TRAINING-FREE VIDEO PARAGRAPH GROUNDING WITH VIDEO SENTENCE LOCALIZERS VIA SEQUENCE CONSISTENCY

Anonymous authors
Paper under double-blind review

ABSTRACT

Video Paragraph Grounding (VPG) jointly localizes the temporal segments described by all sentences in a paragraph. Although paragraph-level descriptions provide semantic context and temporal relations for disambiguating repeated or visually similar events in long videos, existing VPG models can still underperform strong video sentence localizers. Through a diagnostic analysis of existing VPG methods, we find that the sequential organization of paragraph sentences provides a simple yet effective global cue for coordinating independently grounded segments. Based on this insight, we propose Sentence-to-Paragraph Grounding (S2PG), a lightweight and training-free framework that reformulates VPG as a sequence-level candidate selection problem. S2PG first obtains multiple candidate segments and confidence scores for each sentence from an off-the-shelf Video Sentence Grounding (VSG) model, and then applies Sequence-Aware Global Optimization (SAGO) to jointly optimize sentence-level grounding confidence and paragraph-level temporal consistency. Experiments on two representative VPG benchmarks demonstrate that S2PG consistently improves strong VSG baselines and achieves state-of-the-art performance among existing VPG methods.

1 INTRODUCTION

Video Sentence Grounding (VSG) (Hendricks et al., 2017) aims to identify the temporal segment in a video that corresponds to a natural-language description and is a fundamental task in video–language understanding. Most existing methods (Mu et al., 2024; Feng et al., 2025; Hannan et al., 2024; An et al., 2026) are designed for single-sentence queries and ground each query independently. Although this formulation is effective for short videos containing visually distinctive events, it becomes less reliable in long videos, where semantically similar actions may occur repeatedly and the intended moment may be distinguishable only through its surrounding events. Independently grounding individual sentences may therefore overlook cross-sentence dependencies, event ordering, and broader temporal context that are essential for resolving ambiguous queries.

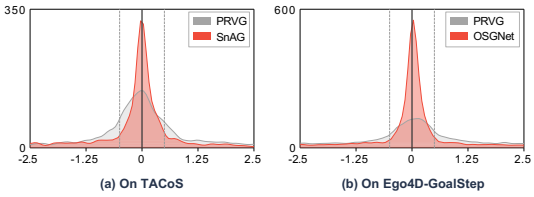


Figure 1: Prediction center drift distributions of the VPG and VSG methods. Drift is the offset between predicted and ground-truth segment centers, normalized by ground-truth segment length. Zero indicates perfect alignment, while the dashed interval denotes predictions centered within the ground-truth segment.

Video Paragraph Grounding (VPG) extends this setting by jointly localizing the temporal segments associated with all sentences in a paragraph (Bao et al., 2021). Compared with an isolated sentence, a paragraph provides complementary semantic cues and temporal relations across events, helping distinguish visually similar moments by their positions in a longer activity sequence. This paragraph-level formulation is particularly relevant to instructional, procedural, and egocentric videos, in which actions are often repeated, closely related, or interpretable only in the context of surrounding steps.

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.

However, leveraging such information requires jointly modeling sentence-level semantics, cross-sentence relations, and temporal dependencies among multiple target segments.

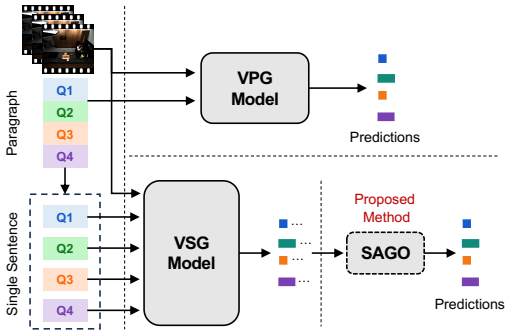


Figure 2: Comparison of VPG and VSG task structures.

Despite the contextual advantages of paragraph-level descriptions, existing VPG models (Shi et al., 2024; Tan et al., 2023) do not necessarily achieve stronger localization accuracy than dedicated VSG models. Figure 1 shows that, on both datasets, the predicted segment centers of some dedicated VSG models are closer to the ground-truth centers than those of the representative VPG models, as reflected by smaller normalized center drift. This observation suggests that strong VSG models may already provide accurate local evidence for individual queries and motivates the following question: **Can their localization capability be effectively transferred to VPG task without training a dedicated paragraph-level model?**

A direct application of existing VSG models is insufficient for VPG. As illustrated in Figure 2, VSG models process one query at a time and localize each sentence independently, preventing them from exploiting the temporal relations encoded in a paragraph. The key challenge is therefore not to replace their sentence-level predictions, but to coordinate them using paragraph-level context. Our analysis of existing VPG methods in the Section titled *What Bridges VSG and VPG?* reveals that the sequential organization of paragraph sentences provides a simple yet effective form of global coordination: individually plausible candidate segments can be further distinguished by whether they form a temporally coherent sequence. This finding naturally bridges independent sentence grounding and paragraph-level joint localization.

Based on this insight, we propose Sentence-to-Paragraph Grounding (S2PG), a lightweight and training-free framework that adapts VSG models to VPG task. S2PG first applies an off-the-shelf VSG model to generate multiple candidate segments and corresponding confidence scores for each sentence. It then formulates paragraph grounding as a sequence-level candidate selection problem and introduces Sequence-Aware Global Optimization (SAGO) method, which jointly considers local grounding confidence and global sequence consistency to identify the globally optimal candidate sequence under the proposed objective.

Unlike conventional VPG methods that introduce paragraph-level reasoning through additional video–text encoding, S2PG performs paragraph-level candidate selection over a compact set of temporal proposals, requiring neither retraining nor modification of the underlying VSG model while avoiding the memory overhead of jointly encoding long videos. Experiments on TACoS and Ego4D-GoalStep show that S2PG consistently improves upon its underlying VSG models and outperforms existing VPG methods. These results demonstrate that S2PG effectively extends the localization capability of VSG models to paragraph-level grounding in a training-free and architecture-agnostic manner.

Our contributions are summarized as follows:

- We present a diagnostic analysis of existing VPG methods and identify paragraph-level temporal order as an effective yet underexplored cue for bridging VSG and paragraph-level joint localization.
- We propose a lightweight, training-free framework that reformulates VPG as a sequence-level candidate selection problem, jointly balancing sentence-level grounding confidence and paragraph-level sequence consistency.
- Experiments on TACoS and Ego4D-GoalStep demonstrate that S2PG consistently improves its underlying VSG models and outperforms existing VPG methods, achieving state-of-the-art performance.

Table 1: Context perturbation results of existing VPG models on TACoS. Abs., Sem., Rel., and Fine denote the target sentence's absolute position, surrounding-sentence semantics, the relative positions of surrounding sentences, and fine-grained sentence order, respectively.

Input Variant	Preserved Contextual Cues				PRVG			DepNet		
	Abs.	Sem.	Rel.	Fine	mIoU	R1@0.3	R1@0.5	mIoU	R1@0.3	R1@0.5
Masked Context Shuffle	×	×	×	×	14.66	21.85	11.45	19.11	28.50	17.46
Masked Context	✓	×	×	×	21.93	32.72	19.58	26.23	39.53	25.09
Target-Fixed Context Shuffle	✓	✓	×	×	26.56	40.60	24.59	27.04	40.68	25.81
Within-Side Context Shuffle	✓	✓	✓	×	29.02	44.96	26.84	27.37	41.15	26.12
Original	✓	✓	✓	✓	29.49	45.40	27.60	27.38	41.14	26.20

2 Related Work

2.1 Video Sentence Grounding.

VSG aims to identify the start and end times of the video segment described by a natural language query, and existing methods (Hu et al., 2024; Lu et al., 2025; Hou et al., 2023; Pramanick et al., 2025) can be broadly categorized into proposal-based and proposal-free approaches. Proposal-based methods (Zhang et al., 2020b; Gao et al., 2017) first construct candidate segments using sliding windows or exhaustive enumeration, and then rank these candidates according to their cross-modal correspondence with the query. In contrast, proposal-free methods (Lei et al., 2021; Lin et al., 2023) directly localize target moments from video–text features. Some (Zhang et al., 2020a) estimate boundary probabilities over the video sequence, while others employ multi-scale temporal representations to generate predictions across multiple temporal scales (Mu et al., 2024; Feng et al., 2025). More recently, Multimodal Large Language Models (MLLMs) have been adapted to VSG by incorporating timestamp-aware representations (Li et al., 2025; Zeng et al., 2025; Chen et al., 2024; Huang et al., 2024; Ren et al., 2024), which provide explicit temporal awareness and timestamp generation capabilities. Despite these advances, existing VSG methods are limited to individual-sentence queries and cannot jointly ground multiple consecutive sentences in a paragraph.

2.2 Video Paragraph Grounding.

VPG jointly localizes a temporal segment for each sentence in an ordered paragraph. Conventional VPG methods (Bao et al., 2021; Shi et al., 2024; Tan et al., 2023; Jiang et al., 2022a; Zhong et al., 2025; Jiang et al., 2022b) generally exploit paragraph-level context through dedicated architectures. In contrast, S2PG preserves the local predictions of an off-the-shelf VSG model and introduces paragraph-level sequence consistency only during candidate selection. Beyond conventional VPG methods, HiERO-StepG (Zenotto et al., 2026) transfers a pipeline developed for Step Grounding to the joint localization of ordered queries. However, Step Grounding generally assumes sequential, non-overlapping procedural steps (Peirone et al., 2025; Dvornik et al., 2021; Chang et al., 2019), whereas S2PG builds on VSG models without this constraint and is therefore better aligned with VPG.

3 What Bridges VSG and VPG?

3.1 Context Perturbation Design

To identify which paragraph-level cues can bridge VSG and VPG, we conduct a controlled diagnostic analysis of the contextual information exploited by existing VPG models. Specifically, we investigate whether paragraph-level localization depends on the semantics and fine-grained ordering of surrounding sentences or can be largely guided by simpler sequence cues.

For each target sentence, we consider four types of contextual information: (1) the target sentence's absolute position in the paragraph; (2) the semantics of surrounding sentences; (3) the relative position of each surrounding sentence with respect to the target, i.e., whether it appears before or after it; and (4) their fine-grained inter-sentence order. Based on these cues, we construct four controlled

input variants, as described below. We evaluate them using multiple baselines on two representative VPG benchmarks. Due to space limitations, we report the results on TACoS dataset in Table 1 and provide the corresponding results on Ego4D-GoalStep in the *Appendix*.

- **Masked Context Shuffle.** Each non-target sentence is replaced with a sequence of `[UNK]` tokens matching its original length, and the target sentence is swapped with a uniformly sampled sentence in the paragraph. This disrupts all four contextual cues.
- **Masked Context.** Each non-target sentence is masked. This preserves the target's absolute position while masking the other three contextual cues.
- **Target-fixed Context Shuffle.** All surrounding sentences are randomly shuffled while the target sentence remains at its original index. This preserves the semantic content of the surrounding sentences and the target's absolute position, but disrupts both their fine-grained ordering and their positions relative to the target.
- **Within-side Context Shuffle.** The preceding and following sentences are shuffled independently. This preserves the semantic content and whether each sentence occurs before or after the target, while disrupting the fine-grained ordering within each side.

3.2 Context Perturbation Analysis

3.2.1 Role of Absolute Position.

We evaluate whether the benefits of paragraph-level input can be attributed solely to the target sentence's absolute position. As shown in Table 1, compared with Masked Context Shuffle, Masked Context preserves the target sentence's original absolute position and improves mIoU by 7.27% for PRVG (Shi et al., 2024) and 7.12% for DepNet (Bao et al., 2021). This indicates that absolute position provides a useful cue even when the surrounding context is masked. However, Masked Context still reduces mIoU by 7.56% for PRVG and 1.15% for DepNet relative to the original input. Thus, absolute position alone cannot fully explain the benefits of paragraph-level input.

3.2.2 Contributions of Sequence Cues.

We first examine the contribution of surrounding-sentence semantics. Compared with Masked Context, Target-fixed Context Shuffle restores the semantic content of surrounding sentences and improves mIoU by 4.63% for PRVG and 0.81% for DepNet, showing that such semantics remain useful even when the original ordering is disrupted. We then assess coarse target-relative order. Compared with Target-fixed Context Shuffle, Within-side Context Shuffle additionally preserves whether each sentence appears before or after the target, yielding further gains of 2.46% and 0.33%, respectively. This indicates that coarse before–after relations help coordinate paragraph-level localization. In contrast, preserving exact inter-sentence order provides only marginal additional benefits: despite shuffling sentences within the preceding and following contexts, Within-side Context Shuffle differs from the original input by only 0.47% for PRVG and 0.01% for DepNet. Thus, fine-grained ordering contributes substantially less than surrounding-sentence semantics and coarse before–after relations.

3.2.3 Key Findings.

These results show that existing VPG models benefit from surrounding-sentence semantics and coarse paragraph-level order, particularly the before–after relations between sentences, while exact inter-sentence ordering provides only marginal additional gains. The results also suggest that the target sentence's absolute position serves as a useful positional cue. Together with our earlier observation that strong VSG models provide accurate sentence-level localization, these findings reveal a natural bridge between VSG and VPG: retaining the strong local predictions of VSG models while enforcing paragraph-level sequence consistency during candidate selection. This observation forms the basis of our proposed S2PG framework.

4 Method

Our perturbation analysis shows that simple sequence cues can effectively guide paragraph-level localization without explicitly modeling fine-grained inter-sentence structure. Motivated by this

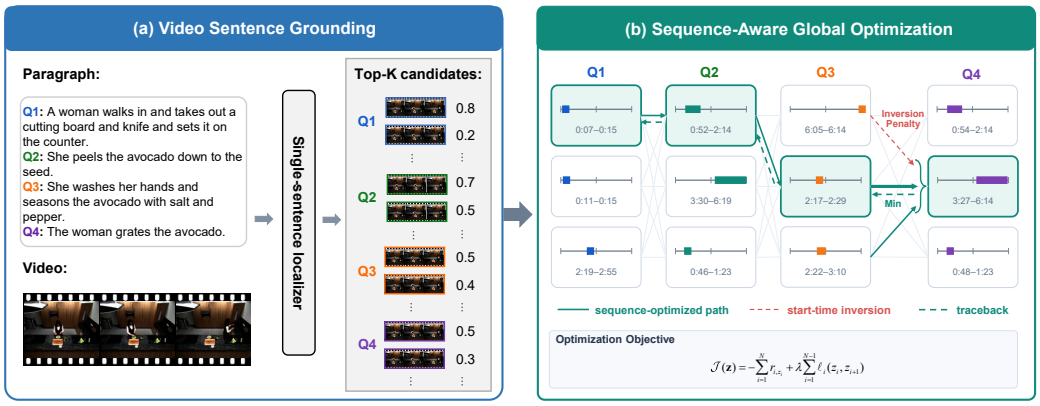


Figure 3: Overview of our S2PG framework. (a) A video sentence localizer independently generates top-K candidate segments and their confidence scores for each sentence. (b) SAGO jointly selects one candidate per sentence by balancing candidate confidence against penalties for start-time inversions, encouraging temporal coherence while allowing temporal overlap.

finding, we introduce S2PG, a training-free framework that combines the accurate local predictions of an off-the-shelf VSG model with paragraph-level sequence consistency. As illustrated in Figure 3, S2PG first generates multiple scored temporal candidates for each sentence independently and then jointly determines the paragraph-level predictions by balancing local confidence and sequence consistency.

4.1 Video Sentence Grounding

Given a video V and an ordered paragraph $Q = \{q_i\}_{i=1}^{N}$, Video Paragraph Grounding aims to predict a temporal segment $\hat{c}_i = (\hat{s}_i, \hat{e}_i)$ in V for each sentence q_i. In the first stage, an off-the-shelf video sentence localizer independently produces K ranked candidate segments $\{c_{i,j}\}_{j=1}^{K}$ for each query q_i. Each candidate is represented as $c_{i,j} = (s_{i,j}, e_{i,j})$, where $s_{i,j}$ and $e_{i,j}$ denote its start and end timestamps, respectively. We use $r_{i,j}$ to denote the corresponding confidence score, with a higher value indicating a more likely localization. S2PG is model-agnostic and can be applied to any video sentence localizer that outputs multiple scored candidates for each query.

4.2 Sequence-Aware Global Optimization

Ordered paragraph descriptions often follow the temporal progression of the corresponding events. Accordingly, the segment associated with an earlier sentence is generally expected to start no later than that associated with a later sentence. We use this property as a soft temporal prior rather than a hard constraint, thereby allowing valid deviations.

Let $\mathbf{z} = (z_1, \ldots, z_N)$ denote a candidate-index sequence for the paragraph, where $z_i \in \{1, \ldots, K\}$ specifies the candidate selected for query q_i. S2PG obtains the optimal sequence $\hat{\mathbf{z}}$ by minimizing

$$\mathcal{J}(\mathbf{z}) = -\sum_{i=1}^{N} r_{i,z_i} + \lambda \sum_{i=1}^{N-1} \ell_i(z_i, z_{i+1}), \tag{1}$$

where the pairwise order penalty between candidates c_{i,z_i} and $c_{i+1,z_{i+1}}$ is defined as

$$\ell_i(z_i, z_{i+1}) = \max\left(0, s_{i,z_i} - s_{i+1,z_{i+1}}\right). \tag{2}$$

The first term in Eqn. (1) favors candidates with high sentence-level confidence scores. The second term penalizes start-time inversions between adjacent queries in proportion to their magnitudes, thereby encouraging paragraph-level sequence consistency. The coefficient $\lambda \geq 0$ controls the trade-off between these two terms. The final prediction for query q_i is $\hat{c}_i = c_{i,\hat{z}_i}$.

Directly minimizing Eqn. (1) requires evaluating K^N candidate sequences and is therefore computationally infeasible as the paragraph length increases. Greedy selection is not guaranteed to produce

Algorithm 1 Sequence-Aware Global Optimization

Input: Candidate segments $\{c_{i,j}\}_{i=1,\dots,N;\, j=1,\dots,K}$, confidence scores $\{r_{i,j}\}_{i=1,\dots,N;\, j=1,\dots,K}$, and penalty coefficient $\lambda \geq 0$
Output: Selected segments $\{\hat{c}_i\}_{i=1}^{N}$

for $j = 1, \dots, K$ **do**
  $C_{1,j} \leftarrow -r_{1,j}$ {initialize the first query}
end for
for $i = 1, \dots, N-1$ **do**
  for $j = 1, \dots, K$ **do**
    $B_{i+1,j} \leftarrow \arg\min_{m \in \{1,\dots,K\}} [C_{i,m} + \lambda \ell_i(m, j)]$
    $C_{i+1,j} \leftarrow -r_{i+1,j} + C_{i,B_{i+1,j}} + \lambda \ell_i(B_{i+1,j}, j)$
  end for
end for
$\hat{z}_N \leftarrow \arg\min_{j \in \{1,\dots,K\}} C_{N,j}$
for $i = N-1, \dots, 1$ **do**
  $\hat{z}_i \leftarrow B_{i+1,\hat{z}_{i+1}}$ {trace back the selected sequence}
end for
for $i = 1, \dots, N$ **do**
  $\hat{c}_i \leftarrow c_{i,\hat{z}_i}$
end for

the global optimum, because a locally preferred candidate may introduce large order penalties with subsequent queries. However, Eqn. (1) has a chain-structured decomposition: each pairwise penalty depends only on the candidates selected for two adjacent queries. This property enables exact optimization using dynamic programming.

Let $C_{i,j}$ denote the minimum cumulative cost for the first i queries, conditioned on selecting candidate $c_{i,j}$ for query q_i. Let $\mathbf{z}_{1:i} = (z_1, \dots, z_i)$ denote the candidate-index sequence for the first i queries. The dynamic-programming state is formally defined as

$$C_{i,j} = \min_{\mathbf{z}_{1:i}: z_i = j} \left[-\sum_{t=1}^{i} r_{t,z_t} + \lambda \sum_{t=1}^{i-1} \ell_t(z_t, z_{t+1}) \right]. \tag{3}$$

Under this definition, the minimum value of the complete objective satisfies

$$\min_{\mathbf{z}} \mathcal{J}(\mathbf{z}) = \min_{j \in \{1,\dots,K\}} C_{N,j}. \tag{4}$$

For the first query, no pairwise order penalty is incurred. The states are therefore initialized as $C_{1,j} = -r_{1,j},\ j \in \{1, \dots, K\}$. For $i \in \{1, \dots, N-1\}$, any candidate sequence ending at $c_{i+1,j}$ must transition from some predecessor candidate $c_{i,m}$. Separating the confidence cost of the current candidate and the pairwise transition penalty from Eqn. (3) yields the recurrence

$$C_{i+1,j} = -r_{i+1,j} + \min_{m \in \{1,\dots,K\}} \left[C_{i,m} + \lambda \ell_i(m, j) \right]. \tag{5}$$

During the forward recursion, for each state $C_{i+1,j}$, we examine all possible predecessor candidates $c_{i,m}$ from the previous query and identify the one that yields the minimum cumulative cost. We store this optimal predecessor in a backpointer:

$$B_{i+1,j} = \arg\min_{m \in \{1,\dots,K\}} \left[C_{i,m} + \lambda \ell_i(m, j) \right]. \tag{6}$$

Here, $B_{i+1,j}$ records which candidate of query q_i should precede candidate $c_{i+1,j}$ in the optimal partial sequence ending at $c_{i+1,j}$.

After all N queries have been processed, each $C_{N,j}$ represents the minimum cost of a complete candidate sequence whose final candidate is $c_{N,j}$. We therefore select the final candidate with the smallest cumulative cost:

$$\hat{z}_N = \arg\min_{j \in \{1,\dots,K\}} C_{N,j}. \tag{7}$$

Table 2: Performance comparison on TACoS. Parentheses denote the absolute gains over the underlying localizers.

	Method	R1@0.3	R1@0.5	mIoU
VSG	VSLNet	29.61	24.27	–
	2D-TAN	37.29	25.32	25.19
	MRNet	55.41	38.54	–
	SnAG	55.56	45.14	41.69
	OSGNet	66.36	55.77	50.89
	UniTime	66.91	55.14	50.00
VPG	DepNet	41.14	26.20	27.38
	PRVG	45.40	27.60	29.49
	ED-VTG	46.00	31.50	32.40
	HSCNet	59.74	42.00	40.61
	S2PG (SnAG)	65.51 (+9.95)	52.79 (+7.65)	49.14 (+7.45)
	S2PG (OSGNet)	**72.56** (+6.20)	**60.56** (+4.79)	**55.45** (+4.56)

Table 3: Performance on Ego4D-GoalStep. Results marked with * are taken from DeCafNet. † denotes evaluation under the clip setting and parentheses indicate absolute gains over the corresponding localizer.

	Method	R1@0.3	R1@0.5	mIoU
VSG	VSLNet	12.15	7.67	–
	BVSLNet-v0	18.15	8.97	–
	RGNet*	21.26	15.71	–
	DeCafNet	21.29	17.46	–
	BVSLNet++	23.75	14.41	–
	UniversalVTG	26.75	22.12	–
	OSGNet	44.27	37.37	34.49
VPG	DepNet†	20.25	14.12	14.44
	PRVG†	21.07	10.82	13.76
	HiERO-StepG	48.51	34.58	–
	S2PG (OSGNet)	**54.90** (+10.63)	**45.79** (+8.42)	**42.70** (+8.21)

Starting from $\hat{z}_N$, we then recover the preceding selections by following the stored backpointers in reverse order. Given the selected index $\hat{z}_{i+1}$ for query q_{i+1}, the selected index for the preceding query q_i is recovered as

$$\hat{z}_i = B_{i+1,\hat{z}_{i+1}}, \qquad i = N-1, \ldots, 1. \tag{8}$$

This backtracking step recovers the globally optimal sequence $\hat{\mathbf{z}} = (\hat{z}_1, \ldots, \hat{z}_N)$.

5 Experiments

5.1 Experimental Settings

Datasets. Following prior VPG work (Bao et al., 2021; Peirone et al., 2025), we conducted experiments on TACoS and Ego4D-GoalStep. **TACoS** (Regneri et al., 2013) contains 127 cooking videos with dense sentence-level temporal annotations. Following prior work (Zhang et al., 2020b; Mu et al., 2024), we used the TACoS$_{\tan}$ split, which includes 9,790 training, 4,436 validation, and 4,001 test queries, and reported results on the test set. Each video is associated with 143.5 queries on average. **Ego4D-GoalStep** (Song et al., 2023) is an egocentric benchmark for localizing procedural steps described in natural language. We used the official split of 851 videos, containing 31,566 training, 7,696 validation, and 5,540 test video–query pairs. Following prior work (Plou et al., 2025; Feng et al., 2025), we reported results on the validation set. Each video is associated with 52.6 queries on average across the three splits.

Baselines. We compared S2PG with a broad range of VSG and VPG baselines. The VSG baselines include VSLNet, 2D-TAN, MRNet, SnAG, RGNet, DeCafNet, OSGNet, Bayesian-VSLNet-v0 (BVSLNet-v0), Bayesian-VSLNet++ (BVSLNet++), UniversalVTG, and MLLM-based method UniTime, while the VPG baselines include DepNet, PRVG, HSCNet, HiERO-StepG, and MLLM-based method ED-VTG (Pramanick et al., 2025).

Evaluation Metrics. Following prior work (Gao et al., 2017; Zhang et al., 2020a), we evaluated temporal localization using mean Intersection over Union (mIoU) and R1@θ. For each query, temporal IoU measures the overlap between the predicted and ground-truth segments. mIoU is the average IoU over all evaluated queries, while R1@θ is the percentage of queries whose predicted segments achieve an IoU greater than θ.

Implementation Details. On TACoS, we followed the inference protocol used in prior VPG work (Shi et al., 2024; Bao et al., 2021). Each paragraph is divided into ordered, non-overlapping sub-paragraphs containing at most eight queries. On Ego4D-GoalStep, following prior work (Zenotto et al., 2026), we used full untrimmed videos and complete paragraphs as the default setting. For comparison with prior VPG methods (Shi et al., 2024; Bao et al., 2021), which are

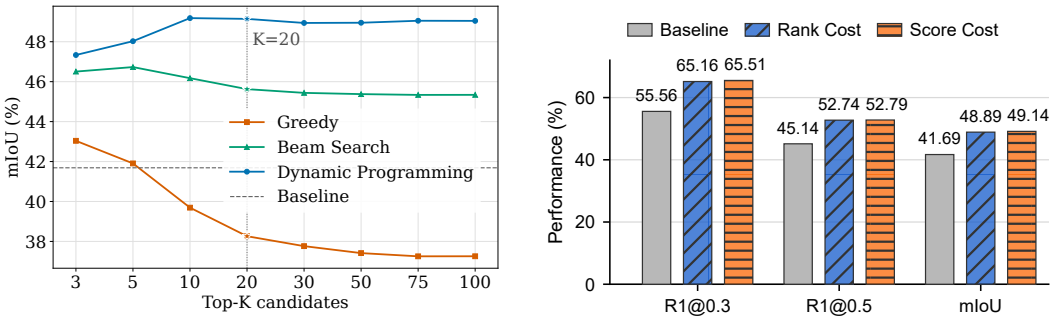


Figure 4: Ablation studies on TACoS using candidates generated by SnAG. Left: Effect of candidate-set size K on mIoU for different selection algorithms. Right: Comparison of different confidence costs.

not designed to process full-length videos and long paragraphs directly, we evaluated them under the clip setting. In this setting, each video is restricted to a 480-second clip, and each paragraph is divided into ordered, non-overlapping sub-paragraphs containing at most eight queries. We used candidate timestamps $s_{i,j}$ and $e_{i,j}$ in seconds and the base localizers' confidence scores $r_{i,j}$ after sigmoid activation. For each query, we retain the top-K candidate segments after non-maximum suppression (NMS). We used $K = 20$ candidates per query and set the penalty coefficient $\lambda = 20$. Parameter sensitivity is analyzed in Appendix C.3.

5.2 Performance Comparison

As shown in Table 2, S2PG brings substantial improvements across all metrics on TACoS when applied to both SnAG and OSGNet. Specifically, it improves mIoU by 7.45% and 4.56% over the corresponding SnAG and OSGNet baselines, respectively. Compared with existing VPG methods, S2PG (SnAG) outperforms HSCNet, the strongest prior method in terms of mIoU, by 8.53%, while S2PG (OSGNet) achieves the best overall performance.

As shown in Table 3, S2PG also demonstrates strong effectiveness on Ego4D-GoalStep. S2PG (OSGNet) improves over the OSGNet baseline by 8.42% and surpasses HiERO-StepG by 11.21% on R1@0.5, achieving the best performance across all metrics. The Appendix (Table 6) further shows that S2PG (OSGNet) substantially outperforms DepNet and PRVG when all three methods are evaluated using full untrimmed videos.

The consistent gains on both datasets show that S2PG effectively leverages paragraph-level sequence consistency to improve paragraph-level localization. Its improvements with both SnAG and OSGNet further indicate that the benefit is not tied to a specific video sentence localizer. These results support our diagnostic finding that simple paragraph-level sequence cues can enhance localization without explicitly modeling fine-grained inter-sentence structure. Moreover, the advantages over existing VPG methods demonstrate the effectiveness of combining strong video sentence localizers with lightweight sequence-aware global optimization.

5.2.1 Efficiency Analysis.

Given candidate segments, the runtime of S2PG depends only on the number of queries N and candidates K, rather than the underlying localizer. Under the default setting of $N \leq 8$ queries per sub-paragraph and $K = 20$ candidates per query, S2PG reranks the candidates for all 4,001 queries in the TACoS test set in only 0.21s on an Intel Core i7-12700F CPU. For comparison, using pre-extracted video features, SnAG takes 92.48s to perform inference on the same test set using a single NVIDIA A100 GPU. Thus, S2PG's optimization time accounts for approximately 0.22% of SnAG's inference time, indicating negligible additional overhead.

5.3 Ablation Study

On Selection Strategy. Table 4 compares different optimization strategies under the default setting. The baseline independently selects the top-ranked candidate for each query, whereas the remaining

strategies optimize the cost in Eqn. (1). Greedy sequentially selects the candidate with the lowest incremental cost without revisiting earlier decisions; Beam Search retains the top $B = 5$ partial sequences at each step; and Dynamic Programming preserves the optimal predecessor for every candidate, thereby exactly minimizing the objective over the complete sequence. Greedy reduces mIoU by 3.43% relative to the VSG baseline, indicating that locally optimal decisions can produce suboptimal paragraph-level predictions. Beam Search mitigates this issue and improves mIoU by 3.94%, while Dynamic Programming performs best across all metrics, improving mIoU by 7.45% over the baseline and 3.51% over Beam Search. Figure 4 (left) further examines the effect of the candidate number K. Greedy degrades rapidly as K increases because larger candidate sets amplify the impact of early local errors. Beam Search is more robust but gradually declines when $K > 5$, suggesting that a limited beam may still prune useful partial sequences. In contrast, Dynamic Programming remains stable across different values of K, confirming the advantage of globally optimizing the complete candidate sequence.

Table 4: Ablation of the selection algorithm on the TACoS dataset using predictions from SnAG.

Method	R1@0.3	R1@0.5	mIoU
Baseline	55.56	45.14	41.69
Greedy	50.69	41.54	38.26
Beam Search	60.41	49.21	45.63
Dynamic Programming	**65.51**	**52.79**	**49.14**

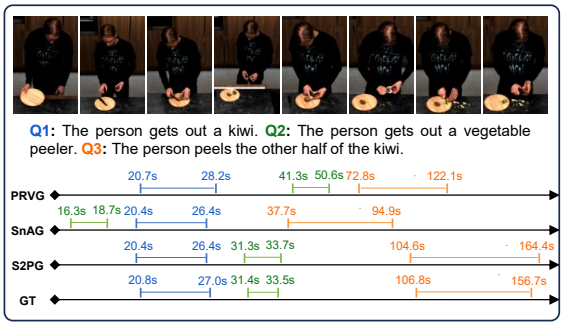


Figure 5: Qualitative comparison on TACoS. Colors indicate the corresponding sentence descriptions.

On Candidate Confidence Cost. We compared score-based and rank-based formulations of the candidate confidence cost. The score-based formulation uses the confidence cost $-r_{i,z_i}$ in Eqn. (1), whereas the rank-based formulation replaces it with the selected candidate rank z_i. As shown in Figure 4 (right), both formulations bring substantial and consistent improvements over the baseline across all three metrics, indicating that S2PG is robust to the specific form of the confidence cost. The score-based formulation further improves mIoU by 0.25% over the rank-based formulation. This suggests that raw confidence scores provide finer-grained information about candidate quality than ranks, which discard the score differences among candidates. We therefore used the score-based formulation as the default in S2PG.

5.4 Visualization

Figure 5 presents a qualitative comparison of SnAG, S2PG (SnAG), and PRVG on an excerpt from a TACoS paragraph. SnAG localizes Q2 before Q1, which is inconsistent with the temporal order of the paragraph. By penalizing start-time inversions in Eqn. (1), S2PG selects a later candidate for Q2, producing a temporally consistent prediction that better matches the ground-truth segment. This example illustrates how S2PG improves the paragraph-level consistency of an underlying video sentence localizer. Compared with PRVG, S2PG (SnAG) also produces temporal boundaries that align more closely with the ground truth, further demonstrating the effectiveness of incorporating sequence consistency into paragraph-level localization.

6 Conclusion

In this work, we revisit how existing VPG methods exploit paragraph context and introduce S2PG. Our analysis shows that existing VPG models have weaker sentence-level localization than strong VSG models and rely mainly on simple sequence cues rather than richer inter-sentence relations. Based on these findings, S2PG provides a training-free framework that combines reliable VSG predictions with paragraph-level sequence consistency through SAGO. Experiments on two benchmarks show that S2PG consistently achieves state-of-the-art performance among existing VPG methods.

AI USE STATEMENT

In this work, we used generative AI tools solely for language polishing. We have not used generative AI tools for other tasks requiring disclosure, and the remaining disclosure categories are not applicable to this work. All AI-assisted language edits were reviewed by the authors. We take full responsibility for the accuracy, originality, and final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 5.1 specifies the datasets, data splits, inference protocols, and hyperparameters. Appendix B.1 derives the dynamic programming recurrence. We also provide the code in the supplementary material to facilitate reproduction of our results.

REFERENCES

Joungbin An, Agrim Jain, and Kristen Grauman. UniversalVTG: A universal and lightweight foundation model for video temporal grounding. *arXiv preprint arXiv:2604.08522*, 2026.

Peijun Bao, Qian Zheng, and Yadong Mu. Dense events grounding in video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 920–928, 2021.

Chien-Yi Chang, De-An Huang, Yanan Sui, Li Fei-Fei, and Juan Carlos Niebles. D3TW: Discriminative Differentiable Dynamic Time Warping for Weakly Supervised Action Alignment and Segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3541–3550, Long Beach, CA, USA, June 2019. IEEE. ISBN 978-1-7281-3293-8. doi: 10.1109/CVPR.2019.00366.

Shimin Chen, Xiaohan Lan, Yitian Yuan, Zequn Jie, and Lin Ma. Timemarker: A versatile video-llm for long and short video understanding with superior temporal localization ability, 2024. URL https://arxiv.org/abs/2411.18211.

Mikita Dvornik, Isma Hadji, Konstantinos G Derpanis, Animesh Garg, and Allan Jepson. Drop-dtw: Aligning common signal between sequences while dropping outliers. *Advances in Neural Information Processing Systems*, 34:13782–13793, 2021.

Yisen Feng, Haoyu Zhang, Meng Liu, Weili Guan, and Liqiang Nie. Object-shot enhanced grounding network for egocentric video. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24190–24200, 2025.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5267–5275, 2017.

Tanveer Hannan, Md Mohaiminul Islam, Thomas Seidl, and Gedas Bertasius. RGNet: A unified retrieval and grounding network for long videos. In *European Conference on Computer Vision*, pp. 1–23, 2024.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5803–5812, August 2017.

Zhijian Hou, Wanjun Zhong, Lei Ji, Difei Gao, Kun Yan, Wk Chan, Chong-Wah Ngo, Mike Zheng Shou, and Nan Duan. CONE: An efficient coarse-to-fine alignment framework for long video temporal grounding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 8013–8028, 2023.

Jingjing Hu, Dan Guo, Kun Li, Zhan Si, Xun Yang, and Meng Wang. Maskable retentive network for video moment retrieval. In *ACM Multimedia 2024*, 2024.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14271–14280, 2024.

Xun Jiang, Xing Xu, Jingran Zhang, Fumin Shen, Zuo Cao, and Xunliang Cai. GTLR: Graph-based transformer with language reconstruction for video paragraph grounding. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2022a. doi: 10.1109/ICME52920.2022.9859847.

Xun Jiang, Xing Xu, Jingran Zhang, Fumin Shen, Zuo Cao, and Heng Tao Shen. Semi-supervised video paragraph grounding with contrastive encoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2466–2475, 2022b.

Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021.

Zeqian Li, Shangzhe Di, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, and Weidi Xie. Universal video temporal grounding with generative multi-modal large language models. *Advances in Neural Information Processing Systems*, 38:64426–64455, 2025.

Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2794–2804, 2023.

Zijia Lu, ASM Iftekhar, Gaurav Mittal, Tianjian Meng, Xiawei Wang, Cheng Zhao, Rohith Kukkala, Ehsan Elhamifar, and Mei Chen. DeCafNet: Delegate and conquer for efficient temporal grounding in long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24066–24076, 2025.

Fangzhou Mu, Sicheng Mo, and Yin Li. SnAG: Scalable and accurate video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18930–18940, 2024.

Simone Alberto Peirone, Francesca Pistilli, and Giuseppe Averta. HiERO: Understanding the hierarchy of human behavior enhances reasoning on egocentric videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19862–19871, 2025.

Carlos Plou, Lorenzo Mur-Labadia, Jose J Guerrero, Ruben Martinez-Cantin, and Ana C Murillo. Temporal video segmentation with natural language using text-video cross attention and Bayesian order-priors. *Computer Vision and Image Understanding*, pp. 104622, 2025.

Shraman Pramanick, Effrosyni Mavroudi, Yale Song, Rama Chellappa, Lorenzo Torresani, and Triantafyllos Afouras. Enrich and detect: Video temporal grounding with multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24297–24308, October 2025.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. In *Transactions of the Association for Computational Linguistics*, volume 1, pp. 25–36, 2013.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. TimeChat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14313–14323, June 2024.

Fengyuan Shi, Weilin Huang, and Limin Wang. End-to-end dense video grounding via parallel regression. *Computer Vision and Image Understanding*, 242:103980, 2024.

Yale Song, Eugene Byrne, Tushar Nagarajan, Huiyu Wang, Miguel Martin, and Lorenzo Torresani. Ego4d goal-step: Toward hierarchical understanding of procedural activities. *Advances in Neural Information Processing Systems*, 36:38863–38886, 2023.

Chaolei Tan, Zihang Lin, Jian-Fang Hu, Wei-Shi Zheng, and Jianhuang Lai. Hierarchical semantic correspondence networks for video paragraph grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18973–18982, 2023.

Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, Yali Wang, Yu Qiao, and Limin Wang. TimeSuite: Improving MLLMs for long video understanding via grounded tuning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Andrea Zenotto, Simone Alberto Peirone, Francesca Pistilli, and Giuseppe Averta. Hiero-stepg @ ego4d step grounding challenge: hierarchical activity understanding enables zero-shot step grounding, 2026. URL https://arxiv.org/abs/2605.31227.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6543–6554, June 2020a.

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2D temporal adjacent networks for moment localization with natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12870–12877, April 2020b. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v34i07.6984.

Yaokun Zhong, Siyu Jiang, Jian Zhu, and Jian-Fang Hu. Context consistency learning via sentence removal for semi-supervised video paragraph grounding. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2025.

A Context Perturbation Analysis on Ego4D-GoalStep

Table 5: Context perturbation results of existing VPG models on Ego4D-GoalStep under the clip setting.

Input Variant	PRVG			DepNet		
	mIoU	R1@0.3	R1@0.5	mIoU	R1@0.3	R1@0.5
Original	13.76	21.07	10.82	14.44	20.25	14.12
Masked Context Shuffle	11.95	18.10	8.91	11.81	16.80	11.12
Masked Context	13.62	20.47	10.78	13.85	19.67	13.42
Target-fixed Context Shuffle	13.50	20.58	10.81	14.37	20.29	13.94
Within-side Context Shuffle	13.55	20.60	10.59	14.34	20.17	13.95

Table 5 reports the corresponding context perturbation results on Ego4D-GoalStep. We first analyze DepNet following the comparisons in the main manuscript. Compared with Masked Context Shuffle, Masked Context preserves the target sentence's original absolute position and improves mIoU by 2.04%. This result shows that DepNet is sensitive to the target sentence's absolute position. However, Masked Context still reduces mIoU by 0.59% relative to the original input. This indicates that absolute position alone cannot fully account for the benefit of paragraph-level input.

Turning to the other contextual cues, Target-fixed Context Shuffle restores the semantics of surrounding sentences and recovers 0.52% mIoU over Masked Context, showing that such semantics remain useful when the original ordering is disrupted. Moreover, Target-fixed Context Shuffle and Within-side Context Shuffle remain close to the original input, with mIoU differences of only 0.07% and 0.10%, respectively. This suggests that target-relative positions and fine-grained inter-sentence ordering provide little additional benefit to DepNet on Ego4D-GoalStep.

PRVG exhibits substantially weaker sensitivity to context on Ego4D-GoalStep. Compared with Masked Context Shuffle, Masked Context improves mIoU by 1.67%, indicating that the target sentence's absolute position also provides a useful cue for PRVG. Masked Context, Target-fixed Context Shuffle, and Within-side Context Shuffle remain close to the original input.

Table 6: Comparison on Ego4D-GoalStep. * denotes the clip setting; other methods use full untrimmed videos.

Method	R1@0.3	R1@0.5	mIoU
DepNet*	20.25	14.12	14.44
PRVG*	21.07	10.82	13.76
DepNet	5.33	2.92	3.99
PRVG	1.52	0.58	1.70
S2PG (OSGNet)	**54.90**	**45.79**	**42.70**

Overall, the Ego4D-GoalStep results further show that existing VPG models do not consistently depend on detailed inter-sentence structure. DepNet benefits mainly from surrounding-sentence semantics and the target sentence's absolute position, while PRVG primarily exploits the latter. These findings likewise support S2PG's design of combining reliable sentence-level predictions with lightweight paragraph-level sequence consistency.

B Theoretical Analysis

B.1 Derivation of the Dynamic Programming Recurrence

For the state $C_{i+1,j}$, replacing i with $i+1$ in Eqn. (3) of the main manuscript gives

$$C_{i+1,j} = \min_{\mathbf{z}_{1:i+1}:z_{i+1}=j} \left[-\sum_{t=1}^{i+1} r_{t,z_t} + \lambda \sum_{t=1}^{i} \ell_t(z_t, z_{t+1}) \right]. \tag{9}$$

Because $z_{i+1} = j$ is fixed, let $m = z_i$ denote the candidate index selected for the preceding query. The final confidence term is $-r_{i+1,j}$, and the final pairwise penalty is $\lambda \ell_i(m, j)$. Separating these

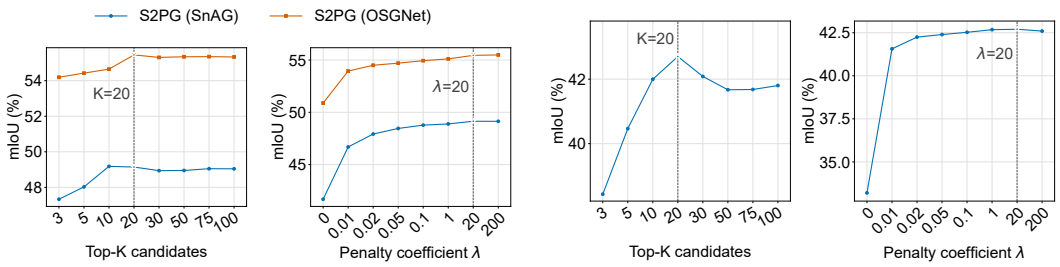


Figure 6: Parameter sensitivity of S2PG regarding candidate-set size K and sequence penalty coefficient λ. Left: Results on TACoS. Right: Results on Ego4D-GoalStep with OSGNet. The vertical dotted lines indicate the shared default settings.

two terms from the remaining prefix cost yields

$$C_{i+1,j} = -r_{i+1,j} + \min_{m\in\{1,\dots,K\}} \left\{ \lambda \ell_i(m,j) + \min_{\mathbf{z}_{1:i}:z_i=m} \left[-\sum_{t=1}^{i} r_{t,z_t} + \lambda \sum_{t=1}^{i-1} \ell_t(z_t, z_{t+1}) \right] \right\}. \tag{10}$$

The inner minimization is exactly $C_{i,m}$ according to Eqn. (3) of the main manuscript. Therefore,

$$C_{i+1,j} = -r_{i+1,j} + \min_{m\in\{1,\dots,K\}} \left[C_{i,m} + \lambda \ell_i(m,j) \right], \tag{11}$$

which recovers the recurrence in Eqn. (5) of the main manuscript.

B.2 Time Complexity

As shown in Lines 4–9 of Algorithm 1, the dynamic program computes K states for each of the $N-1$ remaining queries, and each state examines K predecessors. Its time complexity is therefore $O((N-1)K^2) = O(NK^2)$, while initialization and backtracking contribute only lower-order terms.

C Additional Experiments

C.1 Ground-Truth Temporal Order.

Under the default grouping protocols described above, we measured the proportion of adjacent query pairs with nondecreasing ground-truth start times. This proportion is 100% for the TACoS test sub-paragraphs and 98.32% for the Ego4D-GoalStep validation sequences. These statistics provide empirical support for using start-time order as a temporal prior on the evaluated benchmarks.

C.2 Comparison on Ego4D-GoalStep

To examine the applicability of existing VPG methods to Ego4D-GoalStep, we evaluated DepNet (Bao et al., 2021) and PRVG (Shi et al., 2024) using full untrimmed videos. As shown in Table 6, DepNet and PRVG achieve mIoU scores of only 3.99% and 1.70%, respectively, compared with 14.44% and 13.76% under the clip setting. This substantial degradation illustrates the difficulty of applying existing VPG methods to the considerably longer videos in Ego4D-GoalStep, as these methods are not designed to process full-length videos directly. Therefore, we evaluate prior VPG methods under the clip setting in the main manuscript. In contrast, S2PG (OSGNet) achieves an mIoU of 42.70% using full untrimmed videos, substantially outperforming DepNet and PRVG even when they are evaluated under the simpler clip setting.

C.3 Parameter Sensitivity

Figure 6 analyzes the sensitivity of S2PG to the candidate-set size K and the sequence penalty coefficient λ on TACoS and Ego4D-GoalStep. Increasing K from 3 to 20 improves mIoU, while

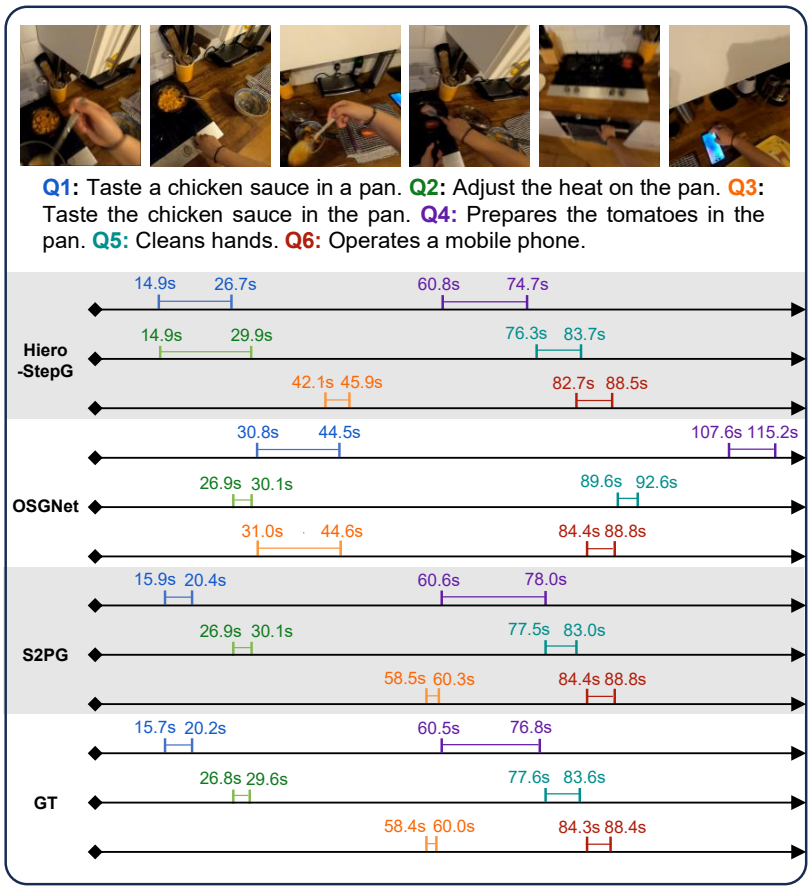


Figure 7: Qualitative comparison on Ego4D-GoalStep. Colors indicate the corresponding sentence descriptions.

further increasing K yields only marginal changes. S2PG also remains stable over a wide range of nonzero λ values. Based on these consistent trends, we use $K = 20$ and $\lambda = 20$ as shared defaults across datasets, without dataset-specific parameter tuning.

C.4 Visualization on Ego4D-GoalStep

Figure 7 presents a qualitative comparison of OSGNet, S2PG (OSGNet), and HiERO-StepG on an Ego4D-GoalStep video. OSGNet localizes Q2 before Q1 and places Q4 after Q5 and Q6, producing a prediction sequence that is inconsistent with the temporal order of the paragraph. By penalizing start-time inversions during candidate reranking, S2PG selects a temporally consistent sequence from the candidates generated by OSGNet. The resulting segments closely match the ground-truth boundaries for all six queries. Compared with HiERO-StepG, S2PG (OSGNet) also produces more accurate temporal boundaries, particularly for Q1–Q3, further demonstrating the effectiveness of incorporating sequence consistency into paragraph-level localization on Ego4D-GoalStep.

D Limitations

By leveraging lightweight paragraph-level sequence consistency, S2PG substantially improves its underlying VSG models and achieves state-of-the-art performance among existing VPG methods. Beyond these gains, the semantics and fine-grained order of surrounding sentences may provide complementary cues for further improving paragraph-level localization. In future work, we will investigate how to leverage these cues to further improve our localization framework.