EvoGround: Self-Evolving Video Agents for Video Temporal Grounding

Minjoon Jung[1,2], Byoung-Tak Zhang[1], Lorenzo Torresani[2]
[1]Seoul National University, [2]Northeastern University

Introduction

Video Temporal Grounding (VTG)

- *Video temporal grounding takes an untrimmed video and a natural-language query as input and localizes the temporal moment that best matches the query.*

Input: video + query → output: time span

Current Challenge

- *Collecting labeled data for VTG is highly labor-intensive, requiring annotators to manually localize temporal moments and write descriptions for untrimmed videos.*
- *Existing models typically rely on such manual labels, making training costly and limiting scalability.*

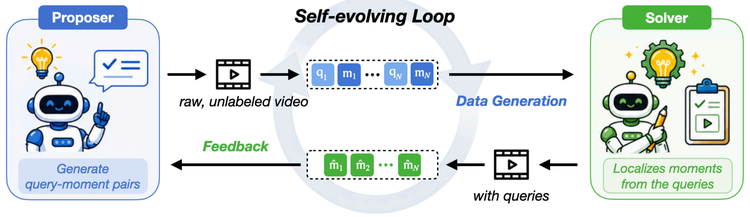
Research Question

- *Can a model reliably learn temporal information in videos without manual labels?*

EvoGround: Self-Evolving Video Agents

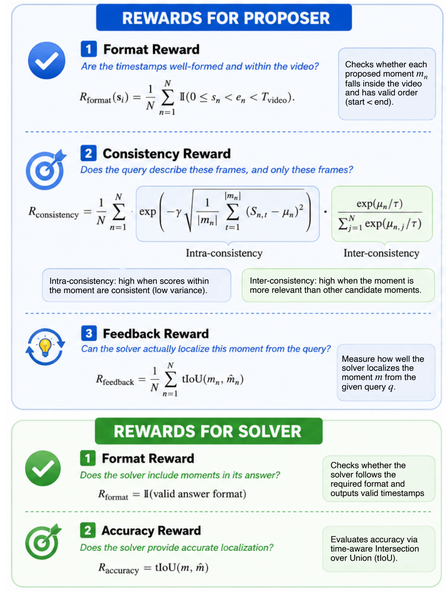
Overview of EvoGround

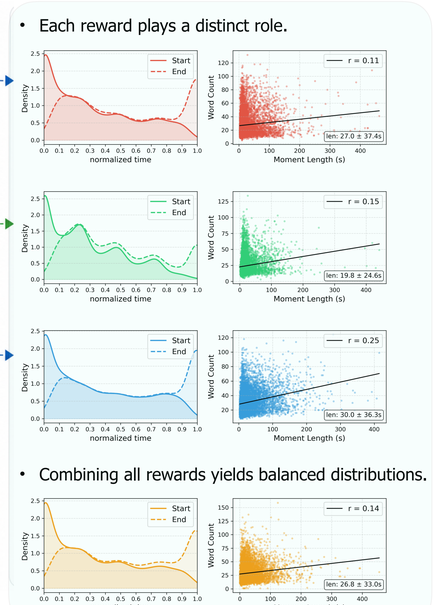
- *EvoGround consists of a proposer and a solver, evolved through an iterative two-stage process via reinforcement learning (RL).*



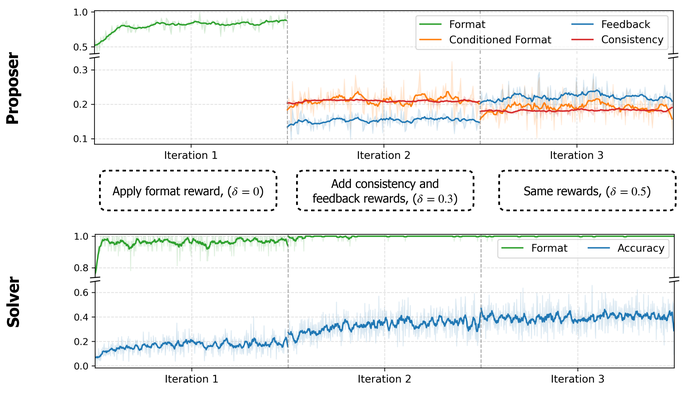
Same backbone. No labels and teacher!

- ***Iterative Optimization.*** *The proposer generates valid, semantically consistent, and solvable query–moment pairs, while the solver is trained to localize temporal moments from the generated queries.*
- ***Self-evolving Loop.*** *Both agents are initialized from the same backbone and progressively improve each other: a stronger proposer generates higher-quality supervision, which in turn trains a stronger solver.*





Optimization



- ***Objective.*** *We use GDPO that normalizes each reward first, combine after. This keeps easy rewards from dominating the signal.*
- ***Curriculum Design.*** *We first warm up the proposer with the format reward only. In later iterations, we apply conditioned format reward:*

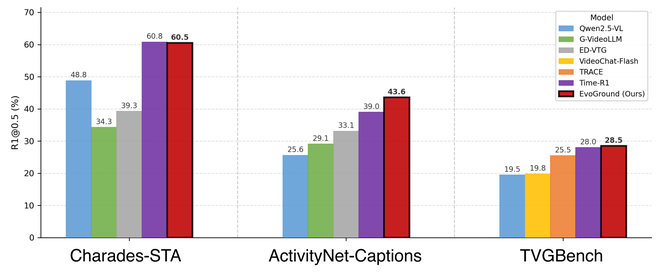
$$\mathbb{I}\big((s_n, e_n) \in \mathbb{V} \wedge \text{tIoU}(m_n, \hat{m}_n) \geq \delta\big),$$

which enforces the solver's accuracy. We progressively introduce consistency and feedback rewards while increasing the localization threshold δ, encouraging more precise temporal grounding over time.

- ***Reward Dynamics.*** *Format rewards quickly saturate, while feedback and accuracy rewards steadily improve across iterations, indicating effective curriculum learning.*

Experiments

- Despite never seeing any manual annotations, EvoGround matches or outperforms existing models across various VTG benchmarks.

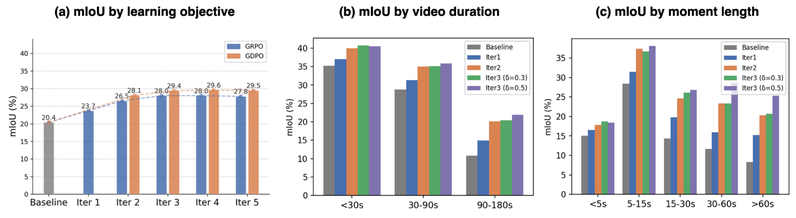


- Beyond grounding, EvoGround emerges as a state-of-the-art fine-grained video captioner, outperforming exiting models by a large margin across all captioning metrics.

Method	TemporalBench						
	Similarity	CIDEr	ROUGE	BLEU_1	BLEU_2	BLEU_3	BLEU_4
Qwen2-VL	51.9	6.9	18.0	12.5	6.1	3.0	1.6
LLaVA-OneVision	50.1	0.3	14.5	11.1	5.1	2.2	1.1
LLaVA-NeXT-Video	50.1	2.3	15.8	18.1	7.0	2.6	1.1
InternLM-XC2.5	52.4	2.3	15.9	17.8	7.1	2.8	1.2
VideoLLaVA	46.0	4.5	16.9	12.6	5.4	2.3	1.0
MiniCPM-V2.6	47.2	1.5	14.2	15.5	5.4	1.9	0.8
Phi-3.5-Vision	42.9	3.7	16.5	20.4	8.4	3.4	1.6
EvoGround (Ours)	**53.8**	**11.4**	**20.5**	**26.9**	**12.5**	**5.6**	**2.6**

Analysis

- GDPO and our curriculum design provide consistent improvements across different video durations and moment lengths, fostering a more robust solver.



- EvoGround scales effectively with data: incorporating an additional 5K videos from Charades-STA yields further improvements, and the method generalizes well across different backbone types and scales.