Appendix for Betrayed by Attention: A Simple yet Effective Approach for Self-supervised Video Object Segmentation

Shuangrui Ding[1*], Rui Qian[1*],
Haohang Xu[2], Dahua Lin[1,3,4†], and Hongkai Xiong[2]

[1] The Chinese University of Hong Kong, Hong Kong, China
[2] Shanghai Jiao Tong University, Shanghai, China
[3] Shanghai Artificial Intelligence Laboratory, Shanghai, China
[4] Centre for Perceptual and Interactive Intelligence (CPII), Hong Kong, China
{ds023,qr021}@ie.cuhk.edu.hk

A More Implementation Details

Dataset. For multi-object segmentation in video, we evaluate our method on one synthetic and two real-world video datasets. **MOVi-E** [5] dataset is a synthetic dataset with granular control over data complexity and comprehensive ground truth annotations. MOVi-E scenes contain up to 23 objects and introduce simple linear camera movement. Our evaluation of learned features extends to real-world datasets **DAVIS-17** [19] and **YouTube-VIS-19** [24]. DAVIS-17, an expansion of DAVIS-16, includes 40 additional video sequences along with multi-object segmentation annotations. We utilize 30 validation videos on DAVIS-2017 for evaluation. As for YouTube-VIS-19, due to the lack of mask annotations in the validation or test set, following previous works [1, 29], we select 300 out of the whole 2,883 videos in the training set for evaluation. For MOVi-E and YouTube-VIS-19, we report the Foreground Adjusted Rand Index (FG-ARI) and mean Intersection over Union (mIoU). Furthermore, for DAVIS-2017, we adhere to the standard protocol [19] and report both Region Similarity ($\mathcal{J}$) and Contour Accuracy ($\mathcal{F}$).

Hierarchical Clustering Algorithm. We present our Hierarchical Clustering based inference in Alg. 1. Given the spatio-temporal attention maps $A_v \in \mathbb{R}^{THW \times THW}$, the algorithm finally outputs the cluster centers $A_c \in \mathbb{R}^{N \times THW}$ and cluster assignments $Z \in \{1, ..., N\}^{T'HW}$ which serve as predicted object segmentation masks. Specifically, each attention map is treated as a separate cluster. The process then cycles through each attention map (or current 'cluster'), calculating distances between it and all other clusters using the KL-divergence metric. It identifies the clusters that are close to it (i.e., those whose distance is less than the threshold) and combines them to form a new, larger cluster, represented by their updated centroid. This updated cluster set then replaces the initial set of

* Equal Contribution
† Corresponding author. Email: dhlin@ie.cuhk.edu.hk

attention maps, and the process continues iteratively until no more clusters can be merged. Finally, the algorithm assigns each original attention map to the cluster whose center it is closest to, yielding the final cluster assignments. Note that executing inference on extensive video sequences with a large T value might cause the self-attention matrix to become redundant, thereby requiring significant computational resources. To address this limitation, we sparsely sample T' frames ($T' << T$) as `key`, with the original densely sampled frames as `query`, and calculate the cross-attention $A'_v \in \mathbb{R}^{THW \times T'HW}$. By applying clustering to more compact A'_v, we linearly reduce memory requirements and maintain stable performance as shown in Sec. C.

Algorithm 1 Hierarchical Clustering

Input: Spatio-temporal attention maps $A_v \in \mathbb{R}^{THW \times T'HW}$, distance threshold τ
Output: Cluster assignment $Z \in \{1, ..., N\}^{THW}$
Initialize cluster centers $A_c \leftarrow A_v$
while the number of clusters in A_c changes **do**
  Initialize updated clusters $A_p \leftarrow \{\}$
  for all $x \in A_c$ **do**
    Compute distances: $\mathcal{M} \leftarrow$ `calculate_distance`(x, A_c)
    Identify proximal members: $\mathcal{I} \leftarrow \{i \mid \mathcal{M}[i] < \tau\}$
    Calculate new cluster centroid: $x \leftarrow \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} A_c[i]$
    Add new centroid to updated clusters: $A_p \leftarrow A_p \cup \{x\}$
    Remove merged attention maps from current set: $A_c \leftarrow A_c \backslash A_c[i], \forall i \in \mathcal{I}$
  end for
  Update the clusters: $A_c \leftarrow A_p$
end while
Compute final distances: $\mathcal{M} \leftarrow$ `calculate_distance`(A_v, A_c)
Compute final cluster assignments: $Z \leftarrow$ `argmin`$(\mathcal{M}, \texttt{dim=1})$

B Unsupervised Single Object Segmentation

Besides multi-object segmentation, our method also works in single-object scenarios. We benchmark on three popular datasets designed for single-object segmentation. **DAVIS-16** [18] consists of 50 high-quality videos, 3455 frames in total. Every frame is annotated with a pixel-level accurate segmentation mask. **SegTrack-v2** [8] contains 14 sequences and 947 fully-annotated frames. Each sequence involves 1-6 moving objects and presents challenges including motion blur, appearance change, complex deformation, occlusion, slow motion, and interacting objects. **FBMS-59** [17] has 59 sequences with greatly varied resolution and annotates every 20th frame. Many sequences contain multiple moving objects. Following previous evaluation metric [22, 27], we merge objects of SegTrackv2 and FBMS-59 into a single one for video object segmentation. We calculate the mean per-frame the Jaccard Index $\mathcal{J}$ over the validation set. In

single-object segmentation benchmarks that annotate all objects collectively, we set the distance threshold to 1.6 to combine all foreground objects into one cluster.

Table 1: Quantitative results on single object video segmentation. The tick(✓) and cross(✗) labels under the RGB and Flow columns indicate whether a method utilizes the corresponding modality during training or inference. We compare per frame mean IoU on DAVIS-16, SegTrack-v2 and FBMS-59 without any post-processing (e.g., spectral clustering, test-time adaptation, CRF [6]).

Model	RGB	Flow	DAVIS	ST-v2	FBMS
NLC [4]	✓	✓	55.1	67.2	51.5
CIS [27]	✓	✓	59.2	45.6	36.8
TokenCut [21]	✓	✓	64.3	60.2	59.6
DyStaB [25]	✓	✓	80.0	74.2	73.2
DeSprite [28]	✓	✓	79.1	72.1	71.8
RCF [11]	✓	✓	80.9	76.7	69.9
SIMO [7]	✗	✓	67.8	62.0	-
MG [23]	✗	✓	68.3	58.6	53.1
EM [16]	✗	✓	69.3	55.5	57.8
OCLR [22]	✗	✓	72.1	67.6	65.4
AMD [12]	✓	✗	57.8	57.0	47.5
SMTC [20]	✓	✗	71.8	69.3	68.4
BA (ours)	✓	✗	75.4	74.8	73.3

We present the quantitative results on unsupervised single object discovery in Table 1. Note that all the compared methods are trained on the target dataset, while our model is only trained on YouTube-VIS-19 and directly transferred to these single object segmentation benchmarks in a *zero-shot* manner. Despite this, our method still achieves the best performance among those only using RGB data. Though SMTC [20] proposes a sophisticated VOS framework based on slot attention, our method outperforms it by approximately 5 points. The superiority demonstrates the generalization ability of our approach simply guided by attention. As for the counterparts that resort to optical flow, some of them achieve very promising performance on three benchmarks [11, 26, 28]. It is because optical flow strongly prioritizes moving areas in videos, making it particularly well-suited for single object segmentation tasks. However, the utility of optical flow may diminish for multi-object setups. For instance, it becomes complicated to distinguish between two objects moving in the same direction based solely on flow information. Moreover, it can be difficult to obtain a reliable flow in complex scenarios. Conversely, our method eliminates the need for any optical prior and can be conveniently adapted to accommodate multi-object scenarios.

C More Ablation Study

Table 2: Ablation on different pretrained backbones. We show the results on various DINO and DINOv2 pretrained ViT encoders with different patch sizes.

	YTVIS-19		DAVIS-17	
Model	FG-ARI	mIoU	$\mathcal{J}\&\mathcal{F}$	FG-ARI
DINO ViT-S/16	42.5	47.2	41.7	38.3
DINO ViT-S/8	44.3	50.1	**43.9**	40.1
DINO ViT-B/8	43.5	**50.2**	42.8	40.7
DINOv2 ViT-S/14	44.1	49.7	43.1	40.5
DINOv2 ViT-B/14	**44.5**	50.1	43.7	**41.6**

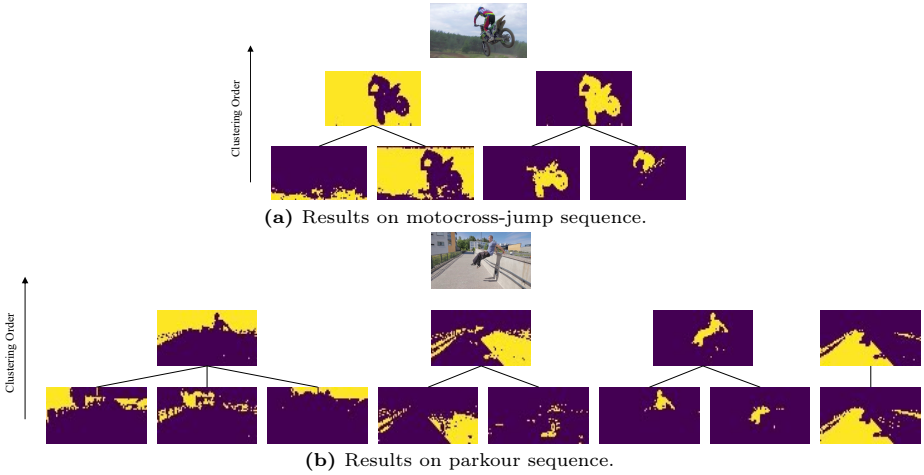
Pretrained backbones. We present the ablation studies on different pretrained backbones in Table. 2. We show the results on both DINO and DINOv2 pretrained ViT encoders with different patch sizes. Generally, our method achieves competitive results on all variants of visual encoders. Comparing the first two lines, i.e., DINO ViT-S/16 vs. DINO ViT-S/8, smaller patch size contributes to notable performance improvements, approximately 2 points on two benchmarks, due to more fine-grained segmentation predictions. Comparing DINO and DINOv2, despite larger patch size, more advanced DINOv2 pretrained backbones reach comparable performance. This reveals the robustness and flexibility of our method to different backbones.

Table 3: Ablation on different numbers of key frames sampled for calculating the spatio-temporal attention matrix. We compare performance under different ratios.

	YTVIS-19		DAVIS-17		Speed
Ratio	FG-ARI	mIoU	$\mathcal{J}\&\mathcal{F}$	FG-ARI	Ratio
0.1	43.5	49.7	42.8	40.0	2.3×
0.2	44.1	50.0	43.4	40.2	2.0×
0.5	44.3	50.2	43.8	40.1	1.4×
1.0	44.3	50.1	43.9	40.1	1.0×

Number of key frames. As stated in the above section, it is feasible to sparsely sample video frames as `key` to reduce computation costs in inference. We present the ablation study in Table 3. We report the multiple object segmentation performance as well as the inference throughput ratio (including the whole feature extraction, attention calculation and clustering process). Interestingly, our findings suggest that promising results for video object segmentation can still be achieved

even when only 10% of the frames are sampled. This sparse sampling approach leads to a remarkable 2.3× speedup in inference. For illustration, given a video with T frames, we uniformly sample $T' = 0.1T$ frames as `key`, with the original T frames as `query`, and calculate the cross-attention $A'_v \in \mathbb{R}^{THW \times T'HW}$. This linearly reduces the channel dimension of each attention map. Then we perform hierarchical clustering on these THW samples with reduced channel dimension and produce the cluster assignments (segmentation masks) for all frames within the video in one shot. The underlying reason is that video frames are highly redundant, sparse sampling could provide an abundant temporal reference for spatio-temporal dependency calculation. Hence, by sampling a small percentage of the entire video, we achieve comparable performance with a substantial reduction in computational cost, leading to faster inference speeds. Furthermore, techniques such as quantization [2, 9, 14] and pruning [3, 10, 13, 15] can provide additional speedups for the method.



(a) Results on motocross-jump sequence.

(b) Results on parkour sequence.

Fig. 1: Visualization of the clustering process. We observe interpretable clustering hierarchies that segment objects at different granularities.

D More Visualization Results

Results of different clustering hierarchies. In our inference stage, the hierarchical clustering algorithm produces different clustering hierarchies. We explore whether there exists an interpretable phenomenon during the clustering process in Fig. 1.

In each hierarchy, we gather the centroids within a distance threshold τ into a new centroid that covers a larger area. Generally, as clustering goes, the hierarchy increases and the result transitions from fine-grained to coarse-grained. For

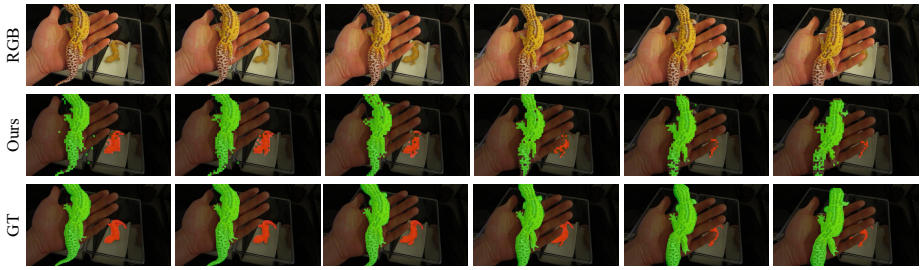
example, the lower hierarchy results in human body parts, and the higher hierarchy results in the whole foreground object like human and motorbike. Smaller τ will force the clustering to stop at lower hierarchies, thus generating fine-grained segmentation. Larger τ will continue to merge the fine-grained cluster centroids, e.g., merge human body parts into a whole human. We find $\tau = 1.0$ works empirically well for various benchmarks.

Interestingly, we observe that our model is able to segment objects at different granularities across hierarchies. Generally, it results in more fine-grained object segmentation in lower hierarchies and vice versa. For example, in Fig. 1a, the model discerns two distinct objects - the motorbike and the human - at a lower hierarchy, subsequently merging them into a cohesive foreground area at a higher hierarchy. Similarly, Fig. 1b shows that the clustering isolates different sections of the human figure at a lower hierarchy, before integrating them to form a holistic human body at a higher hierarchy. Such interpretable hierarchical clustering outcomes yield multi-layered object segmentations, potentially resolving the ambiguities in annotations.

Results on consecutive frames. We additionally show our segmentation results on video sequences with object occlusion, disappearance and reappearance, which is prevalent and challenging in real-world scenarios. In Fig. 2, we present three typical cases. The first is a cat-girl sequence, where there exist mutual occlusions between two objects. Our model is able to accurately segment the object parts despite severe occlusion. The second is a kid-football sequence, where the football disappears in the second frame and reappears in later frames. Since our method refers to the spatio-temporal dependencies across the whole temporal range, it is able to recognize that the ball in the first frame and those in later frames belong to the same instance. This enables our model to process real-world videos with complex temporal dynamics. The third is a very challenging sequence consisting of two lizards, which share very similar colors, body shapes, and textures and only vary in sizes and positions. Moreover, the smaller one is severely occluded by the human hand in the latter three frames. Despite these challenges, our method is still able to distinguish these two lizards and accurately track specific instances over time. These examples demonstrate the applicability of our method to general video scenes.

(a) Results on cat-girl sequence.

(b) Results on kid-football sequence.



(c) Results on lizard sequence.

Fig. 2: Visualization results on video sequences with occlusion. Our model is able to deal with partial or complete object occlusion, where an object disappears in some frames and reappears in later frames.

References

1. Aydemir, G., Xie, W., Guney, F.: Self-supervised object-centric learning for videos. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=919tWtJPXe
2. Ding, L., Fei, W., Huang, Y., Ding, S., Dai, W., Li, C., Zou, J., Xiong, H.: AMPA: Adaptive mixed precision allocation for low-bit integer training. In: Forty-first International Conference on Machine Learning (2024), https://openreview.net/forum?id=HfxFasUfbN
3. Ding, S., Zhao, P., Zhang, X., Qian, R., Xiong, H., Tian, Q.: Prune spatio-temporal tokens by semantic-aware temporal accumulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16945–16956 (2023)
4. Faktor, A., Irani, M.: Video segmentation by non-local consensus voting. In: BMVC. vol. 2, p. 8 (2014)
5. Greff, K., Belletti, F., Beyer, L., Doersch, C., Du, Y., Duckworth, D., Fleet, D.J., Gnanapragasam, D., Golemo, F., Herrmann, C., Kipf, T., Kundu, A., Lagun, D., Laradji, I., Liu, H.T.D., Meyer, H., Miao, Y., Nowrouzezahrai, D., Oztireli, C., Pot, E., Radwan, N., Rebain, D., Sabour, S., Sajjadi, M.S.M., Sela, M., Sitzmann, V., Stone, A., Sun, D., Vora, S., Wang, Z., Wu, T., Yi, K.M., Zhong, F., Tagliasacchi, A.: Kubric: a scalable dataset generator (2022)
6. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001)
7. Lamdouar, H., Xie, W., Zisserman, A.: Segmenting invisible moving objects. In: British Machine Vision Association (2021)
8. Li, F., Kim, T., Humayun, A., Tsai, D., Rehg, J.M.: Video segmentation by tracking many figure-ground segments. In: Proceedings of the IEEE international conference on computer vision. pp. 2192–2199 (2013)
9. Li, H., Li, S., Dai, W., Li, C., Zou, J., Xiong, H.: Frequency-aware transformer for learned image compression. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=HKGQDDTuvZ
10. Li, J., Wang, Y., ZHANG, X., Shi, B., Jiang, D., Li, C., Dai, W., Xiong, H., Tian, Q.: Ailurus: A scalable vit framework for dense prediction. In: Advances in Neural Information Processing Systems. vol. 36, pp. 30979–30996 (2023)
11. Lian, L., Wu, Z., Yu, S.X.: Bootstrapping objectness from videos by relaxed common fate and visual grouping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14582–14591 (2023)
12. Liu, R., Wu, Z., Yu, S., Lin, S.: The emergence of objectness: Learning zero-shot segmentation from videos. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 13137–13152. Curran Associates, Inc. (2021), https://proceedings.neurips.cc/paper/2021/file/6d9cb7de5e8ac30bd5e8734bc96a35c1-Paper.pdf
13. Lu, X., Liu, Q., Xu, Y., Zhou, A., Huang, S., Zhang, B., Yan, J., Li, H.: Not all experts are equal: Efficient expert pruning and skipping for mixture-of-experts large language models. arXiv preprint arXiv:2402.14800 (2024)
14. Lu, X., Zhou, A., Lin, Z., Liu, Q., Xu, Y., Zhang, R., Wen, Y., Ren, S., Gao, P., Yan, J., Li, H.: Terdit: Ternary diffusion models with transformers (2024)
15. Lu, X., Zhou, A., Xu, Y., Zhang, R., Gao, P., Li, H.: SPP: Sparsity-preserved parameter-efficient fine-tuning for large language models. In: Forty-first International Conference on Machine Learning (2024), https://openreview.net/forum?id=9Rroj9GIOQ

16. Meunier, E., Badoual, A., Bouthemy, P.: Em-driven unsupervised learning for efficient motion segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4462–4473 (2022)
17. Ochs, P., Malik, J., Brox, T.: Segmentation of moving objects by long term video analysis. IEEE transactions on pattern analysis and machine intelligence **36**(6), 1187–1200 (2013)
18. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 724–732 (2016)
19. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)
20. Qian, R., Ding, S., Liu, X., Lin, D.: Semantics meets temporal correspondence: Self-supervised object-centric learning in videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16675–16687 (2023)
21. Wang, Y., Shen, X., Yuan, Y., Du, Y., Li, M., Hu, S.X., Crowley, J.L., Vaufreydaz, D.: Tokencut: Segmenting objects in images and videos with self-supervised transformer and normalized cut. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
22. Xie, J., Xie, W., Zisserman, A.: Segmenting moving objects via an object-centric layered representation. In: Advances in Neural Information Processing Systems (2022)
23. Yang, C., Lamdouar, H., Lu, E., Zisserman, A., Xie, W.: Self-supervised video object segmentation by motion grouping. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7177–7188 (October 2021)
24. Yang, L., Fan, Y., Xu, N.: Video instance segmentation. In: ICCV (2019)
25. Yang, Y., Lai, B., Soatto, S.: Dystab: Unsupervised object segmentation via dynamic-static bootstrapping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2826–2836 (June 2021)
26. Yang, Y., Lai, B., Soatto, S.: Dystab: Unsupervised object segmentation via dynamic-static bootstrapping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2826–2836 (2021)
27. Yang, Y., Loquercio, A., Scaramuzza, D., Soatto, S.: Unsupervised moving object detection via contextual information separation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
28. Ye, V., Li, Z., Tucker, R., Kanazawa, A., Snavely, N.: Deformable sprites for unsupervised video decomposition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2657–2666 (2022)
29. Zadaianchuk, A., Seitzer, M., Martius, G.: Object-centric learning for real-world videos by predicting temporal feature similarities. In: Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS 2023) (2023)