Video Instance Segmentation via Bidirectional Mask Propagation with SAM2

Yinyin Hu*
East China Normal University
71275902008@stu.ecnu.edu.cn

Kaining Ying, Henghui Ding
Fudan University
knying24@m.fudan.edu.cn

Abstract

Online Video Instance Segmentation (VIS) methods typically rely on feature-based matching to associate objects across frames, which often struggle with complex scenarios such as occlusions, rapid motion, and appearance changes. In this paper, we propose an approach SAM2BMP that leverages mask propagation from SAM2 for robust cross-frame object association. Unlike traditional feature matching, our method directly utilizes spatial and shape information through bidirectional mask propagation, providing more intuitive and reliable object tracking. Specifically, we introduce a forward propagation phase that uses SAM2 to predict object locations in subsequent frames and matches them with detected instances via mask IoU, and a backward propagation phase that completes object trajectories by filling in early frames where objects may not be prominently detected. Extensive experiments on multiple VIS datasets demonstrate that our approach achieves superior performance compared to state-of-the-art methods across different backbone architectures.

Keywords: ***Video Instance Segmentation, Object Tracking, SAM2, Bidirectional Mask Propagation***

1. Introduction

Video Instance Segmentation (VIS) [51, 39, 58] is a challenging task that requires simultaneously detecting, segmenting, and tracking object instances across video frames. This task is fundamental to a wide range of video understanding applications, such as autonomous driving [41], video surveillance systems [42, 45, 36], augmented reality experiences, and video editing workflows [35]. Existing VIS approaches can be broadly categorized into two paradigms: offline and online methods. Offline methods [6, 19, 46, 24] process entire video sequences holistically, enabling them to model long-term temporal dependencies and jointly optimize segmentation, classification, and temporal association across all frames. In contrast, online methods [59, 58, 23, 30, 18] adopt a sequential processing strategy, performing detection and segmentation on individual frames and then associating instances across frames through temporal matching. While offline methods can leverage global temporal context, online methods offer distinct advantages: they can handle arbitrarily long videos without memory constraints, do not require frame positional priors, and maintain lower computational overhead. These benefits have led to online methods becoming increasingly dominant in recent VIS research. This paper focuses on advancing the online video instance segmentation paradigm.

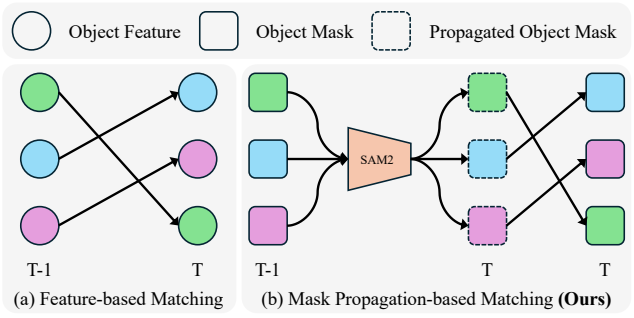

Figure 1. **Comparison of object association strategies in online VIS methods.** The same color represents the same object across different frames. (a) **Feature-based matching** (e.g., CTVIS [58]): Objects are associated by computing similarity between high-dimensional feature embeddings across frames. (b) **Our mask propagation-based approach**: We leverage SAM2 to propagate masks from previous frame (T-1) to current frame (T), then match propagated masks with detected instances via Mask IoU, providing more direct and robust spatial correspondence.

The key challenge in online VIS lies in accurately associating object instances across frames. Current state-of-the-art online methods typically adopt a two-stage pipeline: first performing frame-level instance segmentation, then associating instances across frames through a matching module [58, 47, 59]. Most existing approaches [58, 23, 47] rely on *feature-based matching* (as shown in Figure 1(a)), where each instance is represented as a high-dimensional feature vector, and cross-frame associations are established by computing similarity scores (*e.g.*, cosine similar-

*Corresponding author

ity) between these object feature embeddings. While this paradigm works well in simple scenarios with smooth and static object motion, it faces significant challenges in complex real-world conditions. Specifically, when objects undergo rapid motion, severe occlusions, drastic appearance changes, or temporary disappearances, the discriminative capability of learned object feature embeddings often degrades, leading to incorrect associations, ID switches, or missed detections.

Recent advances in foundation models, particularly the Segment Anything Model 2 (SAM2) [40], have demonstrated remarkable capabilities in video object segmentation (VOS) [38, 9, 11, 20, 21, 49, 55, 54, 10, 32, 22]. SAM2 introduces memory-based mechanisms for temporal consistency and excels at propagating object masks across frames given initial annotations, enabling accurate tracking even under occlusions and appearance changes. This raises an intriguing question: *Can we leverage SAM2's powerful mask propagation capabilities to improve cross-frame object association in video instance segmentation?* However, directly applying SAM2 to VIS is non-trivial. Unlike VOS where objects are provided at initialization, VIS requires *automatic detection* of all object instances throughout the video, including those that appear mid-sequence. SAM2 alone cannot handle this dynamic instance discovery problem, as it requires explicit initialization for each object to track.

To address this challenge, we propose a new framework that synergistically combines image instance segmentation with SAM2's mask propagation for robust VIS, as illustrated in Figure 1(b). Our key insight is to use an image segmentation model (*e.g.*, Mask2Former [7]) to detect potential objects in each frame, and leverage SAM2 to propagate existing tracked instances for cross-frame association. Specifically, in the *forward propagation* phase, we propagate masks of tracked objects to the next frame using SAM2, then match these propagated masks with newly detected instances via Mask Intersection over Union (IoU). This spatial matching based on mask overlap is more direct and robust than abstract feature similarity, enabling accurate association even when objects undergo significant motion or appearance changes. For newly detected instances that do not match any propagated masks, we initialize them as new tracks.

Furthermore, we observe that objects may exist in early frames but remain undetected until they become more prominent later. To address this trajectory incompleteness issue, we introduce a *backward propagation* phase: for each tracked object, we propagate its mask backward in time from its first detection frame, filling in missing detections in earlier frames. This bidirectional design ensures complete trajectories and significantly improves segmentation accuracy, particularly for objects that are initially occluded, small, or low-contrast.

Our main contributions are summarized as follows:

- We propose SAM2BMP, a new online VIS framework that leverages SAM2's mask propagation capability for explicit spatial correspondence, replacing implicit feature-based matching with direct mask-level associations via Mask IoU. This design provides more accurate spatial localization and interpretable temporal associations.
- We introduce a bidirectional mask propagation mechanism combining forward propagation for tracking and backward propagation for trajectory completion. The backward propagation component significantly improves trajectory completeness by recovering objects in frames where they are initially missed.
- We achieve state-of-the-art results on YouTube-VIS 2019/2021 and OVIS datasets.

2. Related Work

2.1. Video Instance Segmentation

Offline VIS Methods. Offline approaches [6, 19, 46, 24, 44, 53, 26, 29, 14, 8] process complete video sequences at once, generating instance masks for all frames simultaneously while exploiting bidirectional temporal context, though at the cost of substantial computational demands. Methods such as Mask2Former-VIS [6] and SeqFormer [46] utilize attention-based architectures to capture spatio-temporal representations and produce sequential instance mask predictions. To address memory constraints when handling exceptionally long video sequences, VITA [19] introduces a strategy that decodes video-level object queries from compact frame-wise object tokens rather than relying on dense spatio-temporal feature representations.

Online VIS Methods. Online VIS approaches [51, 52, 23, 47, 58, 59, 2, 48, 13, 1, 27] are generally constructed on top of image-level instance segmentation frameworks [16, 7, 57]. MaskTrack R-CNN [51] augments Mask R-CNN [16] with an extra tracking module that links instances across frames through heuristic matching strategies. Following the introduction of query-based segmentation architectures [7], leveraging query embeddings for matching rather than manually crafted rules has significantly improved online VIS performance [47, 23]. MinVIS [23] performs instance tracking via frame-by-frame Hungarian matching of queries without requiring video-level training, exploiting the temporal coherence of per-frame instance queries from the image segmentor [7]. IDOL [47] introduces a memory mechanism that stores momentum-averaged embeddings of instances from preceding frames, which are then matched against newly detected foreground embeddings. DVIS [59]

adopts a decoupled design that separates the detection and tracking components. CTVIS [58] leverages contrastive learning [3, 15, 4] to obtain more discriminative embeddings during training, and applies cosine similarity-based matching between object features during inference. Despite their effectiveness in various settings, these embedding-based methods face challenges when dealing with rapid motion, heavy occlusions, and significant appearance variations. Recent approaches [18, 54, 30] explore query propagation for tracking, yet they fundamentally remain matching-based techniques. Our method differs by directly utilizing SAM2's mask propagation mechanism to establish more explicit spatial correspondences.

2.2. Segment Anything Model

The Segment Anything Model (SAM) [28] pioneered foundation models for image segmentation with zero-shot capabilities, while SAM2 [40] extended it to videos through streaming memory for efficient mask propagation in video object segmentation (VOS). Following SAM2's release, several training-free adaptations [50, 12, 43] have emerged for video understanding tasks, but they inherit a fundamental limitation: requiring explicit object initialization, making them suitable only for VOS with predefined objects rather than VIS where new objects must be automatically discovered throughout the video. SAM2MOT [25] introduces a memory mechanism to handle newly appearing objects by combining detectors with SAM2's propagation, but it focuses solely on multi-object tracking with bounding boxes rather than dense segmentation required for VIS. Our work addresses this gap by integrating image instance segmentation with SAM2's mask propagation, enabling automatic object discovery and bidirectional trajectory completion for comprehensive video instance segmentation.

3. Method

In this section, we present the details of our proposed video instance segmentation algorithm SAM2BMP. We first review the framework of online video instance segmentation algorithms and SAM2 in Section 3.1, and then elaborate on our proposed SAM2BMP in Section 3.2. To improve the readability of our method, we provide a concrete example analysis in Section 3.3. Finally, we introduce several practical improvements to our SAM2BMP to further enhance its performance in Section 3.4.

3.1. Preliminaries

Problem Formulation. Given an input video $V = \{I_t\}_{t=1}^{T}$ where $I_t \in \mathbb{R}^{H\times W\times 3}$ denotes the t-th frame, the goal of video instance segmentation is to produce a set of trajectories $\mathcal{T} = \{\tau^j\}_{j=1}^{M}$, where each trajectory $\tau^j = \{(m_t^j, c^j)\}_{t=1}^{T}$ represents a tracked object instance with its category label c^j and a sequence of binary masks $\{m_t^j \in \{0,1\}^{H\times W}\}_{t=1}^{T}$ indicating the object's spatial location in each frame. Note that m_t^j can be an empty set if the object is not visible in frame t.

Online VIS Pipeline. Most online VIS methods [58, 47, 59] adopt a two-stage pipeline that decouples instance segmentation and temporal association. First, an image instance segmentation model $\mathcal{F}_{\text{seg}}$ (*e.g.*, Mask2Former [7]) processes each frame I_t independently to produce per-frame detections:

$$S_t = \mathcal{F}_{\text{seg}}(I_t) = \{(m_t^i, c_t^i, s_t^i)\}_{i=1}^{N_t}, \tag{1}$$

where m_t^i, c_t^i, and s_t^i represent the mask, category, and confidence score of the i-th detected instance, respectively, and N_t is the number of detections in frame t. Then, a matching module $\mathcal{F}_{\text{track}}$ associates instances across consecutive frames to form trajectories. Specifically, given tracked instances $\mathcal{T}_{t-1}$ from previous frames and new detections S_t in the current frame, the matching module computes associations to update trajectories:

$$\mathcal{T}_t = \mathcal{F}_{\text{track}}(\mathcal{T}_{t-1}, S_t). \tag{2}$$

The key challenge is designing $\mathcal{F}_{\text{track}}$ to robustly handle occlusions, appearance changes, and object disappearance-reappearance.

SAM2 for Mask Propagation. SAM2 [40] extends the Segment Anything Model [28] to videos through a streaming memory architecture that enables efficient mask propagation across frames. Given an initial mask m_0 for an object at frame t_0, SAM2 maintains a memory bank $\mathcal{M}$ that accumulates spatial and temporal features from processed frames. For propagating to subsequent frames $t > t_0$, SAM2 uses memory attention to retrieve relevant context from $\mathcal{M}$ and predict the mask at the current frame:

$$m_t = \text{SAM2}(I_t, \mathcal{M}), \tag{3}$$

where the newly predicted mask m_t is then added back to $\mathcal{M}$ to enrich the memory for future propagation. This memory-based mechanism allows SAM2 to maintain object identity and track objects robustly through occlusions, appearance changes, and motion variations. The key advantage of SAM2's propagation is that it directly predicts segmentation masks rather than abstract features, providing explicit spatial information about object locations. In our framework, we leverage this capability to establish correspondences between frames, using SAM2's propagated masks as spatial priors for object association.

3.2. SAM2BMP Algorithm

Overview. Our SAM2BMP framework processes the input video $V = \{I_t\}_{t=1}^{T}$ through two sequential phases: *forward mask propagation and matching* followed by *backward mask propagation completion*. We maintain a memory

bank $\mathcal{M}$ to store tracked instances with their masks, categories, and trajectory information. In the forward phase, for each frame t, we obtain detections S_t using $\mathcal{F}_{\text{seg}}$, propagate existing tracked masks from $\mathcal{M}$ to the current frame using SAM2, and match propagated masks with detections via Mask IoU to update trajectories. In the backward phase, we propagate masks backward from each object's first detection frame to complete missing detections in earlier frames.

Forward Mask Propagation and Matching. The core of our approach is to leverage SAM2's mask propagation for establishing spatial correspondences between consecutive frames. For frame t, suppose we have K tracked objects stored in memory bank $\mathcal{M}$, each represented by its mask m_t^k where $k \in \{1, \ldots, K\}$. We use SAM2 to propagate each tracked mask to the next frame $t+1$:

$$\hat{m}_{t+1}^k = \text{SAM2}(I_{t+1}, \mathcal{M}^k), \tag{4}$$

where $\mathcal{M}^k$ denotes the memory context for object k stored in $\mathcal{M}$. These propagated masks $\{\hat{m}_{t+1}^k\}_{k=1}^K$ represent SAM2's predictions of where tracked objects should appear in frame $t+1$.

Meanwhile, we obtain per-frame detections $S_{t+1} = \{(m_{t+1}^i, c_{t+1}^i, s_{t+1}^i)\}_{i=1}^{N_{t+1}}$ from $\mathcal{F}_{\text{seg}}(I_{t+1})$ and filter high-confidence instances with $s_{t+1}^i > \theta_{\text{conf}}$ to get S_{t+1}^{high}. We then compute an IoU matching matrix $\mathbf{A} \in \mathbb{R}^{K \times N_{t+1}}$ where:

$$\mathbf{A}_{ki} = \text{IoU}(\hat{m}_{t+1}^k, m_{t+1}^i) = \frac{|\hat{m}_{t+1}^k \cap m_{t+1}^i|}{|\hat{m}_{t+1}^k \cup m_{t+1}^i|}. \tag{5}$$

Using this matrix, we perform bipartite matching to associate propagated masks with detections. For each tracked object k, we find the best matching detection $i^* = \arg\max_i \mathbf{A}_{ki}$ if $\mathbf{A}_{ki^*} > 0.5$. If a match is found, we update the trajectory using the detected mask $m_{t+1}^{i^*}$; otherwise, we use the propagated mask $\hat{m}_{t+1}^k$ to maintain trajectory continuity. Unmatched detections in S_{t+1}^{high} are initialized as new trajectories in $\mathcal{M}$.

This mask propagation-based matching offers several advantages over feature matching: **(1)** it directly leverages spatial and shape information, making associations more interpretable; **(2)** SAM2's strong tracking capability ensures accurate mask predictions even under appearance changes; **(3)** using propagated masks maintains trajectory continuity when detections fail in certain frames.

Backward Mask Propagation Completion. After forward processing, we observe that objects may exist in early frames but remain undetected until they become more prominent. For instance, small or partially occluded objects may not produce high-confidence detections initially, resulting in incomplete trajectories. To address this, we introduce a backward propagation phase to complete trajectories before their first detection.

For each tracked object j in the final memory bank $\mathcal{M}$, let t_{first}^j denote its first detection frame with mask $m_{t_{\text{first}}}^j$. We reinitialize SAM2 with this mask as the prompt and propagate backward in time:

$$\tilde{m}_t^j = \text{SAM2}(I_t, \mathcal{M}^j), \quad \text{for } t = t_{\text{first}}^j - 1, \ldots, 1, \tag{6}$$

where masks are propagated in reverse temporal order and $\mathcal{M}^j$ represents the memory context for object j. This backward propagation fills in missing detections before t_{first}^j, significantly improving trajectory completeness. The technique is particularly effective for objects that are initially small, occluded, or have low contrast, as well as for recovering from temporary detection failures.

3.3. Visualization Case Study

Figure 2 illustrates our complete pipeline through a concrete example with both forward and backward propagation. We use Mask2Former to denote $\mathcal{F}_{\text{seg}}$ here.

Forward Propagation (Frames 1-4). Frame 1: $\mathcal{F}_{\text{seg}}$ detects the dog (solid mask). The partially visible soccer ball is missed. We add the dog to $\mathcal{M}$. **Frame 2:** SAM2 propagates the dog's mask from frame 1 (dashed outline). It matches well with the detected dog mask via IoU. The cat appears and is added to $\mathcal{M}$. **Frame 3:** The dog disappears. SAM2 propagates both objects but only the cat is matched. The dog gets an empty mask. The soccer ball becomes visible and is added to $\mathcal{M}$. **Frame 4:** All three objects are detected. SAM2 propagates all masks. IoU matching successfully links them with detections. The dog reappears and is matched correctly.

Backward Propagation Completion. After forward processing, we have first detection frames: dog ($t = 1$), cat ($t = 2$), soccer ball ($t = 3$). We apply backward propagation to the cat and soccer ball. The cat remains absent in frame 1 as expected. SAM2 successfully recovers the soccer ball's mask in frames 1-2. This completes all trajectories, showing how bidirectional propagation handles late detections and temporary disappearances.

3.4. Practical Improvements

While the core bidirectional mask propagation mechanism provides robust tracking, we introduce several practical improvements to further enhance performance and efficiency.

Tracklet NMS. After bidirectional propagation, trajectories may exhibit high spatial-temporal overlap. We apply Tracklet NMS to remove redundant trajectories: for each pair of trajectories with IoU > 0.5 across at least 6 frames, we retain the one with more key frames (frames with high-confidence detections from $\mathcal{F}_{\text{seg}}$) and remove the other. This ensures one-to-one object-trajectory correspondence.

Key Frame Ratio Filter. We observe that unreliable trajectories often have few key frames relative to their length. We

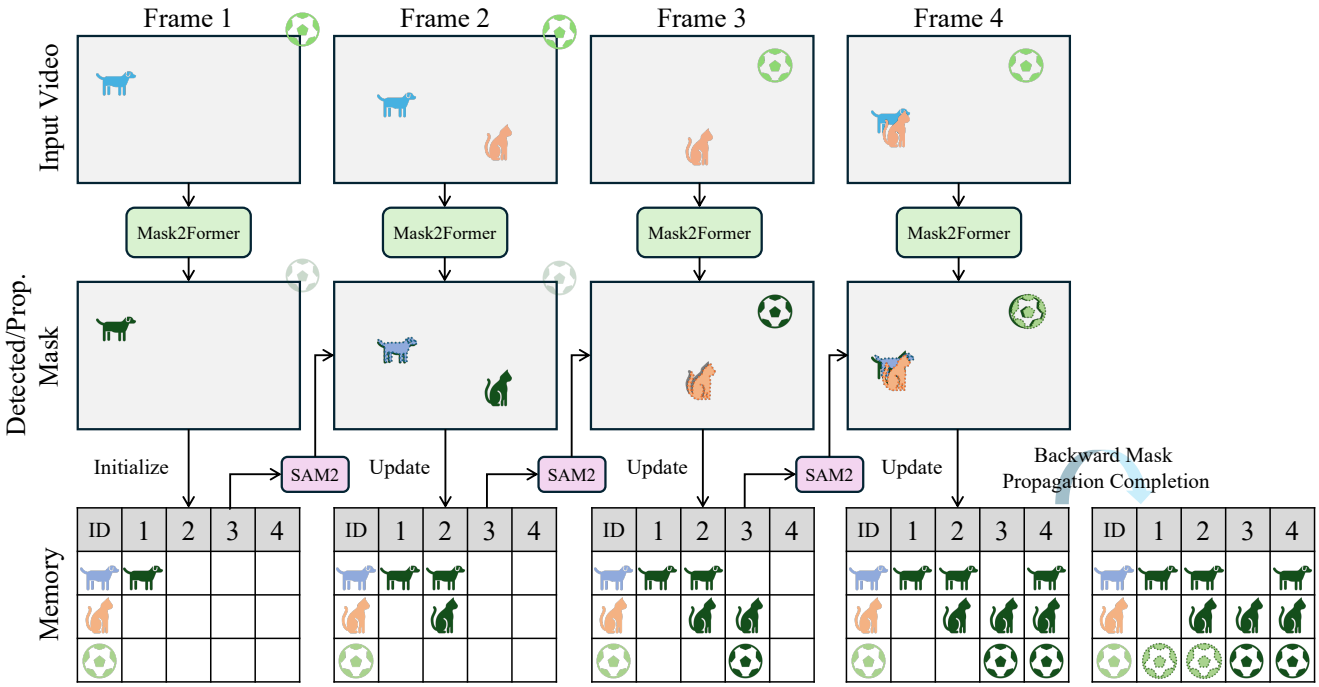

Figure 2. **SAM2BMP algorithm pipeline illustration.** We demonstrate our approach on a video scene with three objects: a dog (blue), a cat (orange), and a soccer ball (green). This scene covers major VIS challenges: **(1)** multiple objects; **(2)** disappearance and reappearance (dog in frame 3-4); **(3)** newly appearing objects (cat in frame 2); **(4)** occlusion (frame 4); **(5)** non-salient objects (soccer ball in frames 1-2). Solid masks indicate detections from Mask2Former, dashed outlines show SAM2 propagated masks, and the memory bank $\mathcal{M}$ tracks object states across frames.

remove trajectories where the ratio of key frames to total frames falls below 0.1. This effectively filters out spurious tracks caused by false propagations while preserving genuine objects that may occasionally lack detections.

Low-Confidence Proposals. Detections with confidence scores in $[\theta_{\text{low}}, \theta_{\text{conf}}]$ are not immediately used for initialization. Instead, we check their spatial overlap with existing tracks. Only non-overlapping low-confidence detections (IoU < 0.3 with all tracked objects) are added as new trajectories. This strategy balances between capturing difficult objects and avoiding false positives from ambiguous detections.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate our proposed method SAM2BMP on three widely-used VIS benchmarks: YouTube-VIS 2019 (YTVIS19) [51], YouTube-VIS 2021 (YTVIS21) [51], and OVIS [39]. YTVIS19 contains 2,238 training videos and 302 validation videos across 40 object categories. YTVIS21 maintains the same 40 categories while expanding the dataset to 2,985 training and 421 validation videos with improved annotation quality. OVIS comprises 607 training and 140 validation videos spanning 25 object categories. Although OVIS has fewer videos, each video is substantially longer, averaging 69.4 frames compared to 27.6 frames in YTVIS19 and 30.2 frames in YTVIS21. Moreover, OVIS samples typically involve more instances and severe occlusion and thus are more challenging.

Implementation Details. Following previous works [59, 61], we adopt Mask2Former [7] as the proposal detector and conduct experimental comparisons using different backbone networks, including ResNet-50 [17], Swin-L [34], and ViT-L [37]. For the tracking model, we adopt SAM2 [40] for mask propagation, experimenting with three different parameter scales: Tiny, Base+ and Large. It is worth noting that we directly use the original pre-trained weights of the SAM2 model without additional fine-tuning on VIS datasets. Following previsou work [58, 59], the Mask2Former model is trained on the target datasets with MS-COCO [31] to adapt to the specific requirements of the video instance segmentation task. Specifically, we train Mask2Former for 60,000 iterations with a batch size of 32. During inference, we set the foreground confidence threshold to 0.95. This threshold choice is based on experimental results on the validation set and can effectively balance detection accuracy and recall rate. Unless otherwise specified, all ablation studies are conducted on the YouTube-VIS 2021 validation set with ResNet-50 as the backbone. All experiments are conducted on 8 NVIDIA A6000 (48GB) GPUs.

Table 1. Performance comparison of SAM2BMP with existing methods on YouTube-VIS 2019 and YouTube-VIS 2021 [51] validation set. The **best** and second best results are highlighted.

Backbone	Method	Publication	YouTube-VIS 2019					YouTube-VIS 2021				
			AP	AP_{50}	AP_{75}	AR_1	AR_{10}	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
ResNet-50 [17]	VITA [19]	[NeurIPS'2022]	49.8	72.6	54.5	49.4	61.0	45.7	67.4	49.5	40.9	53.6
	MinVIS [23]	[NeurIPS'2022]	47.4	49.0	52.1	45.7	55.7	44.2	66.0	48.1	39.2	51.7
	IDOL [47]	[ECCV'2022]	49.5	74.0	52.9	47.7	58.7	43.9	68.0	49.6	38.0	50.9
	DVIS [59]	[ICCV'2023]	51.2	73.8	57.1	47.2	59.3	46.4	68.4	49.6	39.7	53.5
	CTVIS [58]	[ICCV'2023]	55.1	78.2	59.1	51.9	63.2	50.1	73.7	54.7	41.8	59.5
	DVIS-DAQ [61]	[ECCV'2024]	55.2	-	61.9	-	63.7	50.4	-	55.0	-	57.6
	CAVIS [30]	[ICCV'2025]	55.7	78.3	61.7	51.5	63.3	50.5	74.1	54.9	42.6	59.5
	SAM2BMP (ours)	-	**56.4**	**78.7**	**62.4**	**52.0**	**63.9**	**52.7**	**74.5**	**59.5**	**43.7**	**60.4**
Swin-L [34]	VITA [19]	[NeurIPS'2022]	63.0	86.9	67.9	56.3	68.1	57.5	80.6	61.0	47.7	62.2
	MinVIS [23]	[NeurIPS'2022]	61.6	83.3	68.6	54.8	66.6	55.3	76.6	62.0	45.9	60.8
	IDOL [47]	[ECCV'2022]	64.3	87.5	71.0	55.6	69.1	56.1	80.8	63.5	45.0	60.1
	DVIS [59]	[ICCV'2023]	63.9	87.2	70.4	56.2	69.0	58.7	80.4	66.6	46.4	64.6
	CTVIS [58]	[ICCV'2023]	65.6	87.7	72.2	56.5	70.4	61.2	84.0	68.8	48.0	65.8
	DVIS-DAQ [61]	[ECCV'2024]	65.7	-	73.6	-	70.7	61.1	-	68.2	-	66.0
	CAVIS [30]	[ICCV'2025]	66.0	89.5	73.3	56.8	71.4	61.1	84.1	69.2	48.2	66.3
	SAM2BMP (ours)	-	**66.7**	**89.8**	**74.2**	**57.3**	**71.7**	**62.0**	**84.3**	**70.3**	**48.6**	**66.6**
ViT-L [37]	MinVIS [23]	[NeurIPS'2022]	65.6	85.4	72.7	57.5	70.6	59.2	79.9	66.7	47.8	64.1
	DVIS-DAQ [61]	[ECCV'2024]	68.3	-	76.1	-	73.5	62.4	-	70.8	-	68.0
	CAVIS [30]	[ICCV'2025]	68.9	89.3	76.2	58.3	73.6	64.6	85.6	72.5	49.5	69.3
	SAM2BMP (ours)	-	**69.2**	**89.6**	**76.9**	**58.7**	**73.9**	**66.6**	**88.4**	**74.4**	**50.4**	**70.5**

Evaluation Metrics. We adopt standard video instance segmentation evaluation metrics, including AP, AP_{50}, AP_{75}, AR_1, and AR_{10}. Among them, AP is the primary evaluation metric, representing the average across different IoU thresholds (from 0.50 to 0.95 with a step size of 0.05). AP_{50} and AP_{75} represent average precision at IoU thresholds of 0.5 and 0.75, respectively. AR_1 and AR_{10} represent average recall rates when limiting detection to 1 and 10 instances per video, respectively. These metrics comprehensively evaluate the algorithm's overall performance in object detection, segmentation, and tracking.

4.2. Main Results

Table 1 presents comprehensive comparisons on YouTube-VIS 2019 and 2021 validation sets. Our SAM2BMP achieves state-of-the-art performance across all backbone configurations. With ResNet-50, we achieve 56.4% AP on YTVIS19 and 52.7% AP on YTVIS21, surpassing the previous best CAVIS [30] by 0.7 and 2.2 points respectively. The improvements are particularly pronounced in high-precision metrics: our AP_{75} reaches 59.5% on YTVIS21, exceeding CAVIS by 4.6 points, demonstrating superior segmentation quality through SAM2's explicit mask propagation. When equipped with stronger backbones, our method maintains consistent advantages, achieving 66.7%/62.0% AP with Swin-L and 69.2%/66.6% AP with ViT-L on YTVIS19/21. Notably, the consistent gains across different backbones validate that our mask propagation approach is complementary to improved visual representations. The improvements over recent query-propagation methods [30, 58] confirm the effectiveness of explicit spatial correspondence through mask propagation versus implicit feature-based associations. Moreover, our strong AR_{10} scores across all settings indicate that bidirectional propagation successfully maintains trajectory completeness, particularly for objects initially difficult to detect.

Table 2 shows results on the more challenging OVIS dataset, which features longer videos (69.4 frames on average) with heavy occlusions and complex interactions. SAM2BMP demonstrates substantially larger improvements here, achieving 45.4% AP with ResNet-50, 53.0% AP with Swin-L, and 57.1% AP with ViT-L. The gains are particularly significant with ResNet-50: we outperform DVIS-DAQ [61] by 6.7 points and CAVIS [30] by 7.8 points in AP. This substantial margin highlights the robustness of our bidirectional mask propagation in handling challenging long-term tracking scenarios with frequent occlusions. The improvements on AP_{50} are even more remarkable: with

Table 2. Performance comparison of SAM2BMP with existing methods on OVIS [39] validation set.

Backbone	Method	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
ResNet-50	VITA [19]	19.6	41.2	17.4	11.7	26.0
	MinVIS [23]	25.0	45.5	24.0	13.9	29.7
	IDOL [47]	28.2	51.0	28.0	14.5	38.6
	DVIS [59]	30.2	55.0	30.5	14.5	37.3
	CTVIS [58]	35.5	60.8	34.9	16.1	41.9
	DVIS-DAQ [61]	38.7	-	37.6	-	45.2
	CAVIS [30]	37.6	63.4	38.2	16.5	43.5
	SAM2BMP (ours)	**45.4**	**72.2**	**47.6**	**17.3**	**52.3**
Swin-L	VITA [19]	27.7	51.9	24.9	14.9	33.0
	MinVIS [23]	39.4	61.5	41.3	18.1	43.3
	IDOL [47]	40.0	63.1	40.5	17.6	46.4
	DVIS [59]	45.9	71.1	48.3	18.5	51.5
	CTVIS [58]	46.9	71.5	47.5	19.1	52.1
	DVIS-DAQ [61]	49.5	-	51.7	-	54.9
	CAVIS [30]	48.6	74.0	52.5	19.5	53.3
	SAM2BMP (ours)	**53.0**	**79.8**	**57.9**	**19.4**	**58.8**
ViT-L	MinVIS [23]	42.9	65.7	45.4	19.8	46.5
	DVIS-DAQ [61]	53.7	-	58.2	-	59.5
	CAVIS [30]	53.2	75.9	59.1	20.9	58.2
	SAM2BMP (ours)	**57.1**	**81.4**	**63.0**	**21.3**	**62.5**

ResNet-50, we achieve 72.2% versus CAVIS's 63.4% (+8.8 points), indicating superior spatial localization. Across all configurations, our method consistently achieves the best AR_{10} scores, demonstrating strong recall in detecting and tracking multiple instances throughout long videos. The larger improvements on OVIS compared to YouTube-VIS 2021 (6.7 points *vs*. 2.2 points with ResNet-50) suggest that mask propagation-based association is particularly effective for complex scenarios with occlusions and long-term temporal dependencies, where feature-based matching struggles to maintain consistent object identity.

4.3. Ablation Studies

Impact of Tracking Model. For each backbone configuration, we compare three tracking model scales: SAM2-Tiny, SAM2-Base+, and SAM2-Large. As shown in Table 3, using a larger tracking model consistently improves performance, though with slightly higher memory usage and slower inference speed. Specifically, under the ResNet-50 backbone, AP improves by +1.1 (51.3→52.4) when upgrading from SAM2-Tiny to SAM2-Base+, and by an additional +0.3 (52.4→52.7) from SAM2-Base+ to SAM2-Large, with memory increasing from 5.8 to 6.3 GiB. Under the Swin-L backbone, AP improves by +0.5 (60.8→61.3) and +0.7 (61.3→62.0), while memory grows from 7.8 to 8.2 GiB. Under the ViT-L backbone, AP improves by +1.2 (64.0→65.2) and +1.4 (65.2→66.6), with memory usage increasing from 9.8 to 10.3 GiB. These results demonstrate that stronger tracking models enable more accurate object association and improved temporal consistency, particularly in complex video scenes.

Table 3. Ablation study of different backbone and tracking model combinations. Memory (Mem.) usage is reported in GiB.

ID	Mask2Former	SAM2	AP	FPS	Mem.
1	ResNet-50	Tiny	51.3	7.3	5.8
2		Base+	52.4	7.7	6.1
3		Large	52.7	7.1	6.3
4	Swin-L	Tiny	60.8	5.3	7.8
5		Base+	61.3	4.8	8.0
6		Large	62.0	4.4	8.2
7	ViT-L	Tiny	64.0	3.9	9.8
8		Base+	65.2	3.8	10.1
9		Large	66.6	3.5	10.3

Table 4. Ablation study on the threshold θ_{conf} for foreground selection on YouTube-VIS 2021 validation.

θ_{conf}	0.60	0.80	0.93	0.95	0.97
AP	49.9	51.0	51.9	52.7	52.0

Table 5. Ablation results for different configurations.

ID	BMPC	KFRF	LCP	Tracklet NMS	AP
1	✗	✗	✗	✗	43.2
2	✓	✗	✗	✗	48.5
3	✓	✓	✗	✗	51.2
4	✓	✓	✓	✗	52.3
5	✓	✓	✓	✓	52.7

Impact of Backbone Network. As shown in Table 5, fixing the tracking model, stronger backbones consistently yield better segmentation performance, albeit at the cost of inference speed and memory consumption. When using SAM2-Base+, AP increases from 52.4% with ResNet-50 to 61.3% with Swin-L, and further to 65.2% with ViT-L, while FPS decreases from 7.7 to 4.8 to 3.8, and memory usage increases from 6.1 to 8.0 to 10.1 GiB. Similarly, with SAM2-Large, AP increases from 52.7% to 62.0% to 66.6%, with FPS decreasing from 7.1 to 4.4 to 3.5 and memory increasing from 6.3 to 8.2 to 10.3 GiB. The smaller margin of improvement from Swin-L to ViT-L compared with ResNet-50 to Swin-L suggests that the model capacity is approaching saturation under the current dataset scale and complexity.

Speed–Performance Trade-off. As shown in Table 3, inference speed decreases as model complexity increases. With the ResNet-50 backbone, SAM2BMP runs at approximately 7.1–7.7 FPS; with Swin-L, it achieves 4.4–5.3 FPS; and with ViT-L, it drops further to 3.5–3.9 FPS. In com-

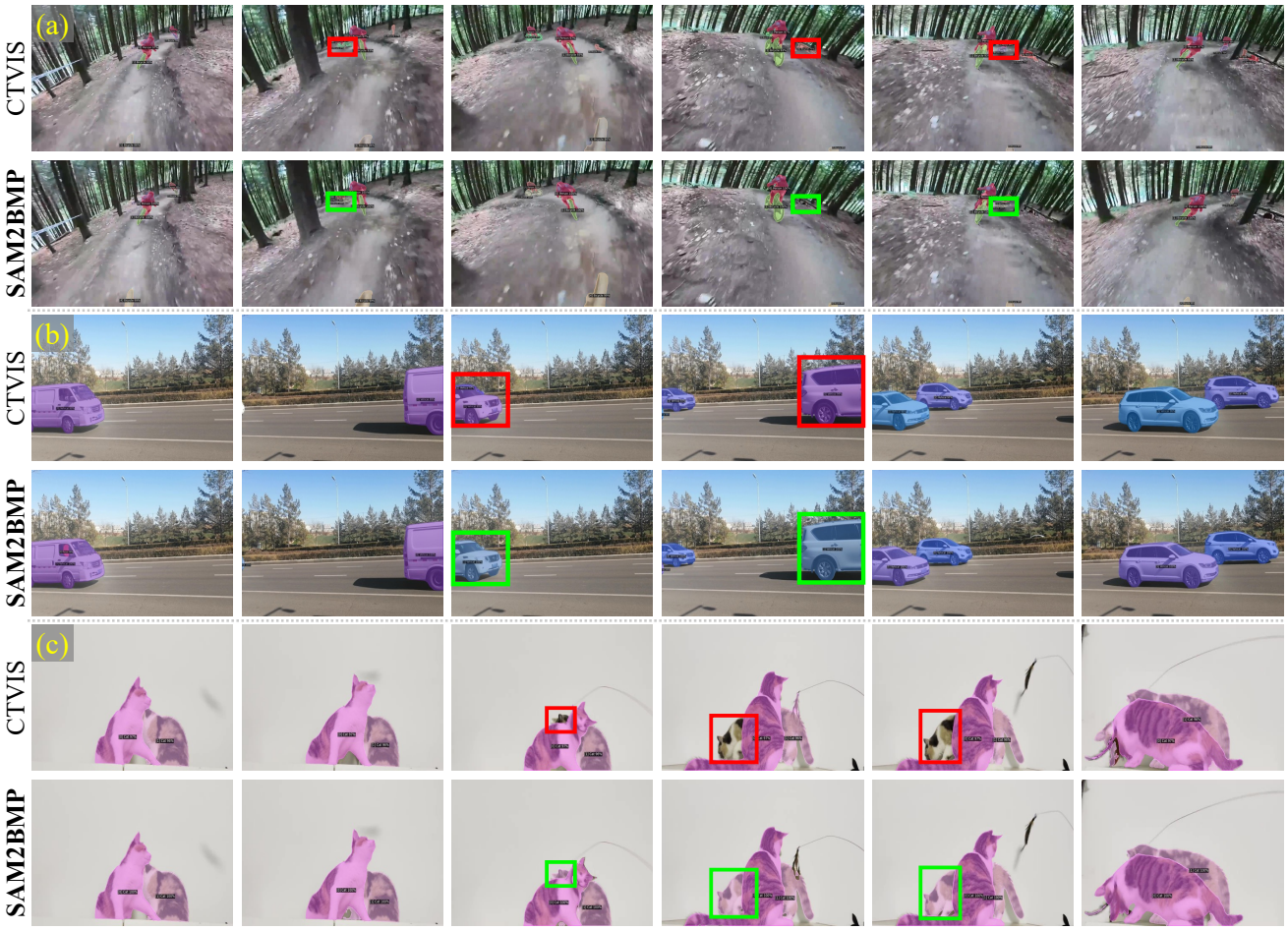

Figure 3. Visualization comparison between CTVIS and SAM2BMP (ours) in complex scenarios.

parison, the current state-of-the-art method CTVIS reports 7.9 FPS when using a ViT-L backbone, showing a clear advantage in speed. This indicates that our method sacrifices some inference efficiency in pursuit of higher segmentation accuracy. However, given the significant improvements in AP achieved by SAM2BMP, this speed–performance trade-off is reasonable and well justified, particularly for applications that prioritize high segmentation quality over real-time processing.

Impact of Foreground Threshold. Table 4 analyzes the confidence threshold θ_{conf} for selecting foreground instances in mask propagation. As the threshold increases from 0.60 to 0.95, AP steadily improves from 49.9% to 52.7%, indicating that stricter filtering reduces false positives and improves association quality by focusing on high-confidence detections. However, when θ_{conf} further increases to 0.97, AP drops to 52.0%, suggesting that overly aggressive filtering excludes valid instances, particularly those with partial occlusions or challenging appearances that naturally have lower confidence scores. We adopt $\theta_{\text{conf}} = 0.95$ as the optimal balance between precision and recall.

Impact of Key Components. Table 5 evaluates the contribution of each technical component using ResNet-50 Mask2Former and SAM2-Large on YouTube-VIS 2021. Starting from a baseline without any improvements (ID 1, 43.2% AP), adding Backward Mask Propagation Completion (BMPC) provides the most significant gain of +5.3 AP (ID 2), demonstrating its critical role in completing fragmented trajectories by propagating masks backward from later frames. The Key Frame Ratio Filter (KFRF) further improves AP by +2.7 points (ID 3), effectively removing spurious tracklets that appear in too few frames. Low-Confidence Proposals (LCP) add +1.1 AP (ID 4) by recovering initially missed instances through backward propagation, while Tracklet NMS contributes +0.4 AP (ID 5) through final refinement to eliminate duplicate tracklets. The full model achieves 52.7% AP, with a cumulative improvement of +9.5 AP over the baseline, validating that all components complement each other and jointly contribute to the overall performance.

4.4. Visualization Analysis

To more intuitively demonstrate the advantages of SAM2BMP, we select some representative video sequences for visualization analysis. In the analysis, we compare SAM2BMP with the previous SOTA method CTVIS [58], which is a classic method based on object feature matching. From the visualization results, we can observe the following key advantages:

1) Handling Complex Motion Scenarios. As shown in Figure 3(a), when multiple bicycles appear simultaneously in the scene and move rapidly with repeated occlusions, CTVIS struggles to maintain consistent tracking of the front bicycle, leading to ID confusion and jumping between different frames. This is mainly because feature-based matching methods tend to fail when object appearance and position change rapidly. In contrast, SAM2BMP, by combining spatiotemporal information and more powerful association mechanisms, can maintain consistent segmentation and tracking of objects in these complex motion scenarios. Even when objects move rapidly, are partially occluded, or temporarily disappear, SAM2BMP can accurately maintain ID consistency, demonstrating stronger object association capabilities and robustness.

2) Handling ID Drift. As shown in Figure 3(b), CTVIS experiences ID drift, incorrectly assigning the ID of the previous white van to the later appearing white SUV, leading to confusion in instance tracking. In contrast, SAM2BMP can accurately distinguish between old and new objects, neither incorrectly assigning original IDs to newly appearing objects nor correctly assigning new IDs to new objects. This capability is crucial for accurate long-term object tracking and instance segmentation, especially in complex scenes containing multiple objects with similar appearances.

3) Occlusion Scenario Handling. As shown in Figure 3(b), CTVIS fails to successfully segment the head region of the background cat due to occlusion. In contrast, SAM2BMP can accurately identify and segment the complete object even under complex occlusion conditions, including the occluded cat's head, demonstrating powerful occlusion handling capabilities. This advantage is particularly evident in complex video sequences containing multiple objects that occlude each other.

These visualization results intuitively demonstrate the robustness and accuracy of SAM2BMP in handling complex video scenes, particularly its superior performance in challenging scenarios such as rapid motion and occlusion.

5. Conclusion

In this work, we present SAM2BMP, a strong video instance segmentation framework that leverages mask propagation from SAM2 for robust object matching. Unlike traditional feature-based matching approaches, our method directly utilizes spatial information and shape features through mask propagation, providing more intuitive and reliable object association across frames. Extensive experiments on three representative datasets demonstrate that SAM2BMP achieves superior performance compared to existing methods across different backbone architectures. Visualization results further confirm the robustness of our approach in handling challenging scenarios including complex motion, occlusion, and ID drift.

Limitations and Future Work. While our method achieves strong performance, several limitations remain. First, mask propagation introduces computational overhead, resulting in lower inference speed compared to some feature-based methods. Second, the approach inherits SAM2's limitations in handling extremely fast motion or severe motion blur. Future work includes: **1)** Developing more efficient propagation mechanisms to improve speed; **2)** Extending bidirectional propagation to handle complex temporal patterns; **3)** Integrating with emerging foundation models (*e.g.*, MLLMs [33, 56, 5, 60]) for enhanced robustness.

References

[1] A. Caelles, T. Meinhardt, G. Brasó, and L. Leal-Taixé. Devis: Making deformable transformers work for video instance segmentation. *arXiv preprint arXiv:2207.11103*, 2022. 2

[2] J. Cao, R. M. Anwer, H. Cholakkal, F. S. Khan, Y. Pang, and L. Shao. SipMask: Spatial Information Preservation for Fast Image and Video Instance Segmentation. *Eur. Conf. Comput. Vis.*, 2020. 2

[3] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *Int. Conf. Mach. Learn.*, pages 1597–1607. PMLR, 2020. 3

[4] X. Chen, H. Fan, R. Girshick, and K. He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3

[5] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 9

[6] B. Cheng, A. Choudhuri, I. Misra, A. Kirillov, R. Girdhar, and A. G. Schwing. Mask2Former for Video Instance Segmentation. *arXiv preprint arXiv:2112.10764*, 2021. 1, 2

[7] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar. Masked-Attention Mask Transformer for Universal Image Segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1290–1299, 2022. 2, 3, 5

[8] A. Choudhuri, G. Chowdhary, and A. G. Schwing. Context-aware relative object queries to unify video instance and panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6377–6386, 2023. 2

[9] H. Ding, C. Liu, S. He, X. Jiang, P. H. Torr, and S. Bai. MOSE: A new dataset for video object segmentation in complex scenes. In *IEEE Int. Conf. Comput. Vis.*, pages 20224–20234, 2023. 2

[10] H. Ding, C. Liu, S. He, K. Ying, X. Jiang, C. C. Loy, and Y.-G. Jiang. Mevis: A multi-modal dataset for referring motion expression video segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[11] H. Ding, K. Ying, C. Liu, S. He, X. Jiang, Y.-G. Jiang, P. H. Torr, and S. Bai. Mosev2: A more challenging dataset for video object segmentation in complex scenes. *arXiv preprint arXiv:2508.05630*, 2025. 2

[12] S. Ding, R. Qian, X. Dong, P. Zhang, Y. Zang, Y. Cao, Y. Guo, D. Lin, and J. Wang. Sam2long: Enhancing sam 2 for long video segmentation with a training-free memory tree. *arXiv preprint arXiv:2410.16268*, 2024. 3

[13] S. H. Han, S. Hwang, S. W. Oh, Y. Park, H. Kim, M.-J. Kim, and S. J. Kim. Visolo: Grid-based space-time aggregation for efficient online video instance segmentation. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2896–2905, 2022. 2

[14] F. He, H. Zhang, N. Gao, J. Jia, Y. Shan, X. Zhao, and K. Huang. Inspro: Propagating instance query and proposal for online video instance segmentation. 2022. 2

[15] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9729–9738, 2020. 3

[16] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *IEEE Int. Conf. Comput. Vis.*, pages 2961–2969, 2017. 2

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 5, 6

[18] M. Heo, S. Hwang, J. Hyun, H. Kim, S. W. Oh, J.-Y. Lee, and S. J. Kim. A Generalized Framework for Video Instance Segmentation. *arXiv preprint arXiv:2211.08834*, 2022. 1, 3

[19] M. Heo, S. Hwang, S. W. Oh, J.-Y. Lee, and S. J. Kim. VITA: Video Instance Segmentation via Object Token Association. In *Adv. Neural Inform. Process. Syst.*, 2022. 1, 2, 6, 7

[20] L. Hong, W. Chen, Z. Liu, W. Zhang, P. Guo, Z. Chen, and W. Zhang. Lvos: A benchmark for long-term video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13480–13492, 2023. 2

[21] L. Hong, Z. Liu, W. Chen, C. Tan, Y. Feng, X. Zhou, P. Guo, J. Li, Z. Chen, S. Gao, et al. Lvos: A benchmark for large-scale long-term video object segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[22] H. Hu, K. Ying, and H. Ding. Segment anything across shots: A method and benchmark. *arXiv preprint arXiv:2511.13715*, 2025. 2

[23] Huang, De-An and Yu, Zhiding and Anandkumar, Anima. MinVIS: A Minimal Video Instance Segmentation Framework without Video-based Training. In *Adv. Neural Inform. Process. Syst.*, 2022. 1, 2, 6, 7

[24] S. Hwang, M. Heo, S. W. Oh, and S. J. Kim. Video Instance Segmentation using Inter-Frame Communication Transformers. *Adv. Neural Inform. Process. Syst.*, 34:13352–13363, 2021. 1, 2

[25] J. Jiang, Z. Wang, M. Zhao, Y. Li, and D. Jiang. Sam2mot: A novel paradigm of multi-object tracking by segmentation. *arXiv preprint arXiv:2504.04519*, 2025. 3

[26] L. Ke, H. Ding, M. Danelljan, Y.-W. Tai, C.-K. Tang, and F. Yu. Video mask transfiner for high-quality video instance segmentation. In *Eur. Conf. Comput. Vis.*, pages 731–747. Springer, 2022. 2

[27] H. Kim, J. Kang, M. Heo, S. Hwang, S. W. Oh, and S. J. Kim. Visage: Video instance segmentation with appearance-guided enhancement. In *European Conference on Computer Vision*, pages 93–109. Springer, 2024. 2

[28] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 3

[29] R. Koner, T. Hannan, S. Shit, S. Sharifzadeh, M. Schubert, T. Seidl, and V. Tresp. Instanceformer: An online video instance segmentation framework. In *AAAI*, volume 37, pages 1188–1195, 2023. 2

[30] S. Lee, J. Seo, K. Han, M. Choi, and S. Im. Context-aware video instance segmentation. In *IEEE Int. Conf. Comput. Vis.*, 2025. 1, 3, 6, 7

[31] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *Eur. Conf. Comput. Vis.*, pages 740–755. Springer, 2014. 5

[32] C. Liu, H. Ding, K. Ying, L. Hong, N. Xu, L. Yang, Y. Fan, M. Gao, J. Chen, Y. Miao, et al. Lsvos 2025 challenge report: Recent advances in complex video object segmentation. *arXiv preprint arXiv:2510.11063*, 2025. 2

[33] S. Liu, K. Ying, H. Zhang, Y. Yang, Y. Lin, T. Zhang, C. Li, Y. Qiao, P. Luo, W. Shao, et al. Convbench: A multi-turn conversation evaluation benchmark with hierarchical capability for large vision-language models. *arXiv preprint arXiv:2403.20194*, 2024. 9

[34] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *IEEE Int. Conf. Comput. Vis.*, pages 10012–10022, 2021. 5, 6

[35] X. Mao, Z. Li, C. Li, X. Xu, K. Ying, T. He, J. Pang, Y. Qiao, and K. Zhang. Yume-1.5: A text-controlled interactive world generation model. *arXiv preprint arXiv:2512.22096*, 2025. 1

[36] J. Meng, Z. Wang, K. Ying, J. Zhang, D. Guo, Z. Zhang, J. Q. Shi, and S. Chen. Human interaction understanding with consistency-aware learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11898–11914, 2023. 1

[37] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabbat, M. Assran,

N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 5, 6

[38] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 2

[39] J. Qi, Y. Gao, Y. Hu, X. Wang, X. Liu, X. Bai, S. Belongie, A. Yuille, P. H. Torr, and S. Bai. Occluded Video Instance Segmentation: A Benchmark. *Int. J. Comput. Vis.*, 130(8):2022–2039, 2022. 1, 5, 7

[40] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2, 3, 5

[41] M. Siam, A. Kendall, and M. Jagersand. Video class agnostic segmentation benchmark for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2825–2834, 2021. 1

[42] C.-H. Tseng, C.-C. Hsieh, D.-J. Jwo, J.-H. Wu, R.-K. Sheu, and L.-C. Chen. Person retrieval in video surveillance using deep learning–based instance segmentation. *Journal of Sensors*, 2021(1):9566628, 2021. 1

[43] J. Videnovic, A. Lukezic, and M. Kristan. A distractor-aware memory for visual object tracking with sam2. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 24255–24264, 2025. 3

[44] Y. Wang, Z. Xu, X. Wang, C. Shen, B. Cheng, H. Shen, and H. Xia. End-to-End Video Instance Segmentation With Transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8741–8750, 2021. 2

[45] Z. Wang, K. Ying, J. Meng, and J. Ning. Human-to-human interaction detection. In *International Conference on Neural Information Processing*, pages 120–132. Springer, 2023. 1

[46] J. Wu, Y. Jiang, S. Bai, W. Zhang, and X. Bai. SeqFormer: Sequential Transformer for Video Instance Segmentation. In *Eur. Conf. Comput. Vis.*, pages 553–569. Springer, 2022. 1, 2

[47] J. Wu, Q. Liu, Y. Jiang, S. Bai, A. Yuille, and X. Bai. In Defense of Online Models for Video Instance Segmentation. In *Eur. Conf. Comput. Vis.*, pages 588–605. Springer, 2022. 1, 2, 3, 6, 7

[48] J. Wu, S. Yarram, H. Liang, T. Lan, J. Yuan, J. Eledath, and G. Medioni. Efficient video instance segmentation via tracklet query and proposal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 959–968, 2022. 2

[49] N. Xu, L. Yang, Y. Fan, D. Yue, Y. Liang, J. Yang, and T. Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 2

[50] C.-Y. Yang, H.-W. Huang, W. Chai, Z. Jiang, and J.-N. Hwang. Samurai: Adapting segment anything model for zero-shot visual tracking with motion-aware memory. *arXiv preprint arXiv:2411.11922*, 2024. 3

[51] L. Yang, Y. Fan, and N. Xu. Video instance segmentation. In *IEEE Int. Conf. Comput. Vis.*, pages 5188–5197, 2019. 1, 2, 5, 6

[52] S. Yang, Y. Fang, X. Wang, Y. Li, C. Fang, Y. Shan, B. Feng, and W. Liu. Crossover Learning for Fast Online Video Instance Segmentation. In *IEEE Int. Conf. Comput. Vis.*, pages 8043–8052, 2021. 2

[53] S. Yang, X. Wang, Y. Li, Y. Fang, J. Fang, W. Liu, X. Zhao, and Y. Shan. Temporally Efficient Vision Transformer for Video Instance Segmentation . In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2885–2895, 2022. 2

[54] K. Ying, H. Ding, G. Jie, and Y.-G. Jiang. Towards Omnimodal Expressions and Reasoning in Referring Audio-Visual Segmentation. In *ICCV*, 2025. 2, 3

[55] K. Ying, H. Hu, and H. Ding. MOVE: Motion-Guided Few-Shot Video Object Segmentation. In *ICCV*, 2025. 2

[56] K. Ying, F. Meng, J. Wang, Z. Li, H. Lin, Y. Yang, H. Zhang, W. Zhang, Y. Lin, S. Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024. 9

[57] K. Ying, Z. Wang, C. Bai, and P. Zhou. ISDA: Position-Aware Instance Segmentation with Deformable Attention. In *Int. Conf. Acoustics, Speech, & Signal Process.*, pages 2619–2623. IEEE, 2022. 2

[58] K. Ying, Q. Zhong, W. Mao, Z. Wang, H. Chen, L. Y. Wu, Y. Liu, C. Fan, Y. Zhuge, and C. Shen. CTVIS: Consistent Training for Online Video Instance Segmentation, 2023. 1, 2, 3, 5, 6, 7, 9

[59] T. Zhang, X. Tian, Y. Wu, S. Ji, X. Wang, Y. Zhang, and P. Wan. DVIS: Decoupled Video Instance Segmentation Framework. *IEEE Int. Conf. Comput. Vis.*, 2023. 1, 2, 3, 5, 6, 7

[60] P. Zhou, K. Ying, Z. Wang, D. Guo, and C. Bai. Self-supervised enhancement for named entity disambiguation via multimodal graph convolution. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1):231–245, 2022. 9

[61] Y. Zhou, T. Zhang, S. Ji, S. Yan, and X. Li. Improving video segmentation via dynamic anchor queries. In *Eur. Conf. Comput. Vis.*, pages 446–463. Springer, 2024. 5, 6, 7