SARL-STG: A Spatially Aware Reinforcement Learning Framework for Refining MLLMs in Spatio-Temporal Video Grounding

Supplementary Material

1. Experiment Configuration

1.1. Dataset and Benchmarks

HC-STVG V2: HC-STVG V2 dataset [8] is designed for the spatio-temporal video grounding (STVG) task, sourced from cinematic scenes with each video clip lasting approximately 20 seconds. The presence of multiple actors performing similar actions within clips presents significant challenges for precise spatio-temporal localization. As an extension of HC-STVG V1, HC-STVG V2 substantially expands the dataset scale, comprising 10,131 training samples, 2,000 validation samples, and 4,413 test samples. Notably, since the ground truth annotations for the test set remain unpublished, all experimental results are reported on the validation set.

VidSTG: Developed for spatio-temporal video grounding, this dataset is derived from VidOR and incorporates comprehensive object relationship annotations [17]. It contains 99,943 video-text pairs, categorized into 44,808 declarative statements and 55,135 interrogative queries. The dataset is partitioned into 80,684 sentence-level queries across 5,436 videos for training, with dedicated validation and test splits. A key characteristic of VidSTG is its restriction to predefined object/relationship categories from the source VidOR dataset in its textual queries.

ST-Align: Introduced in the LLAVA-ST framework [5], this novel dataset enables fine-grained spatio-temporal understanding through three distinct tasks: (1) spatio-temporal video grounding (STVG), (2) event localization and captioning (ELC), and (3) spatial video grounding (SVG). With approximately 4.3 million training samples, the dataset provides 2,000 validation samples for each task to facilitate rigorous evaluation.

Charades-STA: Comprising 6,672 videos, this benchmark offers 16,124 query-moment pairs specifically designed for temporal video grounding tasks [2]. The average video duration is 30.60 seconds, with target moments spanning 8.09 seconds on average. Following conventional evaluation protocols, 3,720 query-moment pairs are allocated for testing purposes.

ActivityNet: As a large-scale video question-answering benchmark [16], it features 5,800 untrimmed long videos (average 180 seconds) paired with 58,000 human-annotated QA pairs. The dataset incorporates structured question templates for motion, spatial, and temporal relationships alongside free-form queries, establishing a robust testbed for both temporal localization and fine-grained video understanding.

RefCOCO: RefCOCO, RefCOCO+, and RefCOCOg [4] are all referring expression datasets built on MSCOCO: RefCOCO uses dialog-style short phrases that include positional terms to describe targets; RefCOCO+ removes positional terms to emphasize appearance attributes; and RefCOCOg employs longer, more complex, written-style expressions with richer context and relations, enabling evaluation of target localization under different linguistic styles and cues.

1.2. Baseline Models

To comprehensively evaluate the proposed method, we compare it with several representative vision-language models and open-world grounding framework. These baselines span general-purpose multimodal large language models (MLLMs) and specialized object grounding models, ensuring a broad and rigorous performance comparison.

Qwen2.5-VL [15] is an advanced vision-language model that integrates a powerful language backbone with high-resolution visual encoders. It supports fine-grained perception tasks such as object localization, image captioning, and visual reasoning, enabled by its multi-stage vision transformer and instruction-tuned multimodal alignment. Its strong generalization ability makes it a widely adopted baseline for multimodal understanding tasks.

Qwen3-VL [10] is the upgraded generation of the Qwen multimodal family, offering improved visual representation learning and tighter cross-modal alignment. It incorporates enhanced spatial reasoning and long-context visual processing, enabling stronger performance in tasks requiring precise regional grounding and detailed visual-textual correspondence. Compared to Qwen2.5-VL, it provides better fine-grained localization and robustness on complex scenes.

GroundingDINO [6] is a state-of-the-art open-world object grounding framework that unifies detection and grounding through a transformer-based architecture. It leverages text-conditioned object queries to directly localize objects referred to by natural language expressions. Its joint localization–classification mechanism enables high accuracy in phrase grounding, open-vocabulary detection, and region-level retrieval.

Together, these models provide a diverse benchmark covering multimodal reasoning, open-world localization, and fine-grained grounding, enabling a comprehensive assessment of our method across different levels of visual-textual understanding.

1.3. Evaluation Metrics

To evaluate the performance of spatio-temporal video grounding models, we adopt three commonly used Intersection-over-Union (IoU) based metrics: temporal IoU (tIoU), spatial IoU (sIoU), and volumetric IoU (vIoU). The temporal IoU assesses the alignment between the predicted and ground-truth temporal segments, defined as

$$\text{tIoU} = \frac{|T_p \cap T_g|}{|T_p \cup T_g|},$$

where $T_p = [t_p^{start}, t_p^{end}]$ and $T_g = [t_g^{start}, t_g^{end}]$ denote the predicted and ground-truth temporal intervals. The spatial IoU measures the frame-wise spatial localization quality and is computed as

$$\text{sIoU}_t = \frac{|B_t^p \cap B_t^g|}{|B_t^p \cup B_t^g|},$$

where B_t^p and B_t^g are the predicted and ground-truth bounding boxes at frame t. The final spatial IoU is obtained by averaging sIoU_t over all frames within the overlapping temporal region:

$$\text{sIoU} = \frac{1}{N} \sum_{t \in T_p \cap T_g} \text{sIoU}_t.$$

Finally, the volumetric IoU jointly evaluates spatial and temporal consistency by computing the IoU over the entire spatio-temporal tube, defined as

$$\text{vIoU} = \frac{\sum_{t \in T_p \cap T_g} |B_t^p \cap B_t^g|}{\sum_{t \in T_p \cup T_g} |B_t^p \cup B_t^g|}.$$

In addition to these instance-level measurements, we further report the mean temporal IoU (m_tIoU), mean spatial IoU (m_sIoU), and mean volumetric IoU (m_vIoU). These metrics are obtained by averaging tIoU, sIoU, and vIoU over all samples in the evaluation set:

$$\text{m\_tIoU} = \frac{1}{M} \sum_{i=1}^{M} \text{tIoU}_i$$

$$\text{m\_sIoU} = \frac{1}{M} \sum_{i=1}^{M} \text{sIoU}_i$$

$$\text{m\_vIoU} = \frac{1}{M} \sum_{i=1}^{M} \text{vIoU}_i,$$

where M denotes the total number of test instances. These averaged metrics provide a comprehensive and robust evaluation of overall temporal accuracy, spatial precision, and spatio-temporal grounding quality across the dataset.

vIoU@0.3 and vIoU@0.5 are thresholded accuracy metrics that measure the proportion of samples where the predicted video segment is considered successful under different strictness levels.

$$\text{vIoU@}\tau = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(\text{vIoU}_i \geq \tau\right), \quad \tau \in \{0.3, 0.5\}.$$

Here, I is an indicator function and N is the number of evaluation instances.

2. Prompt Template for Training and Inference

As shown in Fig. 3, the temporal grounding module in the framework employs a prompt template for training and inference, which is specifically designed for this task.

3. More detail about data synthesis pipeline

3.1. Prompt Template for Caption Generation

This section details the method for generating captions that comply with the Spatio-Temporal Video Grounding (STVG) task for videos without textual queries (e.g., videos obtained through web crawling). Since such videos typically lack manual annotations, we design an automated pipeline based on multimodal large language models (MLLMs) and vision-language reasoning, consisting of the following four key steps:

1. **Target-Relevance Filtering**: Videos crawled from the web are often indexed by keywords and may be weakly relevant or irrelevant to the target task. We first employ high-performance multimodal large language models (e.g., GPT-5, Gemini 2.5 [9], Qwen-VL-3 [10]) to analyze and determine the relevance of video content to potential targets, filtering out candidate clips that are semantically unrelated.
2. **Key Event Timeline Generation**: For the filtered videos, we use the model to infer and generate a timeline of key events, extracting non-redundant and semantically coherent temporal segments that serve as the basis for subsequent spatio-temporal grounding.
3. **Long-Range Event Merging and Refinement**: Based on the extracted key event timeline, we employ a multimodal large language model (MLLM) to merge and refine temporally adjacent or redundant events, producing a more concise and coherent sequence of semantic events.
4. **Optimal Spatio-Temporal Segment Selection**: Finally, the MLLM selects the most suitable spatio-temporal segment from the refined event sequence—one that exhibits clear actions, well-defined temporal boundaries, and strong relevance to the intended query—for which a corresponding caption is generated.

The specific prompts used in each of these four steps are illustrated in Fig. 4.

3.2. More Detail of Timestamp Labeling

This section provides a detailed description of the process used to generate timestamp labels for videos in our data synthesis pipeline. The original datasets, primarily derived from related tasks such as RVOS [14] and VLT [13], share a common limitation: the absence of temporal annotations. Specifically, the events described in the queries are implicitly assumed to occur over the entire duration of the video (from the first to the last frame). Such coarse-grained labeling is not suitable for training or evaluating models for Spatio-Temporal Video Grounding (STVG). To convert these datasets into usable spatio-temporal grounding data, we design and implement a data augmentation and annotation generation pipeline, which consists of the following two key steps:

1. **Query Rewriting and Negative Prompt Generation**: We employ large language models (e.g., GPT or Gemini) to semantically rewrite the original queries, generating new queries with different action semantics but related contexts—effectively creating negative sample queries. The specific prompt template used in this process is illustrated in Fig. 5. These newly generated queries are designed to guide the subsequent video generation model in producing negative sample video frames that are visually and semantically distinct from the original video, based on negative prompts.
2. **Negative Video Generation and Concatenation**: The negative prompts are then fed into video generation models such as SORA or WAN [12]. During generation, we use the first and last frames of the original video as initial prompt images to guide the model in synthesizing video segments that differ significantly in content from the original. The resulting negative sample video frames are concatenated at the beginning and/or end of the original video, thereby extending the temporal scope of the video while introducing clearly distinct content.

Through this process, we annotate the original data with fine-grained temporal information, assigning precise start and end timestamps to queries that originally lacked temporal scope. This enables the construction of high-quality spatio-temporal grounding data that is suitable for model training and evaluation.

3.3. More Detail of Bounding Box Labeling

For datasets such as TAG [3], which provide temporal and textual queries but lack spatial annotations (i.e., bounding boxes), we devised a two-stage pipeline to automatically generate this information. Our approach leverages **ChatRex** [3] for initial detection and **DAM4SAM** [11] for robust tracking. The process is as follows:

1. **Key-Frame Seeding with ChatRex**
   First, we employ ChatRex, a model that decouples traditional visual detection from advanced language reasoning. It infers the subject's location and generates initial bounding boxes (bboxes) for the query subject in three critical frames: the **first, last, and a representative intermediate frame** of the event. These three bboxes serve as the spatial seeds for the subsequent tracking stage.
2. **Trajectory Generation and Fusion with DAM4SAM**
   Next, we utilize DAM4SAM, a tracker featuring a distraction-aware memory mechanism that enhances robustness in complex scenes. The model independently tracks the subject starting from each of the three seed bboxes (first, middle, and last), resulting in three distinct object trajectories. Finally, these trajectories are fused using a frame-wise coordinate averaging algorithm. This fusion mitigates potential tracking drift from any single trajectory and yields a more accurate and stable spatial annotation for the entire event.

3.4. Realism and Utility of Synthetic Data

Addressing realism, we newly report Synthetic vs. Real comparisons for visual fidelity (NIQE) and temporal coherence (LPIPS). Results—4.91 vs. 5.20 and 0.12% vs. 0.05% (Syn. vs. Real)—confirm our synthetic data (22.7% subset, Main Fig. 3) matches real videos in texture and dynamics. In addition, Domain shift is mitigated via human review and post-processing (Supp. Fig. 5). Synthetic data targeted augmentation facilitates semantic negative mining. Consequently, zero-shot performance improves on ST-Align (45.1 $\rightarrow$ 49.5, Main Tab. 7) and ActivityNet (33.1 $\rightarrow$ 36.0, Supp. Tab. 6), validating robust transfer to real-world dynamics.

4. More Ablation Studies

4.1. Impact of Training Step and Learning Rates

Tab. 1 presents the impact of different learning rates and training steps on the performance of the SARL-STG. The experimental results indicate that, when the number of training steps is fixed to 1 epoch, the first training stage achieves the best performance with a learning rate of 1e-4, while the second training stage attains optimal results using a learning rate of 1e-6. Furthermore, the overall results demonstrate that all training configurations achieve their peak performance at the first epoch, and continuing training beyond this point tends to lead to overfitting.

4.2. Impact of Dynamic Reward Coefficient

This section analyzes the impact of the dynamic reward coefficient α in the designed Spatio-Temporal Dynamic Reward on the performance of the SARL-STG model. As shown in Tab. 2, the experimental results indicate that the

Table 1. Impact of Training Step and Learning Rates (m_vIoU).

Training Methods	VidSTG (De)	HCSTVG
Stage 1: Lr: 1e-4 (1 epoch)	34.4	39.5
Stage 1: Lr: 1e-4 (2 epoch)	34.0	39.1
Stage 1: Lr: 1e-5 (1 epoch)	34.2	39.5
Stage 1: Lr: 1e-6 (1 epoch)	34.3	39.2
Stage 2: Lr: 1e-4 (1 epoch)	35.0	40.9
Stage 2: Lr: 1e-5 (1 epoch)	35.3	42.1
Stage 2: Lr: 1e-6 (1 epoch)	**35.5**	**42.5**
Stage 2: Lr: 1e-6 (2 epoch)	35.2	41.0

Table 2. Impact of Dynamic Reward Coefficient (m_tIoU).

Training Methods	VidSTG (De)	HCSTVG	Charades
Stage 1: SFT	50.5	58.6	60.8
Stage 2: $\alpha = 0$	51.3	62.3	60.6
Stage 2: $\alpha = 1$	51.5	63.8	61.1
Stage 2: $\alpha = 2$	**52.3**	**64.2**	**61.3**
Stage 2: $\alpha = 0.5$	50.9	63.7	61.0

Table 3. Impact of Pre-trained Model Selection (m_vIoU).

Training Methods	VidSTG (De)	VidSTG (Ig)	HCSTVG
Qwen2.5VL-3B + GD-T	26.4	20.8	34.2
Qwen2.5VL-3B + GD-B	29.7	22.6	37.0
Qwen2.5VL-7B + GD-T	33.1	27.5	40.4
Qwen2.5VL-7B + GD-B	**35.5**	**29.9**	**42.5**

model achieves the most significant performance improvement after further fine-tuning when the dynamic reward coefficient is set to $\alpha = 2$.

4.3. Impact of Pre-trained Model Selection

This section analyzes the impact of pre-trained model selection on the performance of the SARL-STG framework, with results presented in Tab. 3. We conduct a comparative study using two groups of representative pre-trained models, including Qwen2.5VL-7B and Qwen2.5VL-3B from the Qwen vision-language model series, as well as GroundingDino-B (GD-B) and GroundingDino-T (GD-T) from the object grounding model family. The experimental results demonstrate that the model size of the MLLM significantly affects the final video localization performance, as measured by the vIoU metric. This finding further highlights the critical importance of temporal grounding accuracy in video spatio-temporal grounding tasks—that is, the model's ability to perceive target actions and scene changes along the temporal dimension directly influences the precision and reliability of the overall localization.

4.4. Impact of the Spatial Discriminator Performance on Reinforcement Fine-Tuning

Tab. 4 reports the performance of the SARL-STG framework when using different detection models as the spatial discriminator during reinforcement fine-tuning. As shown in the second row, employing GroundingDINO-B—which has not been fine-tuned on spatio-temporal grounding data—introduces a negative effect, as its inherent weakness in query-based object localization limits its ability to serve as a reliable discriminator. In the second row, although the fine-tuned GroundingDINO-T from Tab. 3 is used as the discriminator, its limited capability in understanding the query results in only marginal improvement. In contrast, the proposed Spatial Grounding Module demonstrates stronger detection performance on video frames that correctly correspond to the query. Consequently, when used as the spatial discriminator, it provides a more accurate assessment of the alignment between the input video and the query, leading to a more effective reinforcement fine-tuning process.

Table 4. Impact of the Spatial Discriminator Performance on Reinforcement Fine-Tuning (m_tIoU).

Training Methods	VidSTG (De)	HCSTVG	Charades
Stage 1	50.5	58.6	60.8
Stage 2: GroundingDino-B	48.9	57.1	59.9
Stage 2: GroundingDino-T (SFT)	49.0	62.0	60.9
Stage 2: Spatial Grounding Module	**52.3**	**64.2**	**61.3**

Table 5. Comparison of GRPO, PPO, and DAPO with and without the think reward (m_tIoU).

Training Methods	VidSTG (De)	HCSTVG	Charades
Stage 1	50.5	58.6	60.8
Stage 2: PPO	51.5	63.3	61.0
Stage 2: DAPO	51.8	63.5	60.9
Stage 2: GRPO	51.7	64.0	61.0
Stage 2: GRPO+Think Reward	**52.3**	**64.2**	**61.3**

4.5. Comparison of Reinforcement Learning Paradigms and the Effectiveness of the Think Reward

This section evaluates the impact of different reinforcement learning paradigms—GRPO, PPO, and DAPO—on the reinforcement fine-tuning stage of our SARL-STG framework. GRPO demonstrates stronger reward sensitivity and more efficient credit assignment across temporal reasoning steps, making it better suited for our multi-stage grounding pipeline. To further enhance reasoning quality, we introduce a think reward, designed to encourage the model to produce more structured, stepwise reasoning that leads to more accurate grounding decisions (As shown in Sec. 5.2). As shown in Tab. 5, GRPO already outperforms PPO and DAPO under identical settings, while the integration of the think reward provides additional, consistent gains.

Table 6. Ablation of training data on ID (in-domain) dataset and OOD (out-of-domain) dataset (m_tIoU).

Training Data	ActivityNet (OOD)	HCSTVG (ID)
Random Real Frames	33.1	63.9
Generated Frames (ours)	**36.0**	**64.2**

Table 7. Comparison of different training paradigms on the HCSTVG benchmark.

Training Methods	HCSTVG	
	m_tIoU	m_vIoU
Qwen2.5-VL+GD	46.7	18.0
Qwen2.5-VL+GD (STVG-Wild SFT)	51.5	30.2
SARL-STG (Stage 1: STVG-Wild SFT)	58.6	38.4
SARL-STG (Stage 2: Spatio-temporal dynamic reward)	64.2	42.5

4.6. Comparison of Negative Sample Generation Methods

We address the absence of query-irrelevant negative frames surrounding the ground-truth (GT) segments in the STVG-Wild dataset by designing two negative-sample construction strategies.

The first strategy samples unrelated frames from other videos and directly appends them to the beginning and end of the GT video. However, this approach introduces abrupt content discontinuities, enabling the model to exploit scene-level shifts for temporal discrimination rather than learning genuine query-aligned spatio-temporal cues. Our second strategy, which constitutes our proposed method, leverages a video generation model to synthesize negative frames that preserve the appearance of the query's subject but remain semantically irrelevant. These frames are inserted before and after the GT segment, producing smoother visual transitions while increasing the difficulty of the task. This encourages the model to focus on learning fine-grained spatio-temporal patterns that truly correspond to the query.

As shown in Tab. 6, the experimental results validate the effectiveness of our proposed approach.

4.7. Effectiveness of the Second-Stage Reinforcement Learning

Despite being tuned on STVG-Wild, the decoupled baseline model (Qwen2.5-VL+GD) still exhibits suboptimal performance (yielding a vIoU of 30.2 on the HCSTVG benchmark, as shown in Tab. 7). In contrast, the second-stage reinforcement learning boosts the performance of SARL-STG from 38.4 to 42.5 (a 4.1% improvement), demonstrating that spatio-temporal knowledge injection is fundamentally more advantageous than mere data scaling.

5. More Implementation Details

5.1. Cold-Start

Before initiating the second-stage GRPO reinforcement fine-tuning, we first constructed a 2.5K CoT-based cold-start dataset using rejection sampling in order to stabilize training and endow the model with preliminary reasoning capability. During rejection sampling, generated responses were filtered using the tIoU metric, and only samples with scores above 0.8 were retained. Additional rule-based filtering was applied to remove low-quality CoT traces exhibiting linguistic inconsistency or excessive repetition. The resulting high-quality CoT dataset was then used to perform a LoRA-based supervised fine-tuning step for cold start. The training setup adopted a learning rate of 1e-6, a LoRA rank of 64, a LoRA α of 128, and a batch size of 32.

5.2. Think Reward

Prior work [1] has demonstrated that explicitly modeling temporal information during reasoning is crucial for effective video understanding. To enhance the temporal grounding capability of the model's chain-of-thought (CoT) reasoning process, we introduce a timestamp-aware reward [7] as an auxiliary supervision signal. This reward encourages the timestamps generated in the reasoning chain to align with those provided in the final answer, thereby improving the consistency between intermediate reasoning steps and the predicted temporal segments. The timestamp reward is defined as:

$$R_T = \mathbb{I}\{T_i^A \in T_i^R\}, \tag{1}$$

where $\mathbb{I}$ denotes an indicator function, T_i^R represents the timestamps produced during the reasoning process, and T_i^A denotes the ground-truth timestamps in the answer. The reward takes the value of 1 if all timestamps in the answer appear within the reasoning trace and 0 otherwise. By incorporating this timestamp-aware reward, we encourage the model to attend to fine-grained temporal cues rather than relying solely on coarse, holistic video semantics. This mechanism significantly improves the model's temporal reasoning ability and contributes to more accurate spatiotemporal grounding performance.

5.3. Detailed Explanation of the Input to the Spatial Grounding Module in Stage 1

In this section, we elaborate on the input design strategy for the Spatial Grounding Module during the first training stage. As introduced earlier (in section 3.3.1 of the main manuscript), rather than providing the spatial module with video frames that exactly match the query, we deliberately feed it ground-truth (GT) video frames with a slight random temporal offset. This design is motivated by two main considerations.

Table 8. Comparison of Temporal Video Grounding results on Charades-STA (ID, in-domain) and ActivityNet (OOD, out-of-domain).

Models	HCSTVG		VidSTG	
	m_tIoU	m_vIoU	m_tIoU	m_vIoU
Pred Input	51.8	25.7	42.4	18.9
GT Input	59.4	38.8	47.2	31.7
GT Input + randomly shift	58.6	39.7	50.5	33.6

First, we use GT frames to ensure training stability and the learning of fundamental capabilities. Effective training of spatial localization abilities relies heavily on the model's ability to correctly perceive and comprehend the content of the input video frames. If the input video frames significantly deviate from the content described by the query — for example, if they are entirely unrelated — it may lead to failure in extracting accurate spatial features or even cause spatial collapse, severely impacting model convergence and training stability. To address this, we choose to use ground-truth (GT) video frames as input during the first training stage. This ensures that the model can learn a stable and correct spatial-text alignment as well as reliable spatial feature representations. Moreover, the primary objective of this stage is to equip the model with basic spatio-temporal localization capabilities. Therefore, during joint training, we encourage the spatial module to focus on accurately learning spatial features, while maintaining effective interaction with the temporal module: The spatial module concentrates on aligning the spatial locations of objects in the video with their corresponding textual descriptions; The temporal module focuses on identifying the temporal scope and the sequence of events. Although the two modules interact jointly, they have distinct and well-defined optimization objectives, which helps ensure that the training loss converges more stably.

Second, we introduce mild random temporal offsets to enhance the model's generalization ability. While using GT frames helps the model learn accurate spatial features, training exclusively on perfectly aligned frames can cause the spatial module to overfit to specific frame positions, thereby reducing its ability to generalize to slight variations in timing or frame content. To mitigate this, we apply a slight random temporal offset to the GT video frames used as input, simulating small, realistic frame misalignments. Specifically, for each training sample, we randomly shift the GT frame by 0 to 64 frames (left or right). This encourages the model to not only learn exact matches but also adapt to localized temporal deviations, thereby improving its robustness to small temporal shifts and enhancing generalization in spatial localization.

Finally, the training outcomes demonstrate the effectiveness of this strategy. As shown in Tab. 8, thanks to the carefully designed input strategy and training mechanism, the model develops more generalized and robust foundational capabilities in spatio-temporal localization during the first stage. These capabilities provide a solid foundation for the subsequent reinforcement fine-tuning phase (Stage 2), ensuring better performance and more reliable spatial-temporal grounding in complex scenarios.

Table 9. Comparative model inference on H800 GPU.

Model	Params(B)	Latency(s)	GPU Mem
SARL-STG	8.55B	4.98	26.9G
LLaVA-ST	8.30B	4.57	31.3G
TA-STVG	234M	0.62	28.4G

5.4. Complexity and Execution Efficiency

SARL-STG (8.55B) adds only 3% parameter overhead (249M) to the Qwen2.5-VL baseline. TEM/GLFFM are both lightweight. Inference results on an H800 (Tab.9) confirm efficiency comparable to MLLMs for STVG (LLaVA-ST). Also, unlike task-specific models (TA-STVG, dense-frame input), SARL-STG enables open-world reasoning without escalating GPU memory requirements.

6. Visualizations

As shown in Fig. 1, Fig. 2, we present multiple qualitative examples to visually demonstrate the spatio-temporal grounding capabilities of SARL-STG. The visualization results across different datasets indicate that our model achieves superior performance in terms of localization accuracy and cross-dataset generalization ability.

7. Dataset Documentation and Usage Statement

Intended use. The STVG-Wild dataset is designed solely for academic research on spatio-temporal video grounding.

Data sources. All videos were collected from publicly accessible online platforms. Due to copyright restrictions, we do not redistribute raw videos; only derived annotations are shared.

User privacy. We filter out content involving private individuals, minors, or identifiable sensitive personal data. No personally identifiable information is released.

Annotation process. Textual queries are generated using Gemini 2.5 with an additional cleaning pipeline. This may introduce annotation bias and limited reproducibility.

Safety. Videos containing violent, hateful, or harmful content were removed through automatic and manual screening.

Limitations. The dataset may still contain distributional biases inherent to online media. Reproduction with strictly

open-source tools is possible but may yield content variations.

8. Ethical Considerations

The STVG-Wild dataset is constructed through a multi-source pipeline that integrates several publicly available video datasets, along with additional web-crawled videos. We follow standard research-oriented fair-use principles for all collected data. For the open-source datasets (RVOS, VLT, TVG, STVG), we strictly comply with their original licenses and redistribution terms. For web-crawled videos, we use them solely for research and do not redistribute any raw video content in order to respect platform copyright policies and content ownership. Only derived annotations that do not contain any identifiable information are released.

Privacy and sensitive content filtering. To minimize privacy risks, we employ automatic filtering rules based on platform metadata and textual descriptions to exclude videos containing minors, identifiable private individuals, personal indoor environments, or sensitive scenarios such as violence, medical procedures, or political events. In addition, a manual audit was conducted on randomly sampled videos to ensure compliance with safety and privacy guidelines. No personally identifiable information (PII) is released, and frames containing human faces are not used for biometric purposes.

Use of proprietary models. The dataset construction pipeline incorporates the proprietary Gemini 2.5 model for query generation and event timeline inference, as well as the Wan2.1 video generation model for synthesizing temporally irrelevant negative samples. Although these models enable high-quality large-scale annotation, the closed-source nature of these systems may introduce potential biases and limit full reproducibility. To mitigate this risk, we apply redundancy removal, semantic consistency filtering, and rule-based quality control to eliminate inappropriate, offensive, or hallucinated outputs. We also provide transparent descriptions of the entire annotation procedure to facilitate future reproduction using open-source alternatives.

Annotation quality and fairness. Our multi-model agent pipeline (Gemini2.5 + ChatRex + DAM4SAM) is designed to reduce human labor while maintaining annotation quality. Nevertheless, automatically generated text queries and bounding boxes may inherit biases from both the source videos and the annotation models. We attempt to reduce these biases through negative sample generation, contrastive temporal construction, and consistency checks across multiple model predictions. Though effective in practice, remaining distributional biases may still reflect societal patterns present in web videos.

Data usage and limitations. The STVG-Wild dataset is intended solely for academic research on spatio-temporal video grounding and related vision-language tasks. It should not be used for surveillance, biometric identification, or applications involving sensitive demographic inference. Since a portion of the videos is sourced from the web under fair-use constraints, we do not release raw videos. Users seeking to reproduce the dataset must re-crawl publicly accessible videos independently, and content availability may vary over time.

Summary. We acknowledge that large-scale dataset construction from online media and proprietary models introduces non-negligible ethical challenges. Through careful filtering, privacy safeguards, bias mitigation, and transparent documentation, we aim to minimize potential risks and ensure that STVG-Wild supports responsible, ethical, and reproducible scientific research.

References

[1] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint*, arXiv:2503.21776, 2025. 5

[2] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, pages 5267–5275, 2017. 1

[3] Qing Jiang, Gen Luo, Yuqin Yang, Yuda Xiong, Yihao Chen, Zhaoyang Zeng, Tianhe Ren, and Lei Zhang. Chatrex: taming multimodal llm for joint perception and understanding. *arXiv preprint*, arXiv:2411.18363, 2025. 3

[4] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing*, pages 787–798, 2014. 1

[5] 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*, pages 8592–8603, 2025. 1

[6] 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*, pages 38–55, 2024. 1

[7] Fuwen Luo, Shengfeng Lou, Chi Chen, Ziyue Wang, Chenliang Li, Weizhou Shen, Jiyue Guo, Peng Li, Ming Yan, Ji Zhang, Fei Huang, and Yang Liu. Museg: Reinforcing video temporal understanding via timestamp-aware multi-segment grounding. *arXiv preprint arXiv:2505.20715*, 2025. 5

[8] 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 TCSVT*, 32(12):8238–8249, 2021. 1

[9] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint*, arXiv:2403.05530, 2024. 2

[10] Qwen Team. Qwen3 technical report, 2025. 1, 2

[11] Jovana Videnovic, Alan Lukezic, and Matej Kristan. A distractor-aware memory for visual object tracking with SAM2. In *CVPR*, 2025. 3

[12] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: open and advanced large-scale video generative models. *arXiv preprint*, arXiv:2503.20314, 2025. 3

[13] Xiao Wang, Xiujun Shu, Zhipeng Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng Wu. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13763–13773, 2021. 3

[14] Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. Visa: reasoning video object segmentation via large language models. In *ECCV*, pages 98–115, 2025. 3

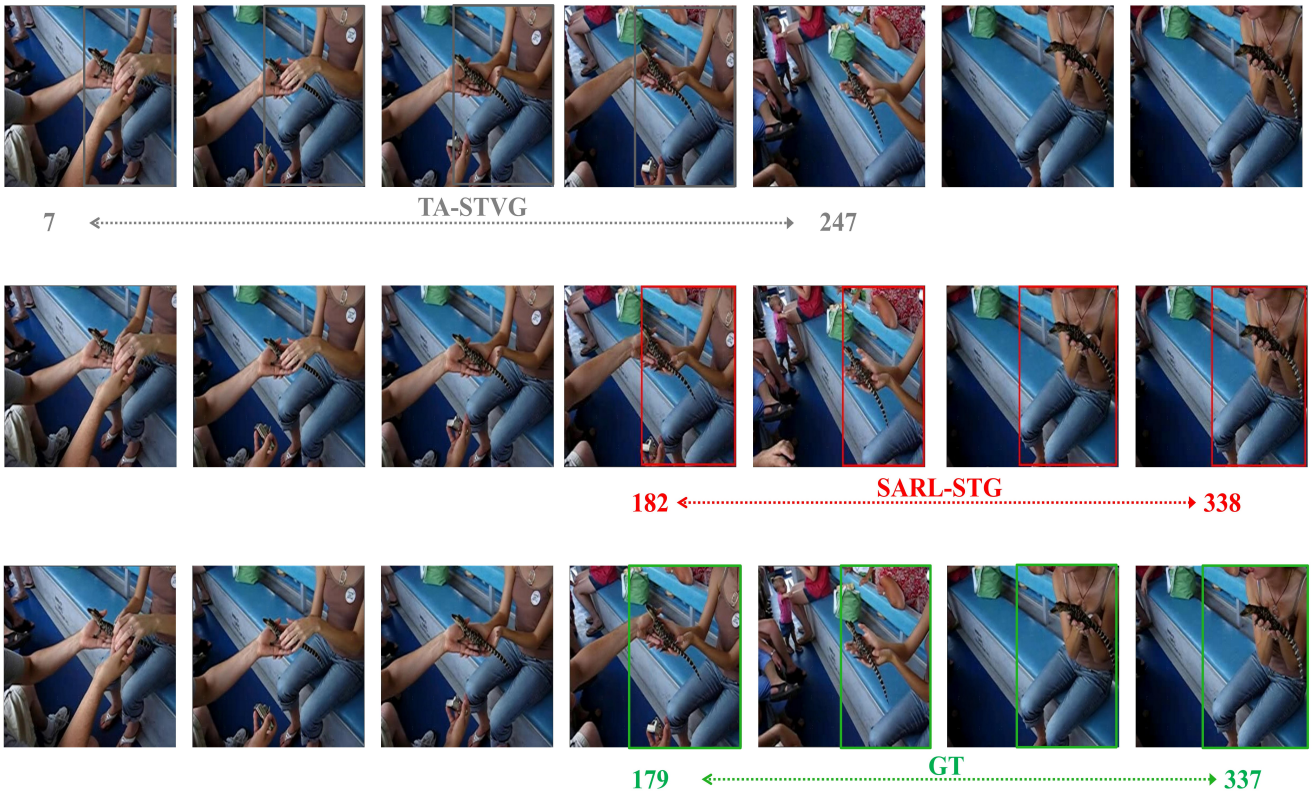
[15] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint*, arXiv:2412.15115, 2024. 1

[16] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, pages 9127–9134, 2019. 1

[17] 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 *IJCAI*, pages 10668–10677, 2020. 1

Query: The man in black and white keeps rubbing his hands and then puts his hands on his hips.

Figure 1. Qualitative Analysis of SARL-STG and TA-STVG on the HCSTVG Dataset.



Query: an adult wearing light blue jeans carries a crocodile.

Figure 2. Qualitative Analysis of SARL-STG and TA-STVG on the VidSTG Dataset.

```
<video>
You are a skilled video understanding assistant. Your task is to precisely localize a specific event within a given video based on a concise event description.
For the event "[EVENT]" described below, determine the exact temporal interval (start and end times in seconds, precise to two decimal places) during which the event occurs in the video.

To complete this task, follow these steps:
1. Carefully analyze the video content and the given event description "[EVENT]".
2. In the <think> </think> tags, provide a clear and logical thought process that explains how you identify the event.
   Your analysis may include references to specific timestamps (e.g., "At 3.20, a person enters the room") or time ranges (e.g., "From 5.10 to 7.80, the dog is chasing a ball").
   Use these timestamps to justify your final time interval prediction.

3. In the <timestep></timestep> tags, you may optionally include intermediate reasoning about key moments or actions related to the event .
4. Finally, in the <answer></answer> tags, output the precise start and end times of the event in the format: "start time to end time" (e.g., "4.50 to 9.22").
   Ensure the times are accurate to two decimal places and reflect the exact moment the event begins and concludes.

Example output format (do not copy the example values):
<think>
[Your detailed reasoning about the video and event, with optional timestamps]
</think>
<timestep>
[Optional intermediate time analysis]
</timestep>
<answer>
start time to end time
</answer>
```

Figure 3. Prompt Template for Training and Inference.

```
You are a professional video understanding assistant. Your task is to analyze the following video and complete the four tasks described below:

### Task 1: Target Object Detection
Determine whether the object **"[{target_object]}"** appears in the video and answer the following:

- `Object Detected: Yes/No`

### Task 2: Key Event Timeline Generation
Based on the video content, generate a concise and non - redundant timeline of key events. The requirements are as follows:

1. **Segmentation Rules:** Split events **only when one of the following conditions occurs**:
    * **Change of the main subject**
      (e.g., the focus shifts from "the woman in red" to "the woman in blue").
    * **Significant change in the nature of the subject's action**
      (e.g., from "arguing" to "turning away and leaving").
    * **Time Precision:** Each event must include a **start and end timestamp (in seconds)**, **accurate to one decimal place (X.X s)**.

2. **Description Requirements:**
   Each event description must include **actions, participants, and objects**.

3. Events must be listed in chronological order without overlapping.
   **Ensure all events are non - redundant.**

Return format:
`[start_time - > end_time] : event description`

### Task 3: Long - Sequence Event Merging and Abstraction
Based on the output of **Task 2**, merge events and refine their descriptions according to the following rules:

1. **Merging Rules:**
   **Only merge** events that:
   * are driven by the **same core subject**, and
   * occur in a **continuous time span**, and
   * form a **continuous or related chain of actions**.

2. **Description Abstraction:**
   For each merged long - duration event, generate a new coherent Chinese description that emphasizes the **subject's action chain and emotional progression**.
   **Events involving different core subjects must be treated as separate long - sequence events.**
   **Maintain the chronological order of the original timeline.**

Return format:
` - - - Merged Events - - - `
`[merged_start_time - > merged_end_time] : refined long - sequence event description`
`(List all merged long - sequence events in order, strictly following subject consistency and action - chain coherence.)`

### Task 4: Optimal Segment Selection for Spatio - Temporal Grounding
From the merged long - sequence events in **Task 3**, select **one segment that is most suitable for the spatio - temporal grounding task**. Selection must satisfy the following criteria:

1. **Single Subject (Primary Criterion):**
   The selected segment **must describe only one main subject**.
   (e.g., "the woman in the blue floral shirt" is valid;
   "two women" or "three people" is invalid.)

2. **Action Salience:**
   The segment must contain a clear and specific **action chain**
   (e.g., "turns, points, and accuses").

3. **Distinctiveness:**
   The segment should be clearly distinguishable from other time periods.

**Selection Priority:**
Single subject > Action salience > Distinctiveness

Return format:
`Optimal Segment (Based on Task 3):`
`[start_time - > end_time] : refined event description`
`Reason: <brief explanation of why it meets the three criteria>`
```

Figure 4. Prompt Template for Caption Generation.

```
You are an expert creative writer specializing in generating highly dynamic and dramatic video concepts.
You are given an original caption:
"[CAPTION]"

Your task is to **rewrite this caption** to generate a compelling, single - sentence script for an AI video model.

**Strict Rules for Rewriting:**

1. **Subject Invariance:** Keep the main subject (the central entity/object) from the original caption **exactly the same**.
2. **Dramatic Action:** Change the described action or situation to something **fundamentally different, highly exaggerated, and visually dramatic** (e.g., sudden explosions, objects violently flying apart, extreme sudden movement like jumping or collapsing).
3. **Dynamic Focus:** The new action must be **vivid, hyper - dynamic, and easy to imagine as a spectacular short video**.
4. **Action Guarantee:** If the original caption contains no main verb or action, you **must invent a visually extreme action** for the subject.
5. **Fluency:** The rewritten caption must be **natural, fluent, and stylistically seamless** English.

**Example of Transformation:**
Original: "A skateboard, characterized by its green wheels."
Rewritten: 'A skateboard featuring green wheels that suddenly break into pieces.'

**Output ONLY the rewritten caption.**
```

Figure 5. Prompt for Generating Captions with Semantically Different Descriptions.