Efficient Track Anything

Supplementary Material

In this supplement, we provide more results to demonstrate competing capabilities of our EfficientTAM for video object segmentation and track anything.

1. Additional Related Work

1.1. Vision Transformers

Combining ViT with convolutions has been explored for fast hybrid models such as MobileViT [13], LeViT [7], EfficientFormer[11], Next-ViT[10], Tiny-ViT[18], Castling-ViT[21], EfficientViT [12], and MobileNetv4 [14]. This line of progression towards building efficient ViTs is orthogonal to our EfficientTAM work towards building efficient video object segmentation. Following SAM [9] and EfficientSAMs [20], we are pursuing plain ViT backbones for efficient video object segmentation and track anything tasks.

1.2. Efficient Attention

Local windowed attention has been applied in [2, 22] for reducing the complexity of self-attention. In [8, 16], a linear dot product approximation is proposed to linearize the softmax matrix in self-attention by heuristically separating keys and queries, which can be viewed as a content only lambda layer [1]. In [4], the Performer model uses random features to approximate self-attention, achieving linear time and memory cost. Nyströmformer in [19] makes use of the Nyström method to approximate self-attention with a linear cost. Linformer [17] shows that self-attention is low-rank, which can be approximated by learning linear projection matrices for the keys and values. The approach of [12, 21] leverages the associative property of matrix multiplication for efficient attentions in vision transformers. This direction has shown success and has achieved decent performance on vision tasks. However, it underperforms when applying to memory cross-attention without considering the underlying structure of memory tokens. We take explicit advantage of the structure of the memory spatial tokens for efficient cross-attention.

2. Efficient Cross-Attention

Assume $\tilde{K}_s$ is a coarser representation of memory spatial keys, K_s, a good surrogate of $K_s \in \mathbb{R}^{n\times d}$ with the same size, $\bar{K}_s \in \mathbb{R}^{n\times d}$ from $\tilde{K}_s \in \mathbb{R}^{\tilde{w}\tilde{h}\times d}$ is constructed by stacking each $\tilde{k}_i, i = 1, \dots, \tilde{w}\tilde{h}$, $l_w \times l_h$ times,

$$\bar{K}_s = [\underbrace{\tilde{k}_1; \dots; \tilde{k}_1}_{l_w \times l_h}; \underbrace{\tilde{k}_2; \dots; \tilde{k}_2}_{l_w \times l_h}; \dots; \underbrace{\tilde{k}_{\tilde{w}\tilde{h}}; \dots; \tilde{k}_{\tilde{w}\tilde{h}}}_{l_w \times l_h}]$$

Each $\tilde{v}_i, i = 1, \dots, \tilde{w}\tilde{h}$, is stacked $l_w \times l_h$ times to make $\bar{V}_s \in \mathbb{R}^{n\times d}$ as a good surrogate of values, $V_s \in \mathbb{R}^{n\times d}$,

$$\bar{V}_s = [\underbrace{\tilde{v}_1; \dots; \tilde{v}_1}_{l_w \times l_h}; \underbrace{\tilde{v}_2; \dots; \tilde{v}_2}_{l_w \times l_h}; \dots; \underbrace{\tilde{v}_{\tilde{w}\tilde{h}}; \dots; \tilde{v}_{\tilde{w}\tilde{h}}}_{l_w \times l_h}]$$

The concatenation of coarse spatial tokens with object pointer tokens is, $\bar{K} = [\bar{K}_s; K_p] \in \mathbb{R}^{(n+P)\times d}$ and $\bar{V} = [\bar{V}_s; V_p] \in \mathbb{R}^{(n+P)\times d}$.

Lemma 1. *For the coarse memory tokens, $\bar{K}$ and $\bar{V}$, queries $Q \in \mathbb{R}^{L\times d}$, we have,*

$$\textit{softmax}\left(\frac{Q\bar{K}^T}{\sqrt{d}}\right)\bar{V} = \textit{softmax}\left(A\right)\tilde{V}, \qquad (1)$$

where $A = [\frac{Q\tilde{K}_s^T}{\sqrt{d}} + \ln\left(l_w \times l_h\right), \frac{QK_p^T}{\sqrt{d}}] \in \mathbb{R}^{L\times(\tilde{w}\tilde{h}+P)}$, $\tilde{V} = [\tilde{V}_s; V_p] \in \mathbb{R}^{(\tilde{w}\tilde{h}+P)\times d}$.

Proof. Denote $Q = [q_1; \dots; q_L]$, where $q_i \in \mathbb{R}^{1\times d}$. The cross-attention matrix, $\bar{C} = \mathrm{softmax}\left(\frac{Q\bar{K}^T}{\sqrt{d}}\right)\bar{V} \in \mathbb{R}^{L\times d}$. The softmax matrix $\bar{S} = \mathrm{softmax}\left(\frac{Q\bar{K}^T}{\sqrt{d}}\right) \in \mathbb{R}^{L\times(n+P)}$ can be formulated as,

$$\bar{S} = D_{\mathcal{S}} \begin{bmatrix} e(\frac{q_1}{\sqrt{d}}\tilde{k}_1^T) & \dots & e(\frac{q_1}{\sqrt{d}}\tilde{k}_1^T) & \dots & e(\frac{q_1}{\sqrt{d}}\tilde{k}_{\tilde{w}\tilde{h}}^T) & \dots & e(\frac{q_1}{\sqrt{d}}K_p^T) \\ \vdots & \dots & \vdots & \dots & \vdots & \dots & \dots \\ e(\frac{q_L}{\sqrt{d}}\tilde{k}_1^T) & \dots & e(\frac{q_L}{\sqrt{d}}\tilde{k}_1^T) & \dots & e(\frac{q_L}{\sqrt{d}}\tilde{k}_{\tilde{w}\tilde{h}}^T) & \dots & e(\frac{q_L}{\sqrt{d}}K_p^T) \end{bmatrix}$$

where $D_{\mathcal{S}}$ is a $L \times L$ diagonal matrix, which normalizes each row of the $\bar{S}$ matrix such that the row entries sum up to 1, and $e(\cdot)$ denotes $\exp(\cdot)$. For each row of the cross-attention matrix, we have,

$$\begin{aligned} \bar{C}_{ij} &= D_{\mathcal{S}_{ii}}(\underbrace{e(\frac{q_i}{\sqrt{d}}\tilde{k}_1^T)\tilde{v}_1 + \dots e(\frac{q_i}{\sqrt{d}}\tilde{k}_1^T)\tilde{v}_1}_{l_w \times l_h} + \dots \\ &+ \underbrace{e(\frac{q_i}{\sqrt{d}}\tilde{k}_1^T)\tilde{v}_{\tilde{w}\tilde{h}} + \dots e(\frac{q_i}{\sqrt{d}}\tilde{k}_1^T)\tilde{v}_{\tilde{w}\tilde{h}}}_{l_w \times l_h} + e(\frac{q_i}{\sqrt{d}}K_p^T)V_p) \\ &= D_{\mathcal{S}_{ii}}(l_w \times l_h \times (e(\frac{q_i}{\sqrt{d}}\tilde{k}_1^T)\tilde{v}_1 + \dots + e(\frac{q_i}{\sqrt{d}}\tilde{k}_1^T)\tilde{v}_{\tilde{w}\tilde{h}}) \\ &+ e(\frac{q_i}{\sqrt{d}}K_p^T)V_p) \\ &= D_{\mathcal{S}_{ii}}(l_w \times l_h \times e(\frac{q_i}{\sqrt{d}}\tilde{K}_s^T)\tilde{V}_s^T + e(\frac{q_i}{\sqrt{d}}K_p^T)V_p) \\ &= D_{\mathcal{S}_{ii}}(e(\ln(l_w \times l_h) + \frac{q_i}{\sqrt{d}}\tilde{K}_s^T)\tilde{V}_s + e(\frac{q_i}{\sqrt{d}}K_p^T)V_p) \\ &= \mathrm{softmax}[\frac{q_i\tilde{K}_s^T}{\sqrt{d}} + \ln\left(l_w \times l_h\right), \frac{q_i\tilde{K}_p^T}{\sqrt{d}}][\tilde{V}_s; V_p] \qquad (2) \end{aligned}$$

Object Pointers	MOSE dev	DAVIS 2017 val	SA-V test
No	75.8	89.0	72.1
Yes	76.5	89.2	74.5

Table 1. Ablation study on the design of memory cross-attention in EfficientTAM.

Pooling	MOSE dev	DAVIS 2017 val	SA-V test
Memory tokens	74.5	87.6	71.7
Spatial tokens only	76.5	88.6	74.0

Table 2. Ablation study on taking care of the memory token structure for efficient cross-attention in EfficientTAM.

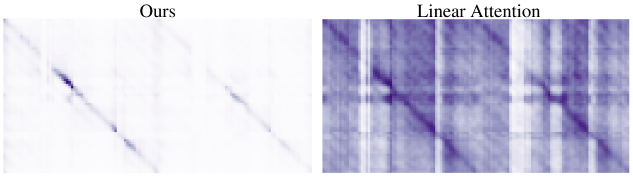

Figure 1. Attention score map.

Cross-Attention	MOSE dev	DAVIS 2017 val	SA-V test
local-windowed	75.4	88.6	72.4
pooling	76.5	88.6	74.0

Table 3. Comparing with local windowed attention.

where $D_{\mathcal{S}_{ii}}$ is the i^{th} diagonal element of the matrix $D_{\mathcal{S}}$. Note that the right side of Eq. (2) is the i^{th} row of softmax $(A)\,\tilde{V}$. It concludes the proof. □

3. Ablation Studies

Impact of the object pointer tokens. In Tab. 1, we ablate the cross-attention with or without the object pointer tokens.
Structure of memory tokens. In Tab. 2, we ablate the impact of memory tokens for efficient cross-attention in the memory module.
Linear cross-attention. Local context is an important component for high-quality segmentation [6]. In Fig. 1, we can see that our attention focuses more on local context while the ability of linear attention to capture local context is reduced. This is also consistent with a recent finding that current linear attentions compromise Softmax attention's ability for local modeling [5]. Therefore, leveraging the underlying token structure for efficient cross-attention is more effective.
Local windowed cross-attention. We adapt local windowed attention for efficient cross-attention by partitioning input tokens into 4 non-overlapping segments (windows), within which we conduct cross-attention. In Tab. 3, we find that local windowed cross-attention underperforms our proposed efficient cross-attention using averaging pooling, 72.4 vs 74.0 $\mathcal{J}\&\mathcal{F}$ on SA-V test dataset. These results demonstrate the effectiveness of our efficient cross-attention by leveraging the strong locality of spatial memory tokens.
Efficient cross-attention. We observe that Eq. (6) in the main paper is close to original cross-attention, visualized in Fig. 2. This suggests that Eq. (6) can serve as a surrogate of the original cross-attention.

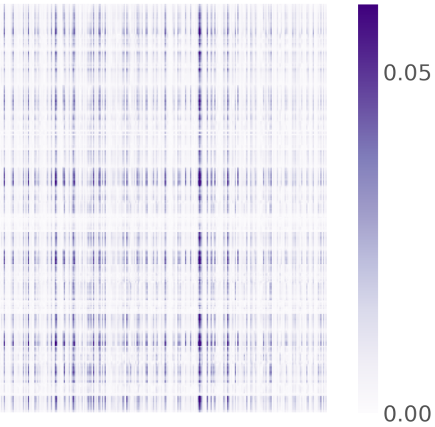

Figure 2. Visualization of the difference between original cross-attention and efficient cross-attention of Eq. (6).

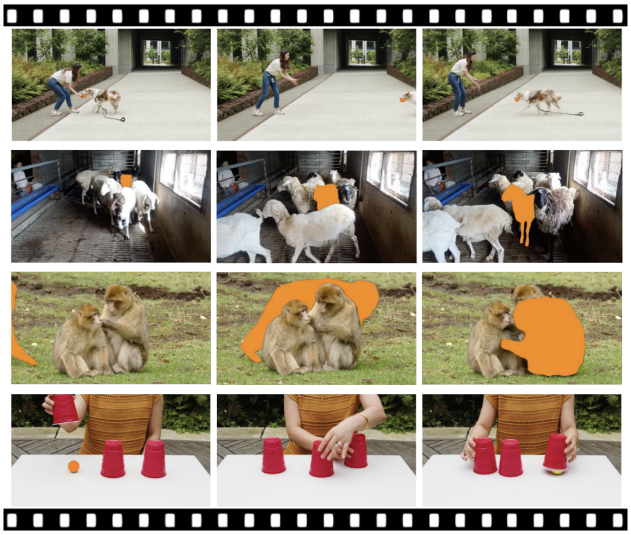

Figure 3. Visualization results on challenging examples. The segmented objects, e.g., toy ball, sheep, monkey, sheep, and croquet ball, are colored in orange.

4. Qualitative Evaluation

We provide more qualitative results of EfficientTAMs for video and image instance segmentation. Fig. 3 shows challenging video examples with small objects and objects with occlusions. We find that our EfficientTAM is able to track small objects and objects with occlusions in Fig. 3. In Fig. 3 (bottom), we also present one interesting failure case of tracking an object after long occlusions (>5 seconds). To address this limitation, long occlusion track anything can be one interesting future direction for performance improvement. For image segmentation, we also observe that our EfficientTAM can generate quality image segmentation results as SAM and SAM 2, shown in Fig. 4. We report the predicted masks with two types of prompts, point and box, and also segment every-

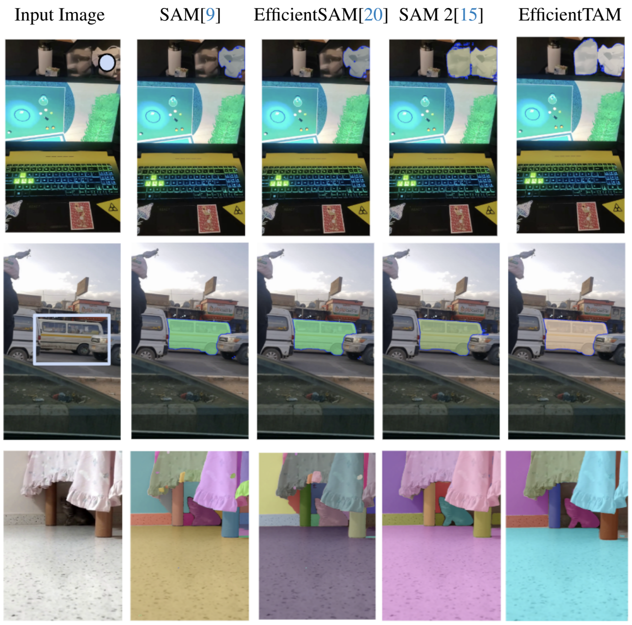

Figure 4. Visualization results on image segmentation with point-prompt, box-prompt, and segment everything for SAM, EfficientSAM, SAM 2, and our EfficientTAM model.

thing results. These results suggest that our EfficientTAMs have similar abilities to SAM 2, while EfficientTAM is more efficient.

5. Discussion

Efficient training techniques. We followed SAM 2[15] for training efficient track anything models using 256 A100-80G GPUs. In our experiments, we note that GPU resources can be reduced by applying efficient training techniques. We find that progressive training can help reduce GPU resources significantly (i.e., 4× fewer GPUs) in our experiments by first training the model on lower resolution, 512×512, and then continuing training on higher resolution, 1024×1024.

Latency-sensitive mobile VOS applications. There is an increasing demand for online video editing tool, which has many use cases such as live streaming [3]. Immediate editing response to interesting objects in streaming frames is necessary for providing a decent online video editing tool. Mobile VOS model is capable of tracking interesting objects across streaming frames, which can serve as an important component of online video editing tool.

References

[1] Irwan Bello. Lambdanetworks: Modeling long-range interactions without attention. In *International Conference on Learning Representations*, 2021. 1

[2] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv:2004.05150*, 2020. 1

[3] Feng Chen, Zhen Yang, Bohan Zhuang, and Qi Wu. Streaming video diffusion: Online video editing with diffusion models. *arXiv preprint arXiv:2405.19726*, 2024. 3

[4] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020. 1

[5] Han et al. Bridging the divide: Reconsidering softmax and linear attention. In *NeurIPS*, 2024. 2

[6] Lin et al. Global-and-local context network for semantic segmentation of street view images. *Sensors*, 2020. 2

[7] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12259–12269, 2021. 1

[8] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020. 1

[9] 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, 3

[10] Jiashi Li, Xin Xia, Wei Li, Huixia Li, Xing Wang, Xuefeng Xiao, Rui Wang, Min Zheng, and Xin Pan. Next-vit: Next generation vision transformer for efficient deployment in realistic industrial scenarios. *arXiv preprint arXiv:2207.05501*, 2022. 1

[11] Yanyu Li, Geng Yuan, Yang Wen, Ju Hu, Georgios Evangelidis, Sergey Tulyakov, Yanzhi Wang, and Jian Ren. Efficientformer: Vision transformers at mobilenet speed. *Advances in Neural Information Processing Systems*, 35:12934–12949, 2022. 1

[12] Xinyu Liu, Houwen Peng, Ningxin Zheng, Yuqing Yang, Han Hu, and Yixuan Yuan. Efficientvit: Memory efficient vision transformer with cascaded group attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14430, 2023. 1

[13] Sachin Mehta and Mohammad Rastegari. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. In *International Conference on Learning Representations*, 2021. 1

[14] Danfeng Qin, Chas Leichner, Manolis Delakis, Marco Fornoni, Shixin Luo, Fan Yang, Weijun Wang, Colby Banbury, Chengxi Ye, Berkin Akin, et al. Mobilenetv4-universal models for the mobile ecosystem. *arXiv preprint arXiv:2404.10518*, 2024. 1

[15] 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. 3

[16] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient attention: Attention with linear complexities. *arXiv preprint arXiv:1812.01243*, 2018. 1

[17] Sinong Wang, Belinda Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020. 1

[18] Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Tinyvit: Fast pretraining distillation for small vision transformers. In *European Conference on Computer Vision*, pages 68–85. Springer, 2022. 1

[19] Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14138–14148, 2021. 1

[20] Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, et al. Efficientsam: Leveraged masked image pretraining for efficient segment anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16111–16121, 2024. 1, 3

[21] Haoran You, Yunyang Xiong, Xiaoliang Dai, Bichen Wu, Peizhao Zhang, Haoqi Fan, Peter Vajda, and Yingyan Celine Lin. Castling-vit: Compressing self-attention via switching towards linear-angular attention at vision transformer inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14431–14442, 2023. 1

[22] Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *arXiv preprint arXiv:2007.14062*, 2020. 1