over the folds. Because folds hold 3 or 4 recordings, the FSH-TAL fold mean (41.3%) differs slightly from its per-recording mean (41.0%, Tab. 2).

	mAP@IoU (%)										
Method	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	Avg.$_{.1-.7}$	Avg.$_{.1-.9}$
MIL classifier	30.6 ±5.9	23.9 ±4.0	15.1 ±2.7	9.8 ±1.7	5.7 ±2.2	2.3 ±0.9	0.2 ±0.1	0.1 ±0.1	0.0 ±0.0	12.5 ±2.4	9.7 ±1.8
MIL + post-proc	44.9 ±6.7	33.7 ±4.3	21.0 ±3.5	10.3 ±1.1	2.6 ±0.6	0.6 ±0.2	0.2 ±0.1	0.1 ±0.1	0.1 ±0.1	16.2 ±2.0	12.6 ±1.6
ASM-Loc [2]	18.5 ±5.1	14.8 ±4.8	11.1 ±4.7	7.1 ±2.6	4.2 ±1.2	2.3 ±0.8	1.1 ±0.3	0.4 ±0.1	0.0 ±0.0	8.4 ±2.7	6.6 ±2.1
HR-Pro [6]	30.8 ±3.8	26.2 ±4.2	19.4 ±4.7	11.0 ±3.1	6.8 ±2.0	3.7 ±1.2	1.2 ±0.5	0.5 ±0.3	0.1 ±0.1	14.1 ±2.7	11.1 ±2.1
ActionFormer [5]	29.6 ±5.0	28.6 ±4.8	27.0 ±4.4	25.8 ±4.5	23.1 ±4.4	19.3 ±3.6	14.6 ±3.1	8.7 ±1.9	3.1 ±0.6	24.0 ±4.2	20.0 ±3.4
FSH-TAL (Ours)	**52.1** ±6.5	**49.7** ±5.9	**47.4** ±5.5	**43.6** ±5.0	**39.1** ±4.9	**33.1** ±4.3	**24.4** ±3.8	**14.0** ±3.6	**4.0** ±0.9	**41.3** ±4.9	**34.1** ±4.1

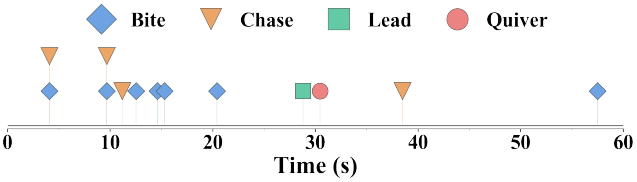


Figure 5. **Annotations example.** Timeline showing an example of clustered behaviors from 60 second snippet of dataset. Stacked behaviors have the exact same annotation time.

4. Ablations

These ablations tune a single hyperparameter or explore how the design decisions in the main text combine. Protocol is identical to the main paper: every localizer, including each ablation, trains on the 13 or 14 recordings outside its fold and is evaluated at its final epoch.

4.1. Target Assignment

The boundary regression and centerness must pick a single event when a timestep is claimed by multiple events. Tab. 3 compares assigning the nearest event center against resolving by detection order. The nearest-center rule is worth **1.8** mAP. The rule only matters where events overlap, which is a significant aspect of our dataset (see main text).

Table 2. **Full mAP@IoU comparison with video-to-video spread.** Every entry is the mean over the 17 held-out videos of mAP (%) at that IoU, followed by ± one standard deviation over the videos. All rows use the same per-recording evaluator.

	mAP@IoU (%)										
Method	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	Avg.$_{.1-.7}$	Avg.$_{.1-.9}$
MIL classifier	29.9 ±8.9	23.5 ±7.3	14.7 ±6.1	9.6 ±4.1	5.4 ±3.7	2.2 ±1.9	0.2 ±0.1	0.1 ±0.1	0.0 ±0.0	12.2 ±4.1	9.5 ±3.2
MIL + post-proc	44.3 ±9.9	33.3 ±8.1	20.7 ±5.9	10.2 ±3.9	2.6 ±1.3	0.6 ±0.3	0.2 ±0.2	0.1 ±0.3	0.1 ±0.2	16.0 ±3.8	12.4 ±3.0
ASM-Loc [2]	18.2 ±8.3	14.6 ±6.9	10.9 ±5.8	7.0 ±3.8	4.1 ±2.4	2.3 ±1.6	1.2 ±1.0	0.3 ±0.3	0.0 ±0.0	8.3 ±4.0	6.5 ±3.1
HR-Pro [6]	30.6 ±8.7	26.1 ±8.3	19.2 ±7.9	10.9 ±5.2	6.7 ±3.9	3.7 ±2.3	1.2 ±0.8	0.4 ±0.5	0.1 ±0.1	14.1 ±5.1	11.0 ±4.0
ActionFormer [5]	29.2 ±11.7	28.2 ±11.5	26.6 ±11.0	25.4 ±10.7	22.7 ±10.5	19.0 ±9.5	14.4 ±7.9	8.6 ±5.2	3.1 ±2.3	23.6 ±10.3	19.7 ±8.7
FSH-TAL (Ours)	**51.7** ±14.1	**49.3** ±13.9	**47.0** ±13.5	**43.2** ±13.2	**38.7** ±12.6	**32.8** ±11.3	**24.1** ±9.8	**13.7** ±7.3	**3.9** ±1.9	**41.0** ±12.4	**33.8** ±10.5

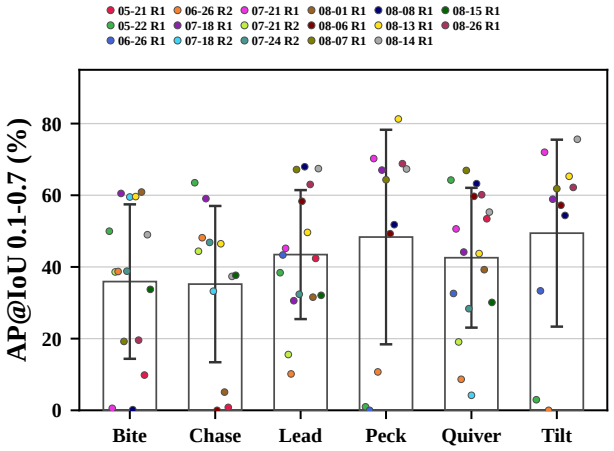


Figure 6. **Per-behavior AP colored by individual video.** Bars show the mean over held-out recordings containing the behavior, whiskers show one standard deviation across those recordings, and colored dots show individual video results.

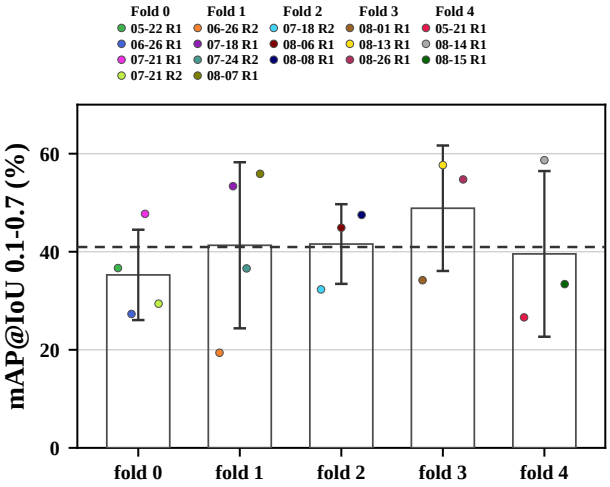


Figure 7. **mAP by fold.** Bars are fold means, whiskers are one standard deviation across that fold's held-out recordings, circles are the individual recordings, and the dashed line shows the overall mean of 41.0%. The legend shows assignment of videos to folds.

4.2. Few-Shot Objective

Tab. 4 displays the effects of the few-shot training paradigm on the frozen MIL used in FSH-TAL results. Without few-shot, examples for each batch are chosen randomly rather than in the N×K method.

4.3. Pseudo-segment Width

The ±2 s window that turns a point into a pseudo-segment is set from extensive empirical observations. Tab. 5 sweeps the supervision on this window: the evaluation reference stays at 4 s while the training target spans 2 to 6 s, capping attainable IoU at $\min(w, 4)/\max(w, 4)$. Within those limits, 2 s is clearly too narrow while 6 s is the best experiment across all ablations at IoU 0.1. However, the 4 s supervision performs the best as an average on our dataset.

Table 3. **Comparison of target assignment.**

	mAP@IoU (%)			
Rule	0.1	0.5	0.7	Avg
iteration order	50.7	37.0	21.2	39.2
nearest center	**51.7**	**38.7**	**24.1**	**41.0**

4.4. Attention Context

Tab. 6 sweeps the attention window: every context from 30 s to 180 s is within 1.9 mAP of 90 s. The 30 s context scores 1.1 mAP higher and is better on 12 of 17 recordings, which is not significant (sign test $p = 0.14$).

4.5. Components

Tab. 7 sweeps the MLP width and provides parameter counts. Tab. 8 compares joint ablations of the centerness head, timespan head, and MLP contraction components.

Table 4. **Comparison of few-shot MIL objective.**

	mAP@IoU (%)			
Few-shot Learning	0.1	0.5	0.7	Avg
✗	43.1	30.0	17.1	32.5
✓	**51.7**	**38.7**	**24.1**	**41.0**

Table 5. **Comparison of supervised width.**

	mAP@IoU (%)			
Span	0.1	0.5	0.7	Avg
±1 s	45.7	31.8	0.0	28.5
±3 s	**54.2**	**39.9**	0.0	38.7
±2 s (ours)	51.7	38.7	**24.1**	**41.0**

Table 6. **Comparison of temporal attention context.**

	mAP@IoU (%)			
Context	0.1	0.5	0.7	Avg
30 s	**52.0**	**40.6**	**25.5**	**42.1**
60 s	51.6	39.3	23.6	41.3
180 s	49.4	37.1	22.5	39.1
90 s (ours)	51.7	38.7	24.1	41.0

Table 7. **Comparison of MLP width.**

		mAP@IoU (%)		
MLP	**#Par.**	0.1	0.7	Avg
2×	12.2M	49.5	**24.2**	39.8
4× (ours)	15.8M	**51.7**	24.1	**41.0**

Table 8. **Beneficial component interactions.**

Centerness Head	**Timespan Head**	**MLP Contraction**	**Avg**
✗	✓	✗	**41.3**
✓	✓	✗	41.0
✗	✓	✓	40.8
✓	✗	✓	40.7
✓	✗	✗	40.0
✓	✓	✓	39.8
✗	✗	✓	38.2
✗	✗	✗	37.3

Exclusion of the centerness head yields the highest average mAP. Despite this, we choose to include all three components in FSH-TAL due to the best variant scoring only **0.3** mAP above the full model and the potential utility of these components on new datasets with variable length behaviors and different ethograms as explained in the main text.

Table 9. **Comparison of object selection methods.**

	mAP@IoU (%)			
Selection	0.1	0.5	0.7	Avg
per-clip persistence	49.1	37.5	22.2	39.0
session-level (ours)	**51.7**	**38.7**	**24.1**	**41.0**

Table 10. **Comparison of SAM3 text prompts.**

	mAP@IoU (%)			
Prompt	0.1	0.5	0.7	Avg
"cichlid"	48.8	36.9	22.0	38.5
"silver or yellow fish"	49.3	37.9	22.7	39.4
"fish" (ours)	**51.7**	**38.7**	**24.1**	**41.0**

4.6. Object Selection

Tab. 9 compares our session-level object selection (Sec. 2.2) with a per-clip rule that keeps, for each 4 s clip, the two objects present in the most of its 16 frames. Both rules select from identical SAM 3.1 output, and the MIL backbone and FSH-TAL are retrained for each. Rejecting stationary objects and carrying fish identities across sessions is worth 2.0 mAP, and is better on 11 of 17 recordings.

4.7. SAM 3.1 Text Prompt

Tab. 10 varies the SAM 3.1 text prompt, again retraining the MIL backbone and FSH-TAL for each prompt. The generic prompt "fish" performs best. A description of our subjects' coloring ("silver or yellow fish") is within 1.6 mAP, while the more specific "cichlid" costs 2.5 mAP and is worse on 13 of 17 recordings.

4.8. No-SAM3 Backbone

The no-SAM3 backbone in Tables 5 and 6 of the main paper replaces the 18 tracked points with a static 16×16 grid of 256 points, and is trained with 5 instead of 6 queries per class to fit in GPU memory.

References

[1] 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. 1

[2] 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. 3, 4

[3] 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. 1

[4] Meta AI. SAM 3.1: Object Multiplex. https://github.com/facebookresearch/sam3, 2026. Release notes, March 2026. 1

[5] Chen-Lin Zhang, Jianxin Wu, and Yin Li. ActionFormer: Localizing moments of actions with transformers. In *ECCV*, pages 492–510, 2022. 3, 4

[6] 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. 3, 4