Conditional Multi-Event Temporal Grounding in Long-Form Video

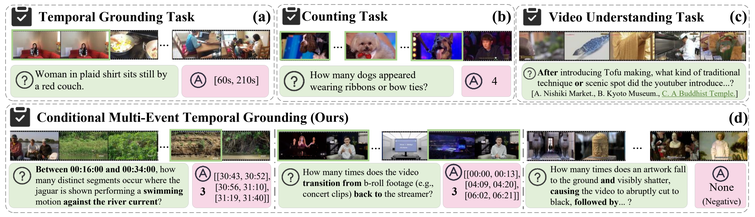
Yuanhao Zou[1] · Arthad Kulkarni[1] · Lucas Toñanez[1] · Lincoln Spencer[1] · Guangyu Sun[1] · Tianxingjian Ding[1] · Andong Deng[1] · Yi Li[2] · Shuangjun Liu[2] · Yuan Li[2]
Dashan Gao[2] · Ning Bi[2] · Taotao Jing[2] · Shuai Zhang[2] · Chen Chen[1] | [1]University of Central Florida [2]Qualcomm AI Research

PROJECT PAGE

Find ALL condition-compliant events. CoMET-Bench unifies counting, temporal grounding and rejection; training-free CoMET-Agent improves GPT-5 F1@0.5 **10.1** → **16.2** and Gemini 3 Flash **14.6** → **19.0** (+6.1 / +4.4 percentage points).

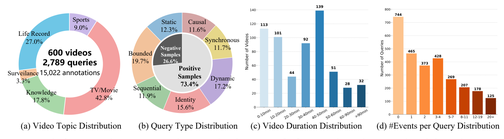
01 TASK: FROM ONE MOMENT TO ALL

Given long-form video V and query Q, output **all qualifying intervals** P = {(s, e)} and count |P|. An empty set is correct only when no event satisfies **every** temporal and spatial condition.



02 CoMET-BENCH: SCALE & COMPOSITION

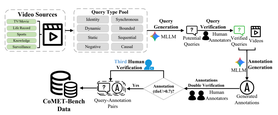
600 videos · 2,789 queries · 15,022 event annotations · 33.8 min/video
Five domains; 73.4% positive / 26.6% negative queries. 111 videos exceed 50 min (32 exceed 90 min); 510 queries have ≥8 events.



(a) Video Topic Distribution (b) Query Type Distribution (c) Video Duration Distribution (d) #Events per Query Distribution

Temporal: causal, sequential, synchronous, bounded. **Spatial:** static, dynamic, identity. Queries combine attributes; event counts range from 0 to 20+.

03 CONSTRUCTION & HUMAN QUALITY CONTROL



Attribute-guided generation → human query verification → MLLM interval drafts → independent refinement → adjudication. Gemini 3 Pro-Preview drafts five queries/video. Compositional sampling favors 2–3 attributes; time references use HH:MM:SS. Humans check relevance, clarity and plausibility, rejecting spatial-only counting.

Human intervention	Reported statistic
Candidate queries	3,000 → 2,789 retained
Queries unchanged / rewritten / discarded	15.4% / 77.6% / 7.0%
Draft intervals removed as spurious	31.7%
Final intervals newly added by humans	26.0%
Queries requiring adjudication	182 / 2,789 (6.5%)

Independent review: a second annotator re-annotates every query; interval agreement requires tIoU > 0.7. Disputes go to a senior reviewer. **11 annotators**, including 3 adjudicators; approximately **791 person-hours** (1,021 s per retained query, excluding tool development/training).

04 EVALUATION: COUNT, GROUND, REJECT

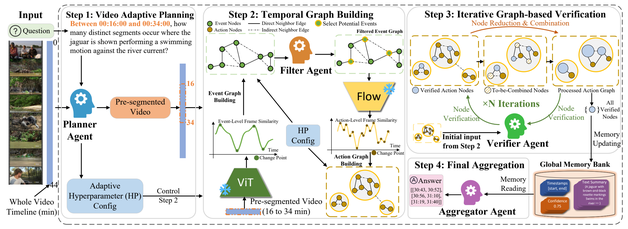
Counting: MAE measures count error; OBO allows |predicted – true count| ≤ 1; Pearson measures count correlation.
Grounding (positive queries): mIoU, R@0.5 and F1@0.5 assess boundary alignment and coverage. F1 uses one-to-one matching at tIoU ≥ 0.5.
Negative recognition: FPR reports hallucination on negatives; **Rejection-F1** balances rejection with non-empty positive coverage.

Rejection-F1 = $2 \times R_{neg} \times C_{pos} / (R_{neg} + C_{pos})$
R_{neg} = fraction of negatives predicted empty; C_{pos} = fraction of positives predicted non-empty. An always-empty model has $C_{pos} = 0$, hence Rejection-F1 = 0.

Release scope: visual-only benchmark; annotations CC-BY-NC-SA-4.0, gated non-commercial academic access. Sources: Video-MME, MLVU, CG-Bench and ≤20 additional YouTube videos; YouTube raw videos are not redistributed.

05 CoMET-AGENT: SEARCH & AGGREGATE

Training-free, frozen components; four MLLM agent roles. Reformulate retrieval as hierarchical search with condition verification and global evidence aggregation.



1 Planner: parse explicit time windows, trim the video and select a duration-adaptive graph profile.
2 Filter: DINOv2 appearance change points build coarse event nodes; discard irrelevant events, then use RAFT motion features for fine action nodes.
3 Verifier: combine candidates with graph neighbors to test all query conditions.
4 Aggregator: consolidate deduplicated memory and resolve cross-moment entity identity before returning intervals and count.

06 VERIFICATION & MEMORY MECHANICS

Context expansion: C(v) = {v} ∪ N_K(v; τ), using direct and indirect temporal neighbors with edge weight > τ.
Selective writes: accepted nodes produce m = <summary, (start, end), confidence>; rejected nodes never enter memory.
Online temporal NMS: when interval overlap exceeds γ = 0.5, keep the higher-confidence entry; otherwise append a new event.
Budget: at most 3 iterations, with early termination when no new candidate is accepted. Aggregation cross-references stored summaries for identity consistency.

07 IMPLEMENTATION & ADAPTIVE PROFILES

Frozen DINOv2-L/14 + RAFT; Qwen3-VL-Embedding-8B for graph edges. Same MLLM instantiates all four roles. Open-source agent setup: 1× A100 80 GB; comparison methods use 4× A40.

Graph parameter	<10 min	10–30 min	>30 min
Event / action fps	2 / 2	1 / 2	0.5 / 1
Event boundary threshold	0.92	0.88	0.82
PELT penalty	0.30	0.50	0.70
Temporal decay / traversal τ	2 / 0.85	3 / 0.82	5 / 0.78
Coarse events / action nodes	3 / 4	4 / 6	6 / 8

08 COMPONENT ABLATIONS

Qwen3-VL variant	F1@0.5	Rej.-F1
Raw backbone	3.4	52.1
w/o adaptive HP configuration	7.5	56.3
w/o temporal graph	6.5	54.9
w/o iterative verification	4.2	52.1
w/o global memory	4.0	53.8
Full CoMET-Agent	**8.8**	**57.4**

09 INFERENCE COST & STAGE PROFILING

Method / backbone	s/query
CoMET / Qwen2.5-VL	27.5
TimeSearch-R / Qwen2.5-VL	49.1
Vgent-adapted / Qwen2.5-VL	87.0
CoMET / GPT-4o	21.3
T*-adapted / GPT-4o	207.8

Stage profile (27.54 s): planner 6.85; graph construction 6.25; verification 12.92 (~47%); aggregator 1.52. Reported wall-clock times depend on hardware/API setup; not hardware-normalized complexity guarantees.

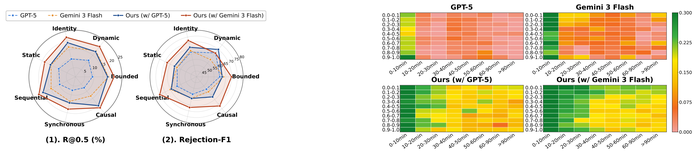
10 MAIN RESULTS: ALL EVALUATED METHODS

Count metrics: MAE↓ / OBO↑ / Pearson↑. Grounding on positives: mIoU↑ / R@0.5↑ / F1@0.5↑. Negative queries: Rej.-F1↑ / FPR↓. All except MAE are reported on the paper's percentage scale.

Method	MAE	OBO	Pear.	mIoU	R@.5	F1@.5	Rej.F1	FPR
GPT-5	3.2	56.8	9.0	9.7	9.5	10.1	60.6	5.3
Gemini 2.5 Pro	3.3	53.6	4.8	10.8	9.9	10.1	60.8	37.3
Gemini 3 Flash	3.4	54.2	6.6	16.5	15.8	14.6	61.5	51.7
LLaVA-Video	3.5	55.3	2.4	0.7	0.3	0.4	48.9	58.5
LLaVA-OV1.5	3.7	47.7	3.9	0.3	0.2	0.2	9.3	2.6
LongVILA-R1	5.7	38.2	11.1	10.8	7.6	4.6	5.6	97.1
MiMO-VL	3.8	48.5	-1.3	0.3	0.1	0.1	8.8	3.8
InternVL3.5	3.6	52.3	3.2	1.9	1.7	1.1	64.7	16.8
Eagle2.5	3.9	48.8	1.0	2.1	0.9	0.7	39.9	14.6
Qwen3-VL	4.8	51.7	9.8	5.1	3.9	3.4	52.1	5.3
Qwen3-VL	3.6	51.5	3.4	3.3	3.0	3.1	47.2	5.9
InternVL3.5	4.0	48.3	13.4	2.6	1.8	1.3	40.3	9.1
VideoMind	4.4	14.8	-0.5	7.6	4.8	2.3	1.2	99.4
Vgent-adapted	6.23	51.9	12.4	2.9	2.6	2.0	48.5	5.4
T*-adapted	3.80	44.2	16.3	1.2	1.0	1.4	34.3	0.0
TimeSearch-R	3.30	48.4	-0.3	8.4	5.5	5.0	26.1	83.7
VideoARM	3.6	52.1	14.5	13.5	12.8	13.0	61.0	9.5
TRACE	3.2	60.9	1.0	4.0	3.3	3.8	2.5	98.8
DisTime	3.3	60.0	-0.6	2.7	0.5	0.5	12.1	93.5
TimeLens	3.3	60.7	0.0	6.2	5.0	5.4	0.0	100.0
LITA	3.7	58.4	0.2	1.8	0.4	0.4	1.7	99.1
Ours (Qwen3-VL)	**3.9**	**54.5**	**11.0**	**9.2**	**8.5**	**8.8**	**57.4**	**8.0**
Ours (GPT-5)	**2.8**	**61.4**	**13.7**	**17.5**	**17.6**	**16.2**	**66.5**	**6.4**
Ours (G3 Flash)	**2.7**	**62.0**	**12.0**	**19.3**	**19.5**	**19.0**	**68.1**	**5.8**

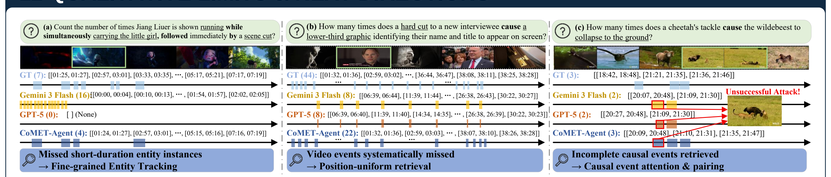
Same-backbone gain: GPT-5 +6.1 pp; Gemini 3 Flash +4.4 pp; Qwen3-VL +5.4 pp in F1@0.5. Best reported F1 remains **19.0%**: this task is far from solved.

11 WHERE GAINS COME FROM — AND FAILURES



Compositional reasoning: GPT-5 sequential R@0.5 improves **7%** → **22%**. Graph verification targets cross-moment conditions.
Duration: baseline mIoU approaches zero beyond 10 min; the agent retains non-trivial performance beyond 90 min across positions.
Scale ≠ grounding: Qwen3-VL 8.3B → 30B reduces MAE 4.8 → 3.6 but F1 drops 3.4 → 3.1 and Rej.-F1 52.1 → 47.2.
Two negative-query failure modes: grounding-specialized models hallucinate (FPR ≥93.5%); LLaVA-OV1.5 / MiMO over-reject (Rej.-F1 9.3 / 8.8) despite low FPR.

12 QUALITATIVE LIMITS & NEXT DIRECTIONS



Entity tracking: short actions and visually similar entities make textual identity summaries ambiguous; integrate persistent object tracklets.
Timeline coverage: only **22 of 44** events recovered in the dense long-video case; use coverage-aware traversal for mid/late regions.
Causal pairing: visually plausible tackles can be retrieved without a resulting collapse; explicitly link cause and consequence nodes.

Beyond this benchmark: audio-visual queries; active-learning-assisted annotation scaling; lightweight verifier/aggregator fine-tuning; joint **(interval, object trajectory)** predictions for spatial-temporal counting and grounding.