SAM-I2V: Upgrading SAM to Support Promptable Video Segmentation with Less than 0.2% Training Cost (*Supplementary Material*)

Haiyang Mei, Pengyu Zhang, Mike Zheng Shou*
Show Lab, National University of Singapore

1. Overview

In this supplementary material, we provide additional experimental results and technical details to complement the main paper. Specifically, we demonstrate the applicability of our proposed SAM-I2V across different SAM variants in section 2. We also present comprehensive evaluations on the semi-supervised video object segmentation (Semi-VOS) task in section 3. Then, we include visual comparisons with state-of-the-art methods in section 4. We further compare and analyze the computational efficiency in section 5 and explore SAM-I2V's scalability in section 6. Finally, detailed descriptions of the modules in our pipeline, including memory encoder, memory attention and mask decoder, are provided in section 7.

2. Applicability of SAM-I2V Across Different SAM Variants

To further validate the versatility and robustness of our proposed SAM-I2V approach, we conducted additional experiments on various SAM variants, upgrading them to promptable video segmentation (PVS) models. Figure 1 demonstrates the performance improvements (under the online, 3-click PVS setting) when upgraded with SAM-I2V across five SAM variants, including TinySAM [9], EdgeSAM [12], MobileSAM [11], SlimSAM [3], and SAM-Base [6]. The results highlight the applicability of SAM-I2V in upgrading these promptable image segmentation models to promptable video segmentation models.

For each SAM variant, SAM-I2V consistently brings performance ($J\&F$ [7]) gains across datasets, including ESD [5], PUMA [1], LV-VIS [10], and SAV-Test [8], as well as the overall accuracy (*OA*) . For example, in the case of TinySAM [9], our approach improves the *OA* from 61.7 to 69.3, representing an improvement of 7.6 points. Similarly, for EdgeSAM [12], SAM-I2V achieves an improvement of 6.6 points in *OA*, demonstrating the robustness of our approach across different SAM models.

*Corresponding Author

These results reaffirm that our SAM-I2V can serve as an efficient and adaptable image-to-video upgradation framework, allowing various SAM models to transition into PVS models without the need for costly training.

3. Comparison on Semi-Supervised Video Object Segmentation

In addition to the online and offline promptable video segmentation results presented in the main paper, we further evaluate our method on the semi-supervised video object segmentation (Semi-VOS) task. For this task, prompts were provided only on the first frame, and the model was tasked with tracking the object through the remainder of the video. This task highlighted the model's capacity for autonomous object tracking without continuous user input, showcasing robustness and generalization in scenarios without ongoing guidance. Table 1 summarizes the performance comparisons under three types of first-frame prompts (*i.e.*, 3-clicks, bounding box, and ground-truth mask) across four benchmark datasets (*i.e.*, ES [5], PU [1], LV [10], and SV [8]).

Despite not including any VOS datasets during training, our method achieves performance comparable to state-of-the-art VOS methods while utilizing fewer model parameters. Specifically, our method, *TinySAM + SAM-I2V*, attains an overall segmentation performance (*OA*) of **70.2**, which is competitive with larger models like *TinySAM [9] + Cutie [4]* (*OA* of 70.6) that require more than double the parameters (45.1M vs. **18.9**M). Moreover, compared to baseline image-to-video upgrade methods such as *TinySAM + TA [8]* and *TinySAM + SA [2] + TA [8]*, our method shows substantial improvements of **7.5** and **4.7** in *OA*, respectively, while maintaining similar computational costs.

Under different first-frame prompt settings, our method consistently outperforms the baselines. For instance, with the *3-click* prompt on the ES [5] dataset, our method achieves an accuracy of **84.7**, surpassing *TinySAM + TA* (80.0) and *TinySAM + SA + TA* (79.0). Similarly, with the *bounding box* prompt on the PU [1] dataset, our method attains **66.1** accuracy, exceeding the baselines by signifi-

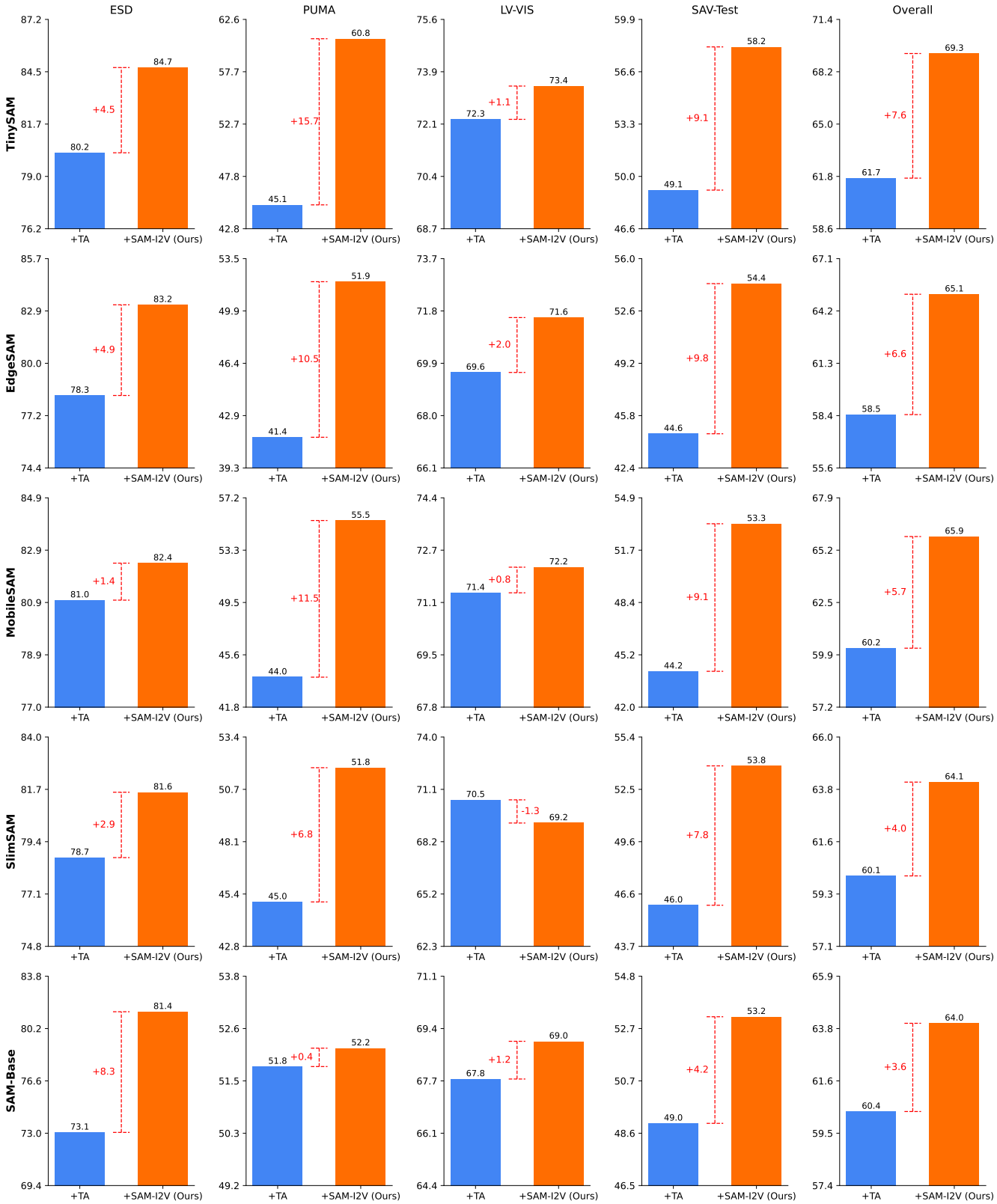

Figure 1. The applicability of our proposed SAM-I2V to upgrade different SAM models for promptable video segmentation (PVS) task under the online, 3-click setting. The comparison includes five SAM variants: TinySAM [9], EdgeSAM [12], MobileSAM [11], SlimSAM [3], and SAM-Base [6].

Methods	Para. (M)	Cost	3-click				bounding box				ground-truth mask†				Average				OA
			ES	PU	LV	SV	ES	PU	LV	SV	ES	PU	LV	SV	ES	PU	LV	SV	
SAM 2.1 [8]	38.9	2.2m	86.6	66.8	78.1	74.4	87.8	75.4	77.7	75.1	90.2	80.9	82.2	76.5	88.2	74.4	79.3	75.3	79.3
SAM [6] + XMem++ [1]	157.0	-	84.5	52.7	71.6	58.4	86.3	63.0	73.0	59.7	90.0	67.9	80.7	61.5	86.9	61.2	75.1	59.9	70.8
SAM [6] + Cutie [4]	129.8	-	84.6	51.0	71.9	61.1	86.1	60.0	73.4	62.8	89.7	62.5	81.4	64.8	86.8	57.8	75.6	62.9	70.8
TinySAM [9] + XMem++ [1]	72.3	-	84.0	59.7	71.3	58.8	84.5	60.2	70.1	58.4	90.0	67.9	80.7	61.5	86.2	62.6	74.0	59.6	70.6
TinySAM [9] + Cutie [4]	45.1	-	84.1	56.7	72.0	61.1	84.3	58.5	70.8	61.3	89.7	62.5	81.4	64.8	86.0	59.2	74.7	62.4	70.6
TinySAM [9] + TA [8]	17.8	4.6k	80.0	45.0	71.2	48.6	80.2	47.3	69.3	48.6	83.1	52.5	77.3	49.6	81.1	48.3	72.6	48.9	62.7
TinySAM [9] + SA [2] + TA [8]	18.7	4.6k	79.0	54.9	71.5	49.8	79.9	57.6	70.5	49.8	83.5	61.1	77.4	50.5	80.8	57.9	73.1	50.0	65.5
TinySAM [9] + **SAM-I2V** (Ours)	18.9	4.6k	84.7	58.1	72.3	57.0	83.2	66.1	69.5	56.5	87.8	68.9	79.3	59.3	85.2	64.4	73.7	57.6	**70.2**

Table 1. Comparisons of the semi-supervised video object segmentation with three types of prompt (*i.e.*, 3-click, bounding box, and ground-truth mask) in the first video frame on four benchmark datasets (*i.e.*, ES [5], PU [1], LV [10], and SV [8]). "-" indicates directly combining existing pre-trained models for inference. "*Cost*" is calculated by *GPU number* × *GPU memory* × *training hours*. "*OA*" denotes the overall performance. "†" indicates the case where we directly use masks as inputs into VOS model without using SAM.

cant margins. When using the ground-truth mask as the prompt, our method still maintains superior performance across datasets. These results clearly validate the effectiveness of our approach in zero-shot semi-supervised VOS performance under various prompt settings.

4. Visual Comparison with State-of-the-Art Methods

We further present visual comparisons of our method with state-of-the-art approaches under the online, 3-click PVS setting. As shown in Figures 4–8, our method, TinySAM + **SAM-12V**, robustly tracks and segments objects with fine details across challenging scenarios, maintaining consistency over time. Specifically, Figure 4 demonstrates our method's ability to handle ***occlusion*** effectively; Figure 5 showcases accurate segmentation of ***small objects***; Figure 6 illustrates robust tracking during ***large spatial movements***; Figure 7 and 8 highlight the segmentation of objects with ***complex shapes***. While differences in segmentation results can be observed across methods, our approach consistently delivers competitive performance, particularly in scenarios requiring fine-grained object representation and robust tracking across multiple frames. This clearly demonstrate the effectiveness of our method.

5. FLOPs Comparison and Analysis

We focus in this work on developing an image-to-video SAM-upgrader to support promptable video segmentation with ***academic-affordable training cost***. For completeness, we report computational efficiency comparisons in Table 2.

First, we can see from Table 2 that our method exhibits superior computational efficiency (87.4G) compared to SAM 2.1 [8] (139.9G) and SAM-based [6] variants (805.2G–811.9G), while maintaining parity with TinySAM-driven [9] approaches (72.9G–103.0G).

Methods	**Para. (M)**	**FLOPs (G)**
SAM 2.1 [8]	38.9	139.9
SAM [6] + XMem++ [1]	157.0	805.2
SAM [6] + Cutie [4]	129.8	811.9
TinySAM [9] + XMem++ [1]	72.3	96.3
TinySAM [9] + Cutie [4]	45.1	103.0
TinySAM [9] + TA [8]	17.8	72.9
TinySAM [9] + SA [2] + TA [8]	18.7	75.0
TinySAM [9] + **SAM-I2V** (Ours)	18.9	87.4

Table 2. Comparisons in terms of model parameters and FLOPs.

SAM-I2V	**Para. (M)**	**FLOPs (G)**
Temporal Feature Integrator	3.7	26.3
Memory Selective Associator: Memory Encoder	1.4	5.8
Memory Selective Associator: Memory Selector	0.0	0.107
Memory Selective Associator: Memory Attention	2.8	14.6
Memory Prompt Generator	0.6	0.9
Overall	**8.5**	**47.707**

Table 3. Analysis of SAM-I2V's computational efficiency.

Second, as shown in Table 3, in our SAM-I2V, the temporal feature integrator (26.3G) and memory selective associator: memory attention (14.6G) collectively account for 85.7% of the upgrader's computational overhead. This is primarily driven by temporal feature extraction and memory-guided attention operations, respectively, establishing them as critical targets for future FLOPs optimization. Besides, our memory selector (0.107G) achieves extended historical frame association (20 frames versus SAM 2's 6 frames) with extra 158 MB GPU memory footprint. This represents 0.64% of the total capacity in modern 24GB GPUs, where memory allocation is dominated by similarity score computation between the current frame and the historical feature buffer.

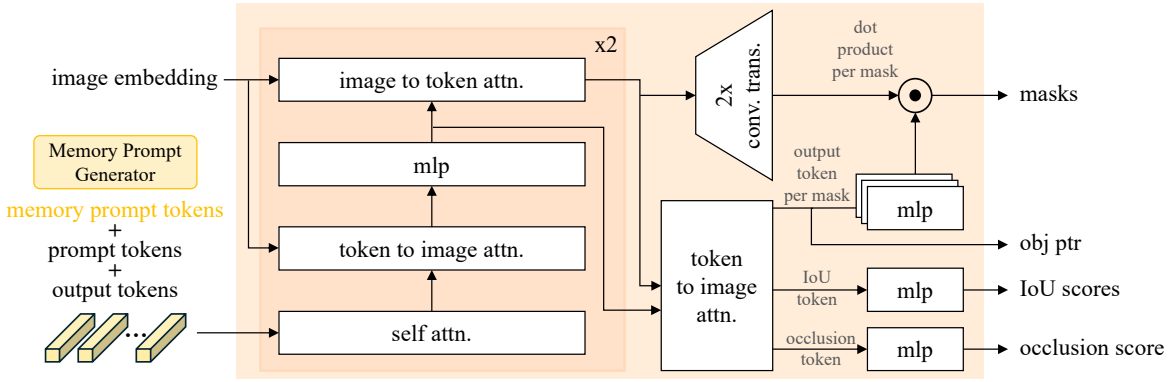

Figure 2. Architecture details of the mask decoder.

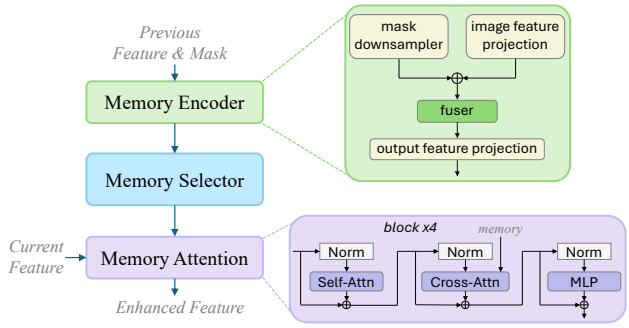

Figure 3. Illustration of the memory selective associator (MSA).

SAM-I2V	Training Configuration (*#Num.* × *#Mem.* × *#Dur.*)	Training Cost	SAV-Test (*J&F*)
(*a*)	8 × 24 G × 24 hours	4.6k	59.3
(*b*)	8 × 24 G × 48 hours	9.2k	62.6
(*c*)	16 × 24 G × 24 hours	9.2k	62.9
(*d*)	32 × 24 G × 24 hours	18.4k	65.2

Table 4. Ablation study on SAM-I2V's scalability.

Third, our SAM-I2V achieves image-to-video upgradation with extra 8.5M parameters and 47.707G FLOPs, demonstrating training feasibility under academic GPU constraints. This establishes a practical foundation for training-resource-efficient PVS model development.

6. SAM-I2V's Scalability

We further explored the scalability of our SAM-I2V when additional GPU resources are available. As shown in Table 4, as we increase the training duration (*b*) or the number of GPUs (*c* and *d*), SAM-I2V's training cost grows proportionally but yields higher SAV-Test performance (59.3 to 65.2), indicating the model's scalability under greater training investments.

7. Architecture Details

Here we further present architecture details, expanding on the model description in the main manuscript.

7.1. Memory Selective Associator

As illustrated in Figure 3, our proposed MSA consists of three sub-networks, *i.e.*, memory encoder, memory selector and memory attention.

The **memory encoder** is a crucial component designed to transform predictions and image encoder embeddings into representations suitable for future frames in the video segmentation. As shown in the top-right of Figure 3, the memory encoder incorporates a combination of downsampled mask features and projected image features, which are fused using convolutional layers. This fusion process ensures that spatial and contextual information from both input sources is effectively integrated. The resulting fused features are then passed through an output feature projection layer to prepare them for the following memory attention.

The **memory attention** is designed to condition the current frame features on the past frames' features and predictions. This conditioning is achieved through a stack of $B = 4$ transformer blocks, where each block consists of three main components: a self-attention layer, a cross-attention layer, and a feedforward multi-layer perceptron (MLP). The self-attention layer processes the current frame's features to capture intra-frame relationships, ensuring a comprehensive understanding of spatial dependencies within the frame. The cross-attention layer enables interaction between the current frame and the selective memories, which include features from both prompted and unprompted previous frames. The memories are stored in the memory bank and are selectively retrieved based on relevance. Each attention block is normalized before and after the attention operations to maintain stability during training. 2D spatial Rotary Positional Embedding (RoPE) is utilized

within self-attention and cross-attention layers to enhance the spatial correspondence of features. Following the attention stages, the MLP further refines the fused features to enhance representational capability. This modular design allows the model to integrate temporal context across video frames, ensuring robust segmentation predictions.

Overall, the memory selective associator plays a pivotal role in enabling the model to maintain and utilize useful temporal information across video frames, facilitating robust mask propogation for accurate video segmentation.

7.2. Mask Decoder

The mask decoder is designed to segment objects based on image embeddings and prompt tokens. Our architecture extends the design of SAM's mask decoder, incorporating additional memory prompts for enhanced segmentation across video frames.

In Figure 2, the prompt tokens represent input guidance such as clicks, bounding boxes, or masks, while memory prompt tokens encode temporal information from previous frames to guide the segmentation in current frame. The inclusion of memory prompt tokens enhances temporal consistency by leveraging the historical context of target objects. The mask decoder employs a sequence of two-way transformer blocks to enable bidirectional attention between image embeddings and tokens. Specifically, the following attention mechanisms are employed:

1. **Self-attention**: Applied to prompt tokens to learn interactions within the token space.
2. **Token-to-image attention**: Enables token embeddings to query relevant image features.
3. **Image-to-token attention**: Aggregates image feature responses into token representations.

After feature fusion through the transformer blocks, the decoder outputs multiple predictions per frame to handle prompt ambiguities (*e.g.*, a click on the tire of a bicycle could correspond to either the tire or the entire bicycle). To disambiguate these predictions, we propagate only the mask with the highest predicted Intersection over Union (IoU) score. Additionally, following [8], the decoder features an *occlusion prediction head*, implemented as an MLP, which predicts whether the target object is visible in the current frame. This is crucial for handling frames where the object is partially or fully occluded.

References

[1] Maksym Bekuzarov, Ariana Bermudez, Joon-Young Lee, and Hao Li. Xmem++: Production-level video segmentation from few annotated frames. In *ICCV*, 2023. 1, 3, 6, 7, 8, 9, 10

[2] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *NeurIPS*, 2022. 1, 3, 6, 7, 8, 9, 10

[3] Zigeng Chen, Gongfan Fang, Xinyin Ma, and Xinchao Wang. 0.1% data makes segment anything slim. *NeurIPS*, 2024. 1, 2

[4] 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, 3, 6, 7, 8, 9, 10

[5] Xiaoqian Huang, Kachole Sanket, Abdulla Ayyad, Fariborz Baghaei Naeini, Dimitrios Makris, and Yahya Zweiri. A neuromorphic dataset for object segmentation in indoor cluttered environment. *arXiv:2302.06301*, 2023. 1, 3

[6] 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 *ICCV*, 2023. 1, 2, 3, 6, 7, 8, 9, 10

[7] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017. 1

[8] 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:2408.00714*, 2024. 1, 3, 5, 6, 7, 8, 9, 10

[9] Han Shu, Wenshuo Li, Yehui Tang, Yiman Zhang, Yihao Chen, Houqiang Li, Yunhe Wang, and Xinghao Chen. Tinysam: Pushing the envelope for efficient segment anything model. *arXiv:2312.13789*, 2023. 1, 2, 3, 6, 7, 8, 9, 10

[10] 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, 3, 7, 8, 9, 10

[11] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv:2306.14289*, 2023. 1, 2

[12] Chong Zhou, Xiangtai Li, Chen Change Loy, and Bo Dai. Edgesam: Prompt-in-the-loop distillation for on-device deployment of sam. *arXiv:2312.06660*, 2023. 1, 2

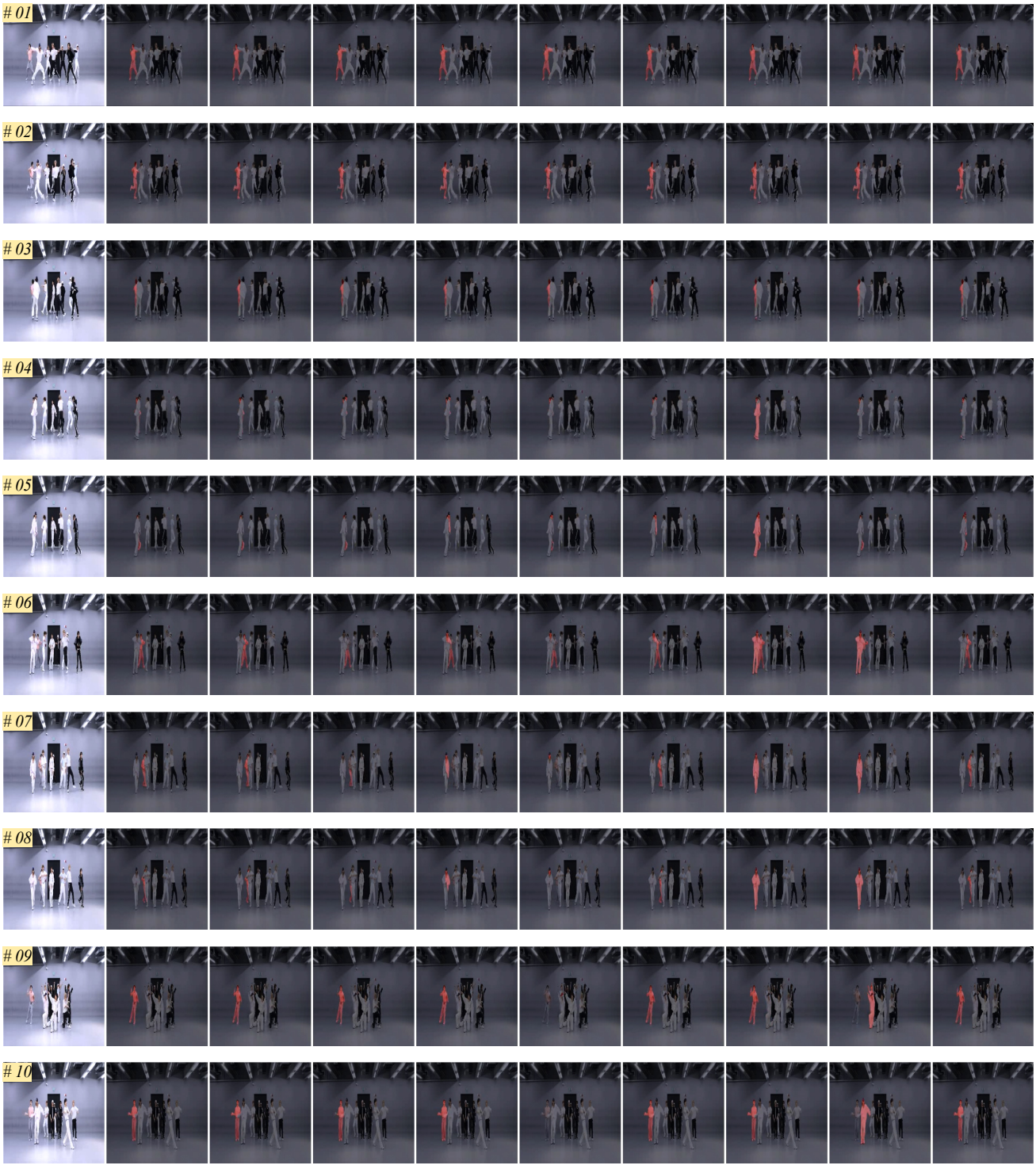

Figure 4. Visual comparison with state-of-the-art methods on a challenging video sequence from PUMaVOS [1] dataset. The methods compared include SAM 2.1 [8], SAM [6] + XMem++ [1], SAM [6] + Cutie [4], TinySAM [9] + XMem++ [1], TinySAM [9] + Cutie [4], TinySAM [9] + TA [8], and TinySAM [9] + SA [2] + TA [8]. Our method, TinySAM + **SAM-12V**, demonstrates robust tracking and fine-grained segmentation on video frames, offering competitive performance across challenging scenarios.

Figure 5. Visual comparison with state-of-the-art methods on a challenging video sequence from LV-VIS [10] dataset. The methods compared include SAM 2.1 [8], SAM [6] + XMem++ [1], SAM [6] + Cutie [4], TinySAM [9] + XMem++ [1], TinySAM [9] + Cutie [4], TinySAM [9] + TA [8], and TinySAM [9] + SA [2] + TA [8]. Our method, TinySAM + **SAM-12V**, demonstrates robust tracking and fine-grained segmentation on video frames, offering competitive performance across challenging scenarios.

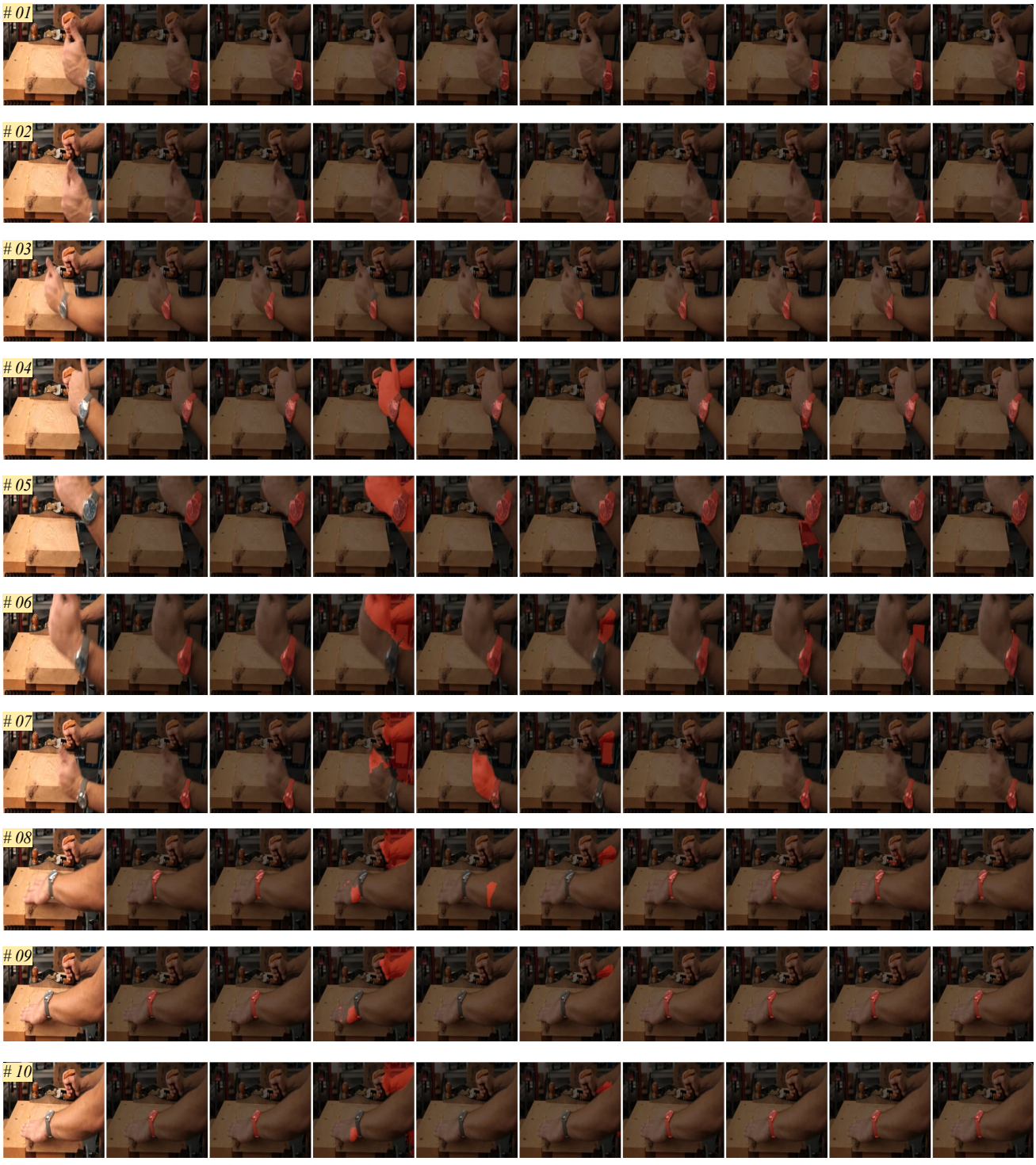

Figure 6. Visual comparison with state-of-the-art methods on a challenging video sequence from LV-VIS [10] dataset. The methods compared include SAM 2.1 [8], SAM [6] + XMem++ [1], SAM [6] + Cutie [4], TinySAM [9] + XMem++ [1], TinySAM [9] + Cutie [4], TinySAM [9] + TA [8], and TinySAM [9] + SA [2] + TA [8]. Our method, TinySAM + **SAM-12V**, demonstrates robust tracking and fine-grained segmentation on video frames, offering competitive performance across challenging scenarios.

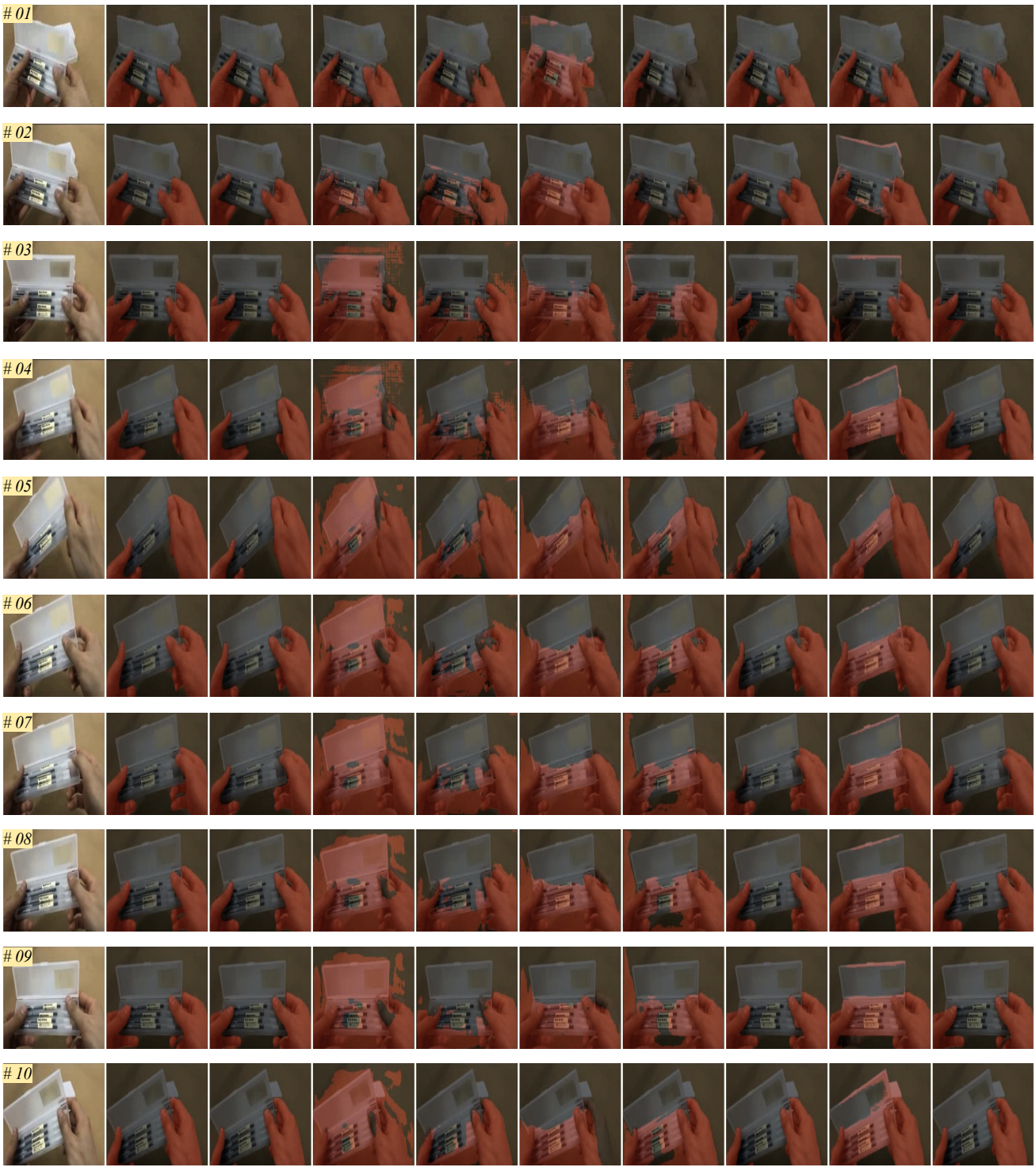

Figure 7. Visual comparison with state-of-the-art methods on a challenging video sequence from LV-VIS [10] dataset. The methods compared include SAM 2.1 [8], SAM [6] + XMem++ [1], SAM [6] + Cutie [4], TinySAM [9] + XMem++ [1], TinySAM [9] + Cutie [4], TinySAM [9] + TA [8], and TinySAM [9] + SA [2] + TA [8]. Our method, TinySAM + **SAM-12V**, demonstrates robust tracking and fine-grained segmentation on video frames, offering competitive performance across challenging scenarios.

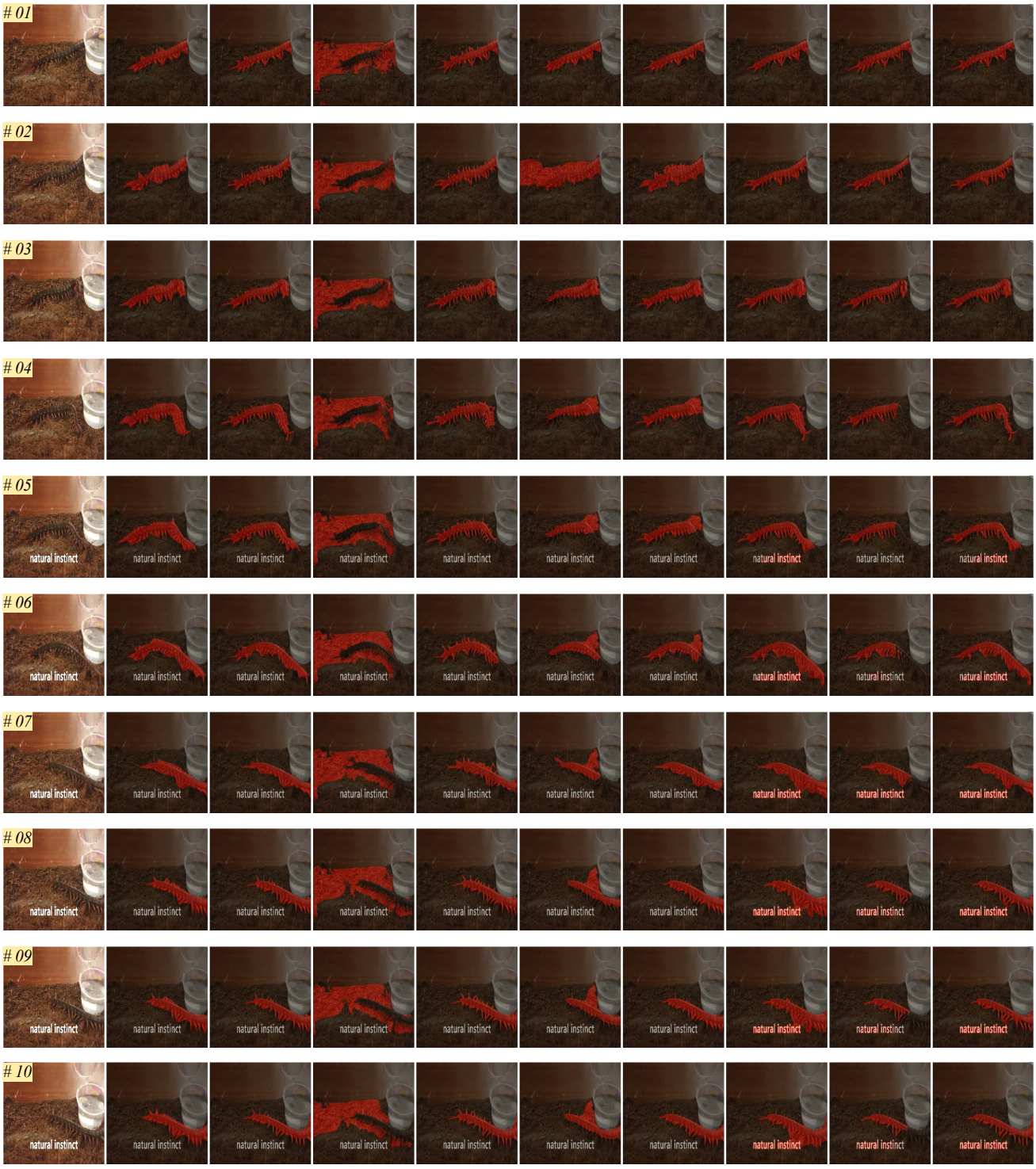

Figure 8. Visual comparison with state-of-the-art methods on a challenging video sequence from LV-VIS [10] dataset. The methods compared include SAM 2.1 [8], SAM [6] + XMem++ [1], SAM [6] + Cutie [4], TinySAM [9] + XMem++ [1], TinySAM [9] + Cutie [4], TinySAM [9] + TA [8], and TinySAM [9] + SA [2] + TA [8]. Our method, TinySAM + **SAM-12V**, demonstrates robust tracking and fine-grained segmentation on video frames, offering competitive performance across challenging scenarios.