PIVOTS-Bench: Evaluating Fine-Grained Interpersonal Relationship Reasoning in Multimodal Large Language Models

Shuxiang Zhang* Yiting Yin* Wenxuan Song Yuhang Wu† Miao Liu‡

Sun Yat-sen University (S. Zhang) University of Michigan (Y. Yin) Tsinghua University (W. Song, Y. Wu, M. Liu)

* Co-first authors, equal contribution † Project lead ‡ Corresponding author

Project page

ciossayin.github.io/pivots-bench

Asymmetric perception:
the missing piece in MLLM relationship reasoning.

PIVOTS tests whether multimodal LLMs can perceive a relationship from each person's perspective.

1 Two people, two perceptions

"She likes me, but I don't like her", so we need to rate **each person's view** on every PIVOTS dimension, separately.

2 Social-IQ 2.0: limited scenes

Mostly **in-car selfie videos of friends and family** and **scripted movie clips**.

3 Psychological foundations

Dimensions drawn from **relationship science** (Wish et al., 1976; Hinde, 1976) and **emotion research** (Russell, 1980), toward **social intelligence in multimodal LLMs**.

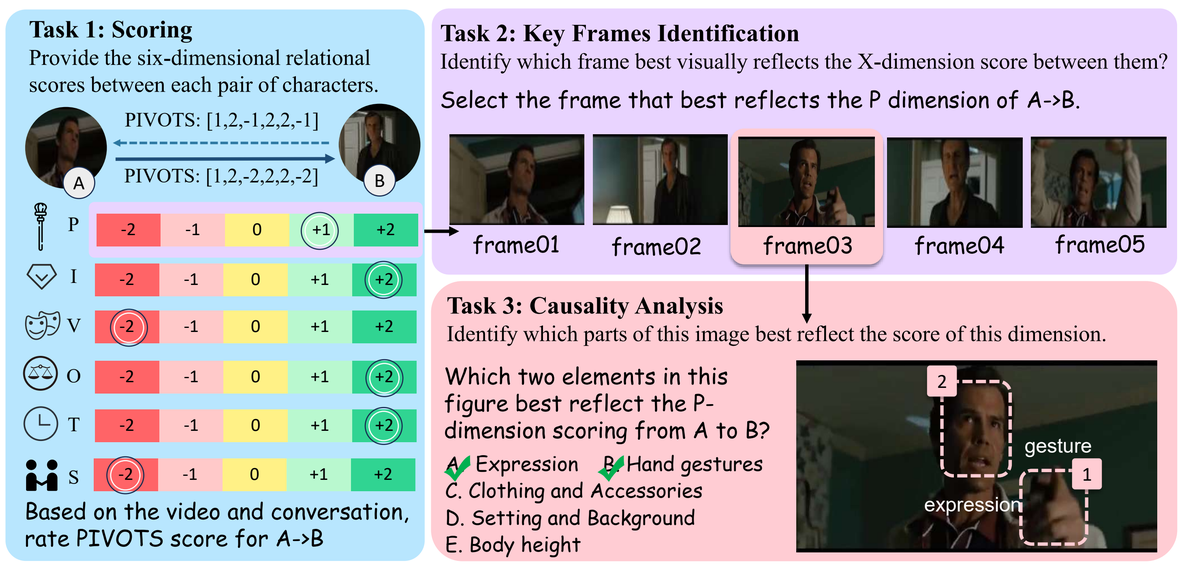
1 The PIVOTS Framework

Wish, Deutsch & Kaplan (1976) found four dimensions people use to perceive relationships. We add **Valence** (affect; Russell, 1980) and **Permanence**, because the same roles play out differently over time (Hinde, 1976). Every level is anchored in behavioral and psychological cues.

Bipolar (Wish et al., 1976) → PIVOTS: graded −2 … +2

P Power	I Involvement	V Valence (NEW)	O Objective	T Permanence (NEW)	S Stance
−2 −1 0 +1 +2	−2 −1 0 +1 +2	−2 −1 0 +1 +2	−2 −1 0 +1 +2	−2 −1 0 +1 +2	−2 −1 0 +1 +2
dominated … dominant	superficial … intense	hostile … pleasant	task … socioemotional	one-time … lifelong	competitive … cooperative

2 Task Definition and Statistics



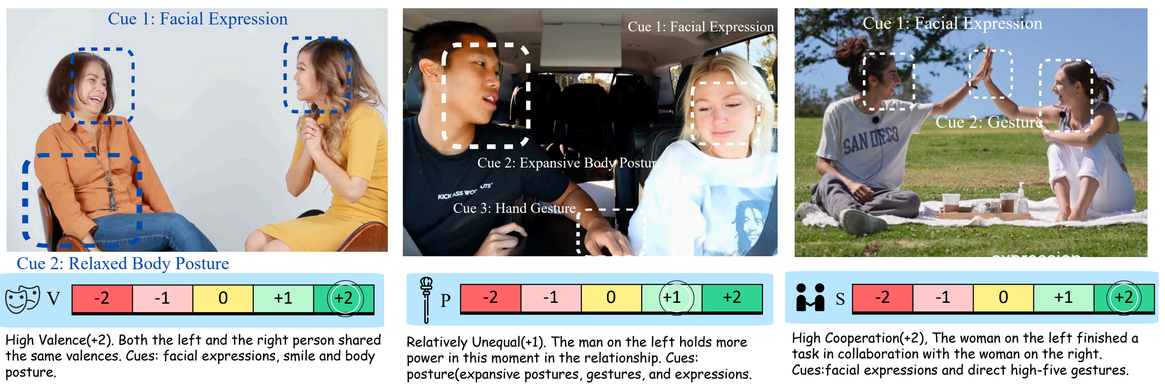
T1 Six-dimensional scoring: rate each person's view — chance 20%

T2 Key-frame identification: pick the frame that shows the score — 20%

T3 Visual-cue causal analysis: pick the 2 cues behind it — 40%

191 clips · 765 ratings per dimension · inter-rater α = .81

3 Demonstration: Cue-Grounded Judgments



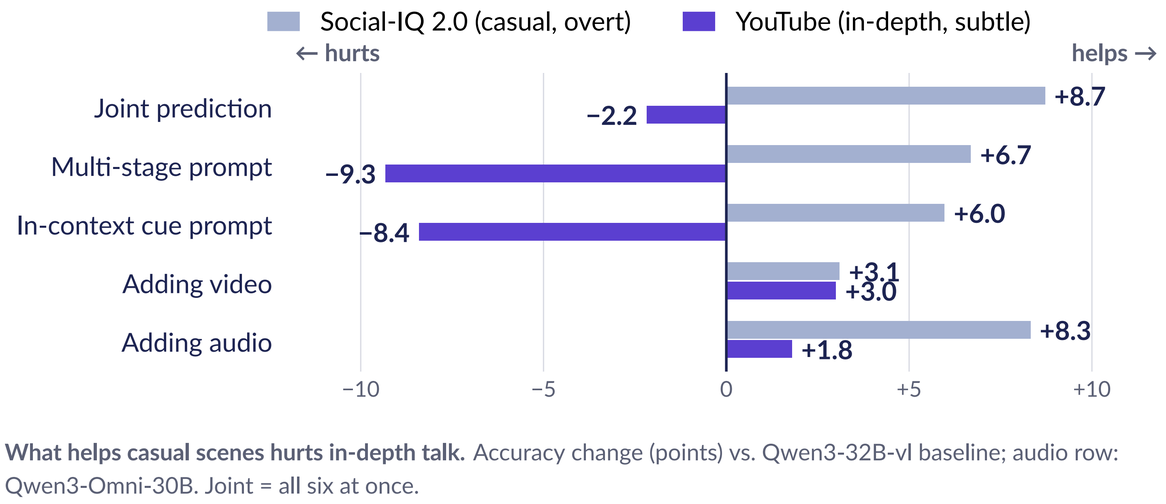
Relationships show in faces, posture and gestures: every rating is tied to the visual cues behind it.

4 Benchmark Results

	T1 Avg	P	I	V	O	T	S	T2	T3
Social-IQ 2.0									
Gemini-2.5-pro	47.7	61.3	50.8	33.8	**51.6**	47.0	41.9	46.6	48.0
GPT-5	**56.9**	**65.1**	**60.3**	**55.7**	42.3	**56.6**	**61.4**	**62.2**	**60.3**
Qwen3-4B-vl	24.3	51.1	20.8	24.0	13.7	25.9	10.5	28.6	30.2
Qwen3-8B-vl	27.9	59.7	22.2	22.9	18.6	26.9	17.1	34.8	31.0
Qwen3-32B-vl	34.5	62.6	23.8	33.8	23.8	30.3	31.5	35.3	33.5
YouTube									
Gemini-2.5-pro	54.7	69.5	**55.1**	31.8	61.0	54.2	56.4		
GPT-5	**58.9**	**73.7**	50.4	**48.3**	**61.1**	**60.7**	**58.9**		
Qwen3-4B-vl	30.0	72.0	16.8	33.6	6.9	29.3	21.6		
Qwen3-8B-vl	35.2	66.7	26.9	31.5	18.5	32.4	35.2		
Qwen3-32B-vl	40.8	67.9	32.4	34.9	29.5	35.9	43.1		

Valence and Objective are the hardest dimensions. Exact-match accuracy (%), **bold** = best per column. T2/T3 annotated on Social-IQ 2.0 only; all Qwen3 models fall below chance on T3. GPT-5 drafted the initial labels, so its scores may be partly inflated.

5 Ablation: Prompting and Input Modalities



What helps casual scenes hurts in-depth talk. Accuracy change (points) vs. Qwen3-32B-vl baseline; audio row: Qwen3-Omni-30B. Joint = all six at once.

6 Takeaways

1 **Scenario-dependent prediction setting.** In the everyday, overt interactions of Social-IQ 2.0, cues across dimensions are complementary, so joint scoring helps; in YouTube's in-depth, interview-style interactions, holistic modeling interferes, and independent scoring better captures dimension-specific cues (joint: +8.7 on Social-IQ 2.0 vs. −2.2 on YouTube).

2 **Relationships are conveyed by global context.** In YouTube clips, relationships emerge mainly from the overall conversational context and shared background rather than salient local cues (audio: +8.3 vs. +1.8).

3 **No prompting strategy generalizes.** The same strategy can help in one setting and hurt in another (multi-stage: +6.7 vs. −9.3), so MLLMs should be evaluated on real, in-depth relationships.

Contributions: Y. Yin designed the PIVOTS framework and rating rubric and built the video dataset; S. Zhang led annotation and experiments. Poster by Y. Yin.

Contact

Yiting Yin
Presenter · University of Michigan
ciossayi@umich.edu

Open to collaboration on **social intelligence** and **human-centered AI**.

WeChat

EMNLP 2026 Main Conference · Budapest, Hungary