A MULTI-DOMAIN ADAPTATION BENCHMARK FOR VIDEO GROUNDING IN VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language Models (VLMs) have shown strong performance in Spatio-Temporal Video Grounding (STVG), yet they are still evaluated mostly in a zero-shot manner on general-purpose benchmarks of everyday scenes. This creates a critical disconnect from real-world applications in specialized domains, where models inevitably encounter rare visual or textual concepts. Since exhaustive pre-training across infinite data distributions is infeasible, the ability to adapt to novel domains with limited data is essential. To bridge this gap, we introduce AnyGroundBench, a domain-adaptation benchmark designed to shift the STVG evaluation paradigm from static zero-shot testing to rigorous domain adaptation. Targeting five specialized domains (animal, industry, sports, surgery, and public security), AnyGroundBench pairs newly captured, expert-annotated videos with established datasets, unifying them through dense, high-fidelity spatio-temporal annotations. Crucially, the benchmark provides dedicated limited training subsets, enabling systematic evaluation of domain adaptability under limited training data. We benchmark 23 state-of-the-art VLMs in the zero-shot setting and further evaluate five adaptation strategies, spanning training-free and fine-tuning-based approaches, on representative models, assessing their zero-shot generalization and adaptation capacity. Our results show that current VLMs remain far from practical performance in the zero-shot setting, that training-free adaptation produces highly variable effects depending on the model and domain, and that fine-tuning-based adaptation, though more effective, still falls short of real-world requirements, with gains varying markedly across domains. These findings expose fundamental limitations in current VLMs' spatio-temporal reasoning, pointing to concrete directions for future research. The code and dataset are available at our Anonymous GitHub and Hugging Face.

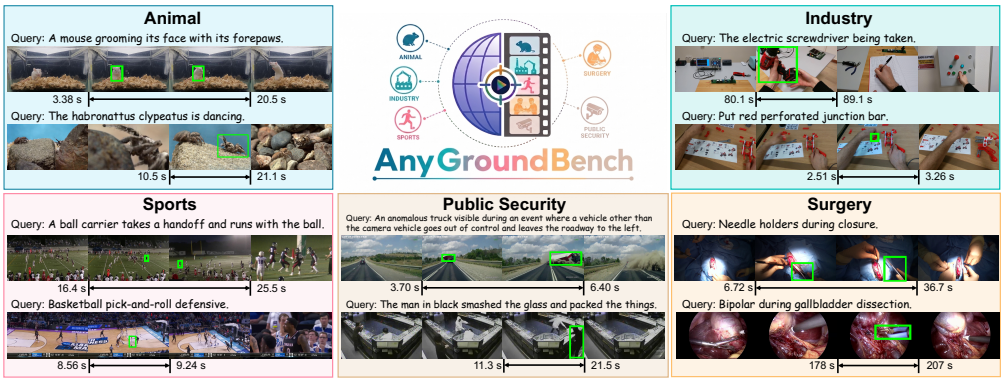


Figure 1: **AnyGroundBench examples across five specialized domains.** AnyGroundBench integrates newly captured, expert-annotated videos with established public datasets, unifying them through dense, new high-fidelity spatio-temporal annotations and language queries.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Spatio-Temporal Video Grounding (STVG) is a fundamental task in the video perception field that requires localizing a target object across both spatial and temporal dimensions in the video based on natural language queries. Its importance is underscored by its role in high-stakes applications, such as video retrieval (Gu et al., 2024), video reasoning (Cheng et al., 2025; Meng et al., 2025), and autonomous driving (Zeng et al., 2025). Recently, the field has transitioned from specialized model architectures (Su et al., 2021; Tang et al., 2022; Wasim et al., 2024; Zhang et al., 2020; Kim et al., 2023; Gu et al., 2025a; 2026) toward general Vision-Language Models (VLMs) (Ahmad et al., 2025; Gu et al., 2025b; Li et al., 2025; 2024; Team et al., 2026; Wang et al., 2026; Zhang et al., 2026; Heo et al., 2025; Pramanick, Shraman and Mavroudi, Effrosyni and Song, Yale and Chellappa, Rama and Torresani, Lorenzo and Afouras, Triantafyllos, 2025). By incorporating the grounding capability in a single unified model, VLMs can extend this capability to tasks that require both perception and reasoning, such as grounded video question answering (Xiao et al., 2024; Shen et al., 2025; Yan et al., 2025), grounded reasoning (Meng et al., 2025), and agentic tool-calling (Li et al., 2026).

Despite recent progress, most established benchmarks (Chen et al., 2019; Gao et al., 2026; Kurita et al., 2023; Li et al., 2025; Liang et al., 2025; Shi et al., 2026; Tang et al., 2022; Xu et al., 2025; Yamaguchi et al., 2017; Yao et al., 2026; Zhang et al., 2020) remain predominantly focused on everyday scenarios and general objects, as summarized in Table 1. However, real-world deployments inevitably extend beyond such settings. For example, the ability to spatio-temporally localize specific mouse behaviors, such as scratching, is crucial for dermatology or neurology research (Segalin et al., 2021). In such highly specialized domains, both the visual appearance and the textual representation deviate significantly from those found in everyday activities. This naturally raises a fundamental question: *Can existing VLMs effectively adapt to ground anything in these rare domains?*

Addressing this question is hindered not only by the lack of benchmarks for specialized domains but also by evaluation protocols, which rely excessively on zero-shot performance (Yang et al., 2025). While VLMs have demonstrated remarkable foundational power, it is practically infeasible for any model to *pre-learn* the near-infinite variety of data distributions encountered, or to be encountered in the future, in the wild. As novel domains and tasks continuously emerge in the real world, such as the aforementioned laboratory or industrial settings (Ragusa et al., 2023; 2024), a model's capability cannot be measured solely by its static pre-trained knowledge. Instead, the ability to seamlessly adapt to a rare domain becomes a strict requirement for real-world deployment. Therefore, to understand whether a model can perceive data from specialized domains, the evaluation paradigm must shift from testing zero-shot performance to measuring adaptability. Moreover, due to the inherent difficulty of collecting data in specialized domains, adaptation should be conducted under *limited-data* conditions.

To bridge these two crucial gaps in domain coverage and adaptability evaluation, we introduce **AnyGroundBench**, a domain-adaptation benchmark that shifts the evaluation paradigm for STVG from general zero-shot evaluation to adaptation to specialized domains, as illustrated in Figure 1. AnyGroundBench comprises 2,042 videos, natural language queries, and corresponding spatio-temporal bounding boxes. AnyGroundBench offers two key advantages. (1) **Diverse and High-Quality Data in Specialized Domains.** AnyGroundBench targets five specialized domains: animal, industry, sports, surgery, and public security. For data sourcing, we aggregate videos from entirely novel captures (including American football from the sports domain and medical expert-curated mouse scratching from the animal domain) alongside new annotations added to existing public datasets. To ensure high-fidelity annotations, we manually provide dense spatio-temporal bounding boxes for ungrounded videos and transform existing labels for the STVG task if applicable. (2) **Dedicated Training Sets for Domain Adaptation.** Unlike previous benchmarks that rely exclusively on general-domain training, AnyGroundBench provides tailored training subsets for each domain. This design explicitly enables researchers to systematically evaluate VLMs' limited-data adaptability: how effectively they adjust to the unique spatio-temporal dynamics inherent to specialized domains.

We evaluate 23 VLMs on our benchmark, including both open-source and proprietary models with diverse architectures and model sizes. We further decompose STVG tasks into isolated Spatial Video Grounding (SVG), which focuses on spatial localization within ground-truth time spans, and Temporal Video Grounding (TVG), which targets temporal boundary detection, thereby pinpointing precisely where current spatio-temporal reasoning breaks down. Furthermore, to serve as a reference for future benchmark users, we evaluate a suite of adaptation strategies spanning training-free

Table 1: **Comparison with existing video grounding datasets.** AnyGroundBenchis the only dataset that spans multiple specialized domains while providing a training set for spatio-temporal grounding adaptation. Public Dataset, Internet-Collected, Self-Collected.

Dataset	Domain	# Videos	Total Duration	# Queries	Data Source	SVG	TVG	STVG	Train Set
General Domain									
TACoS (Regneri et al., 2013)	Cooking	0.1K	10.1h	18.2K			✓		✓
Charades-STA (Gao et al., 2017)	Activity	6.7K	57.1h	16.1K			✓		✓
YouCook2 (Zhou et al., 2018)	Cooking	2.0K	176h	13.8K			✓		✓
VidSTG (Zhang et al., 2020)	General	6.9K	69.1h	99.9K		✓	✓	✓	✓
HC-STVG (Tang et al., 2022)	Human Action	5.7K	31.4h	5.7K		✓	✓	✓	✓
VUE-STG (Vidi Team et al., 2025)	General	0.9K	205h	1.6K		✓	✓	✓	
OmniGround (Gao et al., 2026)	General	3.5K	17.6h	3.5K		✓	✓	✓	
Specialized Domain									
TUMTraffic (Zhou et al., 2025b)	Roadside	1.0K		5.7K		✓	✓	✓	✓
UAV-SVG (Zhan & Yuan, 2025)	Aerial / Drone	3.6K	18.7h	17.8K		✓			✓
QV-M^2 (Cao et al., 2025)	Vlog / News	1.3K	55.9h	2.2K			✓		
AnyGroundBench(Ours)	Animal / Industry / Sports Surgery / Public Security	2.0K	18.3h	3.5K		✓	✓	✓	✓

approaches and fine-tuning-based approaches, covering both practical backpropagation-free settings and higher-performing but more resource-intensive settings.

We reveal three critical findings about VLMs' grounding capability in specialized domains:

- **Limited Spatio-Temporal Grounding Capability of Current VLMs:** Even the latest proprietary VLMs (*e.g.*, GPT-6-Astra) fall short of practical STVG performance, while most open-source VLMs remain near zero in most domains, lagging far behind their proprietary counterparts. This highlights the need to improve the STVG capability of VLMs in general, and of open-source VLMs in particular.
- **Bottleneck Shifts from Spatial to Joint Spatio-Temporal Grounding:** For open-source and earlier proprietary models, SVG lags far behind TVG in most domains, making spatial grounding the primary bottleneck. The strongest proprietary models perform well on both SVG and TVG, yet their STVG remains far below both, indicating that jointly aligning spatial and temporal predictions becomes the limiting factor.
- **Limited and Inconsistent Gains from Adaptation:** Training-free adaptation yields inconsistent effects across models and domains. Fine-tuning improves temporal and spatial grounding individually, with gains that vary greatly across domains. However, these improvements lead to only small gains in STVG. This suggests the need for new adaptation methods that robustly improve grounding across domains.

2 Related Work

Benchmark for Spatio-Temporal Video Grounding. Spatio-temporal video grounding (STVG) aims to localize a target object specified with the text query in both the space and time axes. Progress in this field has been heavily driven by large-scale datasets (Chen et al., 2019; Kurita et al., 2023; Liang et al., 2025; Xu et al., 2025; Yamaguchi et al., 2017; Yao et al., 2026; Zhang et al., 2020). Traditionally, benchmarks like VidSTG (Zhang et al., 2020) (built upon VidOR (Shang et al., 2019)) and HC-STVG (Tang et al., 2022) (sourced from AVA (Gu et al., 2018) and YouTube) have served as primary testbeds, focusing mainly on generic objects and human-centric activities. More recently, LLaVA-ST (Li et al., 2025) enhanced VidSTG with LLM-generated text queries, and OmniGround (Gao et al., 2026) was introduced to cover a wider range of object categories by collecting videos from Pexels[1] and RVOS (Ding et al., 2023; Seo et al., 2020).

However, existing benchmarks share a major limitation: owing to their reliance on general internet platforms and existing generic datasets, they remain heavily confined to common daily-life domains. This narrow scope leaves a critical void for evaluating models in highly specialized, real-world scenarios. To bridge this gap, AnyGroundBench explicitly targets specialized fields. Crucially, we differentiate our work by combining *newly captured* videos curated from experts with established domain-specific datasets with unified, high-fidelity annotations. This exclusive combination provides a benchmark for evaluating STVG domain-adaptation capabilities.

[1] https://www.pexels.com

Table 2: **Overview of AnyGroundBench.** Our benchmark emphasizes domain diversity and high-fidelity manual annotations for specialized scenarios. Avg. Dur. is the mean video length per vid. 🎥: Newly captured videos, ↔: Temporal span annotation, ⧉: Spatial box annotation, ✎: Text query annotation.

Domain	# Videos	Avg. Dur. (sec)	# QA Pairs		Additional Efforts			
			Train	Test	🎥	↔	⧉	✎
Animal	286	47	583	157	✓	✓	✓	✓
Industry	254	55	410	169			✓	✓
Sports	557	20	460	163	✓	✓	✓	✓
Surgery	193	79	559	216				✓
Public Security	752	16	558	250			✓	

Spatio-Temporal Video Grounding via VLMs. Recently, VLMs have dominated not only video understanding tasks (Fu et al., 2025; Zhou et al., 2025a) but also video perception tasks, such as STVG (Ahmad et al., 2025; Li et al., 2024; 2025; Team et al., 2026; Wang et al., 2026; Zhang et al., 2026). Yet, current evaluation protocols for these models rely almost exclusively on zero-shot performance on the general-domain benchmarks (Yang et al., 2025). While measuring foundational knowledge is useful, it ignores the reality of real-world deployments. Since it is impossible to pre-train a model on every conceivable domain, practical systems must rapidly adapt to unseen, specialized data. Currently, the field lacks standardized and unified benchmarks designed to explicitly measure domain adaptability (*e.g.*, via In-Context Learning (ICL) (Brown et al., 2020; Dong et al., 2026; Fujii et al., 2026; Kim et al., 2025; Rubin et al., 2022; Xie et al., 2022; Xue et al., 2026; Yu et al., 2024)). AnyGroundBench is the first benchmark to fill both of these critical gaps by providing highly specialized target domains equipped with standardized adaptation training sets.

3 AnyGroundBench

We introduce AnyGroundBench, a benchmark designed to evaluate STVG capability as well as the adaptation across five specialized domains that present distinct challenges and high real-world relevance: animal, industry, sports, surgery, and public security. To evaluate the adaptability, AnyGroundBench provides training subsets for each domain. An overview of the selected domains and statistics are summarized in Table 2. Details are provided in Appendix C.

3.1 Benchmark Tasks

AnyGroundBench evaluates models on three interconnected tasks defined over a video and a corresponding text query in a zero-shot and via adaptation with limited data. The same VLM is used across all three tasks; the task is induced solely by a task-specific system prompt. We first introduce a unified notation to formalize each task.

Notation. Let $V = \{I_t\}_{t=1}^{T}$ denote a video of T frames with $I_t \in \mathbb{R}^{H\times W\times 3}$, and let Q be a natural-language query referring to a target object and its associated event. A per-frame axis-aligned bounding box[2] is written as $b_t \in \mathbb{R}^4$, and a spatio-temporal tube on an interval $[t_s, t_e] \subseteq [1, T]$ as

$$\tau \;=\; \big\{\,(t, b_t)\,\big\}_{t=t_s}^{t_e}. \tag{1}$$

Each test instance is paired with a ground-truth tube $\tau^* = \{(t, b_t^*)\}_{t=t_s^*}^{t_e^*}$. We denote the evaluated VLM by $\mathcal{F}_\theta$ with parameters θ. All three tasks invoke the same $\mathcal{F}_\theta$,

$$\hat{y} \;=\; \mathcal{F}_\theta(V, Q;\, p), \tag{2}$$

where $p \in \{p_{\text{STVG}},\, p_{\text{SVG}},\, p_{\text{TVG}}\}$ is the task-specific system prompt (full templates in Appendix B) and $\hat{y}$ is the output type appropriate to the task. The three tasks correspond to estimating different subsets of τ^* from (V, Q).

[2]Several VLMs (*e.g.*, the Gemini or Qwen series) emit bounding-box predictions only at a discrete subset of timestamps $\{t_l\}_{l=1}^{L} \subseteq [\hat{t}_s, \hat{t}_e]$ rather than at every frame.

Spatio-Temporal Video Grounding (STVG) requires localizing the target object matching the query in both the spatial and temporal axes. Given (V, Q) and the system prompt p_{STVG}, the model predicts the full tube as:

$$\hat{\tau} \;=\; \mathcal{F}_\theta(V, Q;\, p_{\text{STVG}}) \;=\; \big\{\, (t, \hat{b}_t) \,\big\}_{t=\hat{t}_s}^{\hat{t}_e}, \tag{3}$$

i.e., both the temporal interval $[\hat{t}_s, \hat{t}_e]$ and the bounding box $\hat{b}_t$ in every frame inside it.

Spatial Video Grounding (SVG) focuses on spatial localization within a temporally trimmed video, isolating the spatial component of grounding from temporal-boundary estimation. Given the trimmed video $V_{[t_s^*, t_e^*]} = \{I_t\}_{t=t_s^*}^{t_e^*}$, the query Q, and the system prompt p_{SVG}, the model predicts a per-frame bounding-box sequence as:

$$\big\{\hat{b}_t\big\}_{t=t_s^*}^{t_e^*} \;=\; \mathcal{F}_\theta\big(V_{[t_s^*, t_e^*]},\, Q;\, p_{\text{SVG}}\big)\,. \tag{4}$$

Temporal Video Grounding (TVG) aims to determine the correct temporal boundaries of the queried event, effectively performing temporal segment localization. Given (V, Q) and the system prompt p_{TVG}, the model predicts only the temporal interval

$$[\hat{t}_s, \hat{t}_e] \;=\; \mathcal{F}_\theta(V, Q;\, p_{\text{TVG}}). \tag{5}$$

3.2 Adaptation Protocol

AnyGroundBench provides per-domain training sets $\mathcal{T}_{\mathcal{D}}^{\text{train}} = \{(V_j, Q_j, \tau_j^*)\}_{j=1}^{K}$ that enable evaluation of *adaptation* alongside zero-shot generalization. Let $\mathcal{F}_\theta$ be a base VLM with parameters θ, and let $\mathcal{A}$ denote any adaptation operator (*e.g.*, PEFTs (Hu et al., 2022; Liu et al., 2024a; Jia et al., 2022), or ICL (Brown et al., 2020; Kim et al., 2025) that produces an adapted predictor g from $\mathcal{F}_\theta$ and $\mathcal{T}_{\mathcal{D}}^{\text{train}}$,

$$g \;=\; \mathcal{A}\big(\mathcal{F}_\theta,\, \mathcal{T}_{\mathcal{D}}^{\text{train}}\big)\,, \qquad g : (V, Q) \mapsto \hat{y}, \tag{6}$$

where $\hat{y} \in \{\hat{\tau},\, \{\hat{b}_t\},\, [\hat{t}_s, \hat{t}_e]\}$ is the task-specific output selected by the system prompt p as in Section 3.1. The benchmark is agnostic to the choice of $\mathcal{A}$: any method that consumes $\mathcal{T}_{\mathcal{D}}^{\text{train}}$ to produce a predictor for $\mathcal{T}_{\mathcal{D}}^{\text{test}}$ (whether by augmenting the input with retrieved demonstrations, updating a subset of θ, or full fine-tuning) is directly comparable under AnyGroundBench's metrics.

3.3 Domain and Data Source

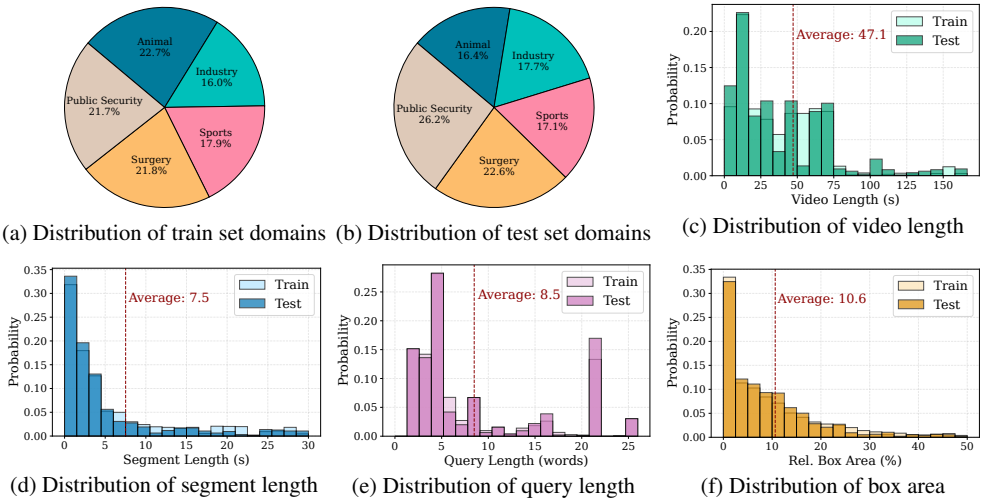
Animal. To advance the understanding of non-human behaviors, we include Animal Kingdom (Ng et al., 2022), capturing diverse species in natural environments. We complement this macroscopic view with a curated Mouse Scratching dataset, collected in a university medical department. This dataset focuses on fine-grained, high-frequency motions (*e.g.*, distinguishing rapid scratches from continuous bouts), presenting challenges for precise spatio-temporal localization in clinical contexts.

Industry. Industrial workflows demand precise, multi-step human-object interactions. We use two egocentric datasets, MECCANO (Ragusa et al., 2023) and ENIGMA-51 (Ragusa et al., 2024), to capture professional activities like mechanical assembly and electrical repair using specialized tools.

Sports. The high-velocity sports domain draws on MultiSports (Li et al., 2021), covering disciplines like basketball and gymnastics, where fine-grained coordination is central. To elevate complexity, we introduce a novel American Football dataset capturing highly tactical, domain-specific plays.

Surgery. The surgical domain requires extreme precision and strict procedural adherence. We encompass two fundamental modalities: EgoSurgery (Fujii et al., 2022; 2024a;b) for direct manual interactions in egocentric open surgery, and CholecTrack20 (Nwoye et al., 2025) for instrument-mediated interventions in laparoscopic procedures. This combination covers the full spectrum of modern surgical practice, from physical manipulation to technology-assisted techniques.

Public Security. The public security domain focuses on critical incidents in public and transit environments. We utilize UCA (Yuan et al., 2024) to provide an exocentric, multimodal perspective on urban surveillance, linking visual anomalies with semantic interpretations. This is complemented by DoTA (Yao et al., 2023), which addresses highly dynamic traffic accidents from an egocentric driving perspective. Together, these benchmarks span the diverse challenges of security and risk monitoring, bridging fixed urban observation and mobile accident perception.



(a) Distribution of train set domains (b) Distribution of test set domains (c) Distribution of video length

(d) Distribution of segment length (e) Distribution of query length (f) Distribution of box area

Figure 2: **Representative statistics on AnyGroundBench**, including distributions of training set domains in (a), test set domains in (b), video length (in seconds) in (c), temporal segment length (in seconds) in (d), textual query length (in words) in (e), and box area in (f).

3.4 Data Annotation

Annotators. To ensure high-quality annotations, we assembled a diverse team tailored to the specific requirements of each dataset. Annotation and data processing for the existing datasets were conducted by graduate students specializing in computer vision. For domains requiring highly specialized knowledge, we engaged domain experts. Specifically, the Mouse Scratching dataset was annotated by *two medical experts to guarantee the clinical fidelity of the captured behaviors*. The newly captured American Football dataset was annotated by *individuals with years of active playing experience*, ensuring the accurate identification and localization of complex, domain-specific actions.

Annotation Process and Quality Control. Our annotation pipeline utilizes advanced vision foundation models with rigorous human verification to ensure high-fidelity spatio-temporal grounding. The annotation process consists of the following key stages. (1) *Textual Query and Corresponding Time Span*: For datasets equipped with existing temporal grounding annotations, we directly utilize their provided text queries. For datasets featuring other types of annotations, such as action category or anomaly localization, we manually paraphrase or annotate textual queries based on their respective labels. For datasets without grounding annotations, annotators provide the temporal span and corresponding textual query. (2) *Spatial Bounding Boxes*: We take a detection-then-tracking approach to label the bounding boxes within the annotated time spans in a fully automated, semi-automated, and manual manner, based on the zero-shot grounding capability of the open-vocabulary object detection (Grounding DINO (Liu et al., 2024b)) and tracking (SAM2 (Ravi et al., 2025)) models for each data source. The fully automated approach applies Grounding DINO to the first annotated frame with the text query input and tracks the box until the annotated end frame. In the semi-automated approach, the human annotator manually labels the corresponding box to the first frame and applies tracking. In the manual approach, the human annotators manually label the bounding box per frame. For datasets that already possess dense spatial tracking annotations, we directly adapt their existing annotations to fit our unified format. (3) *Manual Refinement*: All generated spatio-temporal boxes undergo a comprehensive manual inspection. Human annotators visually review the bounding boxes and manually correct any inaccurate bounding boxes, tracking drifts, or missing frames. (4) *Quality Control*: As a final quality control measure, refined samples are independently double-checked by at least one additional annotator, ensuring that all corrections are accurate and guarantee high reliability, annotation consistency, and fidelity in the benchmark. The details are in Appendix C.

3.5 Benchmark Statistics

We display representative statistics in Figure 2. From Figure 2 (a) and (b), the 3,525 queries are evenly distributed across five specialized domains, providing a balanced testbed for domain adaptation.

Table 3: **Zero-shot results on Spatio-Temporal (STVG), Temporal (TVG), and Spatial Video Grounding (SVG) tasks.** Each cell reports STVG / TVG / SVG, where STVG uses vIoU@0.3, TVG uses tIoU@0.3, and SVG uses sIoU@0.3. **Bold** and underlined fonts represent the best and second-best results, respectively.

Models	Animal	Industry	Sports	Surgery	Public Security
Proprietary VLMs					
GPT-4o	0.00 / 16.6 / 7.64	2.96 / 8.88 / 14.8	0.61 / 17.2 / 0.00	0.00 / 13.4 / 0.00	0.80 / 56.0 / 4.40
GPT-5.1	3.18 / 19.7 / 55.4	5.33 / 9.47 / 28.4	0.00 / 25.8 / 1.84	1.39 / 22.7 / 0.67	4.80 / 61.2 / 35.6
GPT-5.4	6.37 / 21.7 / 65.6	4.73 / 12.4 / 33.1	0.61 / 32.5 / 5.52	1.39 / 38.9 / 2.68	2.00 / 68.0 / 53.6
GPT-5.6-Sol	15.3 / 25.5 / 92.4	**10.7** / 24.3 / 47.9	**6.13** / 44.2 / 45.4	7.87 / 32.4 / 27.5	20.8 / 67.2 / 88.4
GPT-5.6-Terra	13.4 / 22.3 / 86.0	6.51 / 17.2 / 43.8	4.91 / 31.9 / 31.3	4.17 / 32.4 / 10.1	18.4 / 56.8 / 83.2
GPT-6-Astra	**19.1** / 36.3 / **97.5**	7.10 / 24.9 / **53.8**	4.91 / **46.0** / **58.3**	**25.9** / **47.7** / **61.7**	18.8 / **76.0** / **93.6**
Gemini-2.5-Flash	2.55 / 26.1 / 15.3	0.59 / 21.3 / 4.73	1.84 / 31.9 / 3.68	0.00 / 22.7 / 0.67	0.40 / 51.2 / 2.00
Gemini-2.5-Pro	8.28 / 36.9 / 20.4	1.18 / **39.6** / 17.2	0.61 / 37.4 / 8.59	1.39 / 31.5 / 3.36	4.00 / 65.9 / 22.8
Gemini-3-Flash	14.6 / 36.3 / 52.2	5.92 / 30.8 / 20.1	2.45 / 29.4 / 2.45	0.93 / 38.0 / 13.4	11.2 / 66.8 / 45.6
Gemini-3.1-Pro	16.6 / 37.6 / 70.7	7.69 / 21.9 / 41.4	1.23 / 26.4 / 16.6	4.17 / 32.9 / 34.2	22.8 / 69.5 / 52.0
Gemini-3.5-Flash	16.6 / **41.4** / 66.9	5.92 / 19.5 / 47.3	4.91 / 30.1 / 15.3	4.17 / 30.1 / 29.5	22.4 / 61.6 / 52.0
Gemini-3.8-Flash	17.2 / 38.9 / 66.2	7.69 / 20.1 / 37.3	4.29 / 36.2 / 12.9	9.26 / 35.6 / 41.6	**28.4** / 75.2 / 56.0
Open-source Specialized VLMs					
LLaVA-ST	12.1 / 19.7 / 53.5	0.00 / 9.47 / 12.4	0.79 / 8.59 / 3.68	0.00 / 12.0 / 0.00	0.80 / 35.6 / 13.2
Open-source General-Purpose VLMs					
Qwen3-VL-4B	5.73 / 25.5 / 42.0	0.00 / 4.14 / 13.0	0.00 / 8.59 / 0.00	0.00 / 13.4 / 0.00	0.40 / 39.2 / 0.00
Qwen3-VL-8B	3.82 / 19.7 / 17.8	0.00 / 10.7 / 7.69	0.00 / 11.7 / 0.00	0.00 / 15.7 / 0.00	0.80 / 46.0 / 3.60
Qwen3.5-4B	2.55 / 30.6 / 13.4	0.00 / 14.8 / 7.69	0.00 / 12.9 / 0.00	0.00 / 28.2 / 1.34	0.40 / 49.2 / 2.00
Qwen3.5-9B	4.46 / 35.0 / 20.4	0.59 / 17.2 / 12.4	0.61 / 14.7 / 0.00	0.00 / 28.2 / 1.34	0.40 / 50.4 / 4.80
Qwen3.8-27B	12.7 / 26.1 / 12.7	5.33 / 18.3 / 28.4	0.00 / 11.0 / 1.23	0.00 / 35.2 / 10.1	13.6 / 56.4 / 2.80
Eagle2.5-8B	0.00 / 24.8 / 1.27	0.00 / 7.69 / 2.37	0.00 / 7.98 / 0.00	0.00 / 15.3 / 0.00	0.00 / 47.6 / 0.40
InternVL3-8B	0.64 / 15.3 / 3.82	0.00 / 5.92 / 1.18	0.00 / 6.79 / 0.00	0.00 / 10.6 / 0.00	0.00 / 4.40 / 0.80
InternVL3-14B	0.64 / 17.8 / 7.64	0.00 / 5.33 / 1.18	0.00 / 7.36 / 0.00	0.00 / 10.6 / 0.00	0.00 / 13.2 / 0.00
InternVL3.5-8B	0.00 / 11.5 / 0.00	0.00 / 2.96 / 1.18	0.00 / 6.13 / 0.00	0.00 / 7.41 / 0.00	0.00 / 3.60 / 0.00
Gemma4-12B	5.10 / 27.4 / 3.82	4.14 / 11.2 / 18.3	0.00 / 9.20 / 0.00	2.31 / 25.0 / 2.01	1.60 / 43.2 / 0.00

From Figure 2 (c) and (d), the average video length per query is 47.13s while the target segment is 7.51s (16% of the video length), requiring precise temporal localization within videos. Furthermore, Figure 2 (f) shows that the average relative box area is merely 10.63%, demonstrating the challenge of AnyGroundBench in grounding small, specialized visual concepts.

4 Experiments

4.1 Experimental Setup

Models. To ensure a comprehensive evaluation, we benchmark a diverse set of VLMs, including *Proprietary models*: the GPT series (GPT-4o (Achiam et al., 2023), GPT-5.1 (Singh et al., 2025), GPT-5.4 (OpenAI, 2026a), GPT-5.6-Sol/Terra (OpenAI, 2026b), and GPT-6-Astra (OpenAI, 2026c)) and the Gemini series (Gemini-2.5-Flash/Pro (Comanici et al., 2025), Gemini-3-Flash (Google DeepMind, 2025), Gemini-3.1-Pro (Google DeepMind, 2026a), Gemini-3.5-Flash (Google DeepMind, 2026b), and Gemini-3.8-Flash (Google DeepMind, 2026c)), *Open-source Specialized VLMs*: LLaVA-ST (Li et al., 2025), and *Open-source General-Purpose VLMs*: the Qwen series (Qwen3-VL-4B/8B (Bai et al., 2025), Qwen3.5-4B/9B (Qwen Team, 2026a), and Qwen3.8-27B (Qwen Team, 2026b)), Eagle2.5-8B (Chen et al., 2026), the InternVL series (InternVL3-8B/14B and InternVL3.5-8B (Wang et al., 2025)), and Gemma4-12B (Gemma Team, 2026). Detailed inference configurations and full prompt templates are provided in Appendix A and Appendix B, respectively.

Adaptation Methods. We investigate five adaptation strategies spanning training-free adaptation (ICL, as detailed in Appendix A.3, and GEPA (Agrawal et al., 2026)) and fine-tuning-based adaptation (SFT via LoRA (Hu et al., 2022) and full fine-tuning, and GRPO (Shao et al., 2024) with LoRA using IoU-based rewards). ICL and GEPA are prompt-level and applicable to all models, so we evaluate them on both Gemini-3.1-Pro and Qwen3-VL-8B, while SFT, and GRPO require parameter access and are therefore evaluated only on Qwen3-VL-8B.

Table 4: **Adaptation results on Spatio-Temporal (STVG), Temporal (TVG), and Spatial Video Grounding (SVG) tasks.** Each cell reports STVG / TVG / SVG, where STVG uses vIoU@0.3, TVG uses tIoU@0.3, and SVG uses sIoU@0.3. For each model, the first row shows the zero-shot baseline, followed by adaptation using the provided per-domain training subsets. Shaded rows denote fine-tuning-based adaptation, while unshaded adaptation rows are training-free. **Purple** denotes performance improvements or ties and **red** denotes degradations relative to the zero-shot baseline.

Models	Animal	Industry	Sports	Surgery	Public Security
Proprietary VLM (Gemini-3.1-Pro)					
Zero-shot	16.6 / 37.6 / 70.7	7.69 / 21.9 / 41.4	1.23 / 26.4 / 16.6	4.17 / 32.9 / 34.2	22.8 / 69.5 / 52.0
+ICL	12.7 / 33.1 / 60.5	11.8 / 27.8 / 39.6	1.23 / 39.9 / 7.36	9.72 / 42.1 / 34.2	22.4 / 77.2 / 41.6
+GEPA	17.2 / 34.4 / 60.5	13.0 / 38.5 / 51.5	2.45 / 39.3 / 12.9	21.8 / 39.8 / 45.6	19.2 / 78.4 / 53.2
Open-source VLM (Qwen3-VL-8B)					
Zero-shot	3.82 / 19.7 / 17.8	0.00 / 10.7 / 7.69	0.00 / 11.7 / 0.00	0.00 / 15.7 / 0.00	0.80 / 46.0 / 3.60
+ICL	3.82 / 28.0 / 20.4	3.55 / 14.8 / 15.4	0.00 / 23.3 / 0.61	0.93 / 34.3 / 13.4	4.00 / 65.6 / 3.20
+GEPA	3.18 / 17.2 / 35.7	0.00 / 11.8 / 11.8	0.00 / 17.8 / 0.61	0.00 / 39.4 / 0.00	0.40 / 48.8 / 4.80
+SFT LoRA	8.92 / 21.0 / 26.1	0.59 / 7.69 / 21.3	0.00 / 19.0 / 1.23	0.46 / 36.1 / 12.8	4.00 / 64.8 / 0.40
+SFT Full Fine-tuning	10.2 / 16.6 / 28.7	2.96 / 7.69 / 24.3	0.61 / 12.9 / 5.52	4.63 / 27.8 / 32.9	8.00 / 71.2 / 16.0
+GRPO LoRA	11.5 / 34.4 / 31.2	0.59 / 18.3 / 25.4	0.61 / 33.1 / 3.07	3.70 / 54.2 / 36.2	7.20 / 83.6 / 2.40

Evaluation Metrics. In accordance with established protocols (Huang et al., 2024; Li et al., 2025; Ren et al., 2024; Yang et al., 2022), we report vIoU@0.3 for STVG, tIoU@0.3 for TVG, and sIoU@0.3 for SVG. Results for additional threshold-based metrics, including vIoU, sIoU, and tIoU at multiple thresholds, are detailed in Appendix G.

4.2 Zero-Shot Results

As shown in Table 3, current VLMs remain far from practical STVG performance in specialized domains. Even the best model in each domain achieves at most 28.4 vIoU@0.3 (public security, Gemini-3.8-Flash), and only 6.13 in sports. Open-source VLMs lag far behind: all except Gemma4-12B and Qwen3.8-27B score below 1.0 STVG outside animal, and scaling helps only partially, as the largest, Qwen3.8-27B, improves in animal, industry, and public security but still scores 0.00 in sports and surgery. These results highlight the need to improve the STVG capability of VLMs, particularly open-source ones. Decomposing STVG reveals where grounding breaks down. For open-source and earlier proprietary models, SVG lags far behind TVG in most domains (*e.g.*, 56.4 TVG vs. 2.80 SVG for Qwen3.8-27B in public security), making spatial localization the primary bottleneck. In contrast, the strongest proprietary models perform well on both sub-tasks, *e.g.*, GPT-6-Astra reaches 76.0 TVG and 93.6 SVG in public security, yet only 18.8 STVG. This indicates that for these models the bottleneck shifts to jointly aligning spatial and temporal predictions.

4.3 Adaptation Results

As shown in Table 4, training-free adaptation yields inconsistent effects across models and domains. For Gemini-3.1-Pro, ICL and GEPA improve TVG in four of five domains but often degrade SVG, and GEPA substantially improves STVG in surgery (4.17→21.8) while degrading it in public security. For Qwen3-VL-8B, ICL improves nearly all metrics, whereas GEPA yields mixed results, and STVG gains remain at most 4.00. Fine-tuning is more effective, particularly GRPO LoRA, whose TVG in surgery and public security surpasses all zero-shot models including GPT-6-Astra. The gains, however, vary greatly across domains: GRPO LoRA raises SVG from 0.00 to 36.2 in surgery but not in public security (3.60→2.40). Moreover, these sub-task gains translate into only small STVG improvements, *e.g.*, only 3.70 STVG in surgery despite the above TVG and SVG. These results call for adaptation methods that robustly improve joint spatio-temporal grounding across domains.

4.4 Additional Analysis

Sensitivity to Temporal and Spatial Scales. We further analyze sensitivity to two fundamental sources of difficulty in STVG: ground-truth event duration and target bounding box size, as shown in Figure 4. For this analysis, we group sampled evaluation examples into comparable temporal-duration

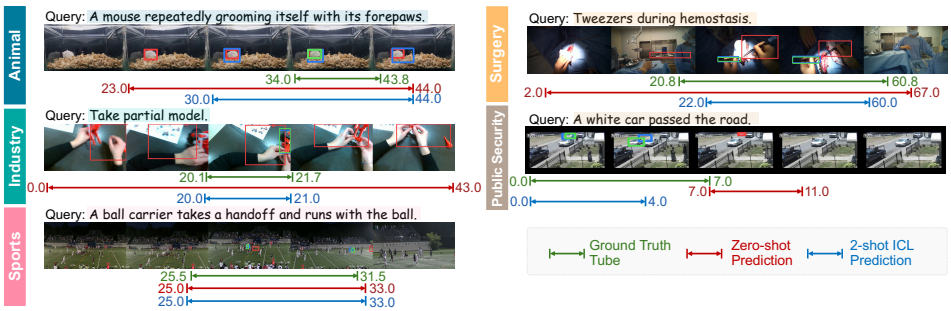


Figure 3: **Qualitative STVG results of Gemini-3.1-Pro across five specialized domains on AnyGroundBench.** Each example compares the ground-truth tube, 2-shot ICL prediction, and zero-shot prediction for the same query. The temporal boundaries are shown in seconds. 2-shot ICL can improve localization on some samples, but the gains are inconsistent, and spatial grounding remains fragile in specialized domains.

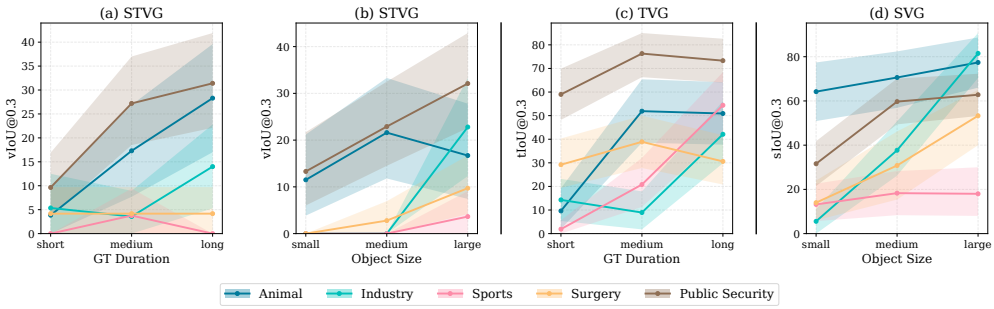


Figure 4: **Sensitivity analysis of temporal and spatial scales.** Within each domain, temporal events are split into terciles by ground-truth duration (short/medium/long), and targets are split into terciles by relative box area (small/medium/large). Error bars denote bootstrap 95% confidence intervals computed per domain. All results use Gemini-3.1-Pro.

and object-size bins. Longer events tend to be easier to ground: both TVG and STVG generally improve as the ground-truth duration increases ((a), (c)). Likewise, larger targets tend to be easier to ground: both SVG and STVG generally improve as the target box size increases ((b), (d)). These trends, however, do not always carry over to STVG, since it requires both the temporal interval and the spatial boxes to be correct. In Sports, for instance, temporal localization improves with event duration, but spatial localization remains poor regardless of duration, so that vIoU@0.3 stays near zero and the duration trend is not reflected in the metric. Additional analyses are in Appendix H.

5 Conclusion

This paper presented **AnyGroundBench**, a domain-specialized benchmark for STVG in VLMs. By spanning five specialized domains and providing both training and evaluation subsets per domain, AnyGroundBench enables evaluating the zero-shot grounding and adaptation capability of VLMs. Experiments with 23 VLMs show that specialized-domain STVG remains far from practical performance. Spatial localization is the main bottleneck for most models, whereas the strongest ones struggle to align spatial and temporal predictions jointly. Neither training-free nor fine-tuning-based adaptation robustly improves STVG across domains.

Broader Impacts. AnyGroundBench provides a standardized benchmark for STVG on five specialized domains, aiming to evaluate how VLMs adapt to specialized domains. By releasing AnyGroundBench with expert-annotated data in a unified format, we expect researchers to develop robust domain-adaptation techniques for video grounding in VLMs under limited training data.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish our writing, including improving phrasing, grammar, and clarity of the manuscript text. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat our dataset, support qualitative and thematic data analysis, or interpret results; these are not applicable to this work. We have reviewed all AI-assisted work: three of the authors independently reviewed and edited the AI-polished text to ensure accuracy, clarity, and faithfulness to the original content and claims. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Eight of the ten components of AnyGroundBench are existing public datasets, for which we release derived annotations only; we do not redistribute their videos, and users must obtain each source from its official provider under the original terms (Appendix E). These sources differ in their handling of personal data: some are de-identified by their providers (*e.g.*, EgoSurgery), while others consist of unmodified surveillance, dashcam, or broadcast footage. We rely on the de-identification, consent, and access arrangements established by each provider and do not alter them. The surgical sources (EgoSurgery, CholecTrack20) additionally constitute medical data. We use all of these as secondary data under each provider's license and data-use agreement, which we confirmed permits third-party re-annotation and the release of derived annotations. We did not attempt to re-identify or link any individual, and collected no demographic or clinical attributes.

Two datasets were collected by us. The Mouse Scratching recordings were obtained observationally from animals already enrolled in an ongoing study at the authors' university school of medicine, approved by its institutional animal experimentation committee and conducted under the institution's animal experimentation regulations. No additional procedure or animal was introduced for this benchmark. The American Football recordings were filmed at public university-level games, with approval subsequently obtained from the host university. Both are released under CC BY-NC-SA 4.0.

One aspect of our own processing warrants note. In the public-security subsets we add bounding boxes to people appearing in surveillance and dashcam footage; a box marks the primary object referred to by the query and does not assert that any depicted individual committed an offence. These sources contain violent incidents and real collisions, which we retain as they constitute the events being evaluated.

AnyGroundBench supports research on spatio-temporal grounding and domain adaptation; it is not evidence of clinical, operational, or forensic validity. Accuracy here does not indicate readiness for intraoperative use, deployed surveillance, judgements about specific individuals, or validated behavioural scoring.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we release our code via an Anonymous GitHub repository and our benchmark data via an anonymized Hugging Face dataset repository. The code includes our data processing pipeline, implementations of the five adaptation strategies evaluated in this work, and the evaluation scripts used to produce the reported results. Machine-readable dataset metadata in the Croissant format is available from the Hugging Face dataset repository. Details of the model configurations, adaptation strategies, and evaluation protocols are provided in the main text and appendix.

REFERENCES

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.

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. GEPA: Reflective Prompt Evolution Can Outperform Reinforcement Learning. In *ICLR*, 2026.

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.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language Models are Few-Shot Learners . *NeurIPS*, 2020.

Zhuo Cao, Heming Du, Bingqing Zhang, Xin Yu, Xue Li, and Sen Wang. When One Moment Isn't Enough: Multi-Moment Retrieval with Cross-Moment Interactions. In *NeurIPS*, 2025.

Guo Chen, Zhiqi Li, Shihao Wang, Jindong Jiang, Yicheng Liu, Lidong Lu, De-An Huang, Wonmin Byeon, Matthieu Le, Max Ehrlich, Tong Lu, Limin Wang, Bryan Catanzaro, Jan Kautz, Andrew Tao, Zhiding Yu, and Guilin Liu. Eagle 2.5: Boosting Long-Context Post-Training for Frontier Vision-Language Models. In *NeurIPS*, 2026.

Zhenfang Chen, Lin Ma, Wenhan Luo, and Kwan-Yee Kenneth Wong. Weakly-Supervised Spatio-Temporally Grounding Natural Sentence in Video. In *ACL*, 2019.

Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. V-STaR: Benchmarking Video-LLMs on Video Spatio-Temporal Reasoning. *arXiv preprint arXiv:2503.11495*, 2025.

Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20:37 – 46, 1960.

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.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. MeViS: A Large-scale Benchmark for Video Segmentation with Motion Expressions. In *ICCV*, 2023.

Yuhao Dong, Shulin Tian, Shuai Liu, Shuangrui Ding, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Jiaqi Wang, and Ziwei Liu. Demo-ICL: In-Context Learning for Procedural Video Knowledge Acquisition. *arXiv preprint arXiv:2602.08439*, 2026.

Alvan R. Feinstein and Domenic V. Cicchetti. High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549, 1990.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. In *CVPR*, 2025.

Ryo Fujii, Ryo Hachiuma, Hiroki Kajita, and Hideo Saito. Surgical Tool Detection in Open Surgery Videos. *Applied Sciences*, 2022.

Ryo Fujii, Masashi Hatano, Hideo Saito, and Hiroki Kajita. EgoSurgery-Phase: A Dataset of Surgical Phase Recognition from Egocentric Open Surgery Videos. In *MICCAI*, 2024a.

Ryo Fujii, Hideo Saito, and Hiroki Kajita. EgoSurgery-Tool: A Dataset of Surgical Tool and Hand Detection from Egocentric Open Surgery Videos. *arXiv preprint arXiv:2406.03095*, 2024b.

Ryo Fujii, Hideo Saito, and Ryo Hachiuma. VIOLA: Towards Video In-Context Learning with Minimal Annotations. *arXiv preprint arXiv:2601.15549*, 2026.

Hong Gao, Jingyu Wu, Xiangkai Xu, Kangni Xie, Yunchen Zhang, Bin Zhong, Xurui Gao, and Min-Ling Zhang. OmniGround: A Comprehensive Spatio-Temporal Grounding Benchmark for Real-World Complex Scenarios. In *CVPR*, 2026.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal Activity Localization via Language Query. In *ICCV*, 2017.

Gemma Team. Gemma 4 Technical Report. *arXiv preprint arXiv:2607.02770*, 2026.

Google DeepMind. Gemini 3 Flash Model Card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf, 2025.

Google DeepMind. Gemini 3.1 Pro Model Card. https://deepmind.google/models/model-cards/gemini-3-1-pro/, February 2026a.

Google DeepMind. Gemini 3.5 Flash Model Card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-5-Flash-Model-Card.pdf, 2026b.

Google DeepMind. Gemini 3.8 Flash Model Card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-8-Flash-Model-Card.pdf, September 2026c.

Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions. In *CVPR*, 2018.

Xin Gu, Heng Fan, Yan Huang, Tiejian Luo, and Libo Zhang. Context-Guided Spatio-Temporal Video Grounding. In *CVPR*, 2024.

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. In *ICLR*, 2025a.

Xin Gu, Haoji Zhang, Qihang Fan, Jingxuan Niu, Zhipeng Zhang, Libo Zhang, Guang Chen, Fan Chen, Longyin Wen, and Sijie Zhu. Thinking With Bounding Boxes: Enhancing Spatio-Temporal Video Grounding via Reinforcement Fine-Tuning. *arXiv preprint arXiv:2511.21375*, 2025b.

Xin Gu, Bing Fan, Jiali Yao, Zhipeng Zhang, Yan Huang, Cheng Han, Heng Fan, and Libo Zhang. Towards Long-Form Spatio-Temporal Video Grounding. In *ECCV*, 2026.

Kilem L. Gwet. Computing Inter-Rater Reliability and Its Variance in the Presence of High Agreement. *The British journal of mathematical and statistical psychology*, 61 Pt 1:29–48, 2008.

Miran Heo, Min-Hung Chen, De-An Huang, Sifei Liu, Subhashree Radhakrishnan, Seon Joo Kim, Yu-Chiang Frank Wang, and Ryo Hachiuma. Omni-RGPT: Unifying Image and Video Region-level Understanding via Token Marks. In *CVPR*, 2025.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*, 2022.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. VTimeLLM: Empower LLM to Grasp Video Moments. In *CVPR*, 2024.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual Prompt Tuning. In *ECCV*, 2022.

Allan V Kalueff, Adam Michael Stewart, Cai Song, Kent C Berridge, Ann M Graybiel, and John C Fentress. Neurobiology of Rodent Self-Grooming and Its Value for Translational Neuroscience. *Nature Reviews Neuroscience*, 17(1):45–59, 2016.

Dahye Kim, Jungin Park, Jiyoung Lee, Seongheon Park, and Kwanghoon Sohn. Language-Free Training for Zero-Shot Video Grounding. In *WACV*, 2023.

Kangsan Kim, Geon Park, Youngwan Lee, Woongyeong Yeo, and Sung Ju Hwang. VideoICL: Confidence-based Iterative In-context Learning for Out-of-Distribution Video Understanding. In *CVPR*, 2025.

Shuhei Kurita, Naoki Katsura, and Eri Onami. RefEgo: Referring Expression Comprehension Dataset from First-Person Perception of Ego4D. In *ICCV*, 2023.

J. Richard Landis and Gary G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, 1977.

Chenglin Li, Qianglong Chen, Feng Han, Yikun Wang, Xingxi Yin, Yan Gong, Ruilin Li, Yin Zhang, and Jiaqi Wang. VideoThinker: Building Agentic VideoLLMs with LLM-Guided Tool Reasoning, 2026.

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 *CVPR*, 2025.

Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. MultiSports: A Multi-Person Video Dataset of Spatio-Temporally Localized Sports Actions. In *ICCV*, 2021.

Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, YiQing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, Zhida Huang, and Tao Wang. GroundingGPT: Language Enhanced Multi-modal Grounding Model. In *ACL*, 2024.

Shuo Liang, Yiwu Zhong, Zi-Yuan Hu, Yeyao Tao, and Liwei Wang. Fine-grained Spatiotemporal Grounding on Egocentric Videos. In *ICCV*, 2025.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-Decomposed Low-Rank Adaptation. In *ICML*, 2024a.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. In *ECCV*, 2024b.

Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, Anran Wang, Zhiyang Teng, Yujing Wang, and Zhuochen Wang. Open-o3 Video: Grounded Video Reasoning with Explicit Spatio-Temporal Evidence. *arXiv preprint arXiv:2510.20579*, 2025.

Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal Kingdom: A Large and Diverse Dataset for Animal Behavior Understanding. In *CVPR*, 2022.

Chinedu Innocent Nwoye, Kareem Elgohary, Anvita Srinivas, Fauzan Zaid, Joël L. Lavanchy, and Nicolas Padoy. CholecTrack20: A Multi-Perspective Tracking Dataset for Surgical Tools. In *CVPR*, 2025.

OpenAI. GPT-5.4 Thinking System Card. https://deploymentsafety.openai.com/gpt-5-4-thinking/gpt-5-4-thinking.pdf, March 2026a.

OpenAI. GPT-5.6 System Card. https://deploymentsafety.openai.com/gpt-5-6, June 2026b.

OpenAI. GPT-6 Astra System Card. https://deploymentsafety.openai.com/gpt-6-astra, 2026c.

Pramanick, Shraman and Mavroudi, Effrosyni and Song, Yale and Chellappa, Rama and Torresani, Lorenzo and Afouras, Triantafyllos. Enrich and Detect: Video Temporal Grounding with Multimodal LLMs. In *ICCV*, 2025.

Qwen Team. Qwen3.5: Towards Native Multimodal Agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026b. URL https://qwen.ai/blog?id=qwen3.8.

Francesco Ragusa, Antonino Furnari, and Giovanni Maria Farinella. MECCANO: A Multimodal Egocentric Dataset for Humans Behavior Understanding in the Industrial-like Domain . *CVIM*, 235:103764, 2023. ISSN 1077-3142. doi: https://doi.org/10.1016/j.cviu.2023.103764.

Francesco Ragusa, Rosario Leonardi, Michele Mazzamuto, Claudia Bonanno, Rosario Scavo, Antonino Furnari, and Giovanni Maria Farinella. ENIGMA-51: Towards a Fine-Grained Understanding of Human Behavior in Industrial Scenarios. In *WACV*, 2024.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. SAM 2: Segment Anything in Images and Videos. In *ICLR*, 2025.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding Action Descriptions in Videos. *TACL*, 2013.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *EMNLP-IJCNLP*, 2019.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. TimeChat: A Time-sensitive Multimodal Large Language Model for Long Video Understanding. In *CVPR*, 2024.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning To Retrieve Prompts for In-Context Learning. In *NAACL*, 2022.

Cristina Segalin, Jalani Williams, Tomomi Karigo, May Hui, Moriel Zelikowsky, Jennifer J Sun, Pietro Perona, David J Anderson, and Ann Kennedy. The Mouse Action Recognition System (MARS) software pipeline for automated analysis of social behaviors in mice. *eLife*, 10:e63720, 2021.

Seonguk Seo, Joon-Young Lee, and Bohyung Han. URVOS: Unified Referring Video Object Segmentation Network with a Large-Scale Benchmark. In *ECCV*, 2020.

Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating Objects and Relations in User-Generated Videos. In *ICMR*, 2019.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Xiaoqian Shen, Min-Hung Chen, Yu-Chiang Frank Wang, Mohamed Elhoseiny, and Ryo Hachiuma. Zoom-Zero: Reinforced Coarse-to-Fine Video Understanding via Temporal Zoom-in. *arXiv preprint arXiv:2512.14273*, 2025.

Jiapeng Shi, Junke Wang, Zuyao You, Bo He, and Zuxuan Wu. VideoLoom: A Video Large Language Model for Joint Spatial-Temporal Understanding. In *ICML*, 2026.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. OpenAI GPT-5 System Card. *arXiv preprint arXiv:2601.03267*, 2025.

Rui Su, Qian Yu, and Dong Xu. STVGBert: A Visual-Linguistic Transformer Based Framework for Spatio-Temporal Video Grounding. In *ICCV*, 2021.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-World Anomaly Detection in Surveillance Videos. In *CVPR*, 2018.

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. *TCSVT*, 2022.

Vidi Team, Chia-Wen Kuo, Chuang Huang, Dawei Du, Fan Chen, Fanding Lei, Feng Gao, Guang Chen, Haoji Zhang, Haojun Zhao, Jin Liu, Jingjing Zhuge, Lili Fang, Lingxi Zhang, Longyin Wen, Lu Guo, Lu Xu, Lusha Li, Qihang Fan, Rachel Deng, Shaobo Fang, Shu Zhang, Sijie Zhu, Stuart Siew, Weiyan Tao, Wen Zhong, Xiaohui Shen, Xin Gu, Ye Yuan, Yicheng He, Yiming Cui, Zhenfang Chen, Zhihua Wu, and Zuhua Lin. Vidi2.5: Large Multimodal Models for Video Understanding and Creation. *arXiv preprint arXiv:2511.19529*, 2026.

Vidi Team, Celong Liu, Chia-Wen Kuo, Chuang Huang, Dawei Du, Fan Chen, Guang Chen, Haoji Zhang, Haojun Zhao, Lingxi Zhang, Lu Guo, Lusha Li, Longyin Wen, Qihang Fan, Qingyu Chen, Rachel Deng, Sijie Zhu, Stuart Siew, Tong Jin, Weiyan Tao, Wen Zhong, Xiaohui Shen, Xin Gu, Zhenfang Chen, and Zuhua Lin. Vidi2: Large Multimodal Models for Video Understanding and Creation. *arXiv preprint arXiv:2511.19529*, 2025.

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. *AAAI*, 2026.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. InternVideo2: Scaling Foundation Models for Multimodal Video Understanding. In *ECCV*, 2024.

Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. VideoGrounding-DINO: Towards Open-Vocabulary Spatio-Temporal Video Grounding. In *CVPR*, 2024.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can I Trust Your Answer? Visually Grounded Video Question Answering. In *CVPR*, 2024.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An Explanation of In-context Learning as Implicit Bayesian Inference. In *ICLR*, 2022.

Qi'ao Xu, Tianwen Qian, Yuqian Fu, Kailing Li, Yang Jiao, Jiacheng Zhang, Xiaoling Wang, and Liang He. ToG-Bench: Task-Oriented Spatio-Temporal Grounding in Egocentric Videos. *arXiv preprint arXiv:2512.03666*, 2025.

Zihui Xue, Ami Baid, Sangho Kim, Mi Luo, and Kristen Grauman. Personal Visual Context Learning in Large Multimodal Models. *arXiv preprint arXiv:2605.10936*, 2026.

Masataka Yamaguchi, Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Spatio-Temporal Person Retrieval via Natural Language Queries. In *ICCV*, 2017.

Ziang Yan, Xinhao Li, Yinan He, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. VideoChat-R1.5: Visual Test-Time Scaling to Reinforce Multimodal Reasoning by Iterative Perception. *arXiv preprint arXiv:2509.21100*, 2025.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. TubeDETR: Spatio-Temporal Video Grounding with Transformers. In *CVPR*, 2022.

Zaiquan Yang, Yuhao LIU, Gerhard Petrus Hancke, and Rynson W. H. Lau. Unleashing the Potential of Multimodal LLMs for Zero-Shot Spatio-Temporal Video Grounding. In *NeurIPS*, 2025.

Jiali Yao, Xin Gu, Xinran Deng, Mengrui Dai, Bing Fan, Zhipeng Zhang, Yan Huang, Heng Fan, and Libo Zhang. OmniSTVG: Toward Spatio-Temporal Omni-Object Video Grounding. In *ICLR*, 2026.

Yu Yao, Xizi Wang, Mingze Xu, Zelin Pu, Yuchen Wang, Ella Atkins, and David J. Crandall. DoTA: Unsupervised Detection of Traffic Anomaly in Driving Videos. *TPAMI*, 45(1):444–459, 2023.

Keunwoo Peter Yu, Zheyuan Zhang, Fengyuan Hu, Shane Storks, and Joyce Chai. Eliciting In-Context Learning in Vision-Language Models for Videos Through Curated Data Distributional Properties. In *EMNLP*, 2024.

Tongtong Yuan, Xuange Zhang, Kun Liu, Bo Liu, Chen Chen, Jian Jin, and Zhenzhen Jiao. Towards Surveillance Video-and-Language Understanding: New Dataset, Baselines, and Challenges. In *CVPR*, 2024.

Shuang Zeng, Xinyuan Chang, Mengwei Xie, Xinran Liu, Yifan Bai, Zheng Pan, Mu Xu, and Xing Wei. FutureSightDrive: Thinking Visually with Spatio-Temporal CoT for Autonomous Driving. In *NeurIPS*, 2025.

Yang Zhan and Yuan Yuan. Where Does It Exist from the Low-Altitude: Spatial Aerial Video Grounding. In *NeurIPS*, 2025.

Xiaowen Zhang, Zhi Gao, Licheng Jiao, Lingling Li, and Qing Li. STVG-R1: Incentivizing Instance-Level Reasoning and Grounding in Videos via Reinforcement Learning. In *ICLR*, 2026.

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 *CVPR*, 2020.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. MLVU: Benchmarking Multi-task Long Video Understanding. In *CVPR*, 2025a.

Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards Automatic Learning of Procedures from Web Instructional Videos. In *AAAI*, 2018.

Xingcheng Zhou, Konstantinos Larintzakis, Hao Guo, Walter Zimmer, Mingyu Liu, Hu Cao, Jiajie Zhang, Venkatnarayanan Lakshminarasimhan, Leah Strand, and Alois Knoll. TUMTraf VideoQA: Dataset and Benchmark for Unified Spatio-Temporal Video Understanding in Traffic Scenes. In *ICML*, 2025b.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *arXiv preprint arXiv:2504.10479*, 2025.