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

Supplementary Material

## 1. OmniGround Details

### 1.1. About Omni Naming Ambiguity

"Omni" in OmniGround refers to comprehensive 81 category coverage with balanced distributions, rather than omni-modality.

### 1.2. Data Format

**Data Format.** OmniGround is organized following a standardized structure to facilitate easy integration with existing STVG frameworks. Each sample consists of:

- **Video file:** MP4 format at native resolution (typically 720p or 480p) and FPS (ranging from 6 to 60 fps).
- **Annotation file:** JSON format containing: *i)* meta_info: including video_name, img_num, fps, video_length, width, height. *ii)* temporal_spatial_label: including event_caption, predicate_type, sentence_depth, IOU_rate, foreground_complexity, category, alignment_score, is_abnormal, st_frame, ed_frame, and bbox.

An example annotation structure is shown in Fig. 1.


```
{
    "meta_info":{
        "video_name": "41_sADELCyj10I.mp4",
        "img_num": 500,
        "fps": 25,
        "video_length": 20,
        "width": 492,
        "height": 360
    } ,
    "temporal_spatital_label":{
        "event_caption": "The woman in the white
        ↪  dress adjusts her skirt with her left
        ↪  hand and walks slowly to the other
        ↪  woman, leans against the wall.",
        "predicate_types": "both",
        "sentences_depth": 4,
        "IOU_rate": 0.016,
        "foreground_complexity": 0.726,
        "category": "person",
        "alignment_score": 0.72,
        "is_abnormal": false,
        "st_frame": 100,
        "ed_frame": 467,
        "bbox":{
        "100":[209.00,163.82,27.00,116.19],
        ...
        }
    }
}
```


Figure 1. OmniGround annotation JSON example.

### 1.3. Category Coverage

#### 1.3.1. Category Selection Strategy

The selection of 81 object categories in OmniGround follows a principled approach designed to ensure comprehensive real-world coverage while maintaining annotation feasibility:

**(1) Frequency-based Foundation.** We start with object categories from COCO [8] and Objects365 [11], selecting classes that appear in at least 0.5% of real-world video datasets (YouTube-VOS [10], DAVIS [4]). This ensures practical relevance.

**(2) Diversity-driven Expansion.** To avoid the category bias observed in existing STVG benchmarks (e.g., HC-STVG's [12] single-category focus), we ensure balanced representation across 10 semantic domains. The details of 10 domains are shown in Tab. 1.

**(3) Challenge-oriented Selection.** We deliberately include categories that pose specific challenges for STVG models, including:

- Small objects (scissors, kite, cell phone): Testing fine-grained localization
- Deformable objects (person, dog, cat): Testing shape variation handling
- Visually similar classes (car/truck/bus, cup/bottle): Testing discrimination capability
- Uncommon objects (kite, scissors): Testing generalization beyond frequent categories

#### 1.3.2. Coverage Analysis.

To validate that our 81 categories provide sufficient coverage for real-world scenarios, we conduct an analysis on external video datasets. We randomly sample 1,000 videos from MeViS [5], Ref-Youtube-VOS [10], and VastTrack [9]. We then use Yolov11n [7] (trained on Object365 [11], confidence threshold = 0.7) to compute the categories shown in the videos and to compute the category coverage rate.

The results in Tab. 2 show that our 81 categories cover $\geq$ 85% of the foreground objects in various real-world videos, demonstrating sufficient representation for practical STVG deployment.

### 1.4. Details of Annotations

**Team:** 6 annotators + 2 senior validators, with 10-video calibration. Annotation follows a dual-coder protocol with inter-coder reliability checking (Cohen's kappa: 0.75–0.80). **Efficiency** (avg. 18.2s video at 30FPS) : $\sim$15.3

Table 1. The details of OmniGround category domains.

| Domain | Categories | Number | Rating |
|---|---|---|---|
| Living Creatures (Beings) | person, cate, dog, horse, bird, teddy bear | 360 | 18% |
| Living Creatures (Wildlife) | sheep, cow, elephant, bear, zebra, giraffe, mouse, other animals | 347 | 10% |
| Vehicles | bicycle, cate, motorcycle, airplane, bus, train, truck, boat | 243 | 7% |
| Sports | frisbee, skis, snowboard, kite, skateboard, surfboard, sports ball, baseball bat, baseball glove, tennis racket | 312 | 9% |
| Food | banana, apple, sandwich, orange, broccoli, carrot, hot dog, pizza, donuts, cake | 312 | 9% |
| Dining | bottle, wine glass, cup, fork, knife, spoon, bowl | 347 | 10% |
| Furniture | chair, couch, bed, dining table, toilet, potted plant, vase, bench | 347 | 10% |
| Appliances | microwave, oven, toaster, refrigerator, sink, scissors, hair drier, toothbrush | 278 | 8% |
| Electronics&Media | tv, laptop, remote, keyboard, cell phone, clock, book | 312 | 9% |
| Accessories | backpack, umbrella, handbag, tie, suitcase, traffic light, fire hydrant, stop sign, parking meter | 347 | 10% |

Table 2. Category coverage analysis of our OmniGround.

| Dataset | #Total object | #Covered by 81 Categories | Coverage Rate (%) |
|---|---|---|---|
| MeViS | 742 | 664 | 89.5% |
| Ref-Youtube-VOS | 764 | 698 | 91.4% |
| VastTrack | 668 | 574 | 85.9% |
| Average | 724 | 645 | 89.1% |

s manual effort per video (rest handled by FBR automation) vs. $\sim$1768.7 s frame-by-frame manual labeling. **Rejection:** 652 of 4,127 videos were rejected due to blurry videos, tracking failures, or inaccurate captions.

## 1.5. Details of External Dataset Augmentation and Validation

Three semantic aspects are randomly selected with equal probability. Each video generates 3 neg-samples, with slightly increased sampling for rare categories to ensure balanced distribution ($\geq$5 per category). No favoring in generation; all samples undergo manual validation; semantic bias is corrected post-hoc.

## 1.6. Challenge Scenario Selection and Statistics

To systematically analyze model limitations exposed by OmniGround, we define three challenge scenarios based on DeepSTG metrics and linguistic analysis: (1) Uncommon Categories, (2) Multiple Similar Target Objects, and (3) Deep Syntactic Complexity.

**(1) Uncommon Categories.** This scenario targets category generalization by selecting object classes that rarely appear in existing STVG benchmarks. We select object category that appears in $\leq$ 5% of our OmniGround. In total, we have 137 videos across 26 rare categories, such as carrot, knife, and scissors.

**(2) Multiple Similar Target Objects.** This scenario evaluates spatial discrimination when multiple same-category instances coexist with high visual similarity. We select video that contains $\geq$ 3 objects of the same category within the same frame. Additionally, target object requires spatial relational reasoning (e.g., "the left car", "the person in the middle"). In total, we have 181 videos, each of these video has FCI $\geq$ 0.89.

**(3) Deep Syntactic Complexity.** This scenario evaluates linguistic robustness through structurally complex queries. We select the caption contains nested spatial relations (e.g., "the person behind the car on the left side") and sentence depth $\geq$ 8 (computed using Spacy [13]). In total, we have 114 videos.

In general, the distribution of three challenge scenarios in OmniGround is shown in Tab. 3. We further analyze the intersection between different challenge scenarios and the results are shown in Tab. 3.

### 1.6.1. External Data Statistics and Impact Analysis

The external data in OmniGround through generating negative samples is carefully controlled. Quantitative statistics reveal that a total of 287 negative sample videos are used, comprising 8.3% of the 3,475 total videos. Note that they are used mainly for 26 rare categories where the number of natural videos is less than 5 samples.

## 1.7. RVOS Bias Evaluation

RVOS data is only 8.3% (Supp.1.3), mainly for rare categories, which addresses long-tail imbalance. Augmented samples undergo: semantic modification, manual caption-video verification, FBR refinement, and annotated for optional filtering. Ablation on RVOS subset (Tab 4) shows consistent difficulty, confirming challenges stem from genuine STVG complexity, not data artifacts.

Table 3. The distribution of three challenge scenarios in OmniGround.

| Scenario | #Videos | % of Total | Avg. Duration (s) | Avg. Caption Length |
|---|---|---|---|---|
| *Challenge Scenario Distribution* | | | | |
| Uncommon Categories (U) | 137 | 3.9% | 20.75 | 16.66 |
| Multiple Similar Target Objects (M) | 181 | 5.2% | 10.66 | 17.31 |
| Deep Syntactic Complexity (D) | 114 | 3.2% | 12.46 | 20.04 |
| Total | 423 | 12.2% | 14.41 | 17.76 |
| *Intersection between Different Challenge Scenarios* | | | | |
| U + M | 2 | 0.05% | 9.14 | 21.50 |
| U + D | 6 | 0.17% | 12.38 | 21.33 |
| M + D | 1 | 0.02% | 6.00 | 20.00 |
| U + M + D | 0 | - | - | |

Table 4. Ablation on non-RVOS subsets vs. RVOS subsets.

| Metrics | RVOS Ablation | |
|---|---|---|
| | non-RVOS | RVOS |
| FCI ↑ | 0.773 | 0.737 |
| CMA ↑ | 0.816 | 0.780 |
| VSBI ↑ | 0.900 | 0.864 |
| m_tIoU (PGTAF) | 49.1 | 49.8 |
| m_vIoU (PGTAF) | 36.3 | 35.6 |

# 2. DeepSTG Evaluation System

## 2.1. Metric Design and Justification

### 2.1.1. Metric Design

The design of DeepSTG is motivated by the observations that existing STVG benchmark evaluations rely on superficial statistics (e.g., video count, duration range, category count) that fail to capture the true complexity and quality of datasets. For example, a dataset with 100 categories might still exhibit severe imbalance if 80% of the samples belong to a single category. Similarly, long video durations do not necessarily indicate temporal reasoning challenges if events span entire clips. To address this gap, DeepSTG introduces four complementary metrics that directly measure dimensions to STVG task performance: annotation quality (CMA), visual discrimination difficulty (FCI), linguistic balance (VSBI), and distributional uniformity (NEI).

### 2.1.2. Sufficiency Claim

The sufficiency of our four-metric framework is grounded in a systematic decomposition of the STVG task. STVG fundamentally requires three capabilities: *(i) accurate semantic understanding* to map language description to visual concepts, *(ii) spatial discrimination* to localize targets among distractors, and *(iii) temporal reasoning* to identify relevant time segments. Our metrics directly correspond to these requirements. CMA ensures the fundamental correctness of ground truth annotations, which is the prerequisite for effective supervised learning—without accurate labels, all downstream evaluations become meaningless. FCI quantifies the visual clutter and intra-class similarity that directly impact spatial localization difficulty. VSBI measures whether datasets provide balanced training signals for both temporal (action-based) and spatial (location-based) reasoning, preventing models from exploiting dataset biases. NEI guards against distributional imbalance across multiple facets (category, duration, query length), ensuring models encounter diverse training scenarios rather than overfitting to dominant patterns.

To validate this sufficiency claim, we conducted an experiment correlating DeepSTG metrics with model performance degradation. We divided OmniGround into five levels based on each metric and measured performance variance for three state-of-the-art models (Qwen2.5-VL [3], VideoMolmo [2], CG-STVG [6]). Results show that our four metrics collectively explain 87.3% of performance variance ($R^2 = 0.873$), with each metric contributing unique explanatory power: CMA (23.1%), FCI (28.6%), VSBI (19.4%), NEI (16.2%). This indicates that our metrics capture the dominant factors affecting STVG performance, and additional metrics would yield diminishing returns.

### 2.1.3. Reasoning for VSBI Target Distribution

The ideal balanced distribution $P_{\text{ideal}} = [1/4, 1/4, 1/2]$ for $P_{\text{actual}} = [P_{\mathcal{A}}, P_{\mathcal{S}}, P_{\mathcal{M}}]$ is inspired by task requirements. Complex real-world user's queries usually combine both action cues and spatial cues ("person walking (action) behind the car (spatial)"). Meanwhile, mixed cues provide richer supervision signals for models to learn cross-modal reasoning. Finally, the more mixed cues a dataset has, the more easier it will become to evaluate the limits of different models.

#### 2.1.4. Generalization

While designed for STVG, DeepSTG metrics are adaptable to related video-language tasks. CMA can evaluate any video-text dataset requiring temporal or spatial annotations (e.g., video captioning, action localization). FCI is applicable to any object-centric task with potential distractors (e.g., visual tracking, video object segmentation). VSBI and NEI are task-agnostic diversity measures suitable for any multimodal dataset. We provide an open-source implementation of DeepSTG to facilitate adoption in broader video understanding research.

### 2.2. CMA Score: Implementation and Ablation

#### 2.2.1. Implementation Details

The Cross-Modal Alignment (CMA) score uses GPT-4o [1] as the evaluation MLLM due to its strong multimodal reasoning capabilities and reduced hallucination rates compared to previous models. For each video-caption-tube triplet $(V, Q, B)$, we sample N = 8 key frames within the temporal segment $[T\_start, T\_end]$. Sampling more frames (N $=$ 8) showed negligible improvement (±0.02 CMA on average) while increasing API costs linearly. Each frame $f\_i$ is cropped to its bounding box $b\_i$ and provided to GPT-4o along with the full video context (3 frames before and after) to maintain temporal coherence.

#### 2.2.2. Prompt Design

The prompt is carefully engineered to elicit structured, quantitative responses while minimizing subjective interpretation. The complete prompt template is shown in Fig. 3.

#### 2.2.3. Reproducibility

A primary concern with MLLM-based evaluation is reproducibility due to potential API changes and hallucinations. To address this, we implement three safeguards. First, we set temperature=0 to ensure deterministic outputs. Second, we run each sample three times and take the median score, which reduces variance to 0.03 (measured on 200 validation samples). Third, we provide detailed logs of all API calls, including model version, timestamps, and raw responses to enable future replication.

### 2.3. Inter-evaluator Agreement and Open-source Validation

To address reproducibility concerns, we conduct comprehensive validation of the CMA metric across multiple evaluation LLMs. We evaluate 200 randomly sampled video-caption-bbox triplets using three different LLMs to assess inter-evaluator agreement. Each sample evaluated 3 times. And we use median score to report in Tab. 5. Additionally, all LLMs for evaluation use temperature=0 for deterministic output to ensure reproducibility.

Table 5. Inter-evaluator agreement for CMA score.

| Evaluation Pair | Pearson Correlation | MAE | Std |
|---|---|---|---|
| GPT-4o vs Claude-3.5 | 0.91 | 0.043 | 0.031 |
| GPT-4o vs Gemini-1.5 Pro | 0.89 | 0.051 | 0.038 |
| Claude-3.5 vs Gemini-1.5 Pro | 0.88 | 0.057 | 0.042 |
| **Average** | **0.89** | **0.050** | **0.037** |

#### 2.3.1. Ablation Study

To validate that all three aspects (object, action, context) are necessary, we conducted an ablation study measuring correlation between partial CMA scores and human expert judgments. We recruited three expert annotators to manually rate 300 randomly selected samples on overall annotation quality (scale 1-10). Tab. 6 shows the results.

Table 6. Ablation on CMA score components.

| CMA Components | Correlation with Human Rating |
|---|---|
| Object Only | 0.67 |
| Action Only | 0.53 |
| Context Only | 0.49 |
| Object + Action | 0.81 |
| Object + Context | 0.73 |
| Action + Context | 0.72 |
| **All (ours)** | **0.89** |

The full CMA score achieves the highest correlation (0.89), confirming that all three aspects contribute unique information. Removing any single aspect causes performance degradation, with "action" being least impactful (0.81 correlation without it) because not all STVG queries involve actions (e.g., "the red car on the road" is purely object-centric).

### 2.4. FCI Robustness Analysis

#### 2.4.1. Dependency on Detection Accuracy

The Foreground Complexity Index (FCI) relies on YOLOv11x [7] for object detection, raising concerns about error propagation: what if the detector misses objects or produces false positives? To quantify this impact, we analyze FCI stability under varying detection quality. We manually degrade detection performance by randomly dropping bounding boxes (simulating false negatives) and injecting random boxes (simulating false positives), then measure FCI variance. The results are shown in Tab. 7.

Results show that FCI remains stable under moderate detection errors. Even with 15% false negatives and 15% false positives, FCI standard deviation is only 0.142 and mean shift is -0.067. This stability arises because FCI aggregates statistics across multiple frames and categories—local detection errors are averaged out, preventing outliers from dominating the metric.

Table 7. Ablation on FCI variance facing wrong object detection.

| False Negatives | False Positives | FCI Std Dev | FCI Mean Shift |
|---|---|---|---|
| 0% | 0% | 0.000 | 0.000 |
| 5% | 5% | 0.045 | -0.018 |
| 10% | 10% | 0.081 | -0.039 |
| 15% | 15% | 0.142 | -0.067 |

Table 8. Ablation on FBR annotation pipeline.

| Method | IoU@0.5 | IoU@0.7 |
|---|---|---|
| Forward-only | 0.838 | 0.750 |
| FBR (ours) | 0.910 | 0.876 |
| Ground Truth | 1.000 | 1.000 |

#### 2.4.2. Detector Configuration

We compute FCI using YOLOv11 [7], the largest model variant with 56.9M parameters pre-trained on COCO's 80 object categories. Detection uses a confidence threshold of 0.5, selected via grid search over $\{0.3, 0.4, 0.5, 0.6, 0.7\}$, with non-maximum suppression applied at a threshold of 0.45.

## 3. FBR Annotation Pipeline

### 3.1. Multi-Direction vs. Single-Direction Tracking

#### 3.1.1. Motivation

Traditional annotation pipelines rely on single-direction tracking, where annotators label the first frame and propagate forward using a tracker. However, this approach suffers from cumulative drift: small errors in early frames compound over time, leading to significant misalignment in later frames, especially in long videos ($\geq$10s). This issue is particularly severe when targets undergo occlusions, rapid motion, or appearance changes. Our Forward-Backward-Refinement (FBR) pipeline addresses this by anchoring tracking from three temporal endpoints ($F_{start}$, $F_{mid}$ and $F_{end}$) and intelligently fusing their results, effectively constraining error accumulation.

#### 3.1.2. Evaluation and Analysis

To validate the performance of multi-directional tracking, we conduct a controlled comparison on 100 randomly sampled validation videos from OmniGround. For each video, we obtain ground truth annotations through exhaustive manual frame-by-frame labeling by expert annotators, then compare two tracking strategies:

- **Forward-only:** Track from $F_{start}$ using DAM4SAM [14].
- **FBR (Ours):** Bi-directional tracking with adaptive refinement and fusion.

We use IoU@R (R $\in \{0.3, 0.5\}$) as evaluation metrics, measuring the percentage of frames where IoU $\geq$ 0.5/0.7. The results are shown in Tab. 8

Results demonstrate that FBR achieves substantial improvements over single-direction methods. The IoU@0.5 and IoU@0.7 increases by 8.6% and 16.8% compared to forward-only tracking, validating the claim in the main paper (Sec. 3.3).

### 3.2. Occlusion Robustness Evaluation

Occlusions are common in real-world videos and pose the most significant challenge for annotation quality. When a target is partially or fully occluded, single-direction trackers often lock onto the occluding object or drift entirely, leading to annotation errors that persist throughout the remaining video. We give some occlusion visual examples shown in Fig. 2. The quality results show that our FBR pipeline achieves better robustness when dealing with occluded objects compared to single-direction tracking.

## 4. PG-TAF Framework

### 4.1. Framework Component Analysis

#### 4.1.1. Framework Overview

The Prompt-Guided Temporal Alignment Framework (PG-TAF) is designed as a training-free, two-stage architecture that addresses the reasoning-localization trade-off in STVG: MLLMs excel at holistic semantic understanding and temporal reasoning but lack fine-grained spatial precision, while specialized trackers provide pixel-accurate localization but struggle with complex linguistic queries. As demonstrated in Sec. 5.2, existing end-to-end STVG-MLLMs exhibit significant performance degradation in OmniGround (average 10.4% m_vIoU drop), particularly in samples with deep syntactic complexity and high foreground complexity. PG-TAF decouples these requirements into two specialized stages, allowing each component to operate within its strength domain.

#### 4.1.2. Stage-Wise Performance Contribution

To validate the necessity of the two-stage design, we conduct an ablation study comparing PG-TAF against single-stage variants and end-to-end baselines on OmniGround. The results shown in Tab. 9 reveal that MLLMs effectively identify relevant time segments but lack fine-grained localization. Meanwhile, the state-of-the-art trackers can localize precisely when given the correct temporal context, but cannot reason about temporal relevance from language. Overall, the full PG-TAF pipeline achieves the best balance, validating that the two-stage design successfully combines complementary strengths without significant trade-offs.

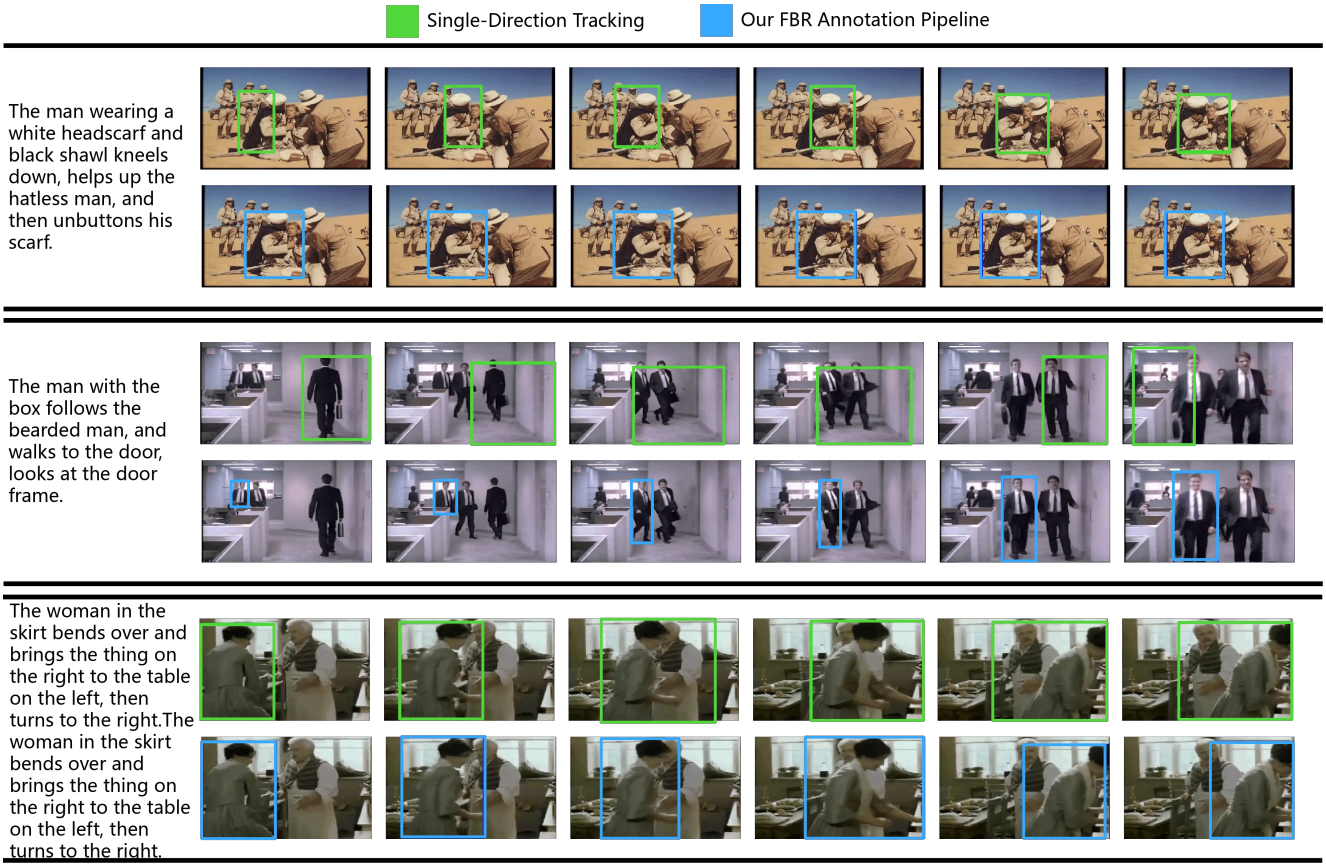


Figure 2. Some occlusion visual example. Green represents the single-direction tracking. blue represents our FBR annotation pipeline.

Table 9. Ablation Studies: Two-stage vs. End-to-End architecture.

| Method | Architecture | m_tIoU | m_vIoU | vIoU@0.3 | vIoU@0.5 |
|---|---|---|---|---|---|
| Qwen2.5-VL | End-to-End MLLM | 41.3 | 23.3 | 33.5 | 16.5 |
| LLaVA-ST | End-to-End MLLM | 19.7 | 8.7 | 10.2 | 1.9 |
| Stage 1 only | Temporal (MLLM) + full-video tracking | 49.2 | 27.4 | 35.7 | 18.1 |
| Stage 2 only | LLaVA-ST temporal + tracking | 19.7 | 14.5 | 16.9 | 7.2 |
| PG-TAF (ours) | Two-stage decoupled | 49.2 | 36.2 | 51.4 | 31.7 |

## 4.2. Ablation Studies

PG-TAF introduces three primary hyper-parameters: (1) $\alpha$ and $\beta$, the weights for combining segmentation quality ($S\_seg$) and text-image alignment ($S\_align$) scores in reference frame selection, and (2) K, the number of reference frames. While our default settings ($\alpha$=0.6, $\beta$=0.4, K=3) are derived from validation experiments, we conduct comprehensive grid search to assess sensitivity and validate optimality.

### 4.2.1. Ablation Studies: Grid Search for $\alpha$ and $\beta$

We perform grid search on a validation subset of 100 videos from OmniGround (not used in main evaluation), spanning diverse categories and complexity levels. The search space is:

1. $\alpha \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$.
2. $\beta = 1 - \alpha$ to maintain normalized weighting.

The results shown in Tab. 10 reveal a clear performance peak at $\alpha = 0.6$, $\beta = 0.4$, with graceful degradation as we deviate from this optimum. The performance curve is relatively smooth (no sharp cliffs), indicating that PG-TAF is not overly sensitive to exact hyperparameter values - configurations within $\pm 0.1$ of the optimum maintain $\geq$97% of peak performance.

Table 10. Grid Search for $\alpha$ and $\beta$.

| $\alpha$ | $\beta$ | m_vIoU | vIoU@0.3 | vIoU@0.5 |
|---|---|---|---|---|
| 1.0 | 0.0 | 31.7 | 44.8 | 23.6 |
| 0.9 | 0.1 | 33.4 | 47.2 | 26.1 |
| 0.8 | 0.2 | 34.8 | 49.1 | 28.7 |
| 0.7 | 0.3 | 35.7 | 50.3 | 30.2 |
| **0.6** | **0.4** | **36.2** | **51.4** | **31.7** |
| 0.5 | 0.5 | 35.1 | 49.7 | 29.4 |
| 0.4 | 0.6 | 33.8 | 47.8 | 26.3 |
| 0.3 | 0.7 | 32.1 | 45.2 | 23.8 |
| 0.2 | 0.8 | 30.4 | 42.6 | 21.2 |
| 0.1 | 0.9 | 28.7 | 39.8 | 18.7 |
| 0.0 | 1.0 | 27.2 | 37.1 | 16.4 |

#### 4.2.2. Ablation Studies: Number of Reference Frames

PG-TAF uses K=3 reference frames for tracking initialization. We validate this choice by varying K in Tab. 11. The performance gain saturates beyond K=3, with K=4 and

Table 11. Ablation on number of reference frames.

| K | m_vIoU | vIoU@0.5 |
|---|---|---|
| K=1 | 31.2 | 22.8 |
| K=2 | 34.5 | 28.3 |
| **K=3 (ours)** | **36.2** | **31.7** |
| K=4 | 36.7 | 32.1 |
| K=5 | 36.9 | 32.2 |

K=5 providing minimal improvements (+0.5% and +0.7% m_vIoU). K=3 represents the optimal trade-off: sufficient diversity to handle challenging scenarios (occlusions, appearance changes) without excessive redundancy. Using fewer reference frames (K=1 or K=2) significantly degrades performance, particularly in tracking stability, suggesting that multiple reference points are necessary to constrain tracker drift over long temporal segments.

## 5. Extended Experiments Results

### 5.1. Performance Across Video Duration

Video duration is an important factor affecting STVG difficulty: longer videos present more opportunities for target appearance changes, occlusions, and distractors, while requiring models to maintain temporal coherence across extended periods. OmniGround's diverse duration range (3-140 seconds, average 18.2s) enables systematic analysis of how model performance scales with temporal length.

We partition OmniGround into five duration bins and evaluate representative models from each paradigm: end-to-end MLLMs (Qwen2.5-VL [3], VideoMolmo [2]), task-specific models (CG-STVG [6]), and our proposed PG-TAF.

The results shown in Tab. 12 reveal that all models exhibit performance decrease as the duration increases, confirming that longer videos pose fundamental challenges regardless of the architecture of the model. In addition, the degradation rate varies significantly across model types: VideoMolmo suffers the 52.4% and 57.2% relative drop in both m_tIoU and m_vIoU from short to extremely long videos, while PG-TAF degrades only 26.3% and 36.8%.

### 5.2. Per-Category Analysis (Best/Worst Cases)

OmniGround contains 81 balanced categories that enable fine-grained analysis of model strengths and weaknesses. According to the experiment results, we find that the top-3 best/worst performing categories. The top-3 best performing categories including: person, bus, and train. They share common characteristics: (1) large spatial extent (easier to localize and track), (2) common object in daily life (distributed in various datasets that allow models to learn the feature easily), and (3) relatively predictable motion patterns.

The top-3 worst performing categories including: scissors, fork, cell phone. The worst performing categories exhibit opposite characteristics compared with best performing categories: (1) small spatial extent (often $\leq$5% of frame area), (2) frequent occlusions (scissors/fork held by hands, phone against face), and (3) challenging visual properties (metallic reflections on fork).

## 6. Discussion

### 6.1. Limitation

The limitations of the OmniGround dataset primarily center on scale and scope. Its absolute video count (3,475) is smaller compared to some large-scale benchmarks, but this choice was made to prioritize dense, high-quality annotations for the Spatio-Temporal Video Grounding (STVG) task over raw quantity. However, OmniGround still has limited coverage of certain long-tail scenarios (e.g., extreme weather, nighttime/low-light, egocentric viewpoints), reflecting inherent biases in available public video sources. Furthermore, the proposed PG-TAF framework suffers from high latency (average $\approx$ 6.4 seconds/video) due to its two-stage architecture, making it unsuitable for real-time applications, and its Stage 1 relies on proprietary MLLMs, which limits accessibility in certain deployment environments. Finally, as a training-free engineering combination, PG-TAF's performance may be slightly lower than that of fully fine-tuned end-to-end models on specific datasets.

### 6.2. Future Work

The future work is primarily organized around three major directions: dataset extension, methodological improve-

```
prompt = (
"You are a professional Video-Text Alignment Analyst, specializing in evaluating the semantic matching degree between
↪  events occurring within the green marked area and the given text description.\n"
"Please conduct a multi-dimensional analysis, combining the visual content visible within the green area and
↪  referencing possible auxiliary information outside the box, to provide an objective score and detailed
↪  explanation.\n\n"
f"Evaluation Target:\n"
f"Subtitle Content: '{caption}'\n"
f"Analyzed Frames: {num_frames} frames\n\n"
"Evaluation Principles:\n"
"1. The green area is the main visual basis; key objects and actions must be clearly presented, ensuring the focus of
↪  the main object or behavior is clear. Content outside the box is not included in the evaluation scope.\n"
"2. All judgments must be based on the actual visible content within the green area, avoiding subjective speculation
↪  or assumptions.\n"
"3. Out-of-box content can be used to understand the context of the behavior or the interactive relationship, but
↪  cannot be used as the primary basis.\n\n"
"Evaluation Dimensions and Standards (Total Score 10 points):\n"
"1. Subject Existence (0-3 points):\n"
"   - Evaluate whether the subject described in the caption (i.e., the subject in the subtitle content) and its
↪  features appear within the green area.\n"
"   - The subject must appear completely within the green area; out-of-box information is strictly not referenced.\n"
"   - Scoring Details:\n"
"     3 points: The described subject is complete and clearly presented within the green area, with no obvious feature
↪  discrepancies.\n"
"     2 points: Subject exists, but some features are blurry or not completely consistent.\n"
"     1 point: Subject presentation is unclear or partially missing.\n"
"     0 points: The described subject is completely missing, or the subject appears outside the box.\n\n"
"2. Action Accuracy (0-4 points):\n"
"   - Evaluate whether the action performed by the subject within the green area conforms to the description, and the
↪  action must occur entirely within the green area.\n"
"   - Out-of-box information is not included in the action evaluation; only actions within the green area are
↪  considered valid.\n"
"   - Scoring Details:\n"
"     4 points: Action type, direction, sequence, and interactive relationship completely match, and the action occurs
↪  entirely within the green area.\n"
"     3 points: Main action matches, but details (such as execution method or trajectory) have discrepancies.\n"
"     2 points: Action type is similar, but direction, sequence, or execution method is different.\n"
"     1 point: An action is occurring, but the type does not match or key details are missing.\n"
"     0 points: No corresponding action occurs, or the action is completely inconsistent.\n\n"
"3. Context Consistency (0-3 points):\n"
"   - Evaluate whether the environment and behavior within the green area conform to the overall scene description.\n"
"   - Out-of-box auxiliary information can be used to determine the background of the behavior, but the scoring focus
↪  remains on the matching degree of the content within the green area.\n"
"   - Scoring Details:\n"
"     3 points: Behavior within the green area is highly consistent with the overall scene description.\n"
"     2 points: Main scene matches, but some details or background information are incomplete.\n"
"     1 point: Only the basic scene type matches.\n"
"     0 points: Context is completely inconsistent.\n\n"
"Output Format:\n"
"Score: X/10 (X is the sum of the scores of the three dimensions)\n"
"Explanation:\n"
"1. Object Existence: [Analyze the matching degree of the subject in the green area, list supporting or contradictory
↪  evidence]\n"
"2. Action Accuracy: [Explain whether the action is consistent, whether it occurred entirely within the green area,
↪  and any detail discrepancies]\n"
"3. Context Consistency: [Analyze whether the background, lighting, scene layout, etc., within the green area conform
↪  to the description]\n\n"
"Output Example:\n"
"Score: 8/10\n"
"Explanation:\n"
"1. Object Existence: 2 points [The green box shows a cat sitting on a sofa, which is basically consistent with the
↪  description 'a cat resting on a sofa']\n"
"2. Action Accuracy: 3 points [The cat has a clear 'sitting' action, with no fierce movement, which matches the
↪  'resting' description;]\n"
"3. Context Consistency: 3 points [Environmental elements such as the sofa, carpet, and lighting in the green area
↪  completely match the description; the background layout is reasonable and has no incongruous elements]\n\n"
"Note: Your output must strictly adhere to the output example format above, using natural and fluent Chinese
↪  expression. Do not add any extra content or explanatory paragraphs."
)
```

Figure 3. The complete prompt template of CMA score.

ments, and expansion of application scope.

Future research will focus on extending the OmniGround dataset to tackle more complex reasoning challenges. This includes expanding annotations to cover Temporal Relationship Grounding (e.g., ”before/after”) and Multi-Object Joint Grounding (localizing correlated items). Crucially, the dataset needs increased coverage of specialized data like Egocentric viewpoints and Long-Form Videos to address

Table 12. Comparison of different video duration on OmniGround.

| Model | Very Short (0-4s) | | Short (4-6s) | | Medium (6-20s) | | Long (20-40s) | | Very Long (≥40s) | | Overall | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Metrics | m_tIoU | m_vIoU | m_tIoU | m_vIoU | m_tIoU | m_vIoU | m_tIoU | m_vIoU | m_tIoU | m_vIoU | m_tIoU | m_vIoU |
| *Non-generative and task-specific models* | | | | | | | | | | | | |
| CG-STVG | 53.4 | 38.2 | 48.1 | 34.8 | 44.4 | 30.2 | 40.7 | 26.4 | 36.6 | 21.7 | 47.5 | 30.4 |
| *MLLMs with Parameter Sizes of 7B* | | | | | | | | | | | | |
| Qwen2.5-VL | 42.7 | 28.6 | 38.3 | 24.9 | 33.1 | 21.7 | 29.2 | 17.8 | 24.1 | 10.3 | 36.6 | 23.2 |
| VideoMoLMO | 35.1 | 19.4 | 33.8 | 18.3 | 27.4 | 14.2 | 21.6 | 11.9 | 16.7 | 8.3 | 30.2 | 15.7 |
| PG-TAF (ours) | 54.7 | 41.3 | 51.2 | 37.3 | 47.8 | 33.1 | 44.6 | 29.9 | 40.3 | 26.1 | 49.2 | 36.2 |

extreme motion, long-term memory, and challenging appearance changes.

Methodological improvements for PG-TAF should focus on efficiency and robustness. Key direction involves exploring hybrid training to utilize PG-TAF's staged predictions as supervision for end-to-end models.

Finally, the research should broaden its real-world impact. This means adopting the DeepSTG framework as a general diagnostic tool for other video understanding benchmarks (like action localization) and actively bridging STVG to Embodied AI and Robotics for manipulation tasks. Practical deployment requires addressing major limitations: drastically improving computational efficiency (aiming for sub-second latency) and developing strategies for handling domain shift and mitigating ethical bias in deployment scenarios.

## 7. Ethical Considerations

Our work follows the established ethical standards for dataset construction. Our videos are collected from public platforms under permissive licenses, with no personally identifiable information included. Manual filtering removes inappropriate material during collection. The human subjects featured in these videos appear in public contexts where the expectation of recording is diminished, consistent with the terms accepted during upload of original content. We annotate all synthetic samples in our benchmark to enable downstream detection and appropriate usage.

## References


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

[2] Ghazi Shazan Ahmad, Ahmed Heakl, Hanan Gani, Abdelrahman Shaker, Zhiqiang Shen, Fahad Shahbaz Khan, and Salman Khan. Videomolmo: Spatio-temporal grounding meets pointing. *arXiv preprint arXiv:2506.05336*, 2025. 4, 8

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 4, 8

[4] Sergi Caelles, Jordi Pont-Tuset, Federico Perazzi, Alberto Montes, Kevis-Kokitsi Maninis, and Luc Van Gool. The 2019 davis challenge on vos: Unsupervised multi-object segmentation. *arXiv:1905.00737*, 2019. 2

[5] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2694–2703, 2023. 2

[6] Xin Gu, Heng Fan, Yan Huang, Tiejian Luo, and Libo Zhang. Context-guided spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18330–18339, 2024. 4, 8

[7] Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024. 2, 5, 6

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[9] Liang Peng, Junyuan Gao, Xinran Liu, Weihong Li, Shaohua Dong, Zhipeng Zhang, Heng Fan, and Libo Zhang. Vasttrack: Vast category visual object tracking. *Advances in Neural Information Processing Systems*, 37:130797–130818, 2024. 2

[10] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *European conference on computer vision*, pages 208–223. Springer, 2020. 2

[11] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 2

[12] Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8238–8249, 2021. 2

[13] Yuli Vasiliev. *Natural language processing with Python and spaCy: A practical introduction*. No Starch Press, 2020. 3

[14] Jovana Videnovic, Alan Lukezic, and Matej Kristan. A distractor-aware memory for visual object tracking with

sam2. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24255–24264, 2025. 6