Bad Idea or Good Prediction? Comparing VLM and Human Anticipatory Judgment

Anonymous Author(s)

Abstract

Anticipatory reasoning – predicting whether situations will resolve positively or negatively by interpreting contextual cues – is crucial for robots operating in human environments. This exploratory study evaluates whether Vision Language Models (VLMs) possess such predictive capabilities through two complementary approaches. First, we test VLMs on direct outcome prediction by inputting videos of human and robot scenarios with outcomes removed, asking the models to predict whether situations will end well or poorly. Second, we introduce a novel evaluation of anticipatory social intelligence: can VLMs predict outcomes by analyzing human facial reactions of people watching these scenarios? We tested multiple VLMs, including closed-source and open-source, using various prompts and compared their predictions against both true outcomes and judgments from 29 human participants. The best-performing VLM (Gemini 2.0 Flash) achieved 70.0% accuracy in predicting true outcomes, outperforming the average individual human (62.1% ± 6.2%). Agreement with individual human judgments ranged from 44.4% to 69.7%. Critically, VLMs struggled to predict outcomes by analyzing human facial reactions, suggesting limitations in leveraging social cues. These preliminary findings indicate that while some VLMs show promise for anticipatory reasoning, their performance is sensitive to model selection and prompt design, with current limitations in social intelligence that warrant further investigation for human-robot interaction applications.

Keywords

social competence; robot error; anticipation; VLMs; human-AI collaboration

ACM Reference Format:

Anonymous Author(s). 2026. Bad Idea or Good Prediction? Comparing VLM and Human Anticipatory Judgment. In *Proceedings of A ACM/IEEE International Conference on Human-Robot Interaction (HRI) (HRI'26).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Vision-Language Models (VLMs) are increasingly being deployed in human-robot interaction systems, with applications ranging from task planning and navigation to manipulation and multimodal perception [4, 7, 13]. These models enable robots to interpret contextual cues and execute tasks aligned with human intentions by processing and correlating visual data from cameras with linguistic inputs. As robots become more prevalent in shared human spaces, recent works have focused on leveraging these models for social intelligence [3, 5, 6, 12]. For example, in Sasabuchi et al. [10], a combination of VLMs for scene understanding and LLMs for judgment performed well on a task of deciding when to initiate interaction.

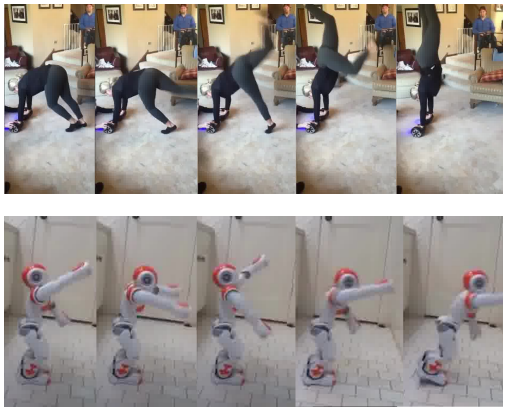

Figure 1: In Parreira et al. [9], human participants were shown videos displaying a variety of scenarios featuring robots and humans. Videos ended before the outcome was displayed, and participants were asked to anticipate the outcome of each video. The same video dataset was used in this study.

However, a critical gap remains in understanding whether VLMs possess **anticipatory reasoning** capabilities for HRI contexts. We frame anticipatory reasoning as a predictive process: observing an unfolding scenario and forecasting its outcome by interpreting contextual cues. While anticipatory thinking has been explored in language models for task planning [8, 11, 14], whether VLMs can perform similar predictive reasoning in visual scenarios – particularly those involving social contexts where outcomes depend on human behavior, intentions, and norms, which we deem as *anticipatory social intelligence* – remains unexplored. This capability is fundamental to human social cognition [1], enabling us to constantly evaluate whether actions are "wise" by processing environmental and social cues. For robots operating in shared spaces, such predictive capabilities are particularly critical in safety-sensitive scenarios where systems must proactively prevent errors rather than reactively respond to them.

This paper presents an **exploratory evaluation of VLM anticipatory reasoning capabilities** using two complementary approaches. Drawing on the "*Bad Idea?*" human study from Parreira et al. [9], we first evaluated how well VLMs predict scenario outcomes by directly analyzing videos with outcomes removed, comparing their predictions against both ground truth and human judgments from 29 participants. Critically, we then introduced a novel evaluation: can VLMs leverage *human anticipatory reactions*

HRI'26, Edinburgh, Scotland, UK

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

Figure 2: Research questions and datasets used in this exploratory study. In Parreira et al. [9], humans observe scenarios and predict outcomes. RQ1 investigates VLMs performance on the same task, and RQ2 investigates the ability to predict outcomes based on human anticipatory reactions to the scenarios.

to predict outcomes? Rather than analyzing scenarios directly, we tested whether VLMs can interpret human facial expressions and reactions while people watch these videos, using those social cues to infer their anticipated outcome. This second approach probes anticipatory social intelligence – the ability to interpret human anticipatory states and use them as predictive signals.

Our investigation reveals several preliminary findings that, while non-comprehensive, contribute an important datapoint as VLMs become increasingly deployed in HRI systems. We found that: (1) some VLMs can exceed average human performance on direct outcome prediction, though with high sensitivity to model and prompt configuration; (2) open-source models substantially underperformed closed-source alternatives, with some exhibiting severe prediction bias; and (3) when analyzing human facial reactions to predict outcomes, VLMs show low performance, suggesting limitations in leveraging social cues. These exploratory findings highlight critical limitations – including brittleness to prompt variation and potential gaps in social intelligence – that warrant careful consideration as these models are integrated into safety-critical HRI applications.

2 Methods

We describe the dataset, models and methods used in this study, which tests the *anticipatory reasoning* (ability to predict outcomes based on anticipatory context) of VLMs. We investigated this across two datasets: the **scenario dataset** includes a set of videos which display a scenario, but the outcome is cut off; the **reactions dataset** is composed of human participant reactions to those videos, and respective outcome prediction (Figure 2). For brevity, we only report top performing methods. Additional methodology and results can be consulted on the study repository [1].

We explored the following research questions:

- **RQ1:** What are the anticipatory reasoning capabilities of VLMs (i.e., predicting outcomes on the *scenario dataset*)?
- **RQ2:** What is the anticipatory social intelligence of VLMs (i.e., predicting anticipated outcomes based on the *reactions dataset*)?

2.1 Dataset

User study: In Parreira et al. [9], the authors conducted an online study to collect anticipatory human reactions to stimulus videos through Prolific, in line with previous work [2]. Participants were shown a series of scenarios through short videos. The protocol included a "warm-up" round of 3 videos, followed by a main round of data collection (30 videos). Each video was shown twice: first, a shorter version of the video, which stopped before the video action was fully resolved (e.g., someone swinging from a rope, approaching a tree branch). After watching this video, participants were asked *"You think this situation ends..."* with the options "well" or "poorly". After this, participants were able watch a longer version of the video, featuring the resolution of the video action. Full details on the study protocol, participants and ethics can be found in the original paper.

Stimulus dataset: A set of 30 videos, which includes scenarios where humans and robots are featured and there are good and bad outcomes, was used (Figure 1). The full list can be found on Parreira et al. [9] study repository [2].

2.2 Prompts

We tested multiple prompts to evaluate VLMs' anticipatory reasoning capabilities. For direct outcome prediction on the *scenario dataset*, we employed two primary prompts that directly query the model's judgment:

Prompt A (Direct): *"Given the scenario shown on the video, you think this situation ends well or poorly? (Use only one word to answer)"*

Prompt B (Human-like): *"Given the scenario shown on the video, you think this situation ends well or poorly as if you are a human watching the video? (Use only one word to answer)"*

For RQ2, which evaluates outcome prediction based on human anticipatory reactions, we adapted the prompts to focus on interpreting facial expressions and reactions rather than the scenario itself.

Prompt C (Reaction-based): *"Given the human reaction shown in the image, do you think the situation observed by that human ends well or poorly? (Use only one word – well or poorly – to answer)"*

[1] ANONYMOUS

[2] https://github.com/IRL-CT/badidea

Table 1: VLM performance compared to ground truth outcomes, for Prompts A and B. Human baseline shown as $M \pm SD$ across 29 participants. PR: ratio of "Poorly" predicted outcomes to total predictions (ground truth ratio is 0.433). Best aggregation method for open-source models was MODE.

Model	Acc	Prec	Rec	F1	PR
Closed-Source VLMs					
GPT-4o (A)	0.433	0.375	0.462	0.414	0.467
GPT-4o (B)	0.467	0.400	0.462	0.429	0.500
Qwen (A)	0.500	0.450	0.692	0.545	0.333
Qwen (B)	0.533	0.476	0.769	0.588	0.300
Gemini (A)	**0.700**	0.625	0.769	0.690	0.467
Gemini (B)	0.633	0.562	0.692	0.621	0.467
Open-Source VLMs					
LLaVA-Llama3 (A)	0.567	0.500	0.231	0.316	0.200
LLaVA-Llama3 (B)	**0.633**	0.750	0.231	0.353	0.133
DeepSeek-OCR (A)	0.567	0.000	0.000	0.000	0.000
DeepSeek-OCR (B)	0.567	0.000	0.000	0.000	0.000
Gemma3 (A)	0.433	0.433	1.000	0.605	1.000
Gemma3 (B)	0.433	0.433	1.000	0.605	1.000
Human Average	0.621 ±0.062	0.575 ±0.086	0.599 ±0.091	0.579 ±0.056	0.433

2.3 Models Tested

We evaluated both closed-source and open-source VLMs to assess anticipatory reasoning capabilities across different deployment scenarios. Closed-source commercial models (accessed via API) typically demonstrate stronger performance due to extensive fine-tuning and larger training resources. However, open-source models offer critical advantages for HRI applications: they can be deployed locally for privacy-sensitive data (such as participant facial reactions), operate without internet connectivity, and provide accessible alternatives for resource-constrained deployments. We used Ollama [3] to run local models. Since many VLMs are designed for image input rather than video, we employed frame extraction strategies detailed below.

2.3.1 RQ1: Anticipatory reasoning of VLMs. For evaluating anticipatory judgement of commercial VLMs on the scenario dataset, we tested three state-of-the-art closed-source models on scenario *videos*: GPT-4o, Gemini 2.0 Flash, Qwen2.5-vl (72b). We also tested publicly available models, that can be deployed locally, namely: DeepSeek-OCR (3b), Gemma 3 (4b), LLaVA-LLama 3 (8b).

The open-source models process images, not video. Thus, we extracted the *frames* from each video (10 fps) and obtained predictions for each. The final prediction was determined by the **mode** (most common prediction) across all extracted frames within the window or the prediction from the **last** frame in the sequence.

2.3.2 RQ2: Anticipatory Social Intelligence (Reactions Dataset). For evaluating VLMs' ability to interpret human anticipatory reactions, we employed the open-source models from RQ1 on the *reactions dataset*. These models analyzed recordings of participants facial expressions and reactions while watching scenario videos, with the task of predicting their anticipated outcome. We tested two context window configurations: **1 second** and **3 seconds** before the reaction video ended, with frames extracted at 2 fps.

[3] https://ollama.com/

Table 2: VLM agreement with individual human predictions, for Prompts A and B. Values shown as $M \pm SD$ across comparisons with 29 participants.

Model	Acc	Prec	Rec	F1
Closed-Source VLMs				
GPT-4o (A)	0.489 ± 0.076	0.457 ± 0.126	0.522 ± 0.091	0.482 ± 0.105
GPT-4o (B)	0.444 ± 0.072	0.408 ± 0.102	0.440 ± 0.074	0.418 ± 0.082
Qwen (A)	0.605 ± 0.087	0.553 ± 0.125	0.794 ± 0.071	0.644 ± 0.101
Qwen (B)	0.639 ± 0.087	0.574 ± 0.119	0.871 ± 0.077	0.684 ± 0.095
Gemini (A)	**0.697±0.070**	0.651 ± 0.129	0.755 ± 0.076	0.691 ± 0.087
Gemini (B)	0.692 ± 0.069	0.647 ± 0.130	0.750 ± 0.074	0.686 ± 0.087
Open-Source VLMs				
LLaVA-Llama3 (A)	0.522 ± 0.083	0.469 ± 0.189	0.201 ± 0.075	0.278 ± 0.102
LLaVA-Llama3 (B)	**0.558±0.090**	0.586 ± 0.248	0.170 ± 0.071	0.261 ± 0.107
DeepSeek-OCR (A)	0.535 ± 0.097	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
DeepSeek-OCR (B)	0.535 ± 0.097	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
Gemma3 (A)	0.465 ± 0.097	0.465 ± 0.097	1.000 ± 0.000	0.628 ± 0.095
Gemma3 (B)	0.465 ± 0.097	0.465 ± 0.097	1.000 ± 0.000	0.628 ± 0.095

2.4 Evaluation Metrics

We evaluated model performance using standard binary classification metrics: accuracy, precision, recall, and f1-score, as well as prediction ratio (ratio of outcomes predicted as "Poorly" to total predicted outcomes), to understand if there are any biases in predictions.

We computed two types of evaluations:

(1) **Ground Truth Agreement:** Model predictions compared against actual video outcomes; we include human performance on the same task from Parreira et al. [9] for reference
(2) **Human Agreement:** Model predictions compared against individual human participant judgments, i.e. the outcomes predicted by each participant (reported as mean ± standard deviation across all participants)

3 Results

RQ1: Anticipatory reasoning of VLMs We evaluated the VLMs using two prompt variants on the scenario dataset. Table 1 presents model performance against ground truth outcomes, while Table 2 shows agreement with individual human predictions.

Key findings from RQ1 include: **(1)** the best-performing closed-source VLM (Gemini 2.0 Flash with Prompt A) achieved 70.0% accuracy on predicting true outcomes, exceeding the average human performance (62.1% ± 6.2%), though performance varied substantially across models (43.3% to 70.0%), **(2)** open-source models demonstrated competitive but generally lower performance, with the best configuration (LLaVA-Llama3 with Prompt B) reaching 63.3% accuracy and approaching human-level performance; some models exhibited severe prediction bias (e.g., DeepSeek-OCR and Gemma3 which predicted only one type of outcome), **(3)** prompt engineering impacted performance even within the same model, with variations up to 6.7 percentage points (Gemini Prompt A vs. B) for commercial models and **(4)** models showing higher agreement with individual human predictions did not necessarily achieve

Table 3: VLM performance on predicting outcomes from human anticipatory reactions. Performance represents agreement with human predictions (not ground truth). Values shown as $M \pm SD$ across 29 participants. Best aggregation method (MODE or LAST) shown for each model. PR: ratio of "Poorly" predicted outcomes to total predictions (human baseline PR is 0.464).

Model	Window	Acc	Prec	Rec	F1	Method	PR
DeepSeek-OCR	1s	0.535 ± 0.098	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	MODE	0.000
DeepSeek-OCR	3s	**0.538±0.104**	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	MODE	0.000
LLaVA-Llama3	1s	0.462 ± 0.101	0.411 ± 0.176	0.510 ± 0.285	0.425 ± 0.188	MODE	0.549
LLaVA-Llama3	3s	0.479 ± 0.136	0.444 ± 0.192	0.524 ± 0.293	0.439 ± 0.186	LAST	0.538
Gemma3	1s	0.450 ± 0.104	0.454 ± 0.106	0.945 ± 0.100	0.608 ± 0.112	MODE	0.969
Gemma3	3s	0.445 ± 0.106	0.451 ± 0.109	0.946 ± 0.111	0.603 ± 0.115	LAST	0.968

higher accuracy on true outcomes, suggesting that humans and models may share similar biases or systematic errors in anticipatory reasoning.

RQ2: Anticipatory Social Intelligence We evaluated open-source VLMs' ability to predict outcomes by analyzing human facial reactions while watching scenario videos. Table 3 shows model alignment *with human predictions* – the model prediction is mapped to what that participant predicted, not the ground truth outcome, as we want to test the models ability to map anticipatory reactions to respective anticipated outcomes.

Key findings from RQ2 include: **(1)** VLMs performed poorly at predicting outcomes from human anticipatory reactions, with accuracy ranging from 44.5% to 53.8% **(2)** several models exhibited severe prediction bias, predicting "poorly" or "well" for nearly all scenarios, suggesting difficulty in discriminating between different outcome types, **(3)** frame aggregation method (MODE vs LAST) impacted performance, and **(4)** no significant performance difference emerged between 1-second and 3-second temporal windows across models. These results hint at limitations in current VLMs' ability to interpret human facial expressions and anticipatory reactions as predictive signals, suggesting a critical gap in social intelligence capabilities beneficial for human-robot interaction.

4 Discussion

This exploratory study provides initial evidence that some VLMs can perform anticipatory reasoning – the predictive process of forecasting outcomes from contextual cues – with potential applications for safety-critical HRI scenarios. Gemini 2.0 Flash achieved 70.0% accuracy on scenario prediction, exceeding average human performance (62.1%), while the best open-source model (LLaVA-Llama3) approached human-level performance at 63.3%. However, considerable fragility emerged, with performance varying up to 26.7 percentage points across models and up to 6.7 percentage points across prompts within some models. To what extent this capability of sophisticated pattern matching on visual cues can genuinely function as applied anticipatory reasoning remains an open question, requiring larger and more diverse datasets.

The performance gap between closed-source (70.0%) and open-source (63.3% at best) models is unsurprising given differences in scale, training resources, and fine-tuning. Several open-source models exhibited severe prediction bias, defaulting to single outcome predictions, suggesting fundamental challenges in discriminating between outcome types. However, given the rapid evolution of open-source VLMs and the exploratory nature of this work, these observations represent snapshots rather than definitive assessments of model capabilities.

Further, we note VLMs' apparent difficulty with anticipatory social intelligence (RQ2), where models analyzing human facial reactions achieved only 47.9% agreement with human predictions, substantially below both the direct scenario prediction performance (63-70%) and demonstrating limitations in leveraging social cues (note that DeepSeek-OCR performance is higher only due to imbalance in the dataset ground truth, as it only predicted one type of label, "well"). This finding, while preliminary, suggests a critical gap in VLMs' ability to interpret human anticipatory states and interpret subtle emotional signals – capabilities essential for effective HRI. If replicated in larger studies, this limitation could significantly constrain VLM utility in collaborative settings where robots must understand and respond to human anticipatory behaviors.

These observations contribute an important datapoint as VLMs become increasingly integrated into HRI systems. Future work can explore hybrid approaches combining scenario and reaction information, or VLM-LLM pipelines [10], potentially leveraging complementary strengths of visual understanding and reasoning capabilities. Such investigations could help establish whether current limitations reflect fundamental performance constraints of current models or suboptimal system configurations.

Limitations. This work is explicitly exploratory with several constraints on interpretation. Our evaluation uses only 30 videos, insufficient for generalizable conclusions. The human baseline comes from a specific online population and may not represent broader human performance. We tested a limited selection of models and only a handful of prompt formulations; different model versions or systematic prompt engineering might substantially change results. Critically, we lack a human baseline for RQ2 (how accurately humans predict outcomes from others' reactions), preventing direct performance comparison. Our frame aggregation analysis examined only two simple strategies (mode and last frame) without testing different temporal horizons or sophisticated aggregation methods; these choices have important computational implications for real-time deployment, and more sophisticated approaches could yield different patterns. Finally, our binary outcome framing (well/poorly) oversimplifies real-world anticipatory reasoning. These results should be viewed as preliminary observations that motivate, rather than answer, questions about VLM anticipatory capabilities.

References

[1] A. Amos-Binks and D. Dannenhauer. Anticipatory thinking: A metacognitive capability, 2019. URL https://arxiv.org/abs/1906.12249.

[2] A. Bremers, M. T. Parreira, X. Fang, N. Friedman, A. Ramirez-Aristizabal, A. Pabst, M. Spasojevic, M. Kuniavsky, and W. Ju. The bystander affect detection (bad) dataset for failure detection in hri, 2023.

[3] F. Bu, M. Tsai, A. Tjokro, T. Bhattacharjee, J. Ortiz, and W. Ju. Using vision-language models as proxies for social intelligence in human-robot interaction, 2025. URL https://arxiv.org/abs/2512.07177.

[4] J. Fan, Y. Yin, T. Wang, W. Dong, P. Zheng, and L. Wang. Vision-language model-based human-robot collaboration for smart manufacturing: A state-of-the-art survey. *Frontiers of Engineering Management*, 12:177–200, 03 2025. doi: 10.1007/s42524-025-4136-9.

[5] X. Fan, X. Zhou, C. Jin, K. Nottingham, H. Zhu, and M. Sap. Somi-tom: Evaluating multi-perspective theory of mind in embodied social interactions, 2025. URL https://arxiv.org/abs/2506.23046.

[6] G. Hou, W. Zhang, Y. Shen, Z. Tan, S. Shen, and W. Lu. Egosocialarena: Benchmarking the social intelligence of large language models from a first-person perspective, 2025. URL https://arxiv.org/abs/2410.06195.

[7] S. Liu, J. Zhang, R. X. Gao, X. Vincent Wang, and L. Wang. Vision-language model-driven scene understanding and robotic object manipulation. In *2024 IEEE 20th International Conference on Automation Science and Engineering (CASE)*, pages 21–26, 2024. doi: 10.1109/CASE59546.2024.10711845.

[8] Z. Liu, H. Hu, S. Zhang, H. Guo, S. Ke, B. Liu, and Z. Wang. Reason for future, act for now: A principled framework for autonomous llm agents with provable sample efficiency, 2024. URL https://arxiv.org/abs/2309.17382.

[9] M. T. Parreira, S. G. Lingaraju, A. Ramirez-Artistizabal, A. Bremers, M. Saha, M. Kuniavsky, and W. Ju. "bad idea, right?" exploring anticipatory human reactions for outcome prediction in hri. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, pages 2072–2078, 2024. doi: 10.1109/RO-MAN60168.2024.10731310.

[10] K. Sasabuchi, N. Wake, A. Kanehira, J. Takamatsu, and K. Ikeuchi. Agreeing to interact in human-robot interaction using large language models and vision language models, 2025. URL https://arxiv.org/abs/2503.15491.

[11] H. Wang, T. Li, Z. Deng, D. Roth, and Y. Li. Devil's advocate: Anticipatory reflection for llm agents, 2024. URL https://arxiv.org/abs/2405.16334.

[12] T. Williams, C. Matuszek, R. Mead, and N. Depalma. Scarecrows in oz: The use of large language models in hri. *J. Hum.-Robot Interact.*, 13(1), Jan. 2024. doi: 10.1145/3606261. URL https://doi.org/10.1145/3606261.

[13] Z. Yu, B. Wang, P. Zeng, H. Zhang, J. Zhang, L. Gao, J. Song, N. Sebe, and H. T. Shen. A survey on efficient vision-language-action models, 2025. URL https://arxiv.org/abs/2510.24795.

[14] Q. Zhao, S. Wang, C. Zhang, C. Fu, M. Q. Do, N. Agarwal, K. Lee, and C. Sun. Antgpt: Can large language models help long-term action anticipation from videos?, 2024. URL https://arxiv.org/abs/2307.16368.