Novelty Assessment Report

Paper: OmniSTVG: Toward Spatio-Temporal Omni-Object Video Grounding
PDF URL: https://openreview.net/pdf?id=azcQJtcYTE
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

We introduce spatio-temporal omni-object video grounding, dubbed $\textbf{OmniSTVG}$, a new STVG task aiming to localize spatially and temporally all targets mentioned in the textual query within videos. Compared to classic STVG locating only a single target, OmniSTVG enables localization of not only an arbitrary number of text-referred targets but also their interacting counterparts in the query from the video, making it more flexible and practical in real scenarios for comprehensive understanding. In order to facilitate exploration of OmniSTVG, we propose $\textbf{BOSTVG}$, a large-scale benchmark dedicated to OmniSTVG. Specifically, BOSTVG contains 10,018 videos with 10.2M frames and covers a wide selection of 287 classes from diverse scenarios. Each sequence, paired with a free-form textual query, encompasses a varying number of targets ranging from 1 to 10. To ensure high quality, each video is manually annotated with meticulous inspection and refinement. To our best knowledge, BOSTVG, to date, is the first and the largest benchmark for OmniSTVG. To encourage future research, we present a simple yet effective approach, named $\textbf{OmniTube}$, which, drawing inspiration from Transformer-based STVG methods, is specially designed for OmniSTVG and demonstrates promising results. By releasing BOSTVG, we hope to go beyond classic STVG by locating every object appearing in the query for more comprehensive understanding, opening up a new direction for STVG. Our benchmark and code will be released.

Core Task Landscape

This paper addresses: **spatio-temporal omni-object video grounding**

A total of **50 papers** were analyzed and organized into a taxonomy with **25 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Spatio-Temporal Video Grounding and Localization**
- **Weakly-Supervised and Unsupervised Video Object Localization**
- **Video-Language Models and Foundation Models**
- **Video Object Detection and Tracking**
- **Video Scene Understanding and Representation**
- **Multimodal Video Understanding**
- **Video Segmentation and Manipulation**
- **Specialized Video Analysis Tasks**
- **Spatio-Temporal Modeling Foundations**

Complete Taxonomy Tree

- spatio-temporal omni-object video grounding Survey Taxonomy
- Spatio-Temporal Video Grounding and Localization
  - Multi-Object and Omni-Object Grounding ★ (4 papers)
  - [0] OmniSTVG: Toward Spatio-Temporal Omni-Object Video Grounding (Anon et al., 2026) View paper
  - [4] SVAG-Bench: A Large-Scale Benchmark for Multi-Instance Spatio-temporal Video Action Grounding (Hannan, 2025) View paper
  - [8] Described spatial-temporal video detection (Ji Wei, 2024) View paper
  - [23] Where Does It Exist: Spatio-Temporal Video Grounding for Multi-Form Sentences (Zhu Zhang, 2020) View paper
  - Single-Object Grounding (3 papers)
  - [33] Weakly-supervised video object grounding by exploring spatio-temporal contexts (Xun Yang, 2020) View paper
  - [36] Video-GroundingDINO: Towards Open-Vocabulary Spatio-Temporal Video Grounding (Syed Talal Wasim, 2023) View paper
  - [46] Video Object Grounding using Semantic Roles in Language Description (Arka Sadhu, 2020) View paper
  - Action-Centric Grounding (2 papers)
  - [7] Interacted object grounding in spatio-temporal human-object interactions (Liu Xiao-yang, 2025) View paper
  - [13] DORi: Discovering Object Relationships for Moment Localization of a Natural Language Query in a Video (Cristian RodrÃguez-Opazo, 2021) View paper
  - Temporal Moment Localization (5 papers)
  - [5] Universal Video Temporal Grounding with Generative Multi-modal Large Language Models (Li, 2025) View paper
  - [19] Unloc: A unified framework for video localization tasks (Shen Yan, 2023) View paper
  - [20] Gaussian Mixture Proposals with Pull-Push Learning Scheme to Capture Diverse Events for Weakly Supervised Temporal Video Grounding (Sunoh Kim, 2023) View paper
  - [30] Grounding-MD: Grounded Video-language Pre-training for Open-World Moment Detection (Zhuang Weijun, 2025) View paper
  - [43] ProTÃ©GÃ©: Untrimmed Pretraining for Video Temporal Grounding by Video Temporal Grounding (Lan Wang, 2023) View paper

- Weakly-Supervised and Unsupervised Video Object Localization
  - Weakly-Supervised Localization with Class Activation (3 papers)
  - [1] Colo-cam: Class activation mapping for object co-localization in weakly-labeled unconstrained videos (Belharbi, 2025) View paper
  - [3] Tcam: Temporal class activation maps for object localization in weakly-labeled unconstrained videos (Soufiane Belharbi, 2023) View paper
  - [27] Leveraging transformers for weakly supervised object localization in unconstrained videos (Murtaza, 2024) View paper
  - Unsupervised Object Discovery and Co-Localization (2 papers)
  - [22] Efficient video object co-localization with co-saliency activated tracklets (Koteswar Rao Jerripothula, 2018) View paper
  - [38] Unsupervised object discovery and tracking in video collections (Suha Kwak, 2015) View paper
  - Zero-Shot Video Grounding (2 papers)
  - [24] Objects2action: Classifying and localizing actions without any video example (Mihir Jain, 2015) View paper
  - [28] Zero-Shot Video Grounding With Pseudo Query Lookup and Verification (Yu Lu, 2024) View paper
- Video-Language Models and Foundation Models
  - General Video-Language Foundation Models (2 papers)
  - [11] General object foundation model for images and videos at scale (Junfeng Wu, 2024) View paper
  - [18] Video owl-vit: Temporally-consistent open-world localization in video (Georg Heigold, 2023) View paper
  - Video Large Language Models with Grounding (4 papers)
  - [32] 1+ 1> 2: Detector-Empowered Video Large Language Model for Spatio-Temporal Grounding and Reasoning (Shida Gao, 2025) View paper
  - [34] VideoRefer Suite: Advancing Spatial-Temporal Object Understanding with Video LLM (Yuqian Yuan, 2024) View paper
  - [35] Object-centric Video Question Answering with Visual Grounding and Referring (Wang Haochen, 2025) View paper
  - [37] Know-Show: Benchmarking Video-Language Models on Spatio-Temporal Grounded Reasoning (Chinthani Sugandhika, 2025) View paper
  - Open-Vocabulary and Open-World Grounding (2 papers)
  - [26] Grounding Language in Images and Videos (Unknown, 2024) View paper
  - [31] GMIS: Marrying Grounding-DINO and Motion Iterative Segment Anything Model for Referring Video Object Segmentation (Junchi Zhang, 2025) View paper
- Video Object Detection and Tracking
  - Video Object Detection with Temporal Aggregation (2 papers)
  - [6] Memory Enhanced Global-Local Aggregation for Video Object Detection (Yihong Chen, 2020) View paper
  - [21] Localizing spatially and temporally objects and actions in videos (Kalogeiton, 2018) View paper
  - Multi-Object Tracking (1 papers)
  - [2] Temporal-Spatial Feature Interaction Network for Multi-Drone Multi-Object Tracking (Han Wu, 2025) View paper
  - Query-Guided and One-Shot Localization (2 papers)
  - [9] QDETRv: query-guided DETR for one-shot object localization in videos (Yogesh Kumar, 2024) View paper
  - [12] Sketch-based video object localization (Sang-Min Woo, 2024) View paper
- Video Scene Understanding and Representation
  - Video Scene Graph Generation (3 papers)
  - [10] Videos as space-time region graphs (X. Wang, 2018) View paper
  - [44] Panoptic Video Scene Graph Generation (Jingkang Yang, 2023) View paper
  - [47] Video Scene Graph Generation with Spatial-Temporal Knowledge (Tao Pu, 2023) View paper
  - Video Captioning with Object Grounding (1 papers)
  - [42] Grounded Objects and Interactions for Video Captioning (Ma, 2017) View paper
  - Video Pre-Training with Spatial-Temporal Alignment (1 papers)
  - [48] Structured Video-Language Modeling with Temporal Grouping and Spatial Grounding (Xiong, 2023) View paper
- Multimodal Video Understanding
  - Audio-Visual Object Localization (3 papers)
  - [14] Egocentric audio-visual object localization (Chao Huang, 2023) View paper
  - [15] Crab: A unified audio-visual scene understanding model with explicit cooperation (Henghui Du, 2025) View paper
  - [29] Joint audio-video object localization and tracking (Norbert Strobel, 2002) View paper
- Video Segmentation and Manipulation
  - Unsupervised Video Object Segmentation (1 papers)
  - [17] Reciprocal Transformations for Unsupervised Video Object Segmentation (Sucheng Ren, 2021) View paper
  - Video Inpainting and Object Manipulation (2 papers)
  - [16] Deep Video Inpainting Localization Using Spatial and Temporal Traces (Shujin Wei, 2022) View paper
  - [25] LoVoRA: Text-guided and Mask-free Video Object Removal and Addition with Learnable Object-aware Localization (Zhihan Xiao, 2025) View paper
- Specialized Video Analysis Tasks
  - Video Forensics and Authenticity Detection (1 papers)
  - [39] DAVID-XR1: Detecting AI-Generated Videos with Explainable Reasoning (Gao Yifeng, 2025) View paper
  - Efficient Large-Scale Video Querying (1 papers)
  - [41] LOVO: Efficient Complex Object Query in Large-Scale Video Datasets (Yu-Xin Liu, 2025) View paper
  - Video Domain Generalization (1 papers)
  - [40] Diversifying Spatial-Temporal Perception for Video Domain Generalization (Lin, 2023) View paper
  - Motion-Based Salient Object Localization (1 papers)
  - [45] A particle filtering approach to salient video object localization (Charles M. Gray, 2014) View paper
- Spatio-Temporal Modeling Foundations
  - Temporal Covariance and Attention Pooling (1 papers)
  - [50] Temporal-attentive Covariance Pooling Networks for Video Recognition (Gao Zilin, 2021) View paper
  - Spatio-Temporal Gaussian Process Models (1 papers)
  - [49] Learning Temporal Evolution of Spatial Dependence with Generalized Spatiotemporal Gaussian Process Models (Lan Shiwei, 2019) View paper

Narrative

Core task: spatio-temporal omni-object video grounding seeks to localize and track multiple objects in video sequences based on natural language descriptions, integrating spatial bounding boxes with temporal dynamics. The field's taxonomy reveals several complementary branches. Spatio-Temporal Video Grounding and Localization focuses on methods that jointly reason about when and where objects appear, often employing transformer-based architectures to align linguistic queries with video regions across frames. Weakly-Supervised and Unsupervised Video Object Localization explores techniques that reduce annotation burden by learning from noisy or incomplete labels. Video-Language Models and Foundation Models leverage large-scale pretraining to build general-purpose representations that transfer across diverse grounding tasks. Meanwhile, Video Object Detection and Tracking emphasizes robust instance-level tracking, and Video Scene Understanding branches into holistic scene graphs and relational reasoning. Multimodal Video Understanding integrates audio, text, and visual cues, while Video Segmentation and Manipulation addresses pixel-level precision and editing. Specialized Video Analysis Tasks cover domain-specific challenges such as egocentric or sketch-based localization, and Spatio-Temporal Modeling Foundations provides core architectural innovations like memory-enhanced detection and temporal attention mechanisms.

Within the Multi-Object and Omni-Object Grounding cluster, recent works tackle the challenge of grounding diverse object categories and multiple instances simultaneously. OmniSTVG[0] exemplifies this direction by proposing a unified framework for omni-object grounding that handles varied object types and complex temporal dependencies. Nearby efforts such as SVAG-Bench[4] introduce comprehensive benchmarks to evaluate grounding across diverse scenarios, while Universal Video Grounding[5] aims for generalization across object classes and query styles. Described Spatial-Temporal Detection[8] emphasizes fine-grained alignment between descriptive language and spatio-temporal tubes, contrasting with earlier memory-based approaches like Memory Enhanced Detection[6] that prioritize long-range temporal consistency. These works collectively highlight trade-offs between model generality, annotation efficiency, and temporal reasoning depth, with OmniSTVG[0] positioned as a holistic solution that bridges multi-object tracking with flexible language-driven localization.

Related Works in Same Category

The following **3 sibling papers** share the same taxonomy leaf node with the original paper:

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

Authors: Hannan, Tanveer, Tanveer Hannan, Weber, Mark, et al. (25 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

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 obje...

Relationship Analysis

Both papers belong to the Multi-Object and Omni-Object Grounding category, focusing on localizing multiple objects simultaneously in videos based on textual queries with spatio-temporal tubes. They overlap in addressing the limitation of single-object grounding by enabling multi-instance localization with spatial and temporal annotations. The key difference is that OmniSTVG grounds all objects mentioned in a query (including interacting counterparts) with 10,018 videos covering 287 classes, while SVAG-Bench specifically focuses on action-centric grounding where objects are localized based on their actions, with 688 videos and 903 unique verbs emphasizing fine-grained action understanding.

---

2. Described spatial-temporal video detection

Authors: Ji Wei, Liu Xiangyan, Wei Ji, Sun Ying-fei, Xiangyan Liu, et al. (19 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Detecting visual content on language expression has become an emerging topic in the community. However, in the video domain, the existing setting, i.e., spatial-temporal video grounding (STVG), is formulated to only detect one pre-existing object in each frame, ignoring the fact that language descriptions can involve none or multiple entities within a video. In this work, we advance the STVG to a more practical setting called described spatial-temporal video detection (DSTVD) by overcoming the a...

Relationship Analysis

Both papers belong to the Multi-Object and Omni-Object Grounding category, addressing the challenge of localizing multiple objects mentioned in textual queries within videos using spatio-temporal tubes. They overlap in their core motivation to extend beyond single-object STVG by grounding multiple query-mentioned targets simultaneously. The key difference is that the original paper (OmniSTVG) aims to ground ALL objects mentioned in the query (omni-object approach) with a larger benchmark (10,018 videos, 287 classes), while the candidate paper (DVD-ST) focuses on grounding multiple objects of potentially the same class with a smaller dataset (2,750 videos, 163 classes) and emphasizes handling varying numbers of objects (zero to many) across frames.

---

3. Where Does It Exist: Spatio-Temporal Video Grounding for Multi-Form Sentences

Authors: Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, et al. (6 authors total) | **Year/Venue**: 2020 • Computer Vision and Pattern Recognition | **URL**: View paper

Abstract

In this paper, we consider a novel task, Spatio-Temporal Video Grounding for Multi-Form Sentences (STVG). Given an untrimmed video and a declarative/interrogative sentence depicting an object, STVG aims to localize the spatio-temporal tube of the queried object. STVG has two challenging settings: (1) We need to localize spatio-temporal object tubes from untrimmed videos, where the object may only exist in a very small segment of the video; (2) We deal with multi-form sentences, including the dec...

Relationship Analysis

Both papers belong to the Multi-Object and Omni-Object Grounding category, addressing spatio-temporal video grounding with multiple objects. They overlap in tackling the challenge of localizing multiple objects in videos using spatio-temporal tubes and handling both declarative and interrogative queries. The key difference is that the original paper (OmniSTVG) explicitly aims to ground ALL objects mentioned in the query simultaneously with a dedicated benchmark (BOSTVG, 10,018 videos), while the candidate paper (VidSTG) focuses on grounding a single queried object per sentence from multi-form sentences (declarative/interrogative) with a smaller dataset (6,924 videos from VidOR).

Contributions Analysis

Overall novelty summary. The paper introduces OmniSTVG, a task for localizing all text-mentioned targets and their interacting counterparts in videos with spatio-temporal tubes. It resides in the Multi-Object and Omni-Object Grounding leaf, which contains four papers including this one. This leaf sits within the broader Spatio-Temporal Video Grounding and Localization branch, distinguishing itself from Single-Object Grounding (three papers) and Action-Centric Grounding (two papers) siblings. The relatively small cluster size

suggests this is an emerging research direction rather than a saturated area, though the parent branch encompasses multiple active subcategories addressing temporal moment localization and action-centric methods.

The taxonomy reveals neighboring work in closely related directions. Single-Object Grounding methods focus on localizing one primary target per query, while Action-Centric Grounding emphasizes events and their interacting objects. Temporal Moment Localization (five papers) addresses temporal boundaries without spatial boxes, representing a complementary problem formulation. The broader Video-Language Models branch (six papers across three leaves) explores foundation models with grounding capabilities, and Weakly-Supervised Localization (seven papers) reduces annotation requirements. OmniSTVG's emphasis on localizing all query-mentioned objects plus their interacting counterparts positions it at the intersection of multi-object tracking and comprehensive language-driven understanding, diverging from single-target or action-only paradigms.

Among thirty candidates examined, the OmniSTVG task contribution shows one refutable candidate from ten examined, suggesting some prior work addresses similar multi-object grounding formulations within the limited search scope. The BOSTVG benchmark contribution examined ten candidates with zero refutations, indicating the dataset's scale and annotation protocol may offer distinct value. The OmniTube method contribution also examined ten candidates without refutation, though this does not preclude related architectural approaches in the broader literature. The statistics reflect a focused semantic search rather than exhaustive coverage, meaning additional relevant work may exist beyond the top-thirty matches analyzed here.

Based on the limited search scope of thirty semantically similar papers, the work appears to advance multi-object grounding with a comprehensive benchmark and method. The task formulation shows some overlap with prior efforts, while the dataset and approach contributions exhibit less direct precedent among examined candidates. The taxonomy context suggests this research direction remains relatively sparse compared to single-object or temporal-only localization, though the analysis cannot definitively assess novelty beyond the top-K semantic neighborhood explored.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: OmniSTVG task for spatio-temporal omni-object video grounding

Description: The authors propose a new task called OmniSTVG that extends classic spatio-temporal video grounding by localizing all objects mentioned in a textual query, including both targets of interest and their interacting counterparts, rather than only a single target.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Zero-shot Natural Language Video Localization

URL: View paper

Brief Assessment

Zero-shot Natural Language[58] addresses zero-shot natural language video localization without paired annotations, focusing on temporal moment localization with single queries. This differs from OmniSTVG's goal of localizing all objects mentioned in a query within videos.

---

2. VideoGLaMM : A Large Multimodal Model for Pixel-Level Visual Grounding in Videos

URL: View paper

Brief Assessment

VideoGLaMM[57] focuses on pixel-level visual grounding (segmentation masks) in videos, while the original paper addresses spatio-temporal bounding box localization for all mentioned objects. These are distinct technical approaches to video grounding.

---

3. Knowing Your Target: Target-Aware Transformer Makes Better Spatio-Temporal Video Grounding

URL: View paper

Brief Assessment

Knowing Your Target[53] focuses on improving single-target spatio-temporal video grounding through target-aware query generation, not on localizing all objects mentioned in queries as in OmniSTVG.

---

4. WINNER: Weakly-supervised hIerarchical decompositioN and aligNment for spatio-tEmporal video gRounding

URL: View paper

Brief Assessment

WINNER[54] addresses weakly-supervised spatio-temporal video grounding with hierarchical decomposition, but focuses on a single target tube per query rather than localizing all objects mentioned in the text. The candidate does not challenge the novelty of grounding multiple objects simultaneously.

---

5. Unleashing the Potential of Multimodal LLMs for Zero-Shot Spatio-Temporal Video Grounding

URL: View paper

Brief Assessment

Unleashing Multimodal LLMs[52] focuses on zero-shot STVG using multimodal large language models with decomposed query strategies, not on proposing a new task for localizing all objects mentioned in queries. The candidate addresses single-target STVG methodology, not the multi-target omni-object grounding task proposed in the original paper.

---

6. Stpro: Spatial and temporal progressive learning for weakly supervised spatio-temporal grounding

URL: View paper

Brief Assessment

Stpro[51] focuses on weakly supervised spatio-temporal video grounding (wstvg) for single-target localization without bounding box supervision, not on localizing all objects mentioned in queries as in OmniSTVG.

---

7. TubeDETR: Spatio-Temporal Video Grounding with Transformers

URL: View paper

Brief Assessment

TubeDETR[56] focuses on single-object spatio-temporal video grounding, not multi-object localization. The paper explicitly states it aims to localize 'a spatio-temporal tube' (singular) for 'the target object' (singular), which differs fundamentally from OmniSTVG's goal of localizing all objects mentioned in a query.

---

8. Weakly-supervised video object grounding by exploring spatio-temporal contexts

URL: View paper

Brief Assessment

Weakly-supervised Video Grounding[33] focuses on weakly-supervised grounding of objects mentioned in sentences, not on the comprehensive localization of all targets including interacting counterparts as in OmniSTVG.

---

9. VoCap: Video Object Captioning and Segmentation from Any Prompt

URL: View paper

Brief Assessment

VoCap[55] focuses on video object segmentation and captioning from various prompts (text, box, mask), producing spatio-temporal masks with captions. This differs fundamentally from OmniSTVG, which localizes all text-mentioned objects with bounding box tubes from free-form queries without requiring mask-level annotations or object captioning.

---

10. Described spatial-temporal video detection

URL: View paper

Prior Art Analysis

Described Spatial-Temporal Detection[8] demonstrates that prior work exists on localizing multiple objects from text queries in videos. The candidate paper explicitly introduces 'described spatial-temporal video detection (dstvd)' which addresses the same core limitation identified by the original paper: that existing STVG methods only detect one object per query. The candidate paper states that STVG 'is formulated to only detect one pre-existing object in each frame, ignoring the fact that language descriptions can involve none or multiple entities within a video' and proposes DSTVD to overcome this limitation by supporting 'grounding from none to many objects.' This directly challenges the novelty claim that OmniSTVG is the first to extend STVG to localize all mentioned objects, as the candidate demonstrates a similar conceptual advancement with their DSTVD task and DVD-ST benchmark.

Evidence

Evidence 1 - **Rationale**: Both papers identify the same limitation of existing STVG (single-object localization) and propose extensions to handle multiple objects. The candidate's DSTVD task directly addresses the same problem space as OmniSTVG, challenging the claim of being first to propose multi-object spatio-temporal grounding. - **Original**: we introduce spatio-temporal omni-object video grounding, dubbedomnistvg, a new stvg task aiming to localize spatially and temporallyalltargets mentioned in the textual query within videos. compared to classic stvg locating only a single target, omnistvg enables localization of not only an arbitrary... - **Candidate**: in the video domain, the existing setting,i.e., spatial-temporal video grounding (stvg), is formulated to only detect one pre-existing object in each frame, ignoring the fact that language descriptions can involve none or multiple entities within a video. in this work, we advance the stvg to a more ...

Evidence 2 - **Rationale**: The candidate's DVD-ST benchmark explicitly supports 'grounding from none to many objects,' which directly parallels the original paper's goal of localizing 'all targets mentioned in the given query.' This demonstrates prior work on the same conceptual task. - **Original**: to mitigate aforementioned limitations of current stvg, the key is to have the ability to locateeverytarget mentioned in the textual query, much like how humans do. thus motivated, we introduce a new type of stvg task, dubbedspatio-temporalomni-objectvideo-grounding (oromnistvg). different from exis... - **Candidate**: dvd-st supports grounding from none to many objects onto the video in response to queries and encompasses a diverse range of over 150 entities, including appearance, actions, locations, and interactions. the extensive breadth and diversity of the dvd-st dataset make it an exemplary testbed for the i...

Contribution 2: BOSTVG benchmark dataset for OmniSTVG

Description: The authors introduce BOSTVG, a large-scale benchmark dataset containing 10,018 videos with over 10 million frames across 287 object categories. Each video is paired with a free-form textual query and manually annotated with spatio-temporal tubes for all mentioned targets.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Human-centric spatio-temporal video grounding with visual transformers

URL: View paper

Brief Assessment

Human-centric Grounding[71] focuses on human-centric spatio-temporal video grounding with a single target person per query, whereas BOSTVG is designed for omni-object grounding with multiple targets (1-10 objects) across 287 diverse object categories, not limited to humans.

---

2. Scene-text grounding for text-based video question answering

URL: View paper

Brief Assessment

Scene-text Grounding[74] focuses on text-based video question answering with scene-text grounding, not general spatio-temporal object video grounding. The candidate addresses scene text regions for QA tasks, while the original paper targets arbitrary object categories for comprehensive video understanding.

---

3. A survey on video temporal grounding with multimodal large language model

URL: View paper

Brief Assessment

Survey Video Grounding[76] is a survey paper on video temporal grounding with multimodal large language models, not a benchmark dataset. It reviews existing datasets and methods but does not introduce a competing benchmark that would refute BOSTVG's novelty as the first large-scale benchmark for omni-object spatio-temporal video grounding.

---

4. A survey on temporal sentence grounding in videos

URL: View paper

Brief Assessment

Survey Temporal Grounding[69] is a survey paper that reviews temporal sentence grounding methods and datasets, but does not introduce any benchmark dataset itself. It discusses existing datasets for temporal grounding but does not claim to create BOSTVG or any similar multi-object spatio-temporal grounding benchmark.

---

5. Tvqa+: Spatio-temporal grounding for video question answering

URL: View paper

Brief Assessment

TVQA+[75] focuses on video question answering with spatio-temporal grounding for QA tasks, not general spatio-temporal video grounding of all mentioned objects. The datasets serve fundamentally different purposes and task formulations.

6. VideoITG: Multimodal Video Understanding with Instructed Temporal Grounding

URL: View paper

Brief Assessment

VideoITG[73] focuses on instructed temporal grounding for frame selection in video-llms, not spatio-temporal object grounding with tube annotations. The datasets serve fundamentally different purposes and task formulations.

7. Context-Guided Spatio-Temporal Video Grounding

URL: View paper

Brief Assessment

Context-Guided Grounding[64] focuses on single-object spatio-temporal video grounding with context-guided localization, not on multi-object omni-grounding benchmarks. The datasets used (hcstvg-v1/-v2, vidstg) are single-object focused.

8. Weakly-supervised video object grounding by exploring spatio-temporal contexts

URL: View paper

Brief Assessment

Weakly-supervised Video Grounding[33] does not introduce a benchmark dataset. It presents a weakly-supervised framework for video object grounding, which is methodologically distinct from the BOSTVG benchmark's focus on comprehensive spatio-temporal annotations for all mentioned objects.

9. Fine-grained spatiotemporal grounding on egocentric videos

URL: View paper

Brief Assessment

Fine-grained Spatiotemporal Grounding[72] focuses on egocentric video grounding with pixel-level mask annotations, while BOSTVG targets omni-object grounding in general videos with spatio-temporal tube annotations. The datasets serve different tasks and video domains.

10. EgoThinker: Unveiling Egocentric Reasoning with Spatio-Temporal CoT

URL: View paper

Brief Assessment

EgoThinker[70] focuses on egocentric video reasoning with question-answering and hand-object grounding, not spatio-temporal video grounding with free-form textual queries and multi-object tube annotations as in BOSTVG.

Contribution 3: OmniTube method for OmniSTVG

Description: The authors develop OmniTube, a Transformer-based approach specifically designed for the OmniSTVG task. It uses text-guided query generation and a spatio-temporal decoder to simultaneously localize multiple objects mentioned in textual queries from videos.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Learning Feature Semantic Matching for Spatio-Temporal Video Grounding

URL: View paper

Brief Assessment

Learning Feature Semantic[65] addresses traditional single-object STVG with vision-text alignment and spatial mislocalization challenges, not the omni-object grounding task that localizes all mentioned targets simultaneously.

2. Query-dependent video representation for moment retrieval and highlight detection

URL: View paper

Brief Assessment

Query-dependent Video Representation[59] focuses on moment retrieval and highlight detection with query-dependent representations, not on spatio-temporal omni-object video grounding with multiple target localization as in OmniTube.

3. LITA: Language Instructed Temporal-Localization Assistant

URL: View paper

Brief Assessment

LITA[63] focuses on temporal localization in video understanding with language instructions ('when?' questions), while OmniTube addresses spatio-temporal grounding of multiple objects simultaneously. These are distinct tasks with different objectives and architectures.

4. Context-enhanced video moment retrieval with large language models

URL: View paper

Brief Assessment

Context-enhanced Moment Retrieval[61] focuses on video moment retrieval with language queries using LLM-based context enhancement, not on spatio-temporal omni-object video grounding with multiple target localization as in OmniTube.

5. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection

URL: View paper

Brief Assessment

Umt[60] focuses on joint video moment retrieval and highlight detection with multi-modal (visual-audio-text) transformers, not on spatio-temporal object grounding with bounding box tubes as in OmniTube.

6. SA-DETR: Span Aware Detection Transformer for Moment Retrieval

URL: View paper

Brief Assessment

SA-DETR[66] addresses moment retrieval (temporal localization of video segments based on text), not spatio-temporal video grounding with simultaneous spatial and temporal localization of multiple objects. The tasks and technical approaches differ fundamentally.

7. Video graph transformer for video question answering

URL: View paper

Brief Assessment

Video Graph Transformer[67] addresses video question answering with graph-based spatio-temporal reasoning, not spatio-temporal video grounding with text queries for object localization.

---

8. Momentum cross-modal contrastive learning for video moment retrieval

URL: View paper

Brief Assessment

Momentum Cross-modal Contrastive[68] focuses on video moment retrieval (temporal localization of query-matching segments), not spatio-temporal video grounding with simultaneous spatial object localization. The tasks and technical approaches are fundamentally different.

---

9. Context-Guided Spatio-Temporal Video Grounding

URL: View paper

Brief Assessment

Context-Guided Grounding[64] proposes a context-guided approach for single-object STVG using instance context generation and refinement modules, not a method for simultaneously localizing multiple objects mentioned in queries as in OmniSTVG.

---

10. Temporal refinement and multi-grained matching for moment retrieval and highlight detection

URL: View paper

Brief Assessment

Temporal Refinement Matching[62] focuses on moment retrieval and highlight detection tasks with temporal refinement and multi-grained matching mechanisms, not on spatio-temporal omni-object video grounding with simultaneous localization of multiple objects mentioned in textual queries.

Appendix: Text Similarity Detection

Textual similarity detection checked 30 papers and found 2 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Context-Guided Spatio-Temporal Video Grounding

Detected in: Contribution: contribution_2, Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] OmniSTVG: Toward Spatio-Temporal Omni-Object Video Grounding View paper
- [1] Colo-cam: Class activation mapping for object co-localization in weakly-labeled unconstrained videos View paper
- [2] Temporal-Spatial Feature Interaction Network for Multi-Drone Multi-Object Tracking View paper
- [3] Tcam: Temporal class activation maps for object localization in weakly-labeled unconstrained videos View paper
- [4] SVAG-Bench: A Large-Scale Benchmark for Multi-Instance Spatio-temporal Video Action Grounding View paper
- [5] Universal Video Temporal Grounding with Generative Multi-modal Large Language Models View paper
- [6] Memory Enhanced Global-Local Aggregation for Video Object Detection View paper
- [7] Interacted object grounding in spatio-temporal human-object interactions View paper
- [8] Described spatial-temporal video detection View paper
- [9] QDETRv: query-guided DETR for one-shot object localization in videos View paper
- [10] Videos as space-time region graphs View paper
- [11] General object foundation model for images and videos at scale View paper
- [12] Sketch-based video object localization View paper
- [13] DORi: Discovering Object Relationships for Moment Localization of a Natural Language Query in a Video View paper
- [14] Egocentric audio-visual object localization View paper
- [15] Crab: A unified audio-visual scene understanding model with explicit cooperation View paper
- [16] Deep Video Inpainting Localization Using Spatial and Temporal Traces View paper
- [17] Reciprocal Transformations for Unsupervised Video Object Segmentation View paper
- [18] Video owl-vit: Temporally-consistent open-world localization in video View paper
- [19] Unloc: A unified framework for video localization tasks View paper
- [20] Gaussian Mixture Proposals with Pull-Push Learning Scheme to Capture Diverse Events for Weakly Supervised Temporal Video Grounding View paper
- [21] Localizing spatially and temporally objects and actions in videos View paper
- [22] Efficient video object co-localization with co-saliency activated tracklets View paper
- [23] Where Does It Exist: Spatio-Temporal Video Grounding for Multi-Form Sentences View paper
- [24] Objects2action: Classifying and localizing actions without any video example View paper
- [25] LoVoRA: Text-guided and Mask-free Video Object Removal and Addition with Learnable Object-aware Localization View paper
- [26] Grounding Language in Images and Videos View paper
- [27] Leveraging transformers for weakly supervised object localization in unconstrained videos View paper
- [28] Zero-Shot Video Grounding With Pseudo Query Lookup and Verification View paper
- [29] Joint audio-video object localization and tracking View paper
- [30] Grounding-MD: Grounded Video-language Pre-training for Open-World Moment Detection View paper
- [31] GMIS: Marrying Grounding-DINO and Motion Iterative Segment Anything Model for Referring Video Object Segmentation View paper
- [32] 1+ 1> 2: Detector-Empowered Video Large Language Model for Spatio-Temporal Grounding and Reasoning View paper
- [33] Weakly-supervised video object grounding by exploring spatio-temporal contexts View paper

- [34] VideoRefer Suite: Advancing Spatial-Temporal Object Understanding with Video LLM View paper
- [35] Object-centric Video Question Answering with Visual Grounding and Referring View paper
- [36] Video-GroundingDINO: Towards Open-Vocabulary Spatio-Temporal Video Grounding View paper
- [37] Know-Show: Benchmarking Video-Language Models on Spatio-Temporal Grounded Reasoning View paper
- [38] Unsupervised object discovery and tracking in video collections View paper
- [39] DAVID-XR1: Detecting AI-Generated Videos with Explainable Reasoning View paper
- [40] Diversifying Spatial-Temporal Perception for Video Domain Generalization View paper
- [41] LOVO: Efficient Complex Object Query in Large-Scale Video Datasets View paper
- [42] Grounded Objects and Interactions for Video Captioning View paper
- [43] ProTÃ©GÃ©: Untrimmed Pretraining for Video Temporal Grounding by Video Temporal Grounding View paper
- [44] Panoptic Video Scene Graph Generation View paper
- [45] A particle filtering approach to salient video object localization View paper
- [46] Video Object Grounding using Semantic Roles in Language Description View paper
- [47] Video Scene Graph Generation with Spatial-Temporal Knowledge View paper
- [48] Structured Video-Language Modeling with Temporal Grouping and Spatial Grounding View paper
- [49] Learning Temporal Evolution of Spatial Dependence with Generalized Spatiotemporal Gaussian Process Models View paper
- [50] Temporal-attentive Covariance Pooling Networks for Video Recognition View paper
- [51] Stpro: Spatial and temporal progressive learning for weakly supervised spatio-temporal grounding View paper
- [52] Unleashing the Potential of Multimodal LLMs for Zero-Shot Spatio-Temporal Video Grounding View paper
- [53] Knowing Your Target: Target-Aware Transformer Makes Better Spatio-Temporal Video Grounding View paper
- [54] WINNER: Weakly-supervised hIerarchical decompositioN and aligNment for spatio-tEmporal video gRounding View paper
- [55] VoCap: Video Object Captioning and Segmentation from Any Prompt View paper
- [56] TubeDETR: Spatio-Temporal Video Grounding with Transformers View paper
- [57] VideoGLaMM : A Large Multimodal Model for Pixel-Level Visual Grounding in Videos View paper
- [58] Zero-shot Natural Language Video Localization View paper
- [59] Query-dependent video representation for moment retrieval and highlight detection View paper
- [60] Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection View paper
- [61] Context-enhanced video moment retrieval with large language models View paper
- [62] Temporal refinement and multi-grained matching for moment retrieval and highlight detection View paper
- [63] LITA: Language Instructed Temporal-Localization Assistant View paper
- [64] Context-Guided Spatio-Temporal Video Grounding View paper
- [65] Learning Feature Semantic Matching for Spatio-Temporal Video Grounding View paper
- [66] SA-DETR: Span Aware Detection Transformer for Moment Retrieval View paper
- [67] Video graph transformer for video question answering View paper
- [68] Momentum cross-modal contrastive learning for video moment retrieval View paper
- [69] A survey on temporal sentence grounding in videos View paper
- [70] EgoThinker: Unveiling Egocentric Reasoning with Spatio-Temporal CoT View paper
- [71] Human-centric spatio-temporal video grounding with visual transformers View paper
- [72] Fine-grained spatiotemporal grounding on egocentric videos View paper
- [73] VideoITG: Multimodal Video Understanding with Instructed Temporal Grounding View paper
- [74] Scene-text grounding for text-based video question answering View paper
- [75] Tvqa+: Spatio-temporal grounding for video question answering View paper
- [76] A survey on video temporal grounding with multimodal large language model View paper