SAM-MT: Real-Time Interactive Multi-Target Video Segmentation

Ruiqi Shen[1], Chang Liu[2]✉, and Henghui Ding[1] ✉

[1] Institute of Big Data, College of Computer Science and Artificial Intelligence, Fudan University, China
[2] Shanghai University of Finance and Economics, China
hhding@fudan.edu.cn
https://henghuiding.com/SAM-MT/

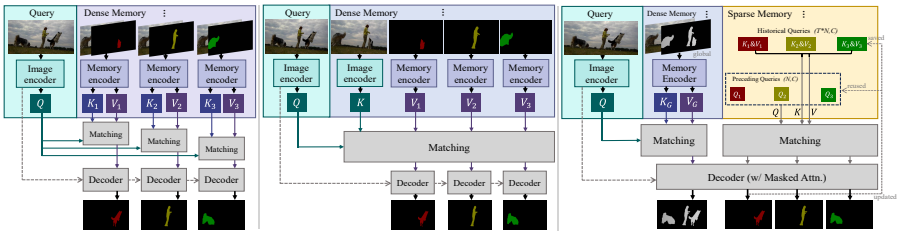
Abstract. Modern Video Object Segmentation (VOS) involves tracking and segmenting user-specified targets. While recent approaches have achieved remarkable performance in single-target scenarios, extending them to multi-target settings typically involves replicating the single-target processing for each individual object, resulting in reduced frame rates (FPS) with unbounded latency as target density scales. Built upon Segment Anything 2 (SAM2), we propose **SAM-MT**, which addresses this by transforming the model into an interactive framework for real-time **M**ulti-**T**arget video segmentation. SAM-MT uses explicit queries propagated and updated across frames to represent different individual targets, in parallel with a shared representation for global context. It employs decoupled masked attention to keep individual identities distinct from cross-target interference, and sparse memory for stable temporal evolution, along with specialized strategies for occlusion handling and overlap prevention. SAM-MT successfully decouples latency from the number of targets, achieving real-time speed on par with single-target baselines (>36 FPS for 10 targets) while maintaining SAM2's robust video segmentation performance.

Keywords: Multi-target video segmentation · Real-time segmentation

1 Introduction

Contemporary Video Object Segmentation (VOS) tracks and segments user-specified objects in open-vocabulary environments, with applications spanning in-the-wild navigation [44] and robotics [21]. Dominant approaches follow the space-time-memory (STM) paradigm [37], exploiting heavy pixel-level representations stored in a dense memory to track and segment target and have achieved state-of-the-art (SOTA) performance across VOS benchmarks [7, 9, 11, 13, 15, 19, 37, 39].

✉ Corresponding authors



(a) STM, SAM2, etc. [10, 19, 27, 37, 39, 43, 48] (b) STCN, XMem, Cutie, etc. [2, 7–9, 11] (c) SAM-MT (Ours)

Fig. 1: Comparison of video segmentation architectures: Single-target methods, implemented via (a) fully independent encoding or (b) feature sharing, essentially rely on object-wise processing, resulting in computational costs and latency that scale with target count. In contrast, (c) SAM-MT models global context with a shared representation and individual targets with lightweight queries, achieving near-single-object efficiency as target count increases while preserving SAM2's segmentation performance.

While highly impressive, most of these approaches, including the powerful SAM2 [39] and its extensions [19, 42, 49], rely on object-wise memory and propagation, making them tailored for single-target processing; therefore, handling multiple targets requires repeating the same computations for every additional instance, as illustrated in Figure 1(a). Even with some earlier attempts to share frame-level image features, they still rely on independent object-wise processing for mask decoding and memory encoding [7, 9, 11] (Figure 1(b)). Consequently, such a straightforward extension results in a substantial rise in computational cost as the number of targets increases, leading to degraded frame rates (FPS) with unbounded latency. One possible workaround is to merge all targets into a single trackable object. However, this approach discards individual target IDs, and is therefore impractical for real-world applications. Furthermore, the merged object often forms irregular shapes that standard VOS models, trained on real-world objects, simply fail to recognize and track (see Section 5.3).

Meanwhile, real-world applications demand real-time tracking and segmentation of multiple targets. For instance, a practical VOS system for autonomous driving must 1) preserve individual target identities during propagation while 2) maintaining real-time speed, even in crowded scenes with dense vehicles and pedestrians. These requirements highlight the need for a multi-target video segmentation framework that could *maintain near-single-object efficiency regardless of target count*, while faithfully *preserving and updating individual identities* during propagation. Recent approaches, including SAM2, struggle to bridge this gap, as their object-wise processing causes computation to grow with target count.

We therefore present *SAM-MT* to address this. As illustrated in Figure 1(c), SAM-MT extends the SAM2 architecture with a hybrid approach: it employs a shared representation to model the global context of all targets, while introducing *target queries* in parallel for representing individual targets. To prevent cross-target interference, we propose *decoupled masked attention*, which restricts attention between queries of different targets, while ensuring their shared access to the global context. For temporal evolution, a *query-based sparse memory*

stores historical queries from individual targets across frames for per-target re-identification. No extra mask decoders or memory encoders are required, allowing SAM-MT to avoid redundant object-wise computation and maintain near-single-object efficiency as the number of targets increases. For training, we adopt strided frame sampling rather than adjacent frames to better handle occlusions under GPU constraints. To further encourage instance-level separation, specialized supervision is introduced to penalize overlaps between different targets.

Comprehensive experiments demonstrate that SAM-MT achieves strong performance in real-world complex scenarios involving dense targets, frequent occlusions, and long-term sequences. It also performs on par with SAM2.1-B+ across six VOS benchmarks, including the challenging MOSEv2 [18] and the long-term LVOSv2 [26]. Remarkably, SAM-MT maintains near-single-object efficiency as target count increases, running at 36+ FPS in crowded scenes with 10 targets. This presents an efficiency advantage over prior methods such as SAM2.1-B+, whose frame rate drops sharply from 37 FPS for a single target to 17.8 and 12.4 FPS for 3 and 5 targets, respectively.

In summary, our contributions are summarized as follows:

- We present *SAM-MT*, an efficient framework for real-time interactive multi-target video segmentation. By decoupling computational costs from target count, SAM-MT achieves *near-single-object efficiency* as target number increases, while maintaining SAM2's robust video segmentation performance.
- We introduce *target queries* to represent individual targets and *decoupled masked attention* to prevent cross-target interference while sharing global context, thereby preserving target identities throughout the sequence.
- We introduce a lightweight *query-based sparse memory* to capture the temporal evolution of individual targets, with strided sampling and overlap prevention for enhanced robustness.
- SAM-MT achieves competitive video segmentation performance across six challenging VOS benchmarks in the interactive setting, while achieving the best efficiency across target densities. For 10 targets, it runs at a real-time speed of 36+ FPS, a 6× speedup over SAM2.1-B+.

2 Related Works

The single-target bottleneck of VOS. The pioneering Space-Time Memory (STM) paradigm is inherently designed for single-target VOS, tracking a specific object by matching its dense pixel-level memory against query frames [37,38,40, 43]. Extending such methods to multiple targets typically requires replicating the single-object pipeline for each target, causing the computational burden to grow with target count. To tackle this, subsequent methods like STCN [11], XMem [9], and Cutie [7] reuse frame-level image features to share affinity computations, but they still rely on independent mask decoding and memory encoding for each object. Despite the success of SAM-family methods in video segmentation [19,39], they still follow an object-wise processing paradigm, requiring the single-object pipeline to be replicated for each target and thus making multi-target streaming costly.

Query-based Segmentation. Query-based transformers have become the dominant architecture for image and video segmentation, with established approaches [5,6,16,17,20,22,28–30,32,34,35,47,52,55] surpassing traditional CNN-based baselines [3,12,23,36]. These methods typically use learnable queries that cross-attend to pixel-level features to produce rich target representations [41], followed by task-specific heads for mask prediction. Recent video segmentation methods propagate queries across frames, but these queries often lack explicit target identities, either serving as generic object slots [4] or foreground-background representations [7]. In contrast, SAM-MT assigns each user-specified target an ID-aware query for identity-consistent propagation across frames.

3 Methodology

3.1 Preliminaries: SAM2

SAM2 [39] extends the Segment Anything [30] paradigm to video by conditioning current-frame features on a dense pixel-level memory. At frame t, the image features $F_t \in \mathbb{R}^{HW\times C}$ are refined by cross-attending to the memory readout $M_{\text{dense},t} \in \mathbb{R}^{T\times HW\times C}$:

$$\tilde{F}_t = \text{softmax}(F_t M_{\text{dense},t}^{\top}) M_{\text{dense},t} + F_t. \tag{1}$$

The mask decoder then generates masks from a set of queries $\mathbf{Q} = [G; P]$, where G and P denote the global queries and encoded user prompts, respectively. These queries are first updated via self-attention and then cross-attend to $\tilde{F}_t$ for contextual refinement and mask prediction. The target is then encoded into a dense pixel-level memory representation, typically with 4096 tokens, along with a compact object pointer as an auxiliary object-level representation.

Since SAM2 maintains independent target representations and dense memories for each object, multi-target tracking thus requires replicated object-wise propagation, causing latency and memory usage to grow with target count.

3.2 Overview of SAM-MT

SAM-MT is a real-time multi-target video segmentation framework, as shown in Figure 2. It adopts point-based interactions, where clicks are associated with target IDs to specify distinct objects.

The input points are encoded as queries and combined with SAM2's global queries in the decoder. To preserve individual target identities, we introduce *decoupled masked attention*, which blocks cross-target interference while allowing each target to interact with the global context. The resulting queries then cross-attend to image features for pixel-level contextual refinement, following [39].

The refined point queries for each target are then consolidated into a unique target query (one query per target), which is used for mask prediction and stored in a unified first-in-first-out (FIFO) *sparse memory* shared across targets. For subsequent frames, an *identity transformer* retrieves the target-specific historical queries from the sparse memory to update the propagated target queries,

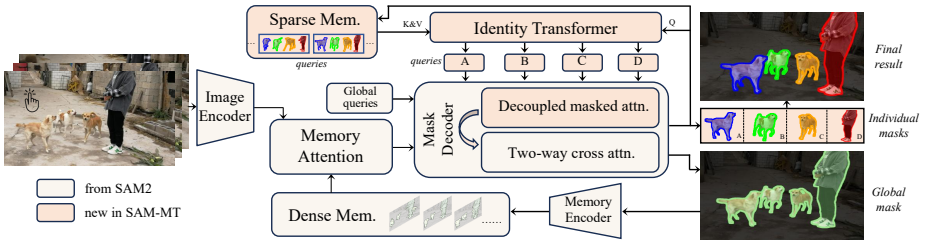


Fig. 2: Overview of SAM-MT. Our framework uses queries to represent individual targets, in parallel with a shared representation for global context. We introduce *decoupled masked attention* to prevent cross-target interference while maintaining access to the global context. A *sparse memory* stores historical queries of different targets, which are processed by an *identity transformer* for robust per-frame target re-identification.

which are then fed into the decoder to repeat the process, enabling consistent segmentation across frames.

In the following, we detail the core components of SAM-MT: decoupled masked attention, query-based sparse memory, and identity transformer. For intra-frame modules, temporal subscripts t are omitted for brevity. We do not claim novelty for modules inherited from SAM2.

3.3 Scalable Target Queries

Existing query-based video segmentation methods typically allocate a fixed number of generic queries (e.g., 16 in [7]), making them difficult to scale to real-world scenes with arbitrary numbers of user-specified targets. As illustrated in Figure 2, SAM-MT adopts a scalable design that assigns explicit target queries to user-specified targets, naturally accommodating varying target counts.

3.4 Decoupled Masked Attention

To prevent cross-target interference, we introduce a *decoupled masked attention* strategy within the self-attention blocks of the decoder. Let $G \in \mathbb{R}^{N_g \times C}$ denote the global queries, $Q_q \in \mathbb{R}^{N_q \times C}$ denote the queries of target q, and k denote the number of targets, we concatenate all queries as $\mathbf{Q} = [G; Q_1; \ldots; Q_k] \in \mathbb{R}^{N \times C}$, where $N = N_g + \sum_{q=1}^{k} N_q$. Here, N_q is the number of user-provided clicks for target q in the initial frame, and simplifies to $N_q = 1$ for all subsequent frames as each target is represented by a single target query (see Section 3.5). The decoupled masked attention with residual connection [24] is then computed as:

$$\mathbf{Q}' = \text{softmax}(\mathbf{Q}K^\top + \mathcal{M})V + \mathbf{Q}, \tag{2}$$

where $\mathcal{M} \in \mathbb{R}^{N \times N}$ denotes the attention mask with entries:

$$\mathcal{M}_{i,j} = \begin{cases} 0 & \text{if } i \in [1, N_g] \text{ and } j \in [1, N], \\ 0 & \text{if } i \in [1, N] \text{ and } j \in [1, N_g], \\ 0 & \text{if } i, j \in \text{Rng}(q), \quad \text{for } q \in \{1, \ldots, k\}, \\ -\infty & \text{otherwise}, \end{cases} \tag{3}$$

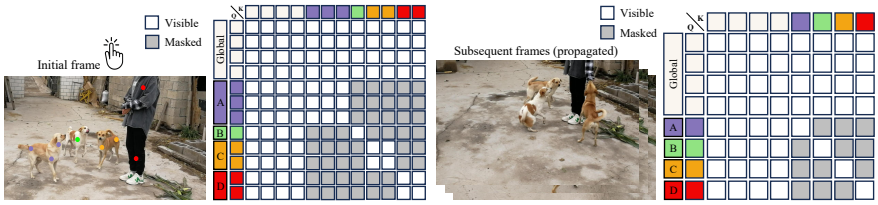


Fig. 3: Visualization of decoupled masked attention. *Left (Initial frame):* An arbitrary number of queries per target (e.g., 3, 1, 2, and 2 for targets A–D) are mutually isolated while sharing the global context. *Right (Subsequent frames):* Propagated target queries (one query per target) follow the same masking rule.

with $\mathrm{Rng}(q)$ defined as the index range of queries for target q:

$$\mathrm{Rng}(q) = \left\{ N_g + \sum_{r=1}^{q-1} N_r + 1, \ldots, N_g + \sum_{r=1}^{q} N_r \right\}. \tag{4}$$

The resulting $\mathbf{Q}'$ aggregates the updated global and individual contexts, from which Q'_q is extracted for target q via $\mathrm{Rng}(q)$. Figure 3 illustrates the effect of decoupled masked attention in both initial and subsequent frames: interference between queries of different targets is blocked, while all target queries share access to the global context. Following this, all queries cross-attend to the image features $\tilde{F}_t$ as in SAM2 [39], extracting pixel-level context for mask prediction.

3.5 Query Consolidation (Initial Frame)

For each target q, users may provide an arbitrary number of N_q points in the initial frame. These points are processed by the decoder into $Q'_q \in \mathbb{R}^{N_q \times C}$ as described in Section 3.4. To obtain a compact representation for the target, we apply a lightweight MLP-based weighting head f_{weight} to compute the importance $w_q \in \mathbb{R}^{N_q}$ of each of its points, aggregating $Q'_q \in \mathbb{R}^{N_q \times C}$ into a single query $\hat{Q}_q \in \mathbb{R}^{1 \times C}$, denoted as the *target query*:

$$w_{q,i} = \mathrm{softmax}(f_{\text{weight}}(Q'_{q,i})), \quad \hat{Q}_q = \sum_{i=1}^{N_q} w_{q,i} \cdot Q'_{q,i}. \tag{5}$$

This process is performed only in the initial frame, producing a unique target query for each target q. For subsequent frames, since each target is represented by its target query that is propagated and updated frame by frame, consolidation is no longer needed and we simply have $\hat{Q}_q = Q'_q$.

3.6 Query-based Sparse Memory

As SAM2 relies on dense pixel-level memory $M_{\text{dense},t} \in \mathbb{R}^{T \times HW \times C}$, duplicating such memory for each target is computationally prohibitive. To maintain real-time efficiency, SAM-MT uses dense memory only for the combined mask of all

targets to capture global context. In parallel, we introduce a *query-based sparse memory* to model the temporal evolution of individual targets.

Formally, let $\hat{Q}_{q,t} \in \mathbb{R}^{1\times C}$ denote the updated query for target q at frame t. We aggregate the queries of all k targets into $\hat{\mathbf{Q}}_t = [\hat{Q}_{1,t}, \ldots, \hat{Q}_{k,t}] \in \mathbb{R}^{k\times C}$ and update the FIFO sparse memory $M_{\text{sparse},t}$, which maintains historical queries over a temporal window size of T frames, including the initial frame and the $T-1$ most recent frames:

$$M_{\text{sparse},t} = [\hat{\mathbf{Q}}_t, \hat{\mathbf{Q}}_{t-1}, \ldots, \hat{\mathbf{Q}}_{t-T+2}, \hat{\mathbf{Q}}_0] \in \mathbb{R}^{T\cdot k\times C}. \tag{6}$$

The sparse memory reduces the frame-level per-target storage cost from HW dense tokens to one target-query token, thus enabling a much longer temporal window that could improve robustness to target occlusions and reappearances.

3.7 Identity Transformer

To maintain identity consistency across frames, we use an identity transformer to update target queries with their corresponding historical queries from the sparse memory. Specifically, at frame t, the target queries from the previous frame $\hat{\mathbf{Q}}_{t-1} \in \mathbb{R}^{k\times C}$ cross-attend to the sparse memory $M_{\text{sparse},t-1} \in \mathbb{R}^{T\cdot k\times C}$. To prevent cross-target interference, an identity-aware mask $\tilde{\mathcal{M}}$ is applied within the update:

$$\mathbf{Q}_t = \text{softmax}(\hat{\mathbf{Q}}_{t-1} M_{\text{sparse},t-1}^{\top} + \tilde{\mathcal{M}}) M_{\text{sparse},t-1} + \hat{\mathbf{Q}}_{t-1}, \tag{7}$$

where the mask $\tilde{\mathcal{M}} \in \mathbb{R}^{k\times(T\cdot k)}$ ensures that each target query only attends to its own history:

$$\tilde{\mathcal{M}}_{i,j} = \begin{cases} 0 & \text{if } j \in \Omega_i, \\ -\infty & \text{otherwise,} \end{cases} \tag{8}$$

with $\Omega_i = \{i + \tau \cdot k \mid 0 \le \tau \le T-1\}$ denoting the indices of historical queries for target i.

By restricting each target to its own history, this design preserves identity consistency across frames and mitigates potential identity drift.

4 Experimental Setup

4.1 Implementation Details

Training Setup. We initialize SAM-MT from the pre-trained SAM2.1-B+ [39] checkpoint and train it on 8 NVIDIA A6000 GPUs. Following [39], we use a learning rate of 3×10^{-6} for the image encoder and 5×10^{-6} for other components. Training strategies are detailed in Section 4.2.

Training Data. We train SAM-MT on a filtered subset of the SA-V training set [39], retaining sequences with at least three concurrent targets ($\approx 35\%$ of the original SA-V training set) to encourage robust multi-target handling.

4.2 Training Strategies

Training Scheme. We employ a two-stage training scheme. Stage I focuses on static image-level training to ensure high-quality one-shot multi-target image segmentation from clicks. Stage II extends the training to video sequences, encouraging temporal identity consistency across frames.
Strided Sampling. We sample 8 frames per sequence by combining consecutive sampling and 4-frame strided sampling (spanning a window of 32 frames). This enables the model to capture both short-term motion and long-term dynamics, such as occlusion and reappearance, under GPU memory constraints.
Point Sampling. To simulate realistic user interactions, we randomly sample 1–5 positive and 0-2 negative points per target in the initial frame of each sequence. We use a combination of grid-based sampling and random sampling to improve spatial coverage.
Overlap Prevention. To reduce spatial ambiguity and encourage each pixel only belongs to one object, we penalize mask overlaps between targets. Let $\mathbf{P}_i \in [0,1]^{H\times W}$ be the predicted probability map for target i, we define its overlap loss to encourage mutually exclusive predictions and alleviate pixel-level ambiguity:

$$\mathcal{L}_{ovl,i} = \text{Mean}\left(\mathbf{P}_i \odot \sum_{j\neq i} \mathbf{P}_j\right). \tag{9}$$

Loss Designs. The overall loss for SAM-MT consists of two parts: supervision for individual masks $\mathcal{L}_I$ and supervision for the global mask $\mathcal{L}_G$. For k individual targets, we apply Focal and Dice losses together with the overlap loss to reduce identity ambiguity. For the global mask, we follow SAM2 [39] and use Focal, Dice, IoU, and object score losses. Loss weights are omitted for brevity:

$$\mathcal{L}_I = \frac{1}{k}\sum_{i=1}^{k}(\mathcal{L}_{focal,i} + \mathcal{L}_{dice,i} + \mathcal{L}_{ovl,i}), \tag{10}$$

$$\mathcal{L}_G = \mathcal{L}_{focal} + \mathcal{L}_{dice} + \mathcal{L}_{iou} + \mathcal{L}_{obj}, \tag{11}$$

$$\mathcal{L}_{total} = \mathcal{L}_I + \mathcal{L}_G. \tag{12}$$

This dual-level supervision encourages coherent global segmentation while improving the accuracy of individual target masks.

4.3 Benchmarks and Metrics

To evaluate SAM-MT's performance in complex real-world scenarios, we conduct quantitative evaluations on six challenging VOS benchmarks, including the validation splits of MOSEv2 [18], MOSEv1 [14], LVOSv2 [26], LVOSv1 [25], as well as SA-V val [39] and SA-V test [39]. Specifically, MOSE provides challenging cases such as frequent occlusions, rapid motion, low-light environments, and crowded scenes with inconspicuous targets, while LVOS focuses on long-term temporal consistency with object disappearances and reappearances. SA-V

Table 1: Quantitative comparisons on VOS benchmarks. The best and second-best performances are marked in red and orange. "–": unavailable due to out-of-memory.

Method	Initialization	MOSEv2-val					MOSEv1-val			LVOSv2-val			LVOSv1-val		
		$\mathcal{J}\&\dot{\mathcal{F}}$	$\mathcal{J}$	$\dot{\mathcal{F}}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
STCN [11]	mask	29.7	28.9	30.5	31.4	30.2	50.8	46.6	55.0	65.3	61.6	68.9	45.8	41.1	50.5
AOT-L [50]	mask	30.2	29.0	31.4	32.9	31.0	57.2	53.1	61.3	63.9	60.0	67.8	59.4	53.6	65.2
R50-AOT [50]	mask	34.9	33.1	36.6	38.9	36.0	58.5	54.4	62.6		–			–	
SwinB-AOT [50]	mask	34.2	32.4	36.0	38.4	35.4	60.2	56.2	64.1		–			–	
DeAOT-L [51]	mask	32.6	30.7	34.5	37.2	33.9	59.4	55.1	63.8	67.0	62.4	71.6	58.2	51.6	64.9
R50-DeAOT [51]	mask	31.6	29.7	33.5	36.1	32.9	59.0	54.7	63.4	71.4	67.8	75.1	64.7	59.1	70.3
SwinB-DeAOT [51]	mask	36.7	34.9	38.6	41.0	37.9	61.7	57.6	65.9	72.6	69.0	76.2	62.6	57.4	67.8
RDE [33]	mask	32.0	30.7	33.3	35.0	32.8	48.8	44.6	52.9	61.5	58.0	65.0	52.9	47.7	58.1
XMem [9]	mask	36.3	34.7	37.9	40.0	37.4	57.6	53.3	62.0	64.7	61.8	67.6	50.0	45.5	54.4
XMem++ [2]	mask	34.2	32.5	35.9	37.9	35.2	56.0	51.5	60.6	61.5	58.6	64.3	42.8	38.6	47.0
DEVA [8]	mask	38.3	36.6	40.0	42.2	39.4	60.0	55.8	64.3	72.4	68.7	76.1	55.9	51.1	60.7
Cutie-base [7]	mask	42.8	41.1	44.4	46.8	43.9	68.3	64.2	72.3	70.1	66.7	73.5	66.0	61.3	70.6
SAM2.1-B+ [39]	click	41.1	39.4	42.8	45.4	42.4	65.1	60.7	69.5	74.6	70.9	78.3	71.3	66.5	76.1
SAM-MT (Ours)	click	43.0	41.4	44.5	47.1	44.2	68.2	64.4	72.0	76.6	73.3	80.0	73.6	69.0	78.3

presents additional challenges in heavy occlusion and part-level object modeling. Following [7, 18, 39], we report $\mathcal{J}\&\mathcal{F}$ for all benchmarks with additional $\mathcal{J}\&\dot{\mathcal{F}}$ for MOSEv2. For multi-target scalability analysis, we report Frames Per Second (FPS) to measure computational efficiency with respect to the number of targets, and VRAM to quantify GPU memory consumption.

4.4 Initialization

For a fair comparison, both SAM-MT and SAM2 are initialized with the same clicks, specifically two positive clicks per target in the initial frame of each sequence. In contrast, all other baselines are initialized with corresponding ground-truth masks, providing stronger initialization than our interactive setting.

5 Experiments

5.1 Quantitative Evaluation

As shown in Table 1, SAM-MT achieves strong video segmentation performance across VOS benchmarks in a *zero-shot* manner. Notably, on the challenging MOSEv2, SAM-MT reaches 43.0 $\mathcal{J}\&\dot{\mathcal{F}}$, improving over previous SOTA baselines including SAM2.1-B+ and Cutie. The gains are more pronounced in long-term scenarios, where SAM-MT surpasses SAM2.1-B+ by 2.0 and 2.3 points in $\mathcal{J}\&\mathcal{F}$ on LVOSv2 and LVOSv1, respectively. Table 2 further shows SAM-MT's robustness on SA-V val/test, achieving competitive $\mathcal{J}\&\mathcal{F}$ on both splits.

5.2 Multi-target Scalability Analysis

Benchmark and Baselines. Existing VOS datasets are dominated by single-target sequences (e.g., 79% in MOSEv2-val). We therefore construct a synthetic benchmark by selecting multi-target sequences from those datasets to assess multi-target scalability under varying numbers of targets. Specifically, the benchmark contains 20 sequences, corresponding to target counts from 1 to 20, with

Table 2: Quantitative comparisons on SA-V validation and test splits. The best and second-best performances are marked in red and orange.

Method	Initialization	SA-V val			SA-V test		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
SwinB-AOT [50]	mask	51.1	46.4	55.7	50.3	46.0	54.6
SwinB-DeAOT [51]	mask	61.4	56.6	66.2	61.8	57.2	66.3
RDE [33]	mask	51.8	48.4	55.2	53.9	50.5	57.3
XMem [9]	mask	60.1	56.3	63.9	62.3	58.9	65.8
DEVA [8]	mask	55.4	51.5	59.2	56.2	52.4	60.1
Cutie-base [7]	mask	60.7	57.7	63.7	62.7	59.7	65.7
SAM2.1-B+ [39]	click	65.6	62.0	69.3	66.1	62.4	69.8
SAM-MT	click	66.1	62.9	69.4	66.4	63.0	69.8

Fig. 4: Example videos from the multi-target synthetic benchmark (initial frames).

each sequence padded or truncated to 100 frames. Example sequences are shown in Figure 4. We compare SAM-MT with SAM2.1-B+ [39], Cutie [7], as well as DeAOT [51], a representative VOS method that also aims to decouple latency from target count. For a fair comparison, SAM2.1-B+ is evaluated with its official multi-target setting, including batch-wise target concatenation and feature reuse, where image features are computed once and shared across targets for acceleration. Cutie and DeAOT include these optimizations by design. All experiments are conducted on a single NVIDIA A6000 GPU with 48GB VRAM.

FPS Comparison. As shown in Table 3 and Figure 5, SAM-MT largely decouples latency from target count, maintaining near-single-object efficiency as target density scales. It achieves the highest FPS across all target counts, despite processing higher-resolution inputs (1024p) than many baselines (480p). In contrast, SAM2.1-B+ and Cutie suffer significant FPS degradation as the number of targets increases; for example, SAM2.1-B+ drops from 37.2 FPS (1 target) to 17.8 FPS (3 targets). Although DeAOT-L and SwinB-DeAOT maintain stable processing speeds for up to 10 targets, their speeds drop sharply beyond this point, and their overall frame rates remain substantially lower than ours.

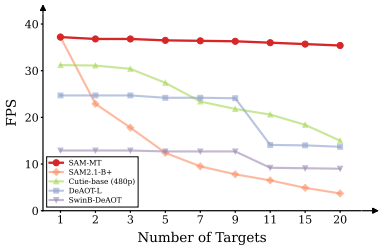


Fig. 5: FPS comparison on the synthetic benchmark.

VRAM Comparison. Table 4 compares VRAM usage under different target counts. SAM-MT keeps memory consumption nearly stable as target count increases, rising only from 3094 MB (1 target) to 3785 MB (20 targets). In contrast, SAM2.1-B+ grows substantially from 3043 MB to 8585 MB, showing the memory efficiency of our design in dense multi-target scenes.

FPS Comparison on VOS Benchmarks. Table 5 reports FPS on standard VOS benchmarks under different target densities. Although these VOS bench-

Table 3: FPS comparison on the synthetic benchmark.

Method	Res.	FPS vs. Number of Targets								
		1	2	3	5	7	9	11	15	20
DeAOT-L [51]	$1.3 \times 480p$	24.7	24.7	24.7	24.2	24.2	24.1	14.1	14.0	13.7
SwinB-DeAOT [51]	$1.3 \times 480p$	12.9	12.9	12.9	12.7	12.7	12.7	9.2	9.1	9.0
Cutie-base [7]	$480p$	31.2	31.1	30.4	27.4	23.4	21.8	20.6	18.4	15.0
Cutie-base [7]	$1024p$	20.3	17.5	14.7	13.2	11.2	9.4	8.9	7.3	6.3
SAM2.1-B+ [39]	$1024p$	**37.2**	22.9	17.8	12.4	9.5	7.8	6.5	4.9	3.7
SAM-MT	$1024p$	**37.2**	**36.8**	**36.8**	**36.5**	**36.4**	**36.3**	**36.0**	**35.7**	**35.4**

Table 4: VRAM (MB) comparison on the synthetic benchmark.

Method / Target Num.	1	3	5	7	9	11	13	15	17	20
SAM2.1-B+	**3043**	3356	3702	4089	4387	4529	4873	5436	6831	8585
SAM-MT	3094	**3247**	**3312**	**3357**	**3430**	**3491**	**3548**	**3627**	**3719**	**3785**

Table 5: FPS comparison on VOS benchmarks. "All": entire validation set; "≥2": subset containing sequences with ≥ 2 concurrent targets, etc.

Method	Res.	MOSEv2				MOSEv1				LVOSv2		LVOSv1	
		All	≥2	≥3	≥5	All	≥2	≥3	≥5	All	≥2	All	≥2
DeAOT-L [51]	$1.3 \times 480p$	22.9	21.5	21.4	21.1	29.7	29.0	28.9	28.5	9.5	9.5	7.2	6.5
SwinB-DeAOT [51]	$1.3 \times 480p$	12.7	12.7	12.6	12.3	15.8	15.8	15.6	14.6	6.9	6.7	5.6	4.9
Cutie-base [7]	$480p$	44.3	41.0	37.9	34.8	42.7	35.9	34.6	27.7	43.4	41.9	44.8	39.6
Cutie-base [7]	$1024p$	21.3	16.4	14.1	10.2	21.9	15.6	12.6	10.0	19.3	15.2	21.1	16.2
SAM2.1-B+ [39]	$1024p$	32.1	21.3	17.3	11.5	30.5	19.0	14.5	10.8	32.0	25.6	32.2	22.1
SAM-MT	$1024p$	36.9	36.2	36.1	35.8	39.2	37.1	36.6	36.1	36.5	35.6	36.7	36.0

marks are dominated by single-target sequences, SAM-MT consistently maintains real-time speed with only minor degradation as target count increases (e.g., 36.9→35.8 FPS on MOSEv2 for ≥ 5 targets). In contrast, SAM2.1-B+ degrades sharply from 32.1 to 11.5 FPS under the same setting, and Cutie shows a similar trend. While DeAOT remains relatively stable, its overall FPS is much lower.

5.3 Qualitative Evaluation

Highly Dense Scenarios. We evaluate SAM-MT in highly dense multi-target scenarios. As illustrated in Figure 6, SAM-MT achieves precise and consistent video segmentation in complex scenes such as a football match (top, 12 targets), a K-pop dance (middle, 10 targets) and a busy highway (bottom, 10 targets), demonstrating its robustness in maintaining and updating distinct identities within highly cluttered environments.

Identity-Ambiguous Scenarios. SAM-MT combines global context and individual identities through decoupled masked attention, while SAM2 processes each target in isolation and may be affected by surrounding distractors. As shown in Figure 7, in crowded scenes with visually similar objects (e.g., pandas, vehicles, billiards, and penguins), SAM2.1-B+ tracks well initially but later suffers

Fig. 6: Qualitative results of SAM-MT in highly dense scenarios ($\geq$ 10 targets). Our method achieves precise tracking and segmentation for all targets while maintaining real-time, near-single-object processing speed.

Fig. 7: Qualitative comparisons in identity-ambiguous scenarios with multiple visually similar targets. SAM-MT maintains robust temporal identity consistency, whereas SAM2 often suffers from tracking loss or identity swaps.

from either tracking loss or identity drift. In contrast, SAM-MT resolves these ambiguities and maintains consistent identities for all targets throughout.

Merging Targets into a Single Mask in SAM2. For VOS models optimized for single-target propagation, such as SAM2, a simple workaround for maintaining real-time single-object efficiency under multiple targets is to merge all targets into a unified mask. However, this not only discards individual IDs but also forms an unnatural shape, making the "target" difficult to track. As shown in Figure 8, merging three cars into one fails to form a meaningful object and thus leads to complete tracking loss. In contrast, SAM-MT tracks and segments individual targets accurately at real-time, near-single-object speed.

Fig. 8: To maintain a single-object efficiency budget, attempting to track the merged target causes SAM2 to easily lose track. In contrast, SAM-MT precisely tracks and segments individual targets with well-preserved IDs under the same efficiency budget.

Table 6: Ablation on Masked Attn. Strategies.

	Strategy	$\mathcal{J}\&\dot{\mathcal{F}}$	$\mathcal{J}$	$\dot{\mathcal{F}}$	FPS
(a)	Full Visibility	37.5	35.8	39.1	**37.1**
(b)	Full Masking	39.3	37.7	40.8	36.9
(c)	**Decoupled**	**43.0**	**41.4**	**44.5**	36.9

Table 7: Ablation on Identity Transformer.

Strategy	Depth	$\mathcal{J}\&\dot{\mathcal{F}}$	$\mathcal{J}$	$\dot{\mathcal{F}}$	FPS
w/o Masking	3	39.1	37.4	40.7	37.1
w/ Masking	1	41.9	40.3	43.4	**37.2**
	3	43.0	41.4	44.5	36.9
	5	**43.3**	**41.6**	**45.0**	36.4

Table 8: Ablation on Window Size.

	Window Size	$\mathcal{J}\&\dot{\mathcal{F}}$	$\mathcal{J}$	$\dot{\mathcal{F}}$	FPS
(a)	8	42.4	40.8	44.0	**37.1**
(b)	12	42.6	41.0	44.2	**37.1**
(c)	16	43.0	41.4	44.5	36.9
(d)	**32**	**43.1**	**41.6**	**44.6**	36.6

Table 9: Ablation on Training Strategies.

	Image Pretrain	Strided Sampling	Overlap Loss	$\mathcal{J}\&\dot{\mathcal{F}}$
(a)		✓	✓	40.3
(b)	✓		✓	41.7
(c)	✓	✓		42.5
(d)	✓	✓	✓	**43.0**

5.4 Ablation Studies

We conduct ablation studies on the MOSEv2 validation set [18]. Following [18], we report $\mathcal{J}\&\dot{\mathcal{F}}$ for segmentation accuracy and FPS for multi-target scalability when applicable. For a fair comparison, all ablation variants are trained with the same protocols described in Sections 4.1, 4.2, and 4.4.

Ablation on Masked Attention Strategies. We compare *decoupled masked attention* with two alternatives: (a) Full Visibility, where all target queries freely interact with each other, and (b) Full Masking, where global queries are completely isolated from target queries. As shown in Table 6(a), allowing unrestricted query interactions causes target queries to lose their identity specificity, leading to a 5.5-point drop in $\mathcal{J}\&\dot{\mathcal{F}}$ from 43.0 to 37.5. In contrast, Table 6(b) shows that isolating global context from target queries weakens global-individual information exchange, resulting in a 3.7-point decrease in $\mathcal{J}\&\dot{\mathcal{F}}$. These results validate the rationale behind our decoupled design, which preserves target-specific identities while enabling effective interactions with global context.

Ablation on Identity Transformer. Table 7 evaluates the identity-aware mask and transformer depth. Without the mask, each query attends to historical queries from all targets, causing cross-target memory pollution and a 3.9-point

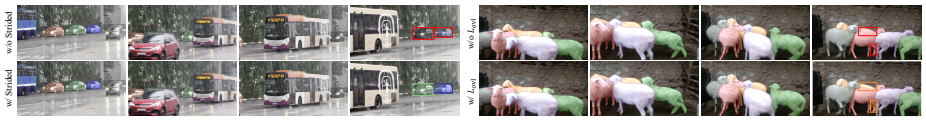


Fig. 9: *Left:* Strided sampling improves robustness to target reappearance. *Right:* Overlap prevention reduces pixel-level identity ambiguity in dense scenes.

Fig. 10: SAM-MT supports target recovery at any frame via simple click(s). (Zoom in for better view)

drop in $\mathcal{J}\&\dot{\mathcal{F}}$. Meanwhile, adding more transformer blocks improves accuracy but lowers FPS. We choose 3 blocks to balance speed and accuracy.

Ablation on Window Size of Sparse Memory. Table 8 evaluates the sparse-memory window size, where our lightweight queries enable SAM-MT to use a longer memory window than SAM2. As shown, increasing the window size from 8 to 32 improves $\mathcal{J}\&\dot{\mathcal{F}}$ by 0.7 points, but reduces FPS by 0.5 due to extra computation. We set window size to 16 for a better accuracy-speed trade-off.

Ablation on Training Strategies. Table 9 evaluates the training strategies. (a) Removing static image pretraining results in a 2.7-point drop in $\mathcal{J}\&\dot{\mathcal{F}}$, showing the necessity of image-level pretraining for one-shot multi-target image segmentation. (b) Without strided sampling (i.e., only using adjacent frames as in SAM2 [39]), $\mathcal{J}\&\dot{\mathcal{F}}$ decreases by 1.3 points; as shown in Figure 9 (left), this often causes identity loss upon target reappearance. (c) As shown in Figure 9 (right), overlap prevention alleviates mask conflicts in dense scenes (e.g., a group of moving sheep), ensuring clearer separation of individual targets.

5.5 Discussions

Target Recovery. SAM-MT allows users to recover tracking loss at any intermediate frame by providing click(s). As shown in Figure 10, three capybaras disappear in frame 2&3. Upon their reappearance in frame 4, one of them is lost from the track (*top*). In this case, a single click can recover the lost target for subsequent tracking and segmentation (*bottom*).

Limitations. To the best of our knowledge, SAM-MT is the first work to tackle the SAM family's multi-target bottleneck at the framework level, achieving near-single-object efficiency with competitive video segmentation accuracy. However, as it inherits SAM's vision-only architecture, it lacks the reasoning capability required for complex high-level tasks. Extending it with multimodal or reasoning [1, 31, 45, 46, 53, 54] is a promising future direction.

6 Conclusion

We present SAM-MT, the first framework to tackle the multi-target bottleneck of the SAM family for real-time interactive video segmentation. SAM-MT represents individual targets using lightweight queries in parallel with shared global context, enabling near-single-object efficiency as the number of targets increases. Extensive experiments show that SAM-MT preserves SAM2's strong segmentation performance across VOS benchmarks, while achieving real-time speed and robust performance in dense in-the-wild scenarios. We hope SAM-MT represents a solid step toward real-time interactive multi-target video segmentation.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62472104, the Science and Technology Commission of Shanghai Municipality under Grant No. 25511103600, and Shanghai Pujiang Program 2025PJA201.

References

1. Bai, Z., He, T., Mei, H., Wang, P., Gao, Z., Chen, J., Zhang, Z., Shou, M.Z.: One token to seg them all: Language instructed reasoning segmentation in videos. Advances in Neural Information Processing Systems **37**, 6833–6859 (2024)
2. Bekuzarov, M., Bermudez, A., Lee, J.Y., Li, H.: Xmem++: Production-level video segmentation from few annotated frames. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 635–644 (2023)
3. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
4. Cheng, B., Choudhuri, A., Misra, I., Kirillov, A., Girdhar, R., Schwing, A.G.: Mask2former for video instance segmentation. arXiv preprint arXiv:2112.10764 (2021)
5. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
6. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. Advances in neural information processing systems **34**, 17864–17875 (2021)
7. Cheng, H.K., Oh, S.W., Price, B., Lee, J.Y., Schwing, A.: Putting the object back into video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3151–3161 (2024)
8. Cheng, H.K., Oh, S.W., Price, B., Schwing, A., Lee, J.Y.: Tracking anything with decoupled video segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1316–1326 (2023)
9. Cheng, H.K., Schwing, A.G.: Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In: European conference on computer vision. pp. 640–658. Springer (2022)

10. Cheng, H.K., Tai, Y.W., Tang, C.K.: Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5559–5568 (2021)
11. Cheng, H.K., Tai, Y.W., Tang, C.K.: Rethinking space-time networks with improved memory coverage for efficient video object segmentation. Advances in neural information processing systems **34**, 11781–11794 (2021)
12. Cheng, T., Wang, X., Huang, L., Liu, W.: Boundary-preserving mask r-cnn. In: European conference on computer vision. pp. 660–676. Springer (2020)
13. Ding, H., Liu, C., He, S., Jiang, X., Loy, C.C.: MeViS: A large-scale benchmark for video segmentation with motion expressions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10156–10166 (2023)
14. Ding, H., Liu, C., He, S., Jiang, X., Torr, P.H., Bai, S.: MOSE: A new dataset for video object segmentation in complex scenes. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 20224–20234 (2023)
15. Ding, H., Liu, C., He, S., Ying, K., Jiang, X., Loy, C.C., Jiang, Y.G.: MeViS: A multi-modal dataset for referring motion expression video segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(12), 11400–11416 (2025)
16. Ding, H., Liu, C., Wang, S., Jiang, X.: Vision-language transformer and query generation for referring segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16321–16330 (2021)
17. Ding, H., Liu, C., Wang, S., Jiang, X.: VLT: Vision-language transformer and query generation for referring segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(6), 7900–7916 (2023)
18. Ding, H., Ying, K., Liu, C., He, S., Jiang, X., Jiang, Y.G., Torr, P.H., Bai, S.: MOSEv2: A more challenging dataset for video object segmentation in complex scenes. arXiv preprint arXiv:2508.05630 (2025)
19. Ding, S., Qian, R., Dong, X., Zhang, P., Zang, Y., Cao, Y., Guo, Y., Lin, D., Wang, J.: Sam2long: Enhancing sam 2 for long video segmentation with a training-free memory tree. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13614–13624 (2025)
20. Ghiasi, G., Gu, X., Cui, Y., Lin, T.Y.: Scaling open-vocabulary image segmentation with image-level labels. In: European conference on computer vision. pp. 540–557. Springer (2022)
21. Griffin, B., Florence, V., Corso, J.: Video object segmentation-based visual servo control and object depth estimation on a mobile robot. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1647–1657 (2020)
22. He, J., Li, P., Geng, Y., Xie, X.: Fastinst: A simple query-based model for real-time instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 23663–23672 (2023)
23. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
25. Hong, L., Chen, W., Liu, Z., Zhang, W., Guo, P., Chen, Z., Zhang, W.: Lvos: A benchmark for long-term video object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13480–13492 (2023)

26. Hong, L., Liu, Z., Chen, W., Tan, C., Feng, Y., Zhou, X., Guo, P., Li, J., Chen, Z., Gao, S., et al.: Lvos: A benchmark for large-scale long-term video object segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
27. Hu, L., Zhang, P., Zhang, B., Pan, P., Xu, Y., Jin, R.: Learning position and target consistency for memory-based video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4144–4154 (2021)
28. Jain, J., Li, J., Chiu, M.T., Hassani, A., Orlov, N., Shi, H.: Oneformer: One transformer to rule universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2989–2998 (2023)
29. Ke, L., Ye, M., Danelljan, M., Tai, Y.W., Tang, C.K., Yu, F., et al.: Segment anything in high quality. Advances in Neural Information Processing Systems **36**, 29914–29934 (2023)
30. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
31. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9579–9589 (2024)
32. Li, F., Zhang, H., Xu, H., Liu, S., Zhang, L., Ni, L.M., Shum, H.Y.: Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3041–3050 (2023)
33. Li, M., Hu, L., Xiong, Z., Zhang, B., Pan, P., Liu, D.: Recurrent dynamic embedding for video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1332–1341 (2022)
34. Liu, C., Ding, H., Jiang, X.: GRES: Generalized referring expression segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 23592–23601 (2023)
35. Liu, C., Jiang, X., Ding, H.: Primitivenet: decomposing the global constraints for referring segmentation. Visual Intelligence **2**(1), 16 (2024)
36. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
37. Oh, S.W., Lee, J.Y., Xu, N., Kim, S.J.: Video object segmentation using space-time memory networks. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9226–9235 (2019)
38. Park, K., Woo, S., Oh, S.W., Kweon, I.S., Lee, J.Y.: Per-clip video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1352–1361 (2022)
39. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. In: International Conference on Learning Representations. vol. 2025, pp. 28085–28128 (2025)
40. Seong, H., Oh, S.W., Lee, J.Y., Lee, S., Lee, S., Kim, E.: Hierarchical memory matching network for video object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12889–12898 (2021)
41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

42. Videnovic, J., Lukezic, A., Kristan, M.: A distractor-aware memory for visual object tracking with sam2. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24255–24264 (2025)
43. Wang, H., Jiang, X., Ren, H., Hu, Y., Bai, S.: Swiftnet: Real-time video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1296–1305 (2021)
44. Wang, J., Liu, D., Chen, J., Da, J., Qian, N., Tram, M.M., Soh, H.: Genie: A generalizable navigation system for in-the-wild environments. IEEE Robotics and Automation Letters (2025)
45. Weihua, Z., Huang, X., Liu, Z., Vangani, T.K., Zou, B., Tao, X., Wu, Y., Aw, A., Chen, N.F., Lee, R.K.W.: Adamcot: Rethinking cross-lingual factual reasoning through adaptive multilingual chain-of-thought. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 33863–33871 (2026)
46. Weihua, Z., Lee, R.K.W., Liu, Z., Kui, W., Aw, A., Zou, B.: Ccl-xcot: An efficient cross-lingual knowledge transfer method for mitigating hallucination generation. In: Findings of the Association for Computational Linguistics: EMNLP 2025. pp. 1768–1788 (2025)
47. Wu, W., Zhao, Y., Li, Z., Shan, L., Zhou, H., Shou, M.Z.: Continual learning for image segmentation with dynamic query. IEEE Transactions on Circuits and Systems for Video Technology **34**(6), 4874–4886 (2023)
48. Xie, H., Yao, H., Zhou, S., Zhang, S., Sun, W.: Efficient regional memory network for video object segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1286–1295 (2021)
49. Yang, C.Y., Huang, H.W., Chai, W., Jiang, Z., Hwang, J.N.: Samurai: Adapting segment anything model for zero-shot visual tracking with motion-aware memory. arXiv preprint arXiv:2411.11922 (2024)
50. Yang, Z., Wei, Y., Yang, Y.: Associating objects with transformers for video object segmentation. Advances in Neural Information Processing Systems **34**, 2491–2502 (2021)
51. Yang, Z., Yang, Y.: Decoupling features in hierarchical propagation for video object segmentation. Advances in Neural Information Processing Systems **35**, 36324–36336 (2022)
52. Ye, M., Oh, S.W., Ke, L., Lee, J.Y.: Entitysam: Segment everything in video. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24234–24243 (2025)
53. Yuan, H., Li, X., Zhang, T., Sun, Y., Huang, Z., Xu, S., Ji, S., Tong, Y., Qi, L., Feng, J., et al.: Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. arXiv preprint arXiv:2501.04001 (2025)
54. Zheng, W., Liu, Z., Chakraborty, T., Xu, W., Gao, X., Tan, B.C.Z., Zou, B., Liu, C., Hu, Y., Xie, X., et al.: Mma-asia: A multilingual and multimodal alignment framework for culturally-grounded evaluation. arXiv preprint arXiv:2510.08608 (2025)
55. Zhu, Y., Shi, C., Wang, D., Tang, J., Wei, Z., Wu, Y., Li, G., Yang, S.: Rethinking query-based transformer for continual image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4595–4606 (2025)