Supplementary Material: Joint Self-Supervised Video Alignment and Action Segmentation

Ali Shah Ali† Syed Ahmed Mahmood† Mubin Saeed Andrey Konin
M. Zeeshan Zia Quoc-Huy Tran

Retrocausal, Inc., Redmond, WA
www.retrocausal.ai

In this supplementary material, we first provide the implementation details of our VAOT and VASOT approaches in Sec. S1. Next, our ablation analysis results on the Pouring dataset are included in Sec. S2, while our sensitivity analysis results of the entropy regularization weight ϵ are presented in Sec. S3. Furthermore, our multi-action video alignment results and our per-video action segmentation results are included in Secs. S4 and S5 respectively. Finally, we present some qualitative results in Sec. S6 and complexity comparison results in Sec. S7.

S1. Implementation Details

For fair comparisons with previous self-supervised video alignment methods [1, 3, 4, 7], our VAOT and VASOT approaches for video alignment utilize a ResNet-50 encoder network. Please refer to Dwibedi et al. [3] for additional details on the encode network. We provide the hyperparameter settings of our VAOT and VASOT approaches for video alignment in Tab. S1. Note that our VASOT approach for video alignment includes both VAOT and ASOT [11] modules, which share the same hyperparameter settings shown in the VASOT column in Tab. S1.

In addition, following state-of-the-art self-supervised action segmentation methods [5, 6, 11], our VASOT approach for action segmentation employs a 2-layer MLP encoder network for fair comparison purposes. Please see Kukleva et al. [5] for more details on the encoder network. The hyperparameter settings of our VASOT approach for action segmentation are included in Tab. S2. Note that both VAOT and ASOT [11] components in our VASOT approach for action segmentation have the same hyperparameter settings presented in the VASOT column in Tab. S2.

S2. Ablation Analysis Results

In addition to the results on IKEA ASM in Sec. 4.1 of the main paper, we present the ablation analysis results of our VAOT approach on the Pouring dataset in Tab. S3.

Effect of Structural Prior. The structural prior imposes temporal consistency on the transport map. Removing it significantly degrades performance across all metrics, highlighting its critical role in our VAOT approach.

Effect of Temporal Prior. Excluding the temporal prior results in performance drops across all metrics, confirming its positive impact on video alignment.

Effect of Balanced Assignment. Using a full unbalanced assignment formulation causes a substantial decrease in performance. This demonstrates the importance of balanced assignment in our VAOT approach.

Effect of Virtual Frame. Virtual frames are added for tackling background/redundant frames and improving robustness. Removing virtual frames results in minor performance drops across all metrics, which is expected given the monotonic nature of the Pouring dataset.

S3. Sensitivity Analysis Results

Sec. 4.2 of the main paper performs sensitivity analyses on different hyperparameters such as r, α, ρ, ζ, w_{seg}, and w_{align}. We now plot the results of the entropy regularization weight ϵ in Fig. S1. We use our VAOT approach and the Pouring dataset for this experiment. From the results, Acc@1.0 remains stable across the studied value range of ϵ. Progress exhibits small fluctuations, whereas τ is the most sensitive metric, steadily increasing from $\epsilon = 0.05$, peaking at $\epsilon = 0.07$, and declining thereafter. Furthermore, we find that $\epsilon = 0.07$ also yields the best results for the remaining datasets.

† indicates joint first author.
{alishah,ahmed,mubin,andrey,zeeshan,huy}@retrocausal.ai.

Hyperparameter	VAOT	VASOT
Number of sampled frames	40 (P), 20 (PA, IA)	40 (P), 20 (PA, IA)
Learning rate	10^{-5}	10^{-4} (P), 10^{-5} (PA, IA)
Weight decay	10^{-5}	10^{-5}
Batch size	2 videos	2 videos
Entropy regularization weight ϵ	0.07	0.07
Virtual frame threshold ζ	0.5	0.5
Gromov-Wasserstein weight α	0.3	0.3
Structural prior radius r	0.02	0.02
Temporal prior weight ρ	0.35	0.35

Table S1. Hyperparameter settings of our VAOT and VASOT approaches for state-of-the-art video alignment comparison. Note that P, PA, and IA represent Pouring, Penn Action, and IKEA ASM respectively.

Hyperparameter	VASOT
Number of sampled frames	256
Learning rate	10^{-3} (B, M, E, DA), 10^{-4} (YTI)
Weight decay	10^{-4} (B, M, E, DA), 10^{-5} (YTI)
Batch size	2 videos
Entropy regularization weight ϵ	0.07
Virtual frame threshold ζ	0.5
Gromov-Wasserstein weight α	0.3 (YTI, M, E, DA), 0.5 (B)
Structural prior radius r	0.02 (DA), 0.04 (B, YTI, M, E)
Temporal prior weight ρ	0.15 (M, E), 0.2 (B, YTI, DA)
Number of epochs	30 (YTI, E), 50 (B), 100 (M, DA)

Table S2. Hyperparameter settings of our VASOT approach for state-of-the-art action segmentation comparison. Note that B, YTI, M, E, and DA denotes Breakfast, YouTube Instructions, 50 Salads (Mid), 50 Salads (Eval), and Desktop Assembly respectively.

	Method	Acc@0.1	Acc@0.5	Acc@1.0	Progress	τ	AP@5	AP@10	AP@15
Pouring	w/o Structural Prior	62.13	88.28	<u>93.68</u>	<u>90.28</u>	72.49	84.45	84.45	84.41
	w/o Temporal Prior	71.57	<u>90.51</u>	91.47	86.94	78.11	86.17	86.29	86.24
	w/o Balanced Assignment	63.29	88.57	92.38	85.73	68.78	82.83	82.71	82.49
	w/o Virtual Frame	<u>86.32</u>	87.22	93.24	88.11	<u>82.65</u>	<u>86.99</u>	<u>86.74</u>	<u>86.56</u>
	All	**91.80**	**92.88**	**94.63**	**91.63**	**88.28**	**91.34**	**90.56**	**90.29**

Table S3. Ablation analysis results. **Bold** and <u>underline</u> denote the best and second best respectively.

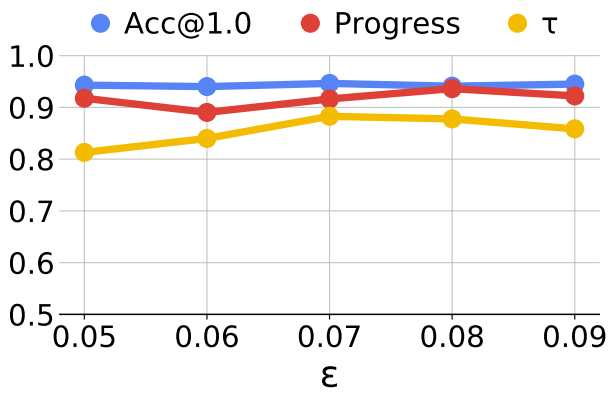


Figure S1. Sensitivity analysis results.

S4. Multi-Action Video Alignment Results

In Sec. 4.3 of the main paper, we train a separate encoder for each action of the Penn Action dataset and report the average results across all 13 actions of the Penn Action dataset, which is not efficient both in terms of time and memory consumption. In this section, we train only a single encoder for all actions of Penn Action and report the multi-action video alignment results in Tab. S4. This is a challenging experiment setting since the shared encoder needs to jointly extract useful features for all actions. We use our VAOT approach for this experiment. It is evident from Tab. S4 that our VAOT approach achieves the best results across all metrics, outperforming all competing methods in this experiment setting. Especially, on Progress, VAOT outperforms previous works by significant margins.

	Method	Acc@1.0	Progress	τ
Penn Action	SAL [8]	68.15	39.03	47.44
	TCN [10]	68.09	38.34	54.17
	TCC [3]	74.39	59.14	64.08
	LAV [4]	78.68	62.52	68.35
	VAVA [7]	80.25	64.82	76.20
	GTCC [1]	73.90	69.70	60.70
	VAOT (Ours)	**83.49**	**79.23**	**77.68**

Table S4. Multi-action video alignment results. **Bold** and underline denote the best and second best respectively.

	Method	Breakfast	YouTube Instructions	50 Salads (Mid)	50 Salads (Eval)	Desktop Assembly
		MoF / F1 / mIoU	MoF / F1 / mIoU	MoF / F1 / mIoU	MoF / F1 / mIoU	MoF / F1 / mIoU
Per-Video	TWF [9]	62.7 / 49.8 / **42.3**	56.7 / 48.2 / -	66.8 / 56.4 / **48.7**	**71.7** / - / -	**73.3** / 67.7 / **57.7**
	ABD [2]	64.0 / 52.3 / -	67.2 / 49.2 / -	**71.8** / - / -	71.22/ - / -	- / - / -
	ASOT [11]	63.3 / 53.5 / 35.9	**71.2** / 63.3 / 47.8	64.3 / 51.1 / 33.4	64.5 / 58.9 / 33.0	73.0 / 68.4 / 47.6
	VASOT (Ours)	**64.5** / **54.3** / 35.8	70.1 / **67.5** / **53.0**	64.7 / **64.2** / 45.1	55.1 / **59.5** / **37.7**	72.1 / **77.2** / 53.2

Table S5. Per-video action segmentation results. **Bold** and underline denote the best and second best respectively.

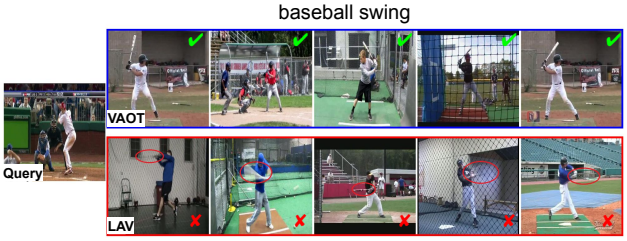
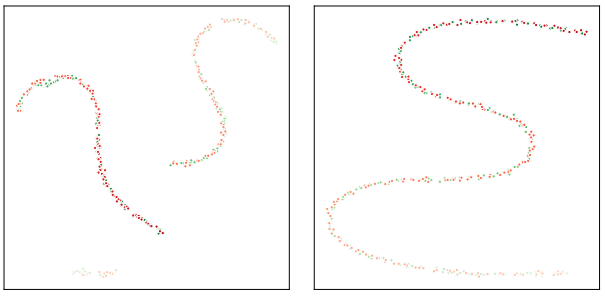


Figure S2. Fine-grained frame retrieval results on Penn Action. The query image is on the left, while on the right are the top 5 matching images retrieved by VAOT (blue box) and LAV (red box).



(a) LAV (b) VAOT (Ours)

Figure S3. t-SNE visualizations of learned frame embeddings of two Pouring videos (green and red). The color opacity represents the temporal frame index from the first frame to the last frame.

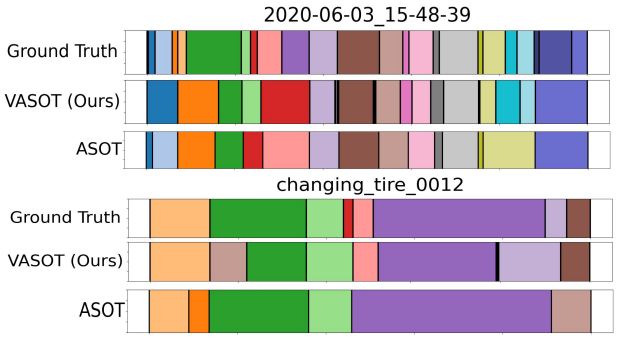


Figure S4. Action segmentation results on Desktop Assembly (top) and YouTube Instructions (bottom).

S5. Per-Video Action Segmentation Results

We perform Hungarian matching over the entire dataset to obtain the full-dataset action segmentation results in Tab. 3 of the main paper. In the following, we benchmark our VASOT approach against previous unsupervised action segmentation methods [2, 9] which conduct Hungarian matching and evaluate per video. Per-video matching and evaluation do not require clusters across all videos and hence tend to yield better results. Tab. S5 presents the results in the per-video matching and evaluation setting. It can be seen from the results that our VASOT approach achieves the best overall performance in this experiment setting. Especially, on F1 score, our VASOT approach consistently performs the best across all datasets.

S6. Qualitative Results

We provide some qualitative results in this section. Firstly, Fig. S2 presents the frame retrieval results of our VAOT approach and LAV [4] on Penn Action. From Fig. S2, our VAOT approach retrieves all 5 correct frames with the same action (*Bat swung back fully*) as the query image, whereas LAV [4] retrieves all incorrect frames (*Bat hits ball*), high-

lighted by red ovals.

Secondly, Fig. S3 illustrates the t-SNE visualizations of the learned frame embeddings of two Pouring videos by our VAOT approach and LAV [4]. In Fig. S3(a), the embeddings by LAV [4] form locally continuous but globally fragmented trajectories, with visible gaps in earlier and later frames. In contrast, for our VAOT approach in Fig. S3(b), the embeddings of corresponding frames from both videos are spatially closer and follow smoother trajectories. This suggests that our VAOT approach can effectively capture both local and global temporal information, resulting in more temporally consistent embeddings.

Lastly, Fig. S4 shows the action segmentation results by our VASOT approach and ASOT [11] on a Desktop Assembly video and a YouTube Instructions video. The Desktop Assembly dataset is relatively balanced, with actions of more uniform durations, whereas the YouTube Instructions dataset is more unbalanced, with some actions being significantly longer or shorter than others. From Fig. S4, our VASOT approach can effectively handle both cases, yielding segmentations which align more closely with ground truth than ASOT [11].

S7. Complexity Comparison Results

We compare the complexity (in terms of model size and training time) of our multi-task VASOT approach and the separate single-task models (VAOT+ASOT) on an Nvidia 3090Ti GPU. Using a ResNet-50 encoder on Pouring, VASOT needs (108 MB, 116 mins) vs. (216 MB, 162 mins) of VAOT+ASOT. Using an MLP encoder on Desktop Assembly, VASOT needs (287 KB, 15 mins) vs. (571 KB, 21 mins) of VAOT+ASOT. The results validate that our multi-task VASOT approach saves both memory and training time as compared to the separate single-task models (VAOT+ASOT).

References

[1] Gerard Donahue and Ehsan Elhamifar. Learning to predict activity progress by self-supervised video alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18667–18677, 2024. 1, 3

[2] Zexing Du, Xue Wang, Guoqing Zhou, and Qing Wang. Fast and unsupervised action boundary detection for action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3323–3332, 2022. 3

[3] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 3

[4] Sanjay Haresh, Sateesh Kumar, Huseyin Coskun, Shahram N Syed, Andrey Konin, Zeeshan Zia, and Quoc-Huy Tran. Learning by aligning videos in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5548–5558, 2021. 1, 3, 4

[5] Anna Kukleva, Hilde Kuehne, Fadime Sener, and Jurgen Gall. Unsupervised learning of action classes with continuous temporal embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12066–12074, 2019. 1

[6] Sateesh Kumar, Sanjay Haresh, Awais Ahmed, Andrey Konin, M Zeeshan Zia, and Quoc-Huy Tran. Unsupervised action segmentation by joint representation learning and online clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20174–20185, 2022. 1

[7] Weizhe Liu, Bugra Tekin, Huseyin Coskun, Vibhav Vineet, Pascal Fua, and Marc Pollefeys. Learning to align sequential actions in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2181–2191, 2022. 1, 3

[8] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016. 3

[9] Saquib Sarfraz, Naila Murray, Vivek Sharma, Ali Diba, Luc Van Gool, and Rainer Stiefelhagen. Temporally-weighted hierarchical clustering for unsupervised action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11225–11234, 2021. 3

[10] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 1134–1141. IEEE, 2018. 3

[11] Ming Xu and Stephen Gould. Temporally consistent unbalanced optimal transport for unsupervised action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14618–14627, 2024. 1, 3, 4