Trokens: Semantic-Aware Relational Trajectory Tokens for Few-Shot Action Recognition

Supplementary Material

In this supplementary, we provide additional quantitative and qualitative results of Trokens. We provide more quantitative analysis in Sec 1. First, we analyze semantic-aware point sampling (Sec 1.1), including the impact of adding SAM mask as refinement [1, 2] and the number of clusters. We then study relational motion modeling (Sec 1.2), comparing early and late fusion strategies and the effect of HoD bin count. Additionally, we ablate the contributions of appearance and motion features (Sec 1.3) to understand how motion and appearance features complement each other.

1. More Quantitative Results

1.1. Analysis on Semantic-aware Point Sampling

Impact of SAM refinement. While Trokens utilizes DINO's self-supervised features for semantic clustering, we explore enhancing our segmentation using the Segment Anything Model (SAM) [1, 2] as a refinement step. Table 2 shows that while performance remains comparable on SSV2-Small, SAM's refined masks yield a 2% improvement on the larger, more diverse SSV2-Full dataset. However, this modest performance gain does not justify the significant computational overhead introduced by SAM, leading us to maintain DINO-based segmentation in our final design.

Impact of number of clusters. Table 1 presents results for varying numbers of semantic clusters used in our point sampling approach. We selected the 12-cluster configuration for our final implementation, as our experiments indicate that performance remains relatively stable when either increasing or decreasing the number of clusters from this value.

1.2. Analysis on Relational Motion Modeling

Early vs. late fusion of motion module. In Table 4, we evaluate the impact of motion information integration timing with trajectory-aligned tokens. We compare our approach of early fusion (adding motion information before the transformer) against late fusion (adding it after the transformer). The results demonstrate that early fusion significantly improves performance, which aligns with our expectations since this approach allows the transformer blocks to process and leverage the motion information throughout their computations directly. This finding validates our architectural design choice and highlights the importance of proper information flow sequencing in multimodal fusion systems.

Table 1. Impact of number of clusters on few-shot accuracy.

	SSV2 Small		SSv2 Full	
Num clusters	1-shot	5-shot	1-shot	5-shot
6	52.4	68.2	61.2	76.9
12	53.4	68.9	61.5	76.7
24	52.9	68.4	62.2	76.8

Varying number of bins used in HoD. Table 3 shows the results of using varying number of bins in HoD. The ablation study investigates the impact of varying the number of bins (8, 16, 32, and 64) in the Histogram of Displacement (HoD) descriptor used in the Intra-motion module. On SSV2 Small dataset, 8 and 32 bins performed equally best for 1-shot learning (53.4%), while on SSV2 Full, a more complex dataset, 16 bins yielded optimal results for 5-shot learning (77.1%). For SSV2 Full 1-shot, 16 bins provided the highest accuracy (62.6%). These results suggest that our method is relatively robust to bin count. A moderate number of bins (16) generally provides better performance on the larger dataset, balancing between too coarse (8) and too fine-grained (64) orientation discretization.

1.3. Analysis on Individual Component

Impact of appearance and motion features. Table 5 examines the contribution of appearance tokens and motion features to few-shot action recognition across three datasets. Using only motion features yields the lowest performance, while appearance-only features perform significantly better, especially on Kinetics (91.4% for 5-shot). Combining both modalities provides the best results on SSV2 Small (53.4%/68.9% for 1/5-shot) and SSV2 Full (61.5%/76.7% for 1/5-shot), with improvements of 3.5% and 2.9% over appearance-only for 1-shot tasks, respectively. Interest-

Table 2. Impact of SAM refinement of segmentation masks on few shot performance.

	SSV2 Small		SSV2 Full	
Segmatation type	1-shot	5-shot	1-shot	5-shot
DINO based (ours)	**53.4**	68.9	61.5	76.7
SAM refinement	53.1	**69.1**	**63.2**	**78.5**

Table 3. Number of bins used in Intra-motion module for HoD descriptor computation.

	SSV2 Small		SSv2 Full	
Num bins	1-shot	5-shot	1-shot	5-shot
8	**53.4**	68.9	61.2	76.3
16	53.1	69.2	62.6	**77.1**
32	**53.4**	68.9	62.5	76.7
64	52.5	69.1	62.5	77.0

Table 4. Analysis between late vs. early fusion for relational motion module.

	SSV2 Small		SSv2 Full	
Fusion type	1-shot	5-shot	1-shot	5-shot
Late	49.0	60.7	57.6	71.2
Early (ours)	**53.4**	**68.9**	**61.5**	**76.7**

Table 5. Impact of appearance and motion features on few-shot action recognition performance across datasets.

		SSV2 Small			SSv2 Full			Kinetics		
Appearance	Motion	1-shot	3-shot	5-shot	1-shot	3-shot	5-shot	1-shot	3-shot	5-shot
✗	✓	31.0	41.6	45.4	47.0	57.6	59.8	26.3	30.4	33.4
✓	✗	49.9	62.1	65.7	58.6	70.8	74.2	82.5	89.8	91.4
✓	✓	53.4	65.3	68.9	61.5	73.8	76.7	82.9	89.9	91.4

ingly, on Kinetics, the performance gain from adding motion features is marginal (0.4% for 1-shot), suggesting that Kinetics actions rely more heavily on appearance cues than on complex motion patterns, unlike the motion-centric SSV2 datasets.

References

[1] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 1

[2] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1