This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

OmniGround: A Comprehensive Spatio-Temporal Grounding Benchmark for Real-World Complex Scenarios

Hong Gao[1,2] Jingyu Wu[2] Xiangkai Xu[2] Kangni Xie[2] Yunchen Zhang[2]
Bin Zhong[2] Xurui Gao[2] Min-Ling Zhang[1*]
[1]Southeast University, China [2]ZTE Corporation, China
{230239517, zhangml}@seu.edu.cn
{wu.jingyu2, xu.xiangkai1, xie.kangni, zhang.yunchen, zhong.bin, gao.xurui}@zte.com.cn

Abstract

Spatio-Temporal Video Grounding (STVG) aims to localize target objects in videos based on natural language descriptions. Despite recent advances in Multimodal Large Language Models, a significant gap remains between current models and real-world demands involving diverse objects and complex queries. We attribute this to limited benchmark scope, causing models to exhibit category bias, oversimplified reasoning, and poor linguistic robustness. To address these limitations, we introduce OmniGround, a comprehensive benchmark with 3,475 videos spanning 81 categories and complex real-world queries. We propose the Forward-Backward-Refinement annotation pipeline that combines multi-directional tracking with intelligent error correction for high-quality labels. We further introduce DeepSTG, a systematic evaluation framework quantifying dataset quality across four complementary dimensions beyond superficial statistics. Evaluations reveal performance average drop of 10.4% on complex real-world scenes, particularly with small/occluded objects and intricate spatial relations. Motivated by these, we propose PG-TAF, a training-free two-stage framework decomposing STVG into high-level temporal grounding and fine-grained spatio-temporal propagation. Experiments demonstrate PG-TAF achieves 25.6% and 35.6% improvements in m_tIoU and m_vIoU on OmniGround with consistent gains across four benchmarks.

1. Introduction

Spatio-Temporal Video Grounding (STVG) aims to localize target objects in both temporal and spatial domains within a video based on natural language descriptions [45]. It finds broad applications in autonomous driving [2, 33], video retrieval [11, 42], intelligent surveillance [45], and human-computer interaction [46]. With the rapid development of Multimodal Large Language Models (MLLMs) [7, 9, 36, 41], the task has greatly benefited from MLLMs' enhanced cross-modal semantic understanding and structured spatial reasoning, substantially outperforming traditional lightweight approaches [5, 14, 18, 37].

Despite these advances, there remains a significant gap between current STVG-MLLMs and real-world demands [2, 17, 35]. As illustrated in Fig. 1, existing models often struggle with complex real-world scenarios: failing to localize some uncommon objects (e.g., "kite", "scissors"); unable to distinguish multiple similar targets ("the right man in red clothes among three people"); and misinterpreting queries with nested spatial relations.

We attribute these limitations to the narrow scope of current benchmarks. Datasets like HC-STVG [30] focus predominantly on a single category (person), while others like VidSTG [26, 45] involve relatively simple scenarios with limited object diversity, which hinders the practical deployment of STVG-MLLMs in complex real-world scenarios.

To address these limitations, we introduce **OmniGround**, a comprehensive benchmark designed to advance STVG-MLLMs toward real-world applicability. OmniGround comprises 3,475 videos (average 18.2s) spanning 81 balanced object categories — twice as many as VidSTG [45] and six times more than ST-Align [17]. Each video is densely annotated with spatio-temporal tubes at native FPS, paired with linguistically complex descriptions (sentence-level NEI: 0.918 vs. 0.659 in VidSTG) that include rich spatial relations and diverse action predicates.

To ensure annotation quality and efficiency, we propose the **Forward-Backward-Refinement (FBR)** pipeline, which uses multi-point bi-directional tracking with fusion-based voting to generate accurate spatio-temporal tubes. Our FBR pipeline reduces annotation time consumption while significantly enhancing continuity and accuracy, and is particularly robust to partial occlusions (Fig. 4).

*Corresponding Author

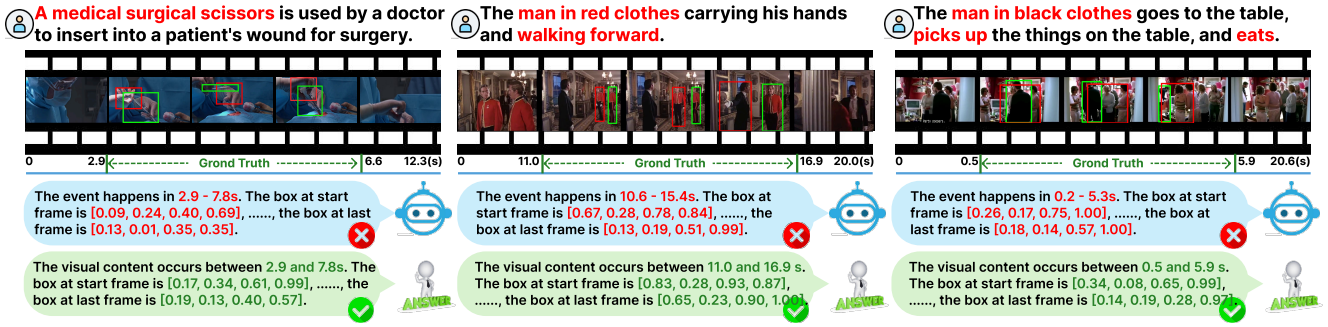


Figure 1. Some examples of STVG MLLMs facing real-world complex scenarios (left: uncommon objects, middle: multiple similar target objects, right: queries with deep syntactic complexity). Content in green and red represents correct and wrong answers, respectively.

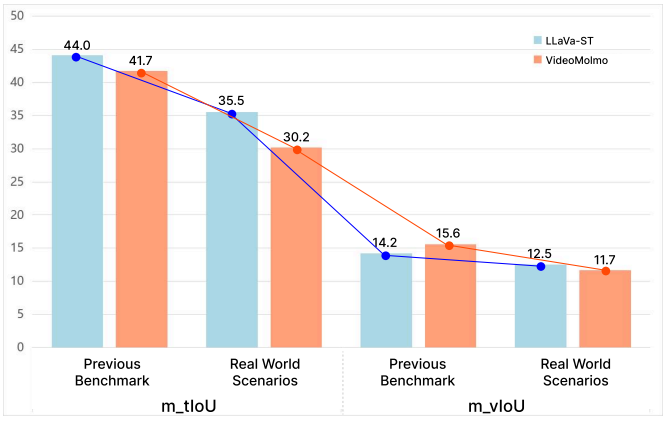


Figure 2. Performance of STVG MLLMs on complex real-world inputs. Inconsistent trends indicate current datasets are insufficient for comprehensive real-world evaluation.

Furthermore, to systematically evaluate dataset quality beyond superficial statistics (e.g., video duration and clip counts), we introduce **DeepSTG**, an evaluation framework that quantifies high-dimensional metrics such as category diversity, foreground complexity, linguistic balance, and cross-modal alignment. A comprehensive analysis with DeepSTG demonstrates that OmniGround significantly outperforms existing benchmarks across all metrics (Tab. 1), thereby establishing a new standard for STVG research.

Based on OmniGround, we conduct extensive evaluations on multiple state-of-the-art models to discover their current performance limitations. Models exhibit performance average drop of 10.4% in m_vIoU compared to simpler benchmarks (Fig. 2), particularly in: *(i)* categories bias (8.42% drop), *(ii)* oversimplified reasoning (13.1% drop), and *(iii)* poor robustness to linguistic complexity (9.42% drop). The results reveal that existing methods struggle with samples exhibiting high foreground complexity and deep semantic structure, particularly when the object is small or heavily occluded, or when spatial relations are intricate.

Motivated by these observations, we propose **PG-TAF**, a training-free two-stage framework that decouples STVG into coarse temporal grounding (via MLLM's holistic reasoning) and fine-grained spatial propagation (via specialized tracking), effectively addressing the reasoning-localization trade-off. Experiments across four benchmarks demonstrate consistent improvements, with 25.6% and 35.6% gains in m_tIoU and m_vIoU on OmniGround. We hope our work can inspire ideas for improving STVG performance under complex real-world scenarios.

In summary, the contributions of our work are as follows:

- We construct **OmniGround**, a STVG benchmark comprising 3,475 videos with 81 balanced categories and complex real-world queries. We further introduce the **FBR pipeline** for efficient annotation.
- We propose **DeepSTG**, a systematic framework for quantifying STVG dataset quality, revealing limitations in category diversity, complex reasoning, and language robustness. Evaluations on OmniGround further show that current models struggle in these complex scenarios due to these constraints in existing datasets.
- We propose **PG-TAF**, a training-free two-stage framework baseline designed to address the challenges exposed by OmniGround. Experiments on four benchmarks compared with several models indicate improvements in our method and a promising direction for future research.

2. Related Work

2.1. Spatio-Temporal Video Grounding Datasets

STVG research relies on comprehensive datasets to improve model performance. HC-STVG [30] is a pioneering large-scale benchmark that localizes spatio-temporal tubes of target persons in untrimmed videos based on textual descriptions. VidSTG [45] provides an untrimmed video benchmark emphasizing multi-form sentence grounding. Li et al.[17] propose ST-Align, using GPT-4-turbo[1] to enhance VidSTG's textual annotations. Ahmad et al.[2] introduce VPos-Bench with 72k video-caption pairs and 100k annotated object points across five real-world scenarios.

EgoMask[19] explores fine-grained object tracking in egocentric videos based on Ego4D [10].

Despite these efforts, existing benchmarks lack comprehensive diagnostic evaluation beyond basic statistics. Current assessments of STVG benchmarks primarily rely on basic statistical metrics, such as object type, domain, and number [28]. While HC-STVG [30] introduces concepts like Ambiguity (BA, TA) and High-level Compositionality (HC), these metrics still can not fully evaluate the benchmark's complexity and generalization demands [38]. Specifically, the ambiguity metrics use binary flags or broad categories that fail to capture the true level of difficulty in the data. HC is often based on simple feature counts instead of balanced syntactic challenges, leading to uneven data distributions that hurt model performance.

Therefore, we introduce **OmniGround**, a comprehensive STVG benchmark designed to address the lack of diversity and complexity in existing datasets. Furthermore, to assess dataset quality beyond superficial statistics, we propose **DeepSTG**, an evaluation framework that enabling objective comparison and rigorous analysis of benchmarks.

2.2. Spatio-Temporal Video Grounding Via MLLM

With the significant progress of MLLMs in video understanding, more research focuses on using MLLMs to effectively address the STVG task [2, 17, 18, 35]. Existing methods can be categorized into two main perspectives.

The first perspective decomposes STVG into multi-stage processes. Methods like VTimeLLM [14] and TRACE [13] focus on temporal localization through boundary-sensitive training or causal event modeling. VideoMolmo [2] focuses on accurately grounding the query in time. Groma [23] designs a localized visual tokenization mechanism for fine-grained region captioning, and GroundingGPT [18] uses MLLM to pinpoint the target object's location spatially.

The second perspective is to let MLLMs simultaneously capture both temporal and spatial information. Recently, LLaVA-ST [17] proposes language-aligned positional embedding and a spatial-temporal packer to improve spatio-temporal multimodal understanding. SpaceVLLM [35] adopts spatio-temporal aware queries and a query-guided space decoder to enhance the MLLMs' performance.

Despite recent advances, existing methods struggle with multi-actor videos, complex foregrounds, and linguistically difficult queries [2, 35]. Because current STVG datasets focus on simple, clean video segments without real-world complexity [17, 30, 45]. Therefore, we propose OmniGround to comprehensively evaluate the models' performance under complex conditions.

3. OmniGround Benchmark

In this section, we introduce the construction pipeline of our OmniGround, which consists of four key stages:1) data acquisition, 2) data manipulation and filtering, 3) high-quality annotation via the Forward-Backward-Refinement pipeline, and 4) dataset augmentation and validation to enhance coverage. Fig. 3 shows the overall statistics of OmniGround.

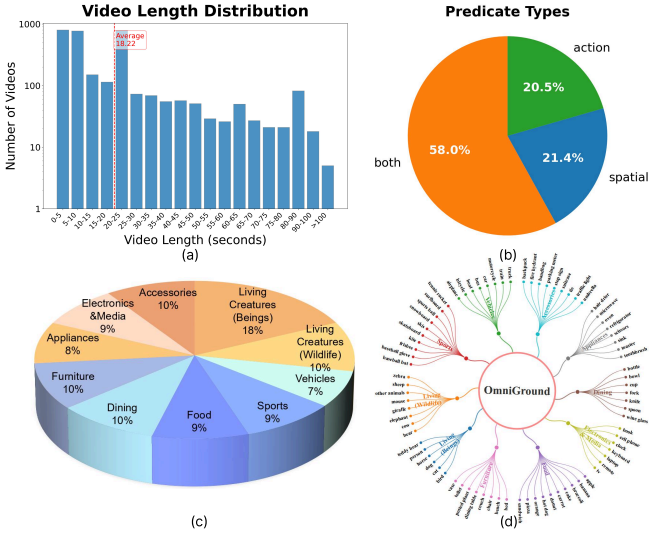


Figure 3. Statistics of our OmniGround benchmark: (a) Video durations range from short clips to 140 seconds. (b) Predicates combine spatial and action elements, reflecting complex object interactions and motion. (c) Balanced category distribution. (d) 81 object categories.

3.1. Data Acquisition

We use a two-source strategy to ensure diversity, real-world relevance, copyright compliance, and annotation feasibility:

Internet Videos. We collect high-quality, copyright-free videos from video-sharing platforms like Pexels, covering diverse real-world scenarios. To ensure category balance, we select at least 5 videos each from 81 categories. We prioritize videos that exhibit: *(i)* natural lighting and camera motion, *(ii)* multiple interacting objects, and *(iii)* complex backgrounds typical of uncontrolled environments.

RVOS Datasets Augmentation. To improve coverage of rare object classes (e.g., "keyboard"), we slightly supplement with Referring Video Object Segmentation (RVOS) videos [8, 25], which are processed (Sec. 3.4) to fit STVG tasks.

3.2. Data Manipulation

Following data acquisition, we establish two temporal constraints to ensure data quality and task validity: (1) Video Duration $\geq$ 3s: Videos shorter than 3s lack sufficient temporal context for STVG queries, making temporal grounding trivial. (2) Event Duration $\geq$ 1s: Target events lasting less than 1s are too fleeting for reliable annotation and evaluation, as they approach the granularity of single frames.

After filtering, OmniGround comprises 3,475 videos with an average duration of 18.2 seconds (ranging from 3s to 140s). Each video contains an independent annotated spatio-temporal event, with caption averaging 16.2 words. The benchmark exhibits high diversity across object categories (81 balanced classes) and rich predicate structures.

3.3. Forward-Backward-Refinement Pipeline

Accurate spatio-temporal tube annotation faces three challenges: high labor cost from frame-by-frame manual labeling; tracking drift due to occlusions and motion blur; and annotation discontinuity from accumulated errors. To address these challenges, we propose the Forward-Backward-Refinement (FBR) pipeline (Fig. 4 (a)), which combines multi-directional tracking with intelligent error correction:

- **Step 1: Multi-Point Initialization.** Instead of relying on a single starting point, annotators first manually label bounding boxes at two key frames: F_{start} and F_{end}. We then initialize bi-directional tracking from F_{start} and F_{end} using DAM4SAM tracker [32], generating forward track T_f and backward track T_b.
- **Step 2: Adaptive Refinement.** The pipeline automatically identifies the frame F_{mid} with maximum IoU discrepancy between T_f and T_b, indicating potential tracking failure. The annotator then provides a corrective bbox at F_{mid}, from which we generate middle track T_m.
- **Step 3: Fusion and Smoothing.** The three tracks are fused via IoU-based voting: bboxes within each frame are clustered (IoU $\geq$ 0.5), and the main cluster's average serves as the fused result. For short gaps where the number of continuously lost frames is ≤ 3, the pipeline recovers them by interpolating between continuous segments using Kalman filtering [27], allowing the track to persist through brief occlusions.

Compared to single-direction tracking, FBR improves IoU consistency by 16.8% (measured on 100 validation samples). The pipeline is particularly robust to partial occlusions, maintaining $\geq$ 0.8 IoU even when the target is 50% occluded. Additionally, we adopt the real-time quality monitoring protocol from [4, 22], automatically pausing annotators when their consistency drops below threshold.

3.4. External Dataset Augmentation and Validation

Videos sourced from RVOS datasets often lack the temporal boundaries and negative samples essential for STVG. This is because RVOS videos typically feature annotations that span the entire duration. To address this, we introduce an automated negative sample generation pipeline (Fig. 4 (b)). For each video with full-duration annotations, we:

1. **Semantic aspect selection.** To create reasonable distractors while ensuring context consistency, we randomly select *one* of three semantic aspects to modify: *i) object* (e.g., "soccer ball" $\rightarrow$ "basketball"), *ii) action* (e.g., "kicking" $\rightarrow$ "holding"), or *iii) scene* (e.g., "on a beach" $\rightarrow$ "in a park"), while preserving the other two aspects.
2. **Counterfactual caption generation.** Using LLMs (Qwen3 [41] or DeepSeek [21]), we generate modified captions based on the selected aspect. For example, "person kicking a soccer ball on a beach" may become "person kicking a basketball on a beach" (object change) while keeping action and scene unchanged.
3. **Negative video synthesis.** We take the counterfactual caption as the input of video generation model WAN [34] to synthesize negative video clips that visually match the original context but exhibit the modified semantic aspect. These generated clips are temporally concatenated to the original video—as a prefix or suffix—creating plausible but incorrect temporal segments.
4. **Manual validation.** All generated samples undergo manual validation to ensure: *(i)* clear semantic distinction from the ground truth, *(ii)* caption-video alignment verification (using Keye-VL [31] for ambiguous cases), and *(iii)* bounding box accuracy (refined via FBR if needed). Samples that fail validation are regenerated with more significant modifications.

4. DeepSTG Evaluation System

In this section, we introduce **DeepSTG**, a systematic evaluation framework that quantifies dataset quality across four complementary dimension. We apply DeepSTG to comprehensively evaluate seven benchmarks, demonstrating OmniGround significantly outperforms across all metrics.

4.1. DeepSTG Evaluation System Design

Our evaluation system is guided by three design principles:

(1) Multi-Faceted Assessment: No single metric can capture dataset quality. DeepSTG combines multiple dimensions, including annotation fidelity (CMA), visual challenge (FCI), linguistic diversity (VSBI), and distributional balance (NEI), to provide holistic evaluation.

(2) Task-Aware Design: Each of four selected metrics directly evaluate a critical part of the STVG task: CMA ensures the fundamental correctness of annotations, which is essential for effective supervised learning. FCI quantifies the difficulty of spatial discrimination, directly impacting the model's localization accuracy. VSBI measures the balance of linguistic cues (temporal/spatial), ensuring a balanced requirement for temporal and spatial reasoning. NEI guards against dataset bias, which is critical for ensuring the model's generalization ability across different scenarios.

(3) Automation and Objectivity: DeepSTG uses automated metrics based on authoritative measures to ensure reproducibility and eliminate subjective bias.

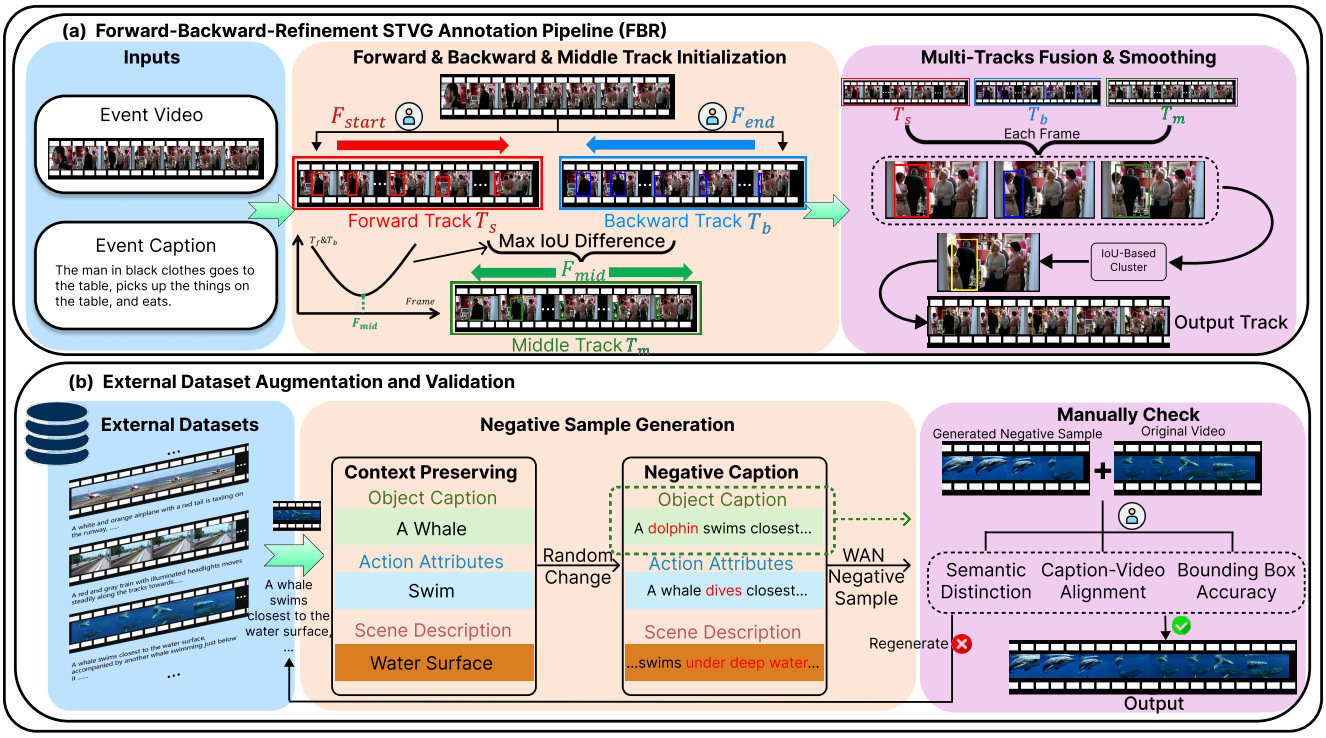


Figure 4. (a) Overview of our Forward-Backward-Refinement pipeline. (b) An illustration of external data augmentation and validation.

Table 1. Results of dataset quality assessment using the DeepSTG evaluation system. '#' indicates the number.

Benchmark	#Category	NEI			FCI	VSBI	CMA Score
		Category	Video Duration	Length of Sentence			
MeVis	14	0.952	0.421	0.635	0.728	0.753	0.624
Ref-YT-VOS	38	0.900	-	0.859	0.719	0.410	0.757
VastTrack	55	0.724	0.520	0.773	0.736	0.679	0.779
HC-STVG	1	-	-	0.778	0.754	0.431	0.678
VIDSTG	40	0.569	0.780	0.659	0.720	0.699	0.701
ST-Align	14	0.628	0.731	0.757	0.718	0.466	0.644
OmniGround (ours)	81	0.992	0.807	0.918	0.770	0.897	0.813

4.1.1. Cross-Modal Semantic Alignment Score

Ensuring consistency between natural language descriptions and their spatio-temporal tubes is crucial for verifying annotation quality. Standard manual verification lacks both scalability and objectivity. To address this, we introduce the **Cross-Modal Semantic Alignment** (CMA) Score, which leverages MLLM reasoning to automatically assess the semantic coherence among the video, caption, and bboxes.

Specifically, for a given video-caption-tube(bbox) triplet (V, Q, B), we sample N key frames $\mathcal{F} = \{f_1, f_2, \ldots, f_N\}$ within the temporal segment $[T_{start}, T_{end}]$ defined by tube B. For each frame f_i and its bbox b_i, we prompt GPT-4o [1] to evaluate three semantic aspects:

(1) Object Presence Score (S_{obj}): Does the target object mentioned in Q exist within b_i?

(2) Action Accuracy Score (S_{act}): Is the action/state described in Q accurately performed by the object in b_i?

(3) Context Consistency Score (S_{ctx}): Are the spatial/temporal relations described in Q consistent with the visual context in f_i and video V?

Each aspect receives a score in $[0, 1]$. The CMA score averages across all frames and aspects, and higher scores indicate better annotation quality and semantic integrity:

$$\mathrm{CMA} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\mathrm{S}_{obj}(f_i) + \mathrm{S}_{act}(f_i) + \mathrm{S}_{ctx}(f_i)}{3}\right] \quad (1)$$

4.1.2. Foreground Complexity Index

STVG models must distinguish targets from visually similar distractors, especially when multiple same-class instances coexist (e.g., the left red car among three red cars). To measure visual clutter, we propose the **Foreground Complexity Index** (FCI), which quantifies intra-class visual similarity.

Given video V, we sample 1 frame per second and detect all foreground objects using YOLOv11 [16], obtaining object set $\mathcal{O} = \{o_i\}_{i=1}^{N}$ with feature vectors $\mathcal{V} = \{v_i\}_{i=1}^{N}$. For each category C, we compute intra-class mean similarity:

$$S_{\text{in}}(C) = \frac{1}{|\mathcal{O}_C|^2 - |\mathcal{O}_C|} \sum_{o_i, o_j \in \mathcal{O}_C, i \neq j} \text{CosSim}(v_i, v_j) \quad (2)$$

where $\mathcal{O}_C$ denotes objects of category C (computed only if $|\mathcal{O}_C| \geq 2$). The FCI averages across all categories:

$$\text{FCI} = \frac{1}{|\mathcal{C}|} \sum_{C \in \mathcal{C}} S_{\text{in}}(C), \text{FCI} \in [0, 1] \quad (3)$$

where higher values indicate objects within the same category are visually similar (harder to distinguish), while lower values suggest diverse appearances (easier grounding).

4.1.3. Verb-Spatial Balance Index

STVG queries require action ("running") and spatial ("in front of") reasoning. Imbalanced datasets biased toward one type hinder model generalization. We introduce **Verb-Spatial Balance Index (VSBI)** to measure linguistic diversity. We categorize each word in captions into three types:

- **Action Cues ($\mathcal{A}$):** Actions ("run") and states ("stand")
- **Spatial Cues ($\mathcal{S}$):** Directional ("towards, away") and locational ("behind, next to")
- **Mixed Cues ($\mathcal{M}$):** Combining both ("climbing up")

Let $P_{\text{actual}} = [P_{\mathcal{A}}, P_{\mathcal{S}}, P_{\mathcal{M}}]$ denote the normalized frequency distribution, where $P_{\mathcal{A}} + P_{\mathcal{S}} + P_{\mathcal{M}} = 1$. The ideal balanced distribution is $P_{\text{ideal}} = [1/4, 1/4, 1/2]$. VSBI $\in [0, 1]$ measures distance from this ideal:

$$\text{VSBI} = 1 - \frac{\|P_{\text{actual}} - P_{\text{ideal}}\|_2}{\sqrt{7/8}} \quad (4)$$

where $\sqrt{7/8}$ is the maximum possible Euclidean distance.

4.1.4. Normalized Entropy Index

To quantify the distributional uniformity across object category, video duration, and text description length, we introduce the Normalized Entropy Index (NEI). The NEI evaluates the balance of data, where a higher score indicates less biased distribution across the dataset:

$$\text{NEI} = \frac{H}{\log N}, H = -\sum_{i=1}^{N} p_i \log p_i \quad (5)$$

where $H, \log N, p_i$, represents the Shannon Entropy [20], Maximum Entropy [39], and relative frequency of an instance falling into a specific range (e.g., a specific category).

4.2. STVG Dataset Analyze and Comparison

We apply DeepSTG to evaluate seven benchmarks: four STVG benchmarks (HC-STVGv2 [30], VidSTG [45], ST-Align [17], OmniGround) and three RVOS benchmarks (MeVis [8], Ref-YouTube-VOS [25], VastTrack [24]). While RVOS datasets are not designed for temporal grounding, they are widely used as source data for STVG research (Sec. 3.1) and provide a valuable baseline for assessing data diversity and visual complexity. Including them enables a more comprehensive understanding of the data landscape from which STVG benchmarks are constructed. Tab. 1 show the evaluation results and exhibit some limitations:

(1) Category Bias: HC-STVG focuses exclusively on persons (category NEI = 0), while VidSTG shows severe imbalance (NEI = 0.569). In contrast, OmniGround achieves near-perfect uniformity (NEI = 0.992), covering 81 categories with balanced representation.

(2) Low Visual Complexity: Most benchmarks have FCI < 0.78, indicating limited visual clutter. ST-Align (FCI = 0.731) and VidSTG (FCI = 0.780) feature relatively simple foregrounds. OmniGround achieves FCI = 0.807, providing more challenging spatial discrimination scenarios.

(3) Linguistic Imbalance: Many datasets skew toward either action-heavy (HC-STVG: VSBI = 0.778) or spatial-heavy (Ref-YT-VOS: VSBI = 0.859) captions. OmniGround attains the highest VSBI (0.918), indicating balanced coverage of action, spatial, and mixed cues.

(4) Annotation Quality: OmniGround's CMA score (0.813) surpasses others, validating the effectiveness of our FBR pipeline. The gap is particularly pronounced compared to VidSTG (0.701) and ST-Align (0.644).

(5) Incomplete Metrics. Some benchmarks exhibit incomplete NEI values due to zero variance: HC-STVG (single category), HC-STVG and Ref-YT-VOS (narrow duration ranges). This underscores the lack of diversity in existing data, motivating OmniGround's construction.

These findings reveal that existing benchmarks are limited on either specific categories (HC-STVG) or imbalanced distributions and low complexity (VidSTG, ST-Align). OmniGround achieves high scores across all dimensions on DeepSTG, creating a new standard for STVG evaluation.

5. Experiments

5.1. Experiments Setup

Datasets. For comprehensive evaluation, we first evaluate state-of-the-art STVG models on OmniGround to analyze their limitations in real-world scenarios. Then, to validate our proposed PG-TAF framework, we conduct experiments on four benchmarks following standard evaluation protocols, including HC-STVG-v2 [30], VidSTG [45] (Declarative/Interrogative), and OmniGround. To ensure fair comparison, we primarily adopt results from original papers

Table 2. Overall comparison on OmniGround and performance across three complex real-world scenarios.

Models	Overall Performance				Uncommon Categories			Multiple Similar Target Objects			Deep Syntactic Complexity		
	m_tIoU	m_vIoU	vIoU@0.3	vIoU@0.5	m_tIoU	m_vIoU	vIoU@0.3	m_tIoU	m_vIoU	vIoU@0.3	m_tIoU	m_vIoU	vIoU@0.3
Non-generative and task-specific models													
TA-STVG	44.6	28.0	40.8	21.5	24.9	14.2	19.3	17.8	11.3	8.7	17.1	9.6	6.0
CG-STVG	47.5	30.4	45.2	23.4	29.5	17.1	23.6	22.3	14.2	15.0	20.6	11.6	8.3
MLLMs with Parameter Sizes of 7B													
LLaVA-ST	19.7	8.7	10.2	1.9	17.8	5.6	2.4	17.6	6.6	2.8	14.7	4.9	1.5
Qwen2.5-VL	36.6	23.2	33.5	16.5	21.4	13.6	18.2	18.9	12.2	11.5	13.9	7.1	3.1
VideoMolmo	30.2	15.7	15.2	7.3	21.2	13.4	18.7	19.1	13.5	13.3	11.1	7.5	2.0

when available. When such results are not available, we assess the models using LMMs-Eval [43] or official scripts.

Baselines. We evaluate two types of models: 1) MLLMs (7B) including Qwen2.5-VL [3], SpaceVLLM [35], LLaVa-ST [17], VideoMolmo [2]. All models are evaluated using their officially released checkpoints fine-tuned on STVG instruction data. 2) Non-generative task-specific models, including CG-STVG [11] and TA-STVG [12]. These models serve as SOTA baselines representing traditional methods.

Metrics. Following [15, 28, 40], we use m_tIoU, m_vIoU, and vIoU@R (R $\in \{0.3, 0.5\}$) as evaluation metrics. m_tIoU evaluates temporal localization, while m_vIoU and vIoU@R measure the accuracy of spatial localization.

5.2. Experimental Results and Analysis

5.2.1. Overall Performance on OmniGround

Tab. 2 shows the overall results, revealing a significant performance drop across all evaluated models. Key observations confirm inherent model weaknesses:

(1) Large Variance Among MLLMs. Qwen2.5-VL (36.6% m_tIoU) signifficantly outperforms LLaVA-ST. This inconsistency suggests that STVG capability is highly sensitive to model architecture and training data rather than being an emergent property of scale.

(2) Low Spatial Precision. All models struggle with high vIoU thresholds (best: CG-STVG at 23.4%), indicating that even when temporal localization is correct, spatial predictions are often imprecise. This confirms the challenge posed by OmniGround's complex foregrounds (high FCI).

(3) Task-Specific Models Outperform MLLMs. Non-generative models like CG-STVG (30.4% m_vIoU) significantly surpass MLLMs (best: Qwen2.5-VL at 23.2%). The gap is particularly pronounced in m_vIoU (+7.2%), highlighting weaknesses in object spatial grounding.

5.2.2. Fine-Grained Analysis Across Complex Scenarios

Tab. 2 show performance across three challenging scenarios, validating the OmniGround design hypotheses:

Uncommon Categories: Models exhibit generalization failure due to category bias. CG-STVG achieves only 29.5% m_tIoU on uncommon objects (vs. 47.5% overall), and Qwen2.5-VL drops by 15.2%. This validates that models are overfitting to frequent categories (e.g., "person"), failing to generalize to balanced categories.

Multiple Similar Targets: This scenario proves challenging across all models. CG-STVG's m_tIoU drops to 22.3% (-25.2% overall), and VideoMolmo to 19.1% (-11.1%). The m_vIoU degradation is even more serious: CG-STVG and Qwen2.5-VL falls to 14.2% (-16.2%) and 13.5% (-11.2%), respectively. This confirms that models struggle to distinguish targets when intra-class similarity is high, reflecting the importance of OmniGround's high FCI.

Deep Syntactic Complexity: Linguistically complex queries (nested spatial relations, multiple clauses) cause the most severe performance decreased. CG-STVG's m_tIoU drops to 20.6% (-26.9%). While MLLMs also perform poorly (Qwen2.5-VL, -9.3%), the decrease is generally less than that of task-specific models, revealing that despite lower absolute localization precision, MLLMs exhibit semantic robustness to linguistic complexity due to their strength in holistic language understanding.

5.3. Prompt-Guided Temporal Alignment

To improve the models' performance and generalization facing real-world scenarios, we propose the **Prompt-Guided Temporal Alignment Framework (PG-TAF)**.

5.3.1. Two-Stage Architecture

The PG-TAF framework is designed as a training-free, two-stage inference architecture that decouples STVG into: *(i)* high-level temporal grounding and *(ii)* fine-grained spatio-temporal propagation (as shown in Fig. 5).

Stage 1: MLLM-Guided Temporal Grounding. The first stage uses an MLLM (Qwen3-VL-8B [41]) to perform coarse temporal localization. We prompt the MLLM with specific query as shown in Fig. 5. Based on MLLMs' ability in holistic video understanding and complex relational reasoning—capabilities that small task-specific models lack—the MLLM can accurately determine the relevant temporal segment, effectively filtering out irrelevant frames and guiding the subsequent spatial localization module.

Stage 2: Agent-Based Spatial Propagation. The second stage performs fine-grained localization within the temporally constrained clip. We use a multi-modal scoring

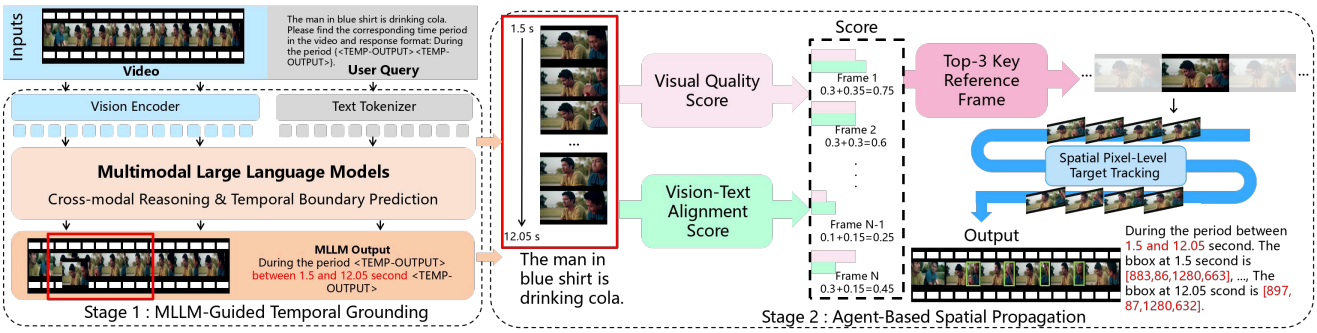


Figure 5. The mainframe of PG-TAF. We decouple STVG into high-level temporal grounding and fine-grained spatio-temporal propagation.

Table 3. Comparison of m_tIou and m_vIoU on four STVG datasets. '-' represents the code is not officially available.

Datasets	HC-STVG		VIDSTG				OmniGround		Average	
Metrics			Declarative sentence		Interrogative sentence					
	m_tIoU	m_vIoU	m_tIoU	m_vIoU	m_tIoU	m_vIoU	m_tIoU	m_vIoU	m_tIoU	m_vIoU
MLLMs with Parameter Sizes of 7B										
LLaVA-ST	21.2	7.6	44.1	14.2	18.5	7.5	19.7	8.7	25.9	9.5
Qwen2.5-VL	22.9	13.0	16.8	10.9	13.8	8.5	36.6	23.3	22.5	13.9
SpaceVLLM	58.0	34.0	47.7	27.4	48.5	25.4	-	-	-	-
VideoMoLMO	44.6	26.8	41.7	15.6	30.2	11.7	30.2	11.7	36.7	16.5
PG-TAF(Ours)	56.2	34.0	48.4	28.1	46.4	24.5	49.2	36.2	49.6	28.2

mechanism to select key reference frames. This mechanism combines semantic segmentation scores S_{seg} from EVF-SAM [44] to ensure visual quality/object clarity, and contrastive text-image similarity score S_{align} from Alpha-CLIP [29] to ensure semantic relevance/match with text description. Subsequently, the top-K frames (K=3 in our experiments) are selected as references based on a composite score: $S_{frame} = \alpha S_{seg} + \beta S_{align}$. We set $\alpha = 0.6, \beta = 0.4$ through grid search on our validation set, which ensure the selected frames have both high visual quality and semantic relevance. We then use pixel-level tracker [6] to propagate object masks across the entire temporal segment, generating the final spatio-temporal tube.

The decoupling addresses the challenge: MLLMs excel at temporal reasoning but lack fine-grained spatial precision, while specialized trackers provide accurate localization but struggle with semantic understanding. PG-TAF combines their complementary strengths in a training-free manner, making it practical for real-world deployment.

5.3.2. Evaluation Results

Tab. 3 shows the performance of PA-TAF across four benchmarks. PG-TAF achieves the highest m_tIoU and m_vIoU (49.2% and 36.2%) on OmniGround, even better than the task-specific models. Notably, PG-TAF achieves these results in a training-free manner, without requiring any fine-tuning on STVG data. This validates our motivations that decoupling temporal reasoning (MLLM's strength) from spatial localization (tracker's strength) effectively addresses the challenges posed by OmniGround's complexity. For other benchmarks, PG-TAF maintains a relative better performance comparing with the MLLMs that need additional traning for specific datasets (SpaceVLLM, LLaVA-ST).

6. Conclusion

In this paper, to bridge the gap between the current STVG MLLMs and real-world demands involving diverse objects and complex queries. We first construct OmniGround, a comprehensive benchmark (3,475 videos, 81 balanced categories) with high-quality FBR annotation pipeline. Second, we introduce DeepSTG, a systematic framework that quantifies dataset quality, revealing critical issues in existing benchmarks like category bias and linguistic imbalance. Extensive experiments show that the current state-of-the-art STVG models drop up to 23.4% on OmniGround. Finally, we propose PG-TAF, a training-free, two-stage framework that decoupling high-level temporal understanding from specialized spatial tracking. We hope that our work could inspire ideas to improve STVG performance in real-world applications.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 5

[2] Ghazi Shazan Ahmad, Ahmed Heakl, Hanan Gani, Abdelrahman Shaker, Zhiqiang Shen, Fahad Shahbaz Khan, and Salman Khan. Videomolmo: Spatio-temporal grounding meets pointing. *arXiv preprint arXiv:2506.05336*, 2025. 1, 2, 3, 7

[3] 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. 7

[4] Alec Burmania, Srinivas Parthasarathy, and Carlos Busso. Increasing the reliability of crowdsourcing evaluations using online quality assessment. *IEEE Transactions on Affective Computing*, 7(4):374–388, 2015. 4

[5] Zhongang Cai, Yubo Wang, Qingping Sun, Ruisi Wang, Chenyang Gu, Wanqi Yin, Zhiqian Lin, Zhitao Yang, Chen Wei, Xuanke Shi, et al. Has gpt-5 achieved spatial intelligence? an empirical study. *arXiv preprint arXiv:2508.13142*, 2025. 1

[6] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3151–3161, 2024. 8

[7] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 1

[8] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2694–2703, 2023. 3, 6

[9] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024. 1

[10] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022. 3

[11] 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, 7

[12] Xin Gu, Yaojie Shen, Chenxi Luo, Tiejian Luo, Yan Huang, Yuewei Lin, Heng Fan, and Libo Zhang. Knowing your target: Target-aware transformer makes better spatio-temporal video grounding. *arXiv preprint arXiv:2502.11168*, 2025. 7

[13] Yongxin Guo, Jingyu Liu, Mingda Li, Qingbin Liu, Xi Chen, and Xiaoying Tang. Trace: Temporal grounding video llm via causal event modeling. *arXiv preprint arXiv:2410.05643*, 2024. 3

[14] 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. 1, 3

[15] Yang Jin, Zehuan Yuan, Yadong Mu, et al. Embracing consistency: A one-stage approach for spatio-temporal video grounding. *Advances in Neural Information Processing Systems*, 35:29192–29204, 2022. 7

[16] Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024. 6

[17] Hongyu Li, Jinyu Chen, Ziyu Wei, Shaofei Huang, Tianrui Hui, Jialin Gao, Xiaoming Wei, and Si Liu. Llava-st: A multimodal large language model for fine-grained spatial-temporal understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8592–8603, 2025. 1, 2, 3, 6, 7

[18] Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Van Tu Vu, et al. Groundinggpt: Language enhanced multi-modal grounding model. *arXiv preprint arXiv:2401.06071*, 2024. 1, 3

[19] Shuo Liang, Yiwu Zhong, Zi-Yuan Hu, Yeyao Tao, and Liwei Wang. Fine-grained spatiotemporal grounding on egocentric videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9385–9395, 2025. 3

[20] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 2002. 6

[21] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 4

[22] Reza Lotfian and Carlos Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483, 2017. 4

[23] Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. In *European Conference on Computer Vision*, pages 417–435. Springer, 2024. 3

[24] Liang Peng, Junyuan Gao, Xinran Liu, Weihong Li, Shaohua Dong, Zhipeng Zhang, Heng Fan, and Libo Zhang. Vasttrack: Vast category visual object tracking. *Advances in Neural Information Processing Systems*, 37:130797–130818, 2024. 6

[25] 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, 6

[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. 1

[27] Dan Simon. Kalman filtering. *Embedded systems programming*, 14(6):72–79, 2001. 4

[28] Rui Su, Qian Yu, and Dong Xu. Stvgbert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1533–1542, 2021. 3, 7

[29] Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alpha-clip: A clip model focusing on wherever you want. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13019–13029, 2024. 8

[30] 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, 6

[31] Kwai Keye Team, Biao Yang, Bin Wen, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, et al. Kwai keye-vl technical report. *arXiv preprint arXiv:2507.01949*, 2025. 4

[32] Jovana Videnovic, Alan Lukezic, and Matej Kristan. A distractor-aware memory for visual object tracking with sam2. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24255–24264, 2025. 4

[33] Joseph Raj Vishal, Divesh Basina, Aarya Choudhary, and Bharatesh Chakravarthi. Eyes on the road: State-of-the-art video question answering models assessment for traffic monitoring tasks. *arXiv preprint arXiv:2412.01132*, 2024. 1

[34] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 4

[35] 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. 1, 3, 7

[36] Shansong Wang, Mingzhe Hu, Qiang Li, Mojtaba Safari, and Xiaofeng Yang. Capabilities of gpt-5 on multimodal medical reasoning. *arXiv preprint arXiv:2508.08224*, 2025. 1

[37] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36:61501–61513, 2023. 1

[38] Weikang Wang, Jing Liu, Yuting Su, and Weizhi Nie. Efficient spatio-temporal video grounding with semantic-guided feature decomposition. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 4867–4876, New York, NY, USA, 2023. Association for Computing Machinery. 3

[39] Nailong Wu. *The maximum entropy method*. Springer Science & Business Media, 2012. 6

[40] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16442–16453, 2022. 7

[41] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 1, 4, 7

[42] Yi Yang, Yiming Xu, Timo Kaiser, Hao Cheng, Bodo Rosenhahn, and Michael Ying Yang. Multi-object tracking retrieval with llava-video: A training-free solution to mot25-stag challenge, 2025. 1

[43] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 881–916, 2025. 7

[44] Yuxuan Zhang, Tianheng Cheng, Lianghui Zhu, Rui Hu, Lei Liu, Heng Liu, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. Evf-sam: Early vision-language fusion for text-prompted segment anything model. *arXiv preprint arXiv:2406.20076*, 2024. 8

[45] 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, 6

[46] Mian Zhou, Angelos Stefanidis, Nan Jiang, Zezhou Sui, and Zhikun Feng. Language-led visual grounding for human computer interaction. *Preprints*, 2023. 1