
SVAG-Bench: A Large-Scale Benchmark for Multi-Instance Spatio-temporal Video Action Grounding

Tanveer Hannan^{1,2*} Shuaicong Wu^{1,2*} Mark Weber^{2,3} Suprosanna Shit⁴
Jindong Gu⁵ Rajat Koner⁶ Aljoša Ošep⁷ Laura Leal-Taixé⁷ Thomas Seidl^{1,2}
¹LMU Munich ²MCML ³Technical University of Munich
⁴University of Zurich ⁵University of Oxford ⁶Amazon ⁷NVIDIA
hannan@dbis.lmu.de, shuaicong.wu@campus.lmu.de

Abstract

Understanding fine-grained actions and accurately localizing their corresponding actors in space and time are fundamental capabilities for advancing next-generation AI systems, including embodied agents, autonomous platforms, and human-AI interaction frameworks. Despite recent progress in video understanding, existing methods predominantly address either coarse-grained action recognition or generic object tracking, thereby overlooking the challenge of jointly detecting and tracking multiple objects according to their actions while grounding them temporally. To address this gap, we introduce **Spatio-temporal Video Action Grounding (SVAG)** — a novel task that requires models to simultaneously detect, track, and temporally localize all referent objects in videos based on natural language descriptions of their actions. To support this task, we construct SVAG-Bench, a large-scale benchmark comprising 688 videos, 19,590 annotated records, and 903 unique verbs, covering a diverse range of objects, actions, and real-world scenes. We further propose SVAGFormer, a baseline framework that adapts state-of-the-art vision-language models for joint spatial and temporal grounding, and introduce SVAGEval, a standardized evaluation toolkit for fair and reproducible benchmarking. Empirical results show that existing models perform poorly on SVAG, particularly in dense or complex scenes, underscoring the need for more advanced reasoning over fine-grained object-action interactions in long videos.

1 Introduction

In recent years, the community has witnessed remarkable progress in fine-grained video-language understanding, driven by standardized datasets and benchmarks targeting increasingly complex grounding tasks. These benchmarks—such as those for temporal action localization [11, 16, 27], referring object tracking [10, 19, 32], and spatio-temporal grounding [8, 36, 28]—have enabled the community to measure progress on localized subproblems. However, each task captures only a partial view of the real-world video understanding problem, where both what is happening (action grounding), who is performing (spatial understanding/actor grounding), and when it occurs (temporal grounding) are interdependent.

As illustrated in Fig. 1, current paradigms treat these components in isolation. The Spatial Video Grounding (SVG) task localizes an object primarily in space based on its visual appearance (e.g., “follow the man in a suit”), as shown in Fig. 1a, typically independent of temporal dynamics. Once

*Equal contribution



Figure 1: Comparison of existing video grounding paradigms with our proposed **Spatio-temporal Video Action Grounding (SVAG)** task. (a) **SVG**: Spatial Video Grounding focuses only on spatial localization and lacks temporal reasoning. (b) **STVG**: Spatio-Temporal Video Grounding jointly localizes objects over time but cannot handle multiple interacting instances. (c) **VTG**: Video Temporal Grounding identifies temporal segments but misses spatial localization. (d) **Ours (SVAG)**: Unifies temporal and spatial grounding to detect and track multiple referent objects performing the queried action across time.

the target is identified, the model can track it frame by frame without further linguistic reasoning. Existing SVG datasets such as LaSOT [6] and Refer-Youtube-VOS [25] predominantly focus on category- or appearance-based tracking, while datasets like GroOT [20] only marginally include action-oriented queries.

Conversely, Video Temporal Grounding (VTG) focuses solely on identifying when a described event occurs, disregarding spatial localization. As shown in Fig. 1c, datasets such as QVHighlights [12], Charades-STA [7], and TACoS [24] restrict the problem to one temporal segment per query, assuming a single implicit actor per event.

To bridge these two perspectives, [36, 28] propose Spatio-Temporal Video Grounding (STVG). A novel dataset and tasks aim to localize who, what, when in a video by grounding object queries by their action and temporal boundaries, as shown in Fig. 1b. Despite the challenges in the proposed STVG, it still assumes a *single* referent per query and largely focuses on simple, short videos with limited motion or scene diversity. As a result, they fail to capture multi-actor, multi-action interactions that are ubiquitous in realistic long-form videos.

To address these limitations, we introduce Spatio-temporal Video Action Grounding (SVAG) — a new benchmark designed to unify and extend prior tasks under a multi-instance, action-centric formulation. Unlike existing settings, SVAG requires the model to (1) detect all objects performing the queried action, (2) track their spatial positions across time, and (3) precisely localize the temporal intervals in which these actions occur. As shown in Fig. 1d, given the query “A person is dancing in the open area,” *multiple* individuals may satisfy the description simultaneously, and the model must recover both the spatial and temporal extents of each.

Dataset	Videos	Queries	Tracks	Queries / Video	Tracks / Video	Distinct Verbs
Refer-Youtube-VOS [25]	3,978	14,952	7,451	3.76	1.87	876
GroOT [20]	1,515	3,567	13,294	2.35	8.77	197
VidSTG [36]	6,924	44,808	35,044	6.47	5.06	246
HC-STVG [28]	5,660	5,660	5,660	1.00	1.00	515
SVAG-Bench (Ours)	688	19,590	9,781	28.47	14.22	903

Table 1: **Comparison of video grounding datasets.** Refer-Youtube-VOS and GroOT belong to the SVG domain, while VidSTG and HC-STVG are STVG datasets. Although VidSTG has the largest overall scale, **SVAG-Bench** achieves the highest *annotation density* (queries and tracks per video) and the broadest *action diversity* (distinct verbs), making it particularly suited for fine-grained, multi-object spatio-temporal grounding.

By introducing SVAG, we aim to move beyond isolated spatial or temporal grounding towards a holistic, action-grounded understanding of videos—one that reflects the true compositional and interactive nature of real-world scenes.

To operationalize the SVAG task, we introduce **SVAG-Bench**, a large-scale benchmark explicitly designed for *action-centric spatio-temporal grounding*. Unlike prior datasets that predominantly rely on static, appearance-based queries, SVAG-Bench focuses exclusively on *what objects do* rather than *how they appear*. This design choice fundamentally shifts the reasoning paradigm—from recognizing static entities to understanding *motion patterns*, *temporal evolution*, and *inter-object interactions*. By compelling models to align natural language descriptions with dynamic events, SVAG-Bench promotes a deeper and more compositional understanding of real-world video content.

The dataset was first manually annotated with 9,781 video–query pairs covering 480 distinct action verbs (including tense variations). To enhance linguistic diversity and coverage, we leveraged GPT-3.5 [21] to rephrase and augment the queries, expanding the dataset to 19,590 records encompassing 903 unique verbs. This extensive verb vocabulary promotes robust generalization across varied action semantics, ranging from atomic actions (e.g., *The person walks inside the boat*, *The cat jumps up at the toy*) to complex multi-actor interactions (e.g., *Horse is fighting with another horse*, *The chicken turns around and chases other chickens*).

A comparison of dataset statistics is presented in Table 1. Datasets such as Refer-Youtube-VOS and GroOT belong to the *spatial grounding* domain, while VidSTG and HC-STVG address *spatio-temporal grounding* with a single referent per query. On average, these benchmarks contain only 3.82 queries per video, indicating sparse supervision and limited linguistic diversity. In contrast, SVAG-Bench provides 28.47 queries per video, offering significantly denser annotations across queries, tracks, and actions. This high annotation density enables fine-grained evaluation of temporal overlap, multi-actor disambiguation, and action compositionality—key aspects of robust video understanding. Further details of the annotation pipeline, statistics, and taxonomy are discussed in Section 3.

Summary of Contributions.

1. **New Task** — Spatio-temporal Video Action Grounding (SVAG): We define a new task that unifies object detection, action understanding, and temporal localization. The goal is to detect and track *multiple referents* performing actions specified in natural language.
2. **New Dataset** — SVAG-Bench: We release a large-scale, *action-centric* dataset featuring diverse scenes, object categories, and fine-grained action descriptions with dense annotations (28.47 queries/video).
3. **New Baseline Model** — SVAGFormer: We propose a modular transformer framework that jointly integrates spatial localization and temporal grounding to address the SVAG task.
4. **New Evaluation Protocol** — SVAGEval: We design a formalized evaluation toolkit for benchmarking *multi-referent spatio-temporal grounding*, providing a unified platform for reproducible and fair comparison in future studies.

2 Related Work

2.1 Spatial Video Grounding

Spatial Video Grounding (SVG) aims to localize one or more objects in a video based on natural language descriptions, typically by producing bounding boxes or pixel-level masks. Existing SVG benchmarks [6, 25, 33] primarily rely on static visual descriptions such as the object’s category, color, or position. These attributes allow the target to be uniquely identified from the very first video frame (e.g., “track the man in a suit”), without requiring temporal reasoning. As a result, conventional detectors or trackers often suffice, and *temporal reasoning* is rarely required.

Recent efforts such as GroOT [20] extend SVG to include action-related queries, but their scope remains narrow, with few verbs or simple interactions. Similarly, subsets of TAO [3] introduce multi-object tracking but focus primarily on appearance-based distinctions rather than motion or intent. Models such as Referring Multi-Object Tracking (RMOT) [32, 35] generalize SVG to multiple referents but remain restricted to *short, domain-specific* videos (e.g., cars and pedestrians). In contrast, **SVAG** emphasizes *action-driven semantics*, where multiple visually similar entities must be distinguished by their dynamic behavior across time—requiring models to reason jointly over spatial and temporal cues.

2.2 Video Temporal Grounding

Video Temporal Grounding (VTG) focuses on localizing temporal segments in untrimmed videos that correspond to natural language queries [23, 14, 2, 34]. Benchmarks such as QVHighlights [12], Charades-STA [7], and TACoS [24] frame this as identifying the start and end times of queried events, typically through *Moment Retrieval (MR)* or *Highlight Detection (HD)*. While these datasets have driven advances in temporal reasoning, they generally annotate only a *single event instance* per query, assuming one relevant action per video.

SVAG extends this formulation to the *object level*, requiring models to jointly perform detection, tracking, and temporal localization of all entities satisfying the query. A single query may thus correspond to multiple distinct objects or time intervals, reflecting real-world scenarios where several actors perform the same action at different times. This multi-instance, temporally compositional structure bridges the gap between event-level and instance-level understanding.

2.3 Spatio-Temporal Video Grounding

Spatio-Temporal Video Grounding (STVG) generalizes SVG by jointly predicting both spatial trajectories and temporal segments for the object referred to in the query [36, 28, 30, 18, 13]. Existing datasets typically contain a single referent per description and focus on short, visually simple clips. Queries are often dominated by static appearances (e.g., “the man in a red shirt”) or coarse action labels (e.g., “a person jumping”), limiting their ability to capture fine-grained, multi-actor behaviors. Models such as STVGFormer [15] primarily rely on visual appearance for spatial grounding, with temporal reasoning based on human-action cues. Moreover, benchmarks like VidSTG [36] derive sentences from fixed triplets in VidOR [26], resulting in restricted linguistic and action diversity.

SVAG advances beyond these constraints through densely annotated, action-centric queries that require reasoning over longer temporal horizons and dense visual scenes. It explicitly demands the spatial and temporal grounding of *all objects* that satisfy the query, enabling comprehensive evaluation of multi-instance, multi-action understanding in realistic, unconstrained videos.

3 Dataset Overview

3.1 Data Collection and Annotation

To support the proposed SVAG task, we construct SVAG-Bench, a comprehensive benchmark designed to cover a broad range of scenes, object categories, and action types. Videos are curated from multiple real-world domains, including crowded urban environments, traffic surveillance, wildlife monitoring, and natural ecosystems. Our goal in building SVAG-Bench is two-fold: to ensure *completeness*, by including diverse interaction patterns and environments, and to ensure

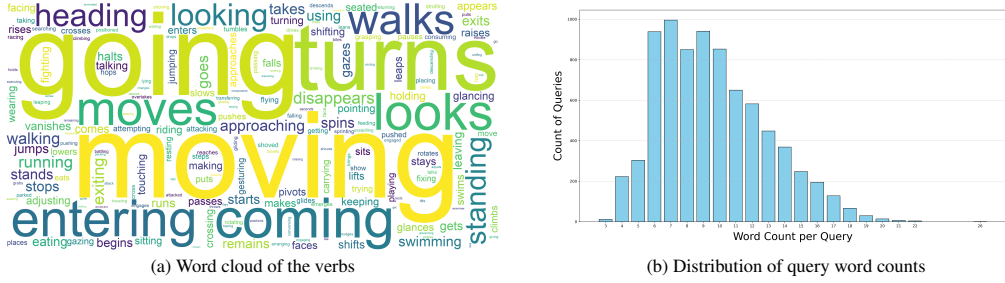


Figure 2: **Statistics of SVAG-Bench.** The majority of queries fall within the range of 6 to 10 words.

discrimination, by featuring multiple visually similar instances of the same category engaged in distinct actions.

The dataset sources videos from established multi-object tracking benchmarks—MOT17 [4], MOT20 [5], and OVIS [22]—selecting sequences where objects of similar appearance perform different actions. Human annotators manually label all visible objects with concise, action-centric natural language descriptions. To enhance linguistic richness and generalization, we further expand these annotations using paraphrases generated by GPT-3.5 [21], followed by human verification to ensure correctness and naturalness.

In total, **SVAG-Bench** comprises approximately **19,590 action-based annotations** across **688 videos**, providing fine-grained ground truth for both spatial and temporal grounding. This combination of dense action coverage and linguistic variation enables robust evaluation of multi-object, multi-action understanding.

3.2 Dataset Statistics

To assess the richness and complexity of SVAG-Bench, we conduct detailed statistical analyses and compare it with representative benchmarks from the *Spatial Video Grounding (SVG)* and *Spatio-Temporal Video Grounding (STVG)* domains. A summary of this comparison is presented in Table 1. Overall, SVAG-Bench exhibits the highest values across key indicators such as *queries per video* (28.47), *tracks per video* (14.22), and *distinct verbs* (903), highlighting its superior annotation density, object diversity, and action coverage.

Unlike prior datasets that emphasize total size (i.e., total number of videos or queries), **per-video annotation density** serves as a more meaningful measure of complexity and reasoning difficulty. High-density videos introduce frequent interactions between multiple entities and actions, requiring models to perform precise spatio-temporal reasoning under dense and overlapping conditions. This property makes SVAG-Bench particularly suitable for evaluating fine-grained video-language understanding beyond appearance-based or single-actor scenarios.

In summary, SVAG-Bench achieves a unique balance between scale and annotation depth: it is compact enough for efficient experimentation yet dense and diverse enough to challenge current vision-language models with realistic, multi-object, action-centric reasoning tasks.

Distinct Verbs (Action). Since our task focuses on object grounding based on actions, the diversity of verb usage in natural language queries is a key metric in evaluating annotation quality. To quantify verb diversity consistently across datasets, we adopt a unified methodology to calculate the number of verbs:

1. We use the **spaCy** library to tokenize and parse all natural language queries, extracting all tokens with a POS tag of VERB.
2. For sentences containing multiple consecutive verbs (e.g., “a person stops walking”), we retain only the main action verb and remove the preceding verbs such as “stops”, “starts”.
3. Verb diversity is computed as the number of unique verb lemmas across all queries.

A word cloud of all annotated verbs is shown in Fig. 2a to visualize the linguistic diversity. The cloud includes verbs in different tenses (e.g., “moving”, “turns”). This level of lexical variability enhances

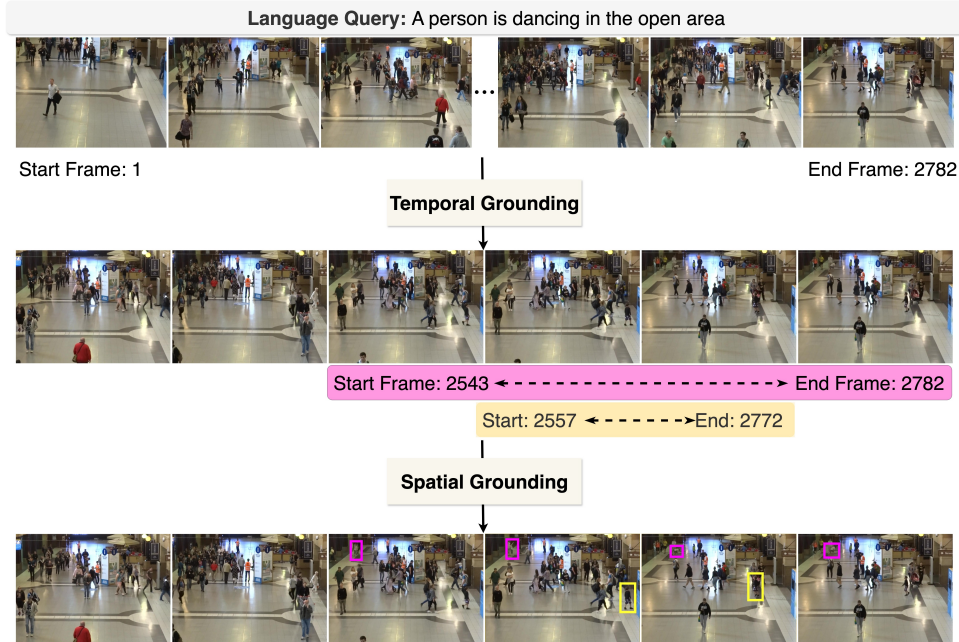


Figure 3: Overview of the **SVAGFormer** pipeline for Spatio-temporal Video Action Grounding (SVAG). Given a natural language query (e.g., “A person is dancing in the open area”), the model first performs temporal grounding to localize the relevant video segment (Start: 2543 → End: 2782), followed by spatial grounding to identify the target person across frames.

the usefulness of this dataset for training and evaluating language-based video understanding models. As can be seen from Table 1, it (903 actions) is more than the other three datasets (GroOT: 197, VidSTG: 246, and HC-STVG: 515), and slightly higher than Ref-Youtube-VOS (876).

Language Query Length. The complexity of natural language queries is another important factor in assessing annotation diversity. We analyze the distribution of query lengths in terms of word count, as depicted in Fig. 2b. The results indicate that the majority of queries fall within the range of 6 to 10 words, striking a balance between conciseness and descriptive richness. The average length of all queries is about 9.58 words. This allows the queries to be both informative and tractable for grounding models.

4 Methodology

We propose SVAGFormer for the SVAG task by decoupling grounding into two modules: Spatial Grounding and Temporal Grounding. The framework is fully modular and built upon off-the-shelf models, leveraging past research to achieve a baseline performance.

The overall pipeline is shown in Fig. 3. Each sub-dataset (OVIS, MOT17, MOT20) is processed separately. For spatial grounding, we employ TempRMOT [35], a state-of-the-art framework for referred multi-object tracking, which builds on TransRMOT [32] and enhances temporal consistency via query memory. This allows robust detection and tracking of arbitrary referents described by action-oriented queries.

For temporal grounding, we adopt FlashVTG [1], a state-of-the-art framework for text-guided video temporal grounding. It incorporates temporal feature layering for multi-scale modeling and adaptive score refinement for improved alignment between queries and video segments.

Since we directly leverage existing models, each module serves as a baseline for its respective sub-task, and the outputs provide a first baseline for SVAG. Details of the architectures and implementations can be found in the original papers of TempRMOT and FlashVTG.

5 Experiments

5.1 Evaluation Metrics

To comprehensively evaluate the performance of our proposed SVAG task, we adopt a set of well-established metrics tailored to the two core subtasks: spatial grounding and temporal grounding. Each subtask requires different aspects of performance to be measured, and thus, the metrics are employed accordingly. We utilize Higher Order Tracking Accuracy (HOTA) [17] to evaluate spatial grounding, i.e., detection in one frame and their temporal association across frames. In referring multi-object tracking (RMOT) [32], predicted tracks corresponding to visible objects not referenced by any query are treated as false positives, ensuring evaluation focuses only on objects relevant to the natural language query. The overall HOTA is computed by averaging per-query HOTA across all sentence queries in the dataset [32]. For the temporal grounding task, we use Recall at 1, 5, and 10 (R1@X, R5@X, R10@X), mean Average Precision (mAP), and mean Intersection over Union (mIoU) as evaluation metrics. We define our evaluation metric based on these popular metrics for spatial [32] and temporal grounding [1, 9, 12] from previous works.

5.2 SVAGEval

We introduce SVAGEval to formalize the evaluation for the SVAG task. This official evaluation codebase will also serve as the benchmark for an ICCV 2025 workshop competition. Unlike existing temporal grounding protocols, SVAGEval supports multiple referents under spatiotemporal constraints. Specifically, spatial and temporal grounding are evaluated separately, with identity mapping strategies ensuring consistent alignment across the two dimensions. The final leaderboard score is computed as the arithmetic mean over OVIS, MOT17, and MOT20. Below, we describe core implementation details of our evaluation pipeline.

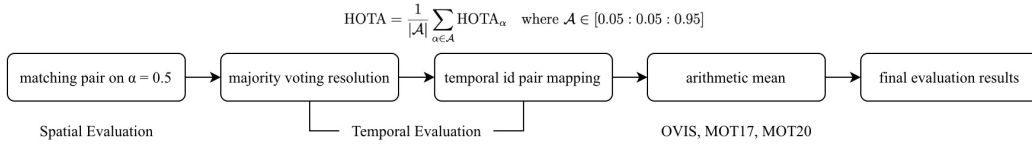


Figure 4: Flowchart for processing evaluation. Spatial and temporal evaluations are conducted separately on the OVIS, MOT17, and MOT20. The results are averaged and combined to form the final result. Threshold α controls the relative importance of detection and association accuracy in HOTA.

Localisation Thresholds. The HOTA [17] score is calculated by averaging over a range of threshold values α that define the matching criteria between predicted and ground truth instances. Its core calculation formula is as follows:

$$\text{HOTA} = \frac{1}{|\mathcal{A}|} \sum_{\alpha \in \mathcal{A}} \text{HOTA}_{\alpha} \quad (1)$$

where $\mathcal{A}$ denotes a set of thresholds ranging from 0.05 to 0.95 in increments of 0.05. In our implementation, we further select the matching result at $\alpha = 0.5$ as the basis for ID mapping, as it provides a balanced trade-off between strictness and flexibility in evaluating spatial matches.

ID Mapping for Temporal Evaluation. Since spatial and temporal grounding are evaluated separately, it is necessary to establish a reliable mapping between predicted and ground truth identities (track_ids) before temporal evaluation. Our strategy proceeds as follows:

1. **Spatial ID Matching:** Using the HOTA matching result at $\alpha = 0.5$, we determine a one-to-one mapping between each ground truth track_id and its most likely predicted counterpart across frames.
2. **Majority Voting Resolution:** Due to inaccurate predictions in spatial grounding, a single ground truth track_id might be associated with multiple predicted track_ids over time. To address this issue, we perform a majority voting scheme. For each ground truth ID, the

Dataset	HOTA	DetA	AssA	DetRe	DetPr	AssRe	AssPr	LocA
OVIS	22.734	11.234	46.625	20.02	19.91	56.555	67.611	82.768
MOT17	0.59603	0.042611	9.1714	0.048201	0.36561	11.659	56.583	75.968
MOT20	0.42845	0.065153	2.9708	0.095114	0.20551	5.5381	13.112	64.431

Table 2: Performance on the different datasets using TempRMOT [35]. The model achieved the highest scores on the OVIS dataset, whereas its performance on MOT was comparatively poor. Association accuracy (AssA) contributed more significantly than detection accuracy (DetA). The MOT datasets have an excessive number of objects and longer videos, which results in worse performance compared to OVIS

number of frames in which each predicted ID appears is taken as the frequency, and the prediction ID with the highest frequency is selected to match the ground truth ID.

3. Temporal Pair Construction: Using this final track_id mapping, we find temporal prediction and ground truth pairs for each referent. These pairs are then passed into the temporal grounding evaluation module.

Since the evaluation metrics for OVIS, MOT17, and MOT20 are calculated independently, we adopt an arithmetic mean of the scores from these datasets to produce the final metric displayed on the competition website. See Fig. 4 for the process.

This design ensures consistent identity alignment across spatial and temporal dimensions, providing a fair basis for evaluating SVAG models on complex multi-object and multi-referent scenarios.

5.3 Implementation Details

The training protocols and model configurations for spatial and temporal grounding are as follows:

Spatial Grounding Settings. We follow the official setup of TempRMOT [35] on Refer-KITTI-V2, setting memory length to 5, using an Adam optimizer with initial learning rate $1e-5$, and a decay by factor 10 after the 40th epoch. We train for 60 epochs on 4 GPUs.

Temporal Grounding Settings. We follow FlashVTG [1], extracting video features via Intern-Video2 [31] and text features via LLaMA [29]. For that, we convert our data into QVHighlights format. All feature dimensions are set to 256 and the fusion module uses 8 attention heads, with $K = 4$. The temporal feature layering has 5 layers. We apply AdamW as optimizer, with the NMS threshold set to 0.7. The maximum visual length parameter is adjusted according to each dataset. We train our model separately on OVIS, MOT17, and MOT20, rather than jointly across all datasets.

5.4 Quantitative Results

We evaluate the performance on three datasets (OVIS [22], MOT17 [4], and MOT20 [5]) separately using two different benchmark frameworks named TempRMOT [35] and FlashVTG [1] based on the experimental setup mentioned above. The results are reported in Tables 2 and 3, respectively. Models perform consistently better on OVIS than on MOT17 and MOT20 across all metrics, suggesting that these models have better generalization capabilities under complex occlusion conditions than scenes with dense objects and very long videos.

Performance on TempRMOT. Table 2 reports the spatial grounding and tracking results on three datasets. The model performs best on dataset OVIS, with a significantly higher HOTA score, indicating better overall tracking performance under occlusion. The AssA (overall association accuracy) scores are relatively high across all datasets, which suggests that the association component of the tracker can continuously associate with one object across frames after it has been detected. In contrast, the DetA scores of MOT17 and MOT20 are extremely low. This may be due to a combination of lots of unlabeled objects that meet the description being treated as false negatives and labeled objects that may not be detected correctly, suggesting that object detection remains the main bottleneck in dense and long-duration videos. These results highlight the importance of improving detection robustness and the annotation density to advance referring multi-object tracking in complex scenarios.

Performance on FlashVTG. Table 3 presents evaluation results under two conditions: with and without non-maximum suppression (NMS, threshold 0.7). The model consistently achieves the best results on OVIS compared to MOT17 and MOT20 across all metrics, confirming the model’s

Dataset	R1			R5			R10			mAP		mIoU	
	@0.1	@0.3	@0.5	@0.1	@0.3	@0.5	@0.1	@0.3	@0.5	@0.1	@0.3		
OVIS	82.37	52.66	33.37	90.91	72.17	56.87	94.68	83.59	71.51	85.25	59.98	42.12	39.27
OVIS [†]	82.37	52.66	33.37	91.57	75.06	59.31	94.9	85.03	75.06	85.63	61.22	43.98	39.27
MOT17	30.86	13.23	6.41	69.74	38.08	14.03	79.76	48.5	21.04	42.45	21.78	8.69	10.48
MOT17 [†]	30.86	13.23	6.41	74.15	39.48	15.63	83.57	48.9	21.24	43.67	21.79	8.97	10.48
MOT20	20.14	9.95	4.17	48.15	22.45	12.73	61.81	31.02	20.83	20.56	10.07	5.47	7.62
MOT20 [†]	20.14	9.95	4.17	48.15	22.45	12.73	61.81	31.02	20.83	20.56	10.07	5.47	7.62

Table 3: Performance on the different datasets using FlashVTG [1]. Unlabeled datasets do not use NMS. Datasets marked with [†] use NMS 0.7. The higher score is highlighted in bold. Applying NMS will slightly improve R@5/10 and mAP for OVIS and MOT17, with no impact on MOT20.

robustness in short videos, despite being occluded. Temporal localization is strongly correlated with video length. MOT videos average hundreds to thousands of frames, and performance degrades significantly. Applying NMS slightly improves R@5/10 and mAP for OVIS and MOT17 by removing redundant overlapping predictions, but it has no impact on R@1, mIoU, or MOT20. This is because R@1 and mIoU are determined solely by the best prediction, while the relatively low baseline detection quality in MOT20 further limits potential gains. These findings highlight that detection quality and redundancy management are critical for temporal action grounding performance.

5.5 Qualitative Results

We provide qualitative visualizations to illustrate the model’s ability in fine-grained spatiotemporal grounding. As shown in Fig. 5, the model successfully localizes subtle actions such as a zebra *tilts its head*, demonstrating spatial sensitivity and temporal precision.

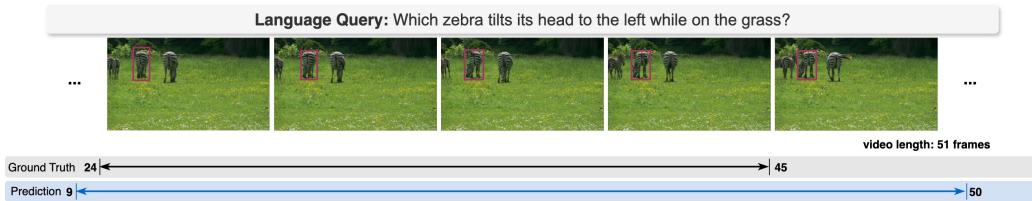


Figure 5: A qualitative visualization example of a zebra performing a fine-grained action: tilting its head to the left across the grass. TempRMOT can localize the object, even with subtle action.

5.6 Results Analysis

To better understand the performance of TempRMOT, we analyze results on the OVIS dataset. We focus on sequences with high performance and identify the top 10 referent categories and actions, as shown in Fig. 6.

The high-performing cases are dominated by animal categories (e.g., dog, rabbit). Conversely, some frequent categories in the dataset (e.g., fish, poultry) are underrepresented, suggesting that action observability and motion diversity are primary drivers of effective grounding.

The distribution of verbs further supports this finding: dynamic actions (e.g., “move”, “fight”, “eat”) dominate successful sequences, though static states (e.g., “remain”) also appear. This indicates that both explicit motion and temporal continuity are leveraged by the model, aligning with the task objective of spatiotemporal grounding based on action queries.

6 Competition

We organize an ICCV 2025 Workshop² dedicated to the SVAG task. At the 8th edition of the BMTT workshop, the focus is on action-aware multi-object tracking, aiming to bridge the gap between vision and language by introducing unified challenges that evaluate both temporal localization and object tracking. To this end, we host a challenging competition where participants are required to

²<https://motchallenge.net/workshops/bmtt2025/>

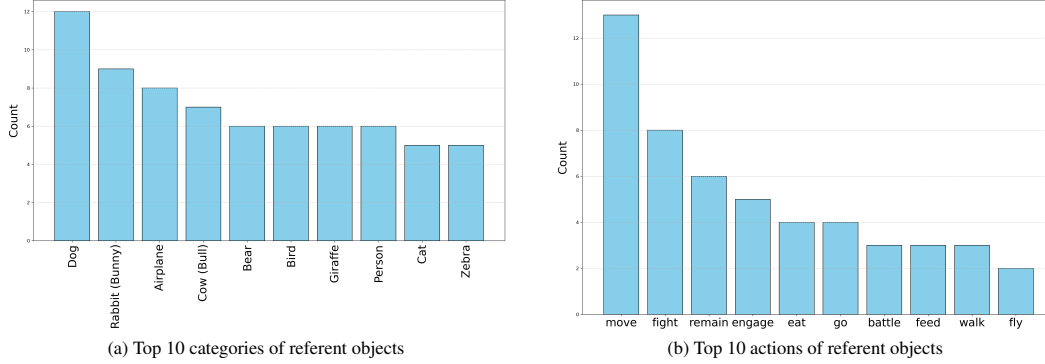


Figure 6: **Statistics on OVIS dataset.** High performance on categories like Dog, Rabbit, and Airplane. Dynamic actions like move, fight lead successful sequences.

develop models to tackle the SVAG task. The competition is hosted on the Codabench platform³, with SVAGEval as the official evaluation benchmark. Several teams have submitted their results, and a summary leaderboard is presented in Tab. 4. We conducted several ablation studies and applied the best-performing spatial and temporal grounding results on the SVAGEval, and the final results are as SVAGFormer recorded in the table. We adopt **m-HIoU** as the primary ranking metric, defined as the arithmetic mean of HOTA and mIoU, which jointly capture temporal and spatial localization quality.

Participant	m-HIoU	HOTA	mIoU	DetA	AssA	R1@0.3
xxcbole	25.417	7.957	42.877	3.167	21.884	47.223
y_squared	20.680	10.734	30.627	4.070	41.693	33.653
gloria	16.114	9.001	23.227	4.048	24.673	29.867
SVAGFormer	14.148	9.159	19.137	4.092	27.698	24.567

Table 4: Competition leaderboard results on Codabench. Teams are ranked in descending order of **m-HIoU**. The highest score is highlighted in bold. The first team improved the overall performance by improving the temporal grounding performance. The second team improved the overall performance by improving the association.

Two teams (Team 1 and Team 3) added additional strategies and techniques to the baseline model to improve performance, and one team (Team 2) used additional models for tracking and retrieval to improve performance. These submissions reflect the growing interest in language-guided video understanding and reveal promising directions for improving spatio-temporal grounding, multimodal alignment, and long-horizon reasoning. We expect the SVAG challenge to serve as a catalyst for future research on scalable, action-aware, and temporally grounded vision-language models.

7 Conclusion

In this work, we introduced Spatio-temporal Video Action Grounding (SVAG), a novel task that unifies object detection, tracking, and temporal localization conditioned on action-specific language queries. To support this task, we proposed the SVAG-Bench dataset, established SVAGFormer as a baseline framework, and released SVAGEval for standardized evaluation. Our analysis highlights that model performance is primarily constrained by spatial grounding quality, thereby affecting the final spatial grounding quality, particularly in dense and long-duration videos. Temporal associations benefit from post-processing strategies. We also show that existing models perform poorly on our dataset. Furthermore, existing models do not yet support multi-instance-level temporal video grounding. Thus, our contributions provide a foundation for advancing video-language research at the intersection of vision, language, and temporal reasoning.

³<https://www.codabench.org/competitions/9743/>

References

- [1] Zhuo Cao, Bingqing Zhang, Heming Du, Xin Yu, Xue Li, and Sen Wang. Flashvtg: Feature layering and adaptive score handling network for video temporal grounding. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 9208–9218, February 2025. 6, 7, 8, 9
- [2] Yi-Wen Chen, Yi-Hsuan Tsai, and Ming-Hsuan Yang. End-to-end multi-modal video temporal grounding. *Advances in Neural Information Processing Systems*, 34:28442–28453, 2021. 4
- [3] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. Tao: A large-scale benchmark for tracking any object. In *European conference on computer vision*, pages 436–454. Springer, 2020. 4
- [4] Patrick Dendorfer, Aljosa Osep, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, Stefan Roth, and Laura Leal-Taixé. Motchallenge: A benchmark for single-camera multiple target tracking. *International Journal of Computer Vision*, 129:845–881, 2021. 5, 8
- [5] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020. 5, 8
- [6] 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. 2, 4
- [7] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017. 2, 4
- [8] Xin Gu, Heng Fan, Yan Huang, Tiejian Luo, and Libo Zhang. Context-guided spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18330–18339, 2024. 1
- [9] Yiyang Jiang, Wengyu Zhang, Xulu Zhang, Xiao-Yong Wei, Chang Wen Chen, and Qing Li. Prior knowledge integration via llm encoding and pseudo event regulation for video moment retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7249–7258, 2024. 7
- [10] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Computer Vision—ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV 14*, pages 123–141. Springer, 2019. 1
- [11] Sangyoun Lee, Juho Jung, Changdae Oh, and Sunghee Yun. Enhancing temporal action localization: Advanced s6 modeling with recurrent mechanism. *arXiv preprint arXiv:2407.13078*, 2024. 1
- [12] 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, 4, 7
- [13] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. *arXiv preprint arXiv:1904.11574*, 2019. 4
- [14] 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. 4
- [15] Zihang Lin, Chaolei Tan, Jian-Fang Hu, Zhi Jin, Tiancai Ye, and Wei-Shi Zheng. Stvgformer: Spatio-temporal video grounding with static-dynamic cross-modal understanding. In *Proceedings of the 4th on Person in Context Workshop*, pages 1–5, 2022. 4

- [16] Shuming Liu, Chen-Lin Zhang, Chen Zhao, and Bernard Ghanem. End-to-end temporal action detection with 1b parameters across 1000 frames. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18591–18601, 2024. 1
- [17] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129:548–578, 2021. 7
- [18] Shu Luo, Jingyu Pan, Da Cao, Jiawei Wang, Yuquan Le, and Meng Liu. Spatial–temporal video grounding with cross-modal understanding and enhancement. *Expert Systems with Applications*, 271:126650, 2025. 4
- [19] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 792–807. Springer, 2016. 1
- [20] Pha Nguyen, Kha Gia Quach, Kris Kitani, and Khoa Luu. Type-to-track: Retrieve any object via prompt-based tracking. *Advances in Neural Information Processing Systems*, 36:3205–3219, 2023. 2, 3, 4
- [21] OpenAI. <https://chatgpt.com/>, 2023. 3, 5
- [22] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *International Journal of Computer Vision*, 130(8):2022–2039, 2022. 5, 8
- [23] Mengxue Qu, Xiaodong Chen, Wu Liu, Alicia Li, and Yao Zhao. Chatvtg: Video temporal grounding via chat with video dialogue large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1847–1856, 2024. 4
- [24] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013. 2, 4
- [25] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 208–223. Springer, 2020. 2, 3, 4
- [26] 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, 2019. 4
- [27] Dingfeng Shi, Qiong Cao, Yujie Zhong, Shan An, Jian Cheng, Haogang Zhu, and Dacheng Tao. Temporal action localization with enhanced instant discriminability. *arXiv preprint arXiv:2309.05590*, 2023. 1
- [28] Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8238–8249, 2021. 1, 2, 3, 4
- [29] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 8
- [30] Jiankang Wang, Zhihan Zhang, Zhihang Liu, Yang Li, Jiannan Ge, Hongtao Xie, and Yongdong Zhang. Spacevllm: Endowing multimodal large language model with spatio-temporal video grounding capability. *arXiv preprint arXiv:2503.13983*, 2025. 4
- [31] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yanan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024. 8

- [32] 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, 4, 6, 7
- [33] Junbin Xiao, Xindi Shang, Xun Yang, Sheng Tang, and Tat-Seng Chua. Visual relation grounding in videos. In *European conference on computer vision*, pages 447–464. Springer, 2020. 4
- [34] Zhe Xu, Kun Wei, Xu Yang, and Cheng Deng. Point-supervised video temporal grounding. *IEEE Transactions on Multimedia*, 25:6121–6131, 2022. 4
- [35] Yani Zhang, Dongming Wu, Wencheng Han, and Xingping Dong. Bootstrapping referring multi-object tracking. *arXiv preprint arXiv:2406.05039*, 2024. 4, 6, 8
- [36] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10668–10677, 2020. 1, 2, 3, 4