

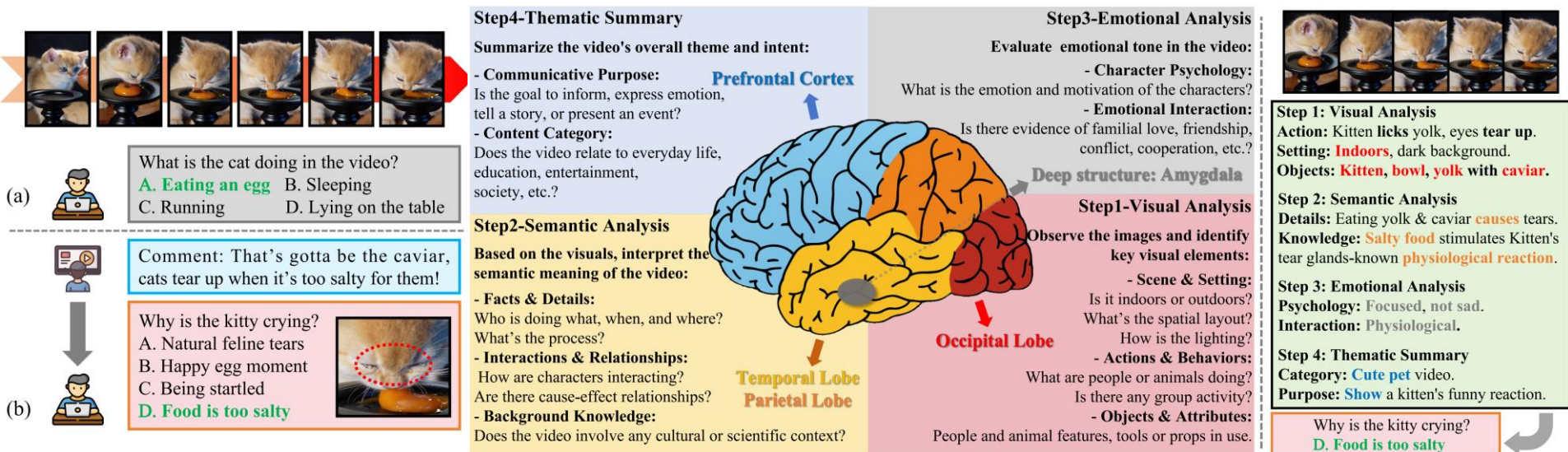
A³Bench: An Audience-Aligned Multilingual Benchmark for Video Audience Insights Understanding

Yiming LEI, Guozhen PENG, Zeming LIU, Hui QIU,
Haitao LENG, Shaoguo LIU, Tingting GAO, Qingjie LIU,
Annan LI, Yunhong WANG

Frontiers of Computer Science, DOI: [10.1007/s11704-026-52159-9](https://doi.org/10.1007/s11704-026-52159-9)

Problems & Ideas

- Problems of conventional MLLMs and video benchmarks:
 - MLLMs often produce responses that are factually correct but misaligned with real audience interests.
 - Existing benchmarks primarily evaluate factual comprehension while neglecting subjective audience insights.
- Ideas: An audience-aligned benchmark and a brain-inspired reasoning framework that take both real-world audience interests and structured cognitive reasoning into account.

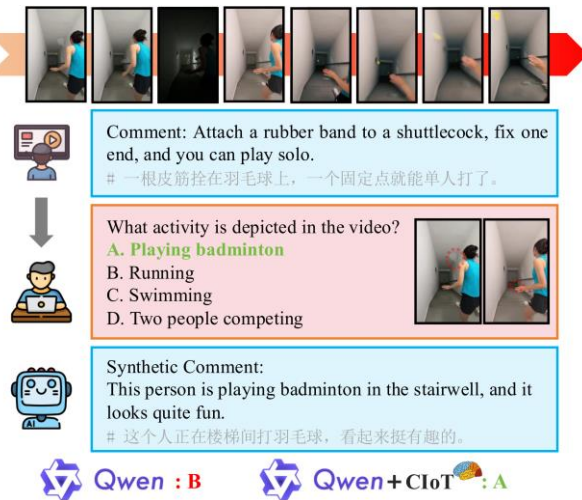


Left: A³Bench shifts from traditional VideoQA to audience-aligned insights by pairing video content with real user comments. Right: CloT emulates the human brain's cognitive workflow for deep reasoning, systematically integrating visual cues, semantic meaning, emotional tone, and high-level thematic intent (Visual → Semantic → Emotional → Thematic).

Main Contributions

- Contributions:
 - A new challenging task shifting from conventional video understanding to audience insights, which reflects a deep comprehension of cognitive processes;
 - A novel audience-aligned benchmark (A³Bench) constructed with large-scale videos and high-quality comments across three languages to evaluate audience insights comprehension;
 - A Cognition Interaction of Thought (CIoT) reasoning framework is proposed to emulate key cognitive processes and enhance MLLMs' comprehension of audience insights.

Model	Lang	Multiple-Choice					True-or-False				
		VS	SE	EM	TH	ALL	VS	SE	EM	TH	ALL
Qwen2.5-VL _{7B}	ZH	72.7	73.8	72.2	78.8	74.4	70.9	60.4	66.2	65.1	65.5
	ID	50.0	72.7	33.3	100.0	58.6	68.2	42.4	54.5	66.7	58.6
	POR	70.6	75.0	65.6	75.8	71.6	57.6	59.7	52.9	68.8	58.3
+Synthetic Comment	ZH	76.9	75.0	87.0	87.6	79.8	71.9	62.8	70.1	61.4	66.5
	ID	60.0	72.7	50.0	100.0	65.5	63.6	47.4	54.6	83.3	58.6
	POR	76.5	72.2	59.4	84.9	74.0	57.6	62.9	61.8	56.3	59.8
+CIoT	ZH	76.2	76.2	87.0	87.6	80.0	69.4	67.5	75.3	69.9	69.4
	ID	60.0	72.7	50.0	100.0	65.5	63.6	57.9	63.6	100.0	65.5
	POR	73.5	77.8	68.8	87.9	76.3	65.2	64.5	50.0	62.5	62.3
InternVL3 _{9B}	ZH	78.3	80.8	77.8	85.0	80.7	65.8	58.4	71.4	61.4	62.9
	ID	60.0	54.5	50.0	100.0	58.6	50.0	36.8	54.5	50.0	46.6
	POR	70.6	72.2	68.8	84.8	73.4	53.3	58.1	38.2	68.8	53.4
+Synthetic Comment	ZH	79.0	76.7	88.9	87.6	81.2	62.8	67.8	68.8	69.9	66.6
	ID	70.0	72.7	50.0	100.0	68.9	50.0	47.4	36.4	83.3	50.0
	POR	77.9	69.4	68.8	87.9	76.3	56.5	54.8	47.1	68.8	55.3
+CIoT	ZH	76.2	78.5	87.0	89.4	81.8	67.3	69.8	77.9	69.9	70.0
	ID	70.0	72.7	50.0	100.0	68.9	59.1	63.2	36.4	83.3	58.6
	POR	79.4	80.6	65.6	90.9	79.3	56.5	61.3	58.8	68.8	59.3



Comment: Attach a rubber band to a shuttlecock, fix one end, and you can play solo.
一根皮筋拴在羽毛球上，一个固定点就能单人打了。

What activity is depicted in the video?
A. **Playing badminton**
B. Running
C. Swimming
D. Two people competing

Synthetic Comment:
This person is playing badminton in the stairwell, and it looks quite fun.
这个人正在楼梯间打羽毛球，看起来挺有趣的。

Qwen : B Qwen + CIoT : A

Performance improvements of MLLMs using synthetic comments and CIoT on A³Bench (Left), alongside a visual example showing how CIoT effectively captures and reasons about continuous video actions (Right).