

EdgeTAM: On-Device Track Anything Model

Supplementary Material

Table 1. **Zero-shot accuracy across 17 video datasets under semi-supervised VOS evaluation using different prompts.** For all prompt types, the annotation is only provided on the first frame. [†]: When the ground-truth mask is available, SAM is not used for XMem++ and Cutie.

Method	1-click	3-click	5-click	bounding box	ground-truth mask [†]
SAM + XMem++ [3]	56.9	68.4	70.6	67.6	72.7
SAM + Cutie [6]	56.7	70.1	72.2	69.4	74.1
SAM 2 [17]	64.3	73.2	75.4	72.9	77.6
SAM 2.1 [17]	64.7	75.3	77.6	74.4	79.3
EdgeTAM	54.4	72.7	75.5	71.3	77.0

1. Video Object Segmentation (VOS)

In our main submission, we follow the standard semi-supervised video object segmentation protocol, where the ground-truth masks on the first frame are available during inference. In Tab. 1, we follow SAM 2 [17] and instead of making the masks on the first frame available, we prompt the object of interest with clicks or boxes on the first frame. Given that XMem++ and Cutie do not support these prompts, we convert the prompt to masks with SAM [15]. We evaluate on 17 zero-shot datasets including EndoVis 2018 [2], ESD [14], LVOSv2 [13], LV-VIS [20], UVO [21], VOST [18], PUMaVOS [3], Virtual KITTI 2 [5], VIPSeg [16], Wildfires [19], VISOR [8], FBMS [4], Ego-Exo4D [11], Cityscapes [7], Lindenthal Camera [12], HT1080WT Cells [10], and Drosophila Heart [9].

In this evaluation suite, except for the 1-click setting, EdgeTAM surpasses the strong baselines, SAM + XMem++ and SAM + Cutie, by 2 to 5 percent. Compared to SAM 2 and SAM 2.1, EdgeTAM still preserves comparable performance especially with more accurate prompts, such as 5-click and ground-truth mask.

2. Implementation Details

We generally follow the original SAM 2 training hyperparameters for image segmentation pre-training [15] and video segmentation training [17]. Here, we highlight only the differences, and the full training details are shown in Tab. 2. First, we do not apply drop path or layer-wise decay in the image encoder. Second, our image pre-training stage adopts a 128 batch size and a total of 175K training steps. In the video training stage, we reduce the maximum number of masks per image from 64 to 32. More importantly, we do not train on the SAM 2 Internal dataset so the total training steps are reduced from 300K to 130K. Finally, our training involves distillation losses in both stages.

3. Speed Benchmark

In the main paper, we provide the throughput FPS on both server GPUs (NVIDIA A100 and V100) and mobile NPU (iPhone 15 Pro Max). The V100 benchmarks are collected from each individual paper and we benchmark with the other two hardware by ourselves. In particular, to optimize the throughput, on A100, we torch compile all the models. For mobile NPU, we convert the model to CoreML format with coremltools [1] and benchmark with the performance report tool of XCode with iOS 18.1 on an iPhone 15 Pro Max. Note that, the speed-up ratios of EdgeTAM v.s. SAM 2 are less pronounced on A100 than on iPhone. To understand the root cause, we monitor the streaming multiprocessor (SM) utilization of both models on A100 and find that even with torch compile, the SM usage of EdgeTAM is less than 50% and the inference is bottlenecked on CPU and IO. We think it is because high-end server GPUs, such as A100, have an enormous amount of parallel executable units (EU) and given the tiny size of EdgeTAM, it cannot occupy all the EUs at the same time. However, the design objective of EdgeTAM is edge devices, such as mobile phones, where we see 22 \times speed-up compared with SAM 2.

4. Video Results

To better show the qualitative results of EdgeTAM, in **5205.mp4**, we provide the tracking results in the video format across several challenging cases.

Table 2. Hyperparameters and details of EdgeTAM image segmentation pre-training and video segmentation training.

(a) Image segmentation pre-training.		(b) Video segmentation training.	
Config	Value	Config	Value
data	SA-1B	data	SA-1B, SA-V, DAVIS, MOSE, YTVOS
steps	~175K	steps	~130K
resolution	1024	resolution	1024
precision	bfloat16	precision	bfloat16
optimizer	AdamW	optimizer	AdamW
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.999$	optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.999$
gradient clipping	type: ℓ_2 , max: 0.1	gradient clipping	type: ℓ_2 , max: 0.1
weight decay	0.1	weight decay	0.1
learning rate (lr)	$4e^{-4}$	learning rate (lr)	backbone: $6e^{-5}$, other: $3e^{-4}$
lr schedule	reciprocal sqrt timescale=1000	lr schedule	cosine
warmup	linear, 1K iters	warmup	linear, 15K iters
cooldown	linear, 5K iters	img. augmentation	hflip
augmentation	hflip	vid. augmentation	hflip, affine (deg: 25, shear: 20), colorjitter (0.1), grayscale (0.05), per frame colorjitter (0.1)
batch size	128	batch size	256
mask losses (weight)	focal (20), dice (1)	mask losses (weight)	focal (20), dice (1)
IoU loss (weight)	ℓ_1 (1)	IoU loss (weight)	ℓ_1 (1)
distill loss (weight)	MSE (1)	occlusion loss (weight)	cross-entropy (1)
max. masks per img.	64	distill loss (weight)	MSE (1) for both \mathcal{L}_{img} and \mathcal{L}_{mem}
# correction points	7	max. masks per frame	image: 32, video: 3
		# correction points	7

References

- [1] Core ml tools. <https://github.com/apple/coremltools>, 2021. 1
- [2] Max Allan, Satoshi Kondo, Sebastian Bodendstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, et al. 2018 robotic scene segmentation challenge. *arXiv preprint arXiv:2001.11190*, 2020. 1
- [3] Maksym Bekuzarov, Ariana Bermudez, Joon-Young Lee, and Hao Li. Xmem++: Production-level video segmentation from few annotated frames. In *ICCV*, 2023. 1
- [4] T Brox, J Malik, and P Ochs. Freiburg-berkeley motion segmentation dataset (fbms-59). In *ECCV*, 2010. 1
- [5] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020. 1
- [6] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *CVPR*, 2024. 1
- [7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 1
- [8] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. In *NeurIPS*, 2022. 1
- [9] Matthew Fishman, Abigail Matt, Fei Wang, Elena Gracheva, Jiantao Zhu, Xiangping Ouyang, Andrey Komarov, Yuxuan Wang, Hongwu Liang, and Chao Zhou. A drosophila heart optical coherence microscopy dataset for automatic video segmentation. *Scientific data*, 2023. 1
- [10] Estibaliz Gómez-de Mariscal, Hasini Jayatilaka, Özgün Çiçek, Thomas Brox, Denis Wirtz, and Arrate Muñoz-Barrutia. Search for temporal cell segmentation robustness in phase-contrast microscopy videos. *arXiv preprint arXiv:2112.08817*, 2021. 1
- [11] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *CVPR*, 2024. 1
- [12] Timm Haucke and Volker Steinhage. Exploiting depth information for wildlife monitoring. *arXiv preprint arXiv:2102.05607*, 2021. 1
- [13] Lingyi Hong, Zhongying Liu, Wenchao Chen, Chenzhi Tan, Yuang Feng, Xinyu Zhou, Pinxue Guo, Jinglun Li, Zhaoyu Chen, Shuyong Gao, et al. Lvos: A benchmark for large-scale long-term video object segmentation. *arXiv preprint arXiv:2404.19326*, 2024. 1
- [14] Xiaoqian Huang, Kachole Sanket, Abdulla Ayyad, Fari-borz Baghaei Naeini, Dimitrios Makris, and Yahya Zweiri. A neuromorphic dataset for object segmentation in indoor cluttered environment. *arXiv preprint arXiv:2302.06301*, 2023. 1
- [15] 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 *CVPR*, 2023. 1

- [16] Jiaxu Miao, Xiaohan Wang, Yu Wu, Wei Li, Xu Zhang, Yunchao Wei, and Yi Yang. Large-scale video panoptic segmentation in the wild: A benchmark. In *CVPR*, 2022. [1](#)
- [17] 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](#)
- [18] Pavel Tokmakov, Jie Li, and Adrien Gaidon. Breaking the” object” in video object segmentation. In *CVPR*, 2023. [1](#)
- [19] Tom Toulouse, Lucile Rossi, Antoine Campana, Turgay Celik, and Moulay A Akhloufi. Computer vision for wildfire research: An evolving image dataset for processing and analysis. *Fire Safety Journal*, 2017. [1](#)
- [20] Haochen Wang, Cilin Yan, Shuai Wang, Xiaolong Jiang, Xu Tang, Yao Hu, Weidi Xie, and Efstratios Gavves. Towards open-vocabulary video instance segmentation. In *ICCV*, 2023. [1](#)
- [21] Weiyao Wang, Matt Feiszli, Heng Wang, and Du Tran. Unidentified video objects: A benchmark for dense, open-world segmentation. In *ICCV*, 2021. [1](#)