OmniVTG: A Large-Scale Dataset and Training Paradigm for Open-World Video Temporal Grounding

Minghang Zheng[1] Zihao Yin[3] Yi Yang[3] Yuxin Peng[1] Yang Liu[1,2,4*]

[1]Wangxuan Institute of Computer Technology, Peking University [2]State Key Laboratory of General Artificial Intelligence, Peking University

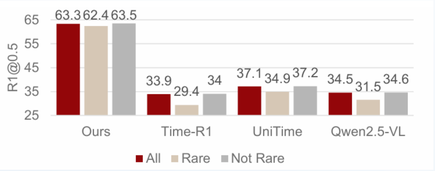
[3]Central Media Technology Institute, Huawei Technologies Ltd. [4]PKU-WUHAN Institute for Artificial Intelligence, Peking University

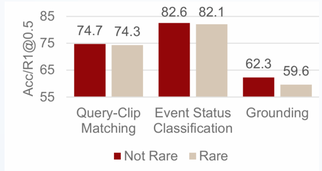
CVPR JUNE 3-7, 2026 DENVER COLORADO

1 Introduction

Task: Video Temporal Grounding
- Localize video segments from text queries
- **Inputs:** Video + Sentence query
- **Outputs:** Target video clip (start and end timestamps)

The Bottleneck: Poor Performance on open-world "**Rare Concepts**"
- **Limitations of Existing Datasets:** Small scale and limited semantic diversity
- **Methodological Gaps:** MLLMs excel at understanding but struggle with precise grounding

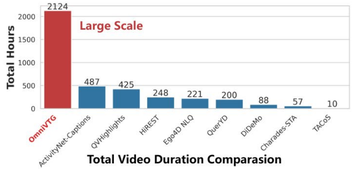




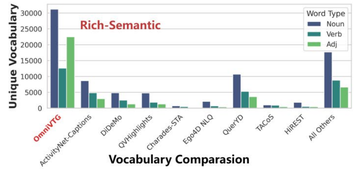
Ours Contributions
- **OmniVTG Dataset:** over **2,000 hours** of video and **rich semantic diversity**
- **Self-Correction CoT:** first make a coarse prediction of the target timestamps, then **zoom in and refine it** using its video understanding capabilities

(a) Our OmniVTG Dataset

(b) Video Domain of Our OmniVTG Dataset



Total Video Duration Comparasion

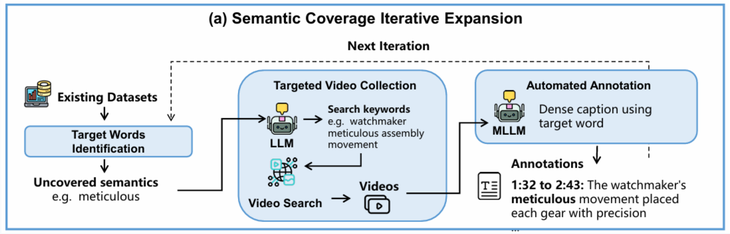


Vocabulary Comparasion

(c) Comparasion with Other Datasets

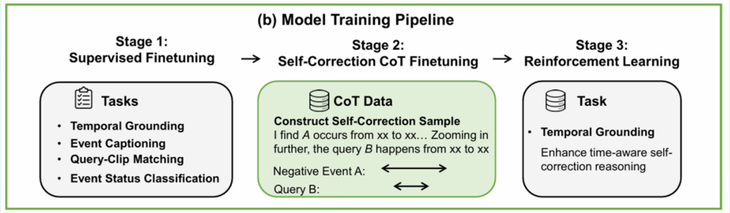
2 OmniVTG Dataset

Iterative Semantic Expansion Strategy
- **How to automate high-quality timestamp annotations?**
  - **Insight:** MLLMs excel at Dense Captioning but are weak at Temporal Grounding
  - **Strategy:** Convert Temporal Grounding into Dense Captioning
- **How to ensure high semantic coverage for long-tail vocabulary?**
  - **Targeted video collection:** Analyze vocabulary gaps in existing datasets and specifically crawled videos for those missing rare concepts
  - Instructed MLLMs to generate **dense captions with timestamps** using those rare concepts



3 Method

Self-Correction CoT
- **Insight:** MLLMs excel at **understanding** but struggle with precise **grounding**
- **The "Predict-Correct" Paradigm**
  - **Predict:** Generate initial, coarse temporal boundaries
  - **Correct:** Reflect on the initial predictions and correct them using its understanding ability
- **Three Stage Training**
  - **Multi-Task SFT:** Build foundational capability
  - **CoT Tuning:** Teach the "predict-correct" reasoning: make a coarse prediction, then zoom in and refine it
  - **RL:** Shift from mimicking fixed-pattern CoT to self-exploring more robust reasoning paths



4 Experiments

Comparison with SOTA
- **Zero-Shot SOTA:** Leading zero-shot performance across Charades-STA, ActivityNet, QVHighlights, and TVGBench.

Method	Charades-STA [7]			ActivityNet [12]			QVHighlights [13]			TVGBench [32]		
	R1@0.3	R1@0.5	R1@0.7	R1@0.3	R1@0.5	R1@0.7	R1@0.3	R1@0.5	R1@0.7	R1@0.3	R1@0.5	R1@0.7
ChatVTG [25]	52.7	33.0	15.9	40.7	22.5	9.4	-	-	-	-	-	-
TimeChat [27]	-	32.2	13.4	36.2	20.2	9.5	-	8.32	4.26	22.4	11.9	5.3
HawkEye [30]	50.6	31.4	14.5	49.1	29.3	10.7	-	-	-	-	-	-
VTimeLLM [11]	51.0	27.5	11.4	44.0	27.8	14.3	-	26.1	11.1	-	-	-
TimeSuite [41]	69.9	48.7	24.0	-	16.6	9.28	-	12.3	9.16	31.1	18.0	8.9
VideoChat-Flash [16]	74.5	53.1	27.6	-	-	-	-	-	-	32.8	19.8	10.4
TRACE [8]	-	40.3	19.4	-	-	-	-	-	-	37.0	25.5	14.6
UniTime [17]	-	59.1	31.9	-	22.8	14.1	-	41.0	31.5	-	-	-
Time-R1 [32]	78.1	60.8	35.3	58.6	39.0	21.4	80.3	66.2	44.8	41.8	29.4	16.4
Qwen2.5-VL-7B [2]	72.5	53.6	28.5	24.4	13.6	6.7	15.9	7.10	4.19	35.3	20.0	12.5
OmniVTG (Ours)	**78.3**	**63.2**	**37.0**	**60.3**	**39.8**	**21.4**	**82.8**	**67.0**	**47.3**	**54.5**	**37.6**	**19.7**

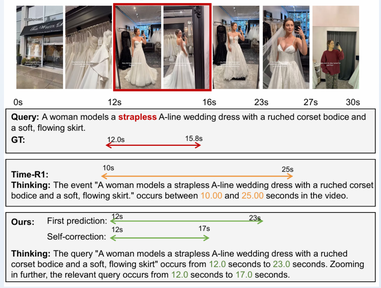
- **Rare Concept Robustness:** Maintains stable accuracy on long-tail queries

Method	OmniVTG Test Set (Ours)						ActivityNet Captions [12]					
	Full			Rare			Full			Rare		
	R1@0.3	R1@0.5	R1@0.7	R1@0.3	R1@0.5	R1@0.7	R1@0.3	R1@0.5	R1@0.7	R1@0.3	R1@0.5	R1@0.7
UniTime [17] [5]	59.9	37.1	15.8	54.2	34.9	12.7	39.9	22.8	14.1	-	-	-
Time-R1 [32]	57.1	33.9	14.7	49.7	29.4	15.7	58.6	39.0	21.4	56.2	36.1	19.3
Qwen2.5-VL-7B [2]	49.0	34.5	16.9	44.7	31.5	15.7	24.4	13.6	6.70	22.3	12.9	4.8
OmniVTG (Ours)	**74.2**	**63.3**	**47.6**	**74.1**	**62.4**	**46.2**	**60.3**	**39.8**	**21.4**	**60.1**	**39.5**	**20.8**

Ablation Studies
- **CoT and RL** is important for Rare Concepts and OOD perfermance
- Performance scales as **data increases**

Model	OmniVTG (Full)	OmniVTG (Rare)	ActivityNet
1. Necessity of Training Stages			
Qwen2.5-VL-7B	34.5	31.5	13.6
+ SFT	62.3	59.6	25.6
+ SFT + COT	62.4	61.3	32.5
+ SFT + RL	62.8	60.6	37.2
+ SFT + COT + RL	**63.3**	**62.4**	**39.8**
2. Impact of SFT Data Scale			
SFT (10% data)	41.9	37.8	15.3
SFT (50% data)	58.7	55.4	21.9
SFT (100% data)	**62.3**	**59.6**	**25.6**
3. Comparison of Reasoning Strategy			
w/o Reasoning	62.3	59.6	25.6
Rule-base reflection	62.4	61.0	37.9
Content-aware reflection	**63.3**	**62.4**	**39.8**

Qualitative Results
- Successfuly correct initial mistakes



Paper Code Project Page