ViewSAM: Learning View-aware Cross-modal Semantics for Weakly Supervised Cross-view Referring Multi-Object Tracking

Jiawei Ge[1,2] **Xintian Zhang**[1] **Jiuxin Cao**[1,3]* **Bo Liu**[1,3] **Fabian Deuser**[4] **Chang Liu**[1]
Wenkang Gong[1] **Siyou Li**[5] **Juexi Shao**[5] **Wenqing Wu**[6] **Chen Feng**[7] **Ioannis Patras**[5]
[1]Southeast University [2]Institute of AI for Industries, Chinese Academy of Sciences
[3]Purple Mountain Laboratories [4]University of the Bundeswehr Munich
[5]Queen Mary University of London [6]Nanjing University of Science & Technology
[7]Queen's University Belfast

Abstract

Cross-view Referring Multi-Object Tracking (CRMOT) aims to track multiple objects specified by natural language across multiple camera views, with globally consistent identities. Despite recent progress, existing methods rely heavily on costly frame-level spatial annotations and cross-view identity supervision. To reduce such reliance, we explore CRMOT under weak supervision by leveraging the capabilities of foundation models. Our empirical study shows that directly applying foundation models such as SAM2 and SAM3, even with task-specific modifications, fails to accurately understand referring expressions and maintain consistent identities across views. Yet, such models remain effective at producing reliable object tracklets that can serve as supervision signal. We therefore repurpose foundation models as pseudo-label generators and propose a two-stage framework for weakly supervised CRMOT, using only object category labels as coarse-grained supervision. In the first stage, we design an Affinity-guided Cross-view Re-prompting strategy to refine and associate SAM3-generated tracklets across cameras, producing reliable cross-view pseudo labels for subsequent training. In the second stage, we introduce ViewSAM, a CRMOT model built upon SAM2 that explicitly models view-aware cross-modal semantics. By formulating view-induced variations as learnable conditions, ViewSAM bridges the gap between view-variant visual observations and view-invariant textual expressions, enabling robust cross-view referring tracking with only approximately 10% additional parameters. Extensive experiments demonstrate that ViewSAM achieves SOTA performance under weak supervision and remains competitive with fully supervised methods.

1 Introduction

The objective of Cross-view Referring Multi-Object Tracking (CRMOT) [1] is to localize and track multiple objects specified by textual descriptions, producing identity-consistent trajectories across time and camera views. By leveraging synchronized observations from multiple cameras, CRMOT extends single-view Referring Multi-Object Tracking (RMOT) [2] to multi-camera scenarios and exploits cross-view complementarities to alleviate long-standing challenges in RMOT, such as severe occlusions and target disappearance. This capability is essential for enabling language-driven perception systems in the real world, particularly autonomous driving [3] and embodied AI [4].

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

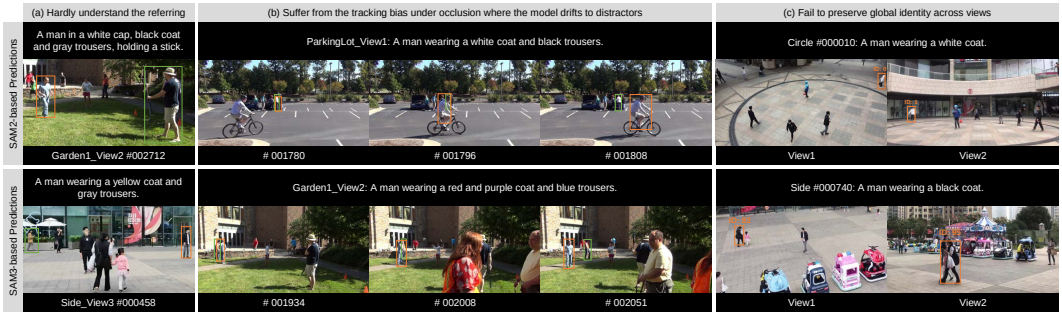

Figure 1: Empirical visualizations (Prediction and GT): (a) hard to understand referring, (b) drift to distractors under occlusion, and (c) fail to preserve cross-view ID with object feature clustering.

Despite its importance, existing CRMOT methods [5, 1, 6, 7, 8, 9] are predominantly developed under a fully supervised learning paradigm. These approaches rely on large-scale datasets annotated with dense frame-level spatial labels and cross-view identity correspondences. However, collecting such fine-grained annotations is extremely expensive, as the labeling cost grows rapidly with both the number of frames and views, making it difficult to scale CRMOT solutions to open-world scenes.

To alleviate such reliance, we explore a new learning paradigm termed **Weakly Supervised Cross-view Referring Multi-Object Tracking (WSCRMOT)**, inspired by recent advances in weakly supervised learning [10, 11, 12]. In this setting, the training data comprise **raw multi-view videos annotated only with object category labels** [13] as coarse-grained supervision, without any task-specific spatial annotations, cross-view identities, or referring expressions. While such weak supervision greatly reduces annotation costs, it also introduces a fundamental challenge: *how can a model learn cross-view referring tracking without explicit spatial or identity supervision?*

Recent progress in large-scale foundation models [14, 15, 16] provides a promising direction. Models such as SAM [16] and its successors SAM2 [17] and SAM3 [18] exhibit remarkable generalization capabilities in segmentation and tracking, and have increasingly been explored as external supervision sources for weakly supervised learning [12, 19, 20]. With their flexible prompting mechanisms and strong capabilities, SAM2 and SAM3 provide a natural starting point for exploring WSCRMOT.

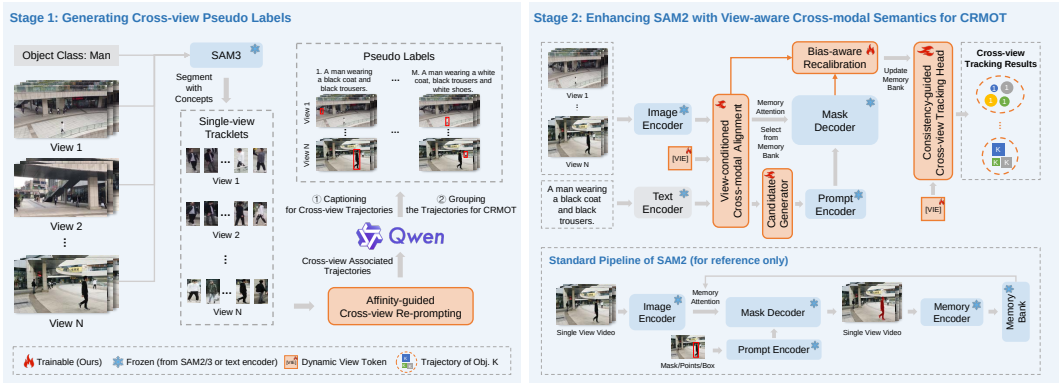

Figure 2: Overview of our two-stage framework for WSCRMOT. In Stage 1, we generate pseudo labels by refining and associating SAM3 tracklets with our Affinity-guided Cross-view Re-prompting strategy. In Stage 2, limited extra parameters are introduced to enhance SAM2 with view-aware cross-modal semantics modeling for CRMOT, where view-induced variations are treated as learnable conditions rather than detrimental factors. The standard SAM2 pipeline is also shown for reference.

However, our empirical study (Fig. 1) reveals key limitations when applying SAM-based models in WSCRMOT. First, they lack robust semantic understanding for language-guided long-term tracking, often suffering from the **tracking bias** where the model drifts to distractors that partially match the textual prompt. Moreover, they fail to maintain global identity consistency, as the same object may exhibit substantial variations in appearance and motion patterns across views. Taken together, these

observations indicate that, while SAM-based models are not well suited for direct application in this task, they can instead be repurposed as effective pseudo-label generators under weak supervision.

Building on these insights, we propose a two-stage framework for Weakly Supervised Cross-view Referring Multi-Object Tracking, as illustrated in Fig. 2. In the first stage, we exploit the strong tracking capability of foundation models to generate cross-view pseudo labels. Starting from single-view tracklets produced by SAM3, we design an **Affinity-guided Cross-view Re-prompting** strategy that iteratively refines object predictions and aligns tracklets across views. By integrating these refined tracklets with trajectory-level descriptions generated by a multimodal large language model (MLLM), we obtain high-quality pseudo supervision for downstream training.

In the second stage, we introduce **ViewSAM**, a SAM2-based CRMOT model that explicitly learns view-aware cross-modal semantics throughout single-view tracking and cross-view association. Instead of treating view variations as factors to be ignored or eliminated, ViewSAM formulates them as learnable semantic conditions. Specifically, we propose a View-conditioned Cross-modal Alignment module to comprehend referring expressions, jointly reasoning over visual and textual modalities through interactions with a learnable dynamic View Token. To mitigate the tracking bias mentioned earlier, we further propose a Bias-aware Recalibration module that encourages the model to refocus on objects that are more consistent with the referring expression. For cross-view association, we further devise a Consistency-guided Cross-view Tracking Head that leverages the learned dynamic view token to modulate tracklet representations and suppress view-induced discrepancies. Together with consistency-guided objectives, this mechanism projects tracklets into a view-invariant space and provides explicit supervision for cross-view tracklet association. Through these designs, ViewSAM bridges the gap between view-variant visual appearances and view-invariant textual expressions, enabling robust cross-view referring tracking. To sum up, our main contributions are as follows:

1. To the best of our knowledge, we propose the first weakly supervised framework for Cross-view Referring Multi-Object Tracking that requires only object category labels for training, substantially reducing annotation costs compared with fully supervised CRMOT methods.
2. We propose an Affinity-guided Cross-view Re-prompting strategy to generate reliable cross-view pseudo labels, which models cross-view affinity for consistent identity association and refines trajectories iteratively.
3. We propose ViewSAM, a CRMOT model built upon SAM2 that explicitly models view-aware cross-modal semantics, achieving robust referring understanding and consistent cross-view tracking with only ~10% additional parameters.
4. Extensive experiments demonstrate that the proposed framework achieves SOTA performance under weak supervision, substantially narrowing the gap to fully supervised methods.

2 Related Work

2.1 Referring Object Tracking

Referring Object Tracking (ROT) aims to localize and track objects in videos according to explicit referring signals such as natural language descriptions or visual prompts. Early studies mainly focus on the single-object setting, often referred to as Vision–Language Tracking (VLT) [21, 22], where visual features are aligned with textual embeddings via cross-modal interaction [23, 24, 10, 25, 26].

Recent work extends this paradigm to multi-object scenarios, known as Referring Multi-Object Tracking (RMOT) [2]. Existing RMOT methods can be broadly categorized into two-stage approaches [27, 28, 29] and end-to-end architectures [2, 30, 31, 32, 33]. However, these methods are limited to single-view settings. To handle multi-camera scenarios, several studies investigate cross-view multi-object tracking by associating identities across cameras using appearance and motion cues [5, 6, 7, 8, 9]. Building upon it, Chen et al. [1] introduce Cross-view Referring Multi-Object Tracking (CRMOT), where objects specified by language expressions are consistently tracked across synchronized camera views. Unlike them, we explore CRMOT under a weakly supervised setting.

2.2 Segment Anything Models

The Segment Anything Model (SAM) [16] is a powerful vision foundation model [34, 35, 36, 37, 38] for generic object segmentation with strong zero-shot generalization [39, 40, 41, 42]. Its promptable

design has made it widely adopted as a universal segmentation backbone or pseudo-label generator for downstream tasks. To extend SAM to videos, SAM2 [17] introduces memory mechanisms for mask propagation across frames. Subsequent works further enhance its tracking robustness [43, 44, 45]. More recently, SAM3 [18] integrates a DETR-based detector to segment all instances of a given concept prompt.

However, SAM-based models are not designed to understand complex referring expressions or maintain identity consistency across views. In this work, we leverage the strong generalization ability of SAM models to generate pseudo supervision and propose ViewSAM, which explicitly models view-induced variations as learnable semantic conditions for robust cross-view referring tracking.

3 The Proposed Method

Problem Setting. Given a set of synchronized multi-view video sequences $\mathcal{V} = \{\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_N\}$ captured from N cameras observing the same scene, and a referring expression $\mathcal{R}$, the goal of cross-view referring multi-object tracking (CRMOT) is to detect and track all objects specified by $\mathcal{R}$. The output is a set of object trajectories, where each target object is assigned a global ID across views.

Unlike fully supervised CRMOT methods that rely on referring expressions, frame-level bounding boxes, and cross-view identity annotations, we consider a weakly supervised setting in which only the object category label $\mathcal{C}$ and raw synchronized multi-view videos $\mathcal{V}$ are available during training, without any additional task-specific supervision.

Overview. As shown in Fig. 3, our framework consists of two stages: In the first stage (Sec. 3.2), we leverage SAM3's concept-level tracking capability to generate cross-view pseudo labels via an Affinity-guided Cross-view Re-prompting strategy. In the second stage (Sec. 3.3), we introduce ViewSAM that incorporates view-aware cross-modal semantics and enables ID-consistency across views. Before elaborating them, we briefly review the architectures of SAM2 and SAM3 in Sec. 3.1.

3.1 Background on SAM2 and SAM3

SAM2 [17] is a foundation model for video object segmentation composed of an Image Encoder, a Prompt Encoder, a Mask Decoder, and a Memory Mechanism for temporal mask propagation.

Memory Mechanism. For each frame, the Image Encoder extracts visual tokens, which interact with historical memory tokens from Memory Bank. Through memory attention, the current frame features attend to previous frames to incorporate temporal context. After mask prediction, a Memory Encoder converts current features and predicted masks into new memory tokens.

Prompt Encoding and Mask Decoding. SAM2 supports sparse prompts (points or boxes) and dense prompts (masks). Sparse prompts are encoded using positional and type embeddings, while dense prompts are fused with image features. The Mask Decoder employs a two-way transformer to decode prompt-conditioned features for initial frame and memory-conditioned features for the subsequent.

SAM3 [18] extends SAM2 toward concept-level segmentation by introducing a **heavy** concept-aware detector while **keeping the underlying video tracking backbone unchanged.** Yet, it is not designed to interpret complex referring expressions. Hence, we adopt SAM2 as the base tracker for efficiency.

3.2 Cross-view Pseudo Label Generation

Since the training data provide only category labels, directly learning a CRMOT model is challenging due to the absence of object locations and cross-view identity annotations. To address this, we first leverage SAM3 to localize candidate objects in each view using the category label as a text prompt. We then associate tracklets across views using an Affinity-guided Cross-view Re-prompting strategy. Based on these preliminary cross-view single-object tracklets, we further refine and expand them into pseudo labels for CRMOT through the LLM-driven Referring-level Multi-object Grouping.

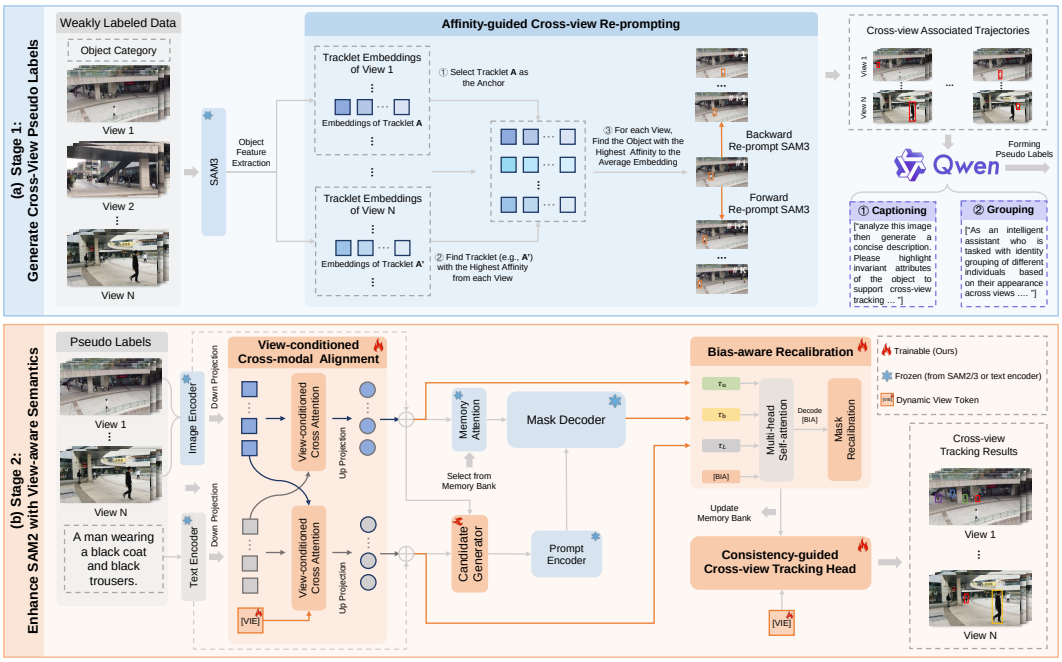

Figure 3: The overall pipeline of our framework. (a) Affinity-guided Cross-view Re-prompting generates cross-view pseudo labels from weakly labeled multi-view videos by extracting tracklets with SAM3, aligning high-affinity trajectories across views, and refining them through forward and backward re-prompting. (b) Using these pseudo labels, we enhance SAM2 with view-aware cross-modal semantics, including a View-conditioned Cross-modal Alignment module that reasons over visual and textual features, followed by the Bias-aware Recalibration to overcome the tracking bias and a Consistency-guided Cross-view Tracking Head to produce the trajectories with global IDs.

3.2.1 Affinity-guided Cross-view Re-prompting.

Our strategy mainly consists of two components: the Affinity-guided Cross-view Association and the Bi-directional Re-prompting. By leveraging the embedding affinity as a semantic constraint, we iteratively establish and refine cross-view correspondences for each object.

Formally, for each camera view i, we first prompt SAM3 with the category label $\mathcal{C}$ to segment candidate object instances, producing a set of single-view candidate tracklets: $\mathcal{T}_i = \{\mathcal{T}_i^1, \mathcal{T}_i^2, \ldots, \mathcal{T}_i^{|\mathcal{T}_i|}\}, \quad i = 1, \ldots, N$. Each tracklet $\mathcal{T}_i^k$ is a temporally ordered mask sequence corresponding to object k: $\mathcal{T}_i^k = \{M_{t,i}^k\}_{t \in \Gamma_i^k}$, where Γ_i^k denotes the set of valid frames containing object k within $\mathcal{T}_i^k$.

Then, we compute a tracklet-level embedding by averaging frame-wise object features:

$$E_i^k = \frac{1}{|\Gamma_i^k|} \sum_{t \in \Gamma_i^k} f_\theta(I_{t,i}, M_{t,i}^k), \qquad E_i^k \in \mathbb{R}^{d_{\text{id}}}, \tag{1}$$

where $I_{t,i}$ is image at frame t and view i, and $f_\theta(\cdot, \cdot)$ denotes an off-the-shelf ReID head [46] that extracts a d_{id}-dimensional embedding.

Affinity-guided Cross-view Association. Given an anchor tracklet $\mathcal{T}_i^k$, we compute cosine similarities between its embedding E_i^k and all candidate tracklet embeddings from every other view $j \neq i$. For each target view, we retain the top-affinity candidate only if its similarity exceeds a reliability threshold; otherwise that view is treated as unmatched. In this way, we form a preliminary cross-view tracklet set for object k. Since this stage serves only as an initialization, duplicated matches across different anchors are allowed and are later corrected by the bi-directional re-prompting strategy.
Bi-directional Re-prompting for Identity Refinement. The preliminary association can still be noisy due to occlusion, appearance variations, or distractors. We further refine it by establishing an aggregated identity prototype from the confidently matched tracklets, which aims to capture semantic

characteristics shared across views while suppressing frame-level noise:

$$\bar{E}^k = \frac{1}{|\mathcal{A}^k|} \sum_{(i',k') \in \mathcal{A}^k} E_{i'}^{k'}, \tag{2}$$

where $\mathcal{A}^k$ denotes the set of matched tracklets retained for the current anchor after affinity filtering. For each view, we search over the candidate object masks predicted by SAM3, and select the mask whose embedding has the highest affinity with the prototype $\bar{E}^k$. This selected mask is used as a renewed mask prompt to initialize SAM3 tracking in both forward and backward temporal directions within that view, producing a refined trajectory with improved temporal consistency. Finally, we obtain a refined set of cross-view single-object trajectories, serving as candidates for pseudo labels.

3.2.2 LLM-driven Referring-level Multi-object Grouping.

While the previous step establishes cross-view correspondences for individual objects, CRMOT still requires supervision at the referring level, where one expression corresponds to multiple objects. To this end, we leverage the MLLM (Qwen3-VL-8B [47]) to generate attribute-centric descriptions for each trajectory, capturing appearance cues such as color, clothing, and accessories. These descriptions are embedded into a shared semantic space using the MLLM encoder. Based on the semantic similarity, trajectories are then grouped under the same referring expression. Finally, each referring expression with grouped cross-view trajectories, forms CRMOT pseudo labels.

3.3 Enhance SAM2 with View-aware Cross-modal Semantics

As illustrated in Fig. 3(b), given multi-view video streams and a referring expression, ViewSAM first aligns visual features with textual embeddings using a **View-conditioned Cross-modal Alignment (VC-CMA)** module guided by the learnable dynamic View Token. Then the view-aware cross-modal representations are fed into the Candidate Generator to produce box prompts that guide mask decoding in SAM2, extending it to multi-object localization. Afterwards, a **Bias-aware Recalibration (BAR)** module dynamically shifts the model's focus on objects that better match the referring expression.

For cross-view association, the predicted tracklets are processed by a **Consistency-guided Cross-view Tracking Head**, which leverages the learned dynamic view token to modulate tracklet representations before association. Under consistency-guided objectives, the tracklet features are projected into a shared view-invariant embedding space, enabling reliable cross-view identity alignment. All the details about the architecture and training losses for each component are provided in C.

3.3.1 View-conditioned Cross-modal Alignment

To mitigate view-induced variations in referring expression comprehension, we introduce a learnable dynamic View Token built upon the popular Adapter framework [48]. Through interactions with visual and textual features, the dynamic View Token enhances multi-modal representations, enabling VC-CMA to learn view-aware cross-modal representations for robust target localization across views.
Learnable Dynamic View Token. For each view i with a clip of T frames, we construct a dynamic view-conditioned token, where a learnable base embedding $g_i \in \mathbb{R}^d$ captures static view priors:

$$\hat{e}_{t,i} = \mathrm{Norm}(g_i + \lambda_{\mathrm{dyn}} \mathrm{Norm}(W_{\mathrm{view}} \mathrm{Pool}_{\mathrm{avg}}(F_{t,i}))), \tag{3}$$

where $F_{t,i} \in \mathbb{R}^{d \times H \times W}$ denotes the visual feature map at frame t and view i, $W_{\mathrm{view}} \in \mathbb{R}^{d \times d}$ is a linear projection, and λ_{dyn} controls the contribution of the frame-level visual context. For temporal stability, we smooth the view token using an exponential moving average (the decay factor $\alpha \in [0, 1]$):

$$e_{t,i} = \mathrm{Norm}(\alpha e_{t-1,i} + (1-\alpha)\hat{e}_{t,i}), \qquad e_{1,i} = \hat{e}_{1,i}. \tag{4}$$

View-conditioned Cross Attention. Let $T_{\mathrm{text}} \in \mathbb{R}^{M \times d}$ denote textual tokens and $X_t \in \mathbb{R}^{HW \times d}$ denote visual tokens obtained from the feature map $F_{t,i}$. Following the Adapter's design [48], tokens are projected into a bottleneck space via $X_t^s = X_t W_v^{\downarrow}$ and $T_{\mathrm{text}}^s = T_{\mathrm{text}} W_t^{\downarrow}$, with dimension $d_s < d$.

To bridge the gap between view-variant visual observations and view-invariant textual expressions, we inject view-related prior into query and key branches via view-conditioned cross attention (VCCA):

$$\mathrm{VCCA}(U, R; e) = U \odot \mathrm{MHA}(U + b_u(e),\ R + b_r(e),\ R), \quad b_u(e) = eW_u, b_r(e) = eW_r, \tag{5}$$

where U denotes the query tokens and R denotes the reference tokens (serving as key and value in standard cross-attention). Here, the Hadamard product $\odot$ acts as a lightweight feature-wise gating mechanism, enabling view-aware conditioning while preserving the original token semantics.

We then update visual tokens using the frame-level view token and project them to the original space:

$$X^t_{\mathrm{ca}} = X_t + \mathrm{VCCA}(X^s_t, T^s_{\mathrm{text}}; e_{t,i})W^{\uparrow}_v. \tag{6}$$

For textual tokens, we instead use a clip-level visual summary $\bar{X}^s$ and a clip-level view token e^{clip}_i:

$$T_{\mathrm{ca}} = T_{\mathrm{text}} + \mathrm{VCCA}(T^s_{\mathrm{text}}, \bar{X}^s; e^{\mathrm{clip}}_i)W^{\uparrow}_t, \quad \bar{X}^s = \frac{1}{T}\sum_{t=1}^{T} X^s_t, e^{\mathrm{clip}}_i = \frac{1}{T}\sum_{t=1}^{T} e_{t,i}. \tag{7}$$

Finally, these view-aware representations are fed into the Candidate Generator to produce box prompts that highlight regions likely to contain targets, which, together with the representations, guide mask decoding in SAM2, hence extending it from single-object to multi-object localization.

3.3.2 Bias-aware Recalibration

SAM-based trackers suffer from the aforementioned tracking bias, where memory-guided decoding amplifies historical errors and causes drift under occlusion. To mitigate this, we propose Bias-aware Recalibration (BAR), which leverages memory-free features to detect failures and re-align predictions. Let $F^u_{t,i}$ and $F^b_{t,i}$ denote pre- and post-memory-attention features (i.e., un-biased vs. biased), with converted soft masks $M^{k,s}_{t,i}$ for $s \in \{u, b\}$. Their object-level tokens are then extracted to form the referring-aware representations:

$$o^k_s = \mathrm{Norm}\left(\frac{\sum_x F^s_{t,i}(x)\, M^{k,s}_{t,i}(x)}{\sum_x M^{k,s}_{t,i}(x)}\right),\ \alpha^k_s = \mathrm{softmax}_k\big(\langle o^k_s, t_{\mathcal{R}}\rangle\big),\ \tau_s = \mathrm{Norm}\left(\sum_k \alpha^k_s o^k_s\right) \tag{8}$$

Here $t_{\mathcal{R}}$ is the referring token derived from textual tokens T_{text}. Afterwards, we introduce a learnable token $[\mathrm{BIA}]$ to assess whether the current predictions are reliable through the self-attention operation $\mathrm{SA}(\cdot)$. Finally, we adaptively fuse the two predictions based on the resulting bias score $p_{\mathrm{bias}} \in [0, 1]$:

$$p_{\mathrm{bias}} = \phi(\mathrm{SA}([\mathrm{BIA}], t_{\mathcal{R}}, \tau_u, \tau_b)), \quad M^k_{t,i} = (1 - p_{\mathrm{bias}})\, M^{k,b}_{t,i} + p_{\mathrm{bias}}\, M^{k,u}_{t,i}. \tag{9}$$

3.3.3 Consistency-guided Cross-view Tracking Head

We propose Consistency-guided Cross-view Tracking (CGCT) head, which is mainly based on CNNs, to learn a view-invariant embedding space for cross-view identity association. It explicitly enforces identity consistency across time and views via view-conditioned modulation and consistency-guided objectives. Given frame t in view i, let $\{F^{\ell}_{t,i}\}^{L}_{\ell=1}$ denote multi-scale features. We first perform masked average pooling over multi-scale features for each candidate mask $M^k_{t,i}$ and fuse them:

$$f^k_{t,i} = \mathrm{Norm}\Big(\mathrm{Fuse}\left(\{\mathrm{Pool}(F^{\ell}_{t,i}, M^k_{t,i})\}^{L}_{\ell=1}\right)\Big), \tag{10}$$

where $\mathrm{Pool}(\cdot)$ denotes masked average pooling, and $\mathrm{Fuse}(\cdot)$ is concatenation followed by an MLP.

View-conditioned Modulation. We incorporate the dynamic view token $e_{t,i}$ via Feature-wise Linear Modulation [49] controlled by modulation strength δ, mitigating view-specific appearance variations:

$$[\gamma_{t,i}, \beta_{t,i}] = \mathrm{MLP}(e_{t,i}),\ z^k_{t,i} = \mathrm{Norm}\left((1 + \delta\tanh(\gamma_{t,i})) \odot f^k_{t,i} + \delta\beta_{t,i}\right),\ \gamma_{t,i}, \beta_{t,i} \in \mathbb{R}^d \tag{11}$$

Consistency-guided Objectives. Let $\mathcal{T}^i_g$ denote the trajectory of identity g within view i, $\bar{z}^i_g$ the view-specific prototype, and $\bar{z}_g$ the global identity prototype aggregated across views. The prototypes are computed as the mean embeddings over instances. Here, we formulate a two-level objective:

$$\mathcal{L}_{\mathrm{CGCT}} = \sum_g \Bigg[\underbrace{\sum_i \sum_{(t,k)\in\mathcal{T}^i_g} \left\| z^k_{t,i} - \bar{z}^i_g \right\|^2_2}_{\text{intra-consistency}} + \lambda \underbrace{\sum_i \left\| \bar{z}^i_g - \bar{z}_g \right\|^2_2}_{\text{inter-consistency}}\Bigg], \tag{12}$$

where the intra-consistency term enforces temporal coherence within each view by compacting trajectory embeddings around the view-specific prototype, while the inter-consistency term aligns view-specific prototypes to a shared global identity prototype for cross-view association.

Table 1: Performance comparison of SOTA methods on the CRTrack benchmark. SAM2 and SAM3 are adapted with a cosine-based clustering for cross-view association. Best results within each setting are highlighted in **bold**. For brevity, metric names are **simplified** here by removing "CVR-" prefix.

In-domain Evaluation										
Method	Supervision	Publication	All		Circle		Gate2		Side	
			IDF1	MA	IDF1	MA	IDF1	MA	IDF1	MA
TransRMOT [2]	Fully	CVPR23	23.30	8.03	18.85	6.94	68.03	28.51	14.33	2.65
CRTracker [1]		AAAI25	54.88	35.97	58.38	42.44	**91.60**	73.40	37.97	14.87
ViewSAM		**Ours**	**57.73**	**43.72**	**60.84**	**48.85**	89.21	**81.39**	**43.02**	**24.33**
SAM2 [17]	Zero-shot	ICLR25	3.05	0.28	1.69	0.16	8.53	0.52	3.05	0.35
SAM3 [18]		ICLR26	**13.44**	**2.89**	**10.52**	**2.62**	**32.40**	**8.78**	**11.02**	**1.30**
CRTracker [1]	Weakly	AAAI25	33.06	21.40	31.54	20.25	71.86	57.10	22.15	11.03
ViewSAM		**Ours**	**38.80**	**26.95**	**35.59**	**27.66**	**72.48**	**61.56**	**31.86**	**14.47**
Cross-domain Evaluation										
Method	Supervision	Publication	All		Garden1		Garden2		ParkingLot	
			IDF1	MA	IDF1	MA	IDF1	MA	IDF1	MA
TransRMOT [2]	Fully	CVPR23	3.66	0.20	2.85	0.01	4.23	0.55	3.87	0.00
CRTracker [1]		AAAI25	12.52	2.32	14.96	2.77	11.87	2.80	10.66	1.30
ViewSAM		**Ours**	**15.86**	**8.35**	**19.51**	**10.08**	**14.22**	**7.98**	**13.83**	**6.95**
SAM2 [17]	Zero-shot	ICLR25	7.78	1.81	**13.93**	**5.06**	4.23	0.25	5.17	0.13
SAM3 [18]		ICLR26	**8.00**	**1.93**	8.58	3.44	**6.90**	**1.06**	**8.63**	**1.29**
CRTracker [1]	Weakly	AAAI25	8.12	1.95	9.02	1.76	7.68	2.19	7.65	1.88
ViewSAM		**Ours**	**11.48**	**5.50**	**14.41**	**6.43**	**9.09**	**4.98**	**11.08**	**5.02**

4 Experiments

4.1 Experiment Setup

Dataset and Metrics. For evaluation, we conduct experiments on the CRTrack [1] benchmark for CRMOT and follow its evaluation protocol with CVR-IDF1 (cross-view referring IDF1) and CVR-MA (cross-view referring matching accuracy), which measure identity consistency across time and views, and correctness of cross-view identity matching, under referring constraints, respectively.

Implementation Details. We use RoBERTa [50] (124.7M) as the text encoder and SAM2 (80.8M) as the base video tracker, both of which are frozen. Only the VC-CMA (~4.6M), Candidate Generator (~12.1M), BAR (~1.1M), and CGCT (~2.3M) are trained (~10% extra parameters). Under the weakly supervised setting, the Candidate Generator is distilled from APTM [51] and trained on ID-agnostic single-view pseudo labels (lr: 4e-5, 12 epochs). VC-CMA and BAR are pre-trained on Ref-YoutubeVOS [52] (lr: 2e-4, 8 epochs), and then fine-tuned on ID-agnostic cross-view pseudo labels (lr: 2e-5, 20 epochs). CGCT is trained on refer-agnostic cross-view pseudo labels (lr: 2e-4, 10 epochs). For inference, clip length T is 8. All experiments are conducted on 2 NVIDIA A800 GPUs.

4.2 State-of-the-art Comparison

As shown in Tab. 1, ViewSAM trained with weak supervision achieves the best overall performance among zero-shot and weakly supervised approaches, with improvements of at least **+3.36 in CVR-IDF1 and +3.55 in CVR-MA** across in-domain and cross-domain evaluations, while remaining competitive with fully supervised models. Compared with foundation models (i.e., SAM2 [17] initialized with VLM-generated box prompts, and SAM3 [18] using object category as prompts), ViewSAM improves both referring understanding and cross-view association. Moreover, CRTracker trained with the same pseudo labels achieves strong performance, **validating the effectiveness of our pseudo labels**, yet ViewSAM still surpasses it, highlighting the benefit of view-aware semantics.

Meanwhile, the fully supervised variant of ViewSAM further pushes the performance boundary, achieving improvements of at least **+2.85 in CVR-IDF1 and +6.03 in CVR-MA** across in-domain

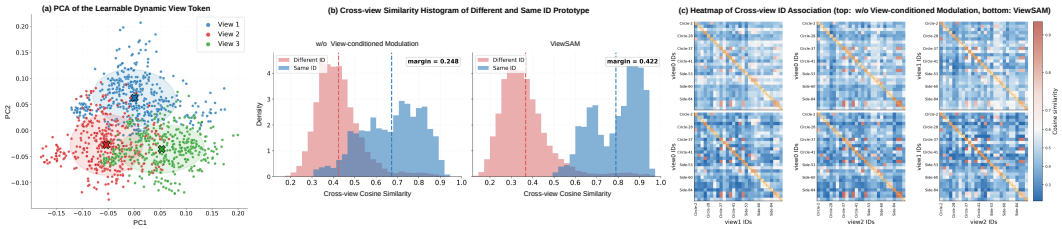

Figure 4: Qualitative analysis of view-aware cross-modal semantics. Please zoom in for better visualization. (a) PCA of the view token shows more separable camera-wise clusters. (b) The cross-view similarity distributions exhibit a larger margin between same-ID and different-ID pairs. (c) The association heatmaps demonstrate stronger diagonal dominance and reduced cross-ID interference.

Table 2: Contribution of main components.

Method	CVR-IDF1	CVR-MA
ViewSAM (full)	**57.73**	**43.72**
w/o VC-CMA	51.43	37.54
w/o CG	38.19	21.30
w/o BAR	53.55	40.02
w/o CGCT	46.50	30.49

Table 3: Detailed design ablation results.

Variant	CVR-IDF1	CVR-MA
w/o VCCA	53.62	38.45
w/o view modulation	52.12	35.77
w/o intra-/inter-loss	50.00	33.95

Table 4: Effect of pseudo-label strategies.

Pseudo Labels	CVR-IDF1	CVR-MA
SAM3 tracklets	39.53	20.33
+ Assoc.	67.87	51.61
+ Assoc. + Bi-RP	**77.96**	**68.44**

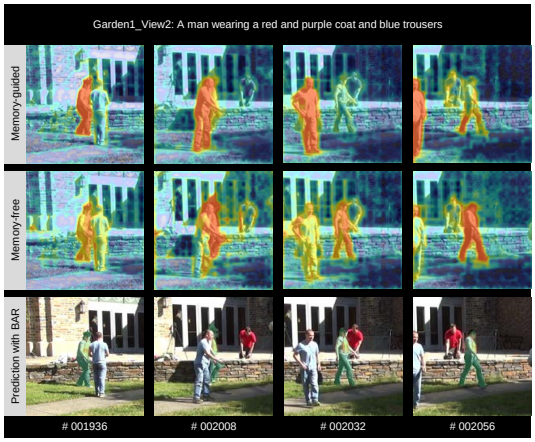

Figure 5: Visualizations on the effect of Bias-aware Recalibration module, which helps the model to re-focus on targets that better align with the referring expression by leveraging the memory-free features.

and cross-domain evaluations. This also confirms that the proposed view-aware design generalizes well across different supervision settings. The qualitative visualization results can be found in A.

4.3 Ablation Study and Analysis

To isolate the effect of individual model components from pseudo-label noise, the model ablations in Tabs. 2 and 3 are conducted under a fully supervised setting using ground-truth training annotations. This provides a cleaner comparison and makes the contribution of each design choice more discernible.

Impact of view-aware cross-modal semantics. We analyze our core contribution from two aspects: (1) the learnable dynamic view token in the VC-CMA and (2) the view-conditioned modulation in the CGCT. Tab. 3 shows that removing either design (i.e., w/o VCCA and w/o view modulation) leads to a clear drop, suggesting that the view-aware semantic learning helps reduce the gap between view-variant visual appearances and view-invariant textual semantics for CRMOT. To gain more insights, we present some qualitative results in Fig. 4. As shown in Fig. 4(a), the dynamic view token yields structured view-wise clusters in the PCA space, indicating that it has learned discriminative view-specific priors as intended. Building upon them, the view-conditioned modulation further aligns instance features across views while preserving identity discriminability. Accordingly, Fig. 4(b) shows that the margin between same-/different-ID similarities increases substantially, from 0.248 to 0.422. This is also evident in Fig. 4(c), where the association matrix exhibits a clearer diagonal pattern with weaker off-diagonal interference, reflecting more reliable cross-view ID matching.

Impact of main components and design details. As shown in Tab. 2, removing the View-conditioned Cross-Modal Alignment while retaining text prompts (w/o VC-CMA) leads to clear degradation, highlighting the importance of view conditioning for cross-modal alignment. Removing the Candidate

Generator (w/o CG) causes a dramatic performance drop, indicating that reliable multi-object proposals are essential for extending SAM2 from single object tracking to multi-object tracking. In contrast, removing the Bias-aware Recalibration (w/o BAR) results in a consistent decline, as also reflected in Fig. 5. Notably, the Consistency-guided Cross-view Tracking head is crucial for cross-view association: removing it (w/o CGCT) significantly degrades performance. As detailed in Tab. 3, discarding consistency objectives also leads to clear degradation.

Quality of cross-view pseudo labels. We select a proxy subset (~25% of all samples), where samples are selected based on the similarity of their referring to those in the original annotated training data. Hence, GTs are available for this subset to enable direct evaluation using CRMOT metrics. As shown in Tab. 4, affinity-guided association (Assoc.) significantly improves CVR-IDF1 from **39.53** to **67.87**, indicating enhanced identity consistency across views. Further introducing bi-directional re-prompting (Bi-RP) boosts both CVR-IDF1 to **77.96** and CVR-MA to **68.44**, demonstrating more accurate and reliable cross-view association. More qualitative results are provided in B.

5 Conclusion

We present the first weakly supervised CRMOT framework using only object category labels, substantially alleviating the reliance on expensive frame-level and cross-view annotations. Built on SAM2, ViewSAM learns view-aware cross-modal semantics and robust cross-view associations to effectively handle appearance variations across cameras. With only ~10% additional parameters, it achieves state-of-the-art performance under weak supervision while remaining competitive with fully supervised methods. Beyond the performance gains, our results demonstrate that CRMOT models can be effectively learned from coarse video-level supervision, opening a scalable and annotation-efficient path toward training on substantially larger collections of multi-view videos.

Acknowledgments and Disclosure of Funding

This work is supported by National Natural Science Foundation of China under Grants No. 62472092, No. 62172089, No. 62106045. National Science and Technology Major Project of China (No. 2026ZD161120). Natural Science Foundation of Jiangsu Province under Grants No. BK20241751. Jiangsu Provincial Key Laboratory of Computer Networking Technology. Jiangsu Provincial Key Laboratory of Network and Information Security under Grants No. BM2003201, and Key Laboratory of Computer Network and Information Integration of Ministry of Education of China under Grants No. 93K-9, Nanjing Purple Mountain Laboratories. We thank the Big Data Computing Center of Southeast University for providing the facility support on the numerical calculations.

Competing Interests. The authors declare no competing interests.

References

[1] Sijia Chen, En Yu, and Wenbing Tao. Cross-view referring multi-object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2204–2211, 2025.

[2] Dongming Wu, Wencheng Han, Tiancai Wang, Xingping Dong, Xiangyu Zhang, and Jianbing Shen. Referring multi-object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14633–14642, 2023.

[3] Tobias Fischer, Yung-Hsu Yang, Suryansh Kumar, Min Sun, and Fisher Yu. Cc-3dt: Panoramic 3d object tracking via cross-camera fusion. *arXiv preprint arXiv:2212.01247*, 2022.

[4] Filippo Ziliotto, Tommaso Campari, Luciano Serafini, and Lamberto Ballan. Tango: training-free embodied ai agents for open-world tasks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24603–24613, 2025.

[5] Shengyu Hao, Peiyuan Liu, Yibing Zhan, Kaixun Jin, Zuozhu Liu, Mingli Song, Jenq-Neng Hwang, and Gaoang Wang. Divotrack: A novel dataset and baseline method for cross-view multi-object tracking in diverse open scenes. *International Journal of Computer Vision*, 132(4):1075–1090, 2024.

[6] Yi Gao, Wanneng Wu, Ao Liu, Qiaokang Liang, and Jianwen Hu. Multi-target multi-camera tracking with spatial-temporal network. In *2023 7th International Symposium on Computer Science and Intelligent Control (ISCSIC)*, pages 196–200. IEEE, 2023.

[7] Yunfei Zhang, Jin Gao, Wenjuan Li, and Weiming Hu. Dual-head feature enhancement for graph-based cross-view multi-object tracking. In *International Conference on Artificial Neural Networks*, pages 643–655. Springer, 2025.

[8] Yihao Zhen, Mingyue Xu, Qiang Wang, Baojie Fan, Jiahua Dong, Tinghui Zhao, and Huijie Fan. Gmt: Effective global framework for multi-camera multi-target tracking. *arXiv e-prints*, pages arXiv–2407, 2024.

[9] Huijie Fan, Yu Qiao, Yihao Zhen, Tinghui Zhao, Baojie Fan, and Qiang Wang. All-day multi-camera multi-target tracking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16892–16901, 2025.

[10] Jiawei Ge, Jiuxin Cao, Xuelin Zhu, Xinyu Zhang, Chang Liu, Kun Wang, and Bo Liu. Consistencies are all you need for semi-supervised vision-language tracking. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1895–1904, 2024.

[11] Chang Xu, Dacheng Tao, Chao Xu, and Yong Rui. Large-margin weakly supervised dimensionality reduction. In *International conference on machine learning*, pages 865–873. PMLR, 2014.

[12] Lianghui Zhu, Junwei Zhou, Yan Liu, Xin Hao, Wenyu Liu, and Xinggang Wang. Weaksam: Segment anything meets weakly-supervised instance-level recognition. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 7947–7956, 2024.

[13] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53, 2018.

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[17] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pages 28085–28128, 2025.

[18] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

[19] Hyeokjun Kweon and Kuk-Jin Yoon. From sam to cams: Exploring segment anything model for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19499–19509, 2024.

[20] Chunming He, Kai Li, Yachao Zhang, Guoxia Xu, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. *Advances in Neural Information Processing Systems*, 36:30726–30737, 2023.

[21] Zhenyang Li, Ran Tao, Efstratios Gavves, Cees GM Snoek, and Arnold WM Smeulders. Tracking by natural language specification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6495–6503, 2017.

[22] Xiao Wang, Xiujun Shu, Zhipeng Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng Wu. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13763–13773, 2021.

[23] Li Zhou, Zikun Zhou, Kaige Mao, and Zhenyu He. Joint visual grounding and tracking with natural language specification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23151–23160, 2023.

[24] Mingzhe Guo, Zhipeng Zhang, Heng Fan, and Liping Jing. Divert more attention to vision-language tracking. volume 35, pages 4446–4460, 2022.

[25] Biao Wang, Wenwen Li, and Jiawei Ge. R1-track: Direct application of mllms to visual object tracking via reinforcement learning, 2025.

[26] Jiawei Ge, Jiuxin Cao, Xiangmei Chen, Xuelin Zhu, Weijia Liu, Chang Liu, Kun Wang, and Bo Liu. Beyond visual cues: Synchronously exploring target-centric semantics for vision-language tracking. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(5):1–21, 2025.

[27] Yunhao Du, Cheng Lei, Zhicheng Zhao, and Fei Su. ikun: Speak to trackers without retraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19135–19144, 2024.

[28] Yunhao Li, Xiaoqiong Liu, Luke Liu, Heng Fan, and Libo Zhang. Lamot: Language-guided multi-object tracking. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6816–6822. IEEE, 2025.

[29] Guangyao Li, Siping Zhuang, Yajun Jian, Yan Yan, and Hanzi Wang. Language decoupling with fine-grained knowledge guidance for referring multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23626–23635, 2025.

[30] Changcheng Xiao, Qiong Cao, Yujie Zhong, Xiang Zhang, Tao Wang, Canqun Yang, and Long Lan. Temporal-enhanced multimodal transformer for referring multi-object tracking and segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[31] Siping Zhuang, Guangyao Li, Qiangqiang Wu, Yang Lu, Hai-Miao Hu, and Hanzi Wang. Cgatracker: Correlation-aware graph alignment for referring multi-object tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[32] Shaofeng Liang, Runwei Guan, Wangwang Lian, Daizong Liu, Xiaolou Sun, Dongming Wu, Yutao Yue, Weiping Ding, and Hui Xiong. Cognitive disentanglement for referring multi-object tracking. *Information Fusion*, page 103349, 2025.

[33] Yizhe Li, Sanping Zhou, Zheng Qin, and Le Wang. Visual-linguistic feature alignment with semantic and kinematic guidance for referring multi-object tracking. *IEEE Transactions on Multimedia*, 2025.

[34] Haoran Wang, Dongliang He, Wenhao Wu, Boyang Xia, Min Yang, Fu Li, Yunlong Yu, Zhong Ji, Errui Ding, and Jingdong Wang. Coder: Coupled diversity-sensitive momentum contrastive learning for image-text retrieval. In *European conference on computer vision*, pages 700–716. Springer, 2022.

[35] Haoran Wang, Bo Zhao, Jinghui Wang, Hanzhang Wang, Huan Yang, Wei Ji, Hao Liu, and Xinyan Xiao. Sega: A stepwise evolution paradigm for content-aware layout generation with design prior. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19321–19330. IEEE, 2025.

[36] Hanzhang Wang, Haoran Wang, Zhongrui Yu, Mingming Sun, Junjun Jiang, Xianming Liu, and Deming Zhai. Fast: Flexibly controllable arbitrary style transfer via latent diffusion models. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(9):1–20, 2025.

[37] Xuelin Zhu, Xiu-Shen Wei, Jiawei Ge, Shuai Xu, and Bing Wang. Rethinking multi-label image classification with deep learning: Taxonomy, challenge, and outlook. *arXiv preprint arXiv:2607.00839*, 2026.

[38] Xuelin Zhu, Jian Liu, Dongqi Tang, Weijia Liu, Jiawei Ge, Bo Liu, and Jiuxin Cao. Multi-label image classification via contrastive co-occurrence learning. *IEEE Transactions on Image Processing*, 2026.

[39] Yuyuan Li, Yizhao Zhang, Weiming Liu, Xiaohua Feng, Zhongxuan Han, Chaochao Chen, and Chenggang Yan. Multi-objective unlearning in recommender systems via preference guided pareto exploration. *IEEE Transactions on Services Computing*, 2025.

[40] Yuyuan Li, Xiaohua Feng, Chaochao Chen, and Qiang Yang. A survey on recommendation unlearning: Fundamentals, taxonomy, evaluation, and open questions. *IEEE Transactions on Knowledge and Data Engineering*, 38(2):781–799, 2025.

[41] Chaochao Chen, Yizhao Zhang, Yuyuan Li, Jun Wang, Lianyong Qi, Xiaolong Xu, Xiaolin Zheng, and Jianwei Yin. Post-training attribute unlearning in recommender systems. *ACM Transactions on Information Systems*, 43(1):1–28, 2024.

[42] Chen Feng, Zhuo Zhi, Zhao Huang, Jiawei Ge, Ling Xiao, Nicu Sebe, Georgios Tzimiropoulos, and Ioannis Patras. Deconstructing the failure of ideal noise correction: A three-pillar diagnosis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 34512–34523, 2026.

[43] Cheng-Yen Yang, Hsiang-Wei Huang, Wenhao Chai, Zhongyu Jiang, and Jenq-Neng Hwang. Samurai: Adapting segment anything model for zero-shot visual tracking with motion-aware memory. *arXiv preprint arXiv:2411.11922*, 2024.

[44] Shuangrui Ding, Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Yuwei Guo, Dahua Lin, and Jiaqi Wang. 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*, pages 13614–13624, 2025.

[45] Junjie Jiang, Zelin Wang, Manqi Zhao, Yin Li, and DongSheng Jiang. Sam2mot: A novel paradigm of multi-object tracking by segmentation. *arXiv preprint arXiv:2504.04519*, 2025.

[46] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3702–3712, 2019.

[47] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[48] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[49] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3942–3951, 2018.

[50] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[51] Shuyu Yang, Yinan Zhou, Zhedong Zheng, Yaxiong Wang, Li Zhu, and Yujiao Wu. Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In *Proceedings of the 31st ACM international conference on multimedia*, pages 4492–4501, 2023.

[52] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *European conference on computer vision*, pages 208–223. Springer, 2020.

A Qualitative Results for Comparing with SOTAs

To provide intuitive insights beyond quantitative comparisons, we present qualitative results of ViewSAM against representative state-of-the-art methods. As illustrated in Fig. 6 and Fig. 7, we visualize cross-view tracking performance under challenging scenarios, including severe occlusion, appearance ambiguity, and large viewpoint variations. Existing methods often suffer from drift to distractors, inaccurate grounding of referring expressions, or inconsistent identity assignment across views.

In contrast, ViewSAM achieves more precise localization and preserves identity consistency across cameras, even in the presence of heavy occlusions and significant viewpoint changes. This highlights the advantage of incorporating view-aware cross-modal semantics to effectively bridge view-variant visual observations with view-invariant textual semantics, leading to more robust cross-view reasoning and association.

Notably, targets with different IDs are annotated in distinct colors for clarity, while **red indicates incorrect predictions**. Specifically, *MISS* denotes missed detections, whereas the remaining red cases correspond to identity assignment errors, such as ID switches or incorrect cross-view associations.

B Qualitative Results of Cross-view Pseudo Labels

We further visualize the generated cross-view pseudo labels to better understand their role in weakly supervised learning. As shown in Fig. 8, the pseudo labels exhibit strong temporal coherence, semantic consistency, and cross-view identity alignment. Specifically, objects maintain stable spatial trajectories across consecutive frames, indicating that the generated tracklets are temporally smooth and robust to motion variations. Moreover, their semantic consistency across views suggests that the pseudo labels capture meaningful object-level cues rather than view-specific artifacts. Importantly, identities are well aligned across different camera views, demonstrating the effectiveness of our cross-view association and re-prompting strategy in mitigating identity ambiguity under weak supervision.

In addition, the pseudo labels remain reliable in challenging scenarios, such as occlusions, scale variations, and viewpoint changes, where naïve applications of foundation models often fail. This highlights the advantage of our affinity-guided cross-view re-prompting mechanism in refining noisy predictions and enforcing structural consistency across views.

Overall, these results demonstrate that our pseudo-label generation pipeline produces high-quality supervision signals that are temporally coherent, semantically consistent, and structurally aligned across views, thereby providing a strong foundation for training the downstream ViewSAM model under the weakly supervised setting.

C Details of Main Components

To complement Sec. 3.3, we provide additional details on the main components of ViewSAM. For consistency with the main paper, this section follows the same order as the inference pipeline: **View-conditioned Cross-modal Alignment (VC-CMA)**, **Candidate Generator**, **Bias-aware Recalibration (BAR)**, and **Consistency-guided Cross-view Tracking Head (CGCT)**. Here, we emphasize only on the architecture and optimization of these modules.



Figure 6: Visualization of comparison results on the in-domain scenes.

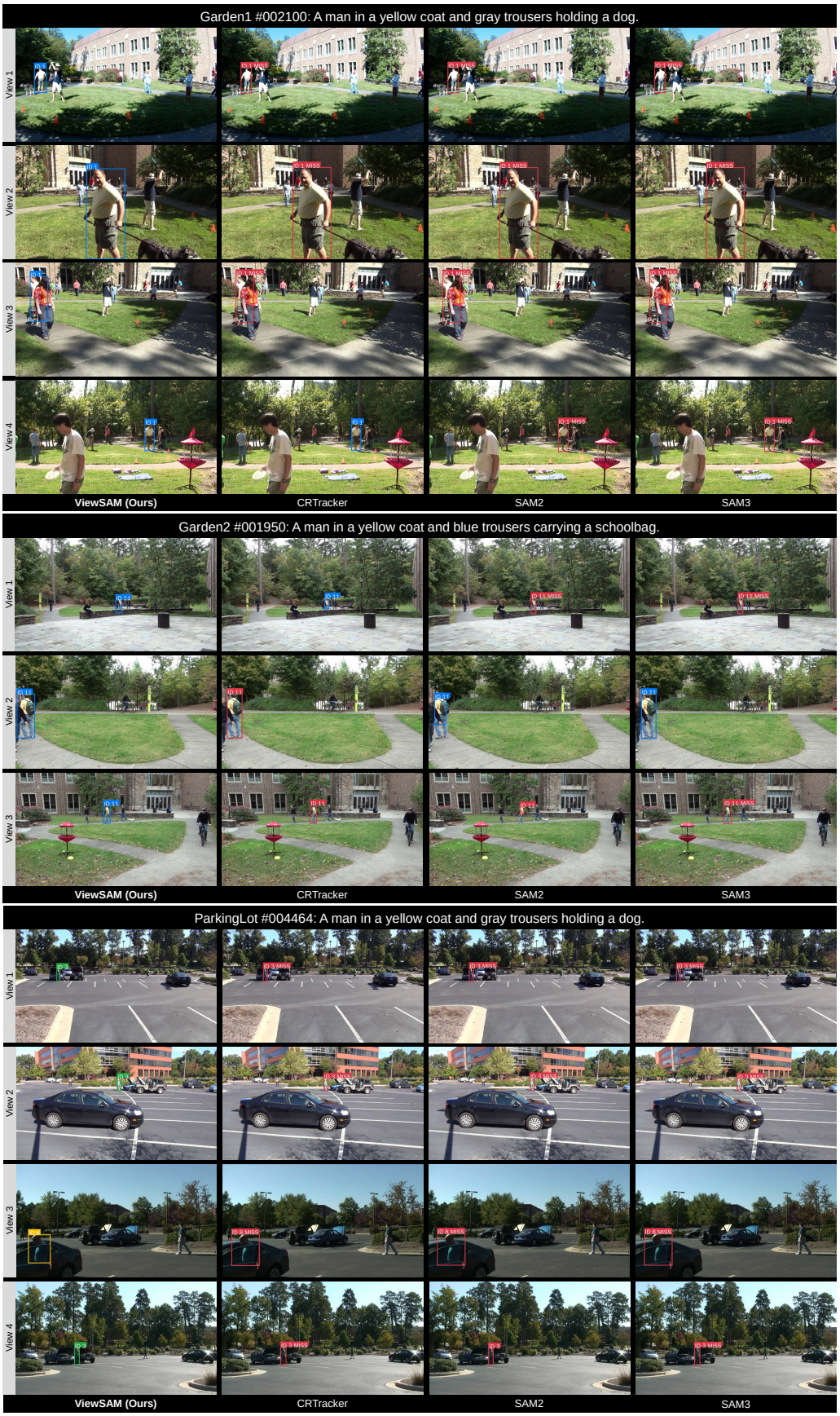

Figure 7: Visualization of comparison results on the cross-domain scenes.



Figure 8: Visualization of generated cross-view pseudo labels for CRMOT, illustrating temporal coherence, semantic consistency, and cross-view identity alignment.

C.1 View-conditioned Cross-modal Alignment

As introduced in Sec. 3.3, VC-CMA is designed to inject view-aware cross-modal semantics into the SAM2-based video tracker. It is implemented as a lightweight adaptation module [48] and inserted into selected fusion stages. At each stage, compressed visual tokens interact with textual tokens through view-conditioned cross-attention, where a dynamic View Token provides view-dependent priors for cross-modal alignment.

The dynamic View Token serves as a conditioning signal rather than an independent prediction head. It does not directly produce localization or identity outputs; instead, it modulates visual–text representations to improve their robustness to viewpoint-induced appearance variations. In this way, VC-CMA bridges the gap between view-variant visual observations and the relatively view-invariant semantics of referring expressions.

Since VC-CMA is an adaptation module within the SAM2 mask decoding pipeline, it is trained through the standard mask supervision used by SAM2. Given the predicted mask $\hat{M}_t$ and the pseudo target mask M_t, we optimize VC-CMA using a combination of focal loss and Dice loss:

$$\mathcal{L}_{\text{VC-CMA}} = \mathcal{L}_{\text{focal}}(\hat{M}_t, M_t) + \mathcal{L}_{\text{dice}}(\hat{M}_t, M_t). \tag{13}$$

This objective encourages the view-conditioned cross-modal features to improve mask prediction under weak supervision. Through end-to-end optimization, VC-CMA learns view-aware visual–text alignment that is beneficial for robust referring segmentation across different camera views.

C.2 Candidate Generator

The Candidate Generator is distilled from APTM [51] and extends SAM2 from single-object tracking to referring multi-object localization. Specifically, we transfer proposal-level localization knowledge by distilling the distribution of proposal quality from APTM, where both proposal confidence

and relative ranking are used as supervision signals. This enables the model to inherit strong object localization priors without relying on manual annotations. To further accommodate the weakly supervised CRMOT setting, the model is trained using ID-agnostic single-view pseudo labels generated in Sec. 3.3.

Given the view-aware features produced by VC-CMA, the Candidate Generator predicts a set of candidate boxes, which are subsequently used as box prompts for SAM2 mask decoding. Following proposal-based referring localization paradigms, it consists of two branches: a proposal generation branch for dense object localization, and a proposal–referring matching branch for measuring compatibility with the referring expression.

Proposal generation branch. The proposal branch adopts a dense prediction formulation with an objectness tower and a regression tower, predicting the center heatmap, centerness, box size, and center offset, respectively. Let $\hat{H}$, $\hat{C}$, $\hat{W}$, and $\hat{O}$ denote the predicted center heatmap, centerness map, box size, and center offset. The objective is defined as

$$\mathcal{L}_{\text{prop}} = \lambda_{\text{ctr}}\mathcal{L}_{\text{focal}}(\hat{H}, H) + \lambda_{\text{ctrness}}\mathcal{L}_{\text{focal}}(\hat{C}, C) + \lambda_{\text{wh}}\mathcal{L}_{\text{s-L1}}(\hat{W}, W) + \lambda_{\text{off}}\mathcal{L}_{\text{s-L1}}(\hat{O}, O), \quad (14)$$

where $\mathcal{L}_{\text{focal}}$ is the sigmoid focal loss and $\mathcal{L}_{\text{s-L1}}$ is the Smooth-L_1 loss.

Proposal–referring matching branch. To bridge proposal localization and language grounding under weak supervision, each proposal is assigned a matching score with respect to the referring expression. Specifically, proposal features are extracted from multi-scale visual representations, fused with the referring embedding and box geometry, and mapped to proposal-level matching logits. Let s_i denote the matching logit of proposal i.

We construct soft supervision targets from pseudo labels:

$$y_i = \text{clip}\left(\frac{\text{IoU}_i - \tau_{\text{neg}}}{\tau_{\text{pos}} - \tau_{\text{neg}}}, 0, 1\right), \quad (15)$$

where IoU_i denotes the overlap between proposal i and the pseudo target, and τ_{pos} and τ_{neg} are positive and negative thresholds. The matching loss is

$$\mathcal{L}_{\text{match}} = \lambda_{\text{match}}\mathcal{L}_{\text{gbce}}(\{s_i\}, \{y_i\}), \quad (16)$$

where $\mathcal{L}_{\text{gbce}}$ denotes a group-balanced binary cross-entropy loss that normalizes positive and negative proposal groups separately.

To further enlarge the margin between valid and invalid proposals, we introduce an auxiliary ranking objective that also facilitates distribution distillation from APTM:

$$\mathcal{L}_{\text{rank}} = \frac{1}{|\mathcal{P}||\mathcal{N}|}\sum_{i\in\mathcal{P}}\sum_{j\in\mathcal{N}}\max(0, m - (s_i - s_j)), \quad (17)$$

where $\mathcal{P}$ and $\mathcal{N}$ denote the positive and negative proposal sets derived from pseudo labels. The final objective is

$$\mathcal{L}_{\text{cand}} = \mathcal{L}_{\text{prop}} + \mathcal{L}_{\text{match}} + \lambda_{\text{rank}}\mathcal{L}_{\text{rank}}. \quad (18)$$

In practice, the Candidate Generator serves as the bridge between view-aware representations and SAM2 mask decoding, transferring proposal distribution knowledge distilled from APTM into reliable box prompts, and enabling the transition from single-object tracking to referring multi-object localization under weak supervision.

C.3 Bias-aware Recalibration

BAR is introduced to mitigate tracking bias caused by error accumulation in memory-guided tracking in SAM2. Memory-guided decoding may amplify early mistakes and gradually drift to distractors, especially under occlusion or severe appearance ambiguity. To address this issue, BAR learns to detect when the memory-guided prediction becomes unreliable by comparing it with a memory-free prediction that does not use the tracking history stored in the Memory Bank.

Specifically, for each target candidate, we compute two masks. The first mask is predicted from the memory-guided branch:

$$\hat{M}_t^{\text{b}} = \mathcal{D}_{\text{dec}}(\mathcal{F}_{\text{mem}}, \rho) > 0, \quad (19)$$

where $\mathcal{F}_{\text{mem}}$ denotes the memory-enhanced features and ρ is the prompt. This prediction corresponds to the standard SAM2 decoding process conditioned on the Memory Bank.

To obtain an unbiased reference, we further compute a memory-free mask:

$$\hat{M}_t^{\text{u}} = \mathcal{D}_{\text{dec}}(\mathcal{F}, \rho) > 0, \tag{20}$$

where $\mathcal{F}$ denotes the features without memory guidance. Since this branch does not access previous tracking context, its prediction is determined mainly by the current-frame visual content and the referring prompt.

Given the two binary masks at frame t, we define the bias supervision label as:

$$y_t = \begin{cases} 1, & \text{if } \hat{M}_t^{\text{u}} \cap \hat{M}_t^{\text{b}} = \emptyset, \\ 0, & \text{otherwise.} \end{cases} \tag{21}$$

Here, $y_t = 1$ indicates that the memory-guided and memory-free branches segment different objects, suggesting that the memory-guided prediction may have drifted to a distractor. We supervise the predicted bias probability $\hat{b}_t$ using a standard cross-entropy loss:

$$\mathcal{L}_{\text{bar}} = -\frac{1}{T} \sum_{t=1}^{T} \left[y_t \log(\hat{b}_t) + (1 - y_t) \log(1 - \hat{b}_t) \right]. \tag{22}$$

Through this supervision, BAR learns to identify frames where memory guidance becomes unreliable and adaptively recalibrates the final mask prediction toward the memory-free branch when necessary.

C.4 Consistency-guided Cross-view Tracking Head

The CGCT head is responsible for cross-view identity association. It is built upon an OSNet [46] to capture appearance features for robust representation learning. As described in Sec. 3.3, it first extracts object-level representations via masked pooling over multi-scale features and projects them into a shared embedding space for identity matching. To alleviate appearance discrepancies across cameras, the dynamic View Token is injected through view-conditioned FiLM modulation, producing embeddings that are view-aware during feature adaptation while encouraging view-invariant representations for identity association.

The core supervision of CGCT follows a consistency-guided formulation that explicitly models both intra-view and inter-view identity structures. Let $z_{t,i}^k$ denote the embedding of instance k at frame t and view i, $\bar{z}_g^i$ the prototype of identity g in view i, and $\bar{z}_g$ the global prototype aggregated across views. The training objective is

$$\mathcal{L}_{\text{cgct}} = \sum_g \left[\sum_i \sum_{(t,k) \in \mathcal{T}_g^i} \left\| z_{t,i}^k - \bar{z}_g^i \right\|_2^2 + \lambda \sum_i \left\| \bar{z}_g^i - \bar{z}_g \right\|_2^2 \right], \tag{23}$$

where the first term enforces intra-view consistency by compacting embeddings belonging to the same trajectory within each camera view, while the second term enforces inter-view consistency by aligning view-specific prototypes toward a shared global identity prototype.

In contrast to conventional association heads that rely solely on appearance similarity, CGCT explicitly models a hierarchical consistency structure, capturing temporal coherence within each view and identity agreement across views. This formulation is particularly suited to the weakly supervised setting, where explicit cross-view identity annotations are unavailable. By leveraging pseudo supervision and consistency constraints, CGCT learns to suppress view-induced discrepancies and establish robust identity associations across cameras.

D Limitations

Despite its effectiveness, our framework has several limitations. First, it relies on pseudo labels generated in Stage 1, whose quality is bounded by the performance of SAM3 and cross-view association; errors may propagate to downstream training, especially in crowded scenes or under severe occlusion. Moreover, the weak supervision setting still assumes synchronized multi-view

videos with category labels as coarse-grained supervision signal, which may limit applicability in more unconstrained scenarios. Meanwhile, the pipeline depends on multiple external components (e.g., SAM3, ReID models, and MLLMs), increasing system complexity and computational cost. Finally, while ViewSAM models view-aware cross-modal semantics, it remains challenged by ambiguous referring expressions and complex multi-object interactions, leaving room for improving language grounding and temporal reasoning.

E Impact

This work contributes to reducing the reliance on dense annotations for cross-view referring multi-object tracking, which may facilitate scalable research and applications in multi-camera perception systems such as intelligent surveillance and embodied AI. However, the ability to track objects across views based on language descriptions also raises potential privacy and misuse concerns, particularly in surveillance scenarios. Moreover, biases inherited from foundation models or pseudo-label generation may affect fairness and reliability. We emphasize that this work is intended for research purposes, and any real-world deployment should comply with ethical standards, privacy regulations, and appropriate human oversight.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately mirror the paper's contributions by clearly stating the development of ViewSAM, its novel WSCRMOT paradigm, and associated performance enhancements, which align with the content and findings detailed throughout the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the current limitations of ViewSAM and prospects for future research in D.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: ViewSAM doesn't involve theoretical results.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have provided detailed technical details necessary for reproducing the experiments.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The source code is not publicly available. However, detailed implementation information is provided in Appendix C to facilitate reproducibility, and all datasets used in this work are publicly available.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We have provided a detailed introduction of various parameters and details in the experiment section.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The experimental errors of ViewSAM are simultaneously affected by different stages. Therefore, due to the constraints of computational costs, we did not provide error bars. However, based on the given intermediate results, the performance of ViewSAM is stable.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have provided the computer resources in the experiment section.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research is fully in compliance with ethical standards.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the impacts of ViewSAM in E.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All the data in the paper are from open-source and accessible datasets.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We will not release new assets.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowd sourcing nor research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowd sourcing nor research with human subjects.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We employ Qwen3-VL-8B for trajectory-level description generation and referring-level grouping during pseudo-label generation, as described in Sec. 3.2.2.