LOMM: Latest Object Memory Management for Temporally Consistent Video Instance Segmentation (Supplementary Material)

Seunghun Lee[1,*] Jiwan Seo[1] Minwoo Choi[1] Kiljoon Han[1]
Jahoon Jeong[1] Zane Durante[2] Ehsan Adeli[2,†] Sang Hyun Park[1] Sunghoon Im[1,†]

[1]DGIST, Daegu, Republic of Korea [2]Stanford University, Stanford, CA, USA

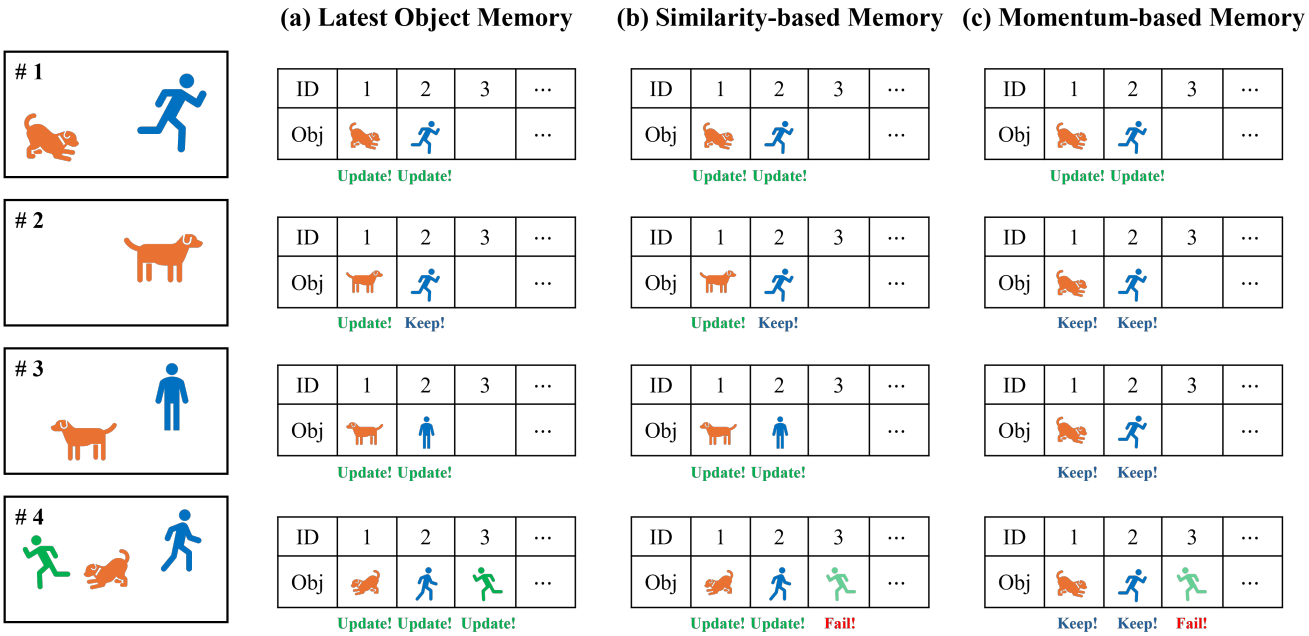

Figure 1. Comparison of three different memory mechanisms

1. Motivation for Memory Design

In this section, we outline the motivation behind the design of LOMM, comparing three different memory systems: (1) Latest Object Memory (LOM), (2) Similarity-based memory, and (3) Momentum-based memory. Fig. 1 and 2 illustrate how each memory mechanism operates.

Latest-state-aware object memory uses the foreground probability of objects as a weight to update the object information at each frame. This means that the most recent and valid information is updated, allowing accurate memory updates even in scenarios where new objects appear, as seen in Fig. 2-(a) at frame #4. Additionally, when objects disappear, the foreground probability is low, so the previous information is largely preserved, maintaining the object information well even in situations like frame #2. This novel mechanism ensures consistent memory updates, both for newly emerging and existing objects, offering robust performance in dynamic scenarios.

Similarity-based memory [16, 19] updates the memory by assessing the similarity between the memory from time $t-1$ and the objects at time t. When the object information at time t shows high similarity to the memory objects, it receives a significant weight during the update; however, if the similarity is low, the update is minimal. This behavior is evident in Fig. 2-(b), where at frame #2, the memory retains information about the person who has disappeared, while the latest information for the dog is updated. As a result, we

* This work was done while visiting Stanford University.
† Corresponding author.

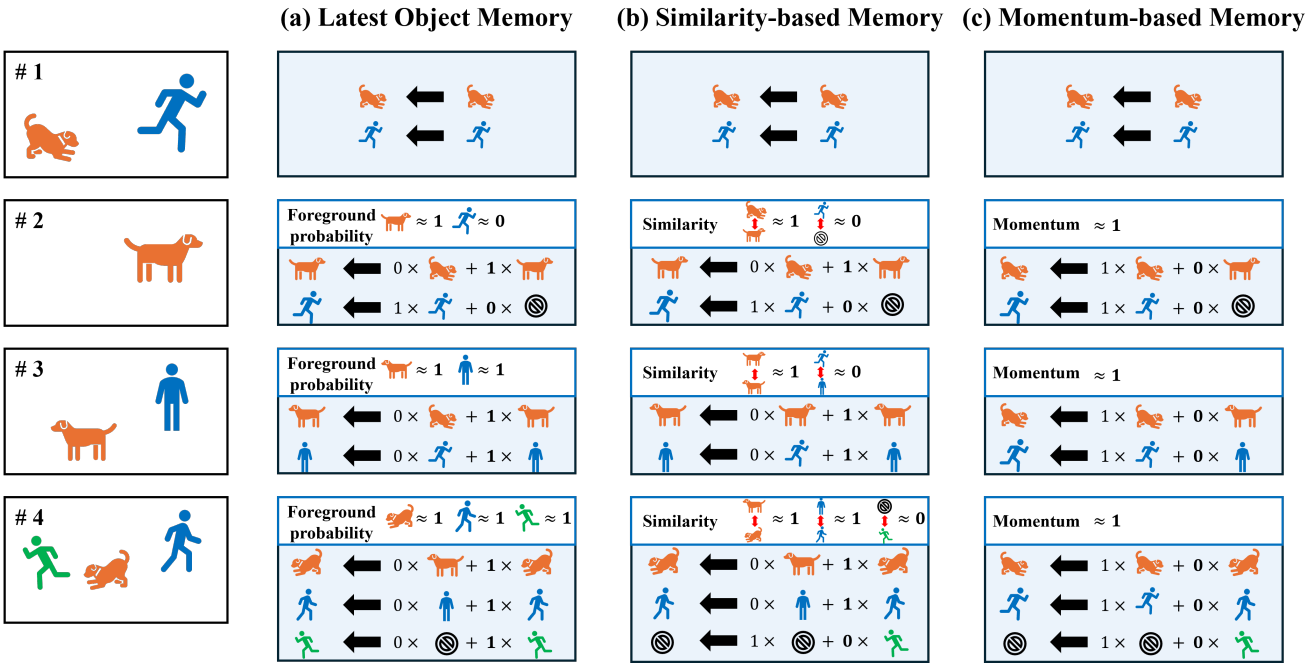

Figure 2. Memory update mechanisms of three different memory systems.

observe a performance improvement over the baseline in Tab. 3**-(b). However, at frame #4, when a new person appears, the memory structure struggles to update effectively due to the low similarity to the existing memory.

Momentum-based memory [4] assigns a high weight (*e.g.*, 99%) to the memory information from time $t-1$ and a low weight (*e.g.*, 1%) to the objects at time t during the update. As a result, the object information from the first frame is retained at a high ratio, while new objects or the latest information are hardly updated.

2. Training Details

2.1. Datasets.

We evaluate the performance of our LOMM using standard benchmark datasets: YouTubeVIS datasets (2019, 2021, 2022) [17] and OVIS [14], as detailed below. Introduced by [17] alongside the pioneering study on the Video Instance Segmentation (VIS) task, the YouTube-VIS datasets consist of high-resolution YouTube videos across 40 categories. The 2019 release includes 2,238 videos for training, 302 for validation, and 343 for testing. In its 2021 update [18], the dataset was expanded to include 2,985 training videos, 421 validation videos, and 453 test videos, allowing for more extensive testing and development of VIS models. The 2022 version includes an additional 71 long videos in the validation set, while the training set remained the same as in the 2021 version. OVIS dataset [14] presents significant challenges with videos that often feature occlusions and long sequences that mirror complex real-world scenarios. This dataset is particularly demanding, with a greater number of objects and frames compared to YouTube-VIS, enhancing the difficulty of segmentation and tracking tasks. OVIS comprises 607 training videos, 140 validation videos, and 154 test videos, providing a robust platform for evaluating the effectiveness of VIS approaches under challenging conditions.

2.2. Implementation Details.

For our segmentation network, we employ the Mask2Former architecture [2] equipped with three distinct backbone encoders: ResNet-50 [5], ViT-L and ViT-H [3]. All backbones are initialized with parameters pre-trained on COCO [12]. To improve memory efficiency with the ViT-L and ViT-H, we incorporate a memory-optimized VIT-Adapter [1], aligning with recent advancements in network efficiency [21]. The segmentation network is further enhanced through pretraining with a contrastive learning approach for better object representation [9, 16, 19, 21]. Our tracking framework integrates two networks $\mathcal{T}_E$ and $\mathcal{T}_A$, each comprising three transformer blocks and enhanced with a referring cross-attention layer [20] for improved accuracy.

For training, our tracking networks are trained with all other parameters frozen as previous studies [10, 20]. We employ the AdamW optimizer [13], initializing with a learning rate of 1e-4 and a weight decay of 5e-2. Training is conducted over 160k iterations, with learning rate reductions

**Green number indicates table in the main paper.

Figure 3. **Limitation.** Our method cannot handle cases where the segmentation network fails to detect objects. However, for detected objects, it demonstrates consistent tracking performance.

scheduled at the 112k mark. We process five frames from each video in a batch of eight during training, adjusting the frame sizes to maintain a shorter side between 320 and 640 pixels, and ensuring the longer side does not exceed 768 pixels. In all experimental settings, we incorporate COCO joint training, as utilized in prior works [6, 7, 15, 19, 20]. For inference, the shorter side of input frames is scaled to 480 pixels, maintaining uniform aspect ratios. We adopt a temporal refiner [20] for our offline model. We empirically set λ_{sim} as 1.0. In the online experiments using the R50 and ViT-L backbones, eight RTX2080 Ti GPUs are employed. For the offline experiments, eight RTX3090 Ti GPUs are used, while the experiments utilizing the ViT-H backbone are conducted with eight RTX A6000 GPUs.

Segmentation network. To achieve distinctive object representation, we employ the following contrastive loss for pretraining the segmentation network $\mathcal{S}$:

$$\begin{aligned}\mathcal{L}_{\text{embed}} &= -\log \frac{\exp\left(\mathbf{v}\cdot\mathbf{k}^{+}\right)}{\exp\left(\mathbf{v}\cdot\mathbf{k}^{+}\right)+\sum_{\mathbf{k}^{-}}\exp\left(\mathbf{v}\cdot\mathbf{k}^{-}\right)} \\ &= \log\left[1+\sum_{\mathbf{k}^{-}}\exp\left(\mathbf{v}\cdot\mathbf{k}^{-}-\mathbf{v}\cdot\mathbf{k}^{+}\right)\right],\end{aligned} \tag{1}$$

where $\mathbf{k}^{+}$, and $\mathbf{k}^{-}$ denote positive embedding and negative embedding from anchor embedding $\mathbf{v}$. This contrastive loss is widely applied in the VIS field [9, 11, 16, 19, 21], learning frame-to-frame associations to create better object representations.

Early training. The initial outputs from the tracking networks $\mathcal{T}_E$ and $\mathcal{T}_A$ are also typically noisy. To address this, we utilize the predictions $\hat{y}$ from $\tilde{Q}_t^*$ for ground truth assignment, formulated as:

$$\hat{\sigma} = \underset{\sigma\in\mathfrak{S}_N}{\arg\min}\sum_{n=1}^{N_{GT}}\mathcal{L}_{\text{Match}}\left(y^{n}_{f(n)}, \hat{y}^{\sigma(n)}_{f(n)}\right). \tag{2}$$

The prediction $\hat{y}$ provides guidance for rapid convergence in the same format as the tracked output of MinVIS [8].

2.3. Additional Qualitative Results

We provide additional comparisons with state-of-the-art models, as shown in Fig. 4, to highlight the robustness of our model in challenging scenarios where objects frequently appear and disappear. Existing methods, including CTVIS [19], DVIS-DAQ [22], and DVIS++ [21], often fail to track accurately by either misidentifying reappearing objects as new or confusing newly appeared objects with existing ones. By utilizing a robust memory mechanism and an effective object association strategy, our model maintains consistently discriminative embeddings, significantly enhancing both segmentation and tracking performance.

2.4. Limitation

Our method adopts a decoupled framework, where the segmentation network is frozen while training the tracking network. As a result, it cannot handle objects that the segmentation network fails to detect. Nevertheless, our approach achieves significant improvements in long-term consistent tracking, a fundamental challenge in VIS. As shown in Fig. 3, even when the segmentation network misses certain objects, our method maintains robust tracking.

References

[1] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022. 2

[2] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 2

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2

[4] Ruopeng Gao and Limin Wang. Memotr: Long-term memory-augmented transformer for multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9901–9910, 2023. 2

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceed-*

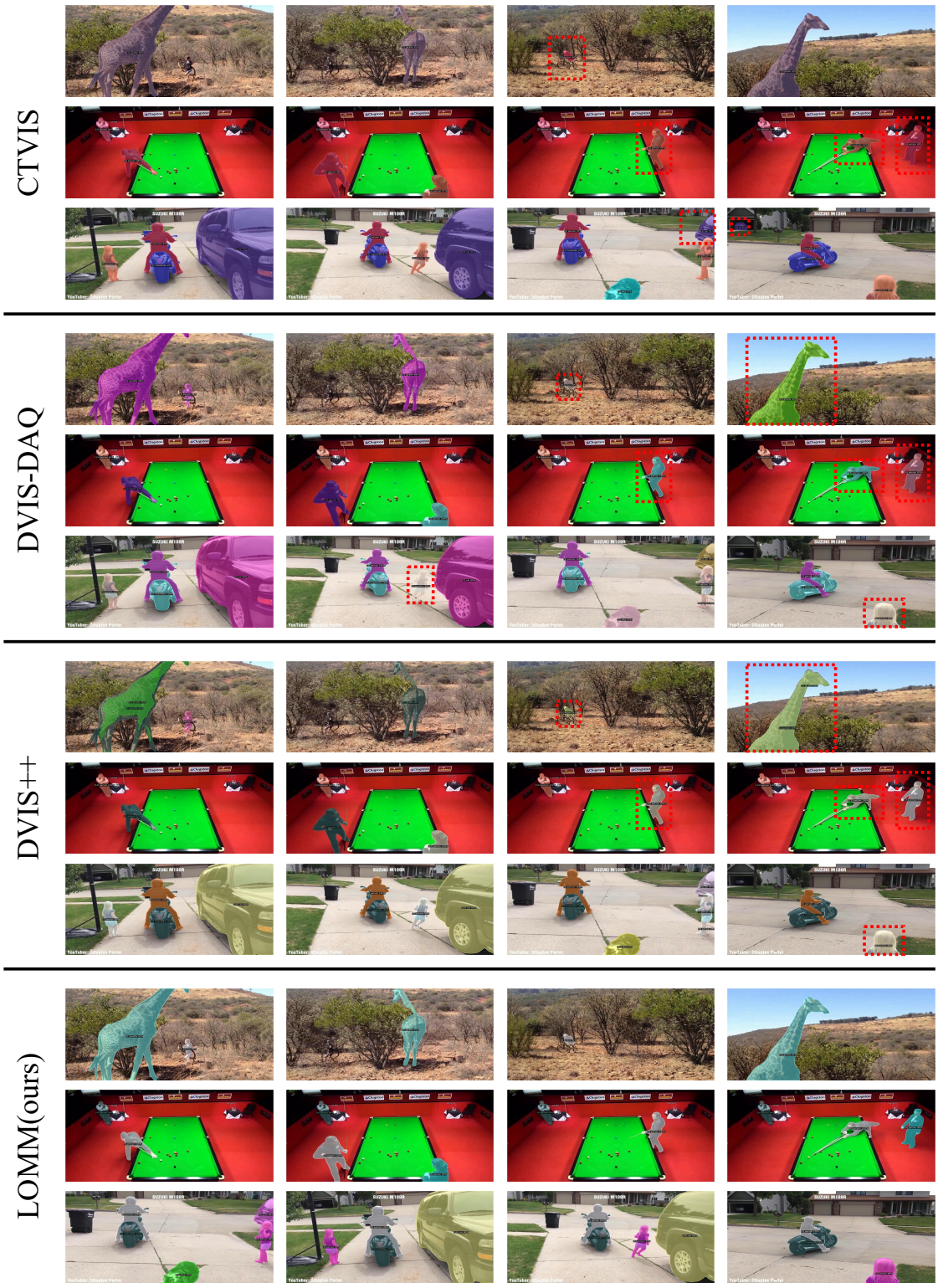

Figure 4. Qualitative comparison of LOMM with CTVIS, DVIS-DAQ, and DVIS++ on challenging scenarios in YTVIS22 dataset.

ings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. 2

[6] Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. Vita: Video instance segmentation

via object token association. *Advances in Neural Information Processing Systems*, 35:23109–23120, 2022. 3

[7] Miran Heo, Sukjun Hwang, Jeongseok Hyun, Hanjung Kim, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. A generalized framework for video instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14623–14632, 2023. 3

[8] De-An Huang, Zhiding Yu, and Anima Anandkumar. Minvis: A minimal video instance segmentation framework without video-based training. *Advances in Neural Information Processing Systems*, 35:31265–31277, 2022. 3

[9] Seunghun Lee, Jiwan Seo, Kiljoon Han, Minwoo Choi, and Sunghoon Im. Context-aware video instance segmentation. *arXiv preprint arXiv:2407.03010*, 2024. 2, 3

[10] Junlong Li, Bingyao Yu, Yongming Rao, Jie Zhou, and Jiwen Lu. Tcovis: Temporally consistent online video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1097–1107, 2023. 2

[11] Minghan Li, Shuai Li, Wangmeng Xiang, and Lei Zhang. Mdqe: Mining discriminative query embeddings to segment occluded instances on challenging videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10524–10533, 2023. 3

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2

[13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2

[14] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *International Journal of Computer Vision*, 130(8): 2022–2039, 2022. 2

[15] Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xiang Bai. Seqformer: Sequential transformer for video instance segmentation. In *European Conference on Computer Vision*, pages 553–569. Springer, 2022. 3

[16] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *European Conference on Computer Vision*, pages 588–605. Springer, 2022. 1, 2, 3

[17] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5188–5197, 2019. 2

[18] Linjie Yang, Yuchen Fan, Yang Fu, and Ning Xu. The 3rd large-scale video object segmentation challenge - video instance segmentation track, 2021. 2

[19] Kaining Ying, Qing Zhong, Weian Mao, Zhenhua Wang, Hao Chen, Lin Yuanbo Wu, Yifan Liu, Chengxiang Fan, Yunzhi Zhuge, and Chunhua Shen. Ctvis: Consistent training for online video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 899–908, 2023. 1, 2, 3

[20] Tao Zhang, Xingye Tian, Yu Wu, Shunping Ji, Xuebo Wang, Yuan Zhang, and Pengfei Wan. Dvis: Decoupled video instance segmentation framework. *arXiv preprint arXiv:2306.03413*, 2023. 2, 3

[21] Tao Zhang, Xingye Tian, Yikang Zhou, Shunping Ji, Xuebo Wang, Xin Tao, Yuan Zhang, Pengfei Wan, Zhongyuan Wang, and Yu Wu. Dvis++: Improved decoupled framework for universal video segmentation. *arXiv preprint arXiv:2312.13305*, 2023. 2, 3

[22] Yikang Zhou, Tao Zhang, Shunping Ji, Shuicheng Yan, and Xiangtai Li. Dvis-daq: Improving video segmentation via dynamic anchor queries. *arXiv preprint arXiv:2404.00086*, 2024. 3