Do Vision–Language Models Understand Human Engagement in Games?

Anonymous ACL submission

Abstract

Inferring human engagement from gameplay video is important for game design and player-experience research, yet it remains unclear whether vision–language models (VLMs) can infer such latent psychological states from visual cues alone. Using the GameVibe Few-Shot dataset across nine first-person shooter games, we evaluate three VLMs under six prompting strategies, including zero-shot prediction, theory-guided prompts grounded in Flow, GameFlow, Self-Determination Theory, and MDA, and retrieval-augmented prompting. We consider both pointwise engagement prediction and pairwise prediction of engagement change between consecutive windows. Results show that zero-shot VLM predictions are generally weak and often fail to outperform simple per-game majority-class baselines. Memory- or retrieval-augmented prompting improves pointwise prediction in some settings, whereas pairwise prediction remains consistently difficult across strategies. Theory-guided prompting alone does not reliably help and can instead reinforce surface-level shortcuts. These findings suggest a perception–understanding gap in current VLMs: although they can recognize visible gameplay cues, they still struggle to robustly infer human engagement across games.

1 Introduction

Player engagement is a central construct in game design, live streaming, esports analytics, and interactive content creation (Boyle et al., 2012). Understanding when and why players become engaged—or disengaged—has direct implications for adaptive game design, automated highlight detection, and player retention modeling.

Recent VLMs (OpenAI, 2023; Gemini Team, Google, 2023; Bai et al., 2023) have demonstrated strong visual perception, but these capabilities do not always translate into robust task performance (Fu et al., 2025; Zhang et al., 2025). A fundamentally different question is whether VLMs can infer how a human feels about what they see—a psychological state requiring contextual reasoning about challenge, progress, and narrative tension.

Games provide an ideal testbed because they decouple visual intensity from engagement. A scoreboard showing a close match requires multi-step reasoning (recognizing the score → understanding competitiveness → inferring tension), while an explosion-filled frame may look intense but correspond to a routine, low-engagement moment.

In this paper, we ask: **Do current VLMs understand human engagement in games?** Specifically, we investigate three research questions:

- **Q1:** Can VLMs predict engagement from gameplay frames, and do few-shot demonstrations or retrieval strategies improve prediction?
- **Q2:** Do theory-aligned probes (Flow, GameFlow, SDT, MDA) help VLMs capture engagement-related dimensions?
- **Q3:** Are VLM predictions consistent across different games and across different models?

Despite growing interest in VLM-based affect recognition, no prior work has systematically evaluated whether VLMs can predict human engagement from gameplay video, nor investigated how prompting strategies—including theory-grounded probing and cross-game retrieval—affect this capability (Table 1). We make the following contributions:

- **Benchmark.** We present the first systematic evaluation of VLMs on engagement prediction in gameplay videos, testing three architecturally distinct VLMs across 9 games under six prompting strategies of increasing complexity.
- **Methodology.** We introduce theory-guided prompting grounded in four engagement frameworks (Flow, GameFlow, SDT, MDA) and cross-game retrieval-augmented strategies using both VLM-derived and CLIP embeddings, enabling structured comparison of how prompting complexity affects affective inference.
- **Findings.** We identify a fundamental perception–

understanding gap: VLMs achieve only ∼57% zero-shot accuracy (below majority-class baselines), conflate visual intensity with engagement, rely on surface-feature shortcuts, and generate confidently wrong rationales. Few-shot demonstrations provide substantial gains (+18pp), while theory guidance alone amplifies biases rather than compensating for them.

We will release all code, prompts, and evaluation scripts to facilitate reproducibility and future research.[1]

2 Background & Related Work

Table 1 positions our work relative to the most related prior studies. Existing VLM evaluations either target non-game affect recognition without game-specific contexts (Lu et al., 2024; Bhattacharyya and Wang, 2025; Liu et al., 2025), or evaluate VLMs in game domains on perceptual tasks (action recognition, game completion) rather than affective inference (Zhang et al., 2025; Paglieri et al., 2024). Conversely, game engagement prediction work has relied on supervised models without evaluating VLMs (Pinitas et al., 2024), or used text-only LLMs without visual input (Melhart et al., 2025). Our work is the first to combine all seven dimensions: visual input, VLM evaluation, game domain, engagement prediction, theory-grounded probes, cross-game transfer, and systematic failure analysis.

2.1 Engagement in Games

Engagement is a multi-dimensional, subjective, and temporally dynamic quality of the user experience (O'Brien and Toms, 2008; Brockmyer et al., 2009): unlike objective visual properties, it is a latent psychological state that must be inferred rather than observed. In games, engagement manifests as a high-arousal, positive-valence state (Juvrud et al., 2022), and multiple studies have modeled it at the moment-to-moment level (Sweetser and Wyeth, 2005; Melhart et al., 2020; Barthet et al., 2024; Pinitas et al., 2023); see Yannakakis and Melhart (2023) for a comprehensive survey.

2.2 LLMs and VLMs for Affect Modelling

From text to vision. LLM-based affective computing has focused primarily on text-based sentiment analysis (Zhang et al., 2024; **?**). Multimodal affect modelling from visual input remains largely unexplored: existing work relies on still images (Lian et al., 2024; Yang et al., 2024; Lu et al., 2024) or evaluates whole scenes rather than time-continuous labelling (Zhao et al., 2025b).

VLM affect evaluation. VLMs consistently lag behind supervised models on affect recognition and exhibit systematic biases (Lu et al., 2024; Bhattacharyya and Wang, 2025; Liu et al., 2025), though careful prompt engineering can partially close this gap (Etesam et al., 2024)—motivating our systematic comparison of six prompting strategies.

Fundamental limitations. Pre-trained LLMs struggle with in-context learning for affect prediction, remaining anchored to knowledge priors (Chochlakis et al., 2024), and exhibit inherent biases on subjective tasks (Mao et al., 2023; Amin et al., 2024). Our work extends Melhart et al. (2025)'s evaluation paradigm by introducing theory-grounded probes, cross-game few-shot transfer with multiple embedding strategies, and systematic failure-mode analysis.

2.3 Theoretical Frameworks as Probes

We use four established engagement theories not as assumptions about engagement, but as structured probing tools to test whether VLMs can extract relevant latent dimensions. Flow Theory (Csikszentmihalyi, 1990) captures the challenge–skill balance; GameFlow (Sweetser and Wyeth, 2005) adapts it for games with dimensions of control, immersion, and social interaction; Self-Determination Theory (SDT) (Ryan and Deci, 2000) targets competence, autonomy, and relatedness; and the MDA Framework (Hunicke et al., 2004) decomposes the experience into mechanics, dynamics, and aesthetics. Each theory yields concrete visual indicators that we operationalize as prompt dimensions in our theory-guided strategies (Section 3.4.2).

2.4 VLMs and Video Understanding

VLMs excel at perceptual tasks such as VQA (Antol et al., 2015) and action recognition (Simonyan and Zisserman, 2014), yet whether they can perform affective reasoning—inferring how a human would feel about a visual scene—remains largely untested (Choi et al., 2023; Kamath et al., 2023).

VLMs in game domains. Recent benchmarks consistently reveal VLM limitations in games:

[1] Code and prompts will be made publicly available upon acceptance.

Table 1: Comparison of our work with the most related prior studies across key evaluation dimensions.

Study	Visual Input	VLM Evaluation	Game Domain	Engagement Task	Theory Probes	Cross-Game Transfer	Failure Analysis
Melhart et al. (2025)	✗	✗	✓	✓	✗	✓	✗
Bhattacharyya and Wang (2025)	✓	✓	✗	✗	✗	✗	✓
Lu et al. (2024)	✓	✓	✗	✗	✗	✗	✗
Pinitas et al. (2024)	✓	✗	✓	✓	✗	✓	✗
Zhang et al. (2025)	✓	✓	✓	✗	✗	✗	✗
Paglieri et al. (2024)	✓	✓	✓	✗	✗	✗	✗
Liu et al. (2025)	✓	✓	✗	✗	✗	✗	✗
Ours	✓	✓	✓	✓	✓	✓	✓

near-zero game completion rates (Zhang et al., 2025), near-random performance from screenshots (Hu et al., 2025), and sometimes worse performance with visual input than text-only descriptions (Paglieri et al., 2024). Our work shows the problem is even more severe when the task requires affective inference beyond perception (Table 1).

3 Experimental Setup

3.1 Task Definition

Given a sequence of n frames $\{f_1, \ldots, f_n\}$ sampled from a 1-second gameplay window, the task is to predict the corresponding binary engagement label $y \in \{\text{High}, \text{Low}\}$, where the ground truth is derived from aggregated human annotations. We evaluate two complementary settings: *pointwise prediction*, where the model predicts the engagement level of each window independently, and *pairwise prediction*, where the model compares two consecutive windows and predicts whether engagement increased or decreased. This task is fundamentally different from standard visual recognition: engagement is a latent psychological state that cannot be directly observed, requiring the model to reason about challenge, progress, and narrative context beyond surface-level visual features.

3.2 Dataset and Labels

We use the public *GameVibe Few-Shot (GVFS)* dataset (Barthet et al., 2024), derived from the GameVibe corpus of 120 audio-visual clips across 30 FPS games. Each clip is annotated by 5 participants (from a pool of 20) who provide continuous engagement traces via the RankTrace interface (Lopes et al., 2017) on the PAGAN platform (Melhart et al., 2019). Following the GameVibe processing pipeline, raw traces are interpolated into 1-second windows and normalized to $[0, 1]$ before aggregation.

Temporal windows and frame sampling. Following Pinitas et al. (2024), we split each 61-second clip into non-overlapping 1-second windows. To account for human reaction lag, we shift the input by 1 second relative to the annotations, yielding 59 valid evaluation windows per clip (window t is paired with annotation $t{+}1$). From each 1-second window we sample 16 frames (i.e., 16 fps), giving 944 frames per game clip. All frames within a window share the same engagement label; the evaluation unit throughout this paper is the 1-second window.

Annotation aggregation and binarization. For each window, we apply trimmed-mean aggregation: discard the maximum and minimum of the five scores and average the remaining three, yielding $e \in [0, 1]$. We binarize: $e \leq 0.5 \Rightarrow$ Low, $e > 0.5 \Rightarrow$ High.

Pairwise sample construction. For pairwise experiments, consecutive windows with engagement change $|\Delta e| > 0.05$ are labeled Increase or Decrease; near-stable pairs are excluded.

Game subcorpora. We evaluate across 9 games spanning diverse visual styles and pacing (Table 2), with DOOM as the source domain for cross-game few-shot demonstrations.

3.3 Evaluation Protocol

Cross-game few-shot setting. Few-shot demonstrations are drawn from DOOM gameplay, while predictions are made on the 9 target games listed in Table 2—testing cross-game generalization across visually distinct titles. No model fine-tuning is performed; all VLMs are evaluated in their pretrained state.

VLMs tested. We evaluate three VLMs without fine-tuning: **InternVL3.5-8B-Instruct** (Wang

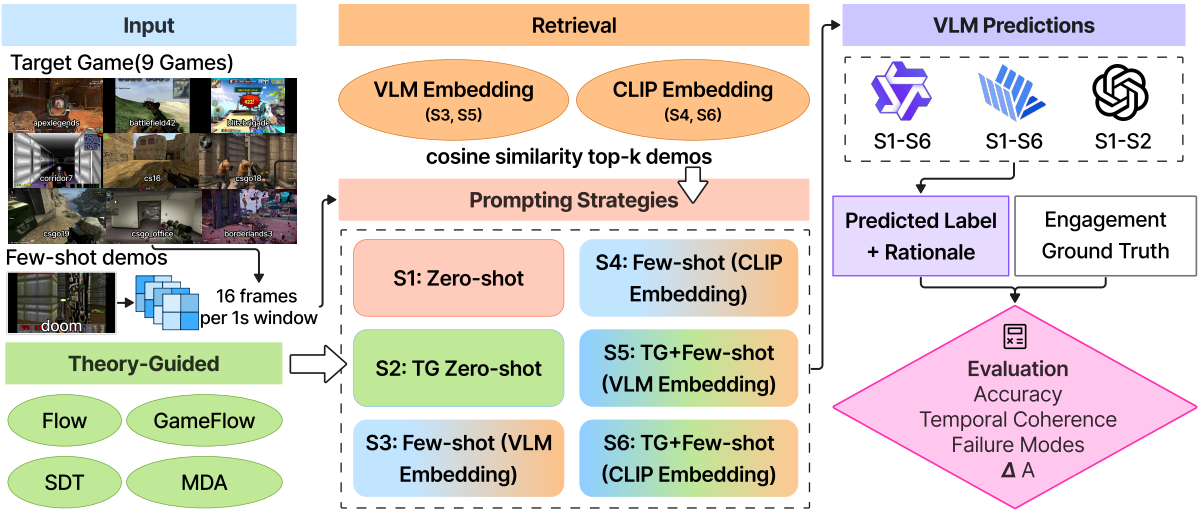

Figure 1: Overview of our experimental pipeline. The GameVibe Few-Shot (GVFS) dataset provides 9 target games, each containing 61 seconds of gameplay sampled at 16 frames per second (59 evaluation windows after reaction-lag shifting). We apply six prompting strategies of increasing complexity (S1–S6). Three VLMs—InternVL3.5-8B-Instruct, Qwen3-VL-8B-Intruct, and GPT-4o—predict binary engagement labels (High/Low). Predictions are evaluated against human annotations using accuracy, ΔA (Eq. 1), temporal coherence, and failure-mode analysis.

Game	Windows	High (%)	Low (%)
Borderlands 3	59	81.4	18.6
CS:GO Office	59	33.9	66.1
Blitz Brigade	59	64.4	35.6
Corridor 7	59	39.0	61.0
Battlefield 42	59	39.0	61.0
Apex Legends	59	55.9	44.1
CSGO19	59	59.3	40.7
CSGO18	59	22.0	78.0
CS 1.6	59	78.0	22.0
DOOM (source)	59	45.8	54.2

Table 2: Label distribution across the 9 evaluation games and the source domain. Each game contributes 61 seconds of gameplay (59 evaluation windows after reaction-lag shifting, 944 frames at 16 fps). All games exhibit varying degrees of class imbalance.

et al., 2025) and **Qwen3-VL-8B-Instruct** (Qwen Team, 2025) (open-source, all strategies S1–S6), and **GPT-4o** (OpenAI, 2024) (proprietary, S1–S2 only due to API cost). Using three architecturally distinct VLMs addresses the concern that observed failures might be model-specific.

Metrics. We report raw accuracy and accuracy relative to the per-game majority-class baseline:

$$\Delta A = \frac{A_{\text{model}} - A_{\text{baseline}}}{A_{\text{baseline}}} \quad (1)$$

where A_{model} is the VLM's accuracy and A_{baseline} is the per-game majority-class accuracy. Positive ΔA indicates performance above the trivial classifier.

ID	Strategy	Theory	Retrieval
S1	Zero-shot	×	×
S2	TG zero-shot	✓	×
S3	Few-shot (VLM)	×	VLM embedding
S4	Few-shot (CLIP)	×	CLIP embedding
S5	TG few-shot (VLM)	✓	VLM embedding
S6	TG few-shot (CLIP)	✓	CLIP embedding

Table 3: Summary of the six prompting strategies.

We additionally measure temporal coherence via flip rate and lag-1 autocorrelation (Section 5.4).

3.4 Prompting Strategies

Figure 1 illustrates the overall experimental pipeline. We design and evaluate six prompting strategies of increasing complexity, summarized in Table 3.

3.4.1 S1: Zero-Shot Prediction

The VLM receives all 16 frames from a 1-second gameplay window and a structured prompt that defines High and Low engagement in terms of observable visual cues (e.g., active combat and visible feedback vs. menus and empty environments), and asks the VLM to consider action intensity, HUD/UI elements, the presence of other entities, and signs of tension or reward (see Appendix F for the full prompt). No demonstrations or theoretical guidance are provided. This establishes the baseline for VLM's innate ability to judge engagement from short gameplay sequences.

3.4.2 S2: Theory-Guided Zero-Shot

We design a structured prompt that operationalizes the four theoretical frameworks from Section 2.3 into concrete visual indicators. Each theory's core dimensions—e.g., challenge and feedback (Flow), control and immersion (GameFlow), competence and relatedness (SDT), sensation and fellowship (MDA)—are mapped to observable cues such as health bars, score displays, teammate visibility, and visual effects (see Appendix G for the full prompt and complete dimension-to-cue mapping). The VLM assesses each dimension, counts how many theories suggest High vs. Low, and outputs a label with a theory-grounded rationale.

3.4.3 S3–S6: Retrieval-Augmented Strategies

Strategies S3–S6 augment VLM predictions with retrieved demonstration examples (Yu et al., 2024; Zhao et al., 2025a), varying along two orthogonal dimensions: *embedding space* and *theory guidance* (Table 3).

Embedding space. We compare two embedding spaces for retrieval. In S3 and S5, we extract embeddings from the VLM's own last-layer hidden states, which capture the model's internal representation of gameplay frames. In S4 and S6, we use CLIP's vision encoder (Radford et al., 2021), which provides a task-agnostic visual-semantic embedding space. Both embeddings are L2-normalized; retrieval is performed via cosine similarity, selecting the top-k most similar cases as demonstrations.

Retrieval sources. For pointwise prediction, S3 and S5 use an in-game error-driven case base: we first run the model on 70% of the windows from a target game, store incorrectly predicted cases in memory, and retrieve from this memory during inference on the remaining 30%. This design is intentional and targets the model's specific weaknesses within a game; it should therefore be interpreted as a memory-based corrective adaptation setting rather than a strict held-out generalization benchmark. For pairwise prediction, demonstrations are drawn from DOOM via difference-vector retrieval: we compute embedding differences between consecutive frame pairs and retrieve DOOM pairs with similar difference vectors, enabling cross-game transfer of engagement change patterns. Pointwise and pairwise retrieval use different sources by design and should not be directly compared.

Theory-guided variants. In S5 and S6, retrieved demonstrations are incorporated into the structured theory prompt (Section 3.4.2), combining concrete examples with theoretical scaffolding.

4 Results: Do VLMs Understand Engagement?

We organize results around three research questions. Tables 4 and 5 present the per-game accuracy breakdown and overall averages for each VLM across all prompting strategies.

4.1 Q1&Q2: Can VLMs Predict Engagement, and Does Prompting Help?

We evaluate all six prompting strategies across 9 games. Note that pointwise and pairwise settings use different retrieval sources (Section 3.4.3) and should be interpreted as complementary evaluations.

Overall accuracy is well below the majority-class baseline. Averaged across 9 games, zero-shot prediction achieves ∼57% for all three models—well below the 67.2% majority-class baseline. GPT-4o achieves virtually identical zero-shot accuracy to the open-source models, suggesting that model scale alone does not resolve this difficulty. Only a handful of game–strategy combinations exceed the per-game majority baseline, and performance varies dramatically: from 83.1% on Borderlands 3 (vs. 81.4% majority) to 30.5% on CS 1.6 (vs. 78.0% majority), echoing the game-dependence observed by Melhart et al. (2025).

Theory-guided prompting provides mixed but often positive effects. Theory-guided zero-shot prompting (S2) alone does not consistently improve over plain zero-shot prompting (S1). However, when combined with few-shot demonstrations (S5), theory-guided prompting often yields additional improvements over few-shot alone. For example, S5 achieves 68.3% accuracy for InternVL and 74.1% for Qwen, suggesting that theory guidance is most effective when paired with concrete examples.

Embedding-based retrieval shows model-dependent effects. Retrieval-augmented prompting produces mixed results across models. For InternVL, retrieval-based strategies provide limited improvement over few-shot prompting. However, Qwen3-VL benefits substantially from CLIP-based retrieval, achieving 67.8% with S4 and

		InternVL						Qwen						GPT-4o	
Game	M	S1	S2	S3	S4	S5	S6	S1	S2	S3	S4	S5	S6	S1	S2
Borderlands 3	81.4	79.7	81.4	**83.4**	79.7	53.5	81.4	83.1	81.4	75.4	49.2	**84.2**	54.2	**83.1**	81.4
CS:GO Office	66.1	61.0	42.4	58.5	55.9	**71.8**	33.9	57.6	44.1	71.1	**86.4**	70.8	78.0	**52.5**	50.8
Blitz Brigade	64.4	66.1	64.4	68.7	**69.5**	62.7	64.4	62.7	64.4	**78.9**	66.1	78.2	66.1	62.7	**64.4**
Corridor 7	61.0	49.2	37.3	**83.5**	64.4	72.9	39.0	66.1	39.0	72.5	55.9	**73.9**	52.5	**57.6**	**57.6**
Battlefield 42	61.0	64.4	52.5	46.8	39.0	**64.8**	39.0	62.7	57.6	**80.3**	61.0	68.3	72.9	**64.4**	45.8
Apex Legends	55.9	45.8	54.2	**64.1**	49.2	54.6	47.5	49.2	55.9	68.3	**69.5**	62.7	62.7	**52.5**	**52.5**
CSGO19	59.3	71.2	59.3	76.8	50.8	**82.0**	64.4	66.1	54.2	**84.2**	55.9	83.8	71.2	55.9	**74.6**
CSGO18	78.0	44.1	25.4	**81.7**	49.2	76.8	23.7	30.5	25.4	66.5	86.4	71.1	**89.8**	22.8	**24.6**
CS 1.6	78.0	30.5	59.3	**78.2**	54.2	75.4	69.5	32.2	55.9	78.2	**79.7**	73.6	**79.7**	**57.6**	**57.6**
Average	67.2	56.9	52.9	71.3	56.9	68.3	51.4	56.7	53.1	75.0	67.8	74.1	69.7	56.6	56.6

Table 4: Per-game pointwise engagement prediction accuracy (%). M = per-game majority-class baseline. Boldface indicates the best result per game within each model.

		InternVL						Qwen						GPT-4o	
Game	M	S1	S2	S3	S4	S5	S6	S1	S2	S3	S4	S5	S6	S1	S2
Borderlands 3	50.1	**54.6**	51.5	53.0	50.0	50.0	25.0	57.6	65.2	53.0	62.1	**87.5**	**87.5**	54.6	**56.1**
CS:GO Office	58.7	42.7	49.3	42.7	**52.0**	23.1	30.8	37.3	45.3	46.7	42.7	69.2	**76.9**	41.3	**45.3**
Blitz Brigade	57.6	56.1	47.0	65.1	57.6	**71.4**	64.3	**68.2**	62.1	51.5	51.5	57.1	42.9	**65.2**	60.6
Corridor 7	64.9	51.4	40.5	51.4	67.6	**70.0**	**70.0**	**73.0**	51.4	35.1	64.9	50.0	50.0	**59.5**	56.8
Battlefield 42	60.7	**57.4**	55.7	55.7	49.2	47.4	47.4	62.3	64.0	54.1	**67.2**	52.6	52.6	54.1	**57.4**
Apex Legends	58.8	67.7	57.4	55.9	61.8	**71.4**	57.1	58.8	**60.3**	57.4	57.4	57.1	50.0	57.4	**66.2**
CSGO19	52.2	43.5	52.2	50.0	41.3	**70.0**	**70.0**	45.7	45.7	43.5	41.3	**50.0**	40.0	43.5	**52.2**
CSGO18	50.7	52.2	50.7	46.4	**53.6**	33.3	16.7	40.6	39.1	47.8	39.1	**50.0**	**50.0**	49.3	**52.2**
CS 1.6	55.1	52.6	43.6	48.7	44.9	**66.7**	**66.7**	48.7	52.6	39.7	42.3	**55.6**	44.4	42.3	**44.9**
Average	56.5	53.1	49.8	51.8	53.1	55.9	49.8	54.7	54.0	47.7	52.1	58.8	54.9	51.9	54.6

Table 5: Per-game pairwise engagement prediction accuracy (%). M = per-game majority-class baseline. Boldface indicates the best result per game within each model.

69.7% with S6, exceeding the majority-class baseline on average. Post-hoc analysis confirms that scene-level visual similarity to the DOOM source domain strongly predicts retrieval-augmented accuracy (Appendix C).

4.2 Q2 (cont.): Qualitative Analysis of Theory-Aligned Rationales

Adding theoretical structure does not consistently improve prediction: rationales tend to repeat surface observations (e.g., "teammates visible ⇒ relatedness ⇒ High") without genuine multi-dimensional reasoning (Section 5.2), suggesting information overload (Chochlakis et al., 2024). However, when theory prompts are combined with few-shot examples (S5, S6), improvements emerge—indicating that theoretical cues provide useful structure only when grounded by concrete demonstrations.

4.3 Q3: Are Predictions Consistent Across Games and VLMs?

Both models consistently over-predict High engagement, but severity varies across games (Table 6): on CSGO18, 90–97% of errors are false positives, while CS:GO Office shows 60–74%. Cross-VLM agreement is substantial overall ($\kappa = 0.67$) but high agreement does not imply correctness—on CSGO18, models agree on 86.4% of frames yet both achieve below-majority accuracy, converging on the same wrong answer. This suggests both architectures rely on similar visual shortcuts. We elaborate on specific failure modes in Section 5.

Game	**Model**	**FN**	**FP**	**Total Err.**
CSGO18	InternVL3	1	32	33
	Qwen3-VL	4	37	41
CS:GO Office	InternVL3	6	17	23
	Qwen3-VL	10	15	25

Table 6: Zero-shot (S1) error distribution on two CS:GO variants.

5 Analysis & Failure Modes

We conduct in-depth analysis on two CS:GO variants (CSGO18 and CS:GO Office), which exhibit contrasting error severities (Table 6), identifying five systematic failure modes.

5.1 Failure Mode 1: Visual Intensity Bias

VLMs systematically conflate visual intensity with engagement. We quantify visual intensity as a composite of color saturation, edge density, brightness variance, and red channel intensity (Appendix B). On CSGO18, VLM predictions correlate significantly with this score ($r = 0.432$, $p < 0.001$), while human engagement labels show no correlation ($r = -0.193$, $p = 0.143$). The VLM performs visual intensity classification rather than engagement prediction: it recognizes "this frame looks intense" but cannot reason that a 15–14 scoreboard implies competitive tension. This bias is most pronounced when visual variation across frames is large; on CS:GO Office, neither VLM predictions nor human labels correlate with visual intensity—consistent with findings that VLMs perform best on action-heavy games (Melhart et al., 2025) and systematically underutilize visual representations for reasoning (Fu et al., 2025).

5.2 Failure Mode 2: Surface Feature Shortcuts

VLMs rely on surface-level visual shortcuts to predict engagement, bypassing genuine reasoning:

Relatedness bias. When teammates are visible or radio chatter indicators appear, the model automatically predicts High engagement. The VLM equates the presence of social cues with high engagement, without considering whether team interaction is actually meaningful in the current context.

Sensation bias. Blood splatter, explosions, and weapon effects trigger High predictions regardless of gameplay context. The model uses visual "spectacle" as a proxy for engagement.

These shortcuts map directly to theory dimensions (SDT's relatedness, MDA's sensation) but in a shallow, stimulus-response manner. The theory-guided prompt amplifies these biases by providing vocabulary for rationalizing surface-level pattern matching, mirroring the sentiment bias in VLM-based emotion recognition (Bhattacharyya and Wang, 2025).

Annotating all 78 errors from S2 on CSGO18 and CS:GO Office reveals that Challenge Misassessment (38.5%) and Relatedness Shortcuts (30.8%) together account for over two-thirds of all errors (Appendix A). A human evaluation of 30 sampled errors found that 40% stem from clear reasoning flaws, while 60% reflect defensible interpretations—highlighting the inherent difficulty of binary engagement classification.

Metric	Dataset	VLM Pred.	Human Labels
Flip Rate ↓	CSGO18	0.310	0.017
	CS:GO Office	0.345	0.190
Autocorr. (lag-1) ↑	CSGO18	0.275	0.950
	CS:GO Office	0.168	0.576

Table 7: Temporal coherence metrics. VLM predictions are dramatically less coherent than human labels across consecutive 1-second windows.

5.3 Failure Mode 3: Post-Match Context Blindness

VLMs cannot distinguish active gameplay from post-match screens. Scoreboards showing close victories are high-engagement moments for human annotators, but VLMs classify them as "static, no active gameplay" and predict Low. On CS:GO Office, 18 of 20 false negatives involve this misclassification—the model equates visual stillness with disengagement, missing that a scoreboard represents the culmination of a competitive experience.

5.4 Failure Mode 4: Temporal Inconsistency

VLM predictions lack temporal coherence: the model may predict High for window t and Low for window $t+1$ even when gameplay context is unchanged. We measure this using flip rate (fraction of adjacent windows with different predictions) and lag-1 autocorrelation (Table 7). On CSGO18, the VLM's flip rate is 0.310—18× higher than the human label flip rate of 0.017—and its autocorrelation (0.275) is far below human labels (0.950). This aligns with Huang et al. (2025)'s finding that Video-LLMs achieve near chance-level temporal consistency.

5.5 Failure Mode 5: Confident but Wrong Reasoning

VLMs generate detailed, fluent, and seemingly well-reasoned rationales that are nonetheless incorrect, performing post-hoc rationalization of surface-level cues rather than genuine engagement reasoning. Strikingly, we find an inverse confidence–accuracy relationship: high-confidence predictions achieve only 38.5% accuracy, while low-confidence predictions achieve 64.3% (Appendix I), confirming that confident VLM reasoning is systematically misleading. Four orderings of few-shot examples yield nearly identical prediction

distributions (Appendix H), confirming this is an intrinsic model bias rather than a positional artifact.

5.6 Analysis: Demonstration Label Balance

Beyond failure modes, we find that few-shot predictions are highly sensitive to the label distribution of retrieved demonstrations. In our error-driven retrieval setup (Section 3.4.3), several games' case bases contain no positive examples, causing systematic over-prediction of the negative class. Adding even a small number of positive cases dramatically improves accuracy (e.g., +15.5pp on Blitz Brigade with just 17 examples; Appendix D). This implies that retrieval-augmented VLM systems for subjective tasks must ensure balanced demonstration selection.

6 Discussion

6.1 The Perception–Understanding Gap

Our failure modes collectively point to a single underlying issue: VLMs treat engagement prediction as visual classification rather than psychological inference, stopping at the first step of the required multi-step reasoning chain (Fu et al., 2025). Games are uniquely diagnostic because they decouple visual salience from the target construct. Our results suggest a capability hierarchy—perception > description > interpretation > inference—where current VLMs succeed at the first two levels but fail at the latter two.

6.2 The Role of Prompting Complexity

The interaction between prompting strategy and performance reveals three key findings. First, few-shot demonstrations provide large, reliable pointwise gains, though these do not extend to pairwise prediction—likely because cross-game DOOM demonstrations are less informative than in-game memory. Second, theory guidance alone hurts performance, likely through shortcut amplification: explicitly naming features like "teammates visible" provides vocabulary for rationalizing surface-level pattern matching rather than enabling genuine reasoning (Section 5.2). Third, concrete examples drive improvements more than theoretical scaffolding: when combined with few-shot demonstrations, theory guidance provides at best modest additional benefit. Cross-game retrieval effectiveness further depends on scene-level visual similarity rather than game-level semantic similarity (Appendix C), and few-shot predictions are sensitive to demonstration label balance (Appendix D).

6.3 Implications

Our findings argue for VLM benchmarks that test affective understanding beyond perception, where game engagement offers a particularly diagnostic testbed. For game analytics, VLMs are not yet suitable without task-specific fine-tuning via LoRA (Hu et al., 2022) or game-specific retrieval (Lewis et al., 2020). For prompt engineering, our results caution that more structured prompts may amplify biases rather than compensate for them when the model lacks genuine understanding.

Recommendations. Future work should (1) decouple perception from inference in evaluation; (2) prioritize concrete examples over elaborate prompts for subjective tasks; (3) incorporate temporal context via memory mechanisms; and (4) calibrate VLM confidence (Section 5.5). We outline concrete architectural directions in Appendix E.

7 Conclusion

We presented the first systematic evaluation of VLMs on human engagement prediction in gameplay videos, testing six prompting strategies across 9 games with three VLMs. Our results show that raw zero-shot predictions are unreliable (∼57%, below the majority-class baseline), but few-shot demonstrations yield substantial pointwise gains (up to 75.0% for Qwen), while theory-guided prompting alone can amplify surface-level shortcuts rather than improve reasoning. Our failure-mode analysis reveals five systematic limitations: VLMs conflate visual intensity with engagement, rely on surface-feature shortcuts, cannot interpret post-match contexts, lack temporal coherence, and generate confident but misleading rationales—collectively exposing a fundamental perception–understanding gap that standard benchmarks would not detect. Games, where visual salience is decoupled from the target construct, provide a uniquely diagnostic testbed for this gap. Notably, pairwise prediction remains near chance level across all strategies, suggesting that using VLMs as automated preference annotators for subjective tasks may be unreliable without substantial task-specific adaptation. We outline concrete architectural directions for bridging this gap in Appendix E.

Limitations

Our evaluation is limited by the GVFS dataset's scale (61 seconds per game, 59 windows each) and genre diversity (FPS only), as well as a small annotator pool (5 students). Binarization of continuous engagement scores may lose meaningful variation; finer-grained approaches could reveal different capabilities. We evaluate two open-source VLMs with all six strategies but only test GPT-4o on S1–S2 due to cost constraints; broader model coverage could provide further insights.

Ethics Statement

This study uses the publicly available GameVibe Few-Shot dataset, which contains gameplay video clips annotated for engagement. No personally identifiable information is involved; all annotations are aggregated and anonymized. Our work evaluates existing pretrained VLMs without fine-tuning and does not introduce new models or training data. The engagement labels reflect subjective judgments from a small annotator pool, and we acknowledge that binarization may oversimplify the nuanced nature of engagement.

References

Mostafa M. Amin, Rui Mao, Erik Cambria, and Björn W. Schuller. 2024. A wide evaluation of ChatGPT on affective computing tasks. *IEEE Transactions on Affective Computing*, 15(4):2204–2218.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Matthew Barthet, Maria Kaselimi, Kosmas Pinitas, Konstantinos Makantasis, Antonios Liapis, and Georgios N. Yannakakis. 2024. Gamevibe: a multimodal affective game corpus. *Scientific Data*, 11:1306.

Sree Bhattacharyya and James Z Wang. 2025. Evaluating vision-language models for emotion recognition. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1798–1820.

Elizabeth A. Boyle, Thomas M. Connolly, Thomas Hainey, and James M. Boyle. 2012. Engagement in digital entertainment games: A systematic review. *Computers in Human Behavior*, 28(3):771–780.

Jeanne H. Brockmyer, Christine M. Fox, Kathleen A. Curtiss, Evan McBroom, Kimberly M. Burkhart, and Jacquelyn N. Pidruzny. 2009. The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology*, 45(4):624–634.

Georgios Chochlakis, Alexandros Potamianos, Kristina Lerman, and Shrikanth Narayanan. 2024. The strong pull of prior knowledge in large language models and its impact on emotion recognition. In *2024 12th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do LLMs understand social knowledge? Evaluating the sociability of large language models with SocKET benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11370–11403. Association for Computational Linguistics.

Mihaly Csikszentmihalyi. 1990. *Flow: The Psychology of Optimal Experience*. Harper & Row, New York.

Yasaman Etesam, Özge Nilay Yalçın, Chuxuan Zhang, and Angelica Lim. 2024. Contextual emotion recognition using large vision language models. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4769–4776. IEEE.

Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. 2025. Hidden in plain sight: Vlms overlook their visual representations. *Preprint*, arXiv:2506.08008.

Gemini Team, Google. 2023. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1321–1330.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Lanxiang Hu, Mingjia Huo, Yuxuan Zhang, Haoyang Yu, Eric P. Xing, Ion Stoica, Tajana Rosing, Haojian Jin, and Hao Zhang. 2025. lmgame-bench: How good are llms at playing games? *Preprint*, arXiv:2505.15146.

Shaoguang Huang, Shuai Li, and 1 others. 2025. On the consistency of video large language models in temporal comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Robin Hunicke, Marc LeBlanc, and Robert Zubek. 2004. MDA: A formal approach to game design and game research. In *Proceedings of the AAAI Workshop on Challenges in Game AI*, volume 4, page 1722.

Joshua Juvrud, Gabriel Ansgariusson, Patrik Selleby, and Magnus Johansson. 2022. Game or watch: The effect of interactivity on arousal and engagement in video game media. *IEEE Transactions on Games*, 14(2):308–317.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What's "up" with vision-language models? Investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9161–9175. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. volume 33, pages 9459–9474.

Zheng Lian, Licai Sun, Haiyang Sun, Kang Chen, Zhuofan Wen, Hao Gu, Bin Liu, and Jianhua Tao. 2024. GPT-4V with emotion: A zero-shot benchmark for generalized emotion recognition. *Information Fusion*, 108:102367.

Zhiwei Liu, Lingfei Qian, Qianqian Xie, Jimin Huang, Kailai Yang, and Sophia Ananiadou. 2025. Mmaffben: A multilingual and multimodal affective analysis benchmark for evaluating llms and vlms. *Preprint*, arXiv:2505.24423.

Phil Lopes, Georgios N. Yannakakis, and Antonios Liapis. 2017. RankTrace: Relative and unbounded affect annotation. In *Proceedings of the 7th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 158–163. IEEE.

Hao Lu, Xuesong Niu, Jiyao Wang, Yin Wang, Qingyong Hu, Jiaqi Liang, Weihao Lin, Jiahao Zhong, Ruichen Liu, Yutong Huang, and 1 others. 2024. GPT as psychologist? preliminary evaluations for GPT-4V on visual affective computing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Rui Mao, Qian Liu, Kai He, Wei Li, and Erik Cambria. 2023. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE Transactions on Affective Computing*, 14(3):1743–1753.

David Melhart, Matthew Barthet, and Georgios N. Yannakakis. 2025. Can large language models capture video game engagement? *Preprint*, arXiv:2502.04379.

David Melhart, Daniele Gravina, and Georgios N. Yannakakis. 2020. Moment-to-moment engagement prediction through the eyes of the observer: PUBG streaming on twitch. In *Proceedings of the 15th International Conference on the Foundations of Digital Games (FDG '20)*, pages 1–10. ACM.

Dávid Melhart, Antonios Liapis, and Georgios N. Yannakakis. 2019. PAGAN: Video affect annotation made easy. In *Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 130–136. IEEE.

Heather L. O'Brien and Elaine G. Toms. 2008. What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, 59(6):938–955.

OpenAI. 2023. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2024. GPT-4o system card.

Davide Paglieri, Ahmad Beirami, Karl Cobbe, Soumith Chintala, Alessandro Lazaric, Junhyuk Oh, and Zhixuan Eric Guo. 2024. BALROG: Benchmarking agentic LLM and VLM reasoning on games. *arXiv preprint arXiv:2411.13543*.

Kosmas Pinitas, Konstantinos Makantasis, and Georgios N Yannakakis. 2024. Across-game engagement modelling via few-shot learning. In *European Conference on Computer Vision*, pages 390–406. Springer.

Kosmas Pinitas, David Renaudie, Mike Thomsen, Matthew Barthet, Konstantinos Makantasis, Antonios Liapis, and Georgios N. Yannakakis. 2023. Predicting player engagement in Tom Clancy's The Division 2: A multimodal approach via pixels and gamepad actions. In *Proceedings of the 25th International Conference on Multimodal Interaction (ICMI '23)*, pages 596–604. ACM.

John Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR.

Richard M. Ryan and Edward L. Deci. 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1):68–78.

Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, volume 27, pages 568–576. Curran Associates, Inc.

Penelope Sweetser and Peta Wyeth. 2005. GameFlow: A model for evaluating player enjoyment in games. *Computers in Entertainment*, 3(3):3.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Qu Yang, Mang Ye, and Bo Du. 2024. EmoLLM: Multimodal emotional understanding meets large language models. *arXiv preprint arXiv:2406.16442*.

Georgios N Yannakakis and David Melhart. 2023. Affective game computing: A survey. *Proceedings of the IEEE*, 111(10):1423–1444.

Shi Yu, Chaoyue Tang, and 1 others. 2024. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*.

Alex L. Zhang, Thomas L. Griffiths, Karthik R. Narasimhan, and Ofir Press. 2025. Videogamebench: Can vision-language models complete popular video games? *Preprint*, arXiv:2505.18134.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906. Association for Computational Linguistics.

Haochen Zhao and 1 others. 2025a. A survey of multimodal retrieval-augmented generation. *arXiv preprint arXiv:2504.08748*.

Jiaxing Zhao, Qize Yang, Yixing Peng, Detao Bai, Shimin Yao, Boyuan Sun, Xiang Chen, Shenghao Fu, Weixuan Chen, Xihan Wei, and Liefeng Bo. 2025b. HumanOmni: A large vision-speech language model for human-centric video understanding. *arXiv preprint arXiv:2501.15111*.

A Error Taxonomy and Human Evaluation

Three annotators independently classified all 78 prediction errors from S2 (InternVL3.5-8B-Instruct) on CSGO18 and CS:GO Office using a two-stage procedure: (1) keyword-based candidate identification from VLM rationales, and (2) manual review and correction. Inter-annotator agreement was moderate (Fleiss' $\kappa = 0.512$); disagreements were resolved by majority vote. Table 8 shows the resulting taxonomy.

Failure Type	CSGO18	CS:GO Office	Total	%
A: Static Scene Misclass.	3	4	7	9.0
B: Relatedness Shortcut	19	5	24	30.8
C: Sensation Shortcut	0	10	10	12.8
D: Challenge Misassess.	16	14	30	38.5
E: Other/Ambiguous	6	1	7	9.0
Total Errors	44	34	78	

Table 8: Distribution of failure types. Challenge Misassessment and Relatedness Shortcuts account for over two-thirds of all errors.

Human evaluation of error severity. Three annotators judged a stratified sample of 30 errors as Model Wrong (MW), Ground Truth Ambiguous (GA), or Both Plausible (BP). Inter-annotator agreement: Fleiss' $\kappa = 0.679$. Of 30 cases, 12 (40%) were MW and 18 (60%) were BP, with no GA cases.

B Visual Intensity Score

We quantify visual intensity for each frame as a composite score:

$$I(f) = \frac{1}{4}\left(\hat{S}(f) + \hat{E}(f) + \hat{V}(f) + \hat{R}(f)\right) \quad (2)$$

where $\hat{S}$, $\hat{E}$, $\hat{V}$, and $\hat{R}$ are the min-max normalized values of color saturation, Canny edge density, brightness variance, and red channel intensity, respectively. On CSGO18, the mean visual intensity for frames predicted as High (0.220) is significantly higher than for Low (0.177), whereas ground-truth High (0.189) and Low (0.211) frames show no such separation (Spearman $\rho = 0.490$, $p < 0.001$ for VLM predictions; $\rho = -0.193$, $p = 0.143$ for human labels).

C Retrieval Similarity Analysis

The effectiveness of retrieval-augmented strategies depends on visual similarity between the source (DOOM) and target game. We compare two CS:GO variants (Table 9): CS:GO Office (indoor, DOOM-like) achieves 63.9% retrieval-augmented accuracy with mean CLIP similarity of 0.657, while CS:GO 2018 (outdoor) achieves only 37.7% with similarity 0.628. This 4.6% gap is highly significant ($t(120) = 5.09$, $p < 10^{-5}$; Cohen's $d = 0.93$) and corresponds to a 26.2pp accuracy difference ($\chi^2 = 7.38$, $p = 0.007$). Per-frame correlation between similarity and correctness is weak (CS:GO Office: $r = 0.14$; CS:GO 2018: $r = 0.01$), suggesting the effect operates at the game level.

Target Game	Avg Sim.	SD	Acc. (%)	r
CS:GO Office	0.657	0.030	63.9	0.14
CS:GO 2018	0.628	0.033	37.7	0.01
Similarity: $t(120) = 5.09$		$p < 10^{-5}$, Cohen's $d = 0.93$		
Accuracy: $\chi^2 = 7.38$		$p = 0.007$		

Table 9: CLIP retrieval similarity and accuracy for two CS:GO variants.

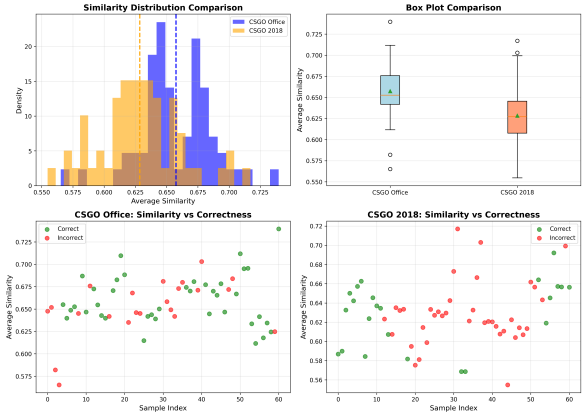

Figure 2: CLIP retrieval similarity analysis for two CS:GO variants. **Top**: Similarity distributions and box plots. **Bottom**: Per-frame similarity vs. prediction correctness.

D Demonstration Label Balance

Few-shot VLM predictions are highly sensitive to the label distribution of demonstration examples. In our few-shot setup, we first run the model on 70% of the frames from a game and store incorrectly predicted cases in an error-driven memory. During inference on the remaining 30%, the model retrieves examples from this memory as few-shot demonstrations. Because the memory is constructed only from the earlier split, no evaluation samples are included in the demonstration set.

For several games, the original memory contained no positive examples (label = 1), resulting in a highly imbalanced demonstration set where the model observes only negative examples during in-context learning, causing it to over-predict the negative class. To investigate this effect, we introduced a small number of positive examples into the memory. Adding only 17 positive cases—corresponding to the average number of positive examples available in other games' memories—consistently improved accuracy (e.g., Blitz Brigade: 62.7% $\rightarrow$ 78.2%; Borderlands 3: 51.4% $\rightarrow$ 66.5%), and increasing to 50 positive examples raised accuracy to 84.2% (Figure 3).

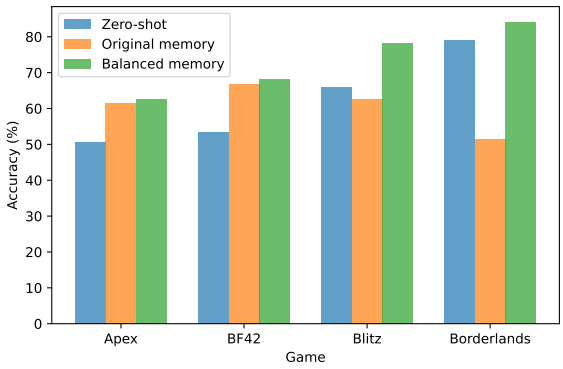

Figure 3: Effect of demonstration label distribution on few-shot prediction accuracy. When the original memory contains no positive examples, the model over-predicts the negative class. Introducing positive examples into the demonstration memory consistently improves accuracy across multiple games.

E Toward Practical Engagement Prediction Systems

Our evaluation identifies clear limitations of raw VLMs for engagement prediction, but also reveals specific capabilities that can be leveraged through architectural augmentation. Below we outline five concrete directions, each grounded in specific failure modes identified in Section 5.

1. Hybrid pipeline: VLM features + supervised classifier. Use the VLM as a feature extractor—prompting it to output structured descriptions of observable game state (score, health, number of visible players, action level, UI state)—and then train a lightweight supervised classifier on these extracted features using the available engagement annotations.

2. Temporal smoothing of per-window predictions. A post-processing temporal model—such as a hidden Markov model, exponential moving average, or a lightweight LSTM operating on the sequence of per-window VLM predictions and extracted features—could enforce temporal smoothness. Given the strong autocorrelation in human engagement labels ($r_1 = 0.576$–0.950), even simple smoothing could substantially reduce flip rate.

3. Decomposed reasoning via chain-of-thought. Decompose the task into a pipeline: (i) identify game state (active gameplay, menu, scoreboard, buy phase); (ii) assess challenge level; (iii) evaluate feedback; (iv) predict engagement conditioned on (i)–(iii).

4. Game-specific calibration. Techniques such as Platt scaling (Platt, 1999) or temperature scaling (Guo et al., 2017), applied to VLM prediction logits using as few as 20–30 labeled windows per game, could correct for game-specific prediction biases.

5. Structured output with verification. Prompt the VLM to extract a structured game-state representation (game phase, score differential, health, visible players, action level), which can then be processed by deterministic engagement logic.

F Zero-Shot Prompt

Below is the complete prompt used in S1 (Section 3.4).

```
Task: Predict whether these game video
    frames
indicate High or Low player engagement.

You are given 16 frames from a 1-second
    gameplay
window. Analyze the sequence as a whole.

Engagement definition:
- High engagement: The player is
    actively involved
  in intense, meaningful, or challenging
      gameplay.
  Visual signs include active combat,
      time pressure,
  dynamic action, visible feedback from
      the game,
  interaction with other players, or
      tense
  situations.
- Low engagement: The player is in a
    passive, idle,
  or low-intensity state. Visual signs
      include
  menus/pause screens, empty
      environments, lack of
  action or feedback, repetitive/
      uneventful scenes,
  or the player appearing disengaged
      from core
  gameplay.

Look at the frames carefully and
    consider:
- What is happening in the scene? (
    action intensity,
  events)
- What HUD/UI elements are visible? (
    health, score,
  timer, kill feed, objectives)
- Are there other players or entities? (
    teammates,
  enemies, NPCs)
- Does the scene suggest tension,
    challenge, or
  reward?

OUTPUT (strict format):
Label: [High/Low]
Rationale: [One sentence explaining the
    key visual
evidence for your prediction]
```

G Full Theory-Guided Prompt

Below is the complete prompt used in S2, S5, and S6 (Section 3.4.2).

```
Task: Predict whether these game video
    frames
indicate High or Low player engagement,
    using game
engagement theories as an analytical
    framework.

You are given 16 frames from a 1-second
    gameplay
window. Analyze the sequence as a whole.

ENGAGEMENT DEFINITION:
- High engagement = the player is deeply
     invested in
  gameplay -- facing meaningful
      challenge, receiving
  feedback, experiencing immersion, or
      interacting
  with others.
- Low engagement = the player is
    disengaged -- idle,
  in menus, facing no challenge,
      receiving no
  feedback, or in an uneventful
      situation.

THEORETICAL FRAMEWORK FOR VISUAL
    ANALYSIS:

1. Flow Theory (challenge-skill balance,
     goals,
   feedback):
   Look for signs of:
   - Active challenge: health not full/
       not empty
     (mid-range), score is close,
         enemies present,
     time pressure visible
   - Clear goals: objective markers,
       mission
     indicators, timers, waypoints on
         screen
   - Immediate feedback: damage numbers,
       hit
     markers, kill notifications, score
         changes,
     reward popups
   -> Low engagement signs: no visible
       objectives,
      no feedback elements, health full
          with no
      threats

2. GameFlow (control, immersion, social
    presence):
   Look for signs of:
   - Player control: crosshair/aim on
       target,
     character in motion, weapon/tool
         actively in use
```

```
    - Immersion: full gameplay view (no
        pause menus,
      no loading screens, no settings),
          rich
      environment detail
    - Social presence: other players
        visible, team
      indicators, chat activity,
          cooperative/
      competitive interaction
    -> Low engagement signs: pause/menu
        screens,
       empty environments, no other
           players or
       entities

3. Self-Determination Theory (competence
    , autonomy,
   relatedness):
   Look for signs of:
   - Competence: positive performance
       indicators
     (high kill count, achievements,
         skill displays),
     successful actions
   - Autonomy: varied equipment/loadout,
        player-
     chosen paths, customized appearance
   - Relatedness: team coordination
       visible, group
     activity, communication indicators
   -> Low engagement signs: poor
       performance shown,
      default/basic equipment, isolated
          player

4. MDA Framework (challenge, sensation,
    fellowship,
   narrative):
   Look for signs of:
   - Challenge: competitive situation,
       close match,
     difficult enemies, survival
         pressure
   - Sensation: visual effects (
       explosions,
     particles, lighting), dynamic
         camera, action-
     heavy scene
   - Fellowship: multiple players
       engaged together,
     team-based activity
   - Narrative: story moments, dramatic
       events,
     critical match situations (final
         round,
     overtime)
   -> Low engagement signs: no visual
       excitement,
      static scene, no narrative tension

DECISION GUIDELINES:
- Count how many theories suggest High
    vs Low
  engagement
- If the frame shows a menu, pause
    screen, or
  loading screen -> almost always Low
- When signals conflict (e.g., active
    scene but
  poor performance), lean toward the
      majority of
  indicators

OUTPUT (strict format):
Label: [High/Low]
Rationale: [One sentence: describe the
    key visual
features you observed and which theory
    dimensions
they satisfy or violate]
```

4-Shot Demonstration Examples. Each example includes a DOOM frame paired with a gold label and a theory-grounded rationale:

- **Example 1 (High):** Low HP clutch with C4 planted and timer under 30s ⇒ active challenge (Flow), high control (GameFlow), competence challenge (SDT), narrative tension (MDA).
- **Example 2 (Low):** Full HP, empty kill feed, no objectives ⇒ no feedback (Flow), low immersion (GameFlow), no competence indicators (SDT), no aesthetic engagement (MDA).
- **Example 3 (High):** Active kill feed with headshot, 4 teammates visible, MVP display ⇒ positive feedback (Flow), social immersion (GameFlow), competence satisfaction (SDT), sensation + fellowship (MDA).
- **Example 4 (Low):** Pause menu visible ⇒ disrupted flow (Flow), broken immersion (GameFlow), no competence demonstration (SDT), eliminated challenge/narrative (MDA).

H Demonstration Order Experiment

We test whether the ordering of 4-shot examples affects predictions (Table 10). All orderings produce nearly identical distributions, confirming that the model's over-prediction of High is an intrinsic bias, not a positional artifact.

Ordering	Pred. High	Pred. Low
Original [L, H, H, L]	12	8
Last-high [L, L, H, H]	12	8
Alternating H-L	10	10
Alternating L-H	12	8

Table 10: Prediction distribution under different 4-shot example orderings (CS:GO Office subset). Distributions are stable across orderings.

A chi-square test of independence confirms no significant association between example ordering and prediction distribution ($\chi^2 = 0.614$, $df = 3$, $p = 0.893$), indicating that the over-prediction of High is an intrinsic model bias rather than a positional artifact of the demonstration examples.

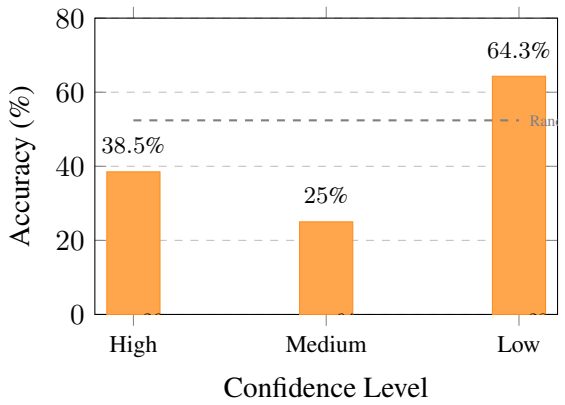

Figure 4: Prediction accuracy by VLM confidence level, showing severe miscalibration. High-confidence predictions are the least accurate (38.5%), while low-confidence predictions are the most accurate (64.3%). The dashed line shows random baseline (52.4%).

I Confidence–Accuracy Analysis

We estimate model confidence from the generated rationale text:

$$C(r) = \frac{|\{w \in r : w \in \mathcal{A}\}|}{|\{w \in r : w \in \mathcal{A} \cup \mathcal{H}\}|} \quad (3)$$

where $\mathcal{A}$ is a set of assertive words ("clearly," "demonstrates," "confirms") and $\mathcal{H}$ is a set of hedging words ("possibly," "minimal," "lacking"). The mean confidence for incorrect predictions (0.597) is higher than for correct predictions (0.453).

J Qualitative Error Examples

The following examples are drawn from real VLM errors in our theory-guided zero-shot (S2) evaluation on CSGO18 and CS:GO Office, organized by failure mode. Each case shows the VLM's actual rationale and the corresponding ground-truth annotation.

Figure 5: CS:GO Office, frame 24–25. The VLM predicts Low due to the static vote screen, but the ground truth is High.

Example 1: Static scene misclassification (Type A). *Ground truth:* High. *VLM prediction:* Low.
VLM rationale: "The scene shows a player in a conference room with no immediate action or visible enemies, and a vote is occurring, indicating a pause in gameplay. Feedback elements like score changes are present, but there's no active challenge or player control evident, suggesting low engagement."
Analysis: The VLM treats the vote screen as a "pause," failing to infer that an ongoing team vote during a close match represents a moment of high strategic engagement. This exemplifies FM3 (Post-Match Context Blindness, Section 5.3): the model equates visual stillness with low engagement.

Figure 6: CS:GO Office, frame 31–32. Blood splatters and weapon animations trigger a High prediction, but the ground truth is Low.

Example 2: Sensation shortcut (Type C). *Ground truth:* Low. *VLM prediction:* High.
VLM rationale: "The frame shows active combat with visible feedback (blood splatter, weapon in use), clear objectives (mini-map, score), and player control (aiming), indicating engagement across multiple theories."
Analysis: Blood splatters and weapon animations trigger a High prediction despite the ground truth being Low. The VLM conflates visual intensity with engagement, illustrating FM1 (Visual Intensity Bias, Section 5.1).

Example 3: Relatedness shortcut (Type B). *Ground truth:* Low. *VLM prediction:* High.
VLM rationale: "The frame shows active combat with multiple players, visible health and score indicators, and immediate feedback from teammate actions, satisfying challenge, feedback, and social presence dimensions of engagement theories."
Analysis: The mere visibility of teammates and HUD indicators (health bars, scores) triggers a High prediction. The VLM uses "social presence"

Figure 7: CSGO18, frame 15–16. Visible teammates and HUD indicators lead to a High prediction, but the ground truth is Low.

as a shortcut for engagement without evaluating whether the moment is genuinely engaging, exemplifying FM2 (Surface Feature Shortcuts, Section 5.2).

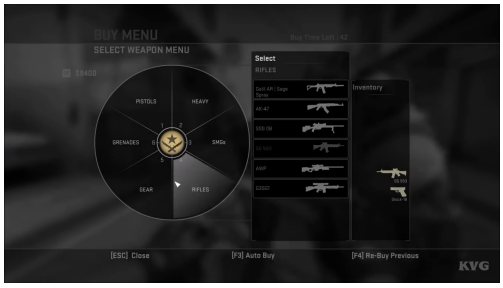

Figure 8: CSGO18, frame 12–13. The buy menu is dismissed as "no active gameplay," but the ground truth is High.

Example 4: Challenge misassessment (Type D). *Ground truth:* High. *VLM prediction:* Low. *VLM rationale:* "The frame shows a buy menu with no active gameplay, indicating a pause in action. This suggests a lack of immediate challenge, goals, and feedback, aligning with low engagement according to Flow Theory and GameFlow." *Analysis:* The buy menu is dismissed as "no active gameplay," yet this phase—selecting equipment under time pressure between rounds—is annotated as High engagement. The VLM fails to recognize that strategic decision-making constitutes engagement, instead requiring visible combat as evidence of challenge.