Tracking and Segmenting Anything in Any Modality

Tianlu Zhang[1], Qiang Zhang[2], Guiguang Ding[3], Jungong Han[1] *

[1]Department of Automation, Tsinghua University
[2]School of Mechano-Electronic Engineering, Xidian University
[3]School of Software, Tsinghua University.
jghan@tsinghua.edu.cn.

Abstract

Tracking and segmentation play essential roles in video understanding, providing basic positional information and temporal association of objects within video sequences. Despite their shared objective, existing approaches often tackle these tasks using specialized architectures or modality-specific parameters, limiting their generalization and scalability. Recent efforts have attempted to unify multiple tracking and segmentation subtasks from the perspectives of any modality input or multi-task inference. However, these approaches tend to overlook two critical challenges: the distributional gap across different modalities and the feature representation gap across tasks. These issues hinder effective cross-task and cross-modal knowledge sharing, ultimately constraining the development of a true generalist model. To address these limitations, we propose a universal tracking and segmentation framework named SATA, which unifies a broad spectrum of tracking and segmentation subtasks with any modality input. Specifically, a Decoupled Mixture-of-Expert (DeMoE) mechanism is presented to decouple the unified representation learning task into the modeling process of cross-modal shared knowledge and specific information, thus enabling the model to maintain flexibility while enhancing generalization. Additionally, we introduce a Task-aware Multi-object Tracking (TaMOT) pipeline to unify all the task outputs as a unified set of instances with calibrated ID information, thereby alleviating the degradation of task-specific knowledge during multi-task training. SATA demonstrates superior performance on 18 challenging tracking and segmentation benchmarks, offering a novel perspective for more generalizable video understanding.

Introduction

Video understanding has witnessed substantial expansion and now encompasses a wide range of tasks, including action recognition, video segmentation, object tracking, and etc. Among these, object tracking and segmentation are dedicated to establishing instance-level or pixel-level correspondences across frames, thereby laying the groundwork for tackling video tasks.

Over the years, the object tracking and segmentation have developed into multiple subtask branches, including Single

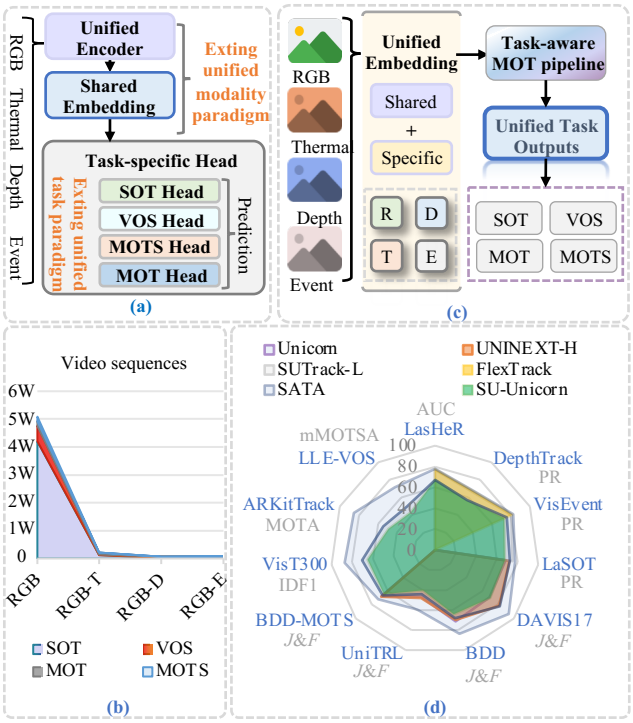

Figure 1: Analysis of data distribution gap and comparison of the proposed SATA with existing strategies. (a) Unified task and modality paradigm obtained by combining existing models. (b) Statistical overview of the data distribution gap during model training. (c) Overview of the proposed SATA framework. (d) Our SATA v.s. the existing methods on 11 challenging benchmarks. Here, SU-Unicorn denotes the combination of SUTrack (Chen, Kang et al. 2025) and Unicorn (Yan, Jiang et al. 2022), Flex-UNINEXT denotes the combination of FlexTrack (Tan, Shao et al. 2025) and UNINEXT (Yan, Jiang et al. 2023).

Object Tracking (SOT) (Huang, Zhao, and Huang 2019), Multiple Object Tracking (MOT) (Li, Ke et al. 2024), Video Object Segmentation (VOS) (Ravi et al. 2024) and Multi-Object Tracking and Segmentation (MOTS) (Yan, Jiang et al. 2023). In addition to the widely used RGB cameras, various sensors have been introduced to enhance the performance of tracking and segmentation in complex scenarios, e.g., depth, thermal, and event data. For a long period,

*Corresponding author: Jungong Han.

research in these typical tracking and segmentation subtasks has adopted a task- and modality-specific paradigm, i.e., designing specialized architectures and loss functions to cater to the unique requirements of different subtasks (e.g., SOT, MOT and VOS) and fixed input modalities (e.g., RGB, RGB-T and RGB-D). The divergent setups require customized methods with carefully designed architectures and hyper-parameters, leading to complex training and redundant parameters.

To mitigate this issue, some recent works have explored the possibility of establishing general models from two perspectives: unified task paradigm and unified modality paradigm. Specifically, the unified task paradigm (Yan, Jiang et al. 2023; Chen, Kang et al. 2025; Wang et al. 2024) breaks through the isolation of task domain knowledge and general knowledge, establishing the implicit collaboration between different tasks. Meanwhile, the unified modality methods (Wu, Zheng et al. 2024; Chen, Kang et al. 2025) consolidate the input of any modality using a unified model with shared parameters. The above two methods respectively establish unified modality representation learning and unified task feature space modeling, thereby reducing the complexity of model design and the need for extensive hyper-parameter tuning across various input modalities and subtask combinations.

Recently, due to the potential to realize Artificial General Intelligence (AGI), these unified models have drawn great attention. A natural idea is: combining the above two paths to achieve a more generalizable video tracking and segmentation model capable of handling any input modality and supporting multi-task inference, as shown in Fig. 1 (a). While this strategy is conceptually straightforward and partially effective, it overlooks the distributional discrepancies across modalities and the representational gaps between tasks—factors that limit the generalist model's ability to fully leverage cross-modal and cross-task knowledge.

Specifically, the distribution gap in multi-modal data arises not only from discrepancies in cross-modal information but also from differences in their underlying representations. Most existing unified modality approaches (Chen, Kang et al. 2025; Tan, Shao et al. 2025) have attempted to learn a cohesive embedding across diverse input modalities, but the effectiveness of modality-specific clues have not been fully exploited. Similarly, while generic representations can be learned with the unified task approaches, they are still constrained by elaborated-designed task-specific heads and isolated multi-stage training strategies across different downstream tasks (Yan, Jiang et al. 2022, 2023; Wang et al. 2024). These fragmented architectures and learning paradigms restrict the model's ability to acquire truly generalizable knowledge. Moreover, as shown in Fig. 1 (b), inconsistencies in the quality and scale of training data across different multi-modal subtasks further exacerbate task and modality biases in unified models.

With this in mind, we present SATA, a unified tracking and segmentation framework that models a unified representation of arbitrary modalities and approaches a broad spectrum of tracking and segmentation subtasks as MOT inference task conditioned on different priors, as shown in Fig. 1 (c). Technically, we introduce a Decoupled Mixture-of-Expert (DeMoE) mechanism to decouple the unified representation learning into parallel modeling of modality-common and modality-specific knowledge, thus achieving a more comprehensive representation of any modality input. In addition, unlike previous unified task models that employ several task-specific heads, we present a Task-aware MOT pipeline to unify all the task outputs (i.e., SOT, VOS, MOT, MOTS) as a unified set of instances with calibrated ID information, thereby alleviating the degradation of task-specific knowledge during multi-task training. With the unified model architecture and totally-shared parameters, SATA can solve 4 type of tracking and segmentation subtasks of 4 combinations of input modality. As shown in Fig. 1 (d), our proposed SATA significantly outperforms the simple combination of existing unified methods on all downstream tasks. We summarize that our work has the following contributions:

- To the best of knowledge, SATA is the first unified framework capable of performing both tracking and segmentation tasks in arbitrary modality input and multi-task joint prediction.
- We propose a Decoupled Mixture-of-Expert (DeMoE) mechanism and a Task-aware MOT pipeline to address the distribution gap in multi-modal data and the feature representation gap across tasks, enabling more effective cross-modality and cross-task knowledge sharing.
- We show superior results of our proposed method on 18 challenging benchmarks from 4 subtasks tasks, all using the same model architecture and parameter set.

Related works

Task- and modality-specific Methods. Over the years, a wide variety of task-specific models have been proposed to continuously improve the tracking and segmenting performance. SOT (Huang, Zhao, and Huang 2019) and VOS (Ravi et al. 2024) specify tracked objects on the first frame of a video using boxes or masks, then require algorithms to predict the trajectories of the tracked objects in boxes or masks, respectively. At present, Transformers based methods have mainstream SOT, which can be categorized into classification- and regression-based (Ye et al. 2022; Lin, Fan et al. 2024; Zhang, Jiao et al. 2024), corner prediction-based (Cui, Jiang et al. 2022; Cui et al. 2023), and sequence-learning-based trackers (Chen, Peng et al. 2023; Bai et al. 2024). Meanwhile, memory-based VOS approaches have become the dominant method in VOS (Ravi et al. 2024; Cheng and Schwing 2022; Zheng, Zhong et al. 2024). Different from SOT and VOS, MOT and MOTS aim to find and associate all instances (Yan, Jiang et al. 2022). The mainstream methods follow the tracking-by-detection paradigm (Yan, Jiang et al. 2022) and tracking-by-query (Chu, Wang et al. 2023) pipeline. With the development of sensor technology, tracking and segmentation tasks have evolved from using only a RGB camera to introduce various auxiliary modalities (Chen, Kang et al. 2025; Gao et al. 2022; Zhang et al. 2025b).

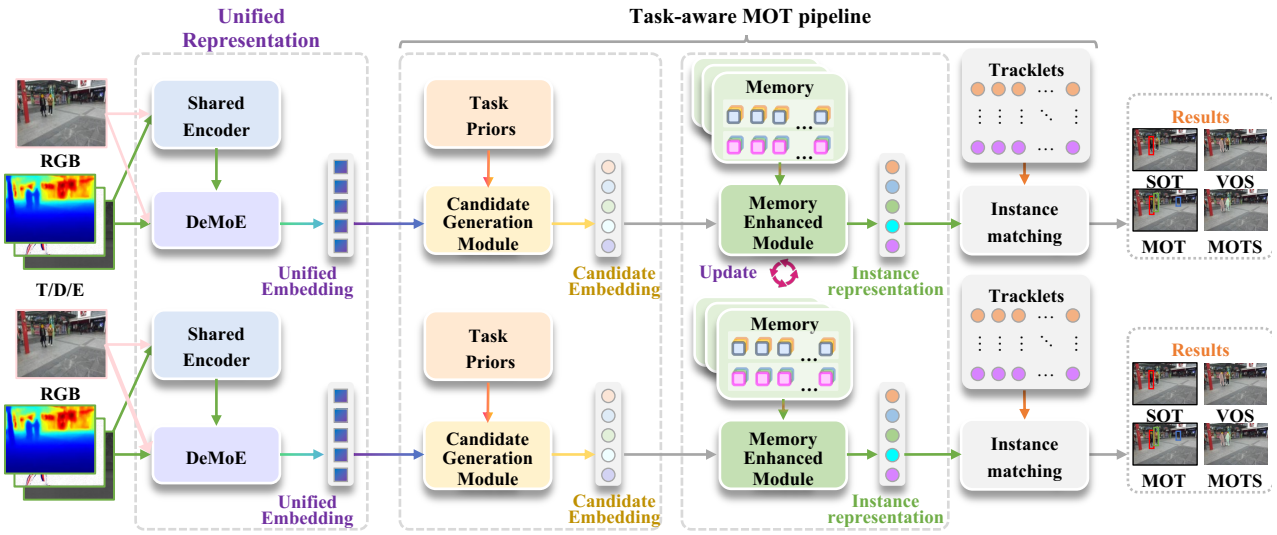

Figure 2: Overview architecture of our SATA, which consists of two core components: the Decoupled Mixture-of-Expert mechanism and the Task-aware MOT pipeline.

Unified Modality Methods. Despite the success of previous multi-modal methods tailored for modality-specific input, the inherent multi-parameter set paradigm limits the flexibility of these models in practical application scenarios. To echo this problem, some works (Zhu et al. 2023; Hong, Yan et al. 2024; Hou, Xing et al. 2024; Feng, Zhang et al. 2025) adopt a architecture-shared design for RGB-X tracking, only activating the modality-specific parameters according to the input modality. Besides, some unified methods (Wu, Zheng et al. 2024; Chen, Kang et al. 2025; Tan, Wu et al. 2025; Tan, Shao et al. 2025) have been proposed to learn the common latent space of any modality input, aiming at achieving more complete unification. *However, the overlook of distribution gap and the lack of multi-task inference capabilities remains incomplete for a powerful unified tracking and segmentation model.*

Unified Task Methods. Meanwhile researchers have undertaken prominent efforts to unify tracking and segmentation tasks within specific modality. Existing unified task methods can be categorized into two main branches: the prompt based methods and the detection based methods. The prompt based methods (Ma, Shou et al. 2022; Yan, Jiang et al. 2022; Wang et al. 2024) apply a shared appearance model for unified representation learning, and employ the delicately designed target prior to solve multiple subtasks. Besides, the detection based methods (Li, Ke et al. 2024; Yan, Jiang et al. 2023; Wang et al. 2025) introduce an external detectors to predict objects and establish the associations for tracking and segmentation. *However, these methods pay less attention to the multi-task gap, result in the degradation of task-specific knowledge during multi-task training.*

Method

Overall Pipeline

The proposed framework, termed SATA, comprises two core components: the Decoupled Mixture-of-Expert (DeMoE) mechanism and the Task-aware MOT (TaMOT) pipeline, as shown in Fig. 2. The DeMoE, combined with a powerful Transformer-based encoder, first yields the unified representation. Subsequently, the designed Task-aware MOT pipeline first generates tracking candidates based on task-specific prior knowledge, then maintains tracking of all candidates, thereby unifying all tracking and segmentation subtasks under the MOT paradigm.

Decoupled Mixture-of-Expert

Given the input video frames of the RGB (R) modality and their corresponding auxiliary modalities (thermal, depth, or event modalities, collectively referred to as TDE), a weight-shared Transformer-based encoder is employed to generate RGB tokens and TDE tokens. To enable the unified model to handle diverse input modalities, the proposed DeMoE is used to replace the feed-forward network (FFN) in each Transformer encoder layer, thereby converting different modality combinations into a unified token embedding format. DeMoE consists of three primary components: a Common-prompt Mixture of Expert (CpMoE), a Specific-activated Mixture of Expert (SaMoE), and Decoupling Learning.

Common-prompt Mixture of Expert. At the l-th Transformer encoder layer, our CpMoE takes the RGB token T_l^{R} and TDE token T_l^{TDE} from the output of the Multi-head Attention (MSA) block as input. It comprises a general router network $\mathcal{R}^{\mathrm{G}}$, a shared expert g^{S}, and N^{G} modality-common experts $\mathcal{G}^{\mathrm{G}} = \{g_1^{\mathrm{C}}, ..., g_{N^{\mathrm{G}}}^{\mathrm{C}}\}$. Specifically, as shown in Fig. 3 (a), the shared expert g^{S} directly copies weights from the corresponding FFN layer of the encoder and remains frozen during training to retain pre-trained general knowledge, yielding TG_l^{G} and TG_l^{TDE}, i.e.,

$$TG_l^{\mathrm{G}} = g^{\mathrm{S}}(T_l^{\mathrm{R}}), \quad TG_l^{\mathrm{TDE}} = g^{\mathrm{S}}(T_l^{\mathrm{TDE}}). \tag{1}$$

Then, the modality-common experts g_n^{C} $(n = 1, ..., N^{\mathrm{G}})$ are employed to generate modality-common prompts, transforming the general knowledge to be more suitable for the

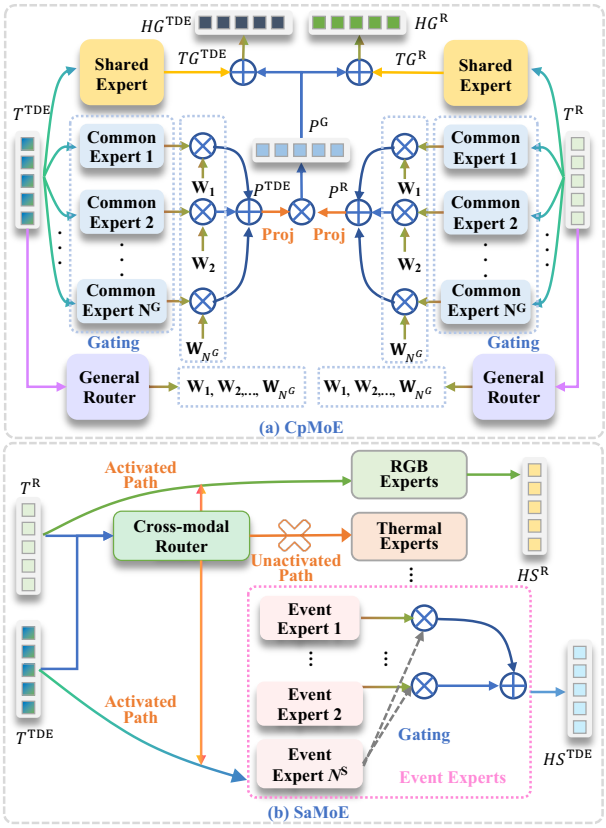

Figure 3: Overview architecture of our CpMoE and SaMoE. (a) CpMoE.(b) SaMoE.

current input modalities. Each modality-common expert g_n^{C} consists of two linear layers and a GELU activation layer. The generated RGB prompt P_l^{G} and TDE prompt P_l^{TDE} are weighted sums of outputs from the top-K activated experts, i.e.,

$$P_l^{\mathrm{R}} = \sum_{n=1}^{N^{\mathrm{G}}} p_n^{\mathrm{R}} g_n^{\mathrm{C}}(T_l^{\mathrm{R}}), \quad P_l^{\mathrm{TDE}} = \sum_{n=1}^{N^{\mathrm{G}}} p_n^{\mathrm{TDE}} g_n^{\mathrm{C}}(T_l^{\mathrm{TDE}}), \tag{2}$$

where p_n^{R} and p_n^{TDE} denote the gating values derived from the function $\mathcal{R}^{\mathrm{G}}$.

Subsequently, the modality-common prompt is obtained by performing element-wise multiplication on the RGB prompt and TDE prompt, which can be formulated as:

$$P_l^{\mathrm{G}} = \mathrm{proj}(P_l^{\mathrm{R}}) \otimes \mathrm{proj}(P_l^{\mathrm{TDE}}), \tag{3}$$

where $\otimes$ denotes the element-wise multiplication operation, and $\mathrm{proj}(*)$ is the projection layer.

Finally, the modality-common prompt is employed to map TG_l^{G} and TG_l^{TDE} into a cohesive token representation through element-wise addition, i.e.,

$$HG_l^{\mathrm{R}} = P_l^{\mathrm{G}} \oplus TG_l^{\mathrm{R}}, \quad HG_l^{\mathrm{TDE}} = P_l^{\mathrm{G}} \oplus TG_l^{\mathrm{TDE}}, \tag{4}$$

where HG_l^{R} and HG_l^{TDE} refer to the prompted tokens for the RGB and TDE modalities, respectively, and $\oplus$ denotes the element-wise addition operation.

Specific-activated Mixture of Expert. SaMoE is designed to capture modality-specific clues within multi-modal data, comprising intra-modal router networks ($\mathcal{R}^{\mathrm{R}}$ and $\mathcal{R}^{\mathrm{TDE}}$), a cross-modal router network $\mathcal{R}^{\mathrm{CM}}$, and several modality-specific experts ($\mathcal{G}^X$=$\{g_1^X, ..., g_{N^{\mathrm{S}}}^X\}(X = \mathrm{R}, \mathrm{TDE})$, where N^{S} denotes the number of modality-specific experts.

As shown in Fig. 3 (b), the cross-modal router network $\mathcal{R}^{\mathrm{CM}}$ activates the specific modality branch, thereby associating the current input modality with the corresponding modality experts. The modality-specific RGB tokens HS_l^{R} and TDE tokens HS_l^{TDE} are weighted sums of outputs from the top-K activated experts for each modality, i.e.,

$$HS_l^X = \sum_{n=1}^{N^{\mathrm{S}}} s_n^X g_n^X(T_l^X), \quad X = \mathrm{R}, \mathrm{TDE}, \tag{5}$$

where s_n^X represents the gating values derived from the function $\mathcal{R}^X$.

Unified Representation. The RGB feature representation F_l^{R} and TDE feature representation F_l^{TDE} are obtained by aggregating these two types of embeddings from CpMoE and SaMoE with residual connections:

$$F_l^X = HG_l^X \oplus HS_l^X \oplus T_l^X, \quad X = \mathrm{R}, \mathrm{TDE}. \tag{6}$$

In the last layer ($L-th$ layer) of the encoder, the final unified feature representation F^{U} is obtained as:

$$F^{\mathrm{U}} = F_L^{\mathrm{R}} \oplus F_L^{\mathrm{TDE}}. \tag{7}$$

Decoupling Learning. To comprehensively model multi-modal representations, the learning processes of CpMoE and SaMoE adhere to two key principles: (a) promoting mutual complementarity among cross-modal experts; (b) avoiding information overlap between specific experts and common experts. Accordingly, we introduce two critical loss functions that act on the DeMoE.

Cross-modal complementary learning: To fully explore the complementary information across modalities, we randomly mask patches of one modality by assigning their values to a learnable token vector, generating masked features $\hat{HG}_l^{\mathrm{R}}$ and $\hat{HG}_l^{\mathrm{TDE}}$. The cross-modal complementary learning loss is then defined as:

$$\mathcal{L}_{\mathrm{CM}} = \sum_{l=1}^{L} \mathrm{MSE}(\hat{HG}_l^{\mathrm{R}}, HG_l^{\mathrm{R}}) + \sum_{l=1}^{L} \mathrm{MSE}(\hat{HG}_l^{\mathrm{TDE}}, HG_l^{\mathrm{TDE}}), \tag{8}$$

where $\mathrm{MSE}(\cdot)$ denotes the mean squared error.

Cross-expert orthogonal learning: Additionally, we introduce an orthogonal loss to encourage independence between common and specific expert representations. This cross-expert orthogonal loss is defined as:

$$\mathcal{L}_{\mathrm{CE}} = \sum_{l=1}^{L} \mathrm{OPL}(\sum_{j=1}^{N^{\mathrm{S}}} \sum_{i=1}^{N^{\mathrm{G}}} g_i^{\mathrm{C}}(T_l^X), g_j^X(T_l^X)), X = \mathrm{R}, \mathrm{TDE}. \tag{9}$$

Here, $g_i^{\mathrm{C}}(*)$ and $g_j^X(*)$ represents the outputs of experts from $\mathcal{G}^{\mathrm{G}}$ and $\mathcal{G}^X$, respectively, $\mathrm{OPL}(*)$ denotes the orthogonal loss (Ranasinghe et al. 2021).

Task-aware MOT pipeline

To achieve the grand unification of tracking and segmentation while mitigating the representation gap across task domains, the proposed Task-aware MOT pipeline integrates all

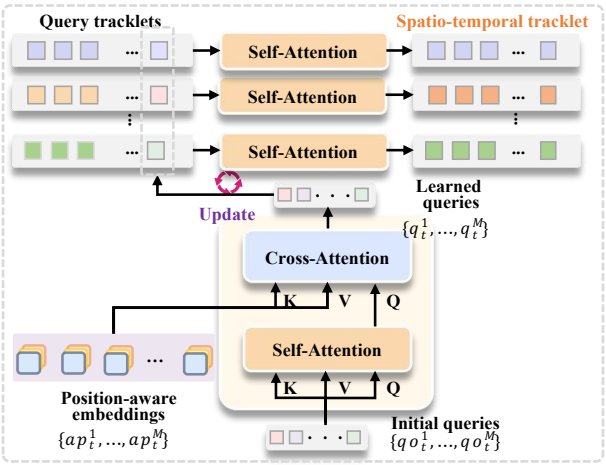

Figure 4: Illustration of the spatiotemporal relationship modeling.

subtasks into the localization and association of multiple objects. Firstly, a Candidates Generation Module (CGM) is utilized to generate potential proposals based on task-specific prior information. Subsequently, the proposed Memory-enhanced Module (MEM) refines the historical features of each candidate, enabling efficient and effective modeling of temporal information. Finally, a simple bi-softmax nearest neighbor search is employed to achieve accurate matching between candidates and trajectories.

Candidates Generation Module. Instead of introducing additional detectors to detect all potential targets in each frame (Yan, Jiang et al. 2023; Wang et al. 2025), we adopt a modified SAM2 (Ravi et al. 2024) as our foundation model to predict candidates based on the prior information of various tasks. Specifically, for SOT and VOS tasks, the initial target information (bounding box or mask) from the first frame can be used to generate the initial prompt token. In subsequent frames, masks predicted by the mask decoder with an affinity score exceeding a preset threshold are regarded as distractors; both the target and these distractors are defined as candidates. For MOT and MOTS tasks, we append a detection head to the foundation model to predict all potential objects in each frame. The box predictions are fed as multi-object prompts to the mask decoder, generating masks for all candidates. Once the positions of candidates in each frame are determined, we extract instance-level features by applying RoI Align. Specifically, given the unified embedding F_t^{U} of $t-th$ frame and candidate coordinate B_t^m corresponding to the $m-th$ tracklet, the corresponding candidate embedding can be acquired via $a_t^m = \mathrm{RoIAlign}(F_t^{\mathrm{U}}, B_t^m)$.

Memory-enhanced Module. The memory-enhanced module comprises two components: (a) generating fine-grained instance embeddings; (b) effectively modeling spatiotemporal relationship, as shown in Fig. 4.

Fine-grained instance embeddings: Candidate embeddings typically contain both foreground and background information at the edges, which leads to inaccurate instance matching when target objects cannot be clearly distinguished from the background. To echo this problem, given

RGB SOT Method	GOT10K			LaSOT		
	AO	$SR_{0.5}$	$SR_{0.75}$	AUC	P_{Norm}	P
Task-specific methods						
GRM (Gao, Zhou, and Zhang 2023)	73.4	82.9	70.4	69.9	79.3	75.8
LMTrack (Xu, Zhong et al. 2025)	80.1	**91.5**	79.0	73.2	83.4	81.0
TemTrack (Xie et al. 2025)	76.1	84.9	74.4	73.1	83.0	80.7
MambaLCT (Li et al. 2025)	76.2	86.7	74.3	73.6	84.1	81.6
SPMTrack-L (Cai et al. 2025)	80.0	89.4	79.9	76.8	**85.9**	84.0
MCITrack-B (Kang, Chen et al. 2025)	77.9	88.2	76.8	75.3	85.6	83.3
DreamTrack-L (Guo et al. 2025)	79.9	88.5	80.3	76.6	85.6	83.1
SAM2.1++ (Videnovic et al. 2025)	81.1	-	-	75.1	-	-
Unified modality methods						
SUtrack-L (Chen, Kang et al. 2025)	81.5	89.5	83.3	75.2	84.9	83.2
Unified task methods						
UTT (Ma, Shou et al. 2022)	67.2	76.3	60.5	64.6	-	67.2
Unicorn (Yan, Jiang et al. 2022)	-	-	-	68.5	76.6	74.1
UNINEXT-H (Yan, Jiang et al. 2023)	-	-	-	72.2	80.7	79.4
OmniViD (Wang et al. 2024)	-	-	-	70.8	79.6	76.9
SAM2.1 (Ravi et al. 2024)	80.7	-	-	70.5	-	-
OmniTracker-L(Wang et al. 2025)	-	-	-	69.1	77.3	75.4
SATA	**81.3**	91.4	**83.7**	**77.3**	85.7	**84.6**

Table 1: SOTA comparisons on RGB SOT.

RGB-T SOT Method	LasHeR		RGBT234	
	PR	SR	PR	SR
Modality-specific methods				
TBSI+ (Li et al. 2024a)	75.5	59.6	91.0	67.0
TransTIH (Zhang et al. 2024b)	72.0	57.2	89.4	66.4
CAFormer (Xiao et al. 2025)	70.0	55.6	88.3	66.4
AETrack (Zhu, Zhong et al. 2025)	74.7	59.6	91.6	68.8
DMD (Hu et al. 2025b)	72.6	57.6	89.3	66.7
Architecture-shared methods				
SDSTrack (Hou, Xing et al. 2024)	66.5	53.1	84.8	62.5
OneTrack (Hong, Yan et al. 2024)	67.2	53.8	85.7	64.2
GMMT (Tang et al. 2024)	70.7	56.6	87.9	64.7
CSTrack (Feng, Zhang et al. 2025)	75.6	60.8	94.0	70.9
STTrack (Hu et al. 2025a)	76.0	60.3	89.8	66.7
Unified modality methods				
UnTrack (Wu, Zheng et al. 2024)	-	-	84.2	62.5
FlexTrack (Tan, Shao et al. 2025)	77.3	-	92.7	69.9
XTrack (Tan, Wu et al. 2025)	73.1	58.7	87.8	65.4
SUTrack-L (Chen, Kang et al. 2025)	76.9	-	93.7	70.3
SATA	**77.8**	**61.7**	**94.3**	**71.5**

Table 2: SOTA comparisons on RGB-T SOT.

the candidate embedding a_t^m, candidate coordinate B_t^m, and their corresponding mask prediction $mask_t^m$, the adjusted mask $\hat{mask}_t^m$ is first generated by down-sampling the predicted mask $mask_t^m$. Then, the fine-grained instance embedding ae_t^m is generated by:

$$ae_t^m = \mathrm{ConV}(\mathrm{ConV}(\hat{mask}_t^m) \oplus a_t^m), \tag{10}$$

where $\mathrm{ConV}(*, \theta_n)$ denotes the convolutional layer.

Spatiotemporal relationship modeling: We first employ several MLP layers on the normalized coordinate $\hat{B}_t^m$ to generate the positional embedding p_t^m. The position-aware embedding can be represented as $ap_t^m = p_t^m \oplus a_t^m$. Then, we introduce the learnable queries q_t^m to capture video spatiotemporal information via the same architecture as the Querying Transformer (Q-Former) (Li et al. 2023). Specifically, the initial queries $\{qo_t^1, qo_t^2, ..., qo_t^M\}$ are fed into the self-attention layer to model spatial interactions. After that, the interacted queries and the position-aware embeddings $\{ap_t^1, ap_t^2, ..., ap_t^M\}$ are fed into the cross-attention layer, obtaining the $t-th$ learned query $\{q_t^1, q_t^2, ..., q_t^M\}$. Finally, we construct the temporal correlations on each query tracklet $\{q_{t-1}^m, q_{t-2}^m, ..., q_{t-T}^m\}(m = 1, ..., M)$ via a single self-attention layer, generating the spatiotemporal tracklet $\{qe_{t-1}^m, qe_{t-2}^m, ..., qe_{t-T}^m\}$.

Finally, the comprehensive representation f_t^M of the $M-th$ instances in $t-th$ frame can be transformer by concatenating the fine-grained instance embedding ae_t^M and the

RGB-D SOT Method	DepthTrack		VOTRGBD	
	PR	Re	EAO	Acc
STTrack (Hu et al. 2025a)	63.2	63.4	-	-
CSTrack (Feng, Zhang et al. 2025)	65.2	66.4	77.4	83.3
OneTrack (Hong, Yan et al. 2024)	60.7	60.4	72.7	81.9
UnTrack (Wu, Zheng et al. 2024)	61.3	61.0	72.1	81.5
FlexTrack (Tan, Shao et al. 2025)	67.1	66.9	78.0	83.8
XTrack-L (Tan, Wu et al. 2025)	65.4	64.3	74.0	82.8
SUTrack-L (Chen, Kang et al. 2025)	66.5	66.4	76.6	83.5
SATA	**67.9**	**67.6**	**78.4**	**84.1**

Table 3: SOTA comparisons on RGB-D SOT.

RGB-E SOT Method	VisEvent		COESOT	
	PR	AUC	PR	SR
CSAM (Zhang et al. 2024a)	81.6	-	76.7	68.3
STTrack (Hu et al. 2025a)	78.6	61.9	-	-
CSTrack (Feng, Zhang et al. 2025)	82.4	65.2	77.4	83.3
SDSTrack (Hou, Xing et al. 2024)	76.7	59.7	-	-
OneTrack (Hong, Yan et al. 2024)	76.7	60.8	-	-
FlexTrack (Tan, Shao et al. 2025)	81.4	64.1	-	-
XTrack-L (Tan, Wu et al. 2025)	80.5	63.3	-	-
SUTrack-L (Chen, Kang et al. 2025)	80.5	63.8	-	-
UnTrack (Wu, Zheng et al. 2024)	76.3	58.7	-	-
SATA	**82.8**	**66.7**	**80.4**	**71.6**

Table 4: SOTA comparisons on RGB-E SOT.

learned query q_t^M, and the comprehensive tracklet can be represented by $\Gamma^n = \{f_{t-1}^n, f_{t-2}^n, ..., f_{t-T}^n\}(n = 1, ..., N)$.

Instance matching. Given the instance f_t^m and the tracklet Γ^n generated by the candidate generation module and the memory-enhanced module, we can obtain the final assignment matrix $A \in \mathbb{R}^{N \times M}$ via bi-softmax nearest neighbor search, as illustrated in the Appendix.[1] Finally, we match the $\epsilon - th$ tracklet with $m - th$ instances via $\text{argmax}(s(f_t^m, \Gamma^1), ..., s(f_t^m, \Gamma^N)) = s(f_t^m, \Gamma^\epsilon)$.

Experiments

Implementation Details

The SATA model is developed using Python 3.8 and PyTorch 1.11. The training process leverages 8 NVIDIA A100 GPUs, while the inference speed is evaluated on a single NVIDIA 3090TI GPU.

Architecture. Our model uses HiViT-L (Zhang, Tian et al. 2022) as the transformer encoder, and we select SAM2 (Ravi et al. 2024) as our foundation model in the Task-aware MOT pipeline. The transformer encoder and the foundation model are initialized with the pre-trained parameters of SAM2 (Ravi et al. 2024).

Training. During the stage of candidate generation, the affinity mask scores, object predictions, and IoU predictions of the mask decoder are optimized by MAE loss, cross-entropy loss, and L_1 loss, respectively (Ravi et al. 2024). In addition, the decoupled MoE loss $L_{\text{MoE}} = \mu L_{\text{CM}} + \lambda L_{\text{CE}}$ in the DeMoE is employed to promote comprehensive learning of the unified embedding. In the stage of instance matching, the partial supervision loss and self-supervised loss in KeepTrack (Mayer et al. 2021) are employed to supervise the assignment matrix A for SOT and VOS datasets, and the cross-entropy loss is applied to optimize SATA for MOT and MOTS datasets.

[1]Please find the appendix in our arXiv version: arxiv.org/abs/2511.19475

RGB MOT Method	BDD		DanceTrack	
	mMOTA	MOTA	HOTA	MOTA
Task-specific methods				
DiffMOT (Lv et al. 2024)	-	-	62.3	**92.8**
Hybrid-SORT (Yang et al. 2024)	-	-	65.7	91.8
Unified task methods				
Unicorn (Yan, Jiang et al. 2022)	41.2	66.6	-	-
UNINEXT (Yan, Jiang et al. 2022)	44.2	67.1	-	-
MASA-SAM-H (Li, Ke et al. 2024)	44.5	-	-	-
SAM2MOT-Co (Jiang et al. 2025)	-	57.5	75.5	89.2
SATA	**46.3**	**67.8**	**76.1**	90.5

Table 5: SOTA comparisons on RGB MOT

RGB-T MOT Method	HOTA	DetA	AssA	MOTA	IDF1
Task-specific methods					
Bytetrack (Zhang, Sun et al. 2022)	53.4	58.2	49.6	72.4	64.9
MOTRv2 (Zhang et al. 2023)	53.1	51.2	**55.8**	63.3	64.4
HGT-Track (Xu, Wang et al. 2024)	54.0	61.3	48.1	71.1	60.9
Unified task methods					
Unitrack (Wang et al. 2021)	46.9	46.1	48.4	57.1	57.4
Unicorn (Yan, Jiang et al. 2022)	49.4	48.3	51.2	59.0	60.6
UnisMOT (Zhang et al. 2025a)	54.2	59.5	49.7	65.7	60.3
SATA	**59.7**	**63.6**	53.7	**75.1**	**65.4**

Table 6: SOTA comparisons on RGB-T MOT

RGB VOS Method	DAVIS 2016 val			DAVIS 2017 val		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Task-specific methods						
XMem (Cheng and Schwing 2022)	92.0	90.7	93.2	87.7	84.0	91.4
OneVOS (Li, Guo et al. 2024)	92.7	91.0	94.3	88.5	84.6	92.4
M3-VOS (Chen, Li et al. 2024)	-	-	-	-	86.0	-
SAM2 (Ravi et al. 2024)	-	-	-	88.9	-	-
TAM-S (Ravi et al. 2024)	-	-	-	89.2	-	-
Unified task methods						
UniTrack (Yan, Jiang et al. 2022)	-	-	-	-	58.4	-
Unicorn-T (Yan, Jiang et al. 2022)]	83.2	83.0	83.4	64.5	62.7	66.3
Unicorn-ConvL (Yan, Jiang et al. 2022)	87.4	86.5	88.2	69.2	65.2	73.2
UNINEXT-R50 (Yan, Jiang et al. 2023)	-	-	-	74.5	71.3	77.6
UNINEXT-H (Yan, Jiang et al. 2023)	-	-	-	81.8	77.7	85.8
OmniTracker-T (Wang et al. 2025)	84.7	84.1	85.3	66.2	64.9	67.5
OmniTracker-L (Wang et al. 2025)	88.5	87.3	89.7	71.0	66.8	75.2
SATA	**93.4**	**91.6**	**95.2**	**89.7**	**86.1**	**93.0**

Table 7: Comparisons on RGB VOS.

State-of-the-Art Comparisons

We compare SATA with state-of-the-art methods on 18 large-scale benchmarks with 4 types of input (i.e., RGB, RGB-T, RGB-D, RGB-E) and 4 subtasks (i.e., SOT, VOS, MOT and MOTS).

SOT. The results of RGB SOT are presented in Tab. 1. Our model achieves 81.3% AO and 77.3% AUC on GOT10K (Huang, Zhao, and Huang 2019) and LaSOT (Fan, Lin et al. 2019), respectively, surpassing the recent RGB tracker SAM2.1++ (Videnovic et al. 2025), SUTrack (Chen, Kang et al. 2025), and LMTrack (Xu, Zhong et al. 2025) by 0.2%/2.2%, 0.3%/2.1%, and 1.2%/4.1%, respectively. In addition, compared with existing unified task models, e.g., UTT (Ma, Shou et al. 2022), Unicorn (Yan, Jiang et al. 2022), and OmniTracker (Wang et al. 2025), SATA achieves performance gains of 12.7%, 8.8%, and 7.9% in AUC on LaSOT, respectively. Besides, SATA sets a new state-of-the-art on 6 multi-modal SOT benchmarks, as illustrated in Tab. 2, Tab. 3, and Tab. 4. On LasHeR (Li et al. 2022), DepthTrack (Yan et al. 2021b), and VisEvent (Wang, Li et al. 2023), SATA surpassing the recent best unified modality trackers, XTrack (Tan, Wu et al. 2025), SUTrack (Chen, Kang et al. 2025) and FlexTrack (Tan, Shao et al. 2025), by 4.7%/2.5%/2.3%, 0.9%/1.4%/2.3%, and 0.5%/0.8%/1.4% in PR, respectively.

MOT. We report the results of RGB MOT in Tab. 5. Our

RGB-T VOS Method	VisT300			VTUAV		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Modality-specific methods						
STM (Oh, Lee et al. 2019)	60.4	57.9	62.8	-	-	-
STCN (Cheng et al. 2021)	71.4	74.4	73.8	65.5	61.0	69.9
STCN-T (Cheng et al. 2021)	72.3	-	-	-	-	-
TBD (Cho, Lee et al. 2022)	70.5	68.3	72.6	-	-	-
CFBI+ (Yang et al. 2021)	74.1	71.8	76.4	-	-	-
AlpahRefine (Yan et al. 2021a)	-	-	-	65.9	59.9	71.9
AOT (Yang, Wei et al. 2021)	76.8	74.0	79.6	-	-	-
AOT-B (Yang, Wei et al. 2021)	-	-	-	81.8	77.7	85.8
AOT-L (Yang, Wei et al. 2021)	-	-	-	82.0	77.5	86.5
XMem (Cheng and Schwing 2022)	75.7	73.3	78.0	69.1	65.1	73.1
XMem-T (Cheng and Schwing 2022)	77.9	-	-	-	-	-
ViTNet (Yang et al. 2023)	81.9	79.2	84.5	76.7	72.9	80.8
Architecture-shared methods						
X-Prompt (Guo, Li et al. 2024)	84.2	81.7	86.7	87.3	82.8	91.8
SATA	**87.4**	**84.5**	**90.3**	**88.5**	**84.4**	**92.6**

Table 8: Comparisons on RGB-Thermal VOS.

RGB-D/E VOS Method	ARKitTrack			VTUAV		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Modality-specific methods						
STCN (Cheng et al. 2021)	53.7	49.8	57.5	47.4	45.0	49.8
ARKitVOS (Zhao et al. 2023)	66.2	62.5	69.8	-	-	-
XMem (Cheng and Schwing 2022)	71.6	68.5	74.6	54.2	59.0	49.4
DeAOT (Yang and Yang 2022)	72.6	70.0	75.3	-	-	-
AOT-L-Swin (Yang, Wei et al. 2021)	77.8	75.0	80.7	-	-	-
AOT-B (Yang, Wei et al. 2021)	-	-	-	62.3	64.9	59.6
DeAOT (Yang and Yang 2022)	-	-	-	63.3	65.3	61.4
LLE-VOS (Li et al. 2024b)	-	-	-	67.8	70.2	65.4
X-Prompt (Guo, Li et al. 2024)	82.1	79.4	84.9	-	-	-
SATA	**85.1**	**82.8**	**87.4**	**71.4**	**73.6**	**69.1**

Table 9: Comparisons on RGB-Depth/Event VOS.

SATA achieves the best MOTA and HOTA scores on DanceTrack (Sun, Cao et al. 2022) and BDD (Yu et al. 2020). Specifically, it outperforms all previous unified task methods, achieving an mMOTA score of 67.8% on BDD, surpassing Unicorn (Yan, Jiang et al. 2022), UNINEXT-H (Yan, Jiang et al. 2023) and SAM2MOT-Co (Jiang et al. 2025) by 1.2%, 0.7%, and 10.3%, respectively. In addition, as shown in Tab. 6, SATA achieves state-of-the-art performance on UniRTL (Zhang et al. 2025a) with RGB-T input, outperforming the previous best method UnisMOT (Zhang et al. 2025a) by 5.3% in HOTA.

VOS. We present a comparison with recent advanced RGB VOS methods in Table 7, reporting accuracy using standard protocols. SATA shows significant improvement over the best existing methods, achieving the highest scores of 93.4% and 89.7% $\mathcal{J}\&\mathcal{F}$ on these DAVIS datasets (Pont-Tuset et al. 2017). Besides, as presented in Table 8 and Table 9, SATA secures the top positions on the RGB-T (Yang et al. 2023; Zhang et al. 2022), RGB-D (Zhao et al. 2023), and RGB-E VOS (Li et al. 2024b) benchmarks.

MOTS. We evaluate SATA's MOTS capability on BDD MOTS (Yu et al. 2020) As shown in Table 10, our approach outperforms existing advanced unified models, i.e., Unicorn (Yan, Jiang et al. 2022) and UNINEXT-H (Yan, Jiang et al. 2023) by noteworthy margins of 8.5% and 2.4% in mMOTSA, respectively.

Ablation Studies

In this section, we conduct component-wise analysis for a better understanding of our method. The methods are evaluated on 5 benchmarks (GOT10K (Huang, Zhao, and Huang 2019), LasHeR (Li et al. 2022), LLE-VOS (Li et al. 2024b), UniRTL (Zhang et al. 2025a), and BDD MOTS (Yu et al. 2020)) from 4 subtasks (SOT, VOS, MOT, MOTS).

RGB MOTS Method	mMOTSA	mMOTSP	mIDF1
Task-specific methods			
MaskTrackRCNN (Fu, Liu et al. 2021)	12.3	59.9	26.2
STEm-Seg (Athar et al. 2020)	12.2	58.2	25.4
QDTrack-mots (Huang et al. 2023)	22.5	59.6	40.8
PCAN (Ke et al. 2021)	27.4	66.7	45.1
VMT (Ke et al. 2022)	28.7	67.3	45.7
MASA-B (Li, Ke et al. 2024)	35.2	-	49.2
MASA-H (Li, Ke et al. 2024)	35.8	-	49.7
Unified task methods			
Unicorn (Yan, Jiang et al. 2022)	29.6	67.7	44.2
UNINEXT-L (Yan, Jiang et al. 2023)	32.0	60.2	45.4
UNINEXT-H (Yan, Jiang et al. 2023)	35.7	68.1	48.5
SATA	**38.1**	**72.3**	**52.4**

Table 10: SOTA comparisons on RGB MOTS tasks.

Method	SOT (AO)	SOT (PR)	VOS ($\mathcal{J}\&\mathcal{F}$)	MOT HOTA	MOTS MOTSA
SATA	**81.3**	**77.8**	**71.4**	**59.7**	**38.1**
DeMoE					
W/o CpMoE	80.8	75.8	70.7	56.2	37.9
W/o SaMoE	81.3	75.3	69.3	55.3	37.7
W/o CpMoE & SaMoE	80.8	74.7	67.9	56.1	37.7
W/o L_{MoE}	80.7	75.2	70.2	58.4	36.9
TaMOT					
W/o CGM	78.5	74.3	68.7	-	-
W/o fine-grained memory	80.7	75.3	69.2	56.7	34.1
W/o spatiotemporal memory	79.7	75.8	67.4	57.5	30.7
W/o MEM	79.5	74.3	67.0	54.2	29.1

Table 11: Ablation studies on DeMoE and TaMOT.

Decoupled Mixture-of-Expert. To investigate the impact of our proposed DeMoE, several versions of our proposed method are provided, including ①: Removing the CpMoE sub-module in DeMoE. ②: Removing the SaMoE sub-module in DeMoE. ③: Removing the CpMoE and SaMoE sub-modules in DeMoE. ④: Removing the MoE loss L_{MoE} in DeMoE. As can be seen in Table 11, the performance degrades significantly after removing CpMoE or SaMoE sub-modules in multi-modal subtasks, which demonstrates the effectiveness of the proposed DeMoE in handling any modality input.

Task-aware MOT pipeline. To further verify the effectiveness of the proposed TaMOT, several variants are designed, including ①: Removing the CGM. ②: Removing the fine-grained memory. ③: Removing the spatiotemporal memory. ④: Removing the MEM. As shown in Table 11, comparative results indicate that keeping track of all potential objects can further improve the robustness of SOT and VOS tasks. Besides, the tracking performance experiences a significant decline upon the ablation of fine-grained memory or spatiotemporal memory, which confirms the necessity of the comprehensive trajectory information incorporated in TaMOT for accurate instance matching.

Conclusion

In this paper, we propose a universal tracking and segmentation framework, referred to as SATA, which is capable of processing inputs from any modality and predicting results for a wide range of tracking and segmentation subtasks with a fully shared network architecture, model weights, and inference pipeline. Extensive experiments on 18 challenging benchmarks demonstrate that SATA achieves superior performance across these tasks. We hope that SATA can lay a solid foundation for future research on AGI.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China under Grant No.62441235, and part by the Beijing Natural Science Foundation under Grant No.L257005.

References

Athar, A.; Mahadevan, S.; Osep, A.; Leal-Taixé, L.; and Leibe, B. 2020. Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In *ECCV*, 158–177. Springer.

Bai, Y.; Zhao, Z.; Gong, Y.; et al. 2024. Artrackv2: Prompting autoregressive tracker where to look and how to describe. In *CVPR*, 19048–19057.

Cai, W.; et al. 2025. SPMTrack: Spatio-Temporal Parameter-Efficient Fine-Tuning with Mixture of Experts for Scalable Visual Tracking. In *CVPR*, 16871–16881.

Chen, X.; Kang, B.; et al. 2025. SUTrack: Towards simple and unified single object tracking. In *AAAI*, volume 39, 2239–2247.

Chen, X.; Peng, H.; et al. 2023. Seqtrack: Sequence to sequence learning for visual object tracking. In *CVPR*, 14572–14581.

Chen, Z.; Li, J.; et al. 2024. M3-VOS: Multi-Phase, Multi-Transition, and Multi-Scenery Video Object Segmentation. *arXiv:2412.13803*.

Cheng, H. K.; and Schwing, A. G. 2022. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *ECCV*, 640–658. Springer.

Cheng, H. K.; et al. 2021. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *NeurIPS*, 34: 11781–11794.

Cho, S.; Lee, H.; et al. 2022. Tackling background distraction in video object segmentation. In *ECCV*, 446–462. Springer.

Chu, P.; Wang, J.; et al. 2023. Transmot: Spatial-temporal graph transformer for multiple object tracking. In *WACV*, 4870–4880.

Cui, Y.; Jiang, C.; et al. 2022. Mixformer: End-to-end tracking with iterative mixed attention. In *CVPR*, 13608–13618.

Cui, Y.; Song, T.; Wu, G.; et al. 2023. Mixformerv2: Efficient fully transformer tracking. *NeurIPS*, 36: 58736–58751.

Fan, H.; Lin, L.; et al. 2019. Lasot: A high-quality benchmark for large-scale single object tracking. In *CVPR*, 5374–5383.

Feng, X.; Zhang, D.; et al. 2025. CSTrack: Enhancing RGB-X Tracking via Compact Spatiotemporal Features. In *ICML*.

Fu, Y.; Liu, S.; et al. 2021. Learning to track instances without video annotations. In *CVPR*, 8680–8689.

Gao, S.; Yang, J.; Li, Z.; Zheng, F.; Leonardis, A.; and Song, J. 2022. Learning dual-fused modality-aware representations for rgbd tracking. In *ECCV*, 478–494.

Gao, S.; Zhou, C.; and Zhang, J. 2023. Generalized relation modeling for transformer tracking. In *CVPR*, 18686–18695.

Guo, M.; Tan, W.; Ran, W.; Jing, L.; and Zhang, Z. 2025. DreamTrack: Dreaming the Future for Multimodal Visual Object Tracking. In *CVPR*, 7201–7210.

Guo, P.; Li, W.; et al. 2024. X-prompt: Multi-modal visual prompt for video object segmentation. In *ACM MM*, 5151–5160.

Hong, L.; Yan, S.; et al. 2024. OneTracker: Unifying Visual Object Tracking with Foundation Models and Efficient Tuning. In *CVPR*.

Hou, X.; Xing, J.; et al. 2024. SDSTrack: Self-Distillation Symmetric Adapter Learning for Multi-Modal Visual Object Tracking. In *CVPR*.

Hu, X.; Tai, Y.; Zhao, X.; Zhao, C.; Zhang, Z.; Li, J.; Zhong, B.; and Yang, J. 2025a. Exploiting multimodal spatial-temporal patterns for video object tracking. In *AAAI*, volume 39, 3581–3589.

Hu, Y.; Shao, Z.; Fan, B.; and Liu, H. 2025b. Dual-level Modality De-biasing for RGB-T Tracking. *IEEE TIP*.

Huang, K.; Lertniphonphan, K.; Chen, F.; Li, J.; and Wang, Z. 2023. Multi-object tracking by self-supervised learning appearance model. In *CVPR*, 3163–3169.

Huang, L.; Zhao, X.; and Huang, K. 2019. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE TPAMI*, 43(5): 1562–1577.

Jiang, J.; Wang, Z.; Zhao, M.; Li, Y.; and Jiang, D. 2025. SAM2MOT: A Novel Paradigm of Multi-Object Tracking by Segmentation. *arXiv:2504.04519*.

Kang, B.; Chen, X.; et al. 2025. Exploring enhanced contextual information for video-level object tracking. In *AAAI*, volume 39, 4194–4202.

Ke, L.; Ding, H.; Danelljan, M.; Tai, Y.-W.; Tang, C.-K.; and Yu, F. 2022. Video mask transfiner for high-quality video instance segmentation. In *ECCV*, 731–747. Springer.

Ke, L.; Li, X.; Danelljan, M.; Tai, Y.-W.; Tang, C.-K.; and Yu, F. 2021. Prototypical cross-attention networks for multiple object tracking and segmentation. *NeurIPS*, 34: 1192–1203.

Li, B.; Peng, F.; Hui, T.; Wei, X.; Wei, X.; Zhang, L.; Shi, H.; and Liu, S. 2024a. RGB-T tracking with template-bridged search interaction and target-preserved template updating. *IEEE TPAMI*.

Li, C.; Xue, W.; Jia, Y.; Qu, Z.; Luo, B.; Tang, J.; and Sun, D. 2022. LasHeR: A Large-Scale High-Diversity Benchmark for RGBT Tracking. *IEEE TIP*, 31: 392–404.

Li, H.; Wang, J.; Yuan, J.; Li, Y.; Weng, W.; Peng, Y.; Zhang, Y.; Xiong, Z.; and Sun, X. 2024b. Event-assisted low-light video object segmentation. In *CVPR*, 3250–3259.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 19730–19742. PMLR.

Li, S.; Ke, L.; et al. 2024. Matching anything by segmenting anything. In *CVPR*, 18963–18973.

Li, W.; Guo, P.; et al. 2024. Onevos: unifying video object segmentation with all-in-one transformer framework. In *ECCV*, 20–40.

Li, X.; Zhong, B.; Liang, Q.; Li, G.; Mo, Z.; and Song, S. 2025. MambaLCT: Boosting Tracking via Long-term Context State Space Model. In *AAAI*, volume 39, 4986–4994.

Lin, L.; Fan, H.; et al. 2024. Tracking meets lora: Faster training, larger model, stronger performance. In *ECCV*, 300–318.

Lv, W.; Huang, Y.; Zhang, N.; Lin, R.-S.; Han, M.; and Zeng, D. 2024. Diffmot: A real-time diffusion-based multiple object tracker with non-linear prediction. In *CVPR*, 19321–19330.

Ma, F.; Shou, M. Z.; et al. 2022. Unified transformer tracker for object tracking. In *CVPR*, 8781–8790.

Mayer, C.; Danelljan, M.; Paudel, D. P.; and Gool, L. V. 2021. Learning Target Candidate Association to Keep Track of What Not to Track. In *ICCV*, 13424–13434.

Oh, S. W.; Lee, J.-Y.; et al. 2019. Video object segmentation using space-time memory networks. In *ICCV*, 9226–9235.

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

Ranasinghe, K.; Naseer, M.; Hayat, M.; Khan, S.; and Khan, F. S. 2021. Orthogonal projection loss. In *ICCV*, 12333–12343.

Ravi, N.; et al. 2024. Sam 2: Segment anything in images and videos. *arXiv:2408.00714*.

Sun, P.; Cao, J.; et al. 2022. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *CVPR*, 20993–21002.

Tan, Y.; Shao, J.; et al. 2025. What You Have is What You Track: Adaptive and Robust Multimodal Tracking. In *ICCV*, 3455–3465.

Tan, Y.; Wu, Z.; et al. 2025. XTrack: Multimodal Training Boosts RGB-X Video Object Trackers. *arXiv:2405.17773*.

Tang, Z.; Xu, T.; Wu, X.; Zhu, X.-F.; and Kittler, J. 2024. Generative-based fusion mechanism for multi-modal tracking. In *AAAI*, volume 38, 5189–5197.

Videnovic, J.; et al. 2025. A distractor-aware memory for visual object tracking with sam2. In *CVPR*, 24255–24264.

Wang, J.; Chen, D.; Luo, C.; He, B.; Yuan, L.; Wu, Z.; and Jiang, Y.-G. 2024. Omnivid: A generative framework for universal video understanding. In *CVPR*, 18209–18220.

Wang, J.; Wu, Z.; Chen, D.; Luo, C.; Dai, X.; Yuan, L.; and Jiang, Y.-G. 2025. OmniTracker: Unifying Visual Object Tracking by Tracking-with-Detection. *IEEE TPAMI*.

Wang, X.; Li, J.; et al. 2023. Visevent: Reliable object tracking via collaboration of frame and event flows. *IEEE TCYB*, 54(3): 1997–2010.

Wang, Z.; Zhao, H.; Li, Y.-L.; Wang, S.; Torr, P.; and Bertinetto, L. 2021. Do different tracking tasks require different appearance models? *NeurIPS*, 34: 726–738.

Wu, Z.; Zheng, J.; et al. 2024. Single-Model and Any-Modality for Video Object Tracking. In *CVPR*.

Xiao, Y.; Zhao, J.; Lu, A.; Li, C.; Yin, B.; Lin, Y.; and Liu, C. 2025. Cross-modulated Attention Transformer for RGBT Tracking. In *AAAI*, volume 39, 8682–8690.

Xie, J.; Zhong, B.; Liang, Q.; Li, N.; Mo, Z.; and Song, S. 2025. Robust tracking via mamba-based context-aware token learning. In *AAAI*, volume 39, 8727–8735.

Xu, C.; Zhong, B.; et al. 2025. Less is more: Token context-aware learning for object tracking. In *AAAI*, volume 39, 8824–8832.

Xu, Q.; Wang, L.; et al. 2024. Heterogeneous graph transformer for multiple tiny object tracking in RGB-T videos. *IEEE TMM*.

Yan, B.; Jiang, Y.; et al. 2022. Towards grand unification of object tracking. In *ECCV*, 733–751.

Yan, B.; Jiang, Y.; et al. 2023. Universal instance perception as object discovery and retrieval. In *CVPR*, 15325–15336.

Yan, B.; Zhang, X.; Wang, D.; Lu, H.; and Yang, X. 2021a. Alpha-refine: Boosting tracking performance by precise bounding box estimation. In *CVPR*, 5289–5298.

Yan, S.; Yang, J.; Käpylä, J.; Zheng, F.; Leonardis, A.; and Kämäräinen, J.-K. 2021b. Depthtrack: Unveiling the power of RGBD tracking. In *ICCV*, 10725–10733.

Yang, J.; Gao, M.; Cong, R.; Wang, C.; Zheng, F.; and Leonardis, A. 2023. Unveiling the Power of Visible-Thermal Video Object Segmentation. *IEEE TCSVT*, 34(7): 5376–5388.

Yang, M.; Han, G.; Yan, B.; Zhang, W.; Qi, J.; Lu, H.; and Wang, D. 2024. Hybrid-sort: Weak cues matter for online multi-object tracking. In *AAAI*, volume 38, 6504–6512.

Yang, Z.; Wei, Y.; et al. 2021. Associating objects with transformers for video object segmentation. *NeurIPS*, 34: 2491–2502.

Yang, Z.; and Yang, Y. 2022. Decoupling features in hierarchical propagation for video object segmentation. *NeurIPS*, 35: 36324–36336.

Yang, Z.; et al. 2021. Collaborative video object segmentation by multi-scale foreground-background integration. *IEEE TPAMI*, 44(9): 4701–4712.

Ye, B.; Chang, H.; Ma, B.; Shan, S.; and Chen, X. 2022. Joint feature learning and relation modeling for tracking: A one-stream framework. In *ECCV*, 341–357.

Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; and Darrell, T. 2020. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2636–2645.

Zhang, L.; Wang, L.; Wu, Y.; Chen, M.; Zheng, D.; Cao, L.; Zeng, B.; and Cai, Y. 2025a. UniRTL: A universal RGBT and low-light benchmark for object tracking. *Pattern Recognition*, 158: 110984.

Zhang, P.; Zhao, J.; Wang, D.; Lu, H.; and Ruan, X. 2022. Visible-thermal UAV tracking: A large-scale benchmark and new baseline. In *CVPR*, 8886–8895.

Zhang, T.; Debattista, K.; Zhang, Q.; Han, J.; et al. 2024a. Revisiting motion information for RGB-Event tracking with MOT philosophy. *NeurIPS*, 37: 89346–89372.

Zhang, T.; He, X.; Luo, Y.; Zhang, Q.; and Han, J. 2024b. Exploring target-related information with reliable global pixel relationships for robust RGB-T tracking. *Pattern Recognition*, 110707.

Zhang, T.; Jiao, Q.; et al. 2024. Exploring Multi-modal Spatial-Temporal Contexts for High-performance RGB-T Tracking. *IEEE TIP*, 1–1.

Zhang, T.; Zhang, Q.; Debattista, K.; and Han, J. 2025b. Cross-Modality Distillation for Multi-modal Tracking. *IEEE TPAMI*.

Zhang, X.; Tian, Y.; et al. 2022. Hivit: Hierarchical vision transformer meets masked image modeling. *arXiv:2205.14949*.

Zhang, Y.; Sun, P.; et al. 2022. Bytetrack: Multi-object tracking by associating every detection box. In *ECCV*, 1–21. Springer.

Zhang, Y.; et al. 2023. Motrv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors. In *CVPR*, 22056–22065.

Zhao, H.; Chen, J.; Wang, L.; and Lu, H. 2023. ARKitTrack: A New Diverse Dataset for Tracking Using Mobile RGB-D Data. In *CVPR*, 5126–5135.

Zheng, Y.; Zhong, B.; et al. 2024. Odtrack: Online dense temporal token learning for visual tracking. In *AAAI*, volume 38, 7588–7596.

Zhu, J.; Lai, S.; Chen, X.; Wang, D.; and Lu, H. 2023. Visual prompt multi-modal tracking. In *CVPR*, 9516–9526.

Zhu, Z.; Zhong, B.; et al. 2025. Adaptive Expert Decision for RGB-T Tracking. *IEEE TCSVT*.