STORM: End-to-End Referring Multi-Object Tracking in Videos

Zijia Lu* Jingru Yi Jue Wang Yuxiao Chen Junwen Chen Xinyu Li Davide Modolo
Amazon
lu.zij@northeastern.edu {jyijingr,juewangn,cyuxiao,chenjunw,xxnl,dmodolo}@amazon.com

Abstract

*Referring multi-object tracking (RMOT) is a task of associating all the objects in a video that semantically match with given textual queries or referring expressions. Existing RMOT approaches decompose object grounding and tracking into separated modules and exhibit limited performance due to the scarcity of training videos, ambiguous annotations, and restricted domains. In this work, we introduce **STORM**, an end-to-end MLLM that jointly performs grounding and tracking within a unified framework, eliminating external detectors and enabling coherent reasoning over appearance, motion, and language. To improve data efficiency, we propose a task-composition learning (TCL) strategy that decomposes RMOT into image grounding and object tracking, allowing STORM to leverage data-rich subtasks and learn structured spatial–temporal reasoning. We further construct **STORM-Bench**, a new RMOT dataset with accurate trajectories and diverse, unambiguous referring expressions generated through a bottom-up annotation pipeline. Extensive experiments show that STORM achieves state-of-the-art performance on image grounding, single-object tracking, and RMOT benchmarks, demonstrating strong generalization and robust spatial–temporal grounding in complex real-world scenarios. STORM-Bench is released at https://github.com/amazon-science/storm-referring-multi-object-grounding.*

1. Introduction

Identifying and localizing objects in videos is essential for various vision applications, including video understanding, human–object interaction analysis, and embodied perception. Precise spatial–temporal localization enables models to reason about how objects move and interact over time. In practice, however, users are typically interested in only a subset of objects relevant to their intent. This motivates the task of *referring object grounding* [35, 42, 59], where the model detects objects that semantically matched with given referring expressions. Language offers a flexible interface for specifying appearance, attributes, and spatial relations, making referring grounding more practical than category-based detection or generic tracking.

Early research efforts [12, 13, 42, 63] primarily addressed referring single-object grounding before extending the idea to tracking, where a model localizes and follows one target based on a textual query. While approaches like referring single-object tracking (RSOT) [52, 55] show strong performance in language-guided tracking, they are inherently limited to a single target and fail to handle multiple interacting objects or relational descriptions. To this end, we explore the more challenging problem of referring multi-object tracking (RMOT) in this work, which seeks to ground and track all objects referenced by textual queries throughout a video.

Prior RMOT approaches [5, 8, 36, 58, 67, 68] typically augment a specialized detector with text encoders, but such systems struggle to interpret complex referring expressions and cannot reason about causal or relational dependencies conveyed in language. Large language models (LLMs), in contrast, excel at understanding nuanced linguistic structures. Most Multi-modal large language models (MLLMs)-based grounding methods [52, 57] operate on static images, and its extensions [2, 11] to videos typically attach an external tracker or detector. However, these methods leverage the LLMs as standalone text encoder and require additional detection module, preventing the model from learning unified spatial-temporal representations. As a result, no existing approach tackle the RMOT in an end-to-end manner.

To fill this gap, we introduce the **S**patial–**T**emporal **O**bject **R**eferential **M**odel (STORM), the first end-to-end multi-modal large language model designed specifically for referring multi-object tracking (RMOT). STORM unifies grounding and tracking within a single MLLM framework, without relying external detectors or trackers. The model utilizes a ViT-based visual encoder to extract spatial features from individual frames, while the LLM leverages these features together with the text query to capture temporal dynamics and reason about cross-modal correspondence between linguistic descriptions and visual entities. It then

*Work conducted during an internship at Amazon.

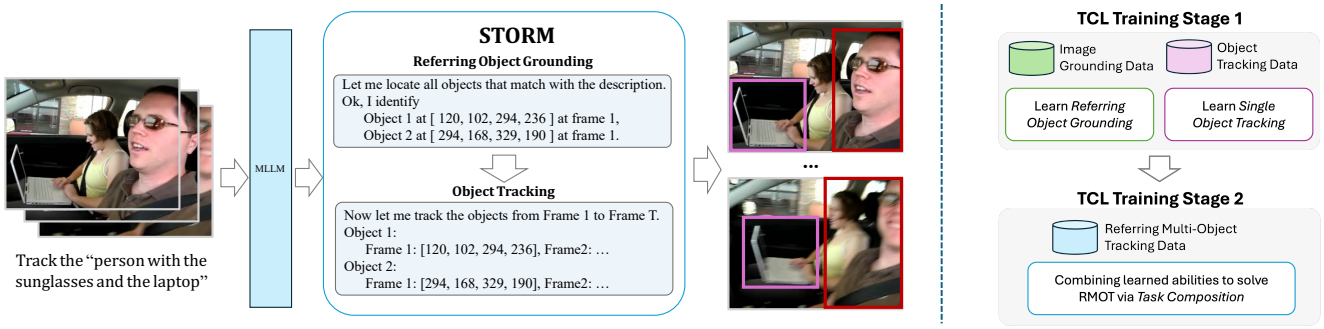

Figure 1. Overview of STORM method. The STORM adopts a LLaVA-style MLLM architecture [41], enabling video understanding on top of a large language model (LLM). Due to the scarcity of RMOT data, we propose a Task-Composition Learning (TCL) recipe. As illustrated in the figure on the right, TCL employs a two-stage training process. Stage 1 transfers subtask knowledge (i.e., object grounding and single-object tracking, SOT) to the referring object tracking task. In Stage 2, the model is fine-tuned on our proposed STORM-Bench dataset, further enhancing performance on the referring multi-object tracking (RMOT) task with limited videos.

directly generates bounding boxes for the text-described objects in a structured plain-text format, allowing the model to fully leverage pretrained language reasoning while avoiding the complexity and computational cost of additional external modules.

As is pointed out in the recent studies [1, 22, 26], multi-model large language models (MLLMs) are known to require massive amounts of high-quality training data to achieve strong performance. However, constructing large-scale datasets for referring multi-object tracking is highly impractical. Inspired by perception learning in the LLM pre-training (PT) stage and skill acquisition in the supervised fine-tuning (SFT) stage [40], we argue that RMOT can likewise be approached by first learning fundamental capabilities from easily accessible data, followed by a small amount of task-specific fine-tuning. To this end, we propose *task-composition learning (TCL)* strategy in STORM, which decompose RMOT as two fundamental tasks: grounding objects according to a textual description and maintaining their identities across time. Specifically, STORM is first trained on large-scale datasets for image grounding and single-object tracking to learn robust cross-modal alignment and temporal consistency. We then fine-tune the model with reasoning-based supervision that guides it to identify all referenced objects in the initial frame and subsequently track them over time. This training paradigm reduces reliance on RMOT-specific annotations and enhances generalization to complex, relational, and ambiguous referring expressions.

Although TCL significantly reduces the need for RMOT-specific data, existing RMOT datasets remain far from satisfactory: they are small in scale, noisy, and lack sufficient diversity. To address these limitations, we construct STORM-Bench, a new dataset featuring accurate multi-object trajectories and diverse, unambiguous referring expressions. STORM-Bench is created through a bottom-up annotation pipeline that first generates and verifies object-level descriptions, and then composes multi-object expressions via controlled language reasoning. This process ensures that the final expressions capture the object attributes, spatial relationships, and temporal interactions essential for referring multi-object tracking.

In summary, our main contributions in this work are as follows:

- We propose STORM, the first end-to-end MLLM framework for referring multi-object tracking that eliminates the need for external grounding modules.
- We introduce a task composition learning (TCL) strategy—a simple yet effective framework that substantially reduces the reliance on RMOT-specific data. TCL enables an MLLM to acquire RMOT capabilities by first learning from widely available subtask datasets and then applying a small amount of RMOT-focused fine-tuning.
- We introduce STORM-Bench, a high-quality and challenging dataset for referring multi-object grounding task, which includes 0.2M diverse referring expressions and 73.7K tracked objects.
- We show STORM achieves the state-of-the-art performance on various benchmarks of object grounding, single object tracking, and referring multi-object tracking.

2. Related Works

Object Grounding. Object grounding is a task that precisely localizes objects referred in a given natural language from images or videos. Object detection-based methods (such as MDETR [25], GLIP [31], Grounding DINO [42]) align object and text features in embedding space and subsequently retrieve object-text pairs. To support timestamp detection, Video Temporal Grounding (VTG) [9, 28, 38]) are proposed to further extend grounding task for videos and support substream tasks such as time retrieval, video summarization, highlight detection. In recent years, LLM-based visual-language models have integrated spatial-temporal grounding (STG) task and serve for zero-shot grounding

with complex languages [20, 23, 55, 62]. Object grounding with precise coordinates are further introduced in MLLM models [2, 12, 13, 46, 55, 56, 66].

Referring Object Tracking. Object tracking associates objects across video frames without specific user guidance. Referring Object Tracking merges object grounding and object tracking and only tracks objects specified in input textual queries [11, 59]. In traditional object tracking, similarity-learning is commonly adopted in single object tracking (SOT) task [7, 29, 30]. Multi-object tracking (MOT) task generally employs tracking-by-detection paradigm and utilizes additional tracker head to associate object bounding boxes [5, 8, 36, 58, 67, 68]. To mitigate the limited object categories and further support referring object tracking, researchers have utilized visual-language models [18, 32, 34, 65, 71] for open-set object association and classification. LLM-based referring object tracking are explored in recent works [52, 57]. In particular, ReasoningTrack [57] incorporated a tracking head in Qwen2.5-VL [2] and iteratively update single object tracklets. ReferGPT [11] proposed a matching module on top of MLLM models to support zero-shot multi-object tracking for cityscape driving scenes. ChatTracker [52] enhances tracking ability with improved data annotations. Elysium [55] constructs a ElysiumTrack-1M dataset to support SOT and Referring SOT (RSOT) tasks on top of MLLM models. Referring MOT (RMOT) [11, 16, 21, 37, 59, 70] is proposed to associate all objects semantically matched with referring expressions.

Referring Object Tracking Datasets. The existing SOT [48, 61] or MOT [19, 45] datasets generally lack referring expressions and have limited categories. Moreover, the small quantity of videos make them inadequate for large-language model finetuning. To support referring understanding, several object grounding datasets are proposed for image-based (Flickr30k [64], RefCOCO [27], RefCOCO+ [27], and RefCOCOg [44]) and video-based (OTB99 [34], Cityscapes-Ref [54], Talk2Car [15], Refer-Youtube-VOS [50]) referring tasks. More recently, RSOT [17] and RMOT [33] datasets incorporate textual annotations but suffer from annotation quality issues. For instance, LaSOT [17, 55] provides text descriptions for single-object tracking, yet videos may contain multiple instances of the same category, and the referring expression fails to specify which instance should be tracked. LaMOT [33] represents one of the earliest efforts to build MOT datasets with referring expressions but still exhibits low annotation diversity and incomplete object coverage. Elysium [55] proposes a larger-scale referring single-object tracking dataset based on WebVid videos [4] while the dataset has bunch of misaligned or erroneous bounding boxes that do not accurately match the referring expression. Existing RMOT works are generally focused on cityscape scenes (Ref-KITTI [59, 70]). In this work, we propose a new dataset STORM-Bench which corrects the ambiguity in existing RMOT datasets and extends RMOT to various domains.

3. Referring Multi-Object Tracking

In this section, we first introduce the problem definition. Next, we present the design of our Spatial-Temporal Object Referring Model (STORM). Finally, we describe the construction of our collected dataset (STORM-Bench) in detail.

3.1. Problem Definition

Referring multi-object tracking (RMOT) aims to track all objects throughout a video that semantically align with specified textual queries or referring expressions. Given a video $\mathcal{V} = \{I_t, t \in \{1, \ldots, T\}\}$ consisting of T frames and a referring expression $\mathcal{R}$ describing one or more objects of interest, the model extracts the spatial locations $\mathcal{B} = \{B_t^k, t \in \{1, \ldots, T\}, k \in \{1, \ldots, K\}\}$ of all corresponding objects K across video frames T.

3.2. STORM

Unlike existing works [3, 11, 39, 47, 59, 70] that decompose grounding and tracking into separate modules, we unify RMOT within a single multi-modal large language model (MLLM) framework, leveraging the strong reasoning capabilities of LLMs and maintain semantic consistency of objects across temporal frames. We name our proposed model as STORM. In addition, due to scarcity of RMOT datasets, we design a task-composition learning (TCL) method that allows model to leverage knowledge learned from sub-tasks such as image grounding and SOT, and thereby generalize to RMOT task.

Architecture. STORM's model architecture follows the common LLaVA-style MLLM design [41]. In particular, STORM extracts the frame-based visual tokens $\mathcal{V}$ through a ViT-based vision encoder. A two-layer MLP projector is followed to map the visual tokens into text space. The constructed visual tokens and referring text query tokens are sent to the a LLaMA-based LLM [41] which autoregressively generates the RMOT outputs $\mathcal{B}$. This formulation enables us to train the model as a next-token prediction task using cross-entropy loss.

Prompt and RMOT Output Format. We adopt a consistent prompting format for $\mathcal{R}$: *"<video> Please locate all objects in the video based on this expression: [referring expression]."*. The RMOT output response is formated as: *"Object 1: Frame 1: [x1, y1, x2, y2], Frame 2: [x1, y1, x2, y2], ...; Object 2: ..."*, where (x_1, y_1, x_2, y_2) denotes the absolute coordinates of a bounding box. When objects are temporarily absent from a frame due to occlusion or camera movement, STORM outputs an empty bounding box to indicate that the object is unobserved in that frame. For

long videos, we split the input into shorter clips and stitch the resulting tracklets by using the predicted boxes from the previous clip as prompts for the next one.

Task-Composition Learning (TCL). Training an MLLM for a new task typically requires large-scale annotated datasets. However, collecting large-scale RMOT training data is prohibitively expensive, as it demands frame-level object annotations paired with natural language descriptions that precisely reference specific objects over time. To address this challenge, we adopt a task-composition strategy that decomposes the complex RMOT task into simpler subtasks with abundant data, allowing the model to be trained on large-scale datasets for each component task. Specifically, we first train the model on extensive object grounding and object tracking datasets to establish strong visual–linguistic and temporal tracking capabilities. We then progressively extend the model's competence to RSOT and ultimately to RMOT using smaller, task-specific datasets. This staged training pipeline enables effective RMOT performance while significantly reducing the need for large-scale RMOT annotations. The detailed TCL training recipe is provided below. In particular, our training process consists of two stages:

Stage1: Pretraining on image grounding and object tracking. We train STORM on large-scale referring image grounding datasets, enabling it to associate textual expressions with spatial visual regions and predict object bounding boxes in single frames. This stage equips the model with strong cross-modal alignment between appearance cues and linguistic descriptions. In addition, we incorporate large-scale single-object tracking (SOT) datasets to help the model learn temporal consistency and motion representations across frames. Finally, we incorporate referring single-object tracking (RSOT) datasets, which help the model integrate its learned visual–linguistic alignment with temporal tracking capabilities, effectively teaching the LLM to perform language-guided single-object tracking.

Stage2: Finetuning with STORM-Bench. Finally, we finetune the model using our curated STORM-Bench dataset. To encourage explicit reasoning, we employ a *Chain-of-Thought (CoT)* training strategy. The model is guided to generate intermediate reasoning traces, first grounding the described objects in the initial frame and then tracking them sequentially throughout the video. This thinking process allows the model to learn the structured composition of the two sub-tasks, leading to improved performance and interpretability. After the reasoning phase, the model outputs object locations following the structured format described above.

This multi-stage training scheme in TCL allows the model to effectively transfer knowledge from abundant image grounding and tracking datasets, while requiring only a modest amount of referring multi-object data for adaptation.

Figure 2. Visualization of videos from our STORM-Bench dataset. Bounding boxes in different colors represent different tracklets. The dataset includes a wide range of challenging scenarios: top—significant scale changes for the same target; middle—multi-object tracking with clear prompts under weaker conditions; and bottom—multi-target tracking of similar object types but different instances under stronger, more specific conditions.

Empirically, we find that TCL substantially improves generalization to complex referring expressions and enhances robustness to ambiguous or previously unseen object descriptions.

4. STORM-Bench

4.1. Dataset Construction

Training and evaluating the proposed STORM model requires a dataset with high-quality referring expressions and accurate object bounding boxes across video frames. While our task-composition learning (TCL) strategy reduces the need for large-scale RMOT training data, using existing RMOT datasets [33] may yield limited performance due to ambiguous and inaccurate descriptions of the target objects [52]. In this work, we introduce a new dataset STORM-Bench, featuring carefully designed referring expressions to support both the training and evaluation of referring multi-object tracking methods.

Bottom-up annotation pipeline. To support tracking tasks in MLLM, existing work Elysium [55] proposed a *top-down* pipeline for million-scale RSOT data annotations: First, it obtains the caption of a video and generates bounding boxes for the nouns in the caption by Grounding-DINO. Second, the generated bbox-object pairs are used to create tracklets of each object across frames by a tracking model e.g. Mixformer. The top-down pipeline is highly sensitive to the performance of the expert detector and the tracker. In contrast, we propose a *bottom-up* data annotation pipeline. We leverage the asymmetry in task difficulty—while locating an ob-

	# Expressions	# Words	# Length	MO	CE
LaSOT [17]	1.4K	9.8K	7.0	×	×
Elysium [55]	1.3M	2.7M	2.1	×	×
Refer-KITTI-V2 [69]	9.8K	63.9K	6.6	✓	×
LaMOT [33]	7.5K	28K	3.8	✓	×
Ours	0.2M	2.4M	12.2	✓	✓

Table 1. STORM-Bench dataset statistics and comparison with other referring tracking datasets. MO: Multi-Objects; CE: Complex-expression.

ject given a description is challenging, generating captions for a given object is a relatively easy task. The bottom-up pipeline first localizes objects and then generates corresponding referring expression for each object. Unlike Elysium [55] that simply uses non-semantic confidence scores for box and tracklets filtering, we employ an LLM-based verification process that leverages the reasoning capabilities of large language models to identify and filter incorrect annotations. The STORM-Bench dataset is built upon the Vidor dataset [51, 53], which provides ground-truth bounding boxes for diverse daily objects, beyond the limited pedestrian–vehicle focus of typical MOT datasets [59, 70]. The bottom-up pipeline consists of two main stages: object-level caption generation and verification, and multi-object referring expression synthesis and validation.

Stage1: Object-level caption generation and verification. Stage1 aims to generate detailed captions for each object in a video. To get the object-level captions, we feed the MLLM [2] model with the video frames and the bounding boxes corresponding to the target object. In addition, we adopt visual prompting techniques [60, 62] to enhance model's attention to the specific object, in which we draw a red bounding box around the target object as a visual marker. We find that visual prompts substantially improve caption quality; however, occasional inaccuracies still arise due to inherent hallucinations in MLLMs.

To address this, we introduce a verification process to detect and filter incorrect captions. In particular, we ask another MLLM to predict whether the generated caption uniquely matches the intended object. During verification, we vary the visual inputs as: (1) the original video with red bounding boxes, (2) videos with Gaussian-blurred backgrounds outside the boxes, and (3) cropped object patches. The three types of input are independently evaluated and only the captions verified across all settings are retained. Notably, we found that blurred or cropped inputs are ineffective during caption generation, as removing contextual cues leads to inaccurate descriptions.

Stage2: Multi-object referring expression synthesis and validation. After obtaining verified captions for individual objects, we proceed to generate referring expressions that jointly describe multiple objects in the same video. These expressions can take two main forms: (1) a single phrase referring to a group of objects sharing a common attribute (*e.g.*, "the instruments on the shelf"), or (2) a conjunction of phrases describing distinct objects (*e.g.*, "the girl and the cat running behind her"). To generate such expressions, we employ a textual LLM with reasoning capabilities. We list all of objects in the video with indices and construct the prompt as the object indice and the corresponding caption pairs. Along with the indices of a randomly sampled subset, we ask LLM to identify 1–5 shared and distinctive attributes and to compose concise referring expressions. If no such attributes are found, the model is allowed to return an empty output. For conjunction-type expressions, the model produces individual object phrases and then combines them into a single sentence.

Similar to Stage I, we apply a verification step to ensure the multi-object referring expressions are clear and correct. We take the video, bounding box coordinates, and the generated expression into an MLLM to check for semantic consistency and filter out the mismatched object/expression pairs. In this stage, we omit visual prompts such as bounding box overlays, as they tend to clutter the scene and obscure objects when multiple targets are present.

In summary, our two-stage bottom-up annotation pipeline enables the generation of high-quality referring expressions grounded in multiple object locations, forming a large-scale dataset suitable for training and evaluating the STORM model.

4.2. Dataset Statistics

Leveraging our annotation pipeline, we curated a new dataset STORM-Bench with 15093 training videos and 714 evaluation videos, with 0.2M diverse referring expressions and 73.7K tracked objects. We visualize example videos in Figure 2. It can be observed that the dataset includes diverse patterns of referring expressions, such as referring objects by spatial-temporal relations or certain unique attributes, or referring to a list of different objects.

We also compare our dataset with recent popular object tracking datasets (LaSOT [17], Elysium [55], LaMOT [33], and Refer-KITTI-V2 [69]) in Table 1. LaSOT and Elysium are designed for referring single-object tracking with simple expressions and are therefore not suitable for our task. Refer-KITTI-V2 is one of the first referring multi-object tracking datasets. However, it is limited to two object classes, traffic scenes and simple expressions. LaMOT is another referring multi-object tracking dataset with diverse scenes. Yet, it is relatively small in scale and many videos use simple object categories as the referring expressions. In contrast, STORM-Bench contains 15.7K videos and 251K images, averages 3 instances per expression, and spans 80 classes with an average expression length of 12.2. Together, these statistics highlight the broader domain coverage and richer relational language in STORM-Bench.

5. Experimental Results

5.1. Experimental Settings

Benchmarks. We adopt a progressive evaluation protocol: beginning with image object grounding, followed by referring single-object tracking, and ultimately referring multi-object tracking, to provide a comprehensive assessment of our method. For the *image object grounding* task, we use the widely adopted RefCOCO [27], RefCOCO+ [27], and RefCOCOg [44] benchmarks. RefCOCO and RefCOCO+ are built on MS-COCO and provide region-level referring expressions for objects in static images, RefCOCO emphasizing spatial reasoning and RefCOCO+ focusing on appearance-based descriptions. RefCOCOg contains longer, more descriptive expressions, making it suitable for assessing language understanding in complex scenes. Following [55], we report results on the *val, testA, and testB* splits. We evaluate the referring single-object tracking (RSOT) task on Elysium [55] test dataset. Elysium [55] contains videos sourced from WebVid-10M [4], the dataset provides 500 test videos with concise referring expressions and tracking annotations. Finally, we evaluate the referring multi-object tracking (RMOT) task on the proposed STORM-Bench dataset. Following standard multi-object tracking (MOT) protocols, we report RMOT performances with HOTA [43], identity-based metrics [49], and CLEAR metrics [6] (MOTA and IDsw).

Implementation Details. Our STORM model is built on an 8B base-model [40], with a NaViT [14] vision encoder. We choose NaViT because it naturally supports arbitrary resolutions and aspect ratios without aggressive resizing, and because it is more token-efficient than common multi-crop ViT alternatives for grounding-heavy video inputs. We use publicly released weights from Hugging Face and finetune the model for image grounding, RSOT, and RMOT following our task-composition training.

5.2. Main Results

STORM builds on MLLM model, leveraging its strong text comprehension and spatial understanding capabilities. In addition, our proposed TCL strategy alleviates the need for large-scale RMOT data and further enhances the RSOT and RMOT performance of STORM. In this section, we present a thorough evaluation of STORM by decomposing its performance across individual subtasks. Unless otherwise noted, we report the best results obtained from models trained on image grounding, SOT, RSOT, and RMOT data.

Image Grounding Results. RSOT and RMOT fundamentally depend on strong spatial grounding, so we first examine whether STORM retains this capability on images, despite being primarily designed for video grounding and tracking. Table 2 presents comparisons with existing MLLMs on RefCOCO, RefCOCO+, and RefCOCOg. Across all datasets and evaluation settings, STORM achieves state-of-the-art performance, demonstrating robust generalization across grounding tasks and visual modalities. Notably, our model is trained on no additional grounding data beyond RefCOCO, the same dataset used by all baselines, suggesting that the image grounding performance also benefits from the additional RSOT and RMOT data used in our training framework.

	RefCOCO val/testA/testB	RefCOCO+ val/testA/testB	RefCOCOg val/test
Shik-7B[13]	87.0/90.6/80.2	81.6/87.4/72.1	82.3/82.2
Shik-13B[13]	87.8/91.1/81.8	**82.9**/87.8/74.4	82.6/83.2
M-GPT2[12]	88.7/91.7/**85.3**	80.0/85.1/**74.5**	**84.4**/84.7
Ferret[63]	87.5/91.4/82.5	80.8/87.4/73.1	83.9/84.8
G-GPT[35]	88.0/91.6/82.5	81.6/ 87.2/73.2	81.7/82.0
Ours	**89.1**/**92.7**/83.8	81.6/**88.6**/73.5	84.0/**85.1**

Table 2. Referring image object grounding performance on RefCOCO, RefCOCO+, and RefCOCOg.

Referring Single-Object Tracking. Following prior work [55], we evaluate RSOT under two settings: *referring tracking* setting, where the model must track an object given a natural-language referring expression, and *standard tracking* setting, where the object is specified by its initial-frame bounding box. We observed that referring tracking performance is sensitive to the quality of the referring expressions. In Elysium [55], most expressions are short noun phrases. To further test the model, we use an LLM to rewrite these phrases into longer and more descriptive expressions, applying the same process to both STORM and Elysium to further test the model's the performance on long and complex prompts. The results (Table 3) show the STORM performs similarly comparing to the Elysium with original short prompts. But with our enriched prompts, the STORM performs strongly in both evaluation settings, improving AUC by 0.8% and 2.1%, respectively. This demonstrates that our task composition strategy enables effective knowledge transfer from object grounding to tracking.

Referring Multi-Object Tracking. We finally evaluate STORM on the STORM-Bench dataset, the first benchmark

	Elysium (RSOT)			Elysium (SOT)		
	AUC	P	P_{norm}	AUC	P	P_{norm}
Elys [55]	83.3	89.1	90.0	88.7	94.6	93.8
Ours	**84.1**	**89.7**	**93.2**	**89.8**	**96.4**	**97.8**
Elys*	**87.5**	94.5	93.7	88.7	94.6	93.8
Ours*	87.4	**95.3**	**97.2**	**89.8**	**96.4**	**97.8**

Table 3. Single-object tracking (SOT) and referring single-object tracking (RSOT) results on Elysium. * denotes the experiments with our prompts (longer and more comprehensive).

	HOTA↑	DetA↑	AssA↑	LocA↑	IDF1↑	IDP↑	IDR↑	MOTA↑	IDsw↓
Grounding DINO [31]†	31.7	17.8	56.5	88.1	27.3	68.5	17.0	15.4	2953
Qwen2.5-VL [2] †	37.9	23.2	62.6	76.6	38.6	77.9	25.5	22.4	2767
VisionLLMv2 [24]†	45.3	35.5	58.9	75.3	44.6	53.9	38.1	18.7	12097
LaMOT [33]	46.7	39.0	56.1	89.2	48.8	71.5	37.1	37.6	7972
Ours*	34.8	21.4	56.8	63.7	36.4	54.2	27.3	15.9	2832
Ours	**66.7**	**55.3**	**82.6**	**81.4**	**78.3**	**78.4**	**78.1**	**57.1**	**183**

Table 4. Referring multi-object tracking (RMOT) performance on STORM-Bench. † indicates that OCSORT is used as the tracker. *denotes the model trained on STORM-Bench only.

specifically designed for RMOT, featuring complex referring expressions and challenging multi-object scenarios that stress-test a model's ability to understand user queries. Because no existing model supports end-to-end RMOT, we selected strong baselines and adapted them to operate in the RMOT setting. Image-grounding capable models such as Grounding DINO [42], Qwen2.5-VL [2], and VisionLLMv2 [24] can produce bounding boxes per frame given a prompt, but cannot track objects over time. For these models, we generate frame-level detections and apply the linking algorithm [10] to form trajectories for fair comparison. LaMOT [33] is a specialist video model that combines an image detector with a temporal tracking module. For a fair comparison, we use its official implementation and pretrained weights, and additionally fine-tune it on the STORM-Bench training split.

As shown in Table 4, Grounding DINO performs poorly in the RMOT setting, reflecting its limitations as an image-only model when handling complex queries. Qwen2.5-VL and VisionLLMv2 are largely image-based MLLMs, and therefore cannot effectively leverage temporal information to produce consistent object trajectories across frames. Compared to QwenVL, VisionLLMv2 additionally incorporates an object detector; however, its LLM only provides object embeddings to the detector and does not directly participate in localization. As a result, it can misinterpret expressions and occasionally detect false-positive objects. LaMOT, being a video model, exhibits stronger temporal consistency, but it lacks an LLM component and thus struggles to interpret complex referring expressions.

In contrast, STORM delivers a substantial leap in performance, surpassing all prior methods by +21.2 HOTA and +31.2 IDF1, demonstrating superior spatial-temporal understanding and RMOT capability. Its end-to-end design allows seamless integration of video context with complex natural-language queries. Moreover, comparing the results of *Ours** and *Ours* shows that incorporating TCL improves HOTA by +31.9 and IDF1 by +41.9, validating that TCL effectively reduces the amount of task-specific data needed to learn new tasks. Overall, despite the difficulty of large-scale RMOT annotation, our task composition strategy and curated dataset contribute to STORM 's robustness and strong generalization in this challenging setting.

ROW	Training Data Scale				Test Task Performance		
	IG	SOT	RSOT	RMOT	SOT	RSOT	RMOT
1	1M	0M	0.1M	0K	83.4	77.4	36.9
2	1M	0M	1M	0K	89.8	84.1	43.3
3	1M	0.9M	0.1M	0K	91.3	83.0	41.8
4	1M	0M	1M	15K	91.3	83.0	66.7

Table 5. Ablation on task-composition training using image grounding (IG), single-object tracking (SOT), and referring single-object tracking (RSOT) data."M" and "K" denote million and thousand, respectively.

RMOT Training Data	5K	10K	12K	15K
RMOT Performance (HOTA)	26.4	53.4	63.3	66.7

Table 6. Effect of the scale of RMOT training data in TCL.

5.3. Ablation Study

Task-Composition Learning. We conduct ablations to compare RSOT and RMOT performance with and without our proposed task-composition learning strategy, as shown in Table 5. Row 1 establishes the baseline using image-grounding data combined with a small amount of RSOT data. Rows 2 and 3 demonstrate that our task composition method effectively transfers knowledge from image grounding and SOT to RSOT. Notably, the task-composition–based model achieves performance comparable to RSOT fine-tuning that uses 10× more RSOT data, indicating that our approach is both budget-efficient (no need large scale RSOT annotation) and effective. A similar trend appears in rows 2 and 4: the task composition strategy also benefits RMOT, and the remaining gap can be closed with only a small amount of RMOT data.

RMOT Scale. Finally, we examine how scaling RMOT data impacts performance, noting that RMOT annotations are extremely costly and difficult to obtain. As shown in Table 6, performance begins to plateau at around 12K samples. Motivated by this trend, we cap our STORM-Bench dataset at 15K samples.

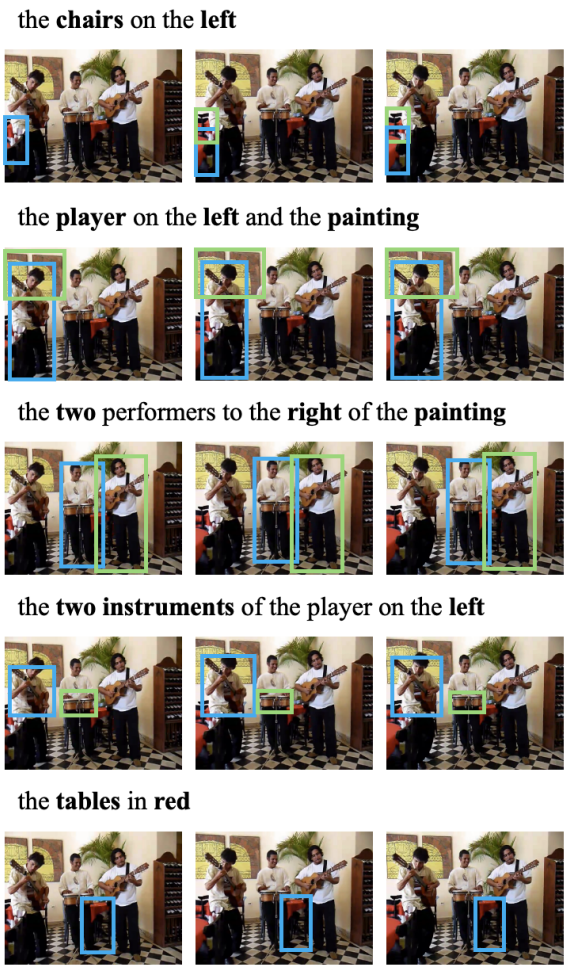

Figure 3. Visualization of complex referring expressions involving attributes and relations (e.g., spatial relations between objects). Our model maintains consistent identities and precise localization across challenging scenarios.

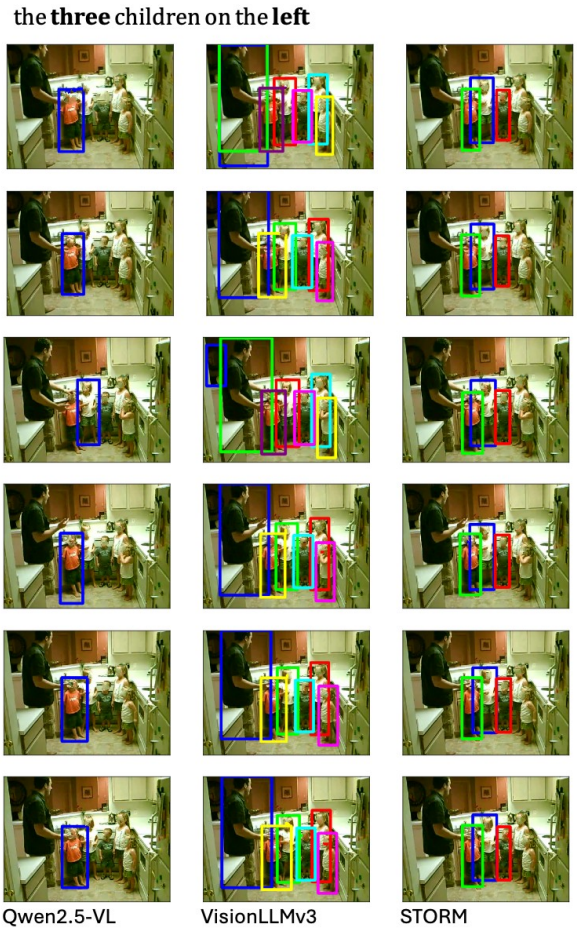

Figure 4. Comparison of STORM, Qwen2.5-VL, and VisionLLMv2 on referring multi-object tracking using examples from STORM-Bench. STORM closely follows the query prompt and accurately localizes and tracks all objects that satisfy the referring expression over time.

5.4. Qualitative Results

First, we demonstrate the generalization ability of STORM by visualizing tracklets produced from different prompts on the same video (Figure 3). In this example, we begin with a simple single-object prompt containing basic spatial cues, and then progressively enrich the prompts with additional spatial and relational conditions to track objects of varying locations and scales. STORM successfully interprets both simple and complex prompts and generates high-quality, consistent tracklets.

We then visualize and compare STORM with Qwen2.5-VL and VisionLLMv2 on their tracking results to better understand the differences between approaches (Figure 4). Qwen2.5-VL [2] cannot fully understand the RMOT prompts, and makes the mistakes in the initial grounding generating the inconsistent bounding boxes, resulting in wrong tacklets. VisionLLMv2 [24] benefits from a specialist detector for spatial localization, yet its outputs often fail to follow the referring expression closely because the detector and LLM operate separately rather than in an end-to-end fashion. In this case, the model failed to identify the "children to the left". In contrast, STORM is a unified end-to-end model that jointly performs detection and temporal association, resulting in more stable and reliable tracking.

6. Conclusion

We present STORM, an end-to-end multimodal large language model for referring multi-object tracking that unifies grounding and tracking without external detectors or trackers. Together with task-composition learning and the new STORM-Bench dataset, STORM achieves state-of-the-art performance across image grounding, single-object tracking, and RMOT benchmarks, demonstrating strong spatial-temporal grounding in challenging real-world videos.

References

[1] Elmira Amirloo, Jean-Philippe Fauconnier, Christoph Roesmann, Christian Kerl, Rinu Boney, Yusu Qian, Zirui Wang, Afshin Dehghan, Yinfei Yang, Zhe Gan, et al. Understanding alignment in multimodal llms: A comprehensive study. *arXiv preprint arXiv:2407.02477*, 2024. 2

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 3, 5, 7, 8

[3] Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Zheng Zhang, and Mike Zheng Shou. One token to seg them all: Language instructed reasoning segmentation in videos. *Advances in Neural Information Processing Systems*, 37:6833–6859, 2024. 3

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1728–1738, 2021. 3, 6

[5] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 941–951, 2019. 1, 3

[6] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008, 2008. 6

[7] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer, 2016. 3

[8] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. Ieee, 2016. 1, 3

[9] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 2

[10] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9686–9696, 2023. 7

[11] Tzoulio Chamiti, Leandro Di Bella, Adrian Munteanu, and Nikos Deligiannis. Refergpt: Towards zero-shot referring multi-object tracking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3849–3858, 2025. 1, 3

[12] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 1, 3, 6

[13] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 1, 3, 6

[14] Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems*, 36:2252–2274, 2023. 6

[15] Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Luc Van Gool, and Marie Francine Moens. Talk2car: Taking control of your self-driving car. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2088–2098, 2019. 3

[16] 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. 3

[17] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5374–5383, 2019. 3, 5

[18] Qi Feng, Vitaly Ablavsky, Qinxun Bai, and Stan Sclaroff. Siamese natural language tracker: Tracking by natural language descriptions with siamese trackers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5851–5860, 2021. 3

[19] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 3

[20] Yongxin Guo, Jingyu Liu, Mingda Li, Dingxin Cheng, Xiaoying Tang, Dianbo Sui, Qingbin Liu, Xi Chen, and Kevin Zhao. Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3302–3310, 2025. 3

[21] Wenyan He, Yajun Jian, Yang Lu, and Hanzi Wang. Visual-linguistic representation learning with deep cross-modality fusion for referring multi-object tracking. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6310–6314. IEEE, 2024. 3

[22] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. 2

[23] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14271–14280, 2024. 3

[24] Wu Jiannan, Zhong Muyan, Xing Sen, Lai Zeqiang, Liu Zhaoyang, Chen Zhe, Wang Wenhai, Zhu Xizhou, Lu Lewei, Lu Tong, Luo Ping, Qiao Yu, and Dai Jifeng. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. *arXiv preprint arXiv:2406.08394*, 2024. 7, 8

[25] A Kamath, M Singh, Y LeCun, I Misra, G Synnaeve, and N MDETR Carion. Modulated detection for end-to-end multi-modal understanding. arxiv 2021. *arXiv preprint arXiv:2104.12763*. 2

[26] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 2

[27] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 3, 6

[28] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34: 11846–11858, 2021. 2

[29] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8971–8980, 2018. 3

[30] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4282–4291, 2019. 3

[31] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975, 2022. 2, 7

[32] Siyuan Li, Tobias Fischer, Lei Ke, Henghui Ding, Martin Danelljan, and Fisher Yu. Ovtrack: Open-vocabulary multiple object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5567–5577, 2023. 3

[33] 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. 3, 4, 5, 7

[34] 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. 3

[35] Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, et al. Groundinggpt: Language enhanced multi-modal grounding model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6657–6678, 2024. 1, 6

[36] Chao Liang, Zhipeng Zhang, Xue Zhou, Bing Li, Shuyuan Zhu, and Weiming Hu. Rethinking the competition between detection and reid in multiobject tracking. *IEEE Transactions on Image Processing*, 31:3182–3196, 2022. 1, 3

[37] Jiacheng Lin, Jiajun Chen, Kunyu Peng, Xuan He, Zhiyong Li, Rainer Stiefelhagen, and Kailun Yang. Echotrack: Auditory referring multi-object tracking for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 2024. 3

[38] Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2794–2804, 2023. 2

[39] Lang Lin, Xueyang Yu, Ziqi Pang, and Yu-Xiong Wang. Glus: Global-local reasoning unified into a single large language model for video segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8658–8667, 2025. 3

[40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2, 6

[41] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2, 3

[42] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 1, 2, 7

[43] J Luiten, A Osep, P Dendorfer, P Torr, A Geiger, L Leal-Taixé, and B Leibe Hota. A higher order metric for evaluating multi-object tracking., 2021, 129. *DOI: https://doi. org/10.1007/s11263-020-01375-2. PMID: https://www. ncbi. nlm. nih. gov/pubmed/33642696*, pages 548–578, 2021. 6

[44] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 3, 6

[45] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 3

[46] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 3

[47] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13009–13018, 2024. 3

[48] Esteban Real, Jonathon Shlens, Stefano Mazzocchi, Xin Pan, and Vincent Vanhoucke. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5296–5305, 2017. 3

[49] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Computer Vision – ECCV 2016 Workshops*, pages 17–35, Cham, 2016. Springer International Publishing. 6

[50] 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. 3

[51] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 279–287. ACM, 2019. 5

[52] Yiming Sun, Fan Yu, Shaoxiang Chen, Yu Zhang, Junwei Huang, Yang Li, Chenhui Li, and Changbo Wang. Chattracker: Enhancing visual tracking performance via chatting with multimodal large language model. *Advances in Neural Information Processing Systems*, 37:39303–39324, 2024. 1, 3, 4

[53] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 5

[54] Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. Object referring in videos with language and human gaze. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4129–4138, 2018. 3

[55] Han Wang, Yongjie Ye, Yanjie Wang, Yuxiang Nie, and Can Huang. Elysium: Exploring object-level perception in videos via mllm. In *European Conference on Computer Vision*, pages 166–185. Springer, 2024. 1, 3, 4, 5, 6

[56] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3

[57] Xiao Wang, Liye Jin, Xufeng Lou, Shiao Wang, Lan Chen, Bo Jiang, and Zhipeng Zhang. Reasoningtrack: Chain-of-thought reasoning for long-term vision-language tracking. *arXiv preprint arXiv:2508.05221*, 2025. 1, 3

[58] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 1, 3

[59] 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. 1, 3, 5

[60] Junda Wu, Zhehao Zhang, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A Rossi, Ruiyi Zhang, et al. Visual prompting in multimodal large language models: A survey. *arXiv preprint arXiv:2409.15310*, 2024. 5

[61] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2411–2418, 2013. 3

[62] Yongliang Wu, Xinting Hu, Yuyang Sun, Yizhou Zhou, Wenbo Zhu, Fengyun Rao, Bernt Schiele, and Xu Yang. Number it: Temporal grounding videos like flipping manga. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13754–13765, 2025. 3, 5

[63] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023. 1, 6

[64] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 3

[65] En Yu, Songtao Liu, Zhuoling Li, Jinrong Yang, Zeming Li, Shoudong Han, and Wenbing Tao. Generalizing multiple object tracking to unseen domains by introducing natural language representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3304–3312, 2023. 3

[66] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. In *European conference on computer vision*, pages 52–70. Springer, 2024. 3

[67] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International journal of computer vision*, 129(11):3069–3087, 2021. 1, 3

[68] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*, pages 1–21. Springer, 2022. 1, 3

[69] Yani Zhang, Dongming Wu, Wencheng Han, and Xingping Dong. Bootstrapping referring multi-object tracking. *arXiv preprint arXiv:2406.05039*, 2024. 5

[70] Yani Zhang, Dongming Wu, Wencheng Han, and Xingping Dong. Bootstrapping referring multi-object tracking. *arXiv preprint arXiv:2406.05039*, 2024. 3, 5

[71] 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. 3